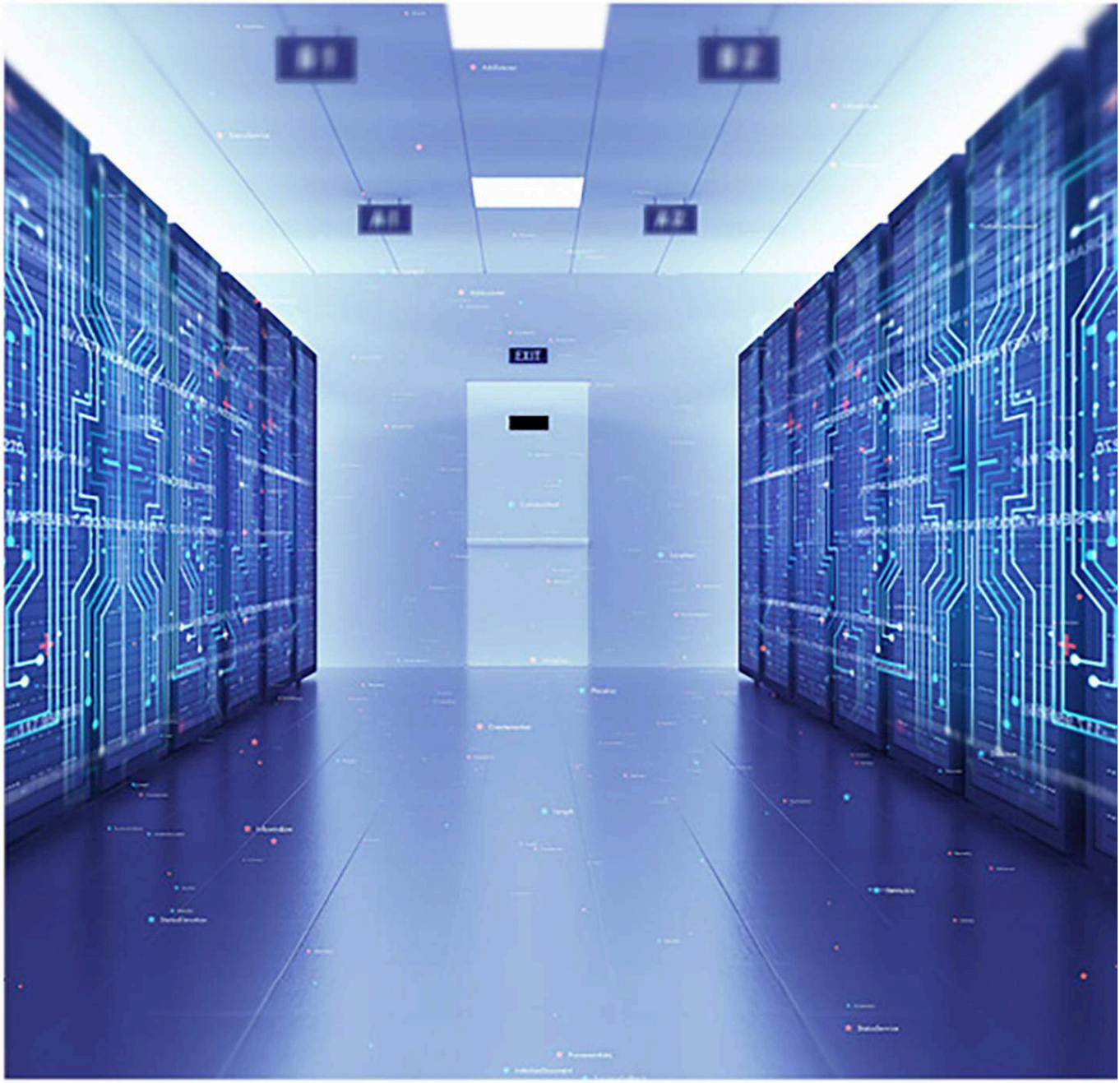
Les nouveaux paradigmes de l'archive

Claire Scopsi, Clothilde Roullier, Martine Sin Blima-Barru, Édouard Vasseur, Olivier Poncet, Ghislaine Chartron, Jean Carrive, Abdelkrim Beloued, Pascale Goetschel, Serge Heiden, et al.

MINISTÈRE
DE LA CULTURE
Liberté
Égalité
Fraternité

ARCHIVES
NATIONALES

Les nouveaux paradigmes de l'archive



haStec

le cnam
Dicen

Publications des Archives nationales

Les nouveaux paradigmes de l'archive

Claire Scopsi, Clothilde Roullier, Martine Sin Blima-Barru et Édouard Vasseur (dir.)

DOI : 10.4000/books.pan.5802
Éditeur : Publications des Archives nationales
Lieu d'édition : Pierrefitte-sur-Seine
Année d'édition : 2024
Date de mise en ligne : 9 février 2024
Collection : Actes
EAN électronique : 978-2-86000-390-2

RÉSUMÉS

De 2019 à 2023, des représentants des Archives nationales, du laboratoire Dicen-IDF du Conservatoire national des arts et métiers et du Centre Jean-Mabillon de l'École nationale des chartes se sont réunis au sein du séminaire « Les nouveaux paradigmes de l'archive », mené avec le soutien du Laboratoire d'excellence HASTEC. Ils y ont accueilli plus de quarante intervenants, venus partager leur vision des mutations amorcées par le numérique dans les pratiques, méthodes et outils du traitement des archives et du patrimoine.

Cet ouvrage prolonge ces échanges et propose des témoignages des évolutions en cours autour de quatre problématiques de conservation : les données volumineuses, l'audiovisuel et ses métadonnées, les données de la recherche et les contenus du web et des réseaux sociaux numériques. Il aborde tour à tour les changements sociétaux, les apports des intelligences artificielles, les architectures des systèmes d'information, les nouvelles compétences et missions des professionnels et leur confrontation avec des contenus de plus en plus volumineux, hétérogènes et fluides. Par son approche volontairement pluridisciplinaire, il offre aux praticiens des archives, aux chercheurs ou aux étudiants auxquels il est destiné d'entrer dans les nouveaux paradigmes de l'archive par la porte qui leur convient le mieux.

Ouvrage publié avec le concours du Laboratoire d'Excellence d'Histoires et Anthropologie des Savoirs, des Techniques et Croyances (ANR-10-LABX-85), de l'École Pratique des Hautes Études – Université PSL.

SOMMAIRE

Partie 3 - Les données de la recherche

Partie 4 - Les contenus des réseaux sociaux

Présentation de l'ouvrage

Le comité scientifique du séminaire « Les nouveaux paradigmes de l'archive »

1 Le séminaire « Les nouveaux paradigmes de l'archive » est issu du partenariat mis en place, en 2019, entre les Archives nationales, le laboratoire Dicen-IDF[1] du Conservatoire national des arts et métiers et le Laboratoire d'excellence HASTEC[2], rejoints en 2020 par le Centre Jean-Mabillon de l'École nationale des chartes. S'adressant à un public hybride composé de chercheurs, jeunes chercheurs, étudiants en archivistique, documentation et *data-management*, ainsi que de professionnels des archives et du *records management*, il aborde les disruptions liées aux mutations numériques du document, aux flux de *data*, au rôle des algorithmes et à l'intelligence artificielle. Toutefois, il ne perd pas de vue la permanence des missions et des pratiques des acteurs du patrimoine et de la connaissance.

2 De 2019 à 2023, une quarantaine d'intervenants, praticiens ou théoriciens, professionnels du patrimoine, archivistes, documentalistes, bibliothécaires, ingénieurs, chercheurs (histoire, anthropologie, sciences de l'information, informatique, etc.), ont croisé leurs approches au cours de dix-huit séances de travail. Leurs compétences et leurs missions les ont conduits à aborder la dimension numérique des archives et du patrimoine sous des facettes diverses : la conservation, l'accès à la connaissance, l'exploitation à des fins de recherche, la structuration, l'éditorialisation ou la valorisation des contenus, ainsi que l'élaboration d'outils.

3 Prolonger ces rencontres par un ouvrage a paru une évidence à la fin de la troisième année. Nous avons alors repris contact avec les intervenants pour leur proposer de le composer avec nous, en reprenant et actualisant la thématique qu'ils avaient traitée au cours du séminaire ou en abordant un nouveau sujet. La publication collective née de cette nouvelle coopération propose un panorama des pratiques et de méthodes émergentes dans le monde des archives. Elle s'adresse aux professionnels des archives, aux étudiants de master et jeunes chercheurs en archivistique et sciences de l'information. Nous avons fait le choix de rejoindre une collection ouverte et gratuite afin que le coût et l'accès ne soient pas un frein à sa consultation. Le format numérique offre également une liberté de forme que nous avons exploitée et les lecteurs rencontreront des textes de tailles et formes diverses : exposés scientifiques, retours

d'expérience, entretiens, discussions ou comptes rendus de séances de séminaires. Certains abordent le numérique sous l'angle de la technique, des outils ou de l'architecture de l'information, d'autres privilégient les usages ou les méthodes des professionnels. Cette variété, qui reflète celle des séances du séminaire, illustre la façon dont le numérique affecte tous les secteurs du traitement et de la conservation des archives. Chaque lecteur est donc assuré de trouver un ou plusieurs articles convenant à ses centres d'intérêt.

4 L'ouvrage se structure autour de quatre types d'archives numériques qui induisent de nouveaux paradigmes de traitement : les données volumineuses, l'audiovisuel, les données de la recherche, les contenus des réseaux sociaux. Pour chacun d'entre eux, nous présentons les enjeux, les technologies et les outils mobilisés, sans omettre les usages.

Partie 1 : Les données volumineuses

5 Dans cette période de cohabitation des données et des documents, les « data » viennent modifier notre perception du réel. Concevoir le monde à travers des instruments de mesure ou accéder à la connaissance *via* des analyses automatiques peut tout à la fois masquer la réalité et ouvrir à de nouvelles opportunités de savoir.

6 Béa Arruabarrena et Armen Katchatourov, dont le texte, issu d'un dialogue mené lors d'une séance du séminaire, s'ancre sur l'exemple des données issues des objets connectés et de la mesure de soi, livrent leur réflexion sur l'influence des données sur les comportements sociaux. Comment modifient-elles le rapport à la norme sociale ? Ils nous rappellent que la volatilité des données, par opposition à la stabilité de l'archive, leur production et leur agrégation souvent obscures ne permettent pas au citoyen de leur porter un regard critique. Ils insistent sur l'importance d'un archivage des données et de la contextualisation de leur constitution.

7 Ghislaine Chartron esquisse un panorama des métiers qui émergent dans le contexte de la gouvernance des données, laquelle occasionne, selon elle, « des évolutions profondes » plutôt que des « transformations radicales ». Elle définit les notions de *data scientist, data manager*, *data stewart* et explique pourquoi de nouvelles compétences en mathématiques, statistiques et algorithmique devront venir conforter les fondamentaux (la rigueur, la qualité, l'objectivité, l'ouverture) des professionnels de l'information.

8 Un exemple concret de l'approche « par les données » d'une analyse de corpus audiovisuels patrimoniaux est donné par le projet ANTRACT qui associe les chercheurs de l'Ina (Jean Carrive, Abdelkrim Beloued, Steffen Lalande), du Centre d'histoire sociale des mondes contemporains (Pascale Goetschel, Franck Mazuet), de l'Institut d'histoire des représentations et des idées dans les modernités (Serge Heiden, Bénédicte Pincemin), d'EURECOM (Pasquale Lisena, Raphaël Troncy) et du Laboratoire d'informatique de l'université du Mans (Sylvain Meignier). Un corpus de 1 262 films d'actualités diffusés entre 1945 et 1969 est « décomposé » en données avec l'aide de diverses technologies : transcription automatique de la parole et identification des entités nommées, analyse visuelle automatique et identification des personnes à l'écran. Ces données font ensuite l'objet d'une analyse textométrique. L'intelligence artificielle rend ainsi possible l'analyse des corpus visuels volumineux.

Partie 2 : L’audiovisuel et ses métadonnées

9 L’audiovisuel numérique (natif ou numérisé) s’invite en masse dans les fonds d’archives sous la forme de films institutionnels ou pédagogiques, de captations ou d’enregistrements d’entretiens qui, au-delà des questions de stockage et de format numériques, interrogent les méthodes, outils et normes archivistiques. Comment décrire des centaines ou des milliers d’heures de vidéo pour guider l’usager non plus vers la pièce, mais vers son contenu ?

10 Sandrine Gill dresse le contexte historique, réglementaire et normatif des fonds audiovisuels versés aux Archives nationales et montre que chaque étape du processus de leur archivage présente des spécificités. Elle se demande comment concilier les normes de description archivistiques issues de l’ISAD(G), orientées vers la contextualisation d’un fonds considéré comme un ensemble, et les pratiques documentaires inspirées de l’Ina, des bibliothèques ou des cinémathèques qui analysent le document individuellement. L’archiviste navigue entre les deux pratiques pour répondre aux attentes des usagers, tout en se rappelant que leur statut juridique d’œuvre rend les archives audiovisuelles plus complexes à communiquer que les autres archives publiques.

11 Le projet d’indexation de la captation du procès des attentats terroristes du 13 novembre 2015, accompli sous l’égide des Archives nationales de septembre 2021 à juin 2022, illustre les défis auxquels peuvent être confrontés les archivistes. Martine Sin Blima-Barru, Claire Scopsi et Aurore Juvenelle décrivent le dispositif de captation et de diffusion audiovisuelles, mis en place par le ministère de Justice dans le prétoire, et les outils méthodologiques et techniques qu’elles ont imaginé pour décrire séquence par séquence les 700 heures des débats, au moment même de leur captation. Ce retour d’expérience pointe les difficultés de l’exercice d’indexation « à la volée » qui demande d’effectuer, sans aucun recul, le choix des termes d’indexation et des séquences à annoter.

12 Un autre défi est celui de la gouvernance des données dans une structure aussi complexe que l’Institut national de l’audiovisuel : comment non seulement centraliser et sécuriser le stockage des éléments audiovisuels, mais aussi centraliser et agréger les masses de données produites par des services aux missions diverses et gérer un modèle de données unique sans contraindre les usages multiples de l’institution ? Gautier Poupeau revient sur les réflexions qui l’ont conduit à concevoir une architecture intégrant une base de données relationnelle, une base de données graphe, une base de données document et un moteur de recherche, et dans laquelle la totalité des données a été migrée. Ce « lac de données » fonctionne comme une boîte noire, à laquelle on accède par des web services qui servent, à la volée, des données mises en forme selon les besoins des différentes applications de l’institution. Cette expérience montre combien les données modifient les approches techniques et organisationnelles.

13 C’est un aspect plus managérial de la gouvernance des données que relatent Rémi Pailhou, Mevena Guillouzic-Gouret et Stéphanie Ange. Le réseau Diazinteregio s’est développé depuis les années 1990 par l’association progressive de cinémathèques, d’archives départementales et de structures patrimoniales françaises dépositaires de fonds de films, dans le but de développer en commun la base de données Diaz. Dans cette application sont versées les métadonnées des collections des membres et un

thesaurus commun est en cours d'élaboration. Le réseau, qui maille la plus grande partie du territoire français, a développé une culture de la coopération poussée et il s'est doté d'un modèle économique et d'instances de décision fondées sur la discussion et la prise en compte des besoins de chacun.

Partie 3 : Les données de la recherche

14 L'ouverture des données de la recherche s'organise sous l'impulsion du ministère de l'Enseignement supérieur et de la Recherche français. Elle concerne particulièrement l'édition scientifique, mais la conservation pérenne, l'accès aux données produites dans le cadre de projets financés sur des fonds publics et leur réemploi, impliquent aussi l'intervention des archivistes et des concepteurs d'infrastructures de dépôt.

15 Désormais les chercheurs des humanités numériques doivent s'habituer à « penser leurs données » en adaptant les méthodes de leur discipline. Gérald Kembellec et Claire Scopsi présentent l'historique récent, les concepts et l'écosystème français de la politique d'ouverture des données de la recherche. Ils soulignent que ce processus oblige les chercheurs en sciences humaines et sociales à adopter le modèle des sciences dures, en structurant et décrivant leur corpus pour en permettre le stockage, la propagation et la consultation. Rendue nécessaire par l'interdisciplinarité, cette modélisation des corpus peut être facilitée par une collaboration pluridisciplinaire et l'intégration des services d'appui à la recherche (centre de documentation ou d'archives, etc.) aux équipes projets.

16 À travers une analyse de la littérature scientifique, Nicolas Sauret, Jean-Luc Minel, Mélanie Bunel et Stéphane Pouyllau proposent un parcours des pratiques de dépôts de données de la recherche et les fonctionnalités et caractéristiques des infrastructures dédiées à leur conservation et à leur exploitation. Trois infrastructures sont présentées : une infrastructure nationale, la *Digital Archiving and Networked Services*, une infrastructure européenne, CLARIN, et l'infrastructure FAIR de l'écosystème Huma-Num reposant sur ISIDORE et NAKALA. L'usage de ces services se développe rapidement et les conduit à évoluer notamment en intégrant des technologies d'intelligence artificielle.

17 Annick Boissel et Véronique Ginouvès livrent les étapes de la rédaction d'un « plan de gestion de données rétrospectif » lié au don du fonds de l'anthropologue Jean-Pierre Olivier de Sardan à la Maison méditerranéenne des sciences humaines [MMSH]. Elles conçoivent ce document, rédigé à l'issue du traitement de chaque fonds, comme un outil destiné aux informaticiens, archivistes et réutilisateurs des données, pour faciliter la compréhension du fonds, constitué des documents versés, associés aux contrats signés, aux documents de contextualisation, aux instruments de recherche rédigés par les archivistes et aux enrichissements apportés par divers acteurs. Le modèle qu'elles ont mis au point comporte neuf parties qui décrivent tant le fond que la forme des données et métadonnées et permet d'anticiper le cycle de vie du fonds.

Partie 4 : Les contenus du web et des réseaux sociaux numériques

18 Par leur volume et leur fluence, les contenus des réseaux sociaux numériques [RSN], des sites web et des plateformes vidéos rompent avec le modèle traditionnel de l'archive. Que l'on considère le « dépôt » légal du web ou les campagnes de sauvegardes thématiques, les contenus ne sont jamais déposés, mais collectés sans recul historique par les archivistes à qui revient le choix des méthodes d'historisation.

19 L'opération « Mémoire de confinement », présentée par Rosine Lheureux et Julia Moro, nous rappelle que la pandémie de Covid-19 a été une période d'intenses collectes de traces numériques. À l'issue du premier mois de confinement, les archivistes et la webmaster des Archives départementales du Val-de-Marne unissent leurs compétences et s'organisent à distance afin de récupérer auprès de la population locale les productions numériques qu'elle juge représentatives de son quotidien confiné. Deux cents dons de traces numériques, de formes et de technologies d'une extrême diversité, témoignent de l'ambiance particulière de cet événement et de l'état d'esprit des habitants du département. Ils sont aussi un support de réflexion pour les archivistes confrontés, dans cette collecte « à chaud », à une temporalité inhabituelle.

20 À l'opposé de cette collecte sélective, la mission de dépôt légal de l'Ina restitue une matière que Boris Blanckemane qualifie d'« omnidonnées massives ». Très exigeante techniquement, la récupération des sites web, RSN, comptes de plateformes vidéos ou *tweets* implique des stratégies de collectes dédiées et des robots adaptés à la spécificité des sources, au volume et à la fréquence des données produites. Le stockage, l'indexation, la manipulation et la mise à disposition de ces omnidonnées massives nécessitent une combinaison complexe d'outils et de concepts que les archivistes doivent appréhender pour mieux communiquer avec l'équipe technique. L'évolution vers un archivage centré sur la donnée s'accompagne ainsi nécessairement de l'acquisition de nouvelles compétences.

21 La BnF et l'Ina, tous deux chargés d'un dépôt légal du web, ont mis au point des stratégies de collectes des RSN adaptées aux domaines pour lesquels elles sont missionnées (le web lié à la radio et la télévision pour l'Ina, les « autres web français » pour la BnF), aux volumes extrêmes, au nombre de publications à suivre et aux attentes des usagers. Alexandre Faye, Jérôme Thièvre et Valérie Schafer proposent, à l'aide d'exemples concrets, une comparaison de leurs méthodes respectives, qui relèvent souvent de la tactique pour contourner les blocages des éditeurs ou réagir rapidement à une actualité. L'hétérogénéité des données, les caractéristiques techniques et la nature même des cibles de la collecte impliquent un travail d'instruction pour chaque plateforme. L'accès à ces ressources est donc complexe pour le chercheur qui doit prendre connaissance des modalités de collecte de chaque institution, mais bénéficie aussi de leur complémentarité. Autour d'un même thème, certaines collections rendent compte de l'apparition d'un thème dans les médias nationaux, d'autres de sa viralité géographique, ou bien, et c'est une tendance nouvelle, des expressions individuelles.

Ghislaine Chartron (Cnam)

Françoise Lemaire (Archives nationales)

Rosine Lheureux (Archives nationales)

Yann Potin (Archives nationales)

Clothilde Roullier (Archives nationales)

Claire Scopsi (Cnam)

Martine Sin Blima-Barru (Archives nationales)

Édouard Vasseur (École nationale des chartes)

NOTES

1. Dispositifs d'information communication à l'ère numérique.
2. Histoire et anthropologie des savoirs des techniques et des croyances.

Du neuf avec de l'ancien : les changements de paradigme des archives

Olivier Poncet

1 Depuis les travaux post-modernes de Michel Foucault (1969) et de Jacques Derrida (1995) et, plus encore, au fur et à mesure du développement des technologies numériques, les archives semblent avoir été totalement reconfigurées (Hamilton, Harris *et al.*, 2022). Elles semblent avoir perdu la rassurante définition qui avait été la leur depuis des siècles, voire des millénaires : un ensemble organisé de documents écrits sur des supports physiques conservés dans des institutions qualifiées du même nom et témoignant de l'activité publique ou privée de personnes morales ou physiques. Pourtant, l'histoire des archives enseigne que la parabole des vieilles outres déclarées inadaptées à recevoir le vin nouveau (Luc 5, 38) ne s'applique pas à elles. Sous l'eau dormante et la conception étroite des dictionnaires, archives et archivistes n'ont cessé de se renouveler dans un cadre anciennement fixé (Ketelaar, 2016). Il n'est pas, dans la langue française, jusqu'à la grammaire qui n'ait rendu compte des mutations de « l'archive » (Anheim, 2004).

2 Depuis qu'il existe un discours réfléchi sur les archives, l'évolution dont celles-ci sont l'objet a été perçue et commentée diversement, que ce soit dans une perspective historique, comme chez Baldassare Bonifacio au XVII^e^ siècle (1632), ou sous l'effet de considérations juridiques et administratives sous la plume d'Armand-Gaston Camus quelques années avant la Révolution française (Outrey, 1953). Plus récemment, au XX^e^ siècle, c'est sur un mode inquiet que les importantes mutations des missions des archivistes ont été discutées (Blaquière, 1963), tandis que le temps digital ouvert à la fin du deuxième millénaire et au début du troisième a installé l'idée qu'un nouveau paradigme des archives s'imposait (Bantin, 1998). Cette accélération des défis et des prises de conscience rappelle que toute archivistique est aussi une futurologie (Hill, 2011). Il n'est pas jusqu'à l'apport de la nouvelle proposition théorique du « *records continuum* » qui n'ait été aussitôt inscrite dans un temps long et non fini (McKemmish & Gilliland, 2018) et n'ait entraîné un paradoxal mouvement perpétuel de la conception

même de paradigme (Gilliland, 2018, p. 6-10). Redisons-le cependant : il n'existe pas, il n'a jamais existé de théorie archivistique reposant sur des principes immuables et reproductibles, quel que soit le contexte (Cook, 1997, p. 229-230).

3 Veut-on chercher à tout prix des changements radicaux dans la vie des archives à l'ère digitale ? Écartons en ce cas d'emblée la fausse piste, d'apparence évidente, du support. Ce dernier change moins de choses qu'on ne le croit, comme l'ont souligné à l'envi les archivistes confrontés dès la fin des années 1990 avec le développement massif des archives numériques (Cox, 1994 ; Duranti, 1994 et 1998 ; Banat-Berger & Nougaret, 2014). Ces dernières décennies, le support numérique, qu'il soit natif ou secondaire (Potin, 2011), a certes modifié la matérialité des archives, dans des proportions que celles-ci n'avaient pas connues depuis la fin du monopole (au moins conservé) des supports minéraux (inscriptions sur la pierre ou le métal, tablettes d'argile ou *ostraca*) et l'apparition des supports végétaux de par le monde, du papyrus antique égyptien aux *olles* ou *lontars* (feuille de palmier) de l'Asie du Sud-Est, en passant par les écorces de bouleau de la Russie médiévale, jusqu'aux multiples avatars du papier au cours des deux derniers millénaires ; l'expérience du parchemin (peau de mouton, chèvre ou veau) antique et occidental semble à cet égard relativement courte sur l'échelle du temps. Les archives numériques restent des archives. Même la possibilité d'une reproduction photographique (avec toutes les extensions technologiques attachées à ces procédés de reprographie) n'a pas modifié fondamentalement la donne et les archivistes ont su se prémunir sans peine des dangers de la reproductibilité à l'infini d'un document avant de s'adapter sans trop de difficultés à la source audiovisuelle, en définitive plutôt aisée à intégrer par le questionnaire diplomatique (Delmas, 2003).

4 L'histoire des archives, comme d'autres domaines de la vie et de l'activité humaines, est scandée par des périodes de définition et de durée variables où les auteurs s'efforcent d'analyser les changements et les innovations qui les distinguent des précédentes. Robert-Henri Bautier a ainsi décelé dans la période moderne, élargie au début du XIX^e^ siècle, « la phase cruciale » de leur histoire (Bautier, 1968) : nouveaux usages (politique et historique), naissance des grands dépôts ministériels, lente constitution d'une science archivistique. La démonstration reste convaincante. Elle est malgré tout en large part téléologique et vise principalement à retrouver dans le passé les fondements de la situation archivistique française au début du dernier tiers du XX^e^ siècle. Elle tend surtout à mettre en exergue des phénomènes qui tranchent effectivement avec la période qui précède immédiatement celle que Bautier étudie et qu'il connaît fort bien, la France du Moyen Âge. Pourtant, sur un temps plus long et moyennant des analogies menées avec prudence et mesure, on retrouverait des exemples pour des périodes plus anciennes qui tendraient à souligner autant d'éléments de continuité que de rupture.

5 L'époque moderne est ainsi à l'origine d'une apparente mutation des typologies conservées qui se révèle moins novatrice qu'on ne le penserait. Au fil des millénaires et surtout des derniers siècles, les cibles des archives ont en effet été affectées de profonds renouvellements, à commencer les correspondances missives. Non créatrices de droit positif, les lettres ont mis longtemps, du Moyen Âge à la Révolution, à gagner une légitimité à être archivées sur le temps long. Elles sont pratiquement absentes des chartriers médiévaux avant de revenir au cœur des nouvelles pratiques archivistiques de l'âge moderne qui saisissent leur caractère emblématique d'un nouvel art de gouverner (Poncet, 2019). Pourtant, elles emplissaient déjà les dépôts mésopotamiens

du deuxième millénaire avant J.-C. où elles reflétaient la réalité d'une souveraineté et d'une capacité à dialoguer ou à imposer. Depuis l'élargissement conceptuel et pratique des temps modernes, la matière épistolaire n'a plus quitté les archives. Elle a désormais acquis une place de choix dans les stratégies de collecte, qu'elle soit disponible sous forme matérielle ou numérique (mails, réseaux sociaux, etc.).

6 Il en va de même des usages des archives. Ceux-ci n'ont jamais été exclusifs les uns des autres et ont plus souvent coexisté qu'ils ne se sont succédé les uns aux autres. Progressivement dégagées de l'emprise exclusive de la preuve de droit, encore sensible dans la définition[1] qu'en donnait Arthur Giry à la fin du XIX^e^ siècle, les archives ont conquis un statut de source, de témoin historique voire de trace. Dès le Grand Siècle, les généalogistes-antiquaires étaient susceptibles de faire flèche de tout bois pour dater ou attribuer un fait ou un événement. La déconstruction que leur fait subir Mabillon dans son *De re diplomatica* dépassait l'opération de *discrimen veri ac falsi* en établissant les règles qui confèrent à tout document une potentielle validité historique (Poncet, 2022, p. 273-274). Mais, dès l'Antiquité égyptienne de Manéthon, l'histoire s'écrivait avec des textes de la pratique articulés à des récits oraux et des mémoires écrits. Et si aujourd'hui l'utilité foncière (aux deux sens du mot) semble, aux yeux des chercheurs, avoir cédé le pas sous l'effet des questionnements historiques et des sciences sociales, le lectorat (physique ou à distance) des services d'archives est encore à la recherche de preuves de droit pour les objets les plus divers (propriété, identité, etc.).

7 L'alternative représentée par la bibliothèque dans nos sociétés digitalisées, comme le traduit, entre autres, l'archivage du web par la Bibliothèque nationale de France, n'est pas davantage une véritable nouveauté. Dans ce dernier cas, comme dans celui de l'Institut national de l'audiovisuel, la notion de dépôt « légal » d'une publication a bien entendu facilité l'archivage par une bibliothèque de documents souvent plus hybrides qu'il y paraît. Mais il y a longtemps que les bibliothèques sont des lieux d'archives dans un processus d'hybridation (Chastang & Lemaigre-Gaffier, 2023) où l'expérience du « temps bibliothécaire » des archives que furent les XVII^e^ et XVIII^e^ siècles (Chapron, 2015) a en réalité été poursuivie sur d'autres modalités au cours des décennies et des siècles suivants jusqu'à aujourd'hui. Les rôles y sont d'autant plus partagés ou inversés en raison de dynamiques différenciées que la nature archivistique a d'une certaine manière horreur du vide : ce que l'un ne recueille pas, l'autre s'en saisira à sa façon. Les récentes typologies documentaires (sites Internet, archives audiovisuelles) ont ainsi en partie « échappé » aux archives instituées, en dépit d'une loi française (1979) très englobante, mais des organismes nombreux et diversifiés les ont pris en charge.

8 En définitive, ce qui a probablement le plus changé dans les époques récentes, c'est la position de l'archiviste dans le dispositif d'archivage entendu comme collecte, conservation et classement. Le temps où l'archiviste était avant tout un agent posté devant ou dans son dépôt, au service de l'organe ou du pouvoir dont il gardait (littéralement) les documents archivés, est plus ou moins révolu. L'archiviste s'est fait le lecteur critique des archives confiées à ses soins, parfois au risque de se fâcher avec un producteur qui n'en demandait pas tant et exigeait plutôt le silence et l'enfouissement que la publicité et l'accueil (Filippini, 2007). Plus récemment, depuis environ un siècle, il a pensé au-delà des missions qui lui étaient implicitement ou explicitement indiquées. En accompagnant avec volontarisme l'appétit des historiens pour des sources que ne recevaient pas ou trop peu les archives publiques, les archivistes ont ouvert les portes de leurs dépôts à des documents provenant de

personnes privées, particuliers, entreprises, associations, etc., jusqu'à leur réserver une place explicite dans les organigrammes fonctionnels et dans les plans de classement officiels : en France, à la création dans les archives départementales, par la circulaire du 15 avril 1944, d'une série J destinée à accueillir les entrées par voie extraordinaire a répondu l'instauration aux Archives nationales d'une sous-section des archives privées, économiques et du microfilm avec quatre nouvelles sous-séries thématiques (AP, AQ, AR, AS) (Nougaret, Éven, Lacousse *et al.*, 2008, p. 10, 14). Non content de compléter par ce picorage dans la masse des archives produites dans le pays les fonds d'administrations publiques dont la collecte s'efforçait, à coup de circulaires, de revêtir un caractère d'automaticité, l'archiviste se donna parfois un rôle de démiurge pour susciter, voire créer des documents d'archives. Des campagnes d'archives orales de plus en plus professionnelles aux opérations plus récentes de grande collecte effectuées à chaud après un événement d'ampleur collective (attentats de 2015, épidémie de 2020), les archivistes ont apporté une contribution littéralement constructive aux archives de leur temps.

9 Quant au traitement à apporter aux archives pour les rendre accessibles, il est le lieu des renouvellements à la fois les plus neufs et les plus fidèles à l'activité conduite par les archivistes au fil des siècles. L'archivistique est une branche à part entière de l'histoire des savoirs et le tournant archivistique, dans son versant universitaire spécialement, y a beaucoup insisté (Corens, Peters & Walsham, 2018). Les enjeux fondamentaux sont bien toujours de retrouver, et de retrouver vite et bien, ce que l'on cherche ou ce que l'on s'attend à retrouver. Enregistrer les entrées, décrire les pièces venues du passé, produire des outils de navigation pertinents, toutes ces démarches documentaires sont constantes dans l'activité des archivistes, qu'elles se nomment analyses, répertoires numériques, états des versements, guides de service, guides de recherche, index, bases de données, métadonnées, lacs de données, etc. L'ouverture au public de portions de plus en plus larges des archives a obligé à partager un peu de ce savoir et surtout forcé de le rendre plus intelligible pour un autre que celui qui conserve et classe. De fil en aiguille et de prises de conscience personnelle en injonctions sociales, l'archiviste est venu à penser sa responsabilité non plus seulement envers ses donneurs d'ordre et employeurs, mais envers une société qui ne se résume plus exclusivement à ses pairs savants ou historiens. Dans la nouvelle ère que d'aucuns qualifient de post-custodiale (Ham, 1981), la démultiplication des points de vue qu'il lui faut conjuguer désormais est proprement énorme (Gilliland, 2018[2]).

10 Le métier d'archiviste a-t-il pourtant si radicalement changé au temps de la *data* triomphante ? Dans les divers acteurs de ce nouveau monde, où le situer ? Probablement quelque part entre le *data manager* et le *data steward*, entre, d'une part, un rôle de conception et de gouvernance de la donnée à haut niveau et, d'autre part, un rôle de police de l'accès et de l'utilisation de la donnée. Mais là encore, ces mots – conception, police – font référence à des missions inscrites depuis longtemps dans l'imaginaire de l'archiviste, un savant doté de pouvoir de contrôle qui empêche de l'enfermer strictement dans un rattachement ministériel, comme nous l'apprend, dans le cas français des XIX^e^-XX^e^ siècles, le passage de l'Intérieur à l'Instruction publique puis à la Culture, avant peut-être de rechercher d'autres tutelles dont on ne soupçonne pas encore l'existence.

11 Les nouvelles perspectives ouvertes par l'intelligence artificielle, l'un des plus grands défis qui attendent l'humanité, pourraient bien, pour une fois, provoquer une lourde

mutation, une transfiguration même de l'économie des archives. Ces technologies, qui s'apparentent pour les esprits pressés à la martingale à laquelle aspirent les chercheurs de tout poil, semblent pouvoir s'affranchir de l'intervention d'un archiviste, désormais inutile à engranger, à classer, à signaler, à conduire et à guider. Toutefois, ce que rappelle l'IA, c'est que les archives sont d'abord et avant tout un discours et un régime de savoir. Foucault, dans tous ses détournements, ses ambiguïtés et les outrances qu'on lui a fait dire, avait vu juste. En décrétant que les archives sont d'abord un « système d'énoncés » (*L'archéologie du savoir*), il anticipe sans le savoir – tout en le pressentant fortement y compris dans ses implications totalitaires – ce que sont les ambitions d'un questionnement soutenu par des technologies surpuissantes (par rapport à l'esprit humain). Créer et recréer des fonds d'archives par la grâce des recompositions infinies de la machine fait certes écho aux fabrications successives des archives dont l'archiviste est en partie l'auteur. Mais ces recompositions sont désormais virtuelles et le basculement est sans doute là comme le pressent Gilles Deleuze :

> Le savoir est affaire d'archive, le pouvoir n'est pas affaire d'archive. Le pouvoir est affaire d'une cartographie, d'une cartographie mouvante, une carte stratégique, toujours remaniable, toujours fluente. Si bien que les deux mots de Foucault pourraient être ceux-ci, au point où nous en sommes : jusqu'à *Surveiller et punir* je suis un archiviste, avec *Surveiller et punir* et *La volonté de savoir*, je suis un cartographe[3]. (Deleuze, 1986)

12 Le rapport des archives aux sciences de l'information, parfois regrettablement conflictuel, d'autres fois (et c'est heureux) coopératif et à l'intérêt bien compris, n'est plus la question principale qui agite les professionnels comme elle fut durant le dernier demi-siècle. Ce qui compte à présent, c'est moins le produit fini et la part qui revient à chacun, que le processus par lequel on parvient à l'établir (Cook, 1997). Après avoir beaucoup douté de leur nature et de leur formation, les archivistes en ont finalement retiré le sentiment que le savoir historique était après tout la meilleure arme pour administrer les archives les plus récentes (Nesmith, 2004). Le débat semble cependant s'être singulièrement déplacé.

13 En effet le web de données, par nature mouvant et changeant, dans lequel l'utilisateur doit apprendre à trier et s'orienter, bouleverse hiérarchies et responsabilités. Les technologies actuelles permettent à tout un chacun de s'affranchir apparemment des professionnels de l'information et des archives, quels qu'ils soient. Le régime de savoir qu'analysait Foucault n'est plus disciplinaire et imposé à l'individu par une norme collective, médiatisée par des archives et des archivistes, mais c'est bien l'individu qui devient le centre et non plus l'objet de la gestion de l'information et de la mémoire. L'individualisme de la norme, où chacun se concocte ses archives pour la gestion de son existence (de l'identité à la santé en passant par tout ce qui peut être imaginé pour se construire des règles de vie), est une perspective qui fait de nous tous des archivistes. Il se pourrait bien que notre monde soit près de succomber à ce caractère hallucinogène pointé naguère par Michel Melot (1986). À moins que l'archiviste, plus que jamais conscient de ce qui se joue, ne sache faire preuve d'un sursaut d'adaptabilité. Une nouvelle fois.

BIBLIOGRAPHIE

Étienne ANHEIM, « Singulières archives. Le statut des archives dans l'épistémologie historique. Une discussion de *La mémoire, l'histoire, l'oubli* de Paul Ricœur », dans *Fabrique des archives, fabrique de l'histoire*, éd. par Étienne ANHEIM et Olivier PONCET, numéro spécial de *Revue de synthèse*, 5e série, t. 125, 2004, p. 153-181.

Françoise BANAT-BERGER et Christine NOUGARET, « Faut-il garder le terme archives ? Des "archives" aux "données" », *La Gazette des archives*, 233, 2014, p. 7-18.

Philip C. BANTIN, « Strategies for managing electronic records : a new archival paradigm ? An affirmation of our archival traditions ? », *Archival Issues*, 23, 1998, p. 17-34.

Robert-Henri BAUTIER, « La phase cruciale de l'histoire des archives : la constitution des dépôts d'archives et la naissance de l'archivistique, XVIe-début XIXe siècle », dans *Archivum*, 18, 1968, p. 139-150.

Henri BLAQUIÈRE, « L'avenir des archives », *La Gazette des archives*, n° 41, 1963, p. 59-65.

Baldassare BONIFACIO, *De archivis liber singularis. Ejusdem praelectiones et civilium institutionum epitome*, Venise, apud J. B. Pinellium, 1632.

Emmanuelle CHAPRON, « The "supplement to all archives". The Bibliothèque royale of Paris in the eighteenth century », dans *Archives and the writing of history*, numéro spécial de *Storia della storiografia*, t. 68, no 2, 2015, p. 53-68.

Pierre CHASTANG et Pauline LEMAIGRE-GAFFIER, « Partages et hybridations : archives et bibliothèques du XIIe au XVIIIe siècle », dans *Archives en bibliothèques (XVIe-XXIe siècles)*, éd. par Emmanuelle CHAPRON et Fabienne HENRYOT, Lyon, ENS Éditions, 2023, p. 37-54.

Terry COOK, « What is past is prologue. A history of archival ideas since 1898 and the future paradigm shift », *Archivaria*, 43, 1997, p. 18-63, rééd. dans *« All shook up ». The archival legacy of Terry Cook*, éd. par Tom NESMITH, Greg BAK et Joan M. SCHWARTZ,Chicago, The Society of american archivists et Association canadienne des archivistes/Association of Canadian archivists, 2020, p. 227-274.

Liesbeth CORENS, Kate PETERS et Alexandra WALSHAM (ed.), *Archives and information in the early modern world*, Oxford, The British Academy et Oxford University Press, 2018 (Proceedings of the British Academy, 212).

Richard J. COX, « The record: is it evolving? », *Records retrieval report*, 10, 1994, p. 1-16.

Gilles DELEUZE, *Sur Foucault, le pouvoir*, transcription de cours à l'université de Vincennes-Saint-Denis, cours du 25 février1986, https://www.webdeleuze.com/textes/276.

Bruno DELMAS, « Donner à l'image et au son le statut de l'écrit : pour une critique diplomatique des documents audiovisuels », Bibliothèque de l'École des chartes, t. 161, 2003, p. 553-601.

Jacques DERRIDA, *Mal d'archive : Une impression freudienne*, Paris, Galilée, 1995.

Luciana DURANTI, « The record, where archival universality resides », *Archival Issues*, 19, 1994, p. 83-94.

Luciana DURANTI, *Diplomatics. New uses for an old science*, Lanham, Scarecrow Press, 1998.

Orietta FILIPPINI, « “Per la fuga non disinteressata di notizie”. Michele Lonigo dall’Archivio Vaticano alle prigioni di Castel Sant’Angelo (1617) : i costi dell’informazione », dans *Offices, écrits et papauté (XIII^e^-XVII^e^ siècle)*, éd. par Armand JAMME et Olivier PONCET, Rome, École française de Rome, 2007 (Collection de l’École française de Rome, 386), p. 705-736.

Michel FOUCAULT, *L’archéologie du savoir*, Paris, Gallimard, 1969.

Anne J. GILLILAND, *Conceptualizing 21st Century Archives*, Chicago, ALA Editions, 2018.

Arthur GIRY, « Archives », dans *La Grande Encyclopédie*, t. III, *Animisme-Arthur*, Paris, H. Lamirault et compagnie, 1887, p. 747-762.

F. Gerald HAM, « Archival strategies for the post-custodial era », *The American Archivist*, 44, 1981, p. 207-216.

Carolyn HAMILTON, Verne HARRIS, Jane TAYLOR, Michele PICKOVER, Graeme REID & Razia SALEH (ed.), *Refiguring the archive*, Dordrecht, Springer-Science, Business Media, 2002.

Jennie HILL, *The future of archives and recordkeeping: a reader*, Londres, Facet Publishing, 2011 (en particulier Victoria Lane & Jennie Hill, « Where do we come from ? What are we ? Where are we going ? Situating the archive and archivists », p. 7-26).

Eric KETELAAR, « Archival turns and returns », dans *Research in the Archival Multiverse*, éd. par Anne J. GILLILAND, Sue MCKEMMISH & Andrew J. LAU, Melbourne, Monash University Publishing, 2016, p. 228-268.

Sue MCKEMMISH et Anne J. GILLILAND, « Archival and recordkeeping research : past, present and future », dans *Research methods. Information, systems and contexts*, éd. par Kirsty WILLIAMSON et Graeme JOHANSON, Cambridge, Kidlington, Chandos Publishing, 2018, p. 85-125.

Michel MELOT, « Des archives considérées comme une substance hallucinogène », *Traverses*, t. 36, 1986, p. 14-19, rééd., Paris, École nationale des chartes, 2023 (Propos).

Tom NESMITH, « What’s history got to do with it ? Reconsidering the place of historical knowledge in archival work », *Archivaria*, 57, 2004, p. 1-27.

Christine NOUGARET, Pascal ÉVEN, Magali LACOUSSE *et al.*, *Les archives privées. Manuel pratique et juridique*, Paris, La Documentation française, 2008, p. 10 et 14.

Amédée OUTREY, « Sur la notion d’archives en France à la fin du XVIII^e^ siècle », *Revue historique de droit français et étranger*, 4^e^ série, t. 31, 1953, p. 277-286.

Olivier PONCET, « Entre patrimoine privé, érudition et État : les vicissitudes des papiers des ministres de la monarchie française (XIV^e^-XVII^e^ siècle) », dans *Recovered voices, newfounds questions. Family archives and historical research*, éd. par Maria DE LURDES ROSA, Rita SAMPAIO DE NÓVOA, Alice BORGES GAGO et Maria JOÃO DA CÂMARA, Coimbra, Universidade de Coimbra, 2019, p. 35-51.

Olivier PONCET, « Au-delà de la preuve. La dramatisation des archives comme discours politique, social et savant (France, XVI^e^-début XVIII^e^ siècle) », dans *Les conflits d’archives. France, Espagne, Méditerranée*, éd. par Stéphane PÉQUIGNOT et Yann Potin, Rennes, Presses universitaires de Rennes, 2022, (Histoire-Archives, histoire et société), p. 259-276.

Yann POTIN, « Institutions et pratiques d’archives face à la “numérisation”. Expériences et malentendus », *Revue d’histoire moderne et contemporaine*, n° 58, 2011, p. 57-69.

NOTES

1. « On désigne sous ce nom les dépôts d'actes, de titres et en général de documents de tous genres ayant un caractère d'authenticité » (Giry, 1887, p. 762).
2. Spécialement p. 37-54, chapitre 2, « Reframing the archive in a digital age; balancing continuity with innovation and responsability with responsabilities » (p. 48-49, table 2.1 : « Examples of multiplying areas of archival engagement over the past fifty years »).
3. En partie cité dans ce volume par Bea Arruabarrena et Armen Khatchatourov.

AUTEUR

OLIVIER PONCET

École nationale des chartes-EHESS

Partie 1 - La mémoire des données

Quels métiers, quelles compétences pour la gestion des Data ?

Ghislaine Chartron

NOTE DE L'ÉDITEUR

Entretien mené par Claire Scopsi, Maître de conférences au Cnam, le 11 juillet 2022 avec Ghislaine Chartron, professeure au Cnam, titulaire de la chaire Ingénierie documentaire, directrice de l'Institut national de sciences et techniques de la documentation [Intd] depuis 2007. Elle a fondé en 2016 le Master Mégadonnées et analyse sociale [MéDAS] du Cnam et nous livre sa vision de l'évolution des métiers et des compétences.

Claire Scopsi : L'Intd a été fondé en 1951. C'est un remarquable observatoire de l'évolution des métiers et des compétences dans le domaine de l'information. Quels sont les outils qui permettent de suivre ces évolutions ?

Ghislaine Chartron : Nous avons un référentiel un peu formel construit avec l'ADBS[1] et les offres d'emplois, les sujets de mémoires, les missions de stages, notamment du titre professionnel de chef de projet en ingénierie documentaire, sont des témoignages intéressants. On voit que nos métiers ont évolué dans le temps, sachant que nous étions identifiés comme des spécialistes de l'indexation, du classement, des archives et de la bibliographie. Tout cela est ancien mais perdure mine de rien. Mais j'ai aussi un prisme personnel pour distinguer les mouvements structurants de l'évolution. Tout d'abord les technologies : il est évident que nos métiers sont des métiers fortement en prise avec une dimension technique. Pour produire des livrables à valeur ajoutée, nous mettons en place des dispositifs dans lesquels la technologie est très structurante, qu'il s'agisse de banques de données, de portails, de moteurs de recherche ou d'archives ouvertes. Au début de ma carrière, il n'y avait que les bases de données bibliographiques, mais nos technologies se sont diversifiées et pourtant, aujourd'hui, faire une base de données bibliographique avec un SGBD[2] relationnel ce n'est pas dépassé. La demande existe toujours.

C. S. : Cela se complexifie, sans abandonner vraiment les approches initiales ?

Gh. C. : Il y a des technologies différentes en fonction des contextes. Dans le contexte associatif, on manipule toujours des petites bases de données documentaires ou des SIGB[3], mais, dans le contexte universitaire, on ne parle plus que d'archives ouvertes, de plateformes de données de recherche et de *data-mining* ! Cette dimension technique, je pense qu'il faut la revendiquer et l'assumer. C'est un atout de notre champ d'activité, par rapport à d'autres professionnels, que de savoir manipuler les outils techniques.

Et puis il y a un deuxième mouvement fort, c'est le web. Le document s'est structuré avec des langages qui permettent de le considérer autrement. Le web a permis l'édition ouverte, l'autopublication, l'éditorialisation, le patrimoine numérique. Tout cela se conjugue avec des technologies nouvelles. Le web et son écosystème ont bouleversé les opportunités et le positionnement de nos métiers : les archives ouvertes et l'édition ouverte, l'enrichissement des documents ainsi que les prises de pouvoir des utilisateurs sur la production et la diffusion des documents… C'est un mouvement majeur depuis vingt ans et cela a changé la valeur ajoutée de nos métiers.

Et la troisième dimension à souligner concerne l'inflation de l'information. C'était si simple avant, quand on n'avait que quelques revues à classer, à archiver… et qu'à raisonner centralement avec des collections d'ouvrages. Aujourd'hui, il y a tellement de sources d'information qu'on ne doit pas ignorer et qui doivent être croisées avec cette information dite « traditionnelle ». C'est ce troisième mouvement qui fait que la veille, à mon avis, reprend aussi de l'essor. Les gens ne savent plus ce qui a de la valeur, ce qui n'en a pas… En même temps, il faut pouvoir s'adapter à des flux différents, aller chercher l'information sur les réseaux sociaux, dans la presse spécialisée ou dans les nombreuses sources audiovisuelles, qu'elles proviennent des médias ou soient auto-diffusées. Tout cela converge sur le web. Cette inflation des sources conduit à revaloriser le rôle de l'analyse, de la veille, c'est-à-dire le rôle de filtrage qui est la dimension originelle de nos métiers.

Enfin, il y a une quatrième transformation forte, la data. Mais nous allons y revenir.

C. S. : Comment définirais-tu les grandes tendances de profils des professionnels ? Est-ce qu'on peut identifier des grands lots de compétences ?

Gh. C. : Si l'on veut rester au niveau macro, je pense qu'il y a toujours des professionnels sur l'activité de développement. Ils travaillent en phase avec l'innovation, ils accompagnent l'aide à la décision, car nos métiers sont des métiers d'appui. De ce côté-là, on trouve les chargés de veille et les *knowledge managers*, voire les *data analysts*. Ils sont sur le *front-office*. D'autres sont positionnés du côté du *back-office*, c'est-à-dire de la gestion. Selon les contextes, ce sera la gestion du risque, de la mémoire. Mais, dans tous les cas, il s'agit de constituer des traces, du patrimoine ou des documents engageants. On y trouve les *records managers*, les chargés d'archivage numérique ou de projets GED. Et puis il y a le contexte dans lequel on exerce ces métiers…

C. S. : Et la proximité du public. Il y a des métiers plus fortement tournés vers des usagers qui ne sont pas forcément dans l'activité d'entreprise. Il y a aussi une activité grand public.

Gh. C. : C'est pour ça que je dirais qu'on peut distinguer trois contextes différents avec des intitulés de postes qui vont varier. Il y a l'entreprise, le domaine scientifique et le savoir, et enfin le domaine de la culture. Pour moi, ces trois grands domaines ont des enjeux différents.

Pour les entreprises, c'est la gestion du risque et de l'innovation qui prime, alors que dans les sciences, c'est la circulation des savoirs et l'innovation. Et dans le secteur culturel on constitue des traces, des archives et de la mémoire.

C. S. : Mais la culture ça peut être aussi l'économie du divertissement… la distinction n'est pas absolue.

Gh. C. : Certes. L'Intd, en fonction des années, a mis le curseur sur un domaine ou sur un autre. En ce moment, nos débouchés, dans la spécificité du Cnam, concernent essentiellement la GED[4] et le *records management*. Nous avons renforcé cet axe, tourné vers les entreprises, car il est en phase avec la demande. Mais il ne comporte pas que des enjeux techniques, il y a également des enjeux normatifs récents. Nos élèves ont plutôt ce profil en ce moment. Mais certains ont aussi des profils de gestionnaires de portail documentaire car le besoin est toujours présent dans les institutions publiques et associatives où il existe des fonds documentaires qu'on veut valoriser. Donc les chefs de projet de système d'information documentaire existent toujours, même si la proximité et la convergence avec d'autres métiers sont observables.

Quant à la data… Pour moi, le master MéDAS n'est pas déconnecté des profils originels de l'Intd. Si j'ai rattaché MéDAS à l'Intd, c'est que je suis convaincue que, même si le paradigme informatique domine, la data tend à englober tout un ensemble de fonctions. Tout devient data. Même s'il y a des documents, même s'il y a de l'*open data*, il y a aussi des données issues du web et des ERP[5], et tout ça, c'est du matériau informatif. Alors, et tant pis si ça fait bondir les puristes du document, pourquoi ne pas considérer le document comme une data particulière ? Je suis persuadée qu'il y aura dans les organisations des services transverses où l'on traitera tout cela ensemble. Aujourd'hui, on voit que ce qui motive la création de postes, c'est ce qui fait gagner de la performance à une organisation. L'automatisation est aussi en marche et se substitue à certaines missions des professionnels de l'information.

C. S. : On est obligé d'être pragmatique, et comme on ne peut pas être exhaustif et tout prendre, on ne va prendre que ce qui est utile ?

Gh. C. : On va sélectionner, mais on doit aussi considérer la diversité des sources. Le contexte d'exercice a bien changé depuis les cadres pensés par Otlet[6]… Par ailleurs, les filières traditionnelles des contenus (la presse, l'audiovisuel, le livre) s'entremêlent désormais. On voit que, sur le web, les radios font de la télé, tout le monde fait de la vidéo, tout le monde fait du réseau social. Et, forcément, le métier est aussi exposé à cette nouvelle donne.

En entreprise, le centre d'intérêt s'est déplacé sur le *reporting* et la valorisation des données internes. Est-ce qu'on est toujours dans le champ de l'information ou pas ? Pour moi, c'est toujours de l'information, des sortes de documents internes, mais sous forme de data. Ces données doivent aider à la décision, à la prévision. C'est un prolongement de l'activité de veille, mais de la veille centrée sur ses propres données. Donc nous ne sommes pas dans des transformations radicales, mais dans des évolutions profondes parce que, si on veut traiter ces données, il faut avoir une culture statistique. Il n'est pas nécessaire de devenir statisticien, mais il faut savoir

manipuler des algorithmes de base pour cartographier un ensemble de données, pour faire de la clusterisation[7], pour faire un peu d'anticipation, un peu de prévision, des régressions[8]. Les stats, les maths et l'algorithmie deviennent importantes dans le traitement de l'information, je pense que ça fait partie des techniques, des technologies qu'il faut désormais avoir dans son bagage.

C. S. : Les maths, ça fait toujours peur quand on n'est pas matheux. Mais est-ce qu'il faut un esprit mathématique ou un esprit logique, c'est-à-dire pouvoir déterminer : « si je veux ça, alors, je dois regrouper mes données comme ça » ?

Gh. C. : Il faut avoir une culture statistique, savoir faire jouer des algorithmes. Il y a dix, douze algorithmes traditionnels qu'on enseigne à MéDAS. L'idée, ce n'est pas de les programmer, c'est de mobiliser des services numériques qui mettent en marche ce genre de traitement sur des données : regrouper les données en voisins proches, faire une régression linéaire pour cibler des variables discriminantes… Ce sont des traitements que les statisticiens de l'école de Benzécri[9] ont appliqués aux sciences humaines dans les années 1970. Mais je pense que ça redevient une actualité majeure, parce qu'on a de plus en plus de données factuelles. Un professionnel de l'information ne peut pas se dire : « d'un côté, il y a des données de documents structurés et puis de l'autre côté… eh bien ! ça je ne sais pas le traiter, parce que c'est trop statistique ».

C. S. : Et on en a besoin aussi pour l'analyse des documents, puisque, de plus en plus, on approche les contenus par l'analyse de termes. Cela arrive aussi quand on pratique la veille : on ne lit pas les documents un par un.

Gh. C. : Oui, pour faire des clusters, des algorithmes sont mobilisés, même si c'est un peu une boîte noire. Ensuite, la qualité des données fait la différence, il n'y a pas de mystère. Par ailleurs, il y a un monde entre traiter des données, des *smart data* dirons-nous, et puis les *big data*[10], qui nécessitent d'ajuster des modèles. Mais on change de domaine, on arrive dans la data science.

C. S. : Justement, comment définis-tu les différents « data métiers » ? Où nous situons-nous, professionnels de l'information ?

Gh. C. : Les *data managers,* les *data analysts*, les *data stewards* concernent nos compétences, parce qu'ils travaillent sur le *sourcing*, la restitution à l'usager et la médiation avec les métiers de l'entreprise. Le *data steward* a une vision globale des métiers, sait recenser les données d'une entreprise et dialoguer avec les métiers autour de leurs données. Ensuite il y a la data science, mais il nous est difficile de jouer dans la cour des *data scientists*.

C. S. : Qui sont les *data scientists ?*

Gh. C. : Ce sont des profils qui sont capables, pour analyser des flots de données, d'affiner des modèles, qui ne sont pas les modèles traditionnels d'analyse statistique descriptive que je citais précédemment. Eux vont mobiliser des réseaux de neurones par exemple, avec des modélisations statistiques complexes à X paramètres, avec des systèmes multicouches pour essayer de trouver des régularités. Ils font de l'apprentissage sur des jeux de données pour parfaire un modèle statistique de ces données. Mais nos élèves de MéDAS n'entrent pas dans cette cour-là, car il faut suivre des master d'écoles d'ingénieurs si l'on vise ces compétences. Par contre, il y a par ailleurs toute une frange de valeurs ajoutées sur la data, parce qu'avant de déployer un traitement mathématique, statistique de haut niveau, il faut pouvoir disposer d'un matériau de qualité. Et cela prend 80 % du temps. Il faut gérer et documenter les

données, faire en sorte qu'il n'y ait pas de données manquantes, ni de données erronées. Il faut savoir collecter et agréger des données de sources différentes et les aligner. Il y a beaucoup d'enjeux sur ce segment pour le moment. Peut-être que dans dix ans les chaînes de production de données seront alignées, mais on n'en est pas encore là.

En ce qui concerne l'analyse, il y a aussi le *reporting* avec des données qui généralement ne sont pas massives, mais sont issues de CRM[11] ou de logs par exemple, et sont analysées avec des statistiques descriptives, sans viser forcément un modèle de prévision. Nos métiers traditionnels du document ont intérêt à se positionner sur ce premier échelon pour prendre en charge la normalisation des données au sein de l'entreprise, leur organisation, les alignements, le traitement des données manquantes. Les référentiels de données, ce qu'on appelle les « données maîtres », et les référentiels de lexiques ressemblent beaucoup à ce qu'on connaît en documentation. Le requêtage en SQL[12] sur des bases de données également. Concernant l'agrégation de données, il faut *a minima* mobiliser des outils comme Excel avancé. C'est la raison pour laquelle les formations de documentalistes doivent intégrer les fonctions avancées d'Excel.

L'échelon suivant concerne la maîtrise de logiciels comme PowerBI qui permet de récupérer des données de différentes sources, de les nettoyer et de faire des visualisations dynamiques. C'est un continuum de pratiques et tout le monde n'ira pas jusque-là, mais beaucoup ont envie d'y aller. Parce qu'elles ont des données à traiter, les organisations créent des départements transversaux où elles positionnent tout sans se préoccuper de savoir si c'est le même métier d'origine. Et finalement, j'ai envie de dire qu'il ne faut pas parler de métier mais parler de compétences. Plus on formera des compétences pour couvrir l'ensemble du spectre, plus nos élèves seront mobilisables et employables.

C. S. : Et dans la formation en ingénierie documentaire, notre formation traditionnelle, est-ce que des élèves ont des velléités d'aller plus près du travail de *data manager* ?

Gh. C. : Pour le moment c'est surtout la data visualisation et l'*open data* qui les intéressent plutôt qu'être *data manager* ou *chief data officer* (ce sont des termes à la mode, en synergie avec la fonction générique de « gouvernance de la donnée »). Cela commence un peu, c'est vrai, parce qu'ils sont souvent conduits à ne pas traiter que du document, mais aussi de la donnée. L'exemple type c'est le domaine de la recherche où les documentalistes scientifiques se positionnent sur les données de la recherche, pour faire des plans de gestion. Il faut documenter les données de la recherche et le plan de gestion des données consiste à faire un document qui accompagne le stockage des données, les décrit pour permettre leur réutilisation et explique tout le contexte de leur prélèvement, comment elles sont construites, dans quel format. Ils documentent la donnée. L'INIST[13] forme au plan de gestion de données. Toutes les bibliothèques veulent se positionner comme des gestionnaires des données de la recherche. Après avoir fait des archives ouvertes, c'est le prochain chantier. Au niveau national, l'ouverture d'une plateforme nationale, sur le modèle de HAL, mais dédié aux données de la recherche, vient d'être annoncée. Il s'agit de *recherche.data.gouv*. Cela dit, c'est encore une forme de centralisation à la française et il faut voir si c'est réaliste ou pas, parce que les communautés ont déjà leurs habitudes. Quoiqu'il en soit, on positionne les bibliothécaires et les conservateurs sur

la gestion des données de la recherche et on leur donne une mission de qualification et de documentation pour assurer la qualité et la réutilisabilité de ces données.

C. S. : Donc tu dirais que dans le domaine de la recherche, il y a un intérêt pour les données. Et ailleurs, de façon transversale dans le monde de la gestion de l'information, on y arrivera en passant d'abord par la datavisualisation. C'est comme ça qu'on va arriver à progressivement prendre en compte les données et développer ses compétences autour des données ?

Gh. C. : Nos métiers sont proches des publics. Or, les publics veulent de la restitution, il est donc logique que la dataviz[14] apparaisse comme le premier enjeu auquel il faut répondre. Avec Béatrice Arrabuerrana[15], nous avons créé au CNAM, en 2019, une unité d'enseignement de « Datavisualisation pour tous ». Il s'agit de démocratiser l'accès à ces techniques de dataviz. S'occuper de la restitution de la data est une entrée importante pour l'appropriation du champ de la donnée pour les documentalistes. C'est très proche des demandes des usagers, car on est de plus en plus dans des modes de communication visuelle voire performative. Il est juste de dire que c'est certainement par la dimension diffusion et restitution que les métiers s'ancrent progressivement dans la data. Après, il y a aussi le *data mining*, la fouille de données, ce sont des fonctionnalités qu'on voit apparaître, notamment dans les activités de veille.

C. S. : Notre cerveau a du mal à analyser tous ces flux de données. Est-ce que cela concerne les données ou plutôt les algorithmes ? On bascule peut-être plutôt vers des outils du type algorithme.

Gh. C. : C'est sûr. Nous assistons au déploiement d'une économie de plus en plus assistée par l'intelligence artificielle [IA] qui se nourrit de données. Nos métiers sont en prise avec cette tendance. On tend effectivement, aujourd'hui, à intégrer de l'IA dans tous les secteurs et l'IA a besoin de données, de données propres. Au final, on est en prise aussi avec ces mouvements de fond, même si on ne s'en rend pas toujours compte.

C. S. : Est-ce qu'on peut dire que la donnée, par rapport aux documents, ce serait une information qu'on va traiter par le biais d'une machine et non pas directement par l'humain ?

Gh. C. : Ce n'est pas faux. Jusqu'à présent, on se nourrissait de documents et on restituait notamment des livrables sous forme de synthèses par exemple. Aujourd'hui, on est fortement assisté par la machine, à cause de la volumétrie et de la rapidité, et cela nous conduit à travailler de plus en plus à la qualité des données pour nourrir des algorithmes, qui nous aideront à restituer et dont nous devrons contrôler les productions. Je pense que la valeur ajoutée de nos métiers se déplace de plus en plus vers la garantie de la qualité des données en amont et la vérification et l'analyse critique de ce qui est produit au final.

C. S. : Cela nous ramène à la question des compétences. Nous avons parlé des compétences mathématiques, des connaissances de la statistique. Alors, quelles sont les compétences *hard*, les compétences vraiment métier, et puis les compétences intellectuelles, les *soft skills* ? Est-ce qu'il faut toujours être rigoureux, organisé et curieux pour gérer l'information ?

Gh. C. : Je pense que les fondamentaux restent. C'est-à-dire la rigueur, la qualité, l'objectivité, l'ouverture. Nous sommes garants du matériau sur lequel d'autres vont pouvoir raisonner ensuite. Pour moi, cette dimension est permanente. C'est pour ça que j'étais un peu énervée de voir parfois des moteurs de recommandation dans des bibliothèques tenir compte prioritairement du fait qu'une information soit ouverte

pour apprécier la qualité des contenus : « C'est ouvert, c'est fermé, c'est bon, c'est pas bon… ». Nous devons rester les garants de la qualité et de la diversité. Ensuite, pour exercer, il faut des compétences qui suivent le cercle de traitement de l'information : la collecte, le traitement, l'analyse, la conservation. Chaque phase renvoie à des compétences particulières qui évoluent avec le contexte des documents structurés, non structurés, etc. Là se situent des compétences techniques plus fortes. La compréhension des techniques informatiques, toujours évolutives, est un enjeu à souligner. Aujourd'hui, comme hier, il faut faire du requêtage, du nettoyage de données avec de nouveaux outils (pas tout à fait ceux d'hier), faire de la restitution avec des outils diversifiés et peut-être aller jusqu'à savoir actionner des intelligences artificielles pour produire du résumé automatique, en conservant un recul critique.

Mais, il faut aussi continuer à développer une compétence juridique. La data et son traitement amplifient, mine de rien, la question de ce qu'on peut faire, de ce qu'on ne peut pas faire. Beaucoup de règlements sur l'IA et les données sont en cours d'élaboration au niveau européen. Tout cela cadre nos métiers et détermine ce qu'on peut faire et ce qu'on ne peut pas faire. Il faut rester en alerte sur tous ces cadres juridiques.

Le numérique déplace de toute façon nos métiers vers l'analyse parce qu'il y a des formes d'automatisation de plus en plus fortes. Les métadonnées, par exemple, sont de plus en plus encapsulées dans les contenus, on n'a donc plus forcément besoin de produire de l'information secondaire puisqu'on va récupérer ces métadonnées par une chaîne de traitement qui l'aura prévu à la source. Notre valeur ajoutée va donc se déplacer vers l'analyse, et c'est tant mieux. On voit, ces derniers temps, de plus en plus d'annonces de postes de veille, ce qui est le signe qu'on n'arrive plus à absorber tout ce qui se diffuse. Et je pense que tous ces métiers qui restituent l'information essentielle, qui éclairent finalement d'autres métiers, continuent à avoir du sens, même si l'on prédit qu'ils seront remplacés par des algorithmes. Pour l'instant, les algorithmes ne sont que des aides, car l'automatisation n'est pas complète et ses résultats ne sont pas toujours bons.

C. S. : « On va faire disparaître les professionnels et on les remplacera par des outils », c'est un serpent de mer. Mais les professionnels trouvent continuellement de nouvelles tâches, en montant d'un cran à chaque fois dans le contrôle et la gestion. Ils finissent par contrôler et gérer les outils qui gèrent ce qu'ils géraient manuellement avant. Il y a toujours, au final, une activité à développer.

Gh. C. : Le contrôle qualité effectivement. Il y a longtemps que les documentalistes de Medline, la grande base de médecine emblématique du métier de documentaliste scientifique, testent des indexations automatiques. Maintenant c'est en production et le rôle des indexeurs est de vérifier ce qui sort et de recadrer.

C. S. : De nourrir les algorithmes…

Gh. C. : Puis de contrôler, parce qu'il y a des choses mal faites, et de corriger, sans cesse corriger, voire enrichir quand l'algorithme n'est pas allé jusqu'au bout. Je crois qu'effectivement on monte à un niveau de contrôle, de qualité, d'enrichissement. C'est peut-être transitoire…

C. S. : Je crois que ce n'est pas définitif. Pourquoi serait-on arrivé au bout de l'évolution du métier ? J'ai une dernière question sur la mémoire des données. Tu as parlé des plans de

gestion de données pour la conservation pérenne des corpus scientifiques. Mais quelle mémoire donne-t-on aux données d'activité des entreprises ?

Gh. C. : Le problème aujourd'hui, c'est qu'on garde trop de données. On ne se pose pas la question de leur valeur. L'Europe veut construire son nuage de données scientifiques et les chercheurs reçoivent des injonctions à déposer et garder toutes leurs données, au risque de fabriquer des cimetières de données. Bien sûr, si on oblige les chercheurs à déposer, en lien avec leurs articles, des données empiriques de valeur, cela a plus de sens. Mais c'est une extension de la fonction éditoriale qui veille à l'intégrité de l'article. Pour les entreprises, faire du *record management* de toutes les données d'activités semble déraisonnable et elles s'orientent plus vers les « lacs de données », sans chercher à sélectionner quitte à partir ensuite à la pêche selon les besoins.

C. S. : Si on veut pouvoir retracer l'activité d'une entreprise, une décision majeure par exemple, on peut imaginer qu'on aurait envie de conserver la datavisualisation, les cartographies ou les tableaux que l'on remet aux décideurs, parce que ce sont les traces d'un processus.

Gh. C. : Pour le contexte des données, je pense que chaque cas est spécifique. Et les données, ça ne suffit pas, il faut aussi faire des *data book* pour les sauvegarder. C'est ce qu'Anna Nesvijevskaia[16] a développé dans sa thèse : il faut documenter les états, il n'y a pas que les données, il y a tout ce qui accompagne le traitement de la donnée qu'il faut aussi documenter. Mais je pense aussi qu'il y a des logiques informatiques qui viennent surplomber tout cela. On peut très bien se dire « on s'en fiche, on garde tout » comme je disais précédemment. Cette inflation de *data centers* n'est pas bonne pour la planète, mais c'est la philosophie du *data lake*, faire du requêtage pour retrouver les données qui éventuellement répondront à des usages qu'on ne peut pas anticiper.

Le Système d'information de l'entreprise, l'ERP[17], fonctionne en temps réel et écrase les données, alors que les structures techniques appelées *datawarehouse*, elles, gardent l'historique des données. Ensuite, avec le *machine learning*, on peut faire de la prévision sur les ventes, voir ce qui a marché, ce qui n'a pas marché. La data science est donc en train de redonner du sens au stockage des données et à l'historisation des données en entreprise. Mais, le plus souvent, on utilise les données des trois dernières années pour formuler des prédictions à N + 1.

C. S. : Et finalement la mémoire peut servir à la prédiction.

Gh. C. : C'est une première réponse à l'historisation des données produites par les entreprises. Mais, à un moment donné, on ne va pas pouvoir tout garder. On va avoir des cadrages différents en fonction du contexte et des usages. Et tout va être fonction de ce qui est, encore une fois, lié au risque. On archivera en fonction du risque et du recours possible.

NOTES

1. Association des professionnels de l'information et de la documentation.
2. Système de gestion de base de données.
3. Système intégré de gestion de bibliothèque.
4. Gestion électronique des documents.
5. *Enterprise Resource Planning* (progiciel de gestion intégré).
6. Paul Otlet (1868-1944), fondateur de l'Office international de bibliographie, a posé les bases et les concepts des méthodes documentaires dans le *Traité de documentation*, éditions du Mundaneum, 1934.
7. Méthode d'analyse statistique consistant à regrouper les données en grappes selon des caractéristiques communes.
8. L'analyse de régression linéaire sert à prévoir la valeur d'une variable en fonction de la valeur d'une autre variable. Les modèles de régression sont utilisés pour l'apprentissage automatique (*machine learning*).
9. Jean-Paul Benzécri (1932-2019) est un mathématicien et statisticien français, professeur à l'Institut de statistique de l'université de Paris [ISUP] et à l'université de Rennes dans les années 1960.
10. Le traitement des *smart data* (données intelligentes) consiste à extraire rapidement les données stratégiques les plus pertinentes parmi les volumes de données produites par l'entreprise. Il se distingue du traitement des *big data* (données massives ou mégadonnées) dans lequel on infère des modèles prédictifs à partir de masses de données.
11. Customer Relationship Management ou gestion de la relation client. Les logiciels de CRM recueillent et stockent les informations sur les clients.
12. *Structured Query Language.*
13. Institut de l'information scientifique et technique. Unité d'appui à la recherche du CNRS dans le domaine de l'accès à l'information scientifique et la valorisation des données de la recherche.
14. Dataviz (ou data visualisation) désigne l'ensemble des méthodes permettant d'aider à la compréhension et à l'interprétation des données en les présentant de façon visuelle.
15. Béatrice Arrabuerrana est enseignante-chercheuse au laboratoire Dicen-IDF du Cnam. Son expertise porte sur l'exploitation scientifique et la visualisation des Data.
16. Anna Nesvijevskaia est data-scientiste et directrice associée Quinten Finance. Sa thèse, soutenue en 2019, a pour titre *Phénomène Big Data en entreprise ; processus projet, génération de valeur et Médiation Homme-Données*.
17. *Enterprise resource planning* ou progiciel de gestion intégré [PGI].

AUTEUR

GHISLAINE CHARTRON

Professeur au Cnam et directrice de l'Intd

Les archives des objets : quelle gestion des traces pour l'internet des objets ?

Entretien avec Béa Arruabarrena et Armen Khatchatourov, propos recueillis par Claire Scopsi

Béa Arruabarrena et Armen Khatchatourov

Introduction

Ce texte rend compte du dialogue organisé entre Béa Arruabarrena et Armen Khatchatourov, tous deux chercheurs en sciences de l'information et spécialistes des technologies connectées, dans le cadre du séminaire Les nouveaux paradigmes de l'archive[1]. Nous leur avons demandé si archiver les données produites par des objets connectés a un sens et s'il faut que les archivistes se préparent à accueillir ces flux de données dans leurs collections. Tous deux ont confronté leurs approches respectives, la sociologie de la quantification de soi pour l'une, la philosophie des techniques pour l'autre, afin de poser les enjeux d'une conservation des traces produites par les objets. Ce compte rendu suit la progression de leur échange. Tout d'abord une présentation de leurs objets d'étude : Béa Arruabarrena étudie particulièrement les pratiques de mesure de soi dans le domaine de la santé ainsi que les pratiques de mesure citoyenne ; Armen Khatchatourov s'interroge sur la possibilité même de la constitution de l'archive, prérequis du savoir, dans les nouvelles conditions du flux perpétuel des données.

Dans un second temps, nous les avons questionnés sur le risque de l'accroissement des injonctions normatives induit par la circulation et le traitement des données de santé. Les deux chercheurs s'accordent pour dire que les faits observés révèlent une forme nouvelle de normativité qui n'est pas seulement disciplinaire, au sens de Foucault, mais qui s'exerce désormais de manière plus complexe.

La discussion s'achève par un exercice de prospective : quelles seront les formes et les fonctions de l'archivage des données issues des objets connectés ? Si les pratiques

émergentes sont motivées par des besoins de preuve ou de contrôle sur les travailleurs par exemple, comment, alors, les individus pourront-ils résister à ces mouvements ? Faut-il explorer la piste d'un archivage citoyen, qui conserverait non seulement les données, mais aussi les contextes de leur production ? L'enjeu de l'archivage des *data* est donc de trouver le moyen de les associer à des formes discursives qui leur donnent du sens.

Approches sociologiques et philosophiques des flux des données

Posons d'abord les contextes de la production de données par des objets connectés en les considérant sous l'angle de leur permanence. En 2022, l'archivage systématique des données des *big data* pose question, ne serait-ce que pour des considérations techniques de capacité de stockage, et tout au plus conserve-t-on les données de mesure le temps de les analyser et d'en dégager des prédictions. Si les débuts des *big data* étaient marqués par l'utopie d'une conservation exhaustive de données toujours plus volumineuses, désormais les *data* se caractérisent par le flux et la connectivité. Au service d'un objectif précis, elles sont vouées à circuler pour alimenter des processus, à se fondre dans d'autres données et à disparaître. Cette fluence complexifie la production de savoirs critiques, car elle ne permet pas de matérialiser les rapports de force et les connaissances qui sous-tendent ces données.

La quantification de soi ou comment les objets connectés agissent sur les personnes

Les travaux de Béa Arruabarrena portent sur les pratiques de mesure par les objets connectés. Elle a abordé en particulier la quantification de soi en analysant le mouvement *Quantified Self*[2] qui a émergé en France au milieu des années 2000 et qui s'inscrit dans un mouvement plus large apparu aux États-Unis dans la même période. Ce mouvement rassemble des praticiens, qui mesurent[3] des paramètres de santé à l'aide d'objets connectés. Certaines expériences à visée plus sensationnelle[4], notamment portées par des artistes, ont donné parfois au mouvement une image sulfureuse, mais il a peu à peu évolué pour converger vers une industrialisation des outils et des pratiques dans le secteur de la santé connectée, en médecine libérale ou à l'hôpital, lorsqu'il s'agit, par exemple, de mesurer le poids, la tension artérielle ou de suivre un patient diabétique. On peut considérer qu'il s'agit à la fois d'une opportunité en termes de nouveaux savoirs pour les praticiens, mais que, dans le même temps, ces nouveaux outils constituent un moyen inédit de captation de données avec les risques que cela comporte sur le plan éthique. La métrologie citoyenne, par exemple pour la mesure de la qualité de l'air, est une autre manifestation de l'émergence des capteurs et des objets connectés qui peut constituer, pour les citoyens, une appropriation nouvelle des données environnementales afin de participer à l'enrichissement des connaissances scientifiques et de peser sur la décision publique.

Les enjeux de la protection des données personnelles liées à ces usages sont bien identifiés et encadrés par le RGPD[5]. En revanche, la protection des personnes, par exemple, qui relève de la bioéthique, reste un domaine encore peu évalué. Or, même si elles présentent certaines problématiques communes, la protection des données et la

protection des personnes diffèrent (Khatchatourov, 2019b). Un objet connecté, même s'il répond à tous les critères de protection des données, peut ainsi nuire « à la personne humaine » parce qu'il intervient dans son comportement ou a un fort impact sur sa vie quotidienne. Ce caractère interventionnel du dispositif connecté est aujourd'hui renforcé par le « design persuasif » et/ou « la captologie » qui relèvent tous d'une forme de design comportemental très répandu dans la conception des dispositifs numériques de mesure. En effet, aujourd'hui, l'enjeu n'est pas simplement de capter des données, mais aussi de les restituer à l'usager, afin de déclencher un nouveau comportement parfois à des fins nobles comme en médecine, mais aussi à des fins commerciales pour inciter à l'achat ou influencer un vote politique comme dans l'affaire Cambridge Analytica[6] par exemple.

Le traitement des données récoltées ne se comprend que lorsqu'on le place dans une perspective temporelle. C'est l'historisation des données comportementales qui permet d'établir des modèles pour prédire les comportements à venir, voire, dans une étape ultérieure, prescrire des comportements. L'historisation est donc un principe fondamental de la quantification et de la mesure : « mesurer, c'est comparer ». Une donnée isolée n'a pas de valeur, c'est en la comparant dans le temps que l'on va pouvoir prédire si un patient est susceptible de grossir, si sa tension va augmenter, etc. À cela s'ajoute le fait qu'aujourd'hui les technologies permettent de faire cela en temps réel. Ces méthodes sont aussi pratiquées dans le monde de la banque ou de l'assurance, pour déterminer si un particulier sera un « bon ou un mauvais client », ou encore calculer un taux d'attrition[7], qui va déclencher des campagnes de publicité ou de communication ciblée pour modifier le comportement du client et l'inciter à rester fidèle au service. Finalement, on peut avancer que l'enjeu de la quantification, ce n'est pas seulement le volume des données disponibles, mais aussi le sens qu'on peut leur donner en analysant « leur histoire », c'est-à-dire dans leur contexte temporel pour agir dans le présent.

La place des objets connectés : entre archive et flux

C'est avec les postures conjointes de la philosophie de Michel Foucault et de Gilles Deleuze et des sciences de l'information et de la communication qu'Armen Khatchatourov analyse le lien entre l'archive et le flux, et la place qu'occupe l'objet connecté dans ce paysage. Il faut d'emblée considérer que l'objet connecté est composé de capteurs, d'algorithmes et éventuellement d'une interface qui interagit avec l'utilisateur, mais ce qui prime sur l'interface ou la forme physique de l'objet, c'est sa connectivité. Cette connectivité rejoue la constitution des savoirs qui découle éventuellement de l'analyse des données ainsi obtenues.

À ce sujet, la distinction que fait Deleuze, à la suite de Foucault, entre l'archive et le diagramme permet de comprendre ce qui relève de l'ordre du savoir et ce qui relève de l'ordre du pouvoir. Le savoir est constitué de strates, des archives sédimentées sous-tendues par des lignes de force et de pouvoir que l'on appelle dans ce cadre « diagramme ». Le diagramme le plus connu est le diagramme panoptique[8], qu'il faut comprendre non pas comme une forme matérielle du bâtiment, mais comme une fonction abstraite, celle du regard omniprésent. Le bâtiment dans lequel le prisonnier ne peut échapper au regard du gardien est une forme particulière de cette surveillance, ce n'est que la traduction matérielle du diagramme panoptique et des forces qui s'y exercent.

Il y a interdépendance entre l'archive et le diagramme. Le savoir constitué, l'archive, n'existe que par les rapports de force évanescents et fluents qui le sous-tendent, tandis que les rapports de force ne se prêtent à notre regard, à notre investigation qu'en tant qu'ils sont actualisés, visibles dans une sédimentation qui est l'archive. Ce rapport de réciprocité entre l'archive et le diagramme, appelons-le *vis-à-vis* (Khatchatourov, 2016). Il est la condition du savoir, car il permet, grâce à l'archive, de porter un regard critique sur le pouvoir. C'est quand on considère l'archive et que l'on regarde comment elle est constituée que l'on comprend les rapports de force qui l'ont rendue possible.

> ...le savoir est affaire d'archives, l'archive étant audiovisuelle, c'est-à-dire archive du voir et archive de l'énoncé, archive du visible à chaque époque, archive de l'énonçable à chaque époque. Le savoir est affaire d'archive, le pouvoir n'est pas affaire d'archive. Le pouvoir est affaire d'une cartographie, d'une cartographie mouvante, une carte stratégique, toujours remaniable, toujours fluente[9]. (Deleuze, 1986).

Les développements récents du phénomène *big data* opèrent un glissement de ce cadre de pensée. Le principe des données volumineuses est de capter et d'analyser des flux, de plus en plus souvent produits par des objets connectés, que ce soient les montres que l'on porte au poignet, les capteurs urbains, les capteurs qui tracent les marchandises d'une *supply chain* ou ceux qui relèvent les données météorologiques. Cette notion de captation a renvoyé, au début du phénomène des *big data,* à l'image de l'entrepôt, le *datawarehouse*, c'est-à-dire à l'idée que l'on pourrait stocker toutes les données pour ensuite les analyser. Ce fantasme d'exhaustivité s'exprime sur deux plans : le plan temporel d'une part, avec l'idée que l'on peut constituer une mémoire dans laquelle on est en droit de revenir sur tout épisode passé, et, d'autre part, le plan de l'extension, dans lequel la connaissance porte, non sur des échantillons représentatifs de données, mais sur une totalité, par exemple, tous les messages échangés dans une population. Dans ce modèle, on conserve des volumes de données sans finalité précise, sans savoir si elles seront analysées. Dans un second temps a émergé le paradigme du *big data stream*, dans lequel l'analyse porte sur le flux même des données, à la volée. Ce modèle correspond bien aux objets connectés, dont les algorithmes traitent les données dès leur *input* et restituent en *output* un résultat immédiat. Dans ce modèle du flux, on ne stocke plus les données, seul le résultat traité par l'algorithme compte.

La dissolution de l'archive ?

L'exemple des capteurs virtuels éclaire le nouveau rapport instauré par le modèle du flux. Dans le milieu industriel, par exemple, des modèles de comportement viennent se substituer aux capteurs réels lorsque ces derniers sont trop coûteux ou trop complexes à mettre en œuvre. Les données calculées à partir de ces modèles circulent et agissent sur les *process* exactement comme si elles avaient été captées par un appareil de mesure[10]. On aboutit donc à une mise en équivalence entre l'événement réel capté par un objet connecté (le capteur physique) et l'événement purement logiciel au sein des *big data stream,* mais cette mise en équivalence n'est pas visible pour l'utilisateur final qui n'est pas concerné par le mécanisme qui produit ces données. Celles-ci rejoignent la chaîne de traitement et sont transmises en entrée à des processus qui ignorent la manière dont elles ont été produites. On voit que dans cette datafication du monde, la question de la conservation exhaustive ou non des données, tant débattue lors de la

première période de *big data warehouse*, s'efface devant la recherche d'une solution juste satisfaisante pour atteindre le but fixé.

Armen Khatchatourov voit dans le *big data stream* et les flux de données des objets connectés un nouveau type de rapport entre l'archive et le diagramme. Il n'y a pas à proprement parler une disparition de l'archive, mais une démission, car le pouvoir peut désormais se conserver et se reproduire sans être actualisé par les savoirs et par l'archive. La réciprocité de l'archive et du diagramme inhérente au vis-à-vis fait place à une forme de soumission. La relation de l'individu à la norme, dans le contexte des *big data,* et la façon dont s'exerce sur lui la nouvelle pression normative – par exemple celle de « ne pas se conformer aux modèles », d'être « fluide » et « multiple » – est le corollaire de cette nouvelle relation entre l'archive et le diagramme (Khatchatourov, 2019a).

Enjeux des objets connectés : normativité et gouvernementalité néo-libérale

Les mesures de santé démultipliées par les objets connectés et publiées sur les réseaux sociaux peuvent-elles conduire à une société toujours plus normative ? Apparemment, les individus jouissent d'une certaine souplesse à l'égard des valeurs de référence, tout en étant incités à ne pas excéder certaines règles. Cette nouvelle forme de contrôle, qui agit en responsabilisant plutôt qu'en contraignant, révèle une nouvelle forme de gouvernementalité.

Injonctions normatives ou normopathie ?

La question de la comparaison des données peut nous aider à cerner les lignes de force en jeu dans les objets connectés. L'historisation permet de comparer, dans le temps et dans l'espace, des données issues de mesures et les prédictions qui en ressortent peuvent être assimilées à des injonctions normatives. Béa Arruabarrena souligne que la quantification peut déboucher sur un risque de « normopathie » (Rouvroy, 2010)[11] où les normes de bonne santé seraient définies selon les critères de performance et d'efficacité fixés par des objets connectés et des algorithmes. La plupart des applications de *quantified self* intègrent une composante de partage et de comparaison *via* une fonction de réseau social, qui peut conduire à la sanction sociale. Pourtant, dans ses enquêtes, elle n'observe cet excès de normativité que dans le monde sportif et la recherche de performances physiques. Dans les pratiques liées à la santé, elle constate plutôt que chacun trouve sa propre norme, en prenant des libertés avec les référentiels proposés. Ainsi en va-t-il pour les applications de comptage de pas. L'Organisation mondiale de la santé recommande d'effectuer 10 000 pas par jour, mais un individu sédentaire s'aperçoit vite qu'il ne peut atteindre ce nombre. En accord avec son médecin ou en discutant, il décidera que 5 000 pas, effectués chaque jour très régulièrement, constituent une norme personnalisée qui lui permet quand même d'agir sur sa santé. La norme commune disparaît alors au profit d'une norme locale. Comme le constate Antoinette Rouvroy[12], cet exemple est une manifestation très simple d'un phénomène plus global d'évolution de la normativité, caractérisé par une place plus importante accordée aux pratiques individuelles et un affaiblissement de la norme collective, avec la possible perte de référentiel commun.

Normativité et gouvernementalité chez Foucault

Dans la société disciplinaire décrite par Foucault, qu'il situe entre les XVII^e^ et XX^e^ siècles, et, pour le dire très schématiquement, le rapport à la norme est un rapport d'acceptation ou d'opposition. Il est tout à fait possible de s'opposer à la norme et cette opposition, franche et frontale, reçoit en retour une sanction. Dans les années 1970 ou 1980, donc avant le numérique, il s'instaure, en Occident, un autre rapport entre l'individu et la norme qui relève, selon le terme de Foucault, de la gouvernementalité néo-libérale. Le comportement des individus n'est plus édicté strictement, mais de façon mouvante, dans un cadre souple qui autorise les oscillations individuelles. La disparition de la norme collective est alors vécue comme une liberté individuelle et conduit à une juxtaposition de normativités, de prises de position locales, tandis que la résistance frontale n'existe plus. Gilles Deleuze décrit cet état comme un moule aux bords déformables, mais bien présents, car, à l'intérieur, l'individu ne jouit que d'une impression de liberté. Le contrôle s'exerce toujours, non en interdisant le franchissement des frontières du moule, mais en autorisant certains parcours à l'intérieur des bords mouvants. Cela se traduit par des effets de persuasion sur les utilisateurs du *quantified self*. Ils sont, par exemple, encouragés à contrôler leur poids, mais à l'intérieur d'une plage de valeurs déterminées, dans laquelle la production de discours et de normes personnelles est valorisée. Cela explique la prolifération de l'expression de soi dans les réseaux : chacun revendique sa liberté en exposant sa différence, mais tous sont finalement semblables dans leur quête d'individualité.

Responsabilité et usages citoyens des objets connectés

Dans la gouvernementalité néo-libérale, la responsabilité de la santé glisse du collectif au particulier ; se mesurer, s'auto-discipliner devient une responsabilité individuelle où chacun est garant de sa santé. Ce sont donc des forces politiques qui s'exercent sur l'individu à travers les objets connectés.

Mais cette situation n'est pas nécessairement une fatalité. Maîtriser ses données et construire la norme peut conduire les citoyens à bousculer le moule et ses bornes normatives sous-jacentes, pour peu qu'on leur donne l'opportunité d'échapper à la circulation permanente de données qui n'autorise aucun recul. Paradoxalement, alors que dans les années 2010 à 2018, les *big data* étaient critiquées pour l'excessif volume de leurs traces, il faut désormais renouer avec l'archive pour recréer du vis-à-vis et la possibilité d'une distance critique. Mais alors que l'archivage des *big data stream* est encore à peine envisagé, on peut douter que l'objectif des premières initiatives soit de développer l'agentivité citoyenne.

Enjeux de la conservation des données de santé

Il est difficile d'envisager en 2022 ce que pourrait-être l'archive des flux des données, tant elle est négligée par les acteurs des objets connectés. En santé, pourrait-elle prendre la forme d'un double numérique de soi, comme il existe des doubles numériques des infrastructures industrielles ? Mais ce sont les logiques commerciales, légales ou managériales, sous des formes aliénantes, qui semblent être les principaux

moteurs. Comment les individus pourront-ils agir pour s'approprier les données archivées ?

Les usages au titre de la preuve

L'enjeu commercial de la conservation des données de santé apparaît plus clairement dans le secteur de l'assurance. Conserver sur le long terme les données de comportement, par exemple le fait d'être fumeur, permet d'y revenir pour analyser un état de santé ultérieur et déterminer les responsabilités. C'est un des leviers qui motive, depuis peu, à opérer des conservations.

La question de la preuve peut être illustrée par l'affaire Ross Compton, un citoyen américain équipé d'un *pacemaker* connecté, dont le domicile a subi un incendie, alors qu'il prétend en avoir été absent. Lorsque Compton a sollicité son assureur, ce dernier a pu démontrer, en analysant les données géo-localisées, produites par son appareil cardiaque et conservées[13], qu'il mentait et avait bien été présent sur le lieu de l'incendie. Ici, on voit les ambigüités de la conservation systématique des données et de leur réutilisation dans des finalités différentes. L'affaire s'est déroulée dans un État où cette utilisation est légale, car il n'y a pas encore d'harmonisation des législations sur les niveaux et les durées de conservation de telles données.

Par ailleurs, il est évident que les données de santé massives – et, espère-t-on, anonymisées – sont déjà conservées pour la recherche médicale, car collecter un ensemble de symptômes conduit à une meilleure connaissance des maladies ou des effets secondaires des substances. C'est donc un enjeu politique et social que de déterminer, non pas s'il faut conserver et exploiter les données issues des objets connectés dans le cadre d'enquêtes, de connaissances scientifiques ou d'intérêt public – car cette pratique semble devoir de développer inéluctablement –, mais de réguler cet usage, d'en fixer les limites et d'en sanctionner les abus.

Les usages managériaux des données de santé

La plupart des utilisations des objets connectés sont présentées comme bénéfiques : la voiture connectée ne démarre qu'après le contrôle de l'alcoolémie du chauffeur, les vêtements connectés surveillent le rythme cardiaque et la température corporelle du travailleur soumis à un environnement extrême, la chaise connectée signale au salarié une mauvaise position afin de préserver son dos.

Mais les objets connectés participent au contrôle que subissent les travailleurs soumis à des surveillances numériques. Pour Ifeoma Ajunwa, professeur à l'université de Caroline du Nord, se dessine un monde où sont contrôlés par des algorithmes non seulement les ordinateurs des salariés, mais aussi leurs postures, leur regard et la manifestation de leurs émotions. Selon elle, dans la nouvelle logique taylorienne, le travailleur quantifié (« *quantified worker* ») est confronté à un management mécanique (« *mechanical management* ») dans lequel les activités de recrutement, d'encadrement et d'évaluation des travailleurs sont de plus en plus déléguées aux technologies d'intelligence artificielle[14].

La législation européenne protège le salarié et l'objectif doit être l'évaluation de l'aptitude du travailleur à effectuer ses tâches pour le protéger d'un accident. Dans cette logique, le port d'objets collecteurs de données de santé devrait être proposé par

le médecin du travail et rester soumis à l'accord du salarié, même si l'on peut douter que le travailleur se sente libre de refuser au risque de mettre en péril son emploi. Mais, comme on le sait maintenant, la frontière entre performance et santé est très mince. Cette ambivalence entre, d'une part, la prévention des maladies ou accidents et, d'autre part, un usage plus coercitif des données de santé rend complexe non seulement l'encadrement juridique des pratiques, mais aussi la posture critique des citoyens à l'égard des usages des données par les organisations.

L'importance d'une contextualisation des données collectées

À l'instar de Bernard Stiegler, c'est dans la participation citoyenne que Béa Arruabarrena et Armen Khatchatourov perçoivent une amorce de solution, c'est-à-dire dans l'émergence de formes auto-organisées, capables de s'approprier les flux et de les baliser, les annoter et leur donner une grammaire. Une telle appropriation collective nécessite une réflexion sur l'articulation des différents imaginaires, à construire ensemble. Par exemple, les mesures d'Airparif et celles des collectifs citoyens ne relèvent pas des mêmes imaginaires, même si elles produisent des données dans les deux cas. Capter ces mesures à 5 mètres d'altitude ou chez soi ne signifie pas la même chose : dans le premier cas, on cherche de la donnée scientifique quantitative, dans le second cas, on cherche de la donnée qualitative qui ait du sens localement, c'est-à-dire dans le milieu de vie de ceux qui la produisent.

Cette nécessaire contextualisation vient interroger les modalités d'archivage des données de santé et leur enrichissement par des métadonnées de contexte.

Conclusion : imaginer un nouveau paradigme d'archivage

Nous vivons un moment charnière, celui de la « dissolution de l'archive » (Khatchatourov, 2016), ce savoir sédimenté dont le rôle est de constituer le vis-à-vis du pouvoir, mais que l'intensification des flux de données menace. Nous avons montré que la captation et l'exploitation immédiate de données ne permettent pas de produire de discours critique sur cette activité, même si des organisations citoyennes tentent d'opposer le résultat de leurs propres captations aux données collectées par les organisations officielles, industrielles ou commerciales.

Cependant, la valeur discursive de la donnée, la possibilité de commenter, d'annoter, n'est pas suffisamment développée aujourd'hui. On capte la donnée pour influencer, à travers son traitement algorithmique, des décisions ou des comportements, mais on ne sait pas dans quel contexte cette donnée a été prise, alors que ce contexte en change considérablement le sens. Il faudrait alors produire une donnée enrichie par des discours. Aujourd'hui, l'enjeu autour de cette nouvelle forme de savoir est de produire et de collecter des discours sur les données volumineuses. Il reste à imaginer le paradigme futur de l'archive des données issues des objets connectés.

BIBLIOGRAPHIE

Ifeoma AJUNWA, *The Quantified Worker: Law and Technology in the Modern Workplace*, Cambridge, Cambridge University Press, 2023.

Béa ARRUABARRENA, « Technologie numérique de quantification des corps à l'épreuve du comportementalisme : vers un design de la médiation homme-données », dans *Corps connectés : figures, fragments, discours*, éd. par Armen KHATCHATOUROV, Olaf AVENAT, Isabelle QUEVAL et Pierre-Antoine CHARDEL, Paris, Presses de Mines, 2021.

Béa ARRUABARRENA, « Objets connectés : penser les enjeux des technologies connectées sous l'angle de la médiation infocommunicationnelle », Tic&Société, vol. 15, n^{os} 1-2, 2^{e} semestre 2021 – 1er semestre 2022/1, p. 9-35.

Béa ARRUABARRENA, Anne BERTHINIER-PONCET, Maryse CARMES et Michel LETTÉ, « Les agencements sociométrologiques de la qualité de l'air : les configurations participatives des Tiers-Lieux », *Les Cahiers du numérique,* 2021.

Éric DAGIRAL, Christian LICOPPE, Olivier MARTIN et Anne-Sylvie PHARABOD, « Le *Quantified Self* en question(s). Un état des lieux des travaux de sciences sociales consacrés à l'automesure des individus », *Réseaux*, 4, 2019, p. 17-54.

Armen KHATCHATOUROV, Olaf AVENATI, Isabelle QUEVAL et Pierre-Antoine CHARDEL (dir.), *Corps connectés : figures, fragments, discours,* Paris, Presses de Mines, 2021.

Armen KHATCHATOUROV, *Les identités numériques en tension : entre autonomie et contrôle*, avec la collaboration de Pierre-Antoine Chardel, Andrew Feenberg et Gabriel Périès, Londres, ISTE Éditions, série « Innovation et recherche responsables », 2019.

Armen KHATCHATOUROV (2019a), « Identités numériques et RGPD », dossier thématique « Le RGPD entre contrainte et innovation : les défis de la mise en conformité », *I2D. Information, données & documents*, juin 2019.

Armen KHATCHATOUROV (2019b), « La question des identités numériques à l'ère du RGPD : *privacy* ou protection des données ? », *I2D. Information, données & documents*, vol. 1, n° 1, 2019, p. 34-39.

Armen KHATCHATOUROV, « Big data entre l'archive et le diagramme », *Études digitales,* n° 2, Paris, Classiques Garnier, 2016.

Cécile PORTIER, « Mémoire personnelle à l'ère numérique – le questionnement des artistes », *Revue de la BNF*, 51, 2015, p. 60-63.

Antoinette ROUVROY et Thomas BERNS, « Le nouveau pouvoir statistique », *Multitudes,* n° 1, 2010, p. 88-103.

NOTES

1. Séance du 24 novembre 2022.
2. La communauté *Quantified Self* est structurée autour de réunions régulières, les *meet-up*. Pour en savoir plus, voir le site de la communauté : https://quantifiedself.com/

3. Il est important, à ce sujet, de distinguer les termes « mesure » et « quantification », même s'ils sont liés : la mesure est un acte ponctuel, tandis que la notion de quantification implique une forme d'historisation.

4. Par exemple le plasticien américain Laurie Frick « se présente comme data-artiste et produit des installations déployant dans l'espace ses "faits et gestes", comme dans cette série où ses déambulations sont systématiquement re-collectées, formant tableau, et, d'une certaine manière, autoportrait » (Portier, 2015, paragraphe 3).

5. Règlement général sur la protection des données.

6. L'entreprise britannique Cambridge Analytica est accusée d'avoir recueilli les données de 30 millions à 70 millions d'utilisateurs de Facebook, sans leur consentement, et de les avoir utilisées à des fins de ciblage électoral pendant la campagne présidentielle américaine de 2016.

7. Le taux d'attrition (ou *churn)* est, au cours d'une période donnée, la proportion de clients perdus ou ayant changé de produit et de service de la même entreprise.

8. Dans *Surveiller et punir* (1975), Michel Foucault consacre tout un chapitre au panoptique, invention du philosophe anglais Jeremy Bentham. Il s'agit d'un dispositif carcéral, dans lequel le surveillant, placé dans une tour centrale, peut observer en permanence la totalité des cellules placées en cercle autour de lui.

9. Extrait de Gille Deleuze, « Sur Foucault, le pouvoir », Transcription de cours à Vincennes – Saint-Denis, Cours du 25 février 1986. https://www.webdeleuze.com/textes/276

10. Autrement dit, on utilise des mesures effectuées par des capteurs physiques pour calculer des mesures théoriques en un point où l'on ne dispose pas de capteur physique. Prenons l'exemple d'un bâtiment dont on mesure la température en des points multiples. Si l'on sait, par des expériences multiples effectuées auparavant, qu'à la proximité d'une certaine machine, la température d'une salle s'élève toujours du même nombre de degrés, on ne mesure plus la température à proximité de la machine, mais on ajoute par calcul le nombre de degrés à la température prise dans d'autres points de la salle. Un exemple d'application des capteurs virtuels dans les locaux de l'université Concordia au Québec est donné dans l'article suivant : « Les capteurs virtuels prennent place dans nos bâtiments », *Le Devoir*, 29 mars 2019 https://www.ledevoir.com/contenu-commandite/551056/les-capteurs-virtuels-prennent-place-dans-nos-batiments (consulté le 23 mai 2023).

11. Antoinette Rouvroy et Thomas Berns, « Le nouveau pouvoir statistique », *Multitudes*, 1, 2010, p. 88-103. Cf. Yves Buuin, « Normopathie », *Passant ordinaire*, n^{os} 45-46 [juin-septembre 2003], http://www.passant-ordinaire.org/revue/45-46-556.asp (consulté le 24 mai 2023).

12. « Les pratiques de quantification dans le domaine de la santé favorisent la micro-gestion individuelle de la santé au détriment d'une appréhension plus collective. Elles font des individus des entrepreneurs d'eux-mêmes responsables de leur bon ou mauvais comportement de santé et peuvent distraire l'attention des causes environnementales ou socioéconomiques des problèmes de santé publique » (Antoinette Rouvroy, Avant-propos à « Le corps nouvel objet connecté : du *quantified self* à la M-santé : les nouveaux territoires de la mise en données du monde », *Cahiers Innovation et Prospective*, 2, 2013, p. 4-5).

13. Pour plus de details, voir : Rand Corporation, "The Internet of Bodies Will Change Everything, for Better or Worse", *The RAND Blog*, October 29, 2020 (consulté le 25 mai 2023).

14. « *In wich the work of hiring monitoring and evaluating workers is increasingly being delegated to AI technologies* » (Ajunwa, 2023, p. 4).

AUTEURS

BÉA ARRUABARRENA

Maître de conférences en sciences de l'information et de la communication, chercheuse au Dice-IDF Cnam Paris. Domaine de recherche : sociologie des objets connectés, quantification de soi (quantified self)

ARMEN KHATCHATOUROV

Maître de conférences en sciences de l'information et de la communication, chercheur au Dicen-IDF, Université Gustave Eiffel. Codirecteur de la revue *Études digitales* (Classiques Garnier). Domaine de recherche : philosophie des technologies

Analyse transdisciplinaire d'un corpus d'actualités filmées

L'environnement d'analyse numérique développé par le projet ANTRACT

Jean Carrive, Abdelkrim Beloued, Pascale Goetschel, Serge Heiden, Steffen Lalande, Pasquale Lisena, Franck Mazuet, Sylvain Meignier, Bénédicte Pincemin et Raphaël Troncy

Ce travail a été soutenu par l'Agence nationale de la recherche [ANR] dans le cadre du projet ANTRACT (ANR-17-CE38-0010) et par le programme de recherche et d'innovation Horizon 2020 de l'Union européenne dans le cadre du projet MeMAD (accord de subvention n° 780069).

Mise en œuvre d'un dispositif de recherche transdisciplinaire sur une collection d'archives cinématographiques : opportunités et défis

1 Le projet ANTRACT[1] réunit des laboratoires de recherche dans une double perspective historique et technologique, ce qui explique le caractère résolument transdisciplinaire du projet. Il s'intéresse à une collection de 1 262 films d'actualités (essentiellement des séquences en noir et blanc) diffusés dans les salles de cinéma françaises entre 1945 et 1969. Ces programmes ont été produits par la société *Les Actualités Françaises* durant une époque qui a pu être qualifiée de Trente Glorieuses. La recherche s'effectue non seulement sur les films proprement dits, mais aussi sur diverses ressources documentaires qui leur sont attachées : tapuscrits des commentaires en voix off contemporains des reportages, scannés et océrisés ; notices documentaires détaillées rédigées par les documentalistes de l'époque et reprises à divers moments par l'INA, avec titres, résumés, mots-clés, description et valeurs des plans, noms des participants, noms de lieux, etc.

2 Afin de mener cette recherche collective, des outils automatiques ont été développés et adaptés à l'analyse de cette collection conservée par l'Institut national de l'audiovisuel : reconnaissance automatique de la parole, classification d'images, reconnaissance faciale, traitement du langage naturel et fouille de textes. Ces logiciels sont utilisés

pour permettre à la communauté scientifique de travailler sur un corpus gérable et cohérent, disponible, à terme, à des fins de recherche.

3 En travaillant sur ces films d'actualités divisés en quelque 20 232 sujets, les historiens et les informaticiens d'ANTRACT collaborent, en définitive, pour optimiser la recherche sur les grands corpus audiovisuels en posant plusieurs questions clés :

- Quelle approche technologique peut efficacement épauler l'étude systématique et exhaustive d'un fonds d'archives multimédias ?
- Quels instruments peuvent compiler, analyser et recouper les données extraites de ces documents ?
- Ces données extraites peuvent-elles être combinées et intégrées dans des interfaces utilisateurs polyvalentes ?
- Avec ces traitements automatisés de sources multiformat, quelles nouvelles possibilités sont offertes aux projets de recherche en sciences humaines ?

4 Afin de mettre en œuvre une coopération solide entre les experts en informatique et les spécialistes de la discipline historique (Deegan et McCarty, 2012), l'objectif principal du projet est donc de fournir des pistes méthodologiques de recherche innovantes adaptées aux questions technologiques et historiques soulevées par ce corpus particulier.

5 **D'un point de vue technologique**, l'objectif est d'adapter les outils d'analyse automatique à la spécificité du corpus des *Actualités Françaises,* c'est-à-dire à son contexte historique, son vocabulaire et son type d'images. L'adaptation des modèles de langue utilisés par les outils de transcription automatique à l'aide de la version tapuscrite des voix off illustre cette orientation. En tant que collection de films comprenant des images, du son et du texte, produits il y a plus d'un demi-siècle, le corpus des *Actualités Françaises* représente un défi sans précédent pour les instruments spécialisés dans l'analyse et l'identification de contenus audiovisuels. Loin de considérer séparément une histoire sociale et culturelle du cinéma, d'une part, et l'utilisation d'outils d'analyse automatique, d'autre part, le projet vise à lier les deux. Ainsi, une bonne compréhension des conditions techniques d'enregistrement du son permet d'améliorer la reconnaissance audio. Tournées en noir et blanc avec un équipement limité et dans des conditions de tournage souvent difficiles, ces anciennes bandes d'actualités ne répondent pas aux normes de qualité fixées par les enregistrements vidéo et audio haute définition qui alimentent les algorithmes de reconnaissance d'images et de parole d'aujourd'hui. De plus, plusieurs bobines de films de la collection, numérisées sous des formats à haute compression, présentent des images pixellisées qui ne peuvent être traitées par les programmes d'analyse et certains des commentaires dactylographiés présentent des défauts d'impression causés par les machines à écrire utilisées pour leur production.

6 À ces obstacles matériels s'ajoute le problème posé par la transformation du contenu des films dans le temps. C'est le cas des personnalités régulièrement filmées par les cameramen de la société tout au long de ses 24 années d'activité. C'est aussi le cas des données topographiques récurrentes prises sur leur pellicule. L'identification automatique de ces éléments en constante évolution enregistrés sur des séquences monochromatiques nécessite des ressources considérables. Dans le cadre de ce processus, les historiens d'ANTRACT ont proposé une liste de 121 personnalités présentes dans les *Actualités Françaises* afin de construire une série de modèles d'extraction.

7 **D'un point de vue historique**, il convient de noter que le corpus des *Actualités Françaises* n'avait pas fait, jusqu'à présent, l'objet d'une analyse historique systématique de son contenu et de ses modalités de production. Aussi le projet ANTRACT vise-t-il à remédier à cette situation tout en cherchant à proposer des analyses historiques de différentes natures. Elles peuvent, dans le sillage des études déjà existantes sur l'histoire des actualités filmées, renvoyer à des considérations politiques et être, en particulier, centrées sur des études de censures (Atkinson, 2011 ; Bartels, 2004 ; Pozner, 2008 ; Veray, 1995). Elles peuvent aussi questionner le rôle de la presse cinématographique comme vecteur de l'histoire sociale, politique et culturelle façonnant l'opinion publique durant la seconde moitié du XX^{e} siècle (Fein, 2004, 2008 ; Althaus, 2018 ; Chambers, 2018 ; Imesch, 2016 ; Lindeperg 2000, 2008). De manière plus innovante, les études historiques effectuées dans le cadre du projet proposent d'autres pistes. En effet, au-delà des films eux-mêmes, d'autres sources liées au corpus filmique sont intéressantes. Les fiches d'observation remplies par les cameramen, les commentaires écrits des journalistes et les archives laissées par la direction donnent un aperçu inédit du contenu d'une collection de films ainsi que de son processus de production. Aussi, les études historiques développées dans le cadre d'ANTRACT conduisent à prendre en compte les conditions de production des informations diffusées.

8 Au-delà, la mise à disposition des résultats des outils de reconnaissance automatique permet de répondre à une question récurrente pour les historiens lorsqu'il s'agit de travailler sur des collections de grande ampleur : comment, parmi les milliers d'heures d'archives filmiques associées à des centaines de fichiers textes produits sur une longue période, travailler sur des objets d'étude particuliers ? Les outils développés par les partenaires du consortium, par le traitement combiné de données extraites, rendent possible une identification de thématiques de recherche historique plus aisée. Ces outils leur permettent donc de fabriquer des « sous-corpus » liés aux objets d'analyse en dépassant une recherche effectuée sur seulement quelques fragments, par exemple : les foules, les fêtes, les personnages influents, les mutations sociales, les prisonniers de guerre ou les mutilés à l'écran... Très concrètement, au-delà du repérage de « sous-corpus » liés à leurs thématiques, les historiens et historiennes peuvent travailler sur la fréquence, la durée, les moments d'apparition des sujets ou leur traitement médiatique.

9 Ainsi, centré sur le contenu des films, le projet s'attache, d'une part, à scruter le processus de production et les différents métiers impliqués dans la réalisation des *Actualités Françaises* en mettant en évidence les ressorts d'une entreprise contrôlée par un État démocratique. D'autre part, ANTRACT nourrit des analyses relatives à l'histoire politique, sociale et culturelle de la France des Trente Glorieuses, prise dans un environnement impérial, mais aussi européen et international. Il alimente les études sur le rôle des médias dans la fabrique des événements (Goetschel et Granger, 2011). Enfin, le projet invite à réfléchir sur le type d'images et de sons offerts aux spectateurs des salles de cinéma : goût pour les nouvelles sensationnelles, extraordinaires et exotiques, mais aussi ordinaires et banales, vie quotidienne de personnalités célèbres (Maitland, 2015), exploits et innovations scientifiques et techniques, mais aussi images évoquant différentes traditions locales, régionales, nationales...

10 À travers les outils d'analyse audio et vidéo (et les plateformes d'analyse historique interactive, cet article présente les résultats du projet, en mettant l'accent sur l'aspect

technologique de la recherche et, plus précisément, sur les appareils et les outils de traitement des données, en lien avec plusieurs exemples d'enquêtes historiques.

Analyse audio automatique

11 Le travail sur la partie audio consiste à détecter les locuteurs, à transcrire la parole en mots (ASR, pour *Automatic Speech Recognition*) et à détecter les entités nommées (EN, *Named Entities*) en utilisant les systèmes que nous avons développés pour les actualités contemporaines de radio et de télévision.

12 L'analyse audio d'un ensemble de données anciennes constitue un défi intéressant pour les systèmes d'analyse automatique. Les appareils d'enregistrement utilisés entre 1945 et 1969 sont très différents des appareils analogiques ou numériques d'aujourd'hui. Les films 35 mm, qui contiennent à la fois le son et l'image, se sont détériorés avant d'être numérisés dans les années 2000. De plus, les modèles acoustiques et linguistiques utilisés par les outils de reconnaissance automatique de la parole sont généralement entraînés sur des données produites entre 1998 et 2012. Ce décalage de 50 ans a des conséquences sur les performances du système.

13 Techniquement, pour ANTRACT, les modèles acoustiques pour l'ASR ont été entraînés sur environ 300 heures tirées de plusieurs sources d'actualités télévisées et radiophoniques françaises[2] associées à des transcriptions manuelles. Les modèles de langage (probabilités de suites de mots) ont été entraînés sur ces transcriptions manuelles, des journaux français, des sites d'information, Google News et le corpus français GigaWord, pour un total de 1,6 milliard de mots. Le vocabulaire du modèle de langage contient les 160 000 mots les plus fréquents. Les modèles EN ont été entraînés uniquement sur un sous-ensemble de transcriptions manuelles[3].

14 En amont du processus de transcription, le signal est découpé en segments de paroles homogènes et groupés par locuteur. Nous appelons ce processus la tâche de Segmentation et regroupement en locuteur [SRL]. Cette tâche est d'abord appliquée au niveau de l'édition (c'est-à-dire d'un journal entier), où chaque enregistrement vidéo est traité séparément. Ensuite, le processus est appliqué au niveau de la collection, sur l'ensemble des 1 262 éditions, afin de relier les locuteurs récurrents qui sont principalement les voix off. Le système développé par le Laboratoire d'informatique de l'université du Mans [LIUM] (Broux, 2018) est destiné à fournir des segments de parole homogènes contenant la parole d'un seul locuteur et des limites de segment précises marquant un changement de locuteur. Cette étape est essentielle au bon fonctionnement de l'ASR pour éviter les erreurs de transcription des début et fin de phrases. Le regroupement des segments par locuteur au niveau de l'édition ou de la collection n'a pas d'influence sur la qualité de la transcription, mais un regroupement précis des tours de parole d'un locuteur facilite la navigation dans la collection et la compréhension. L'élément clé du système de SRL est la caractérisation du signal au moyen d'un réseau de neurones profond qui extrait toutes les 10 ms un vecteur caractéristique du locuteur. Les algorithmes de segmentation et de regroupement reposent sur les méthodes classiques couramment utilisées en traitement acoustique ou d'image.

15 Le système ASR est développé principalement à l'aide de l'outil *open source* Kaldi (Povey, 2011). Il fait intervenir un modèle acoustique pour représenter les unités sonores

élémentaires (les phonèmes), une liste finie de mots potentiels et un modèle de langage qui détermine la probabilité d'une suite de mots. Les modèles acoustiques sont formés à l'aide d'un réseau neuronal profond qui peut traiter efficacement des contextes temporels longs (Povey, 2016). Des modèles de langage calculés avec des contextes de 2 ou 3 mots ont été entraînés sur de vastes corpus audio et textuels. Pour faciliter la lecture, deux systèmes d'étiquetage de séquences ont été entraînés sur des transcriptions manuelles pour ajouter respectivement la ponctuation et les majuscules à la suite de mots générés par l'ASR.

16 Le système neuronal d'extraction d'entités nommées complète l'analyse du texte. Le système, entraîné sur des transcriptions manuelles, détecte huit types d'entités principales : les lieux (la ville de Lyon), les organisations (l'ONU), les fonctions (Général), les personnes (Charles de Gaulle), les produits (un avion Caravelle), les événements (la Fête nationale du 14 juillet), les expressions numériques (1 000 francs) et les expressions temporelles (demain). L'annotation des entités nommées a pour but de mettre en avant les éléments de la transcription qui permet de répondre à des questions factuelles : qui, quoi, quand, où et comment.

17 L'ASR a été réalisée sur la collection complète des 1 262 éditions nationales afin d'alimenter les plateformes Okapi et TXM pour les analyses des historiens (voir les sections « Analyse textométrique interactive » et « Analyse sémantique interactive ») : environ 300 heures de vidéo, résultant en plus de 1,5 million de mots. Un sous-ensemble de 12 éditions de 1945 à 1969 a été transcrit manuellement pour évaluer les systèmes d'analyse audio. En raison de l'écart de 50 ans, les annotateurs humains ont eu quelques difficultés avec l'orthographe des entités nommées [EN], notamment en ce qui concerne les personnes et les EN étrangères. Grâce à Wikipédia et au thésaurus de l'Ina, la plupart des EN ont été vérifiées. En revanche, les locuteurs sont très difficiles à identifier. La plupart d'entre eux sont des voix off masculines. On ne voit jamais leur visage, leur nom est rarement prononcé et n'est pas affiché sur les images. Seuls les journalistes réalisant des interviews et les personnes connues, comme les politiciens, les athlètes et les célébrités, peuvent être identifiés et nommés avec précision.

18 La qualité d'un système ASR est évaluée à l'aide du taux de mots erronés (WER, pour *Word Error Rate*). Cette métrique consiste à compter le nombre d'insertions, de suppressions et de substitutions de mots entre les transcriptions générées automatiquement par le système ASR et les transcriptions humaines prises comme référence. Le WER est d'environ 24 % sur les données ANTRACT en utilisant le système générique ASR entraîné sur des données journalistiques modernes. Le même système évalué sur des données datant de 2010[4] atteint environ 13 %. Il est connu que les systèmes ASR sont sensibles aux variations acoustiques et linguistiques entre le corpus d'entraînement et le corpus de test. Ici, le WER est presque le double. Il est généralement difficile d'exploiter les transcriptions de manière robuste lorsque le WER est supérieur à 30 %. La plupart des erreurs proviennent de mots inconnus (qui ne sont pas répertoriés dans le vocabulaire de 160 000 mots). Ces mots hors vocabulaire sont confondus avec des mots acoustiquement proches, ce qui a un impact négatif sur les mots voisins. En effet, le système sélectionne toujours la séquence de mots la plus probable contenant le mot qui remplace le mot inconnu du système.

19 Des données contemporaines supplémentaires, telles que les notices documentaires et les tapuscrits, s'avèrent utiles pour adapter le modèle linguistique. Par conséquent, les résumés, les titres et les descriptions ont été extraits des notices documentaires. Les

phrases issues des tapuscrits ont été conservées lorsqu'au moins 95 % des mots appartiennent au vocabulaire ASR. Cela a permis de construire un corpus d'entraînement spécifique au domaine, composé de 1,3 million de mots provenant des notices documentaires et de 4,7 millions de mots provenant des tapuscrits. Les 4 000 mots les plus fréquents ont été sélectionnés pour entraîner le nouveau modèle de langage ANTRACT, ce qui a réduit le taux d'erreur de moitié : de 24 % à environ 12 % de WER. La figure 1 montre un exemple de transcription automatique de l'édition du 14 juillet 1955. Le gain est significatif grâce aux transcriptions consignées dans les tapuscrits, qui sont très similaires aux transcriptions manuelles. Ce corpus d'entraînement spécifique va à l'encontre des règles strictes établies pour l'évaluation de systèmes de reconnaissance automatique : les données de test ne doivent jamais être utilisées pour construire un corpus d'entraînement. Cependant, dans notre cas, l'objectif principal est de fournir les meilleures transcriptions possibles aux historiens.

20 Cependant, toute erreur a un impact sur les recherches d'information pour les historiens et constitue une limite à nos travaux. Par exemple, transcrire le mot « poule » lorsque le locuteur a prononcé le mot « foule » ne permet pas de visionner l'ensemble des séquences parlant des foules. En fin de projet, une correction manuelle des transcriptions a été réalisée. La correction est effectuée en deux passes. Les corrections orthographiques et grammaticales des mots communs et de la segmentation (essentiellement les débuts et les fins de segments) sont réalisées dans une première phase. Une seconde passe, qui demande plus de temps que la première, se focalise sur la correction des noms propres. Cette dernière nécessite d'interroger des ressources externes aux projets (Wikipédia, dictionnaire, etc.) ou internes (notice, tapuscrit) pour corriger les noms propres souvent inconnus du vocabulaire du système.

21 En définitive, au fil du projet, les travaux se sont concentrés sur l'amélioration des modèles acoustiques ASR, grâce, notamment, aux données textuelles issues des tapuscrits des commentaires en voix off et au retour d'information des historiennes et historiens qui ont corrigé manuellement des parties de transcription. L'évaluation des entités nommées n'a pas été réalisée sur les données du projet, faute d'annotations. De même, l'évaluation des locuteurs a été difficile en raison de l'absence d'information sur leur identité. Celle-ci s'est donc concentrée sur la détection des commentateurs (voix off) et des intervieweurs. En outre, certaines personnes célèbres, sélectionnées en collaboration avec des historiens pour leur pertinence dans les analyses historiques, sont identifiées, avec l'aide éventuelle du croisement des résultats avec l'analyse d'image, comme décrit dans la section suivante.

Figure 1. Exemple de l'Édition du 14 juillet 1955 des *Actualités Françaises* (de 6:06 à 6:49).

Le sous-titre est généré automatiquement par l'ASR avec un modèle de langue du domaine, une ponctuation automatique et des majuscules.

Analyse visuelle automatique

22 L'identification des personnes apparaissant dans une vidéo est indéniablement un élément clé pour sa compréhension. Savoir qui figure dans une vidéo, quand et où, peut également permettre de découvrir des modèles intéressants de relations entre les personnes pour la recherche historique. De telles annotations liées aux personnes pourraient permettre de créer un contenu à valeur ajoutée. Une archive historique telle que le corpus des *Actualités Françaises* contient de nombreux exemples de célébrités figurant dans le même segment d'actualités, par exemple Charles de Gaulle et Konrad Adenauer (voir Figure 2). Cependant, les annotations produites manuellement par les documentalistes ne permettent pas toujours d'identifier avec précision les individus présents dans les vidéos. D'autre part, le Web offre une quantité importante de photographies de ces personnes, facilement accessibles par les moteurs de recherche en utilisant leur nom complet comme terme de recherche. Dans ANTRACT, l'idée a été d'exploiter ces images pour identifier les visages des célébrités dans les archives vidéo.

Figure 2. De Gaulle et Adenauer ensemble dans une vidéo de 1959.

23 De nombreux progrès ont été réalisés au cours de la dernière décennie concernant la reconnaissance automatique des personnes. Elle comprend généralement deux étapes : il faut d'abord détecter les visages (c'est-à-dire savoir quelle région de l'image est susceptible de contenir un visage de personne), puis les reconnaître (c'est-à-dire savoir à quelle personne chaque visage appartient).

24 L'algorithme de Viola-Jones (Viola, 2004) pour la détection des visages et les caractéristiques des motifs binaires locaux [LBP] (Ahonen, 2006) pour le regroupement et la reconnaissance des visages étaient les techniques les plus utilisées jusqu'à l'avènement de l'apprentissage profond et des réseaux de neurones dits convolutionnels [CNN]. Aujourd'hui, deux approches principales sont utilisées pour détecter les visages dans les vidéos et toutes deux utilisent des CNN. La bibliothèque Dlib (King, 2009) offre de bonnes performances pour les vues de face, mais elle nécessite une étape supplémentaire d'alignement (qui peut également être effectuée à l'aide de la bibliothèque Dlib) avant de pouvoir procéder à la reconnaissance des visages. L'approche plus récente *Multi-task Cascaded Convolutional Networks* [MTCNN] fournit des performances encore meilleures en utilisant une approche image-pyramide et intègre la détection des points de repère des visages afin de réaligner les visages détectés sur la vue de face (Zhang, 2016).

25 Après avoir repéré la position et l'orientation des visages dans les images vidéo, le processus de reconnaissance peut être effectué dans de bonnes conditions. Plusieurs stratégies ont été détaillées dans la littérature pour réaliser la reconnaissance. Actuellement, l'approche la plus pratique consiste à effectuer une comparaison de visages à l'aide d'un espace de transformation dans lequel les visages similaires sont proches les uns des autres, et à utiliser cette représentation pour identifier la bonne personne. De tels espaces de « plongement » (*embeddings*), calculés sur de grandes collections de visages, sont disponibles pour la communauté de recherche (Schroff, 2015).

26 Au sein d'ANTRACT, nous avons développé un système *open source* de reconnaissance faciale des célébrités. Cette application est composée des modules suivants :

- un robot d'exploration du Web qui, étant donné le nom d'une personne, télécharge automatiquement *via* Google un ensemble de k photos qui seront utilisées pour l'entraînement d'un modèle de visage particulier. Dans nos expériences, nous utilisons généralement $k = 50$. Parmi les résultats, les images ne contenant aucun visage ou contenant plus d'un visage sont écartées. En outre, les utilisateurs finaux (par exemple, les experts du domaine) peuvent exclure manuellement les résultats non pertinents, qui, par exemple, ne correspondent pas à la personne recherchée ;
- un module d'entraînement où les photographies récupérées peuvent être converties en noir et blanc, recadrées et redimensionnées afin d'obtenir des images contenant uniquement un visage, en utilisant l'algorithme MTCNN (Zhang, 2016). Un modèle Facenet (Schroff, 2015) préentraîné avec une architecture Inception ResNet v1 entraînée sur VGGFace2dataset (Cao, 2018), est appliqué afin d'extraire les caractéristiques visuelles des visages. Les plongements obtenus sont utilisés pour entraîner un classifieur SVM ;
- un module de reconnaissance qui prend en entrée une vidéo d'actualité et en extrait une image toutes les d images (dans nos expériences, nous avons généralement fixé $d = 25$, soit une image par seconde). Pour chaque image, les visages sont détectés (en utilisant l'algorithme MTCNN) et les *embeddings* sont calculés (Facenet). Le classifieur SVM décide si le visage correspond à l'un des visages des images d'entraînement. *Simple Online and Realtime Tracking* [SORT] est un algorithme de suivi d'objets, qui peut suivre plusieurs objets en temps réel (Bewley, 2016). Son implémentation est inspirée du code de suggestion de Linzaer[5]. L'algorithme utilise la détection de la boîte de délimitation MTCNN et la suit à travers les images. Nous avons introduit ce module pour augmenter la robustesse du traitement. En utilisant ce module, tout en faisant l'hypothèse que les visages ne changent quasiment pas de coordonnées entre deux images consécutives, nous visons à obtenir une prédiction plus cohérente ;
- enfin, le dernier module regroupe les résultats provenant du classifieur et des modules de suivi. Nous observons que même si le visage à reconnaître reste le même sur plusieurs images consécutives, la prédiction du visage change parfois. Pour cette raison, nous sélectionnons pour chaque suivi la prédiction la plus fréquente, en prenant également en compte le score de confiance donné par le classifieur. De cette façon, le système fournit une prédiction commune pour toutes les images impliquées dans un suivi, ainsi qu'un score de confiance agrégé. Un seuil t peut être appliqué à ce score afin d'écarter les prédictions peu fiables. D'après nos expériences, $t = 0{,}6$ donne un bon compromis entre la précision et le rappel.

27 Afin de rendre le logiciel disponible en tant que service, nous l'avons intégré dans une API web RESTful[6]. Le service prend en entrée l'URI d'une ressource vidéo, telle qu'elle apparaît dans Okapi, à partir duquel il récupère l'objet média encodé en MPEG-4. Deux formats de sortie sont disponibles : un format JSON personnalisé et un format de sérialisation en RDF utilisant la syntaxe Turtle et la syntaxe Media Fragment URI (Troncy, 2012), avec la durée de lecture normale exprimée en secondes pour situer les fragments temporels et les coordonnées *xywh* pour définir le rectangle encadrant le visage dans l'image. Un troisième format, toujours selon la syntaxe Turtle, sera bientôt implémenté afin que les résultats puissent être directement intégrés dans le graphe de connaissances Okapi. Un système de cache léger est également mis en place afin de pouvoir fournir des résultats précalculés, sauf si le paramètre *no cache* est activé.

28 Nous avons mené des expériences en utilisant le modèle de visage de Dwight D. Eisenhower sur une sélection de segments vidéo extraits d'Okapi, parmi ceux qui ont été annotés avec la présence du président américain selon les propriétés *ina:imageContient* et *ina:aPourParticipant* dans le graphe de connaissances. En l'absence d'une vérité terrain, nous avons effectué une analyse qualitative de notre système sur trois vidéos. Pour chaque personne détectée, nous avons évalué manuellement si la bonne personne était trouvée ou non. Sur les 90 segments sélectionnés, le système a correctement identifié Eisenhower dans 33 d'entre eux. Cependant, nous ne sommes pas sûrs que Eisenhower soit effectivement présent visuellement dans les 90 segments (il peut avoir été indexé pour une apparition plus tôt ou plus tard dans le même sujet par exemple). Nous avons ensuite produit une vérité terrain qui nous a permis d'évaluer la précision et le rappel du système (Lisena, 2022).

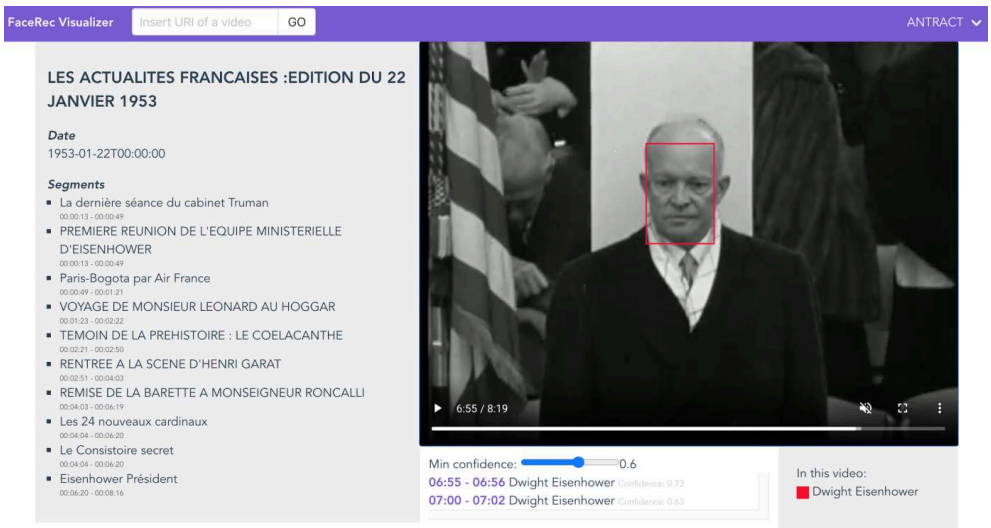
29 En outre, nous avons fait les observations suivantes :

- notre logiciel ne parvient généralement pas à détecter les personnes lorsqu'elles sont en arrière-plan ou lorsque le visage est masqué ;
- lorsque les visages sont parfaitement de face, ils sont plus faciles à détecter. Des améliorations de l'algorithme d'alignement sont prévues dans les travaux futurs.

30 En fixant un seuil de confiance élevé, nous ne rencontrons pas de cas de confusion entre deux célébrités. La plupart des erreurs consistent plutôt à confondre un visage inconnu avec celui d'une célébrité.

31 Afin de visualiser facilement les résultats et de faciliter le retour d'information des historiens, nous avons développé une application web qui affiche les résultats directement sur la vidéo, en tirant parti des fonctionnalités HTML5[7]. L'application fournit également un résumé des différentes prédictions, permettant à l'utilisateur de passer directement à la partie relative de la vidéo où la célébrité apparaît. Un curseur permet de modifier la valeur du seuil de confiance, afin de mieux étudier les résultats jugés peu fiables.

Figure 3. Le visualiseur du système de reconnaissance des visages de célébrités.



Analyse textométrique interactive

32 Dans ANTRACT, l'exploration et l'analyse des données textuelles sont proposées aux historiens selon une approche textométrique (Lebart, 1998). La textométrie combine à la fois des outils quantitatifs – statistiques –, et des outils qualitatifs – de recherche, de lecture et d'annotation de textes. Les fonctionnalités statistiques comprennent des calculs de spécificités lexicales, de cooccurrences (mots associés), de classification hiérarchique et d'analyse des correspondances. Cela représente un gain significatif en matière de possibilités d'analyse par rapport aux fonctions habituelles d'annotation, de recherche et de décompte des logiciels de transcription audiovisuelle tels que CLAN (MacWhinney, 2000) ou ELAN (ELAN, 2018). Quant à l'analyse qualitative, elle est réalisée par des concordances avancées, avec un accès hypertexte aux documents sources, et avec des possibilités d'annotation dynamique du corpus en cours d'analyse. Un tel aspect qualitatif est marginal, voire absent, dans les applications classiques de fouille de textes (Hotho, 2005 ; Feinerer, 2008 ; Weiss, 2015) : la plupart d'entre elles traitent du texte brut, en commençant si besoin par éliminer les marques de structuration du texte, et elles cherchent à produire une visualisation synthétique qui remplace la lecture attentive du texte (au lieu de garder une relation constante au texte original étudié).

33 La textométrie est ici mise en œuvre avec la plateforme logicielle TXM (Heiden, 2010). TXM est développé de façon *open-source* et intègre plusieurs composants spécialisés : R (R Core Team, 2014) pour les calculs statistiques, CQP comme moteur de recherche en texte intégral (Christ, 1994), TreeTagger (Schmid, 1994) pour le traitement du langage naturel (étiquetage morphosyntaxique et lemmatisation). TXM s'inscrit dans les pratiques de la science ouverte au niveau de la standardisation et du partage des données et du code informatique, et a notamment été conçu pour gérer des corpus richement structurés et annotés, tels que des données XML et des textes encodés suivant les recommandations de la TEI[8]. Pour les données textuelles ANTRACT, TXM importe des données tabulées (export Excel de tableaux depuis les bases documentaires de l'INA) et des fichiers au format XML Transcriber fournis par un logiciel de transcription de la parole (voir Section « Analyse audio automatique »). TXM est un outil pour l'analyse textuelle, mais il permet également de gérer les représentations multimédias associées aux textes, qu'il s'agisse d'images scannées des documents sources, d'enregistrements audio ou vidéo : en effet, ces représentations participent à l'interprétation des résultats des traitements en les replaçant dans leur contexte sémiotique complet.

34 En 2018, nous avons commencé par la construction du corpus TXM AF-NOTICES en important les notices documentaires de l'Ina : chaque sujet est représenté par plusieurs champs textuels (titre, résumé, séquences plan par plan) et plusieurs champs lexicaux (listes de descripteurs documentaires de différents types tels que sujets, personnes ou lieux, et générique des noms des personnes montrées ou des cameramen). Chaque sujet est également caractérisé par une dizaine de métadonnées (identifiant Ina, date de diffusion, producteur du film, genre du film, etc.) utiles pour le contextualiser ou le catégoriser.

35 À partir de 2019, nous avons aussi réalisé le corpus TXM AF-VOIX-OFF fondé sur les transcriptions du commentaire audio (voir Section « Analyse audio automatique ») et

synchronisé au mot près aux vidéos. Les champs documentaires des notices Ina sont intégrés au corpus en tant que métadonnées décrivant les transcriptions.

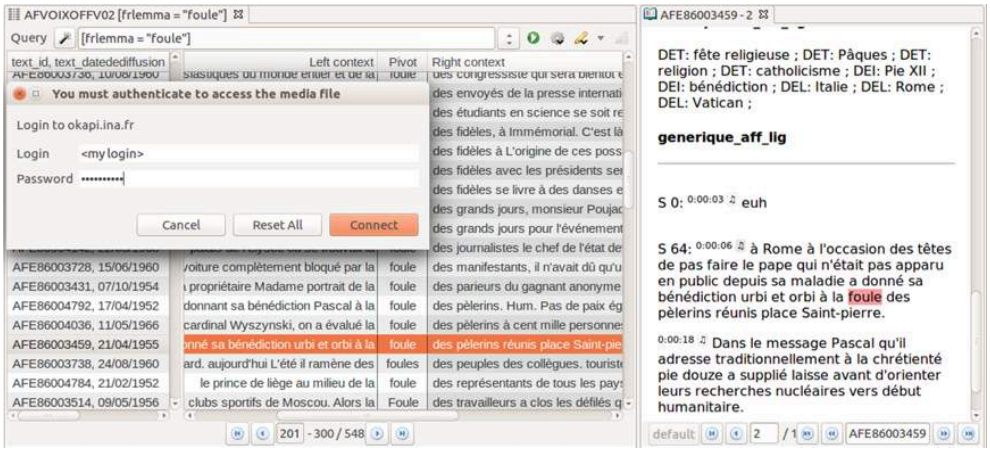
36 Ces corpus pourraient encore être étendus par l'ajout de nouvelles données textuelles : les textes issus des tapuscrits des commentaires qui ont été scannés et océrisés (reconnaissance optique des caractères), les annotations sur les vidéos (annotations manuelles ajoutées par les historiens via la plateforme Okapi) [voir Section « Analyse sémantique interactive »], ainsi que les annotations automatiques générées par les logiciels de reconnaissance d'images (voir Section « Analyse visuelle automatique »), etc.

37 L'une des innovations techniques réalisées dans le cadre du projet a été la consolidation du module de retour à la vidéo de TXM (Pincemin, 2020), de sorte que tout mot ou passage de texte trouvé dans le résultat d'un calcul textométrique puisse être consulté dans la vidéo originale ; nous avons également mis en œuvre un accès contrôlé par mot de passe aux vidéos en ligne sur le serveur média d'Okapi, ce qui s'est avéré être un développement clé pour la mise à disposition des vidéos, étant donné la taille de stockage importante requise pour ces enregistrements et les contraintes de sécurité sur ces archives cinématographiques.

38 Les captures d'écran qui suivent illustrent des étapes types d'une analyse textométrique telle que menée dans le cadre du projet ANTRACT.

39 Dans les figures 4 et 5, nous étudions le contexte d'utilisation du mot « foule » à l'aide d'une concordance. Un double-clic sur une ligne de concordance ouvre une nouvelle fenêtre (à droite) qui affiche la transcription complète dans laquelle apparaît le mot. Ensuite, un clic sur le symbole des notes de musique au début du paragraphe permet de lire la vidéo correspondante. Une boîte de dialogue demande des identifiants pour accéder à la vidéo sur le serveur en ligne d'Okapi. Cette possibilité de confronter l'analyse textuelle à la source audiovisuelle est d'autant plus importante ici que les données textuelles ont été générées par un outil automatique de reconnaissance de la parole, dont la sortie n'a pas pu être entièrement vérifiée. De plus, la vidéo peut apporter des éléments de contexte significatifs qui complètent le contenu textuel.

Figure 4. Confrontation de l'analyse textuelle à la source audiovisuelle, étape 1.



CONCORDANCE du mot « foule » dans le corpus de voix off (fenêtre de gauche), ÉDITION de la page de transcription correspondant à la ligne de concordance sélectionnée (fenêtre de droite), et boîte de dialogue d'authentification permettant d'accéder au serveur vidéo Okapi pour lire la vidéo au temps 0:00:06 (fenêtre supérieure gauche).

Figure 5. Confrontation de l'analyse textuelle à la source audiovisuelle, étape 2.

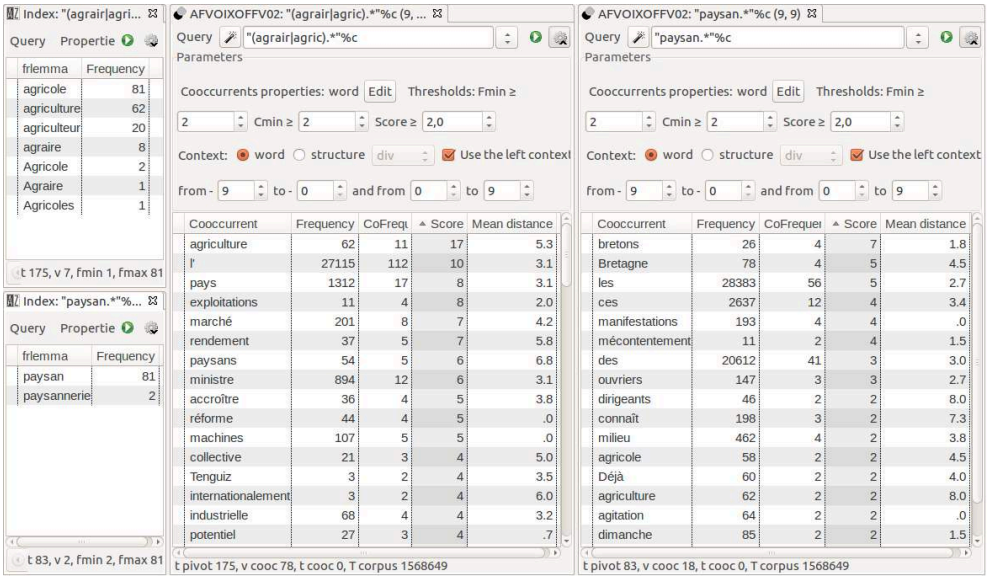
Fenêtres liées entre elles et présentant des résultats pour le mot « foule » : CONCORDANCE (fenêtre de gauche), ÉDITION de la transcription (fenêtre du milieu) et lecture vidéo synchronisée (fenêtre de droite).

40 Notre deuxième exemple concerne la place de l'agriculture et des agriculteurs dans les *Actualités Françaises* et la manière dont ce sujet est abordé. Ce cas illustre comment on peut observer si un mot donné a le même sens dans les notices documentaires et dans les commentaires audio, ou si des mots différents sont utilisés pour traiter du même sujet. Nous obtenons d'abord (Figure 6) un aperçu comparatif de l'évolution quantitative des occurrences de deux familles de mots, les dérivés des radicaux de « paysan » et « agricole »/« agriculture » (voir la liste détaillée des mots dans la Figure 7, fenêtre de gauche). Nous complétons l'analyse par un examen des contextes d'emploi à travers une vue en concordance (voir Figure 6, fenêtre inférieure) et un calcul de cooccurrence (voir Figure 7). Nous remarquons que « paysan » devient moins utilisé à partir de 1952 et qu'il est préféré à « agriculteur » pour parler des individus présents dans les extraits d'actualités ; inversement, « agricole »/« agriculture » sont utilisés de manière plus abstraite, pour traiter des nouveaux équipements agricoles et de la transformation socio-économique de ce secteur d'activité.

Figure 6. Examen des contextes d'emploi de deux familles de mots.

Graphique de PROGRESSION (fenêtre supérieure) et CONCORDANCES associées (fenêtre inférieure), pour comparer deux familles de mots liées à l'agriculture.

Figure 7. Calcul de cooccurrence de deux familles de mots.

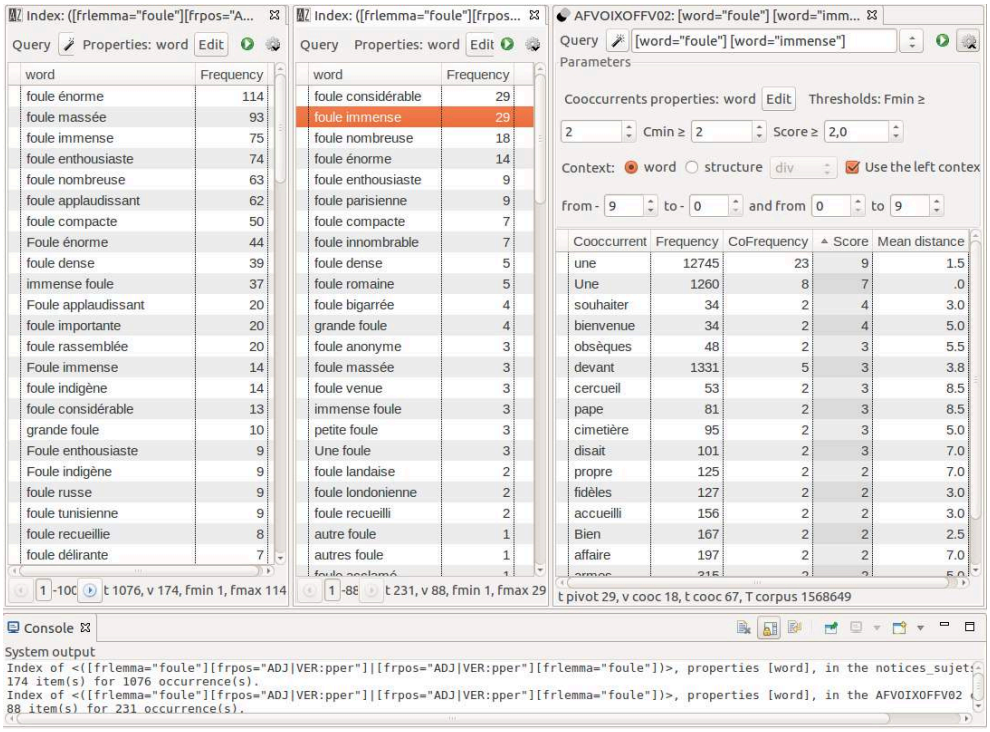


Résultats d'INDEX détaillant le contenu de deux familles de mots (marge gauche) et analyse statistique de COOCCURRENCES pour caractériser leurs contextes.

41 La combinaison des listes de mots (INDEX) et des informations morphosyntaxiques est très efficace pour résumer les formulations dans lesquelles les mots sont employés. Par exemple, dans la figure 8, nous pouvons comparer quels adjectifs qualifient « foule » dans les notices documentaires et quels adjectifs qualifient « foule » dans les discours

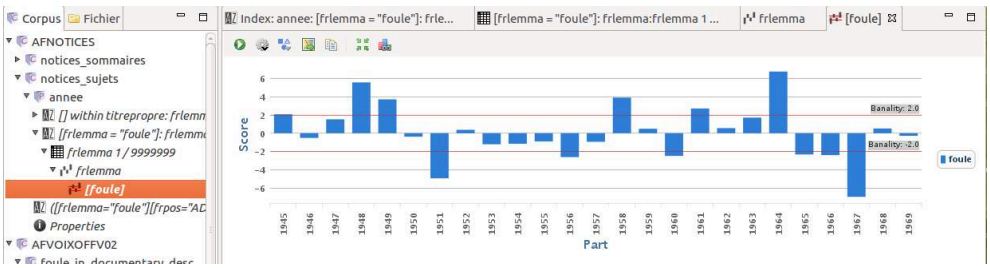
en voix off. Pour une expression donnée (« foule immense ») prononcée dans le commentaire audio, nous calculons ses cooccurrences afin d'identifier dans quels types de circonstances cette expression est généralement utilisée (ici des funérailles et des rassemblements religieux). Dans TXM, la recherche en texte intégral bénéficie du moteur de recherche CQP (Christ, 1994), qui permet des requêtes très précises, y compris avec des conditions de contexte.

Figure 8. Examen de l'emploi d'un mot précédé ou suivi d'un adjectif.



INDEX de « foule » précédé ou suivi d'un adjectif, dans les notices documentaires (fenêtre de gauche) ou dans les transcriptions de la voix off (fenêtre du milieu). COOCCURRENCES de « foule immense » dans les transcriptions de la voix off (fenêtre de droite).

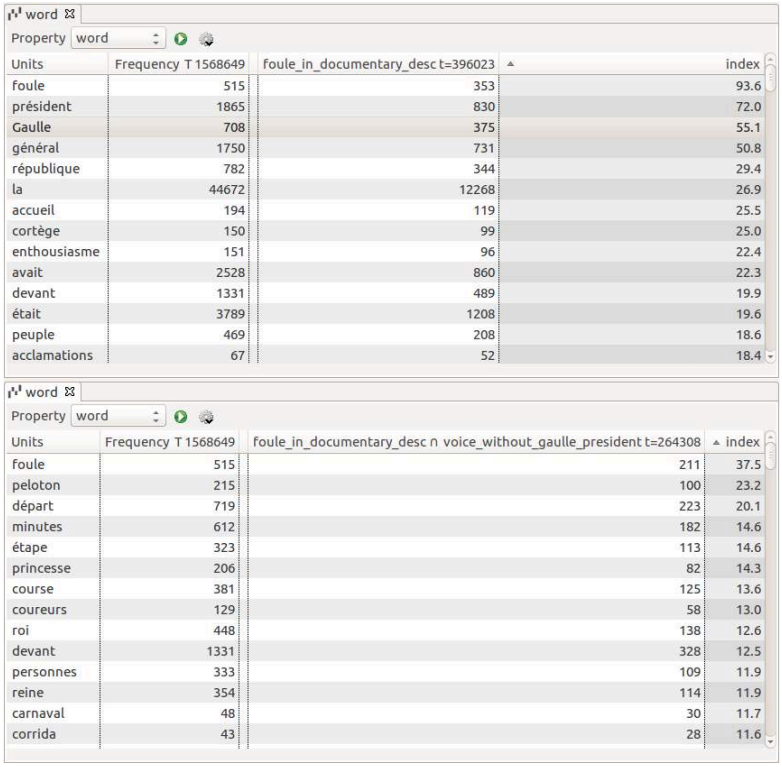
Figure 9. Graphique de SPECIFICITE pour « *foule* » au fil des années.



42 Pour les recherches chronologiques, le logiciel permet de diviser le corpus en périodes de manière très flexible, par exemple année par année ou en définissant des groupes d'années. Toute information disponible codée dans le corpus peut être utilisée pour définir les subdivisions du corpus. Ensuite, la commande SPÉCIFICITÉ mesure statistiquement pour chaque mot l'équilibre de sa répartition à travers les parties et met en évidence ses éventuels sur- ou sous-emplois dans certaines parties. La fonction peut également être utilisée pour lister l'ensemble des termes spécifiques à une période

donnée (ou à une partie quelconque qu'on définit dans le corpus). Par exemple, la figure 9 s'intéresse au mot « foule » au fil des années. Les années aux scores les plus importants révèlent des événements politiques décisifs (par exemple, la Libération de la France après la Seconde Guerre mondiale, l'avènement de la Cinquième République), qui correspondent à la forte exposition du général de Gaulle. Cependant, on note aussi que les moments de plus forte présence du mot ne correspondent pas nécessairement à des bouleversements politiques.

Figure 10. Exemple d'analyse de résonance (Salem, 2004).



word

Property word

Units	Frequency T 1568649	foule_in_documentary_desc t=396023	index
foule	515	353	93.6
président	1865	830	72.0
Gaulle	708	375	55.1
général	1750	731	50.8
république	782	344	29.4
la	44672	12268	26.9
accueil	194	119	25.5
cortège	150	99	25.0
enthousiasme	151	96	22.4
avait	2528	860	22.3
devant	1331	489	19.9
était	3789	1208	19.6
peuple	469	208	18.6
acclamations	67	52	18.4

word

Property word

Units	Frequency T 1568649	foule_in_documentary_desc ∩ voice_without_gaulle_president t=264308	index
foule	515	211	37.5
peloton	215	100	23.2
départ	719	223	20.1
minutes	612	182	14.6
étape	323	113	14.6
princesse	206	82	14.3
course	381	125	13.6
coureurs	129	58	13.0
roi	448	138	12.6
devant	1331	328	12.5
personnes	333	109	11.9
reine	354	114	11.9
carnaval	48	30	11.7
corrida	43	28	11.6

Termes SPÉCIFIQUES dans les commentaires de la voix off pour les sujets montrant une foule (selon les notices documentaires) [fenêtre supérieure] ; puis, termes SPÉCIFIQUES dans les transcriptions de la voix off pour les sujets montrant une foule et n'ayant aucune mention de « De Gaulle » ou de « président » (fenêtre inférieure).

43 Avec la figure 10, nous appliquons une analyse de résonance (Salem, 2004). Lorsqu'une foule est montrée (d'après la notice documentaire), quels sont typiquement les mots prononcés dans le commentaire en voix off ? « Président » et « [le général de] Gaulle » représentent ainsi le principal contexte d'emploi (Figure 10, fenêtre supérieure). Dans un second temps, nous supprimons tous les sujets contenant l'un de ces deux mots et nous nous focalisons sur les sujets restants pour faire émerger de nouveaux types de contextes associés à la mise à l'image d'une foule (Figure 10, fenêtre inférieure), tels que le sport, les commémorations, les manifestations, les fêtes, etc. La mention récurrente de la « foule » dans les commentaires en voix off favorise le sentiment d'appartenance à une communauté de destin. D'un point de vue méthodologique, ce type d'interrogation croisée, combinée à une comparaison statistique entre les notices documentaires et les transcriptions des commentaires audio, permet d'étudier les corrélations ou les divergences entre ce qui est montré à l'image et ce qui est dit dans

les commentaires. Une telle analyse croisée de différents médias est rarement fournie dans les logiciels d'analyse.

44 La figure 11 donne un premier aperçu des résultats d'une analyse des correspondances : nous avons calculé une carte bidimensionnelle des noms des personnes présentes dans plus de 20 sujets, en relation avec les années où elles sont mentionnées. Nous obtenons ainsi une vue synthétique de la relation entre les personnes et le temps dans les sujets des *Actualités Françaises*. En termes de modélisation statistique, comme la textométrie traite souvent de tableaux de fréquences croisant des mots et des parties de corpus (ici nous avons croisé des noms de personnes et des années), elle opte pour l'analyse factorielle des correspondances, car ce type d'analyse multidimensionnelle est particulièrement bien adapté aux tableaux de contingence (Lebart, 1998).

Figure 11. Carte bidimensionnelle des noms des personnes et des années de leur mention.

ANALYSE DES CORRESPONDANCES du tableau de fréquences croisant les années et les noms des 51 personnes qui sont présentes dans au moins 20 sujets.

Analyse sémantique interactive

45 Dans le contexte du projet ANTRACT, les historiens peuvent également explorer le corpus d'une façon complémentaire à l'aide de la plateforme Okapi. Okapi [*Open Knowledge Annotation and Publication Interface*] (Beloued, 2017) est une plateforme en ligne permettant la gestion sémantique de contenus. Elle se situe à l'intersection de trois domaines scientifiques : l'indexation et la description de contenus multimédias, la gestion des connaissances et la médiation de contenus. La plateforme s'appuie sur les langages et normes du web sémantique [RDF : *Resource Description Framework*, RDFS : *Resource Description Framework Schema*, OWL : *Ontology Web Language*] (Motik, 2012) pour représenter le contenu sous forme de graphes de connaissances. Les produits de médiation comme les portails web sont obtenus en appliquant des inférences sémantiques sur ces graphes de données, le processus étant paramétré par un graphe de publication réalisé par l'auteur du portail et spécifiant le lien entre données et objets multimédias.

46 Okapi fournit un ensemble d'outils pour la segmentation et la description sémantique de contenus multimédias (vidéo, image, son, texte et 3D) en s'appuyant sur des ontologies de domaine. Le logiciel fournit également des services pour la constitution de corpus hiérarchiques et thématiques à partir d'extraits annotés pour l'ensemble de ces modalités.

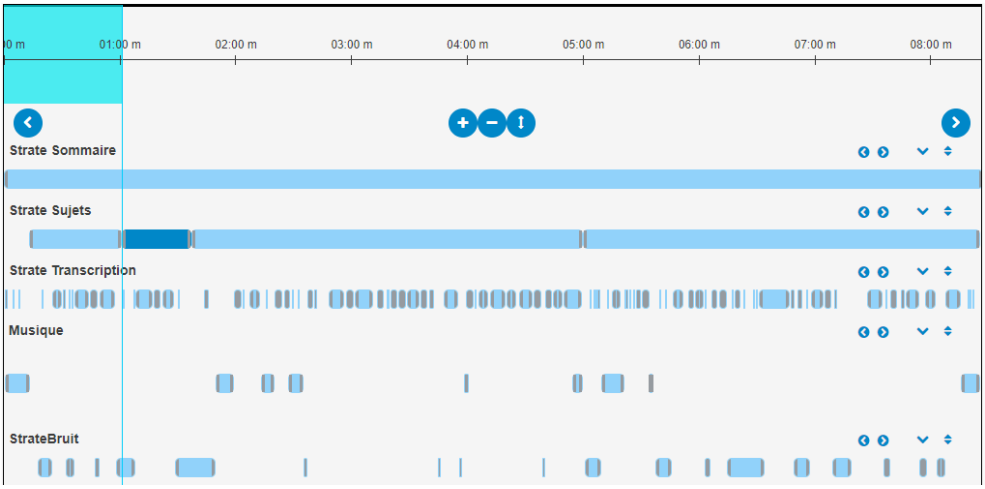
47 Okapi gère les entités nommées comme des graphes de connaissances et fournit des services destinés à les rechercher, les partager et les présenter comme des données ouvertes. Ces entités peuvent être mises en relation avec d'autres entités dans des bases de connaissances comme DBpedia et Wikidata, ce qui rend Okapi interopérable avec l'écosystème des données ouvertes liées [LOD : *Linked Open Data*] (Bizer, 2009). Les entités nommées peuvent être de différents types qui varient en fonction du domaine étudié, par exemple personnes physiques et morales, lieux géographiques, événements, concepts.

48 Le système de gestion de contenu d'Okapi prend en compte les caractéristiques du domaine étudié et les préférences utilisateurs pour générer des interfaces web telles que des portails adaptés au domaine, sans requérir aucune compétence technique de leurs auteurs. Un auteur peut également réaliser une publication thématique sous forme d'un ensemble d'éléments multimédias interconnectés (vidéo, image, son) auxquels il peut ajouter du contenu éditorial. L'outil de publication applique ensuite un ensemble de règles de publication sur ces éléments et génère un mini-portail.

49 Dans le cadre du projet ANTRACT, la plateforme Okapi est utilisée par les historiens pour constituer des corpus thématiques, visualiser et, dans certains cas, corriger les métadonnées originelles issues des notices Ina (l'ancrage temporel des sujets Ina non documentés) ainsi que les résultats issus des algorithmes automatiques (détection et reconnaissance de visages, transcription de la parole, OCR sur les tapuscrits et sur les vidéos, etc.) Les sections suivantes montrent quelques exemples d'utilisation de la plateforme Okapi sur la collection des *Actualités Françaises*.

50 La description d'un média dans Okapi peut être réalisée manuellement par des annotateurs ou automatiquement par des algorithmes d'indexation selon plusieurs axes (thématique, sonore, visuel, etc.) comme le montre la figure 12. La *timeline* Amalia (Hervé, 2015) est utilisée pour visualiser ces axes de description sous forme de strates et représenter la progression temporelle de chacun d'entre eux. Plusieurs types de strates sont proposés pour la représentation temporelle des données ANTRACT, chacune étant dédiée à un type d'annotation : les strates « sommaire » et « sujets » sont consacrées aux métadonnées originelles issues, respectivement, des notices émissions et notices sujets de l'INA, la strate « visage » est dédiée aux métadonnées extraites par l'algorithme de détection et d'identification des visages (Lisena, 2022) [Section « Analyse visuelle automatique »] et la strate « transcription » aux métadonnées issues de l'algorithme de la transcription de la parole (Section « Analyse audio automatique »). Ces métadonnées sont portées par des objets de type « segment » qui délimitent leurs portées temporelles au sein d'une strate (Figure 12) et peuvent être structurées, suivant leur type, en plusieurs annotations. Par exemple, la description thématique (strate intitulée « sujets ») de l'émission *Journal Les Actualités Françaises : émission du 10 juillet 1968* (Figures 12 et 13) consiste à identifier les thèmes abordés dans cette émission, leur portée temporelle et une description détaillée du sujet abordé, des lieux où l'action se déroule et des personnes impliquées.

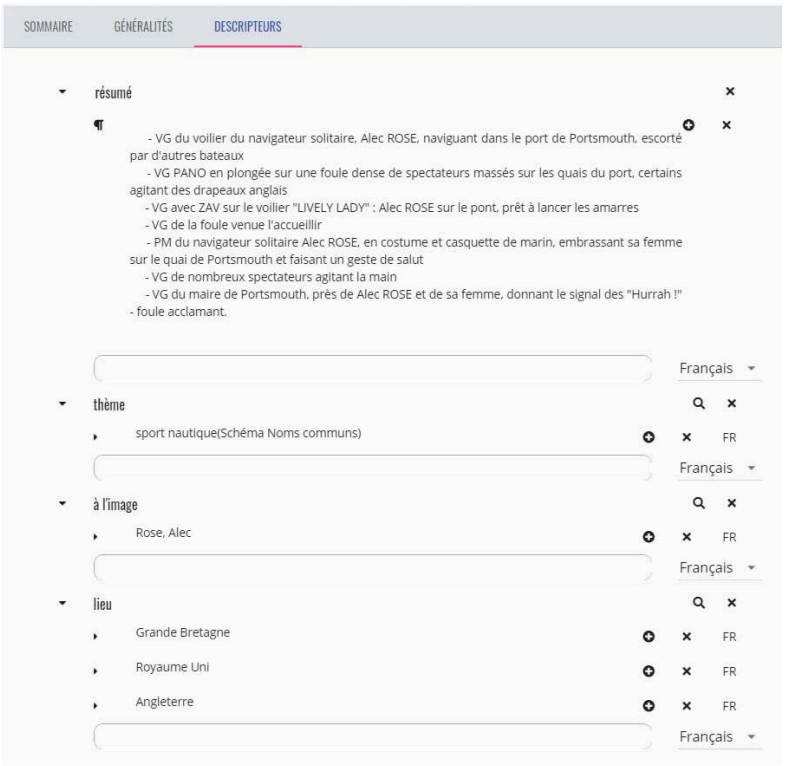
Figure 12. *Timeline* de description.



51 L'utilisateur peut créer une strate pour ajouter une nouvelle dimension de description, identifier les portées temporelles des annotations en créant des segments et renseigner les informations associées. Okapi fournit une boîte à outils pour ajuster finement l'ancrage temporel de chaque segment. Cette boîte à outils a été utilisée notamment dans le projet ANTRACT pour corriger l'ancrage temporel des sujets mal timecodés dans les notices Ina.

52 La description d'un média consiste à affecter un graphe de connaissances à chaque segment de la *timeline*. Cette opération, un peu complexe, a été simplifiée dans Okapi et ramenée à l'édition d'un simple formulaire. Le logiciel Okapi suggère pour chaque champ du formulaire un ensemble pertinent de valeurs en interprétant l'ensemble de l'axiomatique présente dans l'ontologie du domaine. Les formules de l'axiomatique OWL2 sont interprétées comme des contraintes contextuelles qui permettent de réduire cet ensemble de valeurs et donc de réduire la charge cognitive de l'utilisateur en le guidant dans son travail d'annotation. Ces contraintes étant récursives, elles permettent également de définir contextuellement des classes anonymes, cela permet de réduire le nombre de classes nommées et de propriétés devant être déclarées dans l'ontologie du domaine. Prenons comme exemple le deuxième segment de la strate « Sujets » (sélectionné sur la *timeline* en figure 12) où l'on parle des « sports nautiques (concept) » en « Angleterre (Lieu) », en particulier les aventures du navigateur solitaire « Alec Rose (Personne) ». Ces annotations sont structurées autour des thèmes dont on parle, des lieux où cela se passe et des personnes qui y sont impliquées. Ces métadonnées sont représentées par un mini-graphe de connaissances et présentées à l'utilisateur sous forme d'un formulaire éditable (Figure 13). Ces concepts, lieux et personnes sont un sous-ensemble de la base des connaissances qui sont suggérées par Okapi pour compléter la description du segment en interprétant les contraintes posées sur les propriétés « thème », « à l'image » et « lieu ».

Figure 13. Formulaire de métadonnées de segment.

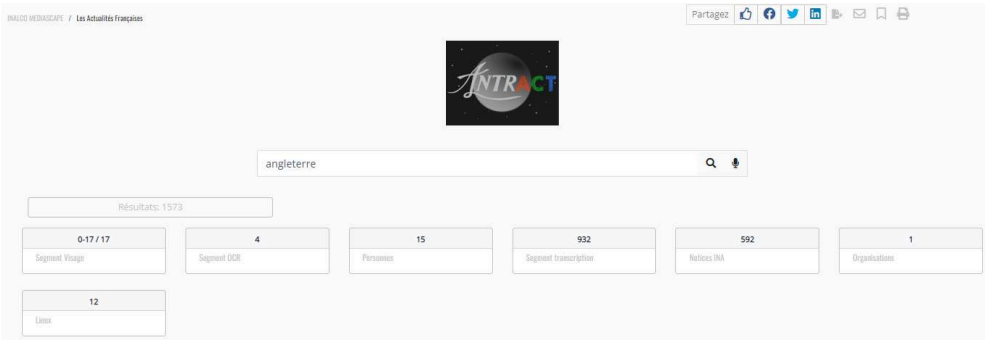


53 Les autres strates de description (transcription, détection de la musique, détection de la parole homme/femme, etc.) sont calculées automatiquement par des algorithmes. Les métadonnées produites viennent enrichir les métadonnées originelles des notices documentaires de l'Ina ou celles créées manuellement par les utilisateurs d'Okapi. L'ensemble de ces métadonnées sont utilisées par la plateforme Okapi pour générer un portail riche qui apporte de la valeur au contenu multimédia et offre plusieurs possibilités d'accès et de navigation dans ce contenu, comme le montre la figure 14.

Figure 14. Page du portail Okapi de l'émission « *Journal Les Actualités Françaises : émission du 10 juillet 1968* ».

54 Ces métadonnées viennent également alimenter les index Okapi pour la recherche en texte intégral sur les objets de la base. Ainsi, les métadonnées de transcription de la parole permettent de rechercher des passages dans le flux audio où l'on prononce certains mots-clés, celles de l'OCR sur les vidéos de retrouver les extraits vidéo où certains mots-clés sont affichés à l'écran, les informations extraites sur les visages de retrouver les extraits vidéo où l'on voit à l'écran le visage d'une personnalité donnée. Les résultats d'une recherche en texte intégral sont classés en fonction de leurs natures dans des catégories différentes (voir Figure 15) pour faciliter leur lecture et compréhension.

Figure 15. Recherche sur le texte intégral.



55 Ces métadonnées peuvent être aussi utilisées comme des critères avancés pour une recherche fine et sémantique de contenus. La figure 16 montre un exemple de recherche avancée d'extraits vidéo dans lequel on parle de « Sports nautiques » en « Angleterre ». Une recherche avancée est réifiée dans Okapi par un objet de type « requête sémantique » qui est stocké dans la base comme un graphe de connaissances et peut être retrouvé à l'aide du moteur de recherche, édité à l'aide d'un formulaire avant d'être transformé en requête SPARQL et exécuté par le serveur. Les résultats de cette requête, illustrés par la figure 17, peuvent être utilisés pour créer et/ou enrichir un corpus.

Figure 16. Exemple d'une requête Okapi.

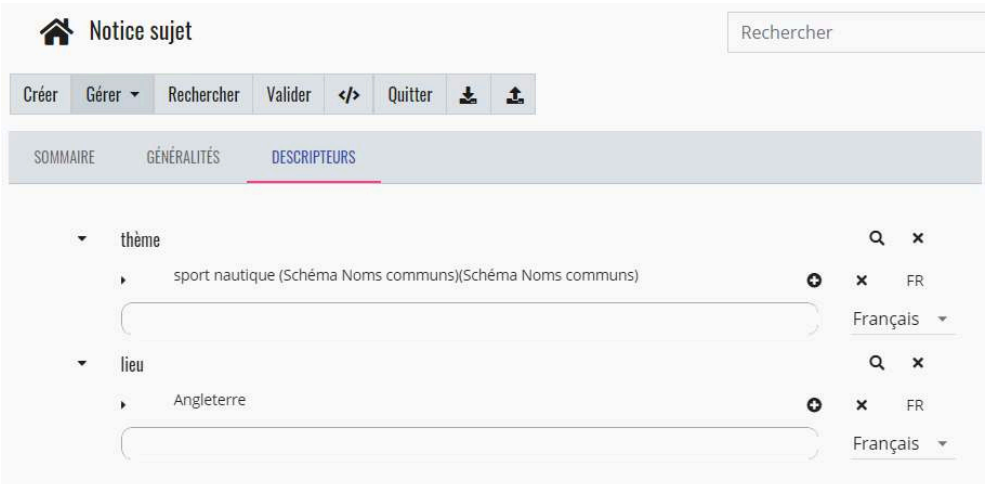


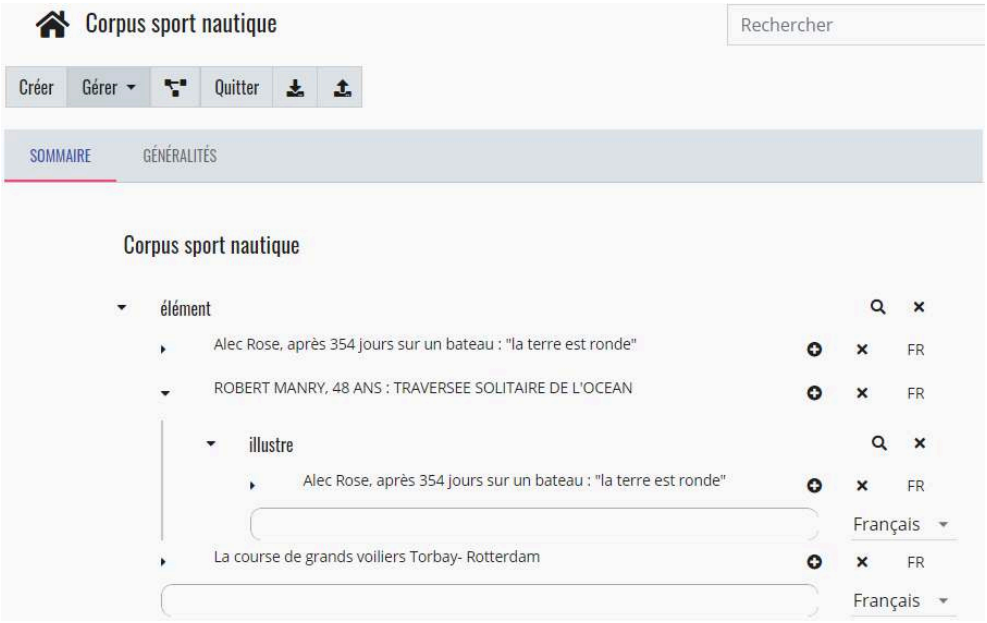
Figure 17. Exemple de résultats d'une requête.

56 Les outils de suggestions de données et de recherche sémantique présentés dans les paragraphes précédents permettent d'alimenter des corpus utilisateurs. Un corpus dans Okapi est un regroupement hiérarchique de contenus multimédias (extraits vidéo/audio, parties d'image, extraits PDF, points de vue caméra d'une scène 3D) qui partagent une certaine thématique. En fonction de cette thématique et de la taille du corpus, celui-ci peut être constitué de plusieurs sous-corpus, chacun abordant par exemple une sous-thématique. Une fonctionnalité de glisser-déplacer (*drag'n drop*) a été mise en place pour faciliter la réorganisation d'un corpus en déplaçant certains éléments entre ses sous-corpus.

57 Un corpus est aussi un objet de la base qui peut être annoté. L'utilisateur peut ajouter de nouvelles métadonnées sur le corpus lui-même ou sur ses éléments. Il peut également tirer des relations rhétoriques (exemplification, illustration, etc.) et discursives entre ces éléments. La figure 18 montre un corpus composé de trois extraits, récupérés à partir de la requête présentée dans le paragraphe précédent. Elle montre également une relation rhétorique entre les deux segments : « Robert Manry, 48 ans : Traversée solitaire de l'océan » qui illustre l'autre segment « Alec Rose, après 354 jours sur un bateau : "la terre est ronde" ». Toutes ces métadonnées peuvent être utilisées pour créer un portail thématique centré sur le contenu du corpus ou intégré à un récit par l'inclusion de contenu éditorial et de parcours de lecture.

58 La plateforme Okapi expose un *endpoint* SPARQL sécurisé et une API qui permet aux autres outils ANTRACT, en particulier à la plateforme TXM (Section « Analyse textométrique interactive »), d'interroger et d'enrichir certains objets de la base de connaissances. Par exemple, un utilisateur TXM peut récupérer un corpus par le biais du point d'accès Okapi, tirer parti des capacités de textométrie de l'outil TXM pour enrichir ce corpus et le renvoyer vers Okapi *via* son API. L'utilisateur peut ensuite reprendre ce corpus sur Okapi pour le compléter, le réorganiser et le publier sur le portail ou sous forme d'une publication auteur.

Figure 18. Corpus thématique « Sports nautiques ».



Conclusion

59 Présenté tout au long de ce chapitre, le défi du projet ANTRACT est de familiariser les chercheurs en sciences humaines et sociales avec l'analyse automatique et les nouvelles possibilités de recherche sur les grands corpus audiovisuels en contexte numérique. En rassemblant des instruments spécialisés dans l'analyse d'images, d'audio et de textes dans un environnement multimodal conçu pour corréler leurs résultats, le projet développe un modèle de recherche transdisciplinaire destiné à ouvrir de nouvelles perspectives dans l'étude de sources mono ou multiformat.

60 Une grande partie des travaux du projet a été consacrée au développement et au réglage des outils d'analyse automatique de contenu ainsi qu'à l'application de leurs résultats à l'organisation et à l'amélioration des données du corpus en lien avec les recherches fournies par les historiens d'ANTRACT (Goetschel, 2019 ; Carrive, Goetschel et Mazuet, à paraître). Des études de cas ont été menées en utilisant la plateforme de textométrie TXM et la plateforme d'annotation et de publication Okapi qui permettent à leurs utilisateurs d'exploiter les données produites par les instruments développés pour le projet.

61 D'un point de vue technologique, la transcription vérifiée réalisée en fin de projet est une nouvelle ressource qui ouvre d'importantes perspectives pour entraîner, adapter et évaluer les outils d'analyse automatique de contenu à la spécificité d'un tel corpus d'archives, comme son contexte historique, son vocabulaire, son format et sa qualité d'image.

62 À la fin du projet, un corpus complet, *Les Actualités Françaises,* complété par ses métadonnées ainsi que les résultats de la recherche obtenus par des outils d'analyse de contenu automatique et des annotations manuelles, ont été mis à la disposition de la communauté scientifique *via* la plateforme en ligne Okapi (pour la consultation et l'analyse en ligne) et *via* l'entrepôt de données Dataset de l'Ina (pour l'accès aux

données et aux corpus TXM)[9]. À cette fin, des didacticiels Okapi ont été réalisés et TXM continue d'être disponible en tant que logiciel libre pour faciliter l'analyse des corpus utilisés dans les nouvelles études de cas. Le code source d'Okapi doit être prochainement distribué en *open source* afin que d'autres développeurs puissent contribuer à son amélioration.

63 En ce qui concerne les sciences humaines et sociales, les outils et la méthodologie d'ANTRACT offrent aux historiens, mais aussi à des spécialistes d'autres disciplines – sociologie, anthropologie, sciences politiques –, la possibilité de disposer d'un corpus enrichi de très grande qualité, composé non seulement des sujets des *Actualités Françaises* entre 1945 et 1969, mais aussi des données obtenues grâce à l'usage des différents outils d'analyse automatique de contenu, considérablement améliorés au fil du projet. Plus globalement, les outils et la méthodologie d'ANTRACT ouvrent de nouvelles perspectives pour l'analyse multidisciplinaire de ce type de corpus.

BIBLIOGRAPHIE

Timo AHONEN, Abdenour HADID et Matti PIETIKAINEN, "Face description with local binary patterns: application to face recognition", *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 28.12, 2006, p. 2037-2041.

Scott ALTHAUS, Kaye USRY, Stanley RICHARDS, Bridgette VAN THUYLE, Isabelle ARON, Lu HUANG, Kalev LEETARU, Monica MUEHLFELD, Karissa SNOUFFER, Seth WEBER, Yuji ZHANG et Patricia PHALEN, "Global News Broadcasting in the Pre-Television Era : A Cross-National Comparative Analysis of World War II Newsreel Coverage", *Journal of Broadcasting and Electronic Media*, 62.1, 2018, p. 147-167.

Nathan S. ATKINSON, "Newsreels as Domestic Propaganda: Visual Rhetoric at the Dawn of the Cold War", *Rhetoric & Public Affairs*, 14.1, 2011, p. 69-100.

Ulrike BARTELS, *Die Wochenschau im Dritten Reich. Entwicklung und Funktion eines Massenmediums unter besonderer Berücksichtigung völkisch-nationaler Inhalte*, Francfort-sur-le-Main, Peter Lang, 2004.

Abdelkrim BELOUED, Peter STOCKINGER et Steffen LALANDE, « Studio Campus AAR : Une plateforme sémantique pour l'analyse et la publication de corpus audiovisuels », dans *Intelligence collective et archives numériques*, Hoboken, NJ, John Wiley & Sons Inc, 2017, p. 85-133.

Alex BEWLEY, Zongyuan GE, Lionel OTT, Fabio RAMOS et Ben UPCROFT, "Simple online and realtime tracking", C*onférence internationale de l'IEEE sur le traitement des images [ICIP]*, 2016, p. 3464-3468.

Chritian BIZER, Tom HEATH & Tim BERNERS-LEE, "Linked data – the story so far", *International Journal on Semantic Web and Information Systems*, 5, 2009, p. 1-22.

Gary BRADSKI, "The OpenCV Library", *Dr. Dobb's Journal of Software Tools*, 2000.

Pierre-Alexandre BROUX, Florent DESNOUS, Anthony LARCHER, Simon PETITRENAUD, Jean CARRIVE et Sylvain MEIGNIER, "S4D : Speaker Diarization Toolkit in Python", *Interspeech,* Hyderabad, Inde, 2018.

Qiong CAO, Li SHEN, Weidi XIE, Omkar M. PARKHI et Andrew ZISSERMAN, “VGGFace2. A dataset for recognising faces across pose and age”, *13e conférence internationale de l’IEEE sur la reconnaissance automatique des visages et des gestes (FG)*, 2018, p. 67-74.

Jean CARRIVE, Pascale GOETSCHEL et Franck MAZUET (dir.), *Pour une histoire outillée d’un corpus d’actualités filmées. Les Actualités Françaises (1945-1969)*. INA, L’Harmattan, coll. « Les Médias en Actes », à paraître en 2024.

Ciara CHAMBERS, Mats JÖNSSON et Roel VANDE WINKEL (dir.), *Researching Newsreels. Études de cas locales, nationales et transnationales*, Global Cinema, Palgrave Macmillan, Londres, 2018.

Jean-Hugues CHENOT et Gilles DAIGNEAULT, “A large-scale audio and video fingerprints-generated database of TV repeated contents”, *12th International Workshop on Content-Based Multimedia Indexing (CBMI),* Klagenfurt, Autriche, 2014.

Oliver CHRIST, “A modular and flexible architecture for an integrated corpus query system”, *in* Ferenc KIEFER *et al.* (dir.), *3rd International Conference on Computational Lexicography*, Research Institute for Linguistics, Hungarian Academy of Sciences, Budapest, 1994, p. 23-32.

Marilyn DEEGAN et Willard MCCARTY, *Collaborative Research in the Digital Humanities*. Ashgate, Farnham, Burlington, 2012.

ELAN (Version 5.2) [Logiciel informatique]. Institut Max Planck de psycholinguistique, Nimègue, 2018. Récupéré sur https://archive.mpi.nl/tla/elan

Seth FEIN, “New Empire into Old: Making Mexican Newsreels the Cold War Way”, *Histoire diplomatique*, 28.5, 2004, p. 703-748.

Seth FEIN, “Producing the Cold War in Mexico. The Public Limits of Covert Communications”, dans Gilbert M. JOSEPH et Daniela SPENSER (dir.), *In from the Cold : Latin America’s New Encounter with the Cold War*, Duke University Press, Durham, 2008, p. 171-213.

Ingo FEINERER, Kurt HORNIK et David MEYER, “Text Mining Infrastructure in R”, *Journal of Statistical Software*, 25.5, 2008, p. 1-54.

Pascale GOETSCHEL et Christophe GRANGER (dir.), « Faire l'événement, un enjeu des sociétés contemporaines », *Sociétés & Représentations*, 32, 2011, p. 7-23.

Pascale GOETSCHEL, « Les Actualités Françaises (1945-1969) : le mouvement d'une époque », *ANTRACT Analyse transdisciplinaire des actualités filmées,* 1, 2019, https://antract.hypotheses.org/127

Davis E. KING, “Dlib-ml: A machine learning toolkit”, *Journal of Machine Learning Research*, 10, 2009, p. 1755-1758.

Serge HEIDEN, “The TXM Platform: Building Open-Source Textual Analysis Software Compatible with the TEI Encoding Scheme”, dans Ryo OTOGURO, Kiyoshi ISHIKAWA, Hiroshi UMEMOTO, Kei YOSHIMOTO, Yasunari HARADA (dir.), *24th Pacific Asia Conference on Language, Information and Computation*, Institute for Digital Enhancement of Cognitive Development, Waseda University, 2010.

Nicolas HERVÉ, Pierre LETESSIER, Mathieu DERVAL et Hakim NABI, “Amalia.js : an Open-Source Metadata Driven HTML5 Multimedia Player”, *Open-Source Software Competition*, ACM Multimedia Conference 2015 (MM), October 2015, Brisbane, Australia.

Andreas HOTHO, Andreas NÜRNBERGER et Gerhard PAASS, “A brief survey of text mining”, *LDV Forum*, 20.1, 2005, p. 19-62.

Kornelia IMESCH, Sigrid SCHADE et Samuel SIEBER (dir.), *Constructions of cultural identities in newsreel cinema and television after 1945*, transcript-Verlag, MediaAnalysis, 17, 2016.

Ludovic LEBART, André SALEM et Lisette BERRY, *Exploring textual data. Text, speech, and language technology,* 4, Kluwer Academic, Dordrecht, Boston, 1998.

Sylvie LINDEPERG, *Clio de 5 à 7 : les actualités filmées à la Libération, archive du futur*, Paris, CNRS, 2000.

Sylvie LINDEPERG, « Spectacles du pouvoir gaullien : le rendez-vous manqué des actualités filmées », dans Jean-Pierre BERTIN-MAGHIT (dir.), *Une histoire mondiale des cinémas de propagande*, Paris, Nouveau Monde Éditions, 2008, p. 497-511.

Pasquale LISENA, Jorma LAAKSONEN et Raphaël TRONCY, "Understanding Videos with Face Recognition: A Complete Pipeline and Applications", *Multimedia Systems*, Special Issue on Data-driven Personalisation of Television Content, 28, 2022, p. 2147-2159.

Brian MCWHINNEY, *The Childes Project. Tools for Analyzing Talk*, L. Erlbaum Associates, Mahwah, N.J., 2000.

Sarah MAITLAND, "Culture in translation. The case of British Pathé News", *Culture and news translation, Perspectives. Études sur la théorie et la pratique de la traduction*, 23.4, 2015, p. 570-585.

Boris MOTIK, Peter PATEL-SCHNEIDER et Bijan PARSIA, *OWL 2 Ontology Language: Structural Specification and Functional-Style Syntax (seconde édition)*, Recommandation du W3C, 11 décembre 2012.

Bénédicte PINCEMIN, Serge HEIDEN et Matthieu DECORDE, "Textometry on Audiovisual Corpora. Experiments with TXM software", *15th International Conference on Statistical Analysis of Textual Data (JADT),* Toulouse, 2020.

Daniel POVEY, Arnab GHOSHAL, Gilles BOULIANNE, Lukáš BURGET, Ondrej GLEMBEK, Nagendra GOEL, Mirko HANNEMANN, Petr MOTLICEK, Yanmin QIAN, Petr SCHWARZ, Jan SILOVSKY, Georg STEMMER et Karel VESELY, "The kaldi speech recognition toolkit", *IEEE 2011 workshop on automatic speech recognition and understanding*, IEEE Signal Processing Society, Hilton Waikoloa Village, Big Island, Hawaii, US, 2011.

Daniel POVEY, Vijayaditya PEDDINTI, Daniel GALVEZ, Pegah GHAHREMANI, Vimal MANOHAR, Xingyu NA, Yiming WANG et Sanjeev KHUDANPUR, "Purely sequence-trained neural networks for ASR based on lattice-free MMI", *Interspeech*, San Francisco, 2016, p. 2751-2755.

Vladimir POZNER, « Les acutalités soviétiques Durant la Seconde Guerre Mondiale : nouvelles sources, nouvelles approches », dans Jean-Pierre BERTIN-MAGHIT (dir.), *Une histoire mondiale des cinémas de propagande*, Paris, Nouveau Monde Éditions, 2008, p. 421-444.

R Core Team, "R : A Language and Environment for Statistical Computing", R Foundation for Statistical Computing, Vienne, Autriche, 2014.

André SALEM, « Introduction à la résonance textuelle », dans Gérald PURNELLE *et al.* (dir.), *7es Journées internationales d'Analyse statistique des Données Textuelle*s, Louvain, Presses universitaires de Louvain, 2004, p. 986-992.

Helmut SCHMIDT, "Probabilistic Part-of-Speech Tagging Using Decision Trees", *Proceedings of International Conference on New Methods in Language Processing*, Manchester, UK, 1994.

Florian SCHROFF, Dmitry KALENICHENKO et James PHILBIN, "Facenet: A unified embedding for face recognition and clustering", *Actes de la conférence de l'IEEE sur la vision par ordinateur et la reconnaissance des formes*, 2015, p. 815-823.

Christian SZEGEDY, Vincent VANHOUCKE, Sergey IOFFE, Jonathon SHLENS et Zbigniew WOJNA, "Rethinking the Inception Architecture for Computer Vision", *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR),* Las Vegas, 2016.

Raphaël TRONCY, Erik MANNENS, Silvia PFEIFFER et Davy VAN DEURSEN, *Media Fragments URI 1.0 (basic),* Recommandation du W3C, 2012.

Laurent VERAY, *Les Films d'actualités français de la Grande Guerre*, Paris, SIRPA/AFRHC, 1995.

Paul VIOLA et Michael JONES, "Robust real-time face detection", *International Journal of Computer Vision*, 57.2, 2004, p. 137-154.

Sholom M. WEISS, Nitin INDURKHYA et Tong ZHANG, *Fundamentals of Predictive Text Mining*, Springer-Verlag, Londres, 2015.

Kaipeng ZHANG, Zhanpeng ZHANG, Zhifeng LI et Yu QIAO, "Joint face detection and alignment using multitask cascaded convolutional networks", *IEEE Signal Processing Letters*, 23.10, 2016, p. 1499-1503.

NOTES

1. ANTRACT : ANalyse TRansdisciplinaire des ACTualités filmées (1945-1969).
2. Corpus ESTER 1 & 2, EPAC, ETAPE, et REPERE disponibles dans les catalogues ELRA (http://www.elra.info/).
3. ETAPE et QUAERO, corpus disponibles dans les catalogues ELRA (http://www.elra.info/).
4. Challenge REPERE, données de test.
5. https://github.com/Linzaer/Face-Track-Detect-Extract.
6. Disponible à l'adresse http://facerec.eurecom.fr/.
7. L'application est accessible au public à l'adresse http://facerec.eurecom.fr/visualizer/?project=antract.
8. Text Encoding Initiative, https://tei-c.org.
9. Voir http://okapi.ina.fr.

AUTEURS

JEAN CARRIVE

Institut national de l'audiovisuel [Ina]

ABDELKRIM BELOUED

Institut national de l'audiovisuel [Ina]

PASCALE GOETSCHEL

Centre d'histoire sociale des mondes contemporains, UMR 8058 (université Paris 1/CNRS)

SERGE HEIDEN

Institut d'histoire des représentations et des idées dans les modernités [IHRIM], UMR 5317 (université de Lyon)

STEFFEN LALANDE

Institut national de l'audiovisuel [Ina]

PASQUALE LISENA

EURECOM

FRANCK MAZUET

Centre d'histoire sociale des mondes contemporains, UMR 8058 (université Paris 1/CNRS)

SYLVAIN MEIGNIER

Laboratoire d'informatique de l'université du Mans [LIUM]

BÉNÉDICTE PINCEMIN

Institut d'histoire des représentations et des idées dans les modernités [IHRIM], UMR 5317 (université de Lyon)

RAPHAËL TRONCY

EURECOM

Partie 2 - L'audiovisuel et ses métadonnées

Les archives audiovisuelles au tamis des archivistes

Sandrine Gill

1 « Pour vous, c'est quoi une archive audiovisuelle ? » Ainsi débutait un dossier thématique de la revue *Archivistes !* au début de l'année 2021[1]. Aussi simple soit-elle, cette interrogation amène une polysémie de réponses, pointant une grande diversité de conceptions et des pratiques chez les archivistes. Au-delà de cette communauté professionnelle, il ne va pas de soi de définir une archive audiovisuelle alors que les films d'archives apparaissent sur nos écrans et semblent à la portée de tous. Avant d'aborder les évolutions actuelles en termes de normes, d'outils, de compétences et d'usages numériques, il est nécessaire de rappeler les périmètres des institutions patrimoniales en charge des archives audiovisuelles. Comment la loi définit-elle les archives audiovisuelles ? Dans quel contexte les services d'archives, en particulier les Archives nationales, ont-ils été amenés à collecter des fonds audiovisuels et quelles sont leurs particularités ? Compte tenu de leur hybridité (supports physiques et données numériques), quelles problématiques soulèvent leur collecte, puis leur gestion, leur communication et leur pérennisation ? Comment les nouveaux usages remettent-ils en cause la dichotomie archivistique/documentaliste ? Les réflexions sont ouvertes, offrant différents champs de possibles.

Le contexte institutionnel et réglementaire des collectes de fonds audiovisuels

Les institutions publiques en charge d'archives audiovisuelles

2 La répartition de l'audiovisuel public s'est faite au fil du temps et de l'évolution des médias. Lors de la réforme de l'audiovisuel mise en place en 1975, l'Institut national de l'audiovisuel [INA] a hérité des archives de radio et de télévision de l'ORTF [Office de radiodiffusion-télévision française], puis s'est vu confier en 1992 le dépôt légal de l'audiovisuel. Cette même année, le Centre national du cinéma et de l'image animée [CNC], dont l'existence remonte à 1946, assure le dépôt légal pour les œuvres

cinématographiques françaises et étrangères, de court métrage et long métrage, diffusées en salles, dès lors qu'elles ont obtenu un visa d'exploitation, ainsi que les films publicitaires ou institutionnels. La Bibliothèque nationale de France [BnF], quant à elle, étend ses missions historiques de dépôt légal instaurées par François Ier aux documents sonores et vidéogrammes, documents multimédias édités, importés ou diffusés en France[2].

3 Cet encadrement législatif, qui s'est constitué progressivement, ne prévoit cependant pas le cas spécifique d'archives audiovisuelles produites par l'administration dont les services d'archives ont pour mission la collecte, le classement, l'inventaire, la conservation et la communication.

Figure 1. Carton d'archives avec des supports audiovisuels.

Fonds audiovisuel du Théâtre national de Chaillot, Arch. nat., 20160438.
© Archives nationales.

Que dit la loi sur les archives audiovisuelles ?

4 Pour éclairer sa lanterne, il est nécessaire de revenir à la définition des archives dans le code du Patrimoine (L211-1) : « Les archives sont l'ensemble des documents, y compris les données, quels que soient leur date, leur lieu de conservation, leur forme et leur support, produits ou reçus par toute personne physique ou morale et par tout service ou organisme public ou privé dans l'exercice de leur activité. » Les archives audiovisuelles n'ont pas de statut juridique spécifique, ce sont des archives sur un support particulier, que ce soit une pellicule, une cassette ou encore une donnée numérique. Les archives de la Justice constituent une catégorie à part (L221-1) : « Les audiences publiques devant les juridictions de l'ordre administratif ou judiciaire peuvent faire l'objet d'un enregistrement audiovisuel ou sonore dans les conditions

prévues par le présent titre lorsque cet enregistrement présente un intérêt pour la constitution d'archives historiques de la justice. »

Figure. 2. Extrait de l'affiche de l'exposition *Filmer les procès, un enjeu social*.

Archives nationales, 15 octobre 2020-14 mai 2021.
© Archives nationales.

5 Ces dernières diffèrent notamment par leur régime de communicabilité. Alors que « les archives publiques sont [...] communicables de plein droit » sauf exception (L213-1), les archives audiovisuelles de la Justice sont soumises à des règles spécifiques (L222-1) : « L'enregistrement audiovisuel ou sonore est communicable à des fins historiques ou scientifiques dès que l'instance a pris fin par une décision devenue définitive. La reproduction ou la diffusion, intégrale ou partielle, de l'enregistrement audiovisuel ou sonore est subordonnée à une autorisation accordée, après que toute personne justifiant d'un intérêt pour agir a été mise en mesure de faire valoir ses droits, par le président du tribunal judiciaire de Paris ou par le juge qu'il délègue à cet effet. Après cinquante ans, la reproduction et la diffusion des enregistrements audiovisuels ou sonores sont libres. »

6 Concernant la reproduction et la diffusion, les archives audiovisuelles du régime général n'en demeurent pas moins couvertes par le code de la Propriété intellectuelle (L112-2).

7 Ainsi, même si la majorité des archives audiovisuelles conservées en services d'archives ont un statut d'archives publiques, le statut et les volontés des auteurs prévalent. D'où l'importance de collecter des contrats de cession de droits, notamment pour les témoignages oraux[3]. Enfin, comme toute autre catégorie d'archives, les archives audiovisuelles sont aussi couvertes par le Code pénal, interdisant, par exemple, la diffamation publique. Cet ensemble de dispositifs législatif demeure cependant

généraliste et insuffisamment adapté, comme nous allons le voir, aux spécificités des archives audiovisuelles.

Une intégration tardive des archives audiovisuelles aux services d'archives

8 Les archives audiovisuelles sont donc des archives avant tout, une affirmation qui ne va pas de soi dès que l'on constate le delta qui existe parfois entre le temps de la production et celui de la collecte. Ainsi le film le plus ancien conservé aux Archives nationales, issu de la cinémathèque de l'Agriculture, date de 1905 ; ce n'est qu'en 1997 qu'il fit son entrée aux Archives nationales[4]. Plusieurs facteurs peuvent expliquer cette situation : chez le producteur, une prise de conscience tardive de l'intérêt archivistique des anciens documents audiovisuels ; dans les services d'archives publiques, l'intégration relativement récente des archives audiovisuelles.

9 Aux Archives nationales, c'est seulement au début des années 1980 que l'on commence à produire des archives orales puis à collecter systématiquement les archives audiovisuelles et électroniques au même titre que les archives papier, majoritaires. Dans un premier temps, sous l'impulsion de Chantal Tourtier-Bonazzi, les Archives nationales se dotent d'une cellule d'archives orales pour collecter des témoignages relatifs à la Seconde Guerre mondiale ou provenant de grands acteurs de la vie politique et culturelle de l'après-guerre. À la même époque, les Archives départementales de la Dordogne[5], du Tarn[6], de la Seine-Saint-Denis[7] et du Val-de-Marne[8] développent un secteur audiovisuel.

Figure. 3. Bobines de films.

Bobines de films de la cinémathèque de l'Agriculture en cours de préparation de versement.
© Archives nationales.

10 L'émergence de l'histoire orale en France n'est pas totalement étrangère à ce mouvement de prise en considération de l'audiovisuel comme source historique[9]. Au début des années 1980, les témoins de la Seconde Guerre mondiale se font plus rares, il devient urgent de recueillir leur parole. Figure pionnière dans ce domaine, la sociologue Dominique Schnapper est chargée en 1975 par le Comité d'histoire de la Sécurité sociale de constituer des archives orales. Dès l'origine, le corpus de témoignages a une visée historique et patrimoniale, aboutissant à un premier versement aux Archives nationales[10]. D'autres ministères reprendront cette pratique, enrichissant petit à petit les fonds d'archives orales.

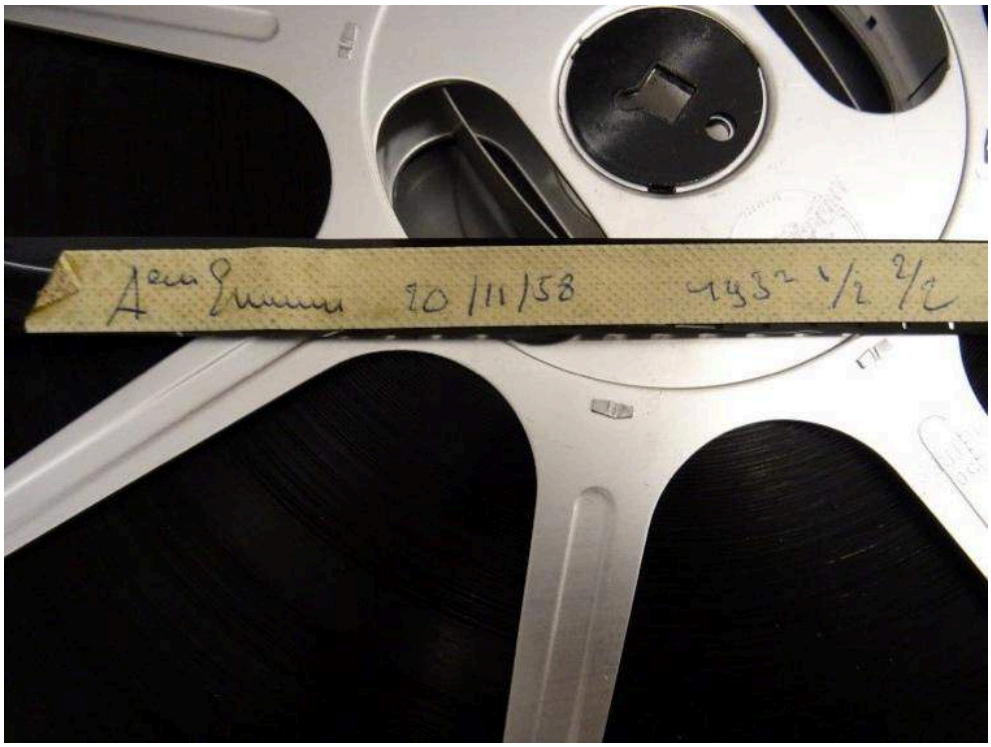
11 Malgré son intérêt manifeste, la considération du document audiovisuel comme une « archive comme une autre » reste cependant fragile dans la sphère archivistique. Pierre Carouge, directeur adjoint des Archives départementales de la Vienne, explique cette marginalité des archives audiovisuelles aussi bien par un manque de visibilité auprès du public qu'au sein de la profession[11]. Du fait de la technicité requise pour traiter un fonds audiovisuel, les archivistes responsables de fonds audiovisuels se forment souvent « sur le terrain » et dans l'interaction avec d'autres professions (techniciens, juristes). La récente réorganisation des Archives nationales tente de pallier cette situation en encourageant la prise en charge des archives audiovisuelles et numériques à un plus grand nombre d'archivistes.

Les spécificités des fonds audiovisuels : des corpus hétérogènes

Dès leur collecte, les archives audiovisuelles se singularisent

12 La difficulté à considérer un document audiovisuel comme une archive à part entière se mesure dès la collecte chez le producteur d'archives. Bien que les entrées de fonds audiovisuels, comme tout autre fonds d'archives, doivent s'inscrire dans une programmation définie en comités, c'est souvent à l'occasion d'un déménagement de locaux – qui peut intervenir d'une semaine à une autre, d'une année à l'autre – que les archivistes « pompiers » sont appelés à la rescousse. Il s'agit alors de collecter des masses de supports audiovisuels fragiles, conservés dans des caves et des greniers, oubliés pendant des décennies…

Figure 4. Détail d'une bobine accompagnée de métadonnées minimales.



Cinémathèque du ministère de la Reconstruction et de l'Urbanisme. Arch. nat., 20170146.

13 Dans le meilleur des cas où un inventaire est disponible, le plus souvent l'archiviste se transforme en archéologue chargé de reconstituer les strates de l'histoire du contexte d'une production. Puis il faut prendre en compte la spécificité des versements audiovisuels : captations ou enregistrements d'origine, rushes, versions et copies, sur toute une génération de supports représentatifs de l'évolution des techniques audiovisuelles, jusqu'au fichier numérique. Par principe de précaution, les Archives nationales collectent tous les supports sans distinction, tant il est difficile de déterminer celui qui conserve le mieux le document d'archives. L'exemple de la captation d'*Ubu Roi*[12], une pièce d'Alfred Jarry mise en scène par Antoine Vitez au Théâtre national de Chaillot en 1985, est éloquent : sur les 18 supports audiovisuels versés, seule une cassette VHS contenait l'intégralité du spectacle. En général, les archivistes s'intéressent en priorité aux supports de format professionnel (comme les Betacam par exemple), les cassettes VHS ne sont pas les meilleures pièces pour la conservation, car il s'agit souvent de copies de travail de piètre qualité. Mais dans ce cas précis, si cette cassette avait été écartée l'archive aurait été irrémédiablement perdue…

14 Par ailleurs, lorsqu'un film ou un enregistrement sonore arrive aux Archives nationales, il n'est pas exclu qu'il se trouve conservé sous une autre forme dans une autre institution chargée du dépôt légal ou de collections constituées à partir de fonds privés ou publics. Le fonds audiovisuel du Théâtre national de Chaillot constitué principalement de captations et d'enregistrements de pièces illustre parfaitement ce phénomène : deux versements ont été effectués aux Archives nationales en 2016 et

2017[13], l'INA a numérisé les supports et en conserve une copie[14], la BnF, quant à elle, compte parmi ses collections un grand nombre d'enregistrements sonores[15]…

De l'entrée à la pérennisation des archives audiovisuelles en service d'archives

15 La préparation en amont d'une entrée d'un versement d'archives audiovisuelles ne constitue qu'une étape de son traitement, elle se poursuit souvent pendant un temps plus ou moins long en service d'archive. C'est à l'archiviste qu'incombe la charge de gérer de façon différenciée les supports physiques et le numérique, de restituer l'authenticité et l'intégrité du fonds sans masquer les manques éventuels. Il importe aussi de prendre en considération la temporalité, le décalage entre le temps de la production et celui du versement des archives, les collectes d'aujourd'hui. Parfois il est difficile de retracer avec précision le chemin parcouru par les archives : une série de bobines de films d'éducation sanitaire du ministère de l'Hygiène sociale des années 1930 est retrouvée dans les caves de la direction de l'Action sociale de l'enfance et de la santé [DASES] de la Ville de Paris en 2021. On présume que les cartons contenant ces bobines furent embarqués dans le déménagement d'un dispensaire sans qu'il soit possible aujourd'hui d'en reconstituer précisément l'historique.

Figure 5. Bobines de films du ministère de l'Hygiène sociale.

Arch. nat., 20220196.

16 La phase de découverte des archives ne s'arrête pas à celle des supports physiques, leur contenu n'étant parfois révélé qu'au moment de la numérisation. En effet, on évite de

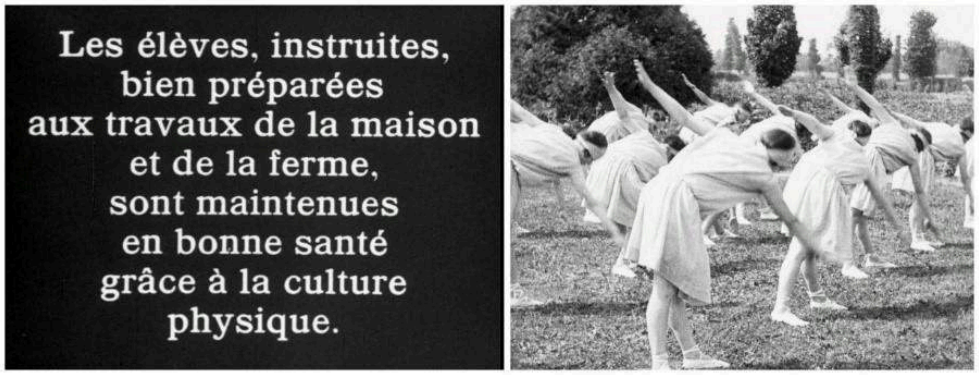
consulter une archive sur support fragile avant l'opération de numérisation pour ne pas risquer de l'endommager et de la perdre à jamais.

17 Enfin les archives nativement numériques posent d'autres problèmes : faut-il les traiter comme des archives audiovisuelles dans la continuité des archives sur supports physiques ou au contraire les fondre dans la masse des données numériques ? Une vigilance particulière s'impose au moment de la collecte pour comprendre le contexte de production et les usages prévus, départager les formats de projets des formats d'archivage, de consultation, de reproduction ou de diffusion. Cette attention participe à la stratégie de pérennisation des Archives nationales, à son engagement à vouloir mettre en place les moyens garantissant la lisibilité et l'exploitabilité des formats dans le temps. Elles participent à la cellule nationale de veille sur les formats, inter-institution, cadre de réflexion et de partage d'expériences[16]. Les connaissances en ce domaine évoluant sans cesse, tous les formats numériques ne peuvent être pris en charge de la même manière à un instant « T ». Les Archives nationales ont choisi de catégoriser les formats : certains sont « acceptés », ils sont préservés avec un audit régulier ; d'autres sont « tolérés », collectés et conservés en l'état, sans engagement sur leur maintenabilité dans le temps ; ou encore « refusés » en raison de problématiques techniques ou en l'absence de valeur juridique ou d'intérêt historique. Un tableau de la politique des formats, régulièrement mis à jour, est à la disposition des professionnels[17]. Parallèlement à cette démarche, les Archives nationales s'interrogent, au sein d'instances collectives, sur la manière de faire évoluer le SEDA, standard généraliste par excellence, afin d'y inclure les métadonnées indispensables à la pérennisation des archives. Elles sont un des membres actifs de la cellule nationale de veille sur les formats[18] et interviennent régulièrement dans des colloques ou journées d'étude[19].

Quels fonds audiovisuels sont conservés aux Archives nationales ?

18 Les problématiques juridiques, matérielles, numériques esquissées, il est temps de s'interroger sur le cœur même des documents audiovisuels. De quoi s'agit-il ? Aux Archives nationales, la collecte initiale d'archives orales dans les années 1980 se complète progressivement d'archives audiovisuelles produites et collectées par les administrations publiques de l'État ou d'origine privée[20]. Le corpus de fonds audiovisuels des Archives nationales, qui ne cesse de s'accroître, représente un ensemble si protéiforme qu'il serait vain de vouloir le synthétiser. En comparaison avec d'autres institutions patrimoniales comme l'INA ou la BnF, il reste relativement modeste (environ 60 000 supports et 80 To de données), les Archives nationales n'étant pas toujours bien identifiées comme institution susceptible de conserver des archives audiovisuelles.

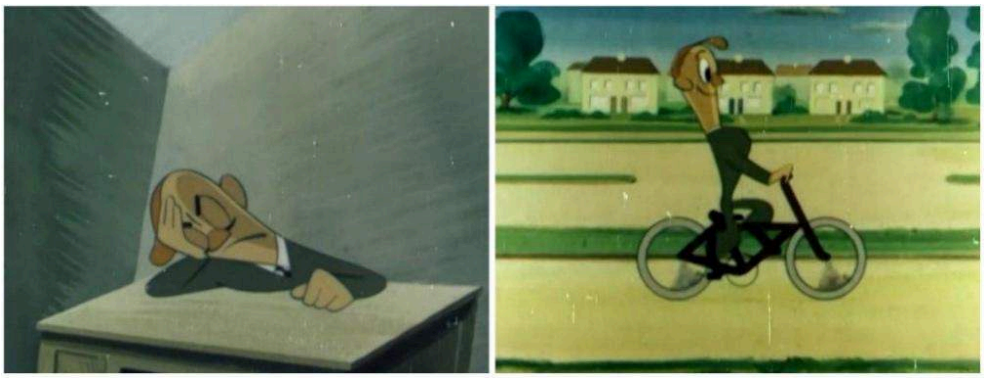
Figure 6. Captures d'écran du film de Jean Benoit-Lévy, *Centre d'enseignement agricole et ménager de Coëtlogon*, 1929, noir et blanc, 20'17", muet.



Cinémathèque du ministère de l'Agriculture, Arch. nat., 19970030/78-19970030/79.
© Archives nationales.

19 De fait, les fonds audiovisuels des Archives nationales sont aussi hétérogènes dans leurs sujets que leurs formes : une grande majorité de documentaires, mais aussi des reportages, des enquêtes, des témoignages, des créations graphiques, des animations, des sons et des images brutes ou montées, en différentes versions. Ils sont attribués à des réalisateurs professionnels ou non, à des équipes techniques des ministères. Parfois il n'est plus possible de retrouver la trace des auteurs. Parmi les corpus les plus significatifs, on citera les cinémathèques créées dans certains ministères, dès 1920 au ministère de l'Agriculture, en 1946 au ministère de la Reconstruction et de l'Urbanisme pour documenter et promouvoir l'action gouvernementale et de manière plus générale s'informer sur l'actualité.

Figure 7. Captures d'écran du dessin animé de John Halas et Joy Batchelor, *New town, la ville heureuse,* 1950, couleur, 8'12.



Cinémathèque du ministère de la Reconstruction et de l'Urbanisme, Arch. nat., 20170146/17.
© Archives nationales.

20 Des thèmes, qui peuvent apparaître aussi peu cinégéniques que l'attribution de subventions ou des certifications de l'État, font l'objet de films pédagogiques (mise en abîme des dossiers de subvention filmés dans *Amélioration de l'habitat rural*[21]), romancés (la vie trépidante d'un contrôleur de semences de luzerne dans *L'Or de la Durance*[22]) ou réaliste (le rude quotidien d'une vieille dame dans un logement insalubre dans *L'aménagement de l'habitat des personnes âgées*[23]). Au-delà du sujet, aucun de ces films ne laisse le spectateur indemne. La saveur des archives se niche aussi parfois là où on ne

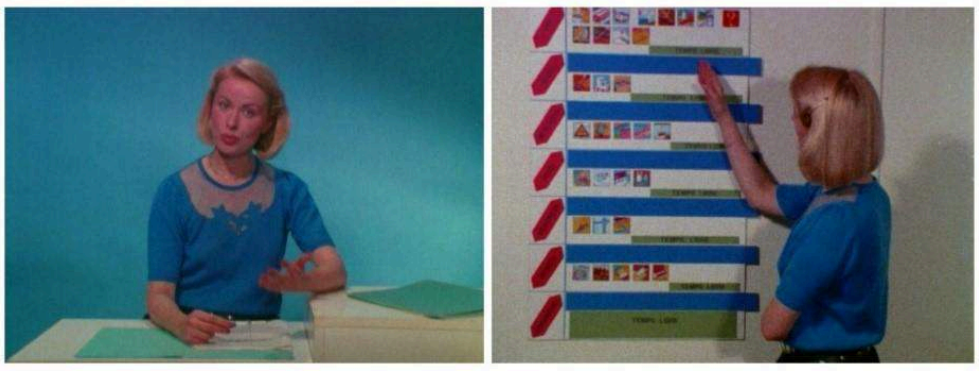
l'attend pas : la sonorité de l'ambiance électrique des rues de Lomé, lors de la visite officielle du président Georges Pompidou au Togo le 22 novembre 1972[24], ne saurait être rendue avec la même acuité par des archives photographiques[25].

Figure 8. Captures d'écran du film de Pierre Franceschi, *Améliorations de l'habitat rural*, 1959, noir et blanc, 14'.

Cinémathèque du ministère de l'Agriculture, Arch. nat., 19960237/23.

21 L'intérêt du film d'archive réside parfois moins dans ses qualités cinématographiques que dans l'angle d'approche si particulier des administrations de l'État. Comment éduquer Martine, 35 ans, femme au foyer, mère de deux enfants, à l'organisation de ses tâches ménagères quotidiennes[26] ? La Caisse nationale de l'assurance maladie des travailleurs salariés [CNAMTS] a produit dans les années 1970 et 1980 une série d'émissions destinées à être diffusées à la télévision et dans les réseaux professionnels. Par le truchement du récit de Martine, elle incite les femmes à adopter les bonnes postures, à gérer leur temps et leur espace familial afin de moins tomber malades… et de maîtriser les dépenses de la Sécurité sociale.

Figure 9. Captures d'écran du film de la série télévisée *Objectif santé*, *Ménagères, ménagez-vous*, vers 1983, couleur, 10'55".



Caisse nationale d'assurance maladie des travailleurs sociaux, Arch. nat., 19990035/26, 19990035/819, 19990035/820 (films préparatoires) ; 19990081/298 (cassette VHS) ; 20130089/14 (fichier numérique issu d'un DVD).

22 Dans ce corpus, les archives audiovisuelles de la Justice conservées exclusivement aux Archives nationales constituent des sources historiques exceptionnelles. Le procès de

Klaus Barbie fut le premier filmé, grâce à la loi du 11 juillet 1985 promue par le ministre de la Justice, Robert Badinter. Depuis lors, d'autres procès portant sur la Seconde Guerre mondiale, la dictature chilienne, le génocide des Tutsi au Rwanda ou les actes de terrorisme sont filmés pour la postérité. Un acte de transparence des débats que retrace l'exposition itinérante, *Filmer les procès, un enjeu social*[27].

La description des archives audiovisuelles

De l'ISAD-G au SEDA, les normes archivistiques

23 Pour décrire et gérer ces fonds, les archivistes sont contraints par des normes spécifiques à leur métier, non à l'audiovisuel. À la différence des documentalistes et bibliothécaires qui ont développé des logiciels spécifiques à l'audiovisuel, les services d'archives publics ne font aucune différence entre les types d'archives, quel que soit leur support. Les archives audiovisuelles sont donc soumises aux mêmes normes de description archivistiques : la norme internationale ISAD(G)[28] publiée par le Conseil international des archives en 1994 (rééditée en 1999) pour les supports physiques, basée sur l'EAD-DTD, standard d'encodage ; le standard SEDA[29], développé en 2006 par les Archives de France et l'ancienne Direction générale de la modernisation de l'État [DGME], pour les données. Dans l'ISAD(G), « l'objet de la description archivistique est d'identifier et d'expliquer le contexte et le contenu des documents d'archives, en vue de faciliter leur accès[30] ».

Figure 10. Captures d'écran de la captation de *Chants de la destinée (Song of pensive Beholding)* du Legend Lin Dance Theatre, novembre 2011.

Fonds audiovisuel du Théâtre national de Chaillot, Arch. nat., 20160438/428.

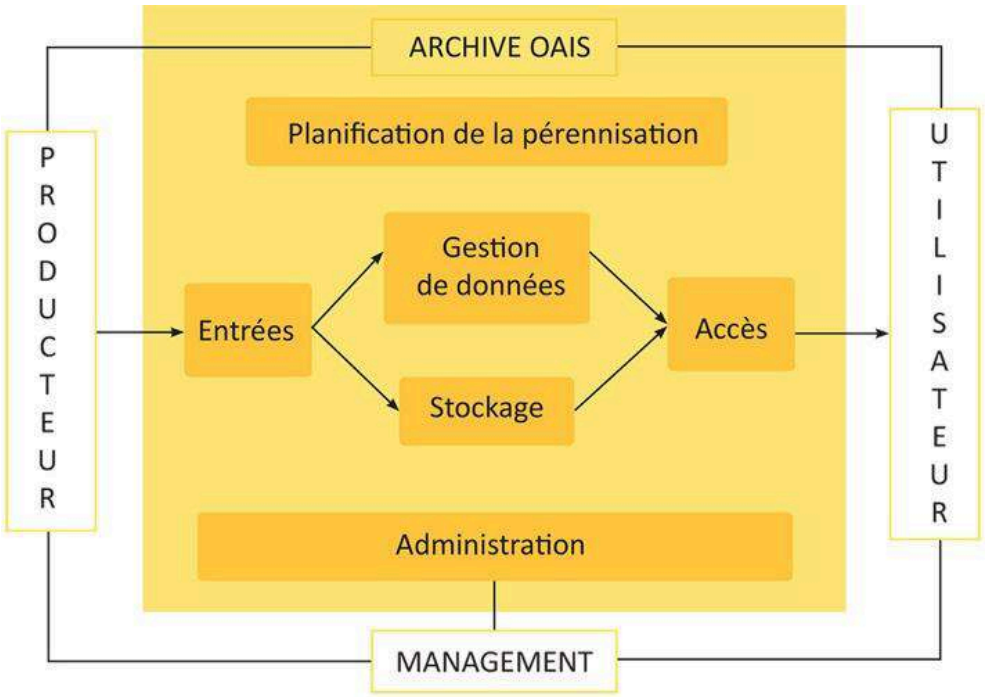
24 Le standard d'échange de données pour l'archivage (SEDA) « qui est autant archivistique que technique, s'inspire de normes existantes et des habitudes archivistiques utilisées dans les procédures papier. Si la description archivistique à plusieurs niveaux du SEDA est issue des normes ISAD-G/EAD, son modèle organisationnel adopte celui de la norme ISO 14721 (OAIS) et la préservation des informations techniques qu'il transporte emprunte les définitions au modèle PREMIS. Le SEDA est techniquement structuré en XML. »

La description des archives audiovisuelles, de l'instrument de recherche au bordereau de transfert

25 Le contexte normatif est sans doute ce qui singularise le plus la description, la gestion, l'accès et la pérennisation des archives audiovisuelles en service d'archives. Les instruments de recherche, inventaires analytiques propres au monde des archives, sont élaborés par les archivistes à l'aide de différents logiciels, puis mis à la disposition des lecteurs. Leur structure est commune aux archives sur tout support, papier, audiovisuel ou photographique. Après une introduction qui permet de contextualiser l'ensemble du fonds ou du sous-fonds, les articles peuvent être regroupés et classés thématiquement, chronologiquement, alphabétiquement, etc., de manière à restituer au plus près l'activité du producteur. Il ne s'agit pas d'un catalogue de films ou d'enregistrements, mais d'un classement archivistique qui restitue à chaque article sa place dans le contexte du fonds. Qu'il soit rédigé dans un document unique ou intégré à l'aide d'un import d'un fichier de récolement enrichi, le corps de l'instrument de recherche conserve sa logique.

26 Pour les archives nativement numériques, le SEDA, standard d'échange, est moins riche que l'ISAD-G en termes de description, mais reprend la même logique arborescente de description d'un fonds, complétée de métadonnées techniques (format, empreinte numérique, etc.) automatiquement récupérées. La description archivistique se fait dans des outils de constitution de paquets d'information[31], qui permettent à l'archiviste de décrire le fonds et les groupements d'articles, comme il le ferait dans un instrument de recherche. La riche contextualisation du fonds, telle qu'elle apparaît dans l'introduction d'un instrument de recherche en ISAD-G, ne trouvant pas sa place dans ce standard, elle est directement renseignée dans un des onglets de la plateforme d'archivage numérique des Archives nationales, fruit du projet ADAMANT, opérationnelle depuis novembre 2018[32].

Figure 11. Schéma des entités fonctionnelles de l'OAIS.



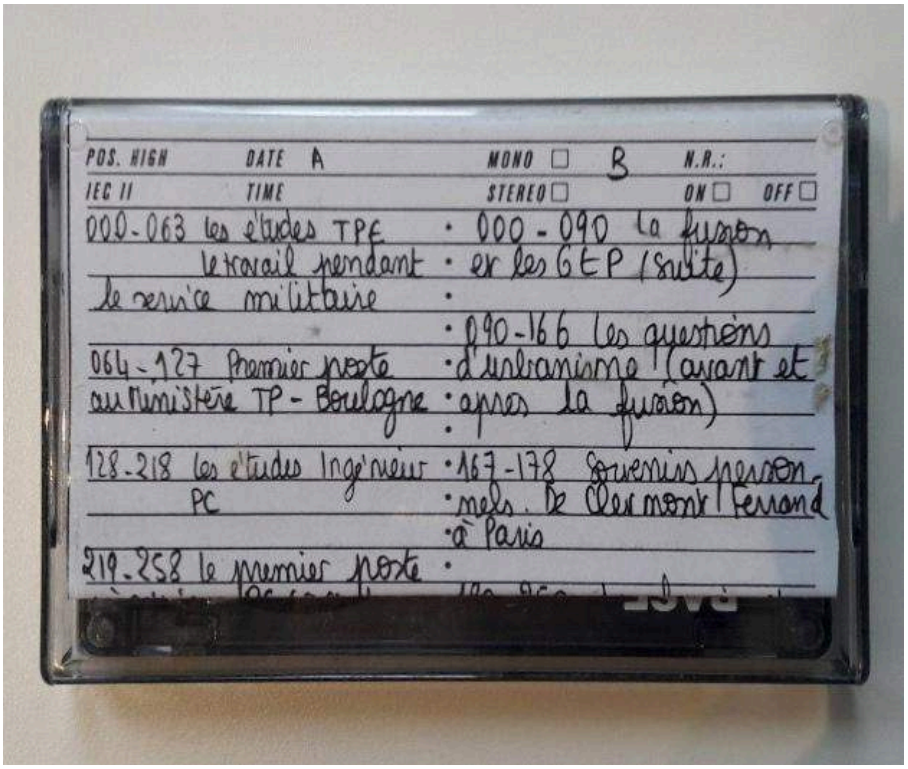
27 Selon la norme OAIS[33], l'entrée, la gestion et la dissémination des archives numériques dans un système d'archivage numérique sont composées de trois phases. Il faut d'abord constituer un SIP [*Submission Information Package*]. Les SIP, une fois élaborés, contiennent à la fois les fichiers d'archives et les métadonnées descriptives, techniques et de transfert dans un fichier XML nommé « manifest ». Ce SIP est ensuite téléchargé dans la plateforme d'archivage numérique. Lors de sa phase de gestion, il est nommé AIP [*Archival Information Package*]. Toute communication à l'extérieur génère un DIP [*Dissimination Information Package*]. De fait, contrairement à l'instrument de recherche qui est immuable une fois publié, le paquet qui contient à la fois les données et les métadonnées change de statut et d'information selon la phase dans laquelle il se situe.

Des normes à adapter aux archives audiovisuelles

28 Comment concilier la description des archives audiovisuelles, et des normes qui n'ont pas prévu d'en intégrer les métadonnées descriptives et techniques spécifiques ? Dans la culture « métier », les archives se conjuguent au pluriel, on traite des masses et non des individualités ou tout à fait exceptionnellement. Un instrument de recherche est adapté à la description d'un fonds d'archives, pas à une pièce unique. Or chaque film ou enregistrement a ses particularités. Aucune balise dans l'ISAD(G) ne permet de décrire de manière satisfaisante des œuvres de l'esprit couvertes par le droit de la propriété intellectuelle. Un utilisateur a besoin de connaître le nom et la fonction des différents auteurs qui ont contribué à l'œuvre, dont les droits (moraux, patrimoniaux, voisins) et la liste diffèrent en fonction du type d'œuvres, du contexte de production et des informations connues : réalisation, production, mise en scène, auteur de l'œuvre adaptée, musique, décor, costume, acteurs, animation, dessins, montage, mixage, scénario, etc.

29 Dans le cas des témoignages oraux, il est important de connaître et de respecter la volonté des témoins et des collecteurs exprimée dans des contrats (autorisations de communication, de réutilisation, de diffusions, conditions particulières énoncées). Le fait qu'une œuvre ait déjà été diffusée (à la télévision, à la radio, sur le web, dans un festival, un événement) a son importance, il permettra de mieux définir ses usages en services d'archives. Par ailleurs, les dates peuvent aussi être distinguées (captation ou enregistrement, copie, diffusion), de même que les lieux (tournage, enregistrement, production, diffusion). Les objectifs de la réalisation d'un film ou d'un document sonore sont tout aussi essentiels pour situer l'œuvre : captation brute à des fins d'archivage, matériaux sonores destinés à composer une émission, reportage ethnographique, document pédagogique.

Figure 12. Cassette audio d'un témoignage.



Archives orales des ministères de l'Équipement et de l'Environnement (1994-2016), Arch. nat., 20170332.

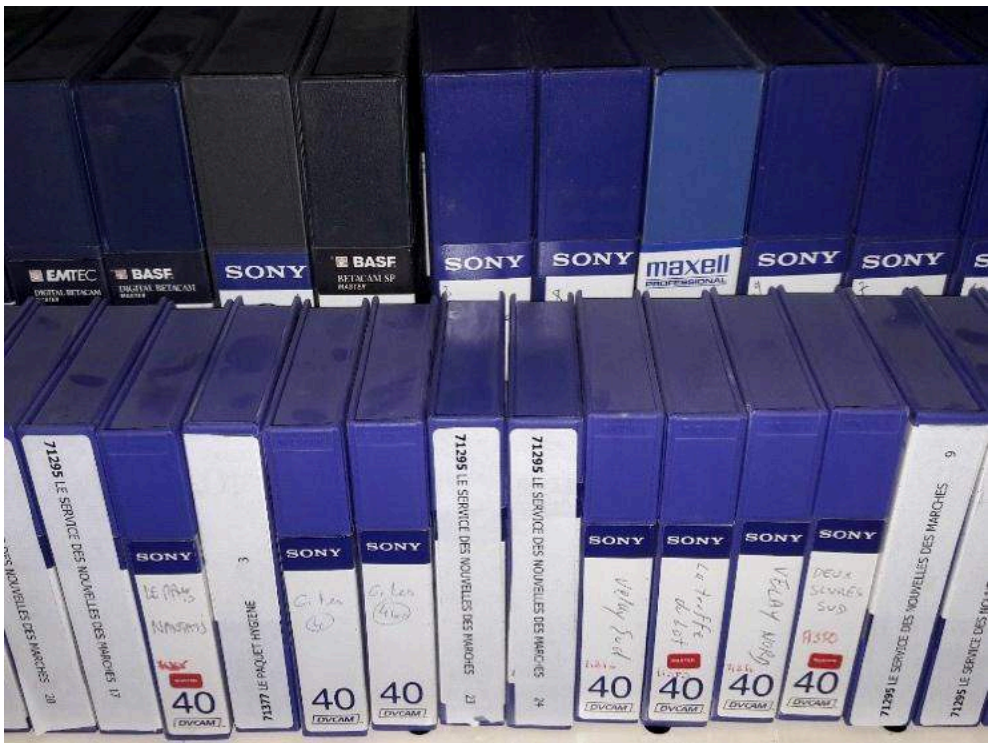
30 Certaines métadonnées descriptives sont spécifiques à des contextes. Pour décrire les fonds audiovisuels du Théâtre national de Chaillot[34], les Archives nationales ont retenu les notions de répétition, d'audition d'un acteur pour un spectacle, du montage d'un décor pour un spectacle ou de la représentation en tournée. Pour des émissions produites par un ministère, il sera pertinent de noter leur appartenance à une série ou à un ensemble identifié.

31 Enfin, il importe de relever les caractéristiques formelles de l'archive (rush, montage, extrait, master, plan large, zoom, travelling, prises de vue directes, animations, reportages, images d'archives intégrées, voix off, musique, etc.), les versions (française, anglaise, sous-titrage), la durée.

32 Ni l'ISAD(G) ni le SEDA n'étant aussi précis et exhaustifs dans les balises qu'ils proposent, la description des archives audiovisuelles, garante aussi bien de leur accès que de leur pérennisation, requiert adaptabilité et rigueur de la part des archivistes. Pour Audrey Clergeau, archiviste aux Archives départementales de Loire-Atlantique, et Aurélien Durr, des Archives départementales de Seine-Saint-Denis, les archives audiovisuelles font figure d'anticonformistes dans le monde des archives, provoquant une grande disparité dans les pratiques de description[35].

33 Aux Archives nationales, des fiches-guides ont été rédigées pour spécifier les attendus de description des archives audiovisuelles en ISAD(G) et en SEDA. Les solutions proposées, à adapter au cas par cas, car chaque versement a ses spécificités, permettent d'établir une certaine cohérence de pratiques. Elles nécessitent de la part des utilisateurs une veille régulière.

Figure 13. Cassettes DvCAM.



Cassettes DvCAM de la cinémathèque de l'Agriculture en cours de préparation de versement.
© Archives nationales.

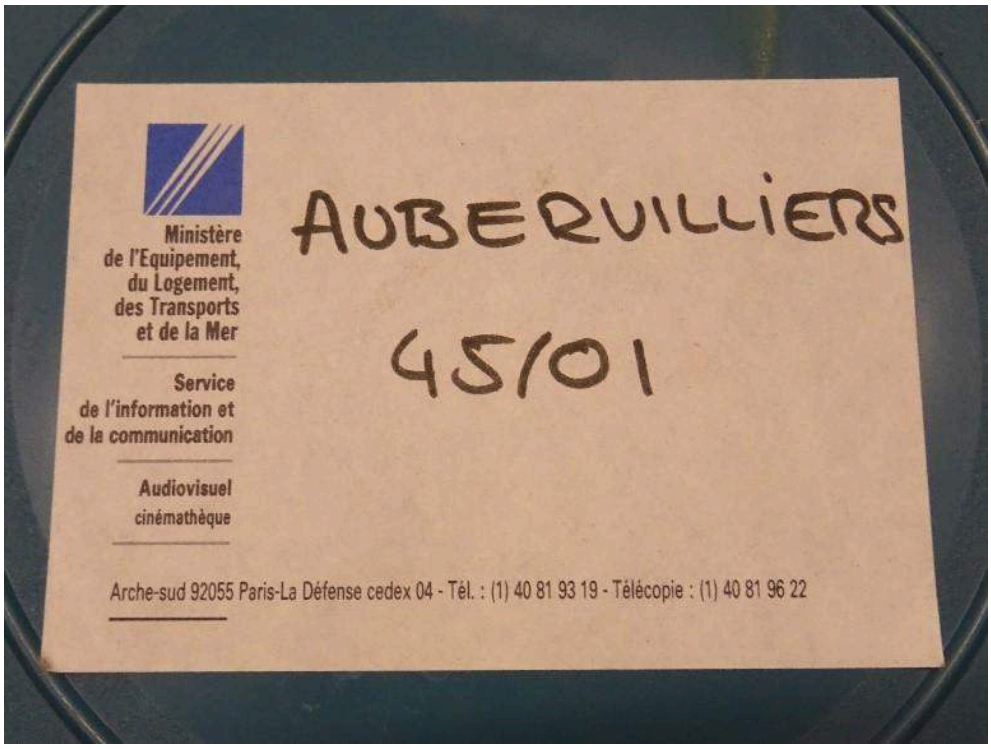
34 Des initiatives de normalisation de la description des archives audiovisuelles ont également été déployées dans les services d'Archives départementales. Audrey Clergeau et Aurélien Durr rapportent l'existence de guides de procédures permettant de faire converger les pratiques de description et de les rapprocher de la norme ISAD(G), mais appellent de leurs vœux la conception d'un standard commun au niveau national. Dans cette perspective, l'Association des archivistes français [AAF] a organisé en octobre 2022 des rencontres professionnelles *La pellicule, la bande magnétique et l'archiviste*[36].

Faut-il choisir entre l'approche archivistique et l'approche documentaire ?

S'affranchir partiellement des normes archivistiques ?

35 Pour le moment, la plupart des services d'archives utilisent des outils « métier » de description archivistique fondés sur l'ISAD(G). Différents logiciels existent sur le marché, mais le principe est le même : décrire un film ou un enregistrement dans le contexte de son fonds d'archives. Chacun en aura fait l'expérience, en fonction de la manière dont un titre a été renseigné par l'archiviste, il est plus ou moins facile de retrouver l'archive, encore moins de la mettre en relation avec sa « jumelle » ou sa « cousine » contenue dans un autre fonds d'archives, les supports de diffusion circulant beaucoup. Il n'est pas rare qu'un même film arrive en service d'archives par le biais de deux versements distincts.

Figure 14. Étiquette d'une bobine 16 mm du film d'Eli Lotar *Aubervilliers*.



Cinémathèque du ministère de la Reconstruction et de l'Urbanisme, Arch. nat., 20170146/8.

36 Par exemple, le documentaire réalisé en 1945 par Eli Lotar, *Aubervilliers*, qui témoigne des misérables conditions d'existence de la classe ouvrière, logée dans des taudis après la Seconde Guerre mondiale, se retrouve aussi bien dans la cinémathèque du ministère de la Reconstruction et de l'Urbanisme[37] que dans celle de l'Agriculture[38]. Dans le premier cas, ce documentaire s'inscrit en toute logique dans le classement chrono-thématique qui précède la Reconstruction, dans le second, il fait partie de films de même nature commandités par d'autres services cinématographiques para-administratifs dont la cinémathèque avait acquis des copies pour en faire bénéficier son

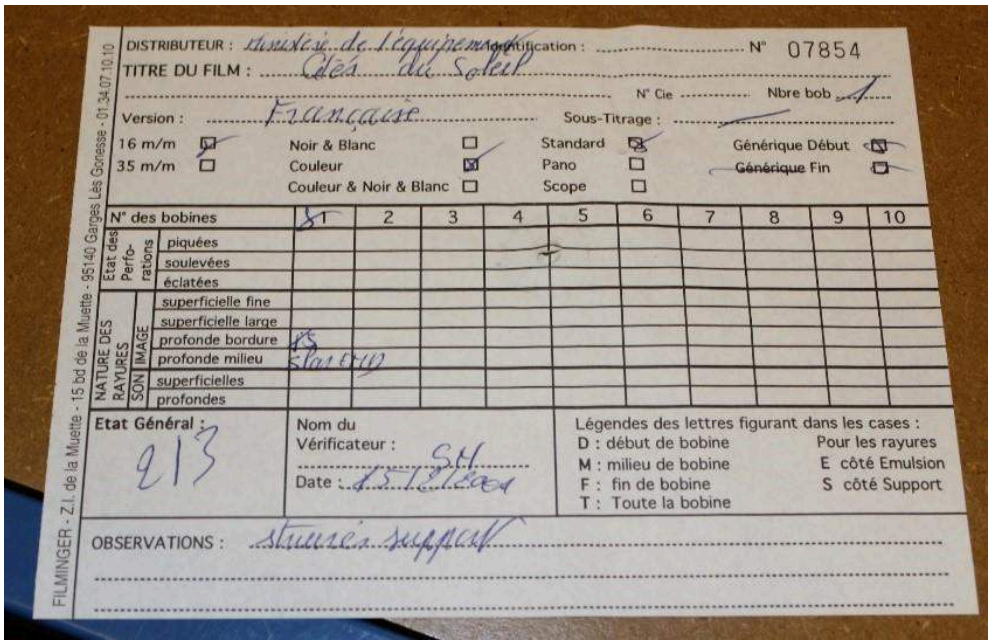
réseau de diffusion. Le classement alphabétique de cet inventaire ne permet pas de contextualiser historiquement ce même film.

37 Pour le lecteur, l'accès au contenu de l'archive se heurte à un autre écueil : on ne consulte pas un film ou des heures d'enregistrement comme on feuillette un carton d'archives. La consultation nécessite *a minima* une version numérique (native ou issue de la numérisation), la numérisation préalable du support le cas échéant, la préparation technique du fichier, l'accompagnement de métadonnées (fiche descriptive, chrono-thématique, etc.), un poste de consultation sécurisé pour éviter toute copie. Car, contrairement à l'image que l'on peut avoir de la diffusion sur Internet, seule une petite proportion d'archives audiovisuelles publiques ou des archives privées acquises par les institutions peut être mise en ligne par des services d'archives : la plupart sont couvertes par le droit d'auteur[39]. De nouveaux outils adaptés au numérique sont en cours de développement[40], mais la consultation de la grande majorité des archives audiovisuelles en services d'archives se fait encore en salle de lecture, selon les moyens mis à disposition dans chaque structure.

D'autres manières d'aborder les archives audiovisuelles

38 Cette conception archivistique se distingue de l'approche documentaire d'autres institutions patrimoniales qui conservent des archives audiovisuelles. Dans les catalogues en ligne de l'INA[41], de la BnF[42], du CNC[43], de l'ECPAD[44], chaque film ou enregistrement fait l'objet d'une notice individuelle. Ainsi, il est nettement plus aisé de retrouver un titre accompagné de ses métadonnées descriptives et techniques. Il est en revanche plus difficile de reconstituer précisément sa redécouverte, son contexte de production, l'histoire des copies, des usages, des prêts, du public destinataire, etc.

Figure 15. Fiche technique d'une bobine de film.



Cinémathèque du ministère de la Reconstruction et de l'Urbanisme, Arch. nat., 20170146.

39 Entre l'une ou l'autre approche, pourquoi nécessairement choisir ? Dans le monde normalisé des archives, pour faire connaître des mines d'or de documentaires audiovisuels (en grande majorité) méconnus, les ministères de la Transition écologique, de la Cohésion des territoires et de la Mer ont opté pour les deux, en faisant, d'une part, des versements réglementaires aux Archives nationales[45] et, d'autre part, en mettant en ligne une sélection de photographies et de films patrimoniaux numérisés dans leur médiathèque numérique TERRA[46]. Les films consultables en ligne sont accompagnés de notices individuelles comprenant des métadonnées descriptives et techniques. Destinés à un usage non commercial, ils peuvent être téléchargés en différents formats sur inscription. Ainsi les mêmes titres sont désormais accessibles de deux manières différentes.

40 Les Archives départementales de l'Ain, quant à elles, n'ont pas hésité à franchir le pas vers le documentaire en adoptant, à travers un partenariat avec la Cinémathèque des pays de Savoie et de l'Ain depuis 2007[47], le logiciel de gestion audiovisuel DIAZ[48]. Les fiches documentaires des archives audiovisuelles du service d'archives sont réalisées par la cinémathèque, permettant ainsi une autre approche des fonds. C'est aussi une manière de déléguer la saisie des métadonnées à des spécialistes de l'audiovisuel, les archivistes professionnels n'étant pas formés à cette catégorie d'archives, ni nécessairement disponibles pour effectuer un travail de fourmi, la gestion d'entités individuelles (des titres de films). L'association Diazinteregio fédère depuis 2010 les centres d'archives et cinémathèques qui utilisent cette base de données comme instrument de gestion de leurs fonds audiovisuels et cinématographiques. Leur intérêt se porte autant sur le cinéma d'archive que sur l'usage d'un outil adapté et partagé, garantissant une interopérabilité dans la description des fonds d'archives audiovisuelles, précisément ce qui fait défaut actuellement dans les pratiques des services d'archives publiques[49].

41 Pour les archivistes, la gestion des archives audiovisuelles ne cesse d'évoluer. Impossible de concevoir les archives nativement numériques, de définir leurs métadonnées garantes de leur communication et de leur pérennisation sans comprendre la logique des supports physiques qui les précèdent et leur contexte spécifique de production. De nouveaux usages bousculent les cadres normatifs, remettent en cause la conception même d'archive audiovisuelle. Les Archives nationales ont été récemment sollicitées par plusieurs associations, *Lieux fictifs*[50], *Passeurs d'images*[51], *Les films de l'Arpenteur*[52], impliquées dans des projets pédagogiques et/ou artistiques auprès de différents publics (jeunes, personnes sous main de justice). Le régime juridique actuel favorise la libre réutilisation des « informations publiques », mais seulement celles sur lesquelles des tiers ne détiennent pas des droits de propriété intellectuelle[53]. Or les archives audiovisuelles sollicitées sont pour la plupart couvertes par des droits d'auteur ; leur mise à disposition nécessite l'établissement d'une convention spécifique. Ces « œuvres transformatives », créées à partir d'œuvres préexistantes, induisent leur adaptation ou leur transformation… en une nouvelle œuvre[54]. Une des pistes d'avenir pour ces archives ?

NOTES

1. « Les archives audiovisuelles », dossier coordonné par Audrey Clergeau, Aurélien Durr et Olivier Meunier, *Archivistes !*, n° 139, janvier-mars 2022, p. 21-33.

2. La BnF est également engagée dans des missions de collecte d'archives privées, que de récentes expositions ont mises en exergue : *Claudine Nougaret, dégager l'écoute*, exposition à la Bibliothèque nationale de France, François-Mitterrand, Galerie des donateurs, 14 janvier-15 mars 2020 : https://www.bnf.fr/fr/agenda/claudine-nougaret-degager-lecoute, consulté le 24 juin 2022. *Yitzhak Rabin/Amos Gitai*, exposition à la Bibliothèque nationale de France, François-Mitterrand, Allée Julien Cain, 19 mai-17 novembre 2021 : https://www.bnf.fr/fr/agenda/yitzhak-rabin-amos-gitai, consulté le 24 juin 2022.

3. Les contrats que signent les témoins sont archivés avec les enregistrements. Leur accord est en effet indispensable pour pouvoir communiquer leurs témoignages qui sont par ailleurs couverts par un mille-feuille juridique : code du Patrimoine, code de la Propriété intellectuelle, Code pénal.

4. *Industrie de la machine agricole en France*, 1905, noir et blanc, 60', 35 mm, Cinémathèque du ministère de l'Agriculture, Arch. nat., 19970030/282-287.

5. https://archives.dordogne.fr/r/38/archives-orales/, consulté le 15 juin 2022.

6. http://archives-av.tarn.fr/Fonds.php, consulté le 15 juin 2022.

7. https://archives.seinesaintdenis.fr/Etat-des-fonds-off/p101/Archives-audiovisuelles, consulté le 15 juin 2022.

8. https://archives.valdemarne.fr/r/287/les-archives-font-leur-cinema/, consulté le 15 juin 2022.

9. Florence DESCAMPS, *Archiver la mémoire. De l'histoire orale au patrimoine immatériel,* Paris, Éditions de l'École des hautes études en sciences sociales, 2019, Chapitre 2, « Les archives orales ou le tournant patrimonial », p. 63-90.

10. Comité d'histoire de la Sécurité sociale [CHSS], première, deuxième et troisième campagne d'enquêtes orales auprès de membres du personnel ou d'acteurs du domaine de la Sécurité sociale, 1975-2010, Arch. nat., 19980606/1-19980606/215, 20070674/1-20070674/7, 20160461/1-20160461/149, 20200094/1.

11. Pierre CAROUGE, « Quelle place pour les archives audiovisuelles en Archives départementales ? », *Archivistes !*, n° 139, janvier-mars 2022, p. 25.

12. Arch. nat., 20160438/43.

13. *Archives audiovisuelles du Théâtre national de Chaillot*, Arch. nat., 20160438 : https://www.siv.archives-nationales.culture.gouv.fr/siv/IR/FRAN_IR_055964, consulté le 24 juin 2022. *Archives sonores du Théâtre national de Chaillot*, Arch. nat., 20170371 : https://www.siv.archives-nationales.culture.gouv.fr/siv/IR/FRAN_IR_056791, consulté le 24 juin 2022.

14. Avant le versement des fonds aux Archives nationales, l'INA a numérisé les supports des captations des pièces de Théâtre national de Chaillot dans une logique de constitution de collections. Les références des films numérisés sont cataloguées dans l'INAthèque : http://inatheque.ina.fr, consulté le 24 juin 2022.

15. Ainsi les Archives nationales conservent l'archive audiovisuelle de la pièce de Peter Handke *Par les villages* mis en scène par Claude Régy en 1983 (Arch. nat., 20160438/31) et la BnF un enregistrement sonore dans le fonds Chaillot (Marie-Madeleine Mervant-Roux, Complémentarité des modes d'archivage modernes du spectacle théâtral : archives sonores / archives

papier / archives audiovisuelles, *Chaillot, lieu de tous les Arts*, publication des Archives nationales, 2020, https://books.openedition.org/pan/2359, consulté le 24 juin 2022).

16. Association Aristote : https://www.association-aristote.fr/cellule-format/, consulté le 8 septembre 2022.

17. La stratégie de pérennisation des Archives nationales : https://www.archives-nationales.culture.gouv.fr/strategie-de-perennisation-des-archives-nationales, consulté le 8 septembre 2022.

18. L'objectif de la cellule format de l'association Aristote est de mutualiser les savoir-faire et expertises et produire des recommandations : https://www.association-aristote.fr/cellule-format, consulté le 21 septembre 2022.

19. Sandrine GILL et Émeline LEVASSEUR, « Enjeux de préservation des archives audiovisuelles aux Archives nationales (France) », *Colloque international sur les archives audiovisuelles à l'ère numérique : préservation, accessibilité et gouvernance*, 27-28 octobre 2022 à Tunis, coorganisé par la Bibliothèque nationale et les Archives nationales de Tunisie, l'École nationale des chartes de Paris-Université Paris Sciences et Lettres : https://aavn2022.sciencesconf.org/, consulté le 21 septembre 2022.

20. Sous la direction de Martine SIN BLIMA-BARRU, *Panorama sur 35 ans de collecte d'archives audiovisuelles aux Archives nationales*. https://www.siv.archives-nationales.culture.gouv.fr/siv/IR/FRAN_IR_052868.

21. *Amélioration de l'habitat rural*, 1959, réalisé par Pierre Franceschi, produit par *Je vois tout*/ Service cinématographique du ministère de l'Agriculture, noir et blanc, 14', 16 mm, Cinémathèque du ministère de l'Agriculture, Arch. nat., 19960237/23.

22. *L'or de la Durance,* 1960, réalisé par Robert Enrico, produit par *Je vois tout*, musique de François de Roubaix, couleur, 26', 16 mm, Cinémathèque du ministère de l'Agriculture, Arch. nat., 19960237/315.

23. *L'aménagement de l'habitat des personnes âgées*, une émission de Fred Orain de la série *Je voudrais savoir* (1969-1974) diffusée sur TF1, réalisée par Guy Loriquet, couleur, 08'27'', 16 mm, Caisse nationale d'assurance maladie des travailleurs salariés, Arch nat., 19990035/1051, 1081 ; 19990081/17 ; 20130089/40.

24. *Allocutions du Président Étienne Eyadema et du Président Georges Pompidou lors de son arrivée à Lomé (Togo),* palais de la Présidence, 22 novembre 1972, réalisateurs anonymes, bande magnétique audio 13 cm, 13'54'', Présidence de la République, Arch. nat., 2AV/163.

25. Reportage photographique du *Voyage officiel du Président Pompidou au Togo (22-24 novembre 1972)*, Arch. nat., AG/5(2)/982/N, Reportage n° 2597. Le déplacement du Président en Haute-Volta puis au Togo est l'occasion pour le service photographique de l'Élysée d'inaugurer la diapositive couleur pour documenter l'activité du chef de l'État.

26. *Ménagères, ménagez-vous*, émission de la série *Objectif santé* diffusée sur TF1 de 1976 à 1984, coordination Roger Schodet, vers 1983, producteur Caisse nationale d'assurance maladie des travailleurs sociaux, Arch. nat., 20130089/14.

27. L'exposition *Filmer les procès, un enjeu social* (commissariat Martine Sin Blima-Barru, Archives nationales, et Christian Delage, Université Paris 8) fut présentée aux Archives nationales du 15 octobre 2020 au 18 décembre 2021 : https://www.archives-nationales.culture.gouv.fr/filmer-les-proces-un-enjeu-social, consulté le 24 juin 2022. Adaptation de l'exposition *Filmer les procès, un enjeu social* : Institut français à Berlin, 3 mai au 25 juin 2021 ; Archives départementales des Yvelines, 22 novembre 2021 au 22 avril 2022 ; Centre Iriba pour le multimedia (Kigali, Rwanda), 27 janvier au 14 juillet 2022 ; Institut français de Kigali, 27 janvier au 18 février 2022 ; Institut français de Brême, 3 mars au 20 avril 2022 ; Université catholique de Lille, 15 mars au 15 mai 2022 ; Archives départementales de la Gironde, 30 mars au 4 novembre 2022 ; Archives du Département du Rhône et de la Métropole de Lyon, 15 septembre 2022 au 24 février 2023 ; Université de Limoges et Archives départementales de Haute-Vienne, 15 septembre 2022 au 18 novembre 2022 ; Institut français d'Allemagne et Université de Ratisbonne, 14 septembre au

28 octobre 2022 ; Institut français de Dresde, novembre 2022 ; Archives départementales de la Corrèze et Association Mémoire en Chemin, 15 mai au 15 juin 2023. Création d'une nouvelle exposition *Juger et filmer les procès pour crimes contre l'humanité, France, Argentine, Chili :* Espace Mémoire et Droits Humains (ESMA, Buenos Aires, Argentine), Archivo de la Memoria (Buenos Aires, Argentine), et Musée de la Mémoire et des Droits Humains (Santiago, Chili), 11 septembre au 10 décembre 2023.

28. ISAD(G), norme générale et internationale de description archivistique : https://www.ica.org/fr/isadg-norme-generale-et-internationale-de-description-archivistique-deuxieme-edition, consulté le 27 mai 2022.

29. SEDA, standard d'échange de données pour l'archivage. https://francearchives.fr/fr/article/88482501, consulté le 27 mai 2022.

30. Introduction de la seconde édition de l'ISAD(G), Ottawa, 2000 : https://www.ica.org/sites/default/files/CBPS_2000_Guidelines_ISAD%28G%29_Second-edition_FR.pdf, consulté le 27 mai 2022.

31. Les Archives nationales préconisent l'utilisation de *ReSIP*, outil conçu par le programme VITAM : http://www.programmevitam.fr/pages/ressources/resip/, consulté le 24 juin 2022.

32. Le projet ADAMANT [Administration Des Archives et de leurs Métadonnées aux Archives nationales dans le Temps] avait pour objectif d'adapter les outils, les procédures et l'organisation de la chaîne archivistique des Archives nationales, pour répondre aux enjeux de conservation de l'information numérique sur le long terme : https://www.archives-nationales.culture.gouv.fr/web/guest/archiver-les-donnees-numeriques-adamant, consulté le 21 septembre 2022. Il s'appuie sur le programme Vitam : https://www.programmevitam.fr/pages/presentation/, consulté le 18 mai 2022.

33. L'*Open Archival Information System* [OAIS], que l'on pourrait traduite comme « système ouvert d'archivage d'information », est un modèle conceptuel destiné à la gestion, à l'archivage et à la préservation à long terme de documents numériques. Commandité par l'ISO [Organisation Internationale de Normalisation], il a été élaboré par le CCSDS [*Consultative Committee for Space Data Systems*/Organisme International de normalisation des agences spatiales]. Une version française est disponible : https://public.ccsds.org/Pubs/650x0m2%28F%29.pdf, consulté le 21 septembre 2022.

34. *Archives audiovisuelles du Théâtre National de Chaillot*, répertoire numérique détaillé du versement 20160438 par Justine Dilien, sous la direction de la Mission des archives du ministère de la Culture et de la Communication, et Sandrine Gill, Archives nationales, département de l'Archivage électronique et des archives audiovisuelles, Archives nationales (France), 2018 : https://www.siv.archives-nationales.culture.gouv.fr/siv/IR/FRAN_IR_055964, consulté le 27 mai 2022.

35. Audrey CLERGEAU et Aurélien DURR, « De l'art de la description des archives audiovisuelles », *Archivistes !,* n° 139, janvier-mars 2022, p. 29.

36. *Rencontres professionnelles du groupe de travail archives audiovisuelles des territoires*, Clermont-Ferrand, 19-21 octobre 2022 : https://www.archivistes.org/Rencontres-professionnelles-du-Groupe-de-travail-Archives-audiovisuelles-des, consulté le 27 mai 2022.

37. Arch. nat., 20170146/8 : 2 bobines de films 16 mm, 2 cassettes U-matic, 3 VHS.

38. Arch. nat., 20000458/30-20000458/31 : 2 bobines de films 16 mm.

39. Parmi les institutions qui diffusent des archives audiovisuelles, on signalera les Archives départementales du Val-de-Marne, dont les fonds audiovisuels sont en partie constitués de dons d'associations et de particuliers : https://archives.valdemarne.fr/r/287/les-archives-font-leur-cinema/, consulté le 15 juin 2022.

40. Le logiciel Annotate-Chrono développé au laboratoire Dicen-IdF pour faciliter la rédaction des analyses chrono-thématiques est notamment utilisé par les Archives nationales pour produire en direct les métadonnées du filmage, à des fins d'archivage historique, du procès des

attentats du 13 novembre 2015 au Palais de justice de Paris : https://fplab.parisnanterre.fr/ateliers/claireScopsi_30092021.html, consulté le 15 juin 2022.

41. INAthèque, Catalogue des programmes collectés et archivés par l'INA : http://inatheque.ina.fr/index/TV-RADIO/, consulté le 16 juin 2022.

42. Catalogue général de la BnF : https://catalogue.bnf.fr/index.do. Gallica : https://gallica.bnf.fr/html/und/videos/documentaires-en-ligne?mode=desktop, consulté le 16 juin 2022.

43. Images de la culture, catalogue du CNC : https://imagesdelaculture.cnc.fr/, consulté le 16 juin 2022.

44. Images défense : https://imagesdefense.gouv.fr/, consulté le 16 juin 2022.

45. *Photothèque du ministère de la Reconstruction et de l'Urbanisme*, Arch. nat., F/14/18261-F/14/18284 : https://www.siv.archives-nationales.culture.gouv.fr/siv/IR/FRAN_IR_050103, consulté le 16 juin 2022 ; *Cinémathèque du ministère de la Reconstruction et de l'Urbanisme, de la Construction et de l'Équipement (1938-2004),* Arch. nat., 20170146/1-20170146/988 : https://www.siv.archives-nationales.culture.gouv.fr/siv/IR/FRAN_IR_058381, consulté le 16 juin 2022. Voir également Sandrine GILL (2023), « La Reconstruction filmée : focus sur le fonds de la cinémathèque du ministère de la Reconstruction et de l'Urbanisme aux Archives nationales », *Actes de la journée d'étude Les fonds iconographiques et audiovisuels de la Reconstruction de 1940 à 1960*, sous la dir. de Christel PALANT et de Boris LABIDURIE, publications des Archives nationales, OpenEdition Books, https://books.openedition.org/pan/4849?nomobile=1.

46. Médiathèque TERRA, un service numérique du ministère de la Transition écologique et du ministère de la Cohésion des territoires et des Relations avec les collectivités territoriales qui gère, traite, conserve et valorise les fonds numériques (photographies, vidéos, documents) produits par les ministères. Ce service est destiné à tous les publics en interne et en externe, à l'exception des organismes utilisant les données pour des produits payants : https://terra.developpement-durable.gouv.fr/, consulté le 15 juin 2022.

47. *Vos images, notre mémoire*, flyer de collecte de fictions, documentaires, films de commande, films amateurs, en partenariat entre Archives départementales de l'Ain et la Cinémathèque des pays de Savoie et de l'Ain : https://www.archives.ain.fr/data/archives_fly_memoire2020.pdf, consulté le 16 juin 2022.

48. La base de données DIAZ a été développée en 2003-2004 par la Cinémathèque de Bretagne et la société Virtualys, dans l'objectif d'une base spécialement adaptée aux besoins des centres de gestion d'archives audiovisuelles inédites.

49. *Diazinteregio* (https://www.facebook.com/Diazinteregio/, consulté le 16 juin 2022).

50. *Lieux Fictifs* est un espace collaboratif de création et d'éducation sur l'image, qui développe des pratiques artistiques dont le sujet et le champ d'intervention sont la « frontière » : https://www.lieuxfictifs.org/, consulté le 16 juin 2022.

51. L'association *Passeurs d'images* a pour objet de fédérer et d'animer le réseau des acteurs de l'éducation aux images : https://www.passeursdimages.fr/, consulté le 16 juin 2022.

52. *Les films de l'Arpenteur* est un collectif de réalisateurs qui se réunit autour de projets de documentaires de création : http://www.lesfilmsdelarpenteur.fr/, consulté le 16 juin 2022.

53. « La loi Valter, qui porte transposition de la directive européenne du 26 juin 2013 relative à la réutilisation des informations du secteur public, et la loi pour une République numérique ont pour objectif de favoriser la réutilisation des informations publiques. Elles élargissent le champ d'application du droit de la réutilisation. Désormais, les établissements et services culturels (donc les services d'archives) et les établissements d'enseignement et de recherche relèvent du droit commun de la réutilisation (alors qu'ils appartenaient auparavant au périmètre dérogatoire défini à l'ancien article 11 de la loi CADA). Les informations produites dans le cadre d'une mission de service public à caractère industriel et commercial rejoignent également le champ d'application du droit de la réutilisation alors qu'elles en étaient précédemment exclues. » *Le*

nouveau régime juridique de la réutilisation des informations publiques : https://siafdroit.hypotheses.org/659, consulté le 16 juin 2022.

54. Il n'existe à l'heure actuelle aucun régime spécifique à cette catégorie d'œuvre à la suite du rapport confié par le ministère de la Culture en 2014 au Conseil supérieur de la propriété littéraire et artistique [CSPLA], ce qui ne ferme pas la porte aux initiatives. Valérie-Laure BENABOU et Fabrice LANGROGNET, *Rapport de la mission du CSPLA. Les « œuvres transformatives »,* Conseil supérieur de la propriété littéraire et artistique, 2014 : https://www.culture.gouv.fr/Thematiques/Propriete-litteraire-et-artistique/Conseil-superieur-de-la-propriete-litteraire-et-artistique/Travaux/Missions/Mission-du-CSPLA-relative-aux-creations-transformatives, consulté le 16 juin 2022. Mireille Maurice, déléguée régionale de l'Ina Méditerranée, et Sandrine Lardeux, juriste à l'INA, font un retour d'expérience éclairant sur la manière dont des films d'archives ont été mis à disposition dans le projet européen *In living Memory*. Mireille MAURICE et Sandrine LARDEUX, « Regardez, c'est moi qui vous parle… Les archives de l'INA dans le projet *In Living Memory* », *Living Memory, Mémoires vivantes cinéma-arts visuels-spectacle vivant*, Erasmus +, Union européenne, 2014, p. 30-35. Livre téléchargeable sur le site :http://inlivingmemory.eu/editions/, consulté le 16 juin 2022.

AUTEUR

SANDRINE GILL

Docteure en histoire de l'art, référente métier dans le domaine des archives audiovisuelles au Département de l'administration des données (DAD), Archives nationales

Le « lac de données », une infrastructure technique pour déployer la gouvernance des données à l'Ina

Gautier Poupeau

Nous remercions Éléonore Alquier, responsable du département Data de l'Ina pour sa collaboration à l'élaboration de ce texte.

1 Depuis huit ans, l'Institut national de l'audiovisuel [Ina] travaille à mettre en place une nouvelle architecture pour son système d'information, autour d'une infrastructure, le « lac de données », qui organise l'ensemble des processus autour de la donnée : traitement, stockage, exploitation et valorisation. En permettant l'émergence d'une véritable gouvernance de données, cette évolution conduit à une nouvelle répartition des rôles et fait émerger de nouveaux profils professionnels.

2 Cet article revient sur les différentes étapes qui ont conduit à définir cette architecture : les problèmes que nous avions à résoudre, les options envisagées et les raisons de nos choix. Il aborde dans un second temps les impacts de cette évolution sur l'organisation.

La donnée : nouvelles perspectives pour le système d'information

3 L'idée du lac de données à l'Ina est née de deux convictions. D'une part, sur le plan de la modélisation des données, il est nécessaire de s'affranchir des « carcans de la pensée documentaire » touchant à la description de la collection[1], de quitter le paradigme de la notice de catalogue pour aller vers celui de la donnée. Cette première exigence implique de dépasser les principes traditionnels des systèmes d'information documentaires, fondés sur l'association d'un document à une notice, pour appréhender de manière cohérente les différents niveaux de granularité qui vont de la collection au

contenu du document (les données) en passant par le document et sa description (les métadonnées). D'autre part, sur le plan technique, il n'existe actuellement aucun système de gestion de données qui réponde de manière parfaite à tous les besoins. L'enjeu principal réside dans le fait de séparer les données des usages, c'est-à-dire penser les données indépendamment de l'utilisation qu'on en fait, en les plaçant au cœur du système.

4 Cette double posture conceptuelle a guidé la réflexion des équipes de l'Ina dans les objectifs de la refonte de son système d'information : pour répondre aux enjeux actuels et futurs des systèmes d'information [SI] documentaires[2], il faut adopter une approche qui implique de déconstruire la notice et de s'appuyer sur la diversité des technologies aujourd'hui à notre disposition.

La donnée au cœur de la nécessaire déconstruction de la notice

5 La notice, au sens qu'on lui donne traditionnellement dans les métiers de l'information, est un document qui regroupe un certain nombre d'informations descriptives, ou métadonnées, concernant un objet d'une collection[3]. Si l'on examine l'évolution de la description documentaire, on constate que la notion de notice a préexisté à la notion de métadonnées. Cette dernière a émergé principalement en raison de l'augmentation du nombre d'informations nécessaires pour décrire un document, d'abord avec la numérisation et l'ajout des métadonnées techniques, administratives, juridiques et de structure, puis avec le besoin croissant de collecter, gérer et exploiter d'autres types de données périphériques : traces d'usage, données produites par les usagers dans le cadre du collaboratif (*User Generated Content*), données extraites du document ou générées automatiquement à partir de son analyse.

6 L'apparition des métadonnées présentait un autre avantage : on pourrait parler de dimension « marketing » de la notion de métadonnées, le choix d'un nouveau terme venant redonner de l'intérêt à un sujet considéré comme dépassé. Pourtant, les premiers formats et modèles qui ont émergé pour encoder les métadonnées, qu'il s'agisse du XML, avec sa structure hiérarchique encadrée par des balises, du Dublin Core[4] simple, avec ses quinze éléments facultatifs et répétables, ou encore de METS, un format *container* qui regroupe toutes les métadonnées et cartographie la structure du document, présentaient un point commun : comme la notice de catalogue avant eux, ils enfermaient la description documentaire à l'intérieur de frontières artificielles, créant de fait un nouveau document, un méta-document dont la fonction était de décrire le document-objet se trouvant à l'intérieur des collections.

7 Ces quinze dernières années, la réflexion autour du modèle de graphe, survenue notamment à travers le concept de web sémantique, a fait émerger un nouveau modèle[5]. L'idée est de structurer l'information autour de la description d'entités du monde réel (objets, personnes, concepts, lieux, événements), de formuler des assertions concernant ces entités au moyen de triplets, puis de relier ces entités entre elles sous la forme d'un graphe. La notion de graphe a conduit à l'idée de « détruire » le méta-document que constituait la notice ou encore de faire « éclater » la notice[6]. Les données ne sont plus enfermées à l'intérieur d'une notice, mais liées les unes aux autres dans un espace global d'information. Dès lors, il devient possible de dépasser la vision documentaire des informations de signalement qui avait émergé avec l'idée de notice, qui limitait notre réflexion et maintenait artificiellement des frontières, en particulier

entre le document et sa description, lorsqu'il s'agissait de lier les informations entre elles.

La donnée pour assurer le lien entre le document, sa description et son contenu

8 La réflexion autour du web sémantique et du modèle du graphe nous a rappelé que la métadonnée est une donnée comme une autre et qu'il n'existe pas, pour le document numérique, de frontière imperméable entre le document dans sa globalité, sa description et son contenu. L'OCR[7] ou la transcription, puis la fouille de texte et de données rendent possible l'analyse, non plus du document, mais du contenu lui-même : on peut désormais repérer une chanson diffusée à l'intérieur d'une émission télévisée, cartographier les lieux cités dans un documentaire, identifier le visage d'une personnalité ou l'apparition d'une marque à l'écran. Le contenu du document se retrouve à son tour partie prenante du signalement : la granularité de la description devient encore plus fine. Les métadonnées s'éloignent du rôle de *métatexte*[8], qui pouvait être joué par la notice, pour être plus intrinsèquement liées aux entités repérables à l'intérieur du document lui-même. À partir de ce moment, la place de la donnée de signalement traditionnelle est inévitablement remise en question.

9 Dans cette vision des choses, la donnée de signalement fait partie du contenu numérique de l'objet et est gérée conjointement avec lui, devenant ainsi en tant que telle un objet d'études potentiel. Par exemple, depuis 2005, l'Ina publie le baromètre Ina'STAT[9], qui fournit une vision de l'actualité établie à travers la documentation des sujets des journaux télévisés [JT]. Cette documentation des JT est devenue une source, de la même manière que le catalogue de la BnF est devenu une source d'analyses massives sur les fonds. Ces exemples marquent la fin de la dichotomie entre notice et document : les données, qu'elles soient issues du catalogage ou extraites du document lui-même, constituent une représentation du contenu et permettent de l'appréhender, de l'exploiter, d'y accéder et de le valoriser.

La donnée au cœur du système d'information

10 Ce nouveau paradigme du signalement, qui implique de traiter de façon cohérente toutes les informations dont on dispose au sujet de la collection et des objets qui la composent, nécessite de placer la donnée au centre du système d'information. Nous souhaitions donc centraliser toutes les données, quels que soient leur origine et leur usage (documentaire, commercial, juridique) concernant les documents qui forment les collections de l'Ina, quel que soit leur type (sons, images télévisées, archives écrites, livres et revues, photographies et même archives web), quelles que soient la nature et la structure des données qu'elles contenaient.

11 Cette approche représentait pour l'Ina un véritable renversement du système d'information traditionnel. Auparavant, comme dans la plupart des organisations, le SI de l'Ina était organisé de manière verticale : à chaque besoin métier correspondait une application qui gérait ses propres données, ce qui avait pour conséquence la création de silos de données sans lien entre elles. Depuis une vingtaine d'années, l'architecture orientée service a permis l'émergence de solutions qui ajoutent une couche pour faire le lien entre les services applicatifs de différents silos, mais l'équivalent n'existe pas

pour faire des liens au niveau de la couche de données. Notre idée était donc de briser ces silos pour permettre d'accéder de manière homogène à toutes les données disponibles. Il nous fallait gérer les données dans une infrastructure technologique unique, chargée de centraliser le traitement des données : c'est le rôle du lac de données.

12 Le lac de données de l'Ina se définit donc comme une infrastructure unique de traitement et de stockage de la donnée structurée ou semi-structurée. Le terme de « lac » s'est imposé progressivement : au début du projet, on privilégiait les appellations de *Business Intelligence*[10], « puits de données » ou « *datawarehouse* » (entrepôt de données). Cependant, l'image du puits présentait des connotations négatives (un puits sans fond d'où l'information ne remonte jamais), aussi lui a-t-on ensuite préféré celle, plus fédératrice, du lac. Nous imaginions un endroit où des bateaux pourraient naviguer et dont on maîtriserait le périmètre, au contraire de petits étangs isolés les uns à côté des autres.

13 Pour autant, ce terme de « lac » peut s'avérer trompeur : en effet, les « *Data Lake* » tels qu'identifiés par le marché ne correspondent pas techniquement au modèle adopté par l'Ina[11]. Ces outils gèrent des fichiers à plat, avec une perte relative de la structure de données ; au contraire, le lac de données de l'Ina préserve cette structure, notamment à travers un modèle conceptuel commun auquel vont se relier toutes les données concernant la collection et son contenu.

Mise en œuvre du lac de données

14 L'histoire de l'informatique n'est qu'une suite de centralisations et de décentralisations ; actuellement, nous sommes dans une phase de recentralisation des systèmes d'information constitués en silos. Au-delà de la centralisation des données, nous étions également convaincus, à l'heure de mettre en place le lac de données, qu'il fallait centraliser techniquement les processus de traitement et de stockage des données. La centralisation est aussi une réponse à un problème de moyens : maintenir un système et maîtriser les données est plus difficile et plus coûteux lorsque les silos se multiplient les uns à côté des autres, car une telle architecture dilue les moyens. L'enjeu était donc de définir une architecture adaptée pour centraliser les données tout en s'affranchissant du paradigme de la notice.

La solution universelle pour stocker et exploiter les données n'existe pas

15 Historiquement, la base de données relationnelle constitue le cœur des systèmes d'information, dès lors organisés autour du modèle relationnel et de la notion de table. Pratiqués depuis quarante ans, ces logiciels robustes gèrent de façon optimale les contraintes d'intégrité des données. Cependant, les limites de ce modèle sont apparues au moment où les moteurs de recherche sont montés en puissance, faisant émerger des applications logicielles dans lesquelles le moteur de recherche est au cœur de l'infrastructure (*search based applications*[12]) et dont le principe consiste à interroger en langage naturel des données structurées ou semi-structurées stockées sous la forme de documents. On peut alors parler de « base de données document[13] ». Ces nouveaux outils rendaient possible la scalabilité horizontale[14], qui n'était pas supportée par les

bases relationnelles, et présentaient donc des avantages en termes de coûts, de performance et de rapidité. Ils permettaient en outre de rechercher en plein texte à l'intérieur de l'ensemble des informations disponibles, métadonnées et contenu, limitant le besoin de connaître finement leur structuration pour formuler des requêtes. En revanche, les moteurs de recherche posent un problème en termes de gestion de données : pour rendre ce service, les mêmes informations doivent être répétées dans chaque document, ce qui oblige à des opérations complexes de réplication des données et peut conduire à les désynchroniser. Le modèle de table de la base de données relationnelle et le modèle orienté document du moteur de recherche présentent donc chacun des avantages et des inconvénients, sans qu'aucun des deux ne puisse résoudre tous les problèmes.

16 Le modèle de graphe survenu par la suite propose une modélisation qui offre une grande souplesse dans la manipulation des données, grâce à une granularité très fine, mais peut aussi rendre très complexe la représentation de certains liens à l'intérieur des documents. En outre, cette notion de graphe s'est avérée difficile à appréhender pour les développeurs, ce qui peut expliquer une adoption relativement limitée dans l'industrie de ces technologies. Les *triplestores*, bases de données de graphes conformes aux technologies du web sémantique, présentent, en outre, de sérieux problèmes de performance et de modélisation. Même si d'autres modèles de graphes, indépendants du web sémantique, ont émergé par la suite avec des produits comme Neo4J, les problématiques de maintenabilité de ce type de technologie ne sont toujours pas pleinement résolues aujourd'hui et nous interdisent de nous reposer entièrement sur elles.

17 Cette rapide comparaison des trois modèles actuels de stockage et d'interrogation des données montre qu'aucun ne présente toutes les caractéristiques qui lui permettraient de supplanter les autres. Différentes contraintes, en fonction du cycle de vie de la donnée et des besoins rencontrés pour son usage, impliquent de faire appel à la base de données relationnelle, à la base de données document, au moteur de recherche ou à la base de données graphe suivant les cas. Ce constat a conduit à la nécessité, pour le lac de données de l'Ina, de disposer de ces différents types d'applications pour gérer des données organisées en tables, en graphes ou en documents et de faire passer les informations de l'une à l'autre en fonction des besoins et des cas d'usage.

Un modèle conceptuel de données au cœur du système d'information

18 Si le web sémantique ne nous fournissait pas les technologies capables d'atteindre notre objectif, son apport conceptuel nous montrait la voie pour réaliser des liens entre les données isolées dans différents silos, construire un graphe de l'ensemble des données et parvenir à les maîtriser[15]. Suivant ce principe, l'interopérabilité des données est assurée au travers des identifiants communs qui se trouvent dans les différents types de bases de données et *via* l'adoption d'un modèle conceptuel [16] commun, suffisamment souple pour couvrir tous les cas de figure.

19 Le modèle conceptuel mis en place pour le lac de données est organisé autour de trois entités principales[17] : le contenu (l'essence de l'objet décrit), le support physique ou numérique, et l'événement (qui décrit l'histoire de l'objet depuis sa création en passant par son utilisation et sa conservation). Les éléments textuels (les attributs tels que les

noms et les titres, ou les textes qui composent le contenu lui-même) sont eux-mêmes traités comme des entités, ce qui donne au modèle une souplesse et une extensibilité optimales. Enfin, chaque donnée est caractérisée par des informations de provenance. Très générique, ce modèle permet de décrire n'importe quel type d'objet et de lui associer toute sorte d'information. Les données représentées suivant ce modèle fondé sur leur logique intrinsèque, indépendamment de leur usage, constituent le cœur du système d'information : on peut dès lors parler de données primaires. Ces données primaires pourront ensuite être dérivées en données secondaires au sein du lac pour fournir des vues spécifiques sur les données, adaptées à tel ou tel type de besoin.

20 Grâce à ce modèle, nous voulions centraliser la couche « data », c'est-à-dire toutes les bases de données de notre système, quel que soit leur type et quelles que soient la nature et la structure des données qu'elles contenaient. Pour cela, une infrastructure de stockage contenant à la fois une base de données relationnelle, une base de données graphe, une base de données document et un moteur de recherche a été mise en place. Au sein de cette couche de données, chaque type de base de données remplit sa fonction propre. La base de données relationnelle gère la donnée primaire s'agissant des données structurées (typiquement, les métadonnées) : le modèle de table est aujourd'hui celui qui assure le mieux les contraintes d'intégrité et la gestion des transactions. Par ailleurs, il s'agit d'une technologie bien connue et qui a fait ses preuves. Les données semi-structurées générées automatiquement (par exemple le résultat d'une transcription, d'une OCR ou d'un processus d'analyse automatique des visages) sont, elles, stockées dans une base de données document, plus adaptée à leur formalisme. Ce sont également des données primaires. Enfin, le graphe de connaissance et le moteur de recherche sont quant à eux alimentés par des données secondaires calculées à partir de ces données primaires.

21 La mise en place du nouveau modèle de données a nécessité une reprise totale des données, transférées depuis toutes les bases existantes au sein de l'Ina vers le modèle unique du lac. L'enjeu de cette opération de longue haleine, conduite en parallèle au déploiement de l'infrastructure technique, était de mettre fin au silotage des données, d'éviter les redondances d'informations et de remettre les données en cohérence.

Une architecture en couches

22 Les avantages de cette organisation des données, qui fait cohabiter plusieurs types de bases, sont clairs : il devient possible de manipuler la donnée de différentes manières suivant les besoins, en combinant les avantages respectifs de chaque type de produit. Les données primaires sont stockées suivant un modèle conceptuel commun qui respecte leur logique intrinsèque, tandis que des données secondaires sont produites pour répondre aux différents cas d'usage qui peuvent se présenter. Cependant, ce stockage multiple implique des risques. Comment contrôler la cohérence des données lors des réplications entre bases, puisque c'est le principe de ce système ? Comment s'assurer que cette cohérence n'est pas remise en cause en cas de panne ou d'interruption de la transaction ?

23 Afin de répondre à cette problématique, nous avons donc ajouté, entre les bases de données et l'interface d'accès aux données, une couche de traitement dont la fonction est de transformer la structure physique des données, telle qu'elle se présente dans les bases, en une structure logique : le modèle de données qu'on expose *via* des *web services*.

La donnée se trouve ainsi séparée des usages. En architecture de systèmes d'information, le modèle appelé MVC [Modèle Vue Contrôleur] distingue les données (le modèle), les services (le contrôleur) et l'interface (la vue). Le lac de données est organisé suivant ce principe, cette couche intermédiaire de services assurant le lien entre le stockage des données et les interfaces d'accès, et opère les traitements nécessaires. Dans cette couche intermédiaire, nous n'avons pas fait de choix dans les logiques de traitements : ils s'exécutent de façon synchrone (sous forme de *web service*), asynchrone (en batch[18]) ou encore en flux. Le choix s'effectue en fonction des besoins. Ensuite, les applications du système d'information de l'Ina qui ont besoin d'interagir avec des données s'appuient sur le lac de données à travers les *web services* offerts par la couche d'interface.

24 Le principal apprentissage que nous avons retiré de notre démarche est que la couche intermédiaire, le module de traitement, est le cœur de notre système : c'est lui qui garantit l'interopérabilité du système en faisant en sorte que le modèle de données utilisé par les applications (notre modèle conceptuel) reste indépendant du stockage physique des données. De ce fait, les maîtrises d'œuvre des applications métiers du SI ne voient jamais la structure physique des données, leur structure réelle, mais accèdent à une structure normalisée produite par cette couche intermédiaire, dont nous sommes en mesure de garantir la permanence : si nous devons modifier la structure des bases, nous faisons évoluer les traitements intermédiaires, pour continuer à fournir aux usagers la même structure logique. De même, il est possible d'introduire un nouveau type de stockage de données sans perturber les services fournis par le lac de données ni sa logique d'ensemble.

L'articulation avec les applicatifs métiers

25 Une fois le *framework* disponible, c'est-à-dire l'infrastructure déployée et le nouveau modèle de données disponible sous forme de *web services*, une diversité d'usages devient possible. Ceux-ci peuvent porter sur la consommation des données *via* des interfaces de consultation, sur l'exploitation des données disponibles pour construire de nouvelles visualisations, ou encore sur la production (manuelle ou automatique) de nouvelles données qui vont venir rejoindre le lac. Ces fonctions sont assurées par des applications métiers qui vont se brancher sur le lac de données en s'appuyant sur l'implémentation de nos flux et de nos *web services*. Les maîtrises d'œuvre de ces applications métiers n'ont pas la main sur le format des données : elles doivent décrire les données dont elles ont besoin, les interactions qu'elles souhaitent mettre en œuvre et les performances attendues. Côté lac de données, la base de données qui va répondre à la requête sera déterminée en fonction de ces besoins. Par exemple, la décision de stocker les données dans une base relationnelle puis de les répliquer dans un moteur de recherche est un traitement envisagé de façon générique, donc rapide à mettre en place. Le modèle de données le plus pertinent est ensuite exposé au demandeur, qui sera libre de modifier à la volée les données fournies et d'établir les profils d'accès des usagers, mais sans stocker les données modifiées dans une autre base : les données doivent rester en interaction avec le système du lac, sinon on retombe dans l'inconvénient de la multitude des « petits étangs ».

26 Cette contrainte crée une dépendance de ces applications métiers à l'égard du lac, puisque c'est lui qui reçoit et stocke les enrichissements produits de leur côté pour

conserver la cohérence de l'ensemble des données. Il est donc important de pouvoir compter sur la souplesse du modèle conceptuel du lac et la maniabilité offerte par l'infrastructure, avec ses différents types de bases et sa couche de traitement intermédiaire, pour limiter cette contrainte. Ainsi à l'heure actuelle, une nouvelle source de données peut être ajoutée en vingt-quatre heures et grâce à l'abstraction et aux possibilités de configuration offertes par le système, la plupart du temps sans recourir à du développement spécifique. De plus, tous les systèmes du lac sont automatisés avec une intégration continue. À partir de l'étape du codage, le déploiement s'effectue sur quatre environnements : développement, intégration, préproduction, production. Tout a été réfléchi pour que le déploiement soit automatisé et le plus simple possible, afin d'assurer la haute disponibilité et la scalabilité de ces quatre environnements.

27 Cette souplesse et cette évolutivité constituent le point fort du système : c'est grâce à elles qu'il a été possible de concevoir cette approche par la structuration et la modélisation de la donnée avant même d'attaquer la question des processus ou de commencer à travailler sur les logiciels métier. Même si le chantier est déjà bien engagé, ce dernier chemin reste encore largement à parcourir : traduire ces cas d'usage en outils utilisables par les utilisateurs finaux, s'appuyant sur une nouvelle organisation adaptée.

Les implications organisationnelles

28 Nous l'avons vu, la mise en place du lac de données constitue un bouleversement pour le système d'information de l'Ina. Au-delà de la mise en place d'un modèle conceptuel nouveau et de la centralisation de toutes les données dans une infrastructure technique commune, l'architecture mise en place impose à toutes les applications métiers de l'Ina d'articuler leur fonctionnement avec le lac de données. La mise en place d'une gouvernance et d'une organisation adéquates était donc nécessaire pour garantir le bon fonctionnement du SI.

Un département Data

29 La réorganisation de l'Ina, effective au 1er septembre 2021, est venue confirmer cette trajectoire en modifiant l'organisation du travail à deux niveaux.

30 Tout d'abord, au sein de l'ancienne direction des systèmes d'information [DSI], la mise en place d'une architecture centrée sur les données a eu pour conséquence de rendre obsolète la division classique de l'activité entre développement (*build*) et production (*run*). Il faut acter que les applications du SI ne sont pas des projets qui, une fois achevés, basculent dans un mode de maintenance minimal, mais des produits en constante évolution. Il devient donc nécessaire, outre l'intégration continue déjà évoquée, de s'organiser en lignes de produits pour gérer l'ensemble du cycle de vie des applications du SI. L'infrastructure du lac de données est elle-même un produit : une équipe pluridisciplinaire a été mise en place pour le gérer, réunissant des compétences allant de l'intégrateur technique, les devOps et les dataOps, à l'expert métier de la donnée, les analystes de la donnée.

31 De même, au niveau de l'entreprise dans son ensemble, maîtriser la gouvernance des données de manière globale implique de regrouper tous les spécialistes de la donnée,

qu'ils soient techniques ou métier, dans une même unité organisationnelle. Le nouveau département Data rassemble l'ingénierie de la donnée, c'est-à-dire des compétences en informatique ciblées sur la gestion de la donnée, et l'ingénierie documentaire, à savoir les compétences nécessaires pour modéliser, produire, gérer et interpréter les données dans un domaine spécifique, à l'Ina, l'audiovisuel.

32 Cette nouvelle organisation, qui a fait fusionner la DSI avec une partie de l'ancienne direction des Collections au sein d'une nouvelle direction dite « Data et technologie », est aujourd'hui en mesure de rendre un meilleur service à l'ensemble de l'organisation. Les besoins métiers de l'entreprise sont dès lors pris en main par d'autres directions, par exemple la direction des Patrimoines, qui regroupe les documentalistes chargés de traiter et valoriser les collections. Au sein de ces directions, les pratiques doivent évoluer pour prendre en compte l'articulation avec le lac de données.

L'introduction de l'intelligence artificielle

33 Disposer d'une infrastructure centralisée pour toutes les données de l'Ina nous ouvre également une nouvelle perspective : celle de pouvoir automatiser, en s'appuyant sur des technologies d'intelligence artificielle, la production d'un certain nombre de données. Certains systèmes sont déjà en production : transcription de la parole en texte [*Speech to text*] des émissions liées à l'actualité comme les JT, extraction d'entités nommées dans les transcriptions, analyse de la parole pour distinguer les voix d'homme et de femme, océrisation des bandeaux sur les chaînes d'information en continu ou analyse des visages sur les mêmes chaînes, segmentation automatique de programmes ou de sujets, etc. D'autres sont encore à l'étude, comme l'indexation automatique par sujet avec la classification automatique opérée à partir du texte transcrit, pour la rapprocher des 900 principaux descripteurs de notre thésaurus de noms communs.

34 L'automatisation progressive d'un certain nombre de traitements répond principalement à deux cas d'usages : assister le travail des documentalistes pour augmenter la couverture des collections que l'on est capable de traiter et la granularité de certaines descriptions ; faire émerger de nouvelles connaissances par l'analyse d'informations extraites massivement de ces collections. Ainsi, les données générées au sujet des JT alimenteront un site qui donnera une vision de l'activité médiatique des différentes chaînes de télévision, y compris des chaînes d'information continue[19]. Ce changement technique est donc aussi un changement culturel, qui suscitera de nouveaux enjeux. Le premier concerne la validation humaine des données issues de l'IA : quelles données souhaitera-t-on valider, suivant quel processus et à quelles fins ? Le second est apparu alors que nous construisions petit à petit l'équipe chargée de piloter ce nouveau système d'information. Les professionnels capables de manipuler autant de types de modèles physiques, logiques, conceptuels, intellectuels, que ce soit au niveau logique ou au niveau des bases physiques, sont encore peu nombreux. Nous sommes donc confrontés à une lacune dans la formation des professionnels des données.

Du professionnel de l'information au professionnel de la donnée

35 Peu à peu, la démarche consistant à penser les organisations par la donnée, non plus à travers la dimension technique, mais à travers la modélisation et le traitement de la donnée, arrive au cœur de la réflexion sur l'architecture des systèmes d'information : il ne s'agit pas seulement d'une spécificité du secteur culturel et patrimonial. On a ainsi vu émerger le profil de *data analyst* : alors que le *data scientist* fait des algorithmes et le data ingénieur des infrastructures techniques, le *data analyst* fait le lien entre cette vision technique et la vision métier des données.

36 Cette évolution n'est pas terminée, mais pose la question de l'identification des bons profils. Le *data analyst* est plus difficile à former que les *data scientists*, qui sont des mathématiciens, ou les data ingénieurs, qui sont des informaticiens. Il s'agit d'un profil hybride, puisque son rôle consiste à appliquer les technologies numériques dans un domaine métier particulier. Il doit donc à la fois être formé dans ce domaine, s'approprier les technologies numériques alors même que celles-ci deviennent de plus en plus diverses et complexes, et apprendre à modéliser : une compétence spécifique qui n'est pas maîtrisée par les ingénieurs informaticiens, qui ne sont pas formés à cela.

37 Il faut donc réaffirmer l'importance des professionnels de l'information dans cette transformation. Tandis que les entreprises supprimaient les centres de documentation, les documentalistes n'ont pas toujours compris l'importance de s'affranchir du document au sens strict du terme et de creuser la question de la donnée, certes technique, mais face à laquelle ils ont des compétences spécifiques à apporter. Cette double tendance a été à l'origine d'une perte de compétence, que l'on pourrait aujourd'hui compenser en formant de nouveaux professionnels de la donnée, héritiers directs des professionnels de l'information ou de la documentation. À l'heure de former des jeunes professionnels à ces métiers, l'appellation « *data analyst* » permet d'ouvrir un dialogue avec les directions informatiques des entreprises et les SSII[20]. L'enjeu est de leur expliquer nos méthodes et nos convictions et, en échange, de leur demander quelles technologies enseigner à nos étudiants pour qu'ils trouvent leur place parmi les professionnels du système d'information, pas seulement en tant qu'assistants à la maîtrise d'ouvrage [AMOA], mais aussi en tant que véritables spécialistes de la donnée. C'est un travail de *lobbying* de longue haleine, d'autant que le marché n'a pas fini d'évoluer et qu'il sera nécessaire de s'adapter au fur et à mesure de la demande.

NOTES

1. Gautier POUPEAU, « Les carcans de la pensée hiérarchique et documentaire » sur le blog *Les petites cases*, 2009, http://www.lespetitescases.net/carcans-de-la-pensee-hierarchique-et-documentaire-1 et http://www.lespetitescases.net/carcans-de-la-pensee-hierarchique-et-documentaire-2, consulté le 1er novembre 2022.

2. Pour une idée des enjeux des SI documentaires, cf. Gautier POUPEAU, « La donnée : nouvelle perspective pour les bibliothèques », dans Emmanuelle BERMÈS (dir.), *Vers de nouveaux catalogues*,

Paris, éd. du Cercle de la librairie, 2016, p. 159-171, https://www.cairn.info/vers-de-nouveaux-catalogues--9782765415138-page-159.htm, consulté le 1er novembre 2022.

3. Cette partie reprend en grande partie les éléments de l'article suivant : Gautier POUPEAU, « Histoire(s) de notices », dans Lisette CALDERAN, Pascale LAURENT, Hélène LOWINGER et Jacques MILLET, *Le document numérique à l'heure du web*, ADBS, p. 25-40, 2012, Sciences et techniques de l'information, https://hal.inria.fr/hal-00740295/, consulté en ligne le 1er novembre 2022.

4. Le Dublin Core est un format descriptif simple et générique créé en 1995 à Dublin (Ohio) par OCLC [Online Computer Library Center] et le NCSA [National Center for Supercomputing Applications], définition issue du site de BNF https://www.bnf.fr/fr/dublin-core.

5. Pour plus d'informations sur l'intérêt du web sémantique pour les professionnels de l'information, cf. Emmanuelle BERMÈS avec la collaboration d'Antoine ISAAC et de Gautier POUPEAU, *Le Web sémantique en bibliothèque*, Paris, éd. du Cercle de la librairie, 2013 ; ou Emmanuelle BERMÈS, « Web de données et bibliothèques : l'évolution du modèle d'agrégation des données », dans *I2D – Information, données & documents*, 2016/2, vol. 53, p. 37, https://www.cairn.info/revue-i2d-information-donnees-et-documents-2016-2-page-37.htm, consulté le 1er novembre 2022.

6. C'est l'approche que le modèle FRBR [Functional Requirements for Bibliographic Records/ Fonctionnalités requises des notices bibliographiques], par exemple, s'est efforcé d'adopter pour s'affranchir d'une vision documentaire attachée à la description d'objets matériels et aller vers une représentation logique des contenus. FRBR est un modèle conceptuel prenant en compte quatre niveaux d'analyse d'un document : l'objet physique (item), les caractéristiques de sa publication (manifestation), les caractéristiques de son contenu intellectuel (expression), les caractéristiques de la création abstraite à laquelle il se rattache (œuvre). Définition inspirée de celle de la BnF (https://www.bnf.fr/fr/modeles-frbr-frad-et-frsad).

7. *Optical Character Recognition* [Reconnaissance optique de caractères].

8. La notion d'intertextualité, développée par Gérard Genette en 1982, désigne les modes de relations existant entre plusieurs textes. La métatextualité est la relation d'un texte qui commente un autre texte. Gérard GENETTE, *Palimpsestes. La Littérature au second degré*, Paris, Seuil, 1982.

9. Ina'STAT classifie en rubrique les sujets des JT et propose un suivi de l'évolution des thématiques abordées. La méthode utilisée s'appuie sur un thésaurus, les termes du sujet de JT sont automatiquement rattachés à la branche du thésaurus correspondant. Voir l'exposé de la méthode, sur le site de l'INAthèque : http://www.inatheque.fr/publications-evenements/ina-stat/ina-stat-methodologie.html.

10. L'informatique décisionnelle, ou *Business Intelligence*, centralise et consolide aussi les données d'entreprise pour créer des aides à la décision pour les dirigeants.

11. Sur les lacs de données *stricto sensu*, se reporter à la captation de la présentation d'Anne Laurent à la séance du séminaire *Les nouveaux paradigmes de l'archive* le 15 décembre 2020, sur le thème « Des lacs de données pour quelles archives ? ». https://nparchive.hypotheses.org/738.

12. Terme utilisé par Exalead, une société française créée en 2000 par François Bourdoncle et Patrice Bertin. Elle proposait une solution logicielle documentaire reposant sur la technologie des moteurs de recherche.

13. En réalité, le moteur de recherche est une base de données document d'un type particulier, optimisée pour la recherche plein texte. Il existe d'autres bases de données document, comme les bases de données XML ou Json par exemple.

14. Possibilité de faire face à une augmentation d'activité en ajoutant un serveur pour répartir la charge sur plusieurs machines, elle s'oppose à la scalabilité verticale qui impose de faire fonctionner l'application sur une seule machine.

15. Gautier POUPEAU, « Au-delà des limites, que reste-t-il concrètement du Web sémantique ? », sur le blog *Les Petites cases*, http://www.lespetitescases.net/au-dela-des-limites-que-reste-t-il-concretement-du-web-semantique, consulté le 1er novembre 2022.

16. Description de la structure des données d'un système d'information et des relations existant entre elles.
17. Il s'agit d'un modèle entité-relation inspiré de Bibframe et du Cidoc-CRM.
18. Traitement informatique dont on programme à l'avance l'exécution, souvent pour l'exécuter la nuit.
19. Gautier POUPEAU, « En quoi l'intelligence artificielle constitue pour l'Ina une opportunité pour renforcer ses missions de valorisation et de conservation du patrimoine audiovisuel », *Les Futurs fantastiques*, Bibliothèque nationale de France, 10 décembre 2021, https://www.bnf.fr/fr/mediatheque/en-quoi-lintelligence-artificielle-constitue-pour-lina-une-opportunite-pour-renforcer, consulté le 1er novembre 2022.
20. Sociétés de service en ingénierie informatique.

AUTEUR

GAUTIER POUPEAU

Directeur de l'infrastructure des données, service du numérique du ministère de la Culture
De 2014 à 2022, Gautier Poupeau a assuré les fonctions de Data Architect puis de responsable adjoint du département Data de l'Institut national de l'audiovisuel.

Diazinteregio : un réseau pour valoriser la mémoire filmique régionale

Rémi Pailhou, Mevena Guillouzic-Gouret et Stéphanie Ange

1 Une idée est souvent le fait d'une seule personne, mais faire de cette idée un projet d'envergure ne peut être que le fruit d'un travail de coopération. Diazinteregio s'inscrit véritablement dans cette ligne pour faire d'une impulsion régionale une coopération à l'échelle nationale. Mais qu'est-ce que Diazinteregio ? Pour répondre à cette question nous allons, dans un premier temps, rappeler comment le réseau s'est formé. Ensuite, nous reviendrons sur les usages de l'outil qui le fédère : la base de données Diaz. Et enfin, nous aborderons le fonctionnement mutualiste du réseau en abordant les différents chantiers qui ont été menés, la méthodologie appliquée ainsi que les projets en cours[1].

Le réseau Diazinteregio

2 Comment est-on passé d'une base de données, c'est-à-dire d'un outil, à la création d'un réseau national ? À l'origine, la problématique était de prendre en charge le film au format substandard, autrement dit amateur. Il faut se remettre dans le contexte : dans les années 1990, le film amateur n'est pas pris en compte, ou seulement de façon marginale, par les outils de gestion documentaire. Certes, le dépôt légal du film existe, mais aucune mesure spécifique ne concerne l'ensemble des supports amateurs, lesquels n'entrent pas dans le cadre de ce dépôt légal.

Émergence du réseau autour d'un outil commun

3 C'est donc dans les années 1990 que la Cinémathèque de Bretagne, en collaboration avec Normandie Images, développe en interne un outil, la base Klaskatao[2], qui pose le premier jalon de la future base de données Diaz. Ce travail de coopération porte prioritairement sur le traitement documentaire, la gestion des droits, la création du

conventionnement avec les déposants et aboutit à la création d'un thésaurus commun, élaboré par les deux structures qui historiquement sont liées par la thématique de l'activité maritime.

4 En 2004-2005, un outil d'indexation performant, baptisé *Diaz*[3], est développé par la Cinémathèque de Bretagne au cours d'un projet soutenu par le FEDER[4], mené avec la région Bretagne et la région Pays de la Loire. En 2006, la Cinémathèque de Bretagne propose à Normandie Images et Ciclic Centre-Val de Loire une licence Diaz leur permettant d'utiliser l'outil ; ces trois structures, une cinémathèque, la Cinémathèque de Bretagne, et deux mémoires audiovisuelles régionales, Normandie Images et Ciclic Centre-Val de Loire, posent alors les jalons du réseau en devenir.

5 En 2010, le trio de membres fondateurs met en place le cadre administratif du réseau en créant l'association Diazinteregio, dont la mission est de pérenniser et de partager les compétences développées. C'est, à l'époque, Gilbert Le Traon[5], le directeur de la Cinémathèque de Bretagne, qui préside Diazinteregio, Emmanuel Porcher[6], le directeur de Ciclic Centre-Val de Loire, en est le secrétaire général, et Denis Darroy[7], directeur de Normandie Images, le trésorier. Le bureau est aujourd'hui toujours composé des trois structures historiques et fondatrices du réseau. Ce réseau a des objectifs communs : créer un standard commun pour l'indexation d'œuvres cinématographiques et audiovisuelles, étudier les perspectives d'évolution de ce logiciel, mutualiser les informations contenues dans les bases des différentes structures, favoriser les coopérations interrégionales et mettre en œuvre toutes les actions visant à assurer la promotion de nos différentes missions. À ces trois structures s'agrègent, dès 2010, d'autres structures dont la Cinémathèque des Pays de Savoie et de l'Ain, la Cinémathèque d'images de montagne et Ciné Archives, qui, elles aussi, ont au préalable acheté une licence pour utiliser Diaz.

La philosophie de Diazinteregio

6 Au fil des années Diazinteregio a développé une philosophie de coopération reposant sur un modèle de partenariat ouvert. Tout d'abord, il s'agit de faciliter l'arrivée des nouveaux entrants. Tous les ans des petites, moyennes ou grandes structures rejoignent le réseau, et l'objectif est de les soutenir, sans distinction, en leur proposant un outil adapté à leurs besoins. Le réseau porte aussi attention aux structures émergentes et c'est la raison pour laquelle l'obtention de Diaz a été simplifiée. Il n'est plus besoin aujourd'hui de s'acquitter d'une licence pour bénéficier de la base. En revanche, le principe d'un engagement financier commun a été mis en place. Chaque structure qui adhère à Diazinteregio paye un droit d'entrée, mais aussi une cotisation à un fonds de mutualisation pour financer les développements.

7 Les réflexions sur les évolutions de l'outil sont communes et structurées par la coordinatrice du réseau, autour de différentes commissions qui se consacrent à des thématiques très précises, comme la technique, les appels à projets ou les aspects documentaires. Le réseau dispose d'un plan de formation pour appréhender l'outil et les nouveaux adhérents, comme les anciens, ont droit à un temps de formation à la prise en main de l'outil ou à ses nouvelles fonctionnalités.

Poids et avenir de Diazinteregio

8 C’est ainsi que Diazinteregio a abouti à un vrai maillage du territoire national, même s’il reste quelques « zones blanches » dans le Sud-Ouest, car certaines structures ont développé leurs propres solutions en interne. C’est le cas, notamment, de la Cinémathèque de Nouvelle-Aquitaine à Limoges, qui a développé le Pill et anime un réseau régional avec le Fonds audiovisuel de recherche [FAR], de Trafic Image à Angoulême et de la Mémoire de Bordeaux Métropole[8]. Les structures qui conservent des films de patrimoine utilisent Garance[9], un outil développé par le Centre national de la cinématographie [CNC] qui échange régulièrement avec Diazinteregio pour que les outils restent complémentaires.

Figure 1. Carte des membres du réseau Diazinteregio au 31 mai 2022.

9 Le maillage autour de Diaz évolue constamment avec l’entrée de nouvelles structures professionnelles ou bénévoles, situées sur le territoire métropolitain, mais aussi en outre-mer et sur certains territoires francophones comme la Suisse ou le Maroc. Chaque région ou presque possède une structure qui utilise la base Diaz, ce qui permet d’élargir le réseau et les différentes collections qu’il traite. Cela permet aussi d’enrichir le savoir-faire autour des fonds audiovisuels, cinématographiques et photographiques, car les structures-membres sont indépendantes et ont leurs méthodes de travail propres. Le réseau est un lieu d’échanges et de discussions autour de leurs pratiques pour atteindre une certaine homogénéisation, par exemple avec le thésaurus. Car Diaz gère une collection unique, constituée de milliers de supports films, de photographies, parfois d’affiches de films ou d’appareils. Avec plus de dix ans d’expérience dans le développement et l’usage d’outils documentaires, Diazinteregio est désormais reconnu par ses pairs, le CNC et l’INA, comme un acteur à part entière du patrimoine audiovisuel.

Usages et évolutions de la base de données Diaz

10 Pour comprendre le fonctionnement communautaire de Diaz, il est important d'appréhender la diversité de ses membres.

La communauté Diazinteregio

11 La communauté Diazinteregio est composée de 18 structures[10] qui présentent différentes formes juridiques. Ce sont majoritairement des associations, car on compte quinze associations de type loi 1901, ainsi que deux établissements publics de coopération culturelle et un centre d'archives municipales, qui ont tous pour particularité de travailler sur des fonds thématiques et principalement sur le film amateur. Bien que les couvertures des collections dont elles sont dépositaires relèvent d'échelles différentes, régionales, départementales ou municipales, elles sont fortement centrées sur le rayonnement territorial. Elles conservent ainsi des films tournés dans leur région ou des films tournés hors de leur région, mais dont les auteurs sont issus de la région. Cependant, trois parmi ces dix-huit structures ont un fonctionnement thématique, comme par exemple le centre audiovisuel Simone de Beauvoir (histoire des femmes et des luttes féministes), Ciné Archives (histoire du Parti communiste et des mouvements ouvriers) ou French Lines et Compagnies (histoire de la Marine marchande française). Les types de collections diffèrent aussi. Elles sont consacrées principalement au film amateur, mais beaucoup de structures conservent également du film professionnel, des bandes son et ce qu'on appelle le « non film », c'est-à-dire des photographies, des appareils, des affiches, des livres, etc.

Usages des films amateurs

12 Rappelons que les fonds cinématographiques patrimoniaux sont un matériau extrêmement intéressant dont les usages sont multiples. La première mission est, bien sûr, la construction mémorielle : conserver et restituer l'image à ses déposants, mais la vente d'images connaît aussi une forte demande. Des réalisateurs ou des sociétés de production achètent des extraits pour les réutiliser dans de nouvelles productions et les expositions utilisent de plus en plus ces fonds dans leurs muséographies. En termes de diffusion culturelle, beaucoup de créations contemporaines, de spectacles, de ciné-concerts et de ciné-théâtres se font aujourd'hui à partir de ces archives et donnent une nouvelle vie à l'image. Le *mashup*[11] est de plus en plus utilisé pour simplifier l'initiation au montage. On y utilise aussi des prises de vue contemporaines, mais l'image d'archive comporte une dimension pédagogique intéressante pour les plus jeunes : on peut, par exemple, à partir de plusieurs films d'archives qui se passent dans une ferme, recréer, par un procédé très simple, la journée type d'un fermier des années 1960. C'est à la fois un travail sur l'histoire et un travail de montage et d'éducation à l'image.

Figure 2. Photogrammes.

A) *Racleurs d'océans*, Anita Conti, 1952, Cinémathèque de Bretagne.
B) *Saint-Léonard, Mont-Saint-Michel*, Fernand Bignon, 1940, Normandie Images.
C) *Pont du Mont-Blanc et rade de Genève*, auteur inconnu, 1945, Autrefois Genève.

13 La base Klaskatao, qui, rappelons-le, a précédé Diaz, était spécialement conçue pour documenter le film amateur. Par la suite, les besoins et les techniques changeant, la base de données a dû constamment évoluer afin de rester en accord avec les usages professionnels de chacune des structures adhérentes. C'est un outil très complet qui permet de gérer toutes les collections des membres ainsi que leur valorisation et même leur traitement administratif. Ainsi, il intègre des « fichiers de collection » qui permettent de traiter tout ce qui concerne les collections, comme les « fiches documentaires » où l'on indique, par exemple, le titre, le résumé, les descripteurs thématiques et de lieu. Les « fiches supports » consignent le format original, le type de coloration, la présence d'une bande son, l'état du support physique, le temps passé à le traiter et toutes les opérations effectuées sur la pellicule, tandis que les « fiches appareils » sont consacrées aux matériels. Diaz intègre aussi des « fichiers de gestion » qui sont des sortes de fiches projets. Ces derniers permettent de garder une trace de l'ensemble des utilisations des films par exemple les diffusions culturelles ou les ventes d'images. Les ateliers sont également documentés dans cet outil.

Figure 3. Photogrammes.

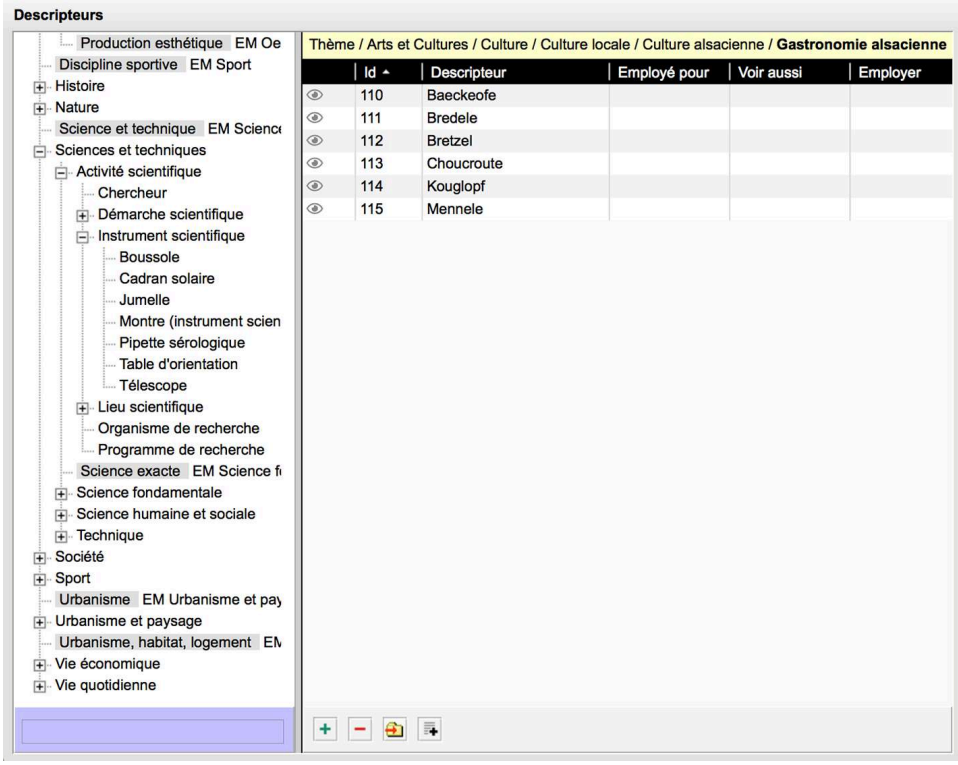
A) *Fête du retour des Poilus à Châteauroux*, Maurice Brimbal, 1919, CICLIC.
B) *Conchette*, Claude Marcelin et Claude Bondier, 1974, Archives départementales de Savoie – CPSA.
C) *Été 1970 à Calais*, Patrice Gobert, Archipop.

Un thésaurus commun

14 Le projet de refonte et d'harmonisation du thésaurus s'est déroulé de 2016 à 2019. Dans les années 1990, avaient eu lieu les premiers travaux d'élaboration d'un thésaurus commun entre la Cinémathèque de Bretagne et Normandie Images. Il comportait trois parties : les lieux (termes géographiques, administratifs et localités), les noms propres et les thèmes, très développés autour de l'activité maritime et agricole, en raison des contenus des films détenus par ces structures à cette époque. Avec le déploiement de la nouvelle base Diaz en 2005, la création de Diazinteregio et l'arrivée de nouvelles structures, le thésaurus initial faisait régulièrement l'objet de modifications et d'ajouts

chez les uns et les autres, car chaque structure avait son propre fichier thésaurus, qui pouvait ainsi évoluer de manière indépendante. La Cinémathèque de Bretagne, par exemple, possède des films de vacances au ski qui sont indexés directement dans la catégorie « loisirs » ou « sports d'hiver » ou encore en « sports ». Mais la Cinémathèque des Pays de Savoie et de l'Ain et la Cinémathèque d'images de montagne [Cimalpes] ont des besoins beaucoup plus précis pour tout ce qui traite du ski. De la même manière, la Cinémathèque de Bretagne a un vocabulaire très précis sur les oiseaux marins ou sur les coiffes, ce qui intéresse relativement peu les autres structures et peut les « noyer » au moment de l'indexation de leur propre collection. C'est alors qu'émerge l'idée d'unifier les bases et de se doter d'un thésaurus commun pour mettre en œuvre, dans le futur, une plateforme commune permettant d'interroger toutes les bases.

Figure 4. Thésaurus.



15 Dès 2013, la démarche est soutenue par le CNC qui incite notamment à la mise en conformité avec la norme européenne sur la description documentaire des films[12]. Puis une étape d'étude des besoins est menée par Diazinteregio dès 2017 auprès de toutes ses structures pour identifier les différents manques et les problèmes qui pourraient survenir avec ce thésaurus. Le projet est confié à la société Ourouk qui travaille sur l'existant. Plus de 24 000 termes sont nettoyés et reclassés, les doublons sont supprimés. Le nouveau thésaurus livré par Ourouk est presque entièrement remodelé, même s'il ressemble encore très fortement à l'ancien. Les classements hiérarchiques sont beaucoup plus cohérents et les particularités locales sont conservées. Les structures migrent vers le nouveau thésaurus en fin d'année 2019.

16 Pour éviter les ajouts de termes excessifs par les membres, les structures créent des termes-candidats en indiquant leur place dans l'arborescence, puis une commission Diazinteregio valide ou non l'ajout de ces mots-clés, ce qui permet de poursuivre le

développement du thésaurus en fonction des besoins de chacun tout en conservant un fonctionnement beaucoup plus harmonisé.

17 Pour faciliter ce fonctionnement mutualiste, Diazinteregio a décidé, à la fin de l'année 2014, d'embaucher Stéphanie Ange en tant que coordinatrice du réseau.

Un fonctionnement mutualiste

18 Au départ, chaque structure était dotée d'une version serveur de Diaz, installée chez elle à demeure, et pouvait donc modifier le code de la base. Chacun était libre de développer l'outil dans le sens qu'il souhaitait. Ce ne pouvait pas être très opérant sur le long terme, si l'on voulait intégrer de nouveaux utilisateurs. Une version *cloud*[13] de l'outil Diaz a donc été élaborée entre 2013 et 2015 pour permettre une base commune qui ne soit plus modifiable par chaque membre, mais qui soit véritablement partagée. Un autre enjeu important a été l'arrivée, entre 2015 et 2017, de pratiques de numérisation plus massives et dématérialisées qui nécessitent de pouvoir indexer des fichiers numériques et leurs espaces de stockage informatiques et non plus des supports analogiques.

Des chantiers d'adaptation aux pratiques

19 Il a donc fallu mener le chantier dans plusieurs directions, d'une part, en travaillant aux fonctions d'indexation des fichiers numériques natifs et issus de la numérisation, tout en ménageant la possibilité de renumériser, car des campagnes de renumérisation se pratiquent désormais. L'outil prévoit donc de renseigner quatre types de fichiers : un fichier numérique natif, un fichier de conservation, un fichier d'exploitation ou pivot, qui permet ensuite de créer, en règle générale, jusqu'à trois formats de fichiers de diffusion, soit un fichier de consultation, un fichier de diffusion en ligne et parfois un DCP[14]. Mais chaque structure est libre d'utiliser les types de fichiers de diffusion comme elle le souhaite. D'autre part, un chantier a été mené en 2018, pour améliorer la gestion des collections de photographies. Initialement, la partie photographie de Diaz a été pensée pour indexer des photogrammes, des photographies de tournage ou des photographies vraiment liées aux collections de films, mais de nouveaux membres, comme la Fabrique de patrimoines en Normandie, Image'Est et Autrefois Genève notamment, conservent d'importantes collections de photographies originales et il a fallu ajouter tous les champs spécifiques à l'indexation de collections photographiques particulières.

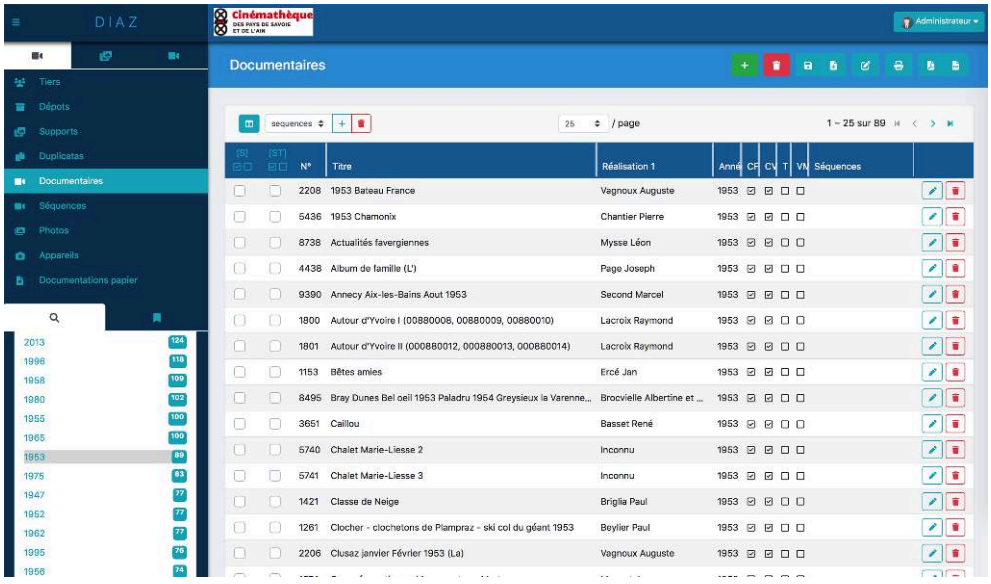
20 Les différences d'organisation que l'on constate, notamment entre les structures territoriales et les structures thématiques aux échelles variées, permettent une grande complémentarité et soulèvent des questionnements qui enrichissent la réflexion de l'ensemble des membres. Par exemple, Ciné Archives travaille sur des films professionnels avec plusieurs copies, alors qu'un grand nombre de structures du réseau travaillent principalement sur des fonds inédits et sur des copies uniques, ce qui conduit à s'interroger sur l'opportunité de créer des champs tels que « support unique » ou « support de référence » quand on a plusieurs supports d'un même contenu.

Concertation, travail en commun et arbitrages

21 Cette question sera d'abord posée à distance, car, comme le réseau s'étend sur le territoire national, il est difficile de se rencontrer en présentiel très régulièrement. Ces premiers échanges ont donc lieu par mail, *via* une liste de diffusion suivie par un responsable de Diaz dans chaque structure. Ils sont suivis, après une première synthèse des retours, par des débats en présentiel, au cours des réunions trimestrielles du réseau, lorsque les synthèses révèlent des différences de vues. Les arbitrages sont plus ou moins simples à effectuer, mais, quoiqu'il en soit, le fonctionnement fait que personne n'est contraint dans sa façon de travailler et l'arbitrage vise surtout à déterminer les priorités.

22 Ces réunions trimestrielles impliquent un fort engagement des personnels des structures. Un correspondant consacre au minimum cinq jours par an au réseau pour répondre aux échanges et assister aux réunions, ce qui peut être étendu à dix jours par an s'il participe à une commission thématique. Il existe actuellement trois commissions : la commission « IHM[15] », à laquelle participent trois membres, concerne la refonte de la base actuelle ; la commission « genre cinématographique » a refondu la liste des termes de genres en prenant en compte les travaux du CNC sur ces termes ; enfin, une commission « financement » contribue au développement de projets. Les temps de formation proposés chaque année aux nouvelles structures adhérentes et aux nouveaux personnels des membres plus anciens sont des moments d'échanges moins formels sur les pratiques. Au cours de l'apprentissage de l'outil, les petites variations d'utilisation ou de conception de certains champs apparaissent et sont discutées pour essayer, toujours, d'harmoniser les fonctionnements.

Figure 5. L'interface Homme-Machine de DIAZ en cours de développement.



23 Au fil du temps Diazinteregio est aussi devenu un service, en offrant la possibilité d'échanger par mail ou par téléphone pour obtenir une assistance à l'utilisation. De plus, depuis le travail sur les fichiers numériques, à la fin de 2018, un serveur partagé et une solution de partage de fichiers ont été ajoutés pour la mise en ligne des contenus et la livraison des fichiers de conformation pour différents usages.

Un projet en cours : la création d'un portail commun

24 Le réseau Diazinteregio permet la création d'un véritable catalogue de la mémoire filmique du territoire national, mais dont l'accès relève encore de la mise en ligne des membres sur leur site régional, local ou thématique. L'objectif est de créer un portail commun de publication en ligne de ces contenus et, pour cela, deux étapes majeures sont programmées en 2022-2023. D'une part, la modernisation de l'infrastructure de l'interface homme-machine permettra de réaliser beaucoup plus vite les formulaires de saisie et de personnaliser les formats de listes. D'autre part, des fonctionnalités seront ajoutées comme le visionnage du film dans la notice documentaire ou la navigation de la notice documentaire à la notice support. La description technique sera également étoffée et les types de plans, l'exposition, la qualité cinématographique pourront être décrits. Ces évolutions anticipent le projet d'un portail commun de mémoire filmique des territoires dont l'ouverture est envisagée en 2023. Cette plateforme commune offrira un axe d'éditorialisation pour valoriser certaines thématiques et certains contenus des membres. Une interrogation croisée des collections conduira à l'ensemble des documents qui traitent, par exemple, d'une même commune. En effet, même si chacun couvre une mission territoriale, une structure conserve parfois des fonds qui intéressent d'autres régions, car un réalisateur breton peut avoir tourné à Saint-Étienne ou inversement. La phase de concertation est entamée en 2022 afin de statuer sur les envies et besoins des membres, et surtout sur l'articulation entre les plateformes individuelles des structures, les plateformes régionales, les plateformes thématiques et ce portail commun.

25 À l'horizon 2024-2025, c'est un rapprochement avec le monde des archives qui est en gestation. Les membres de Diazinteregio travaillent beaucoup avec des archives départementales, notamment MIRA [Mémoire des Images Réanimées d'Alsace] avec les Archives départementales d'Alsace ou la Cinémathèque des Pays de Savoie et de l'Ain avec les Archives départementales de la Savoie, de la Haute-Savoie et de l'Ain. Les archivistes départementaux, qui sont proches de ces cinémathèques, ont déjà l'habitude de travailler sur Diaz pour y renseigner des collections et d'autres Archives départementales se sont montrées intéressées par l'outil. Mais rajouter un outil à ceux qu'utilisent déjà les archivistes, et pour ne traiter qu'une partie de leurs collections, n'est sans doute pas la bonne approche. Ce qui est envisagé est d'obtenir une interopérabilité plus grande entre les outils de type Arkheia ou Ligeo pour la gestion archivistique et Diaz pour l'indexation afin de faciliter la valorisation de leurs fonds audiovisuels.

26 Enfin, le réseau regarde aussi au-delà des frontières, en Italie par exemple, avec *Home Movies* à Bologne, pour défendre et dupliquer le modèle du réseau à l'international. Le modèle est facilement réplicable, mais le questionnement porte sur la traduction du thésaurus : faut-il traduire les milliers de termes communs ou repartir d'un thésaurus vierge ?

NOTES

1. Ce texte est issu de la table ronde qui s'est tenue le 31 mai 2022 dans le cadre du séminaire Les Nouveaux Paradigmes de l'Archive. Nous remercions Gaïd Pitrou, directrice de la Cinémathèque de Bretagne, pour sa contribution à l'élaboration de cette table ronde.
2. Ce qui veut dire « cherche toujours » en breton.
3. Ce qui signifie « base » en breton.
4. Le fonds européen de développement régional [FEDER] intervient dans le cadre de la politique de cohésion économique, sociale et territoriale.
5. Gilbert Le Traon est l'un des fondateurs du Festival européen du film court de Brest en 1986. Délégué général en 1991, puis directeur artistique jusqu'à sa 15e édition en novembre 2000, il est recruté cette même année pour diriger la Cinémathèque de Bretagne. Il rejoint l'équipe de Ciclic, en prenant la responsabilité du pôle patrimoine le 1er septembre 2015.
6. Emmanuel Porcher dirige Centre Images, l'Agence régionale du Centre pour le cinéma et l'audiovisuel de 2006 à 2011. Directeur général délégué Culture, patrimoine et sports, puis directeur général délégué Éducation, égalité des chances et vie citoyenne, au conseil régional Centre-Val de Loire, il est nommé en 2022 directeur général des services de la Ville de Tours.
7. Denis Darroy est producteur et réalisateur de films documentaires et d'animation, ainsi que coordinateur de dispositifs d'éducation au cinéma et à l'audiovisuel. De septembre 2001 à juillet 2009, il a occupé le poste d'inspecteur conseiller pour le cinéma, l'audiovisuel et le multimédia au sein des Drac de Lorraine et d'Alsace. Depuis août 2009, il a pris ses fonctions de directeur du pôle image de Haute-Normandie.
8. https://www.memoirefilmiquenouvelleaquitaine.fr/
9. Garance est le fruit d'une collaboration entre la Cinémathèque de Toulouse, la Cinémathèque française et l'Institut Jean-Vigo.
10. En 2022.
11. Le *mashup* consiste, dans différents domaines artistiques, à créer une œuvre par l'assemblage de fragments d'autres œuvres.
12. Identification des films-Jeu minimal de métadonnées pour œuvres cinématographiques, NF EN 15744.
13. C'est-à-dire complètement hébergée sur le web et accessible *via* Internet.
14. Digital Cinema Package [DCP]. Format numérique destiné à l'utilisation en salle de projection.
15. Interface Homme-Machine.

AUTEURS

RÉMI PAILHOU

Ciclic Centre – Val de Loire. Membre fondateur de Diazinteregio aux côtés de Normandie Images et de la Cinémathèque de Bretagne (histoire du réseau Diazinteregio, de sa base de données, de ses valeurs de mutualisation des savoirs et des coûts).

MEVENA GUILLOUZIC-GOURET

Cinémathèque de Bretagne. Membre fondateur de Diazinteregio (usages de la base de données Diazinteregio et élaboration d'un thésaurus commun)

STÉPHANIE ANGE

Coordinatrice du réseau Diazinteregio (accompagnement de la construction sur mesure de la base de données avec le collectif).

Synchroniser la rédaction des métadonnées et la fabrication des données audiovisuelles numériques : en direct depuis le procès des attentats terroristes du 13 novembre 2015

Claire Scopsi, Martine Sin Blima-Barru et Aurore Juvenelle

1 Du 8 septembre 2021 au 29 juin 2022, le procès des attentats terroristes du 13 novembre 2015 s'est tenu dans la salle d'audience spécialement construite dans la salle des pas perdus du Palais de justice de Paris. Pendant dix mois, quatorze accusés ont été jugés dans une instance judiciaire placée dans le domaine de l'exceptionnel par l'architecture du prétoire, le nombre des parties civiles et des avocats, le nombre de salles de retransmission, la présence, pour la première fois, d'écrans de retransmission dans la salle principale et l'enregistrement de 705 heures d'archives audiovisuelles de la Justice [AAJ]. Autre fait extraordinaire, les Archives nationales étaient présentes tous les jours avec, pour mission, la transformation de l'oralité du débat propre à la cour d'assises en transcription et en métadonnées.

2 Il s'agit ici de réfléchir aux conditions de valorisation des débats, enregistrés d'une façon intégrale, et notamment à la production des métadonnées sémantiques qui donneront accès à leurs contenus pour créer des archives historiques.

3 Les trois auteures de l'article ont participé à des degrés divers à ce projet : Martine Sin Blima-Barru, conservatrice du patrimoine, responsable du département de l'Administration des données, intervient au Tribunal en tant que garante de la pérennisation des films et du respect des préconisations de la production de données numériques, puisque les Archives nationales sont dépositaires des fonds des AAJ et responsables de leur conservation et de leur valorisation. Claire Scopsi, chercheuse en

sciences de l'information, a travaillé en amont du procès à la conception du dispositif d'annotation en direct des débats, qui a été testé et utilisé pendant le déroulement du procès. Aurore Juvenelle, historienne et documentariste, a assisté à la totalité du procès. Elle était chargée de la production des métadonnées descriptives de l'événement pour le compte des Archives nationales.

4 Nous évoquons dans cet article deux types de dispositifs : le dispositif de filmage, élaboré par le ministère de la Justice au titre des archives audiovisuelles de la Justice, et dans lequel notre projet s'intègre, et le dispositif de documentarisation que nous avons expérimenté pour donner accès aux contenus.

5 Il faut comprendre ici le terme dispositif au sens que lui donne Michel Foucault :

> un ensemble résolument hétérogène comportant des discours, des institutions, des aménagements architecturaux, des décisions réglementaires, des lois, des mesures administratives, des énoncés scientifiques, des propositions philosophiques, morales, philanthropiques ; bref, du dit aussi bien que du non-dit. (Foucault, 1994 [1977], p. 299)

6 Il est donc question, dans les deux dispositifs, de la mise en place et de la coordination d'équipes, d'outils, de méthodes, dans des contraintes économiques, réglementaires, spatiales et/ou techniques. Mais les objectifs sont différents.

7 Le dispositif de captation mis en place par le ministère de la Justice vise à produire une trace du procès, c'est-à-dire à donner une permanence à cet événement avec un soin constant de produire un résultat neutre, en limitant les effets narratifs. Le dispositif de documentarisation est mis en place à l'initiative des Archives nationales. La documentarisation traduit, selon Jean-Michel Salaün, la perspective « de la recherche et de l'accès » (2017, paragraphe 2).

> Documentariser, ce n'est ni plus ni moins que traiter un document comme le font, ou le faisaient, traditionnellement les professionnels de la documentation (bibliothécaires, archivistes, documentalistes) : le cataloguer, l'indexer, le résumer, le découper, éventuellement le renforcer, etc. [...] L'objectif de la documentarisation est d'optimiser l'usage à venir du document en permettant un meilleur accès à son contenu et une meilleure mise en contexte. (Salaün, 2017, paragraphe 8)

8 L'article produit en collaboration par les trois auteures présente, dans sa première partie, le contexte exceptionnel du procès, tant en termes de durée que du nombre de parties, de témoins et de publics, caractéristiques qui ont pesé sur le dispositif de captation et sur le déroulement du procès lui-même. Il revient également sur l'intérêt que présente pour le monde scientifique l'étude de cet événement judiciaire hors norme. La deuxième partie présente le déroulement du projet d'élaboration d'outils et de méthodes d'annotation dans les contextes temporels et matériels imposés. Y est présenté notamment le logiciel Annotate-on Event développé pour la circonstance. Enfin, dans une troisième partie, l'expérience d'Aurore Juvenelle témoigne du quotidien du processus de documentarisation tout au long du procès et en montre les limites. Nous concluons sur les résultats provisoires de l'expérience et la suite à donner au projet.

Un contexte exceptionnel

Le contexte légal

9 La loi du 11 juillet 1985, voulue par le garde des Sceaux Robert Badinter, a créé une catégorie spécifique d'archives historiques qu'on appelle les archives audiovisuelles de la Justice. À la date de son adoption, elle est dérogatoire par rapport à l'interdiction opposée aux journalistes de filmer, capter le son ou photographier les procès, portée par la loi du 29 juillet 1881[1]. Le code du patrimoine a intégré la loi sur les archives audiovisuelles de la Justice, mais en lui concernant un régime particulier par rapport à celui des autres archives traitées dans le même livre du code. Les archives audiovisuelles de la Justice sont désignées d'emblée dans la loi comme étant des archives « historiques », ce qui est en soi assez intéressant, alors que dans le vocabulaire métier des archivistes le terme désigne les archives définitives. Certes ces archives audiovisuelles de la Justice ont comme destination une entrée immédiate aux Archives nationales, où n'entrent que des archives historiques ou définitives, mais on peut se demander s'il suffit de légiférer pour « faire » des archives historiques et si des conditions ne sont pas requises au-delà de ce qui est dit dans la loi, laquelle stipule simplement que les archives doivent être intégrales et filmées avec une caméra fixe.

10 Par ailleurs, la loi du 22 décembre 2021 pour la confiance dans l'institution judiciaire, promue par Éric Dupond-Moretti, a en revanche modifié la loi de 1881 en introduisant la possibilité que les audiences de justice civile, pénale, économique ou administrative soient enregistrées ou filmées pour un motif d'intérêt public d'ordre pédagogique, informatif, culturel ou scientifique.

11 À l'occasion du procès des attentats terroristes du 13 novembre 2015, le ministère de la Justice, la cour d'appel de Paris, qui a hébergé dans la salle des pas perdus une salle spécialement construite, et les Archives nationales ont mis en œuvre des conditions spécifiques non seulement pour que le procès se déroule, mais pour qu'il puisse être filmé et que ces archives soient créées. C'est aussi une création qui s'inscrit dans un espace dédié avec une organisation et une gestion architecturale spécifiques.

Contexte matériel et dispositif de filmage

12 La création d'un dispositif spécifique était nécessaire en raison du caractère exceptionnel du procès, lié non seulement à sa durée (dix mois), au nombre de parties civiles (plus de 2 000 représentées par 350 avocats) et à l'intérêt du public et de la presse (150 médias français et étrangers ont suivi l'audience)[2]. Il fallait donc assurer l'accueil et la sécurité du public, dont le nombre s'élevait certains jours à plusieurs milliers de personnes, et garantir la publicité des débats, sans nuire à l'exercice de la justice. Une salle d'audience longue de 45 mètres, large de 15, a été construite pour l'occasion dans la salle des pas perdus du Palais de justice. Cinq cents places y sont réservées au public accrédité : les parties civiles, les avocats et les membres de l'organisation du Palais de justice, ainsi que des journalistes, des chercheurs et les représentants des Archives nationales. Trois autres salles accueillent, l'une, les parties civiles ne pouvant ou ne voulant pas trouver place dans la salle d'audience, une autre, les avocats des parties civiles, et la troisième, les journalistes et les chercheurs. Elles

sont reliées par un réseau audiovisuel à la salle d'audience. Enfin, un espace composé de plusieurs salles accueille le public non accrédité.

13 Dans la salle d'audience, huit caméras assurent la captation des débats selon les règles fixées par le code du patrimoine et mises en application par le ministère de la Justice. Les images produites ont deux fonctions : d'une part, constituer les archives historiques du procès, d'autre part, assurer la publicité des débats depuis les différents écrans installés dans tous les espaces[3]. Des écrans sont disposés dans les différentes salles, y compris dans la salle d'audience où ils offrent au public présent des points de vue diversifiés (des gros plans, les accusés sont vus de face).

14 Un flux audio est diffusé par webradio, auprès des parties civiles qui ne peuvent être présentes au Palais de justice. Cet enregistrement, qui leur est réservé, est diffusé avec un décalage de trente minutes. Aucune image ne sort de l'enceinte du Palais de justice.

Contexte scientifique : un lieu d'étude

15 Le dispositif audiovisuel du procès, innovant et imposant, fait l'objet de plusieurs réflexions dans le domaine des études visuelles. Sylvie Lindeperg[4], historienne du cinéma, analyse la singularité de l'image, notamment des photographies sélectionnées et floutées, diffusées par la Cour pendant les débats judiciaires, alors que les images les plus violentes ont circulé sans précaution sur Internet au lendemain des faits[5]. Romane Gorce pose la question de l'influence des choix de prises de vue sur la perception du procès par le public et surtout par les journalistes, car « relier par la vidéo une salle de retransmission à la salle d'audience nécessite d'opérer des choix sur ce que l'on peut et veut retransmettre » (Gorce, 2022, paragraphe 6).

16 Ces travaux, livrés « à chaud » dans les mois qui ont suivi la clôture du procès, montrent l'enjeu de la valorisation de ces archives sur lesquelles les études en cours pourront s'appuyer pour se développer. Notre problématique est pragmatique : la loi du 11 juillet 1985 vise à constituer des archives historiques audiovisuelles de la Justice et en réglemente la prise de vue et la conservation. Elle statue sur les conditions de leur communication, mais elle ne dit rien de leur exploitabilité. En 2021, Martine Sin Blima-Barru et Christian Delage ont posé les données du problème en ces termes : « Accéder, diffuser ou réutiliser les archives nécessite un encadrement juridique adéquat, mais aussi des conditions scientifiques et techniques rendant matériellement possible leur consultation » (Sin Blima-Barru et Delage, 2021, paragraphe 15). Le procès des attentats du 13 novembre 2015 et ses huit caméras qui produisent pendant dix mois des centaines d'heures de captation sont donc l'occasion d'une recherche-action, qui vise à construire un dispositif de création de points d'entrées permettant d'accéder ensuite aux films. Le projet se déroule en deux étapes qui présentent chacune des temporalités et des contraintes propres. Six mois avant le procès, Claire Scopsi, chercheuse en sciences de l'information, et les Archives nationales travaillent à l'élaboration d'outils et de processus de contextualisation et d'indexation. Aurore Juvenelle, accréditée pour la totalité de la durée du procès, applique ces méthodes et les confronte à la réalité, en lien avec Martine Sin Blima-Barru.

Indexer des vidéos sans vidéos

17 Martine Sin Blima-Barru et Christian Delage formulaient quelques préconisations de bon sens : « définition d'un vocabulaire d'indexation adapté au contexte, qui permette ensuite d'annoter à la volée le flux vidéo une fois le procès terminé, prise d'un verbatim au fur et à mesure du déroulement du procès ou enregistrement sonore qui pourrait ensuite être "travaillé" pour créer une source textuelle » (Sin Blima-Barru et Delage, 2021, paragraphe 20).

18 Apporter des métadonnées de contexte à un enregistrement vidéo peut sembler banal. C'est une opération couramment pratiquée qui ne nécessite pas de mobiliser des chercheurs. Dans le cadre de traitement de fonds documentaires vidéo, les documentalistes travaillent ainsi sur des enregistrements déjà réalisés et conservés dans leurs institutions. Ici, il faut penser qu'il s'agit d'indexer des vidéos qui n'existent pas encore. En avril 2022 nous sommes encore très ignorantes de l'organisation du procès, des dispositifs de filmage et de la teneur des échanges captés. Nous avons cependant quelques données de départ :

- dix mois de captations produisent plusieurs milliers d'heures d'enregistrement[6] ;
- si, à l'issue du procès, aucune partie ne formule d'appel, les enregistrements doivent être communicables immédiatement après le procès pour le public des Archives nationales qui a un motif de recherche scientifique et historique ;
- il n'est prévu aucun système de sténotypie des débats pendant le procès, le principe de l'oralité des débats prévalant en cour d'assises ;
- les films seront versés à l'issue du procès, avec un tableau de suivi technique, réalisé par les opérateurs de l'enregistrement, indiquant l'ouverture de la session et sa suspension et associant chacune de ces séquences au nom du fichier numérique enregistré correspondant.

19 Nous avons donc recherché une solution tenant compte de ces quatre contraintes.

Transcrire ou indexer ? Les choix technologiques et méthodologiques

20 À l'heure de l'intelligence artificielle, les premières réflexions s'orientent naturellement vers l'utilisation d'un logiciel de reconnaissance automatique de la parole qui sera exécuté après le procès et dans l'enceinte des Archives nationales, puisque les enregistrements ne sont pas livrés au fur et à mesure du déroulement du procès.

21 La solution apparaît séduisante, toutefois elle nécessite de livrer l'algorithme à un apprentissage (*deep learning*), à l'aide d'échantillons des corpus à traiter, associés à une relecture et une correction humaines. Le temps d'apprentissage varie en fonction des caractéristiques des corpus, mais, compte tenu du volume à traiter, l'investissement paraît rentable. Mais le corpus de ce procès présente un inconvénient majeur : l'apprentissage s'effectue en général sur chaque voix, car l'algorithme doit s'habituer aux inflexions spécifiques du locuteur pour traduire le son émis en texte. Or le procès donne la parole à de nombreux intervenants et certains ne s'expriment qu'une fois. L'apprentissage ici ne pourra pas être rentabilisé.

22 L'algorithme a d'autant plus de difficultés à reconnaître la parole (et donc, nécessite un apprentissage important) que le son est « propre », sans bruits parasites ni

chevauchements de paroles. Au cours du procès, le son fourni est capté par les micros placés dans le prétoire, ce qui limite les bruits de fond en provenance du public. Au président revient le pouvoir d'activer les micros et ainsi de distribuer la parole, comme le veut le rituel judiciaire. Ainsi, « toute parole inopinée devient-elle [...] inaudible » (Gorce, 2022, paragraphe 3). Le dispositif produit donc un son favorable à la reconnaissance automatique. Cependant, d'autres facteurs entrent en compte. La parole du locuteur peut être brouillée par l'émotion, l'âge ou l'état de santé, ou bien elle présente des accents divers et peut être produite en langues étrangères. Toutes ces circonstances font baisser le taux de reconnaissance.

23 Les efforts à fournir pour optimiser la reconnaissance dépendent de l'usage que l'on projette. Si l'on envisage seulement d'utiliser le texte restitué pour fournir des points d'accès à la vidéo via un moteur de recherche avant de la visionner, une relecture n'est peut-être pas nécessaire si l'on suppose que l'algorithme peut reconnaître suffisamment de termes pour nourrir la recherche. Malheureusement, tous les termes n'ont pas la même valeur. Pour exploiter des archives historiques, les entités nommées (les noms de personnes, de lieux ou d'organisations par exemple) sont essentielles. On recherchera peut-être le nom d'un accusé tout au long d'un procès. Or, si les fiches de greffe indiquent bien le nom de celui qui est invité à s'exprimer à la barre, elles ne proposent évidemment ni le nom de ceux qui demandent la parole au président, ni les noms de ceux qui sont nommés dans les débats. Ces noms propres là sont à extraire du contenu sonore. Transcrire automatiquement un nom de personne demande au système de distinguer l'entité nommée des autres types de termes, d'en transcrire la phonétique et de la rapprocher d'un lexique de noms propres, qui souvent doit être établi spécifiquement en fonction du corpus. Dans le cas présent, la situation se complique d'un grand nombre de noms à consonance arabe, qui sont prononcés de façon très diverse selon que le locuteur est arabophone ou francophone.

24 Si l'on souhaite que le texte puisse aussi être parcouru, une relecture humaine s'avère nécessaire, car l'algorithme produit un certain nombre de transcriptions fautives, variables selon la qualité du son et les caractéristiques de la parole du locuteur. Certains de ces faux sont aisément détectables à la lecture parce qu'ils produisent des absurdités. D'autres échappent à la relecture parce qu'ils seraient sémantiquement corrects tout en trahissant le sens original. C'est le cas par exemple des erreurs liées au genre ou lorsqu'une négation échappe à l'algorithme, ce qui produit une phrase dont le sens est trompeur. Pour repérer ces erreurs, la relecture doit se faire en écoutant le son. Comme il n'est pas envisageable de proposer à un public de chercheurs des archives historiques transcrites automatiquement sans les relire, il faut prendre en compte le temps du traitement automatique de la relecture. On ne peut alors envisager de disposer de cette transcription avant une année[7].

25 Pour toutes ces raisons, et surtout pour disposer de métadonnées dès l'issue du procès, nous avons décidé d'écarter définitivement la solution de la transcription automatique, et d'opter pour une annotation manuelle « à la volée ».

Le dispositif d'annotation manuelle avec Annotate-on

26 Il ne s'agit pas ici, contrairement à la solution de transcription automatique analysée précédemment, de travailler à partir de l'enregistrement, mais à partir de l'événement lui-même pendant qu'il se déroule[8]. On analyse donc l'enregistrement des archives en

train de se faire, tout en ayant une perméabilité plus ou moins grande aux émotions des personnes qui entourent la personne gérant le logiciel d'annotation. C'est un avantage par rapport à la transcription du son qui ne prend en compte que les informations sonores captées et reste impuissante à restituer les sons trop éloignés ou faibles pour être captés, les gestes et les manifestations visuelles et plus largement l'« ambiance » de l'audience.

27 L'objectif est de définir les marques temporelles de début et de fin des moments riches en information, ou lorsque les débats changent de thèmes, et d'y associer des annotations (descriptions, précisions, résumés) et des mots-clés.

28 Le dispositif nécessite, d'une part, l'intervention d'une personne compétente pour comprendre l'événement dans lequel elle est immergée et faire le choix de consigner un moment particulier. Elle doit, en outre, maîtriser les techniques d'indexation et d'analyse documentaires. Ce rôle a été endossé par l'historienne Aurore Juvenelle, pendant toute la durée du procès.

29 Un langage spécifique au procès a été développé par Martine Sin Blima-Barru au cours des mois qui l'ont précédé. Composé de 409 termes au départ et enrichi tout au long du procès, il comporte les termes précis du secteur judiciaire, ceux liés à l'action terroriste, les noms propres à l'événement (noms des lieux, des rues, des cafés et établissements concernés), ainsi que ceux des parties civiles, avocats, témoins ou membres de la cour.

30 Un micro-ordinateur, équipé du logiciel Annotate-on Event[9] complète le dispositif. Annotate est un outil ergonomique d'annotation d'images pour les sciences naturelles, qui permet de décrire et d'annoter des planches d'herbier numérisées. Il a été initialement développé, sous la direction du laboratoire Dicen-IDF du Cnam Paris dans le cadre du projet e-Recolnat du Muséum national d'histoire naturelle[10].

31 Avec le partenariat du Labex Les passés dans le présent[11], un module d'analyse chronothématique a été adjoint en 2020. Le principe est le suivant : pendant qu'un lecteur déroule une vidéo, on la découpe en séquences, identifiées par leur TCIN-TCOUT[12], et on la décrit en langage libre ou on l'indexe à l'aide de langages prédéfinis ou de tags. L'ergonomie d'Annotate-chrono permet d'annoter la vidéo pendant son déroulement sans nécessairement arrêter le défilement. Avec sa gamme de fonctionnalités permettant de décrire et annoter des collections, des photographies et des vidéos, ainsi que des fragments de photographies et des séquences de vidéos, Annotate-on Chrono est un logiciel libre, particulièrement adapté au traitement des archives orales et des collectes patrimoniales[13].

32 De mai à juillet 2021, nous avons fait adapter le module Chrono aux besoins du procès, c'est-à-dire à l'analyse de vidéos, qui n'existent pas encore, mais qui sont en cours d'enregistrement parallèlement au travail d'analyse :

- le lecteur de vidéos (inutile dans ce cas précis) a été supprimé et remplacé par un chronomètre synchronisé sur l'horloge de l'ordinateur ;
- la notion de séquence a été redéfinie. Il ne s'agit plus d'une portion de vidéo comprise entre un TCIN et un TCOUT, mais d'un intervalle temporel défini à la seconde précise par son début et sa fin. Cet intervalle est l'unité décrite et indexée ;
- la fonction de synchronisation entre les descriptions d'intervalles et les vidéos captées est un élément essentiel, car c'est celui qui permet, une fois les vidéos produites et chargées dans Annotate-on Event, de naviguer des métadonnées descriptives aux séquences vidéo.

> Nous avons vérifié au préalable que les caméras de la salle d'audience étaient bien réglées sur le même fuseau horaire que le micro. En théorie, les informations de début et de fin de chaque captation, inscrites automatiquement par la caméra dans les métadonnées IPTC[14] du fichier vidéo, doivent permettre de calculer les bornes temporelles correspondant aux intervalles d'événements décrits dans le logiciel.

Des contraintes liées à la sécurité et à la confidentialité

33 Avant d'aborder plus précisément le rôle des Archives nationales dans la production des points d'accès aux contenus audiovisuels, rappelons les conditions dans lesquelles nous avons construit ce dispositif, car cela a pesé sur son travail.

34 Le délai de développement d'Annotate-on Event a été très court, puisque les premières réunions d'identification de la problématique ont eu lieu en avril 2021. Le cahier des charges du logiciel a été remis au fournisseur en mai, le développement s'est déroulé en juin et juillet et la première version a été livrée et testée mi-août, le procès commençant officiellement le 9 septembre. Comme tout logiciel, Annotate-on Event comportait quelques bugs qui n'ont été décelés qu'à l'usage dans les premiers jours du procès.

35 Pour des raisons de sécurité et de confidentialité, la salle d'audience n'était pas reliée au réseau Internet et il n'était pas possible de téléphoner pendant les audiences. Le seul moyen de communication pour assister Aurore était le SMS, ce qui est très limité pour expliquer un bug ou un « plantage ». L'absence de réseau ne permettait pas non plus de partager l'accès à l'application, qui n'a pu être installée que sur le disque dur d'un ordinateur portable. Cela ne permettait pas à Claire Scopsi, qui n'était pas autorisée à accéder à l'enceinte du tribunal, d'assurer un appui technique en vérifiant le bon fonctionnement de l'application et la qualité des données produites. De ce fait, les premiers jours d'annotation ont été assez inégalement réalisés.

36 Pour les mêmes raisons de confidentialité, aucun échantillon des vidéos captées n'a été communiqué au cours du procès. L'ensemble des vidéos a été remis en une fois aux Archives nationales à l'issue du procès. Le projet s'est donc révélé un long tunnel de dix mois, pendant lesquels des annotations étaient effectuées tous les jours, sans certitude que le dispositif produise le résultat escompté.

Les difficultés de l'annotation à la volée

37 Si le procès a été vécu comme une mystérieuse « boîte noire » par les membres de l'équipe non accrédités, il a aussi été une aventure humaine pour les deux archivistes, les chercheurs et tous ceux et celles qui y ont assisté intégralement.

38 Aurore Juvenelle se définit elle-même comme « une cheville ouvrière de transmission du sens [...] chargée de faire entrer ce procès extraordinaire dans des cases de tableur Excel ou dans des cases du logiciel Annotate[15] » (2022). Les Archives nationales doivent fournir au futur chercheur une indexation, le référencement des contenus des captations ainsi que les conditions de la pérennisation sur le temps long des fichiers numériques. Il s'agit de passer au crible l'événement, tout en en faisant une synthèse, sans toutefois réaliser un *verbatim*. Et c'est bien là que réside une des difficultés de

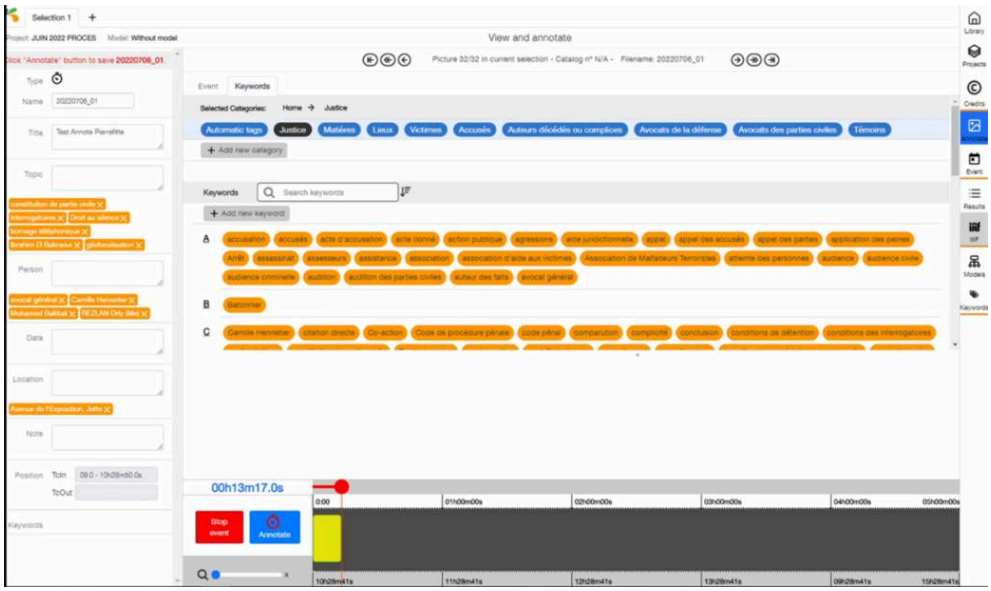
l'annotation en direct : saisir le sens de ce qui se dit, le synthétiser et le reformuler dans un temps très court.

Structurer et contrôler la description à la volée

39 Une deuxième difficulté consiste à structurer et contrôler la description au fur et à mesure de l'inscription. Deux outils sont utilisés au cours du procès :

1. un tableau Excel, dans un premier temps. Mais il s'est avéré trop peu ergonomique pour une saisie de texte. Il a été rapidement remplacé par un fichier Word, dont le contenu est synthétisé dans le tableau Excel à la fin de chaque journée. Cette étape quotidienne de réécriture est encore l'occasion de synthétiser davantage. On y inscrit dans des colonnes séparées ce que l'on voit sur les écrans et ce que l'on entend ;
2. le logiciel Annotate-on qui propose de sélectionner des termes parmi différentes catégories : justice, matière, lieu, victime, accusé, etc. On y crée un événement qui correspond à une session de l'audience, puis on crée une séquence, pour distinguer un moment remarquable, et, au fil de cette séquence, on clique sur des termes pour les inscrire dans un champ de la notice. Cette tâche d'indexation dure jusqu'à la suspension d'audience, qui marque la fin de l'événement dans Annotate. Dans l'exemple ci-dessous, le champ topic désigne ce dont on parle pendant la séquence de procès. On indique aussi qui est le locuteur (ici l'avocat général qui s'oppose à une constitution de partie civile).

Figure 1. Copie d'écran d'Annotate-on Event.



À gauche l'affichage de la notice en cours de création, à droite, l'affichage des termes du langage de la catégorie « justice ».

40 Une grande partie des termes provient du langage conçu par Martine Sin Blima-Barru. La catégorie « justice » de l'exemple concerne les termes liés aux composantes du procès et à son déroulement : accusation, acte d'accusation, constitution de partie civile, audience civile, incident d'audience, ministère public, plaidoirie, suspension, etc. Tous ces éléments sont prévisibles et ont pu être intégrés avant le procès.

41 A pu être anticipé également tout ce qui figurait dans l'ordonnance de mise en accusation, comme les noms des accusés et ceux des avocats. Mais tous les avocats de

parties civiles n'y étaient pas identifiés et il a fallu les saisir dans la liste d'autorité au fil de leur prise de parole.

Gérer l'imprévu

42 Beaucoup de termes de matière ne pouvaient absolument pas être anticipés. Dans cette catégorie où l'on décrit toute la vie du procès figurent des éléments variés, par exemple la désignation d'événements particuliers, comme « attaque du Thalys », car, même si le dossier ne porte pas sur cet événement-là, l'un des accusés y était lié. Pour la seule catégorie « Matières », la liste des thèmes comptait 160 entrées à l'ouverture du procès et 270 à la fin. 110 mots « matière » ont donc été ajoutés au cours des audiences.

43 Si l'on fait un focus sur les mots commençant par A, on voit parmi ces termes ajoutés : « Al-Quaïda », car le groupe a été évoqué, alors même qu'il ne s'agit pas d'un attentat commis par Al-Quaïda. A également été ajouté le terme « Anachid » qui désigne un chant religieux et c'est un terme qui est revenu très souvent puisqu'il évoque des chants de propagande qui tiennent une place très importante dans le procès. Il a posé un problème particulier à Aurore Juvenelle, lorsqu'elle l'a entendu pour la première fois, car elle n'est spécialiste ni du dossier, ni du terrorisme, ni des chants religieux : « quand vous entendez "Anachid" la première fois, vous n'êtes pas sûre que vous avez entendu ça. Et après, vous cherchez l'orthographe et vous vous rendez compte qu'il y a plusieurs orthographes. En attendant, le procès continue, votre logiciel continue à tourner, votre synthèse n'avance pas. Aïe, aïe, aïe ! » (Juvenelle, 2022).

Faire des choix

44 Décrire un événement à la volée revient à écouter, synthétiser et reformuler tout en continuant à écouter pour synthétiser et reformuler la suite. Si le rythme des audiences permet de réaliser approximativement cette tâche sur les deux outils, la longueur du procès et la concentration que cela exige finissent par peser. La fatigue et une certaine saturation psychologique s'installent au fil du procès : « une chose que vous avez entendue il y a cinq, dix secondes, impossible de la retrouver. Impossible. J'avais vraiment beaucoup de mal avec ça » (Juvenelle, 2022).

45 Une autre question est liée à la responsabilité dont dispose Aurore de choisir ce qu'elle décrit, puisqu'elle ne transcrit pas intégralement l'événement. Il lui faut très rapidement identifier si un terme doit être ajouté à la liste d'autorité. Certains thèmes sont abordés au cours du procès, sans qu'elle sache s'ils sont importants puisqu'elle n'est pas spécialiste du terrorisme et qu'elle ne dispose que de très peu de temps pour faire des recherches. Certains thèmes, entendus au début, n'ont pas été notés et c'est au fil des journées et des audiences qu'elle se rend compte qu'ils reviennent souvent et décide de créer une entrée. Mais le terme ne figure pas dans les séquences qui ont précédé sa création et cela sera un inconvénient pour les recherches futures.

46 D'autres termes manquent pour décrire le déroulement formel du procès, malgré le travail anticipé sur les termes professionnels de la justice. Ainsi, « Audio » s'avère un terme important, parce que, dans le Bataclan, une personne qui avait un enregistreur (certainement pour pirater le concert) a tout enregistré. Cela provoque un débat : faut-il ou non l'écouter ? Les parties civiles peuvent-elles avoir accès à l'intégralité de l'enregistrement utilisé à l'enquête ? Moins polémique, le terme « Lecture » n'a pas

paru nécessaire, au départ, pour décrire les nombreux moments où des lettres, des ordonnances, etc. sont lues. Puis, en échangeant avec d'autres personnes dans le public, Aurore se rend compte que ces procédés pourront intéresser des chercheurs. Ce dilemme : « faut-il créer ce terme au risque d'alourdir la liste de termes inutiles ou trop spécifiques ? » est le lot de tous les professionnels de l'information, mais, ici, la solitude et la nécessité de prendre une décision immédiate rendent l'exercice inconfortable.

Gérer les émotions

47 Dans ce lieu et face à ces dispositifs qui se veulent les plus rationnels et distanciés possibles, naît un sentiment de décalage difficilement surmontable pour celle qui annote. Traversée d'émotions variées, de fatigue, de besoin de bouger de son siège où elle est plantée des heures durant face à un écran enchaînant les plans fixes, elle poursuit la mission holistique, voire robotique, qu'elle s'est fixée : mettre le contenu du procès dans un logiciel d'indexation et dans des cases de tableau Excel. En découle un sentiment persistant d'irréalité. Face à l'ampleur de ce qui se passe, les tentatives d'organisation documentaires paraissent dérisoires.

48 Aurore est postée devant un écran, car elle doit décrire ce que captent les caméras, le procès se déroulant derrière elle, deux portes plus loin, dans la salle d'audience principale, alors qu'elle-même se trouve dans la salle de retransmission dédiée aux journalistes et aux chercheurs. Pendant deux semaines, des enquêteurs belges, dont les noms, par sécurité, sont remplacés par des suites de chiffres et dont les voix sont robotisées, rendent compte d'écoutes téléphoniques en énonçant pendant des heures des suites de numéros de téléphone. Que décrire de ces moments-là ? « Je me suis sentie [...]. C'était un sentiment d'irréalité et une espèce de révolte du corps et du cerveau » (Juvenelle, 2022). Il en résulte un sentiment partagé, à la fois de mission accompli et de crainte d'être passée, avec la pauvreté des outils documentaires, à côté de la dimension humaine et universelle de ce procès majeur.

> Finalement, j'ai été confrontée en permanence à deux choses très contradictoires, à savoir que j'étais dans l'obsession de synthétiser, d'élaguer le plus possible et en même temps de ne pas perdre la profondeur, le relief du procès. Je ne sais pas si on peut parler d'une frustration, mais en tout cas, ce qui est évident, c'est que même la plus fine indexation aura toujours des manques. (Juvenelle, 2022)

49 Après la fin du procès, les enregistrements des archives audiovisuelles de la Justice sont versés aux Archives nationales et une analyse sur un échantillon de données est effectuée. Commence alors une autre phase du projet permettant d'expérimenter une autre fonctionnalité du logiciel. Il s'agit d'évaluer la faisabilité d'une synchronisation entre les métadonnées temporelles enregistrées respectivement par les caméras et par Annotate-on Event. Cette synchronisation permet de récupérer les enregistrements et les métadonnées dans Annotate-on Chrono et de les manipuler ensuite comme une très classique analyse de vidéo.

50 Cette opération, un peu ralentie par les contraintes de sécurité, car les enregistrements ne peuvent être communiqués en dehors de l'enceinte des Archives nationales, a révélé une anomalie importante : les *time code* restitués par les caméras et par Annotate (ou plutôt par le micro utilisé par Annotate) ne sont pas alignés et présentent des décalages aléatoires, ce qui va compliquer le travail de synchronisation. C'est encore un inconvénient lié au « tunnel » du processus : aucun échantillon d'enregistrement n'étant livré en cours de procès, l'anomalie n'a pu être repérée et corrigée à temps.

51 Néanmoins, les métadonnées existent bel et bien et, en dépit des inquiétudes d'Aurore, elles seront utiles aux chercheurs. Elles viendront en complément des nombreuses annotations et *verbatim* produits par les équipes de journalistes et de chercheurs accrédités.

52 La suite du projet consistera à tester la récupération des métadonnées dans une application de consultation, réalisée par exemple avec le logiciel Oméka. Nous évaluerons également la possibilité d'utiliser les métadonnées produites pour optimiser l'apprentissage profond d'un logiciel de transcription automatique de la parole, puisque les noms propres ont été transcrits. Ces compléments au projet permettront de statuer sur l'avenir d'Annotate-on Event. Les chercheurs qui accèderont aux fichiers pourront évaluer, par la pratique, la pertinence de l'approche du projet d'annotation à la volée.

53 Le logiciel Annotate-on Event trouvera aussi son utilité dans d'autres contextes pour annoter en direct d'autres types d'événements comme des colloques, des cours ou des conférences, afin de les publier en ligne et de permettre d'y naviguer très rapidement.

BIBLIOGRAPHIE

Michel FOUCAULT *et al.*, « Le jeu de Michel Foucault (entretien réalisé en 1977) », dans *Dits et écrits (1954-1988), Vol. 3 (1976-1979)*, Paris, Gallimard, 1994, p. 298-329.

Romane GORCE, « La publicité des débats à l'épreuve du dispositif audiovisuel », *Politika,* [septembre 2022], consulté le 30 décembre 2022, à l'adresse https://www.politika.io/fr/article/publicite-debats-a-lepreuve-du-dispositif-audiovisuel

Emmanuel-Pierre GUITTET, Antoine MÉGIE et Sharon WEILL, « Ce que la "guerre au terrorisme" fait à la justice », *Cultures & Conflits*, 123-124(3-4), 2021, p. 95-103. https://doi.org/10.4000/conflits.23280

Aurore JUVENELLE, **«** Indexer les procès filmés. Le cas du procès des attentats du 13 novembre 2015 », dans *L'archivage des procès filmés : de la captation à l'accès aux images,* Séminaire Les nouveaux paradigmes de l'archive, 6 juillet 2022, consulté 30 décembre 2022, à l'adresse https://nparchive.hypotheses.org/971

Jean-Michel SALAÜN, « La redocumentarisation, un défi pour les sciences de l'information », *Études de communication*, 30, 2007, p. 13-23. https://doi.org/10.4000/edc.428

Martine SIN BLIMA-BARRU et Christian DELAGE, « Filmer les procès pour l'histoire . La fabrique d'une archive de la justice », *Les Cahiers de la Justice*, 2(2), 2021, p. 297-308. https://doi.org/10.3917/cdlj.2102.0297

NOTES

1. « Dès l'ouverture de l'audience des juridictions administratives ou judiciaires, l'emploi de tout appareil permettant d'enregistrer, de fixer ou de transmettre la parole ou l'image est interdit. Le président fait procéder à la saisie de tout appareil et du support de la parole ou de l'image utilisés en violation de cette interdiction. » Loi du 29 juillet 1881 sur la liberté de la presse, article 38 *ter*.

2. Ces chiffres sont ceux avancés par Radio France, en novembre 2022, dans le podcast « Retour sur le procès des attentats du 13 novembre 2015 » (série « Esprit de justice »), en ligne (https://www.radiofrance.fr/franceculture/podcasts/esprit-de-justice/retour-sur-le-proces-des-attentats-du-13-novembre-9343491, accédé le 30 décembre 2022). Les chiffres varient selon les sources. Le nombre de parties civiles et d'avocats a notamment évolué au cours du procès.

3. Nous nous appuyons, pour cette description, sur l'intervention de Romane Gorce au séminaire « Les Nouveaux Paradigmes de l'Archive » du 6 juillet 2022, en ligne (https://nparchive.hypotheses.org/971) consulté le 31 décembre 2022, ainsi que sur l'article qu'elle a produit à l'issue du procès (Gorce, 2022). Elle y décrit précisément les lieux et l'ensemble des dispositifs audiovisuels. Elle a consacré son mémoire de master d'histoire du cinéma à la mise en récit des procès historiques du terrorisme par la médiation du filmage d'archives et a suivi l'ensemble du procès.

4. Sylvie Lindeperg est professeure à l'université Paris 1. Elle est directrice du Cerhec (Centre d'études et de recherches en histoire et esthétique du cinéma). https://univ-droit.fr/actualites-de-la-recherche/manifestations/46015-le-proces-v13-vu-par-les-sciences-sociales

5. Sylvie Lindeperg, « Procès V13 : À l'épreuve des images ? », *Terrain Social*. Entretien audio, 6 juin 2022, (24 minutes). https://podcast.ausha.co/terrain-social/67

6. Cette estimation résulte d'une simple règle de trois : 9 mois × 20 jours × 7 heures × 8 caméras = 10 080 heures d'enregistrement pour 1 260 heures de procès environ. Notre évaluation était très généreuse, puisque le procès a duré 148 jours, soit environ 700 heures. Le volume final d'enregistrements numériques n'est donc « que » de l'ordre de 4 000 à 5 000 heures.

7. Si l'on considère une situation très favorable, c'est-à-dire un traitement automatique qui traite une heure d'enregistrement en une heure, et une relecture du résultat en écoutant, soit une heure de correction pour une heure d'enregistrement transcrit, on obtient 700 heures (durée des enregistrements) × 2 = 1 400 heures soit un peu plus de 11 mois, sans prendre en compte le temps d'« apprentissage machine ».

8. Le procédé rappelle, avec quelques différences, l'indexation en direct des flux audiovisuels pratiquée par les documentalistes de l'Institut national de l'audiovisuel dans les années 1980, soit depuis les plateaux de journaux télévisés, soit devant un téléviseur. Anna Tible l'évoque rapidement dans Anna Tible, « Mobilisées contre les discriminations salariales de genre. Les documentalistes de l'INA en grève (novembre 1981) », *Le Temps des médias*, 34(1), p. 73-88, 2020.

9. Précisons que le Palais de justice n'autorisait pas l'accès au réseau Internet depuis les salles d'audience. Le logiciel doit donc être installé sur un micro et utilisé en monoposte.

10. https://www.recolnat.org/fr/annotate.

11. Dans le cadre du projet OPAHH [Open Pictures Annotator for Humanities and Heritage], porté par le Dicen-IDF du Cnam Paris et les Archives nationales.

12. C'est-à-dire le *time code* d'entrée et de fin de séquence.

13. Précisons qu'Annotate-on Chrono est une brique logicielle intermédiaire dont les fonctions de consultation et recherche sont limitées au strict nécessaire pour produire des annotations et des analyses dans le cadre d'un projet. Les métadonnées produites sont ensuite exportées dans des formats ouverts pour être chargées dans des logiciels documentaires orientés vers la valorisation des collections.

14. Métadonnées définies par l'International Press Telecommunications Council. Produites automatiquement par l'appareil de captation, elles sont incluses dans le format du fichier audio ou vidéo.
15. D'après l'intervention d'Aurore Juvenelle au séminaire Les Nouveaux Paradigmes de l'Archive, séance 2/2022 « L'archivage des procès filmés : de la captation à l'accès aux images », 6 juillet 2022, https://nparchive.hypotheses.org/971

AUTEURS

CLAIRE SCOPSI

Maître de conférences en sciences de l'information, chercheuse au Dicen-IDF, Cnam/Paris.

MARTINE SIN BLIMA-BARRU

Conservatrice du patrimoine, responsable du département de l'Administration des données des Archives nationales.

AURORE JUVENELLE

Chargée de mission aux Archives nationales pour les archives audiovisuelles de la Justice du procès des attentats terroristes du 13 novembre 2015.

Partie 3 - Les données de la recherche

La recherche ouverte et les données en Lettres, Sciences humaines et sociales (LSHS)

Le cas des GLAM

Gérald Kembellec et Claire Scopsi

Introduction

1 Depuis 2018, la « recherche ouverte » est une des priorités du ministère de l'Enseignement supérieur et de la Recherche français [MESR]. Il s'agit d'un ensemble de mesures progressives impactant les acteurs de la recherche française : chercheurs, personnels des universités et des institutions scientifiques, éditeurs scientifiques. L'objectif est de doter le pays d'une politique cohérente et systémique conduisant à apporter plus de transparence à la recherche financée grâce à des fonds publics et à en restituer largement les résultats. Deux plans triennaux se sont succédé, le second s'achèvera en 2024[1]. De 2018 à 2021, l'accent a été porté sur l'ouverture des publications et notamment l'accès ouvert aux revues scientifiques[2], ainsi que l'élaboration de plans de gestion de données dès la conception des projets de recherche. De 2021 à 2024, les mesures visent l'ouverture des codes source produits dans le cadre de la recherche publique et le partage des données *via* des plateformes dédiées et ouvertes.

2 Nous revenons dans cet article sur ces mesures en montrant de quelle façon elles impliquent les chercheurs, influencent la composition des équipes projet et renouvellent les modalités de publication. Dans une première partie, nous présentons les directives, les instances et les ressources qui constituent l'écosystème de la politique publique de l'ouverture des données de la recherche. Nous regardons également la manière dont les fondamentaux de ces politiques, la conservation et le partage, rencontrent les conditions actuelles d'exercice du métier de chercheur en sciences humaines et sociales, invité à se rapprocher des méthodes des sciences « dures ». Nous constatons qu'elles impliquent inévitablement d'intégrer la dimension pluridisciplinaire dans les projets. Dans la partie suivante, nous nous appuyons sur

l'exemple des GLAM[3], pour montrer que les experts de la discipline, qui travaillent sur des corpus numériques, numérisés ou transformés en données, n'abandonnent pas les méthodes et les bonnes pratiques traditionnelles, mais doivent les prolonger dans l'univers numérique. Cependant, la modélisation des objets de recherche, qui est au centre des humanités numériques, les conduit à s'appuyer sur l'expertise des informaticiens et les archivistes sont des alliés précieux lorsqu'il s'agit d'anticiper le cycle de vie des données par la rédaction d'un plan de gestion des données [PGD]. La troisième partie présente les solutions disponibles pour la publication et le partage des données et de leur documentation.

L'ouverture des données de la recherche

3 Les mesures exposées dans le second plan national pour la science ouverte ont pour objectif d'adopter les principes structurants de la recherche ouverte pour toutes les productions de la recherche publique, c'est-à-dire les mémoires, publications, données, logiciels, etc. Toutes les productions émanant « d'une activité de recherche financée au moins pour moitié par des dotations de l'État, des collectivités territoriales ou des établissements publics, par des subventions d'agences de financement nationales ou par des fonds de l'Union européenne[4] » doivent respecter cette politique et les règles de partages qui y sont associées. Rappelons que ces principes, désignés par l'acronyme FAIR, sont « faciles à trouver, accessibles, interopérables et réutilisables ».

Une vision politique

4 En application de la loi pour une république numérique[5] promulguée en 2016, le deuxième axe, « Structurer, partager et ouvrir les données de la recherche » (MESRI, 2021, p. 12-15), pose trois mesures concrètes : la mise en œuvre de l'obligation de diffusion des données de recherche financées sur fonds publics (mesure 4), la création de la plateforme nationale fédérée *Recherche Data Gouv* (mesure 5) et la promotion de l'adoption d'une politique de données sur l'ensemble du cycle des données de la recherche (mesure 6).

5 Ces orientations se concrétisent par la mise en place d'un « écosystème » constitué de réseaux de correspondants, plateformes et de financement de projet. Sur le thème des données, l'animation de la politique de gestion des données de la recherche repose sur de multiples initiatives :

- le Comité pour la science ouverte [CoSo] a pour mission de définir la politique de sciences ouvertes, d'en coordonner la mise en œuvre et de la développer au niveau national et international. Présidé par le directeur général de la recherche et de l'innovation [DGRI][6], il s'appuie sur un comité de pilotage, composé de représentants des établissements de l'enseignement supérieur et de la recherche et des acteurs de la science ouverte, qui se réunit deux fois par an ;
- la mise en place d'un réseau d'administrateurs au sein des établissements afin d'assurer la coordination. Des ateliers de la donnée, recrutés par appels à projets, sont mis en place dans une quinzaine d'établissements supérieurs[7]. Labellisés par le CoSO et structurés en réseau autour d'un bureau national, ils appuient en local les établissements et les chercheurs pour propager la culture de l'*open science,* mettre en place une offre de formation sur la gestion des données de la recherche ou aider à la rédaction des plans de gestion de données

désormais obligatoire dans la constitution de dossiers de réponse aux appels à projets nationaux et européens ;

- des plateformes de partage de connaissances et de fédération des ressources sur l'*open science* émanant de réseaux de partenaires, d'établissements ou d'associations prolongent les actions et assurent l'information et l'initiation aux techniques d'ouverture des données ;
- la visibilité des données produites est assurée par un entrepôt pluridisciplinaire, *Recherche Data Gouv*[8], confié à l'INRAE et ouvert en 2022. Sa mission est d'héberger des jeux de données et de référencer par moissonnage ceux qui sont hébergés dans d'autres entrepôts institutionnels, disciplinaires ou internationaux de façon à fournir un catalogue complet ;
- le choix de licences libres pour les données et code sources. En matière informatique, la reproductibilité des résultats de la recherche demande non seulement la mise à disposition des données, mais également celle des codes sources des logiciels qui en assurent le traitement, qu'il s'agisse de scripts, de macros Excel ou de logiciels plus élaborés (MESRI, 2022). L'ouverture des codes implique que les détenteurs des droits patrimoniaux sur le code soient identifiés et qu'ils optent pour une licence libre. Le code peut être archivé et décrit dans la plateforme *Software Héritage*[9] et référencé dans Hal ;
- le Fonds national pour la science ouverte [FNSO] est un GIS [groupement d'intérêt scientifique] créé en 2019, constitué de financements publics et privés. Il soutient les projets sélectionnés dans le cadre d'appels thématiques. Il a permis le financement de la plateforme et des ateliers de la donnée de recherche.data.gouv.fr. Un troisième appel à projet a été lancé en 2023.

6 Comme l'indique la mesure 6 du second plan, la politique de données concerne l'ensemble du cycle des données de la recherche. L'Inist[10] s'inspire du modèle de *Research Data Life* de la *UK Data Archive* pour définir ce cycle en six étapes *:*

- traitement des données (*processing data*),
- analyse des données (*analysing data*),
- conservation des données (*preserving data*),
- accès aux données (*giving access to data / data discovery*),
- réutilisation des données (*reusing data*).

7 Toutefois les appels à projet du FNSO sont encore très orientés vers la publication. On peut cependant relever quelques soutiens à des plateforme d'archives ouvertes[11].

8 Nous assistons donc à la mise en place d'un ensemble hétérogène d'initiatives, d'acteurs, d'outils méthodologiques et techniques et d'actions de support qui visent à modifier la posture et les pratiques de tous les chercheurs à l'égard des données qu'ils manipulent. À l'autre bout de la chaîne, sur le terrain même de la recherche, comment les chercheurs abordent-ils cette injonction à l'ouverture et à la conservation des données qu'ils recueillent ou produisent ? En effet, en termes pratiques, penser les données de la recherche peut s'avérer une démarche complexe.

L'impact pour les professionnels de la recherche : pourquoi apprendre à penser les données ?

9 Il est évident que les politiques publiques évoquées précédemment vont avoir un impact sur les métiers des acteurs de la recherche. Si l'on se réfère à une vision socio-anthropologique du chercheur en tant qu'individu socialisé en milieu professionnel, il

est un certain nombre de besoins et d'objectifs auxquels un acteur de la science est confronté.

10 Latour et Woolgar (1986), Bourdieu (1972, 1977) entre autres, tous repris dans une synthèse par Vigni *et al.* (2023) ont théorisé successivement – sur le temps long – la question du « cycle de crédibilité » qui présente les résultats scientifiques comme une « monnaie d'échange » sociale, mobilisée pour obtenir du crédit (social et financier) qui permet au chercheur de poursuivre son travail de recherche. D'autres auteurs, issus des milieux internationaux des *library and information sciences*, ont également participé à la réflexion. En examinant les besoins des chercheurs, ils montrent que dans le cycle du travail scientifique, collecter, analyser, partager et discuter des résultats[12] sont des étapes importantes qui doivent être prises en compte dans l'élaboration des outils et services documentaires (Palmer *et al.*, 2009, Unsworth, 2000, Nakakoji *et al.*, 2015).

11 D'un point de vue pratique, afin de faciliter la circulation des travaux de recherche, des théoriciens du Web comme Berners-Lee[13] (2004, 2010), Shotton (2009), Peroni (2017) ou Kuhn (2017) proposent des principes et des outils d'édition en ligne qui rendent les corpus facilement découvrables et compréhensibles par les machines : des dispositifs de lecture équipée, des moteurs de recherche ou des outils d'extraction et de filtrage automatique (*scraping*). Ces principes sont mis en œuvre, par exemple, dans des systèmes d'édition et d'écriture scientifique comme « Stylo »[14] de l'éditeur canadien Érudit. Ils sont également proposés dans des dispositifs comme Omeka S, inclus dans l'offre de l'infrastructure de recherche française Huma-Num. Ces outils techniques, fondés sur des technologies, des identifiants et des modèles de données standardisés, proposent de relier entre eux les éléments associés à une production scientifique comme les articles, les données, les logiciels éventuels ou les profils des chercheurs. Cette bonne pratique passe par l'observation des règles du concept de *semantic publishing* (Peroni *et al.*, 2017, Shotton, 2009) qui garantit la bonne accessibilité FAIR et donc la bonne visibilité, la diffusion et la ré-exploitation des travaux, des données et des documentations incluses. Toutefois, cette exploitation du Web de données, socle conceptuel et technique du *semantic publishing*, est complexe et requiert, autour du chercheur lui-même, la collaboration d'informaticiens (pour la bonne implémentation technique) et des archivistes[15] (pour garantir les bonnes pratiques de classement, de nommage et de préservation des corpus).

Les humanités numériques : des méthodes de plus en plus « dures »

12 Dans le domaine des humanités numériques, Pierre Mounier a clairement démontré que les évolutions techniques, la mise en données de la société, la compétition, la course aux financements transforment le métier des chercheurs et de leurs équipes dans leurs pratiques quotidiennes. Cette transformation a commencé dans les sciences dites dures, habituées à manipuler des données avec des protocoles stricts. Elle affecte désormais également les sciences humaines et sociales avec des méthodes similaires (Mounier, 2018, p. 9-19). Cette analyse était d'ailleurs partagée par Bruno Latour depuis bien longtemps :

> Vous pouvez bien établir toutes les nuances que vous voulez entre sciences *soft* et sciences *hard*, vous êtes obligé de reconnaître que les sciences sociales se constituent elles aussi par l'intermédiaire de questionnaires, de collections, de banques de données [...]. (Latour, 2001, p. 70)

13 On assiste notamment à une injonction à appliquer des méthodes issues des sciences dites dures aux disciplines des lettres, sciences humaines et sociales [LSHS] pour crédibiliser leurs méthodologies auprès des organismes financeurs des projets, car la transparence et la reproductibilité des recherches sont des gages supposés de crédibilité et de véracité (Mounier, 2018, p. 9-19). Penser la création ou la numérisation de corpus pour en faire des objets numériques manipulables fait partie de ces nouvelles méthodes, transformer des textes issus d'archives ou des œuvres d'art en bases de données également.

Modéliser un corpus : une approche multidisciplinaire

14 Cette transformation s'accompagne d'une réflexion pointue, car les informations contenues dans les corpus sont loin d'être vues de manière unanime au sein d'une même discipline, d'un courant à un autre et même d'un chercheur à un autre. Il y a donc une part importante de subjectivation dans l'appréciation de l'objet étudié. Dans le cadre d'un projet interdisciplinaire, la problématique se complexifie encore. Cela va donc avoir une importance sur la méthode de description du corpus et de ses « documents » – et, à l'instar de Suzanne Briet, nous partons du principe générique que « tout » est document, c'est-à-dire objet descriptible[16] –, c'est ce qu'on appelle la modélisation. L'enjeu de ces questions est à la fois info-documentaire (comment l'on décrit), informatique (la manière d'inscrire, de stocker, de propager et de consulter), linguistique au sens large du terme (les idées derrière les mots et les individus), tout en relevant, bien sûr, de la discipline originale liée au projet de recherche. De prime abord, cette tâche peut sembler rebutante voire insurmontable aux chercheurs des LSHS[17] qui, s'ils sont spécialistes de leur domaine, n'ont pas forcément les connaissances nécessaires pour en réaliser seuls la modélisation. Ils n'ont pas toujours non plus le temps, ni l'envie, de s'investir dans un processus d'apprentissage coûteux. Heureusement, nous verrons que cette étape peut être grandement facilitée par des collaborations interdisciplinaires et/ou par l'aide des nombreux services d'appui à la recherche, qu'ils soient nationaux ou locaux.

Implications interdisciplinaires dans les projets sur corpus numériques : l'exemple des données des GLAM

15 Les recherches liées aux disciplines historiques, à l'histoire de l'art, à la muséologie, à l'archivistique ou encore à la gestion info-documentaire sont un secteur particulièrement intéressant à considérer pour traiter du processus de mise en données et de partage des matériaux de la recherche. Les acteurs de ces disciplines se sont très tôt dotés d'outils adaptés au travail sur de corpus numériques. Rappelons que c'est le centre historique *Roy Rosenzweig*[18] qui a développé le logiciel d'édition bibliographique Zotero et Omeka, le logiciel de publication et de valorisation en ligne de collections numériques ou numérisées. Si, au premier abord, les documents d'archives qui constituent les corpus d'études des historiens semblent bien éloignés de la problématique des *data*, des projets innovants, comme Biblissima[19] (autour des manuscrits numérisés) ou Gloss-e[20] (étude de la Glose), ont su nous convaincre qu'une relation harmonieuse et fructueuse pouvait exister entre le numérique et les études historiques.

16 Quelles sont les conditions d'un dialogue interdisciplinaire autour des données en sciences humaines ? Nous proposons ici d'analyser comment chercheurs, informaticiens et archivistes peuvent associer leurs expertises au sein d'un même projet.

Questions de discipline : valider les données et leurs sources

17 Pour les historiens, comme pour les archivistes, la méthode diplomatique[21] est fondamentale. Ces disciplines se doivent de vérifier l'authenticité et la fiabilité d'un document original par une analyse rigoureuse de sa forme : c'est un élément de la crédibilité du fait historique. Mais les humanités numériques ajoutent des étapes de transformation de cet original (par exemple numérisation, océrisation, extractions de contenus) afin d'y appliquer des outils d'analyse automatique. Il devient nécessaire de s'assurer que ces opérations n'ont pas altéré la structure et le sens de la source originale. C'est la raison pour laquelle les chercheurs des disciplines historiques ne devraient s'appuyer sur des bases de données qu'à condition d'en connaître les étapes de constitution et en d'en vérifier la bonne exécution.

18 Dans l'idéal, le chercheur en humanités ne devrait pas se fier aux données stockées dans une base sans avoir consulté un fac-similé de qualité de la source originale ou même, idéalement, d'avoir compulsé la source originale. Malheureusement cela peut être difficile si ces sources sont conservées dans un lieu éloigné et peut remettre en question l'un des attraits du partage des données numériques de la recherche : donner facilement et rapidement accès à des corpus déjà assemblés.

19 C'est pourquoi différents outils et méthodes doivent être identifiés pour expliquer et garantir la qualité des opérations de constitution des bases :

1. connaître l'emplacement de la source originale, si elle existe encore, afin de pouvoir s'y reporter en toute extrémité et la compulser physiquement ;
2. disposer d'un fac-similé de qualité et pas seulement des données extraites de la source, afin de disposer des précieuses informations de contexte : ratures, disposition originelle du texte, éléments graphiques ou graphologiques, etc. N'oublions pas aussi qu'une base de données peut faire oublier les caractéristiques de l'époque et décorréler une information de son contexte ;
3. une base de données ne dispense pas de représenter visuellement les éléments dans le contexte de leur époque, lorsque l'on élabore un dispositif de filtrage et de visualisation.

20 Ainsi, dans un projet sur l'implantation des immigrés allemands recensés à Paris au milieu du XIX^e^ siècle, il était tout à fait inutile de représenter lesdits immigrés sur une carte issue de « *Google map* » : cela n'a pas de pertinence, il a donc fallu transposer les adresses d'époque en adresses modernes puis en coordonnées cartographiques en intégrant les découpages administratifs et infrastructures d'époque, sur un fond de carte de l'époque. Cela n'aurait pas été possible sans l'apport des Archives nationales, des archives de la Ville de Paris, des fonds de cartes de la BnF et des données du projet Alpage[22] (König *et al.*, 2023). Ces divers matériaux de recherche et archives historiques, numérisés patiemment par d'autres projets de recherche et d'archivistique, ont permis de faire avancer un nouveau projet. C'est là que les données numérisées forment un nouveau paradigme archivistique dans le cadre de la recherche en GLAM.

21 Transposer dans le monde numérique les pratiques et les exigences d'une discipline demande donc de la créativité et parfois une collaboration interdisciplinaire. D'autres outils méthodologiques sont à mobiliser pour faire dialoguer les disciplines autour de l'objet commun.

Élaborer un modèle

22 Comme nous l'avons évoqué, ce genre de projet se trouve obligatoirement à l'intersection de plusieurs disciplines, toutes aussi pointues et exigeantes les unes que les autres. Dans le projet « Bibliographies de critiques d'art francophones[23] », des chercheuses et chercheurs en histoire de l'art, en infocom et en informatique se sont retrouvés pour définir la notion naissante du critique d'art entre 1870 et 1950 en lien avec l'avènement de la presse écrite et des salons artistiques. La méthode incluait la prosopographie : la sociologie des acteurs de la critique, les lieux de production et d'édition des textes ainsi que leurs supports de publication, sans oublier l'impact du contexte historique (Gispert et Méneux, 2020 ; Kembellec, 2020). Ces diverses questions ont été l'objet d'intenses réflexions et d'un dialogue interactif entre historiens, historiens de l'art, archivistes et chercheurs en infocom. Six mois ont été nécessaires pour élaborer, corriger et reconstruire plusieurs fois le modèle conceptuel représentant les acteurs et les objets intervenant dans l'activité de critique ainsi que leurs interactions. Il a fallu confronter les points de vue, tant sur le fond (le fait) que sur la forme (le modèle) et, enfin, la manière de stocker les données, visualiser l'information et proposer des représentations de connaissances[24]. Cette étape de modélisation, que nous nommons « maïeutique de recherche » en hommage au dialogue socratique, pousse les différents interlocuteurs à interroger les autres pour amener chacun à expliciter le plus simplement et le plus clairement possible ses besoins et contraintes afin de « négocier » un dispositif d'accès aux connaissances compilées, produites et vérifiées par les historiens de l'art. Avant d'être mise en œuvre, cette étape de modélisation s'appuie sur des méthodes issues de l'informatique de gestion comme UML ou Merise[25], mais aussi sur la production de métadonnées descriptives et d'une connaissance approfondie de l'objet étudié. Il faut donc impérativement savoir s'entourer pour avancer dans un programme GLAM, sous peine d'avoir au final un dispositif inutilisable car techniquement mal pensé ou scientifiquement approximatif. Dans le cadre de ce dispositif, l'interface éditait dynamiquement des notices exportables vers Omeka, référençables par Zotero, téléchargeables en plusieurs formats et identifiables par les moteurs de recherche pour alimenter le Web de données[26].

Penser le cycle de vie de ses données

23 Une fois le projet de recherche terminé et les données exploitées, il convient (légalement) de rendre accessibles, non seulement les articles, livres et rapports associés, mais aussi les données brutes capitalisées comme nous l'avons expliqué en début de chapitre. Nous avons vu, dans le cadre du projet sur l'immigration allemande à Paris au XIX^e^ siècle, à quel point les données d'autres projets avaient été capitales pour sa bonne mise en œuvre. Le dépôt des nouvelles données peut être vu comme une contrainte, mais ce n'est pas du temps perdu, car il est aussi, à son tour, un don aux futurs chercheurs. C'est aussi l'occasion de faire connaître le projet grâce à la

publication de ses données qui conduit l'usager vers les travaux associés. Du point de vue de l'archiviste, il convient, autant que possible, de documenter et de transférer la qualité de preuve historique de la source à sa version numérisée, tant par la qualité de la numérisation que par la possibilité de localiser la source originale (si elle existe encore) pour vérifier la fidélité de la copie numérique. Enfin, dans le cas d'un dispositif de consultation en ligne, la granularité des métadonnées issues du modèle peut être exposée pour être comprise à la fois par les humains (visuellement) que par les machines (moteur de recherche ou Zotero). Il s'agit là de l'une des règles de base du FAIR : l'interopérabilité. En plus d'être libres d'accès et de droit, dans des formats libres, accessibles avec des outils si possible gratuits, les données doivent être décrites avec des métadonnées explicatives. Elles peuvent aussi bien s'appliquer à l'annotation textuelle avec des liens conceptuels ou bibliographiques qu'à la description de faits historiques, de modèles artistiques au moyen d'annotations d'images. Les pionniers du domaine depuis David Shotton (Shotton, 2009) ont nommé cette méthode de valorisation des archives de données ou de textes scientifiques en ligne « *semantic publishing* ». Ce concept est repris par Tobias Kuhn (2017) et le *genuine semantic publishing*, comme vecteur du FAIR, mais aussi de sérendipité et rend possible le raisonnement conceptuel automatisé, selon des méthodes et des formats techniques également consensuels élaborés dans le cadre du Web de données.

24 Rob Sanderson, directeur du département *Cultural Heritage Metadata* de l'université de Yale, œuvre beaucoup dans la documentation des archives numériques et particulièrement les images numérisées et partagées en contexte GLAM. Il insiste sur l'importance d'assurer la pérennité et de donner accès sur Internet non seulement aux archives elles-mêmes, mais aussi aux annotations qui leurs sont appliquées. Par exemple, grâce à l'*International Image Interoperability Framework* [IIIF], son institution offre de charger à la demande tout ou partie d'une image en diverses tailles et qualités dans un article culturel ou scientifique en ligne *via* une URL paramétrable de manière standardisée. Ce protocole permet également de fournir de manière standardisée toutes les métadonnées de contextualisation disponibles dans le catalogue de la base[27].

25 On l'a compris, la documentation du projet est fondamentale à la réutilisation des fonds d'archives, car il s'agit d'un point d'entrée dans les données du projet pour les générations futures. C'est la raison pour laquelle la politique d'ouverture des données impose, comme condition au financement de la recherche publique, une démarche de plan de gestion des données ou PGD qui ne se limite pas à la conservation des données.

26 Le plan de gestion de données, PGD ou *Data Management Plan* [DMP], est un outil phare de la politique d'ouverture des données. Le décret n° 2021-1572[28] stipule à l'article 6 que « les établissements publics et fondations reconnues d'utilité publique [...] veillent à la mise en œuvre par leur personnel de plans de gestion de données[29] [...] ».

27 Il s'agit d'un document rédigé au commencement d'un projet de recherche, qui doit être mis à jour tout au long du projet. Il est par exemple requis dans les six premiers mois d'un projet financé par l'Agence nationale de la recherche[30]. Puis une version mise à jour doit être transmise à mi-projet et la remise de la version définitive conditionne le dernier versement du financement. Dans les faits, la réflexion sur la gestion des données démarre en amont du projet, puisque la gestion des données est mentionnée dans la réponse à l'appel d'offre.

28 Un PGD type comporte (INRAe, 2021, p. 5) :

- une présentation du projet et/ou de la structure,

- les caractéristiques des données (nature, volume) et leurs modalités de production et de traitement,
- les métadonnées et les documentations qui les accompagnent,
- les modalités de stockage et de sécurisation,
- les informations légales les concernant : propriété des données, respect de l'éthique et du RGPD,
- les conditions d'accès et de partage,
- le plan d'archivage à long terme,
- les responsabilités et les budgets affectés à la gestion de ces données.

29 L'enjeu de la formation est d'importance : il s'agit de sensibiliser et de former tous les chercheurs à la rédaction des PGD et les initiatives de supports de formation à la rédaction des PGD fleurissent sur les sites des établissements de recherche. Nous ne citons que deux initiatives :

- la plateforme DoraNum[31], réalisée par l'Inist-CNRS et le GIS Réseau Urfist, propose depuis 2015 une centaine de ressources et d'outils de formation. Elle est associée depuis juillet 2022 à l'écosystème national *Recherche Data Gouv* en tant que centre de ressources supports pédagogiques et e-formation, afin de mutualiser les ressources pédagogiques issues des initiatives locales, par exemple des Ateliers de la donnée ;
- DMP OPIDor[32] est une plateforme d'aide à la rédaction de PGD. Issue de codes sources ouverts du *Digital Curation Centre* (Royaume-Uni) et de l'*University of California Curation Center* (États-Unis), personnalisés par des chercheurs français. Des dizaines d'exemples et de modèles de PGD élaborés par des établissements français y sont disponibles en téléchargement. Mais surtout, il est possible de créer son propre PGD à partir du modèle de son choix et de le compléter en ligne de façon collaborative.

30 Ces outils qui illustrent et appliquent les principes du FAIR, tout en contribuant à leur propagation, montrent la rapidité des mises en œuvre des PGD dans la recherche française[33]. Ce dynamisme de cette dernière est vertueux, mais, comme nous allons le montrer avec la problématique de la publication, il conduit à un foisonnement d'initiatives complémentaires, parfois difficilement lisibles par les chercheurs. Publier et partager ses données demande de faire des choix, d'identifier les partenaires et les modalités de publication les plus adaptées parmi les offres disponibles.

Les stratégies de publication et de partage en ligne des données de la recherche

31 Concrètement, l'ouverture des données est un acte de publication et de partage dont les modalités ne sont pas imposées, mais sont à choisir au cas par cas en fonction de l'environnement du projet, des types de données et des contraintes réglementaires ou techniques. La plateforme Datapartage[34] de l'Institut national de recherche pour l'agriculture, l'alimentation et l'environnement identifie quatre stratégies de partage de données.

Publier ses données dans un entrepôt

32 Un entrepôt de données de la recherche est une infrastructure publique ou privée qui offre, pour les jeux de données de la recherche, un service de publication, de

référencement, de documentation et d'accès pérenne. Le partage des données de la recherche est une opération de communication qui répond aussi à des injonctions réglementaires et institutionnelles, il faut donc bien choisir son entrepôt[35]. Souhaite-t-on être associé à une société privée ou à un partenaire institutionnel ? Dispose-t-on d'un budget pour le faire ? Quel est le volume de données à partager ? Quelle est la pérennité de l'infrastructure ? Ces questions méritent d'être débattues dans le PGD. Or le choix d'une plateforme peut être difficile, car la mise en place des entrepôts répond à des logiques multiples. Certaines sont mises en place par des éditeurs, et les données seront parfois obligatoirement associées à un article ou un livre porté au catalogue de cet éditeur. Mais ce n'est pas toujours le cas. *Data Mendeley*, entrepôt de l'éditeur scientifique néerlandais Elsevier, par exemple, accepte des données associées ou non à ses publications. D'autres plateformes sont liées à un établissement ou une institution et réservées aux chercheurs qui y sont rattachés. Il existe aussi des entrepôts par discipline comme Nakala, plateforme d'Huma-num, qui accueille des données en SHS, tandis que le CNRS a ouvert *CNRS Research Data*, entrepôt destiné aux communautés scientifiques qui ne disposent pas encore d'entrepôt thématique.

33 L'accroissement de l'écosystème de partage de données s'accompagne de développements de logiciels de valorisation *ad hoc*. Nous donnons ici l'exemple de la plateforme POUNT[36] de l'université de Strasbourg (Witz *et al.*, 2019) dont le code est librement utilisable par d'autres universités. POUNT est conçu pour sauvegarder et versionner les jeux de données, les structurer selon le standard de la discipline, configurer les droits d'accès, visualiser les fichiers de tous types (3D, vidéos, etc.) et les enrichir par des liens ou des informations. La qualité des fonctionnalités offertes par ces logiciels peut aussi peser sur le choix.

Fournir ses données sous la forme de matériel supplémentaire à la publication

34 Les revues scientifiques en ligne en sciences, techniques et médecine [STM] ont pour habitude de publier en annexes les jeux de données qui étayent l'article, en plus de leur méthode d'exploitation (protocoles de calcul et algorithmes associés). Les publications qui respectent le plan traditionnel *Introduction, Methods, Results, and Discussion* [IMRaD], particulièrement usité en science dures, restent en relation étroite avec leurs données qui garantissent la reproductibilité de la démonstration. Pour aller encore plus loin, certaines revues publient directement des articles dits « exécutables » sous la forme de carnets, par exemple les carnets Jupyter (Dombrowski, Gniady et Kloster, 2020), qui incluent de la programmation (des statistiques ou du calcul scientifique). Le résultat est visuellement un article traditionnel mais dont les résultats et les schémas sont calculés dynamiquement depuis les données. Cela permet aux évaluateurs et aux autres chercheurs de valider en toute transparence les résultats et de faire des tests avec d'autres paramètres (choix d'algorithme ou de variables exploitées, sous-ensembles de données étudiés) sans forcément être spécialistes, car ils ont l'accès aux données et au raisonnement.

35 Ces méthodes commencent à être intégrées par les LSHS, puisqu'il existe déjà une revue, le *Journal of Digital History*[37], qui ne publie que des carnets exécutables. Frédéric Clavert, l'un des coordinateurs de cette revue, présente deux échelles de lecture des sources historiennes : d'une part, le *close reading* au plus près du texte/*distant reading*

globalisant[38] et, d'autre part, une échelle lecture humaine/lecture computationnelle (Clavert, 2014). La clé de la lecture et de l'interprétation des sources de l'histoire à l'ère numérique réside dans les allers-retours constants entre *close reading* et *distant reading* et entre appréhension humaine et appréhension computationnelle des sources primaires. Cette approche dite distante permet de prétraiter un plus grand volume de données, plus rapidement et, ensuite, de mettre les résultats intermédiaires à disposition des chercheurs pour une analyse plus qualitative. Les travaux issus de ces nouvelles méthodes de travail en LSHS se doivent de fournir les données étudiées comme matériau complémentaire à l'écrit scientifique.

Publier ses données dans un *Data Paper* (article de données)

36 Les données se doivent d'être publiées, mais il est évident que les données seules, même avec leurs métadonnées, forment un ensemble trop aride pour être exploitable en l'état. Pour utiliser une métaphore triviale, c'est un peu comme proposer un meuble suédois en kit sans y adjoindre de notice d'utilisation : c'est peu utilisable. C'est là qu'intervient le concept de *data paper*. Un *data paper* est un article scientifique, généralement assez court, qui présente un ou plusieurs jeux de données et explique brièvement les objectifs du projet de recherche associé, explicite les méthodes de collecte, de numérisation – voire de calcul –, qui ont amené à la production du jeu de données publié dans l'entrepôt. Il faut bien le distinguer des articles de résultats scientifiques : le *data paper* n'est là que pour documenter les données du corpus. Ce document est indispensable à la bonne compréhension des données, des règles de nommage et apporte tous les éléments de contextualisation utiles. Comme le *data paper* est un article scientifique, il peut être cité et être pris en compte dans l'évaluation de la production scientifique de son auteur[39]. Bien que ce type d'article vienne initialement des sciences dures, il pénètre les LSHS et particulièrement les humanités numériques[40].

37 Un *data paper* en sciences humaines et sociales peut présenter le plan suivant :

- contexte et résumé : succincte description de données produites, leur contexte scientifique ainsi que leurs utilisations potentielles ;
- méthodes : description précise du processus de production des données afin que celui-ci soit reproductible ;
- fichiers de données : description de chaque jeu de données associé avec le *data paper* (variables, noms de fichiers, localisation, formats et taille) ;
- validité des données décrites : analyses ou procédures ayant permis de confirmer la validité des données décrites (confrontation avec différentes sources ou avec des données comparables) ;
- notes d'usage : procédures de réutilisation des données, licence ;
- disponibilité du code (éventuellement) : reproductibilité, un éventuel accès au code de reproduction du jeu de données.

Publier dans le Web des données

38 Outre la publication des données, il est courant en humanités numériques, et plus spécifiquement dans les GLAM, que les équipes de recherche souhaitent partager des fac-similés numériques des documents patrimoniaux (ouvrages, documents administratifs, œuvres d'art, cartes) participant au corpus. En effet, ces documents

numérisés et mis à disposition sur des dispositifs de consultation en ligne sont des sources précieuses pour les futurs projets de recherche.

39 Le Web de données repose sur le fait que des moteurs de recherche, des systèmes d'affichage ou de requête du web s'appuient sur des fichiers contenant des métadonnées structurées selon une norme de modélisation (RDF) et souvent non visibles à l'écran, car intégrées au code source du document. Cela permet de filtrer, regrouper, relier des fichiers du Web selon le sens de leurs contenus. Par exemple, l'accès à une page d'ouvrage rare numérisée pourra être réalisé au moyen de filtres de recherche tels que les lieux ou les entités nommées désambiguïsées qui y sont évoqués, le type de police d'écriture, ou même sur la présence de *marginalia* ou d'enluminures. Ces informations, issues du travail d'enrichissement effectué par les chercheurs, sont les métadonnées figurant dans les fichiers de données structurées que les techniques de documentation du Web des données vont ainsi rendre accessibles aux outils de collecte et de filtrage.

40 Ces quatre procédés n'aboutissent pas au même degré d'ouverture et ne visent pas les mêmes cibles. La « démocratisation de l'accès aux savoirs, utile à l'enseignement, à la formation, à l'économie, aux politiques publiques, aux citoyens et à la société dans son ensemble[41] », préconisée par le gouvernement, passera peut-être davantage par le Web de données que par l'accès aux *data papers* ou aux entrepôts, plus adaptés aux partages entre chercheurs. Ils requièrent aussi des degrés de technicité différents. Un chercheur peut publier un *data paper* en autonomie, les entrepôts proposent des interfaces appropriables par les chercheurs ou les archivistes, mais l'accès au Web de données demande l'aide d'un ingénieur de recherche ou d'un *data librarian*. L'interdisciplinarité reste un atout dans cette étape vers l'ouverture.

Conclusion

41 Nous avons livré dans ce texte un état des lieux, à la fin de 2023, de la politique de FAIRification des données de la recherche publique française qui concerne autant les sciences dures que le domaine des LSHS et plus particulièrement des humanités numériques dont les méthodes s'appuient sur des corpus et des traitements numériques. Pour atteindre l'objectif de partage, de conservation pérenne et de réemploi, les modèles de données doivent être clairement documentés et doivent, le plus possible, respecter des normes ouvertes de structuration. Cela requiert la collaboration de plusieurs profils d'intervenants : informaticiens, professionnels de l'information ou des archives, et spécialistes de la discipline. Même si l'époque actuelle est davantage préoccupée par la mise en place des instances, outils et formations, les enjeux de la collaboration ne doivent pas être négligés, car des collègues dont les statuts, missions et cultures professionnelles diffèrent doivent parvenir à harmoniser leurs langages, leurs méthodes et leurs objectifs au sein de l'équipe du projet. Si ce processus peut sembler rébarbatif ou chronophage, il peut néanmoins être gratifiant, car la publication des jeux de données sur des plateformes ouvertes, liée aux profils des contributeurs, permet aux acteurs des supports à la recherche de valoriser leur travail et de faire reconnaître leur expertise. Le rôle de la recherche ouverte ne se limite pas à la valorisation des données, elle promeut aussi les savoir-faire scientifiques.

BIBLIOGRAPHIE

Baromètre français de la Science Ouverte (décembre 2022), consulté le 3 août 2023, à l'adresse : https://barometredelascienceouverte.esr.gouv.fr/

Zoé ANCION, Francis ANDRE, Francis CADOREL, Romain FERET, Odile HOLOGNE, Kenneth MAUSSANG, Marine MOGUEN-TOURSEL et Véronique STOLL, « *Plan de gestion de données – Recommandations à l'ANR*, 2019, Ministère de l'enseignement supérieur et de la recherche », https://doi.org/10.52949/7

Pierre BOURDIEU, *Esquisse d'une théorie de la pratique*, Genève, Droz, 1972, 10.3917/droz.bourd.1972.01

Pierre BOURDIEU, « La production de la croyance », *Actes de la recherche en sciences sociales*, 1977, 13 (1), p. 3-43.

Suzanne BRIET, *Qu'est-ce que la documentation ?,* Paris, EDIT, 1951.

Frédéric CLAVERT, « Vers de nouveaux modes de lecture des sources », dans Olivier LE DEUFF (dir.), *Le temps des humanités digitales*, Limoges, Fyp Éditions, 2014.

Quinn DOMBROWSKI, Tassie GNIADY et David KLOSTER, « Introduction aux carnets Jupyter », traduction par François Dominic Laramée, *Programming Historian en français* 2, 2020, https://doi.org/10.46430/phfr0014

Marie GISPERT et Catherine MÉNEUX, « Bibliographies de critiques d'art francophones », *Cahiers Octave Mirbeau*, 2020, 27, p. 315-320.

James HENDLER et Tim BERNERS-LEE, « From the Semantic Web to social machines: A research challenge for AI on the World Wide Web », *Artificial intelligence*, 2010, 174 (2), p. 156-161.

INIST, « Le plan de gestion des données », dans *Une introduction à la gestion et au partage des données de la recherche,* (s.d.), consulté le 5 août 2023, à l'adresse https://www.inist.fr/wp-content/uploads/donnees/co/module_Donnees_recherche_26.html

INRAe, « Rédiger un plan de gestion de données (PGD) », OpenClass, 2021, https://ist.inrae.fr/wp-content/uploads/sites/21/2021/11/OpenClass-PGD-Octobre2021.pdf, consulté le 06/08/2023.

Gérald KEMBELLEC et Thomas BOTTINI, « Réflexions sur le fragment dans les pratiques scientifiques en ligne : entre matérialité documentaire et péricope », *20e Colloque International sur le Document Numérique : CiDE.20*, novembre 2017, Villeurbanne, France.

Gérald KEMBELLEC, « Dialogie disciplinaire en Humanités Numériques : vers une percolation épistémique et méthodologique négociée. Le cas de l'analyse des acteurs de la critique d'art (1850-1950) », *Sens public*, 2020, p. 1-31. https://doi.org/10.7202/1079443ar

Gérald KEMBELLEC et Olivier LE DEUFF, « Poétique et ingénierie des data papers », *Revue française des sciences de l'information et de la communication*, 2022, 24, ⟨10.4000/rfsic.12938⟩ ou ⟨hal-03850522⟩

Mareike KÖNIG, Gérald KEMBELLEC et Evan VIREVIALLE, « Data paper en humanités numériques : Adressbuch 1854 » 2023, preprint, à paraître dans *Publier, partager, réutiliser les données de la recherche : les data papers et leurs enjeux*, inPress. https://hal.science/hal-03947294/

G. KUCK, « Tim Berners-Lee's Semantic Web », *South African Journal of information management*, 2004, 6(1).

Tobias KUHN et Michel DUMONTIER, « Genuine semantic publishing », *Data Science*, 1(1-2), 2017, p. 139-154.

Bruno LATOUR et Steve WOOLGAR, « *La vie de laboratoire : la production des faits scientifiques* », Paris, La Découverte, 1986.

Ministère de l'Enseignement supérieur, de la Recherche et de l'Innovation, « *Deuxième Plan national pour la science ouverte 2021-2024* », consulté le 12 novembre 2023, à l'adresse : https://www.enseignementsup-recherche.gouv.fr/sites/default/files/content_migration/document/2e-plan-national-pour-la-science-ouverte-2021-2024-7794.pdf

Ministère de l'Enseignement supérieur et de la Recherche et Université de Lille, « *Science ouverte. Codes et logiciels* », 2022, https://www.ouvrirlascience.fr/wp-content/uploads/2022/10/Passeport_Codes-et-logiciels_WEB.pdf

Kumiyo NAKAKOJI, Yasuhiro YAMAMOTO, Mina AKAISHI et Koichi HORI, « Interaction design for scholarly writing: hypertext representations as a means for creative knowledge work », *The New Review of Hypermedia and Multimedia*, Special issue: Scholarly hypermedia, 2015, vol. 11, nº 1, Taylor et Francis.

Carole L. PALMER, Lauren C. TEFFEAU et Carrie M. PIRMANN, « Scholarly Information Practices in the Online Environment – Themes from the Literature and Implications for Library Service Development », rapport, OCLC Research, 2009.

Silvio PERONI, Francesco OSBORNE, Angelo DI IORIO, Andrea Giovanni NUZZOLESE, Francesco POGGI, Fabio VITALI et Enrico MOTTA, « Research Articles in Simplified HTML: a Web-first format for HTML-based scholarly articles », *PeerJ Computer Science*, 2017, https://doi.org/10.7717/peerj-cs.132

David SHOTTON, « Semantic publishing: the coming revolution in scientific journal publishing », *Learned Publishing*, 2009, vol. 22, nº 2, p. 85-94.

John UNSWORTH, « Scholarly Primitives: what methods do humanities researchers have in common, and how might our tools reflect this? », *Symposium on Humanities Computing: formal methods, experimental practice*, sponsored by King's College, London, 2000 May 1st.

Régis WITZ, Julia SESÉ, Ana SCHWARTZ, Stéphanie CHEVIRON et Vincent LUCAS, « Science Ouverte : sauvegarder, visualiser et partager vos données », *Jres, Journées réseaux de l'enseignement et de la recherche*, Dijon, 16 décembre 2019, https://conf-ng.jres.org/2019/document_revision_5259.html?download, consulté le 8 août 2023.

NOTES

1. https://www.enseignementsup-recherche.gouv.fr/fr/le-plan-national-pour-la-science-ouverte-2021-2024-vers-une-generalisation-de-la-science-ouverte-en-48525

2. 67 % des publications scientifiques françaises parues en 2021 étaient en accès ouvert en décembre 2022 (source : *Baromètre français de la Science Ouverte* (décembre 2022), consulté le 3 août 2023, à l'adresse : https://barometredelascienceouverte.esr.gouv.fr/).

3. Acronyme anglais pour *Galleries, Libraries, Archives and Museums* (en français galeries, bibliothèques, archives et musées).

4. Définition issue du Code de la recherche, article L533-4.

5. Loi n° 2016-1321 du 7 octobre 2016 pour une république numérique, https://www.legifrance.gouv.fr/jorf/id/JORFTEXT000033202746

6. En 2023, il s'agit de Claire Giry.

7. https://recherche.data.gouv.fr/fr/page/ateliers-de-la-donnee-des-services-generalistes-sur-tout-le-territoire

8. https://recherche.data.gouv.fr/fr

9. https://www.softwareheritage.org/

10. Une introduction à la gestion et au partage des données de la recherche, https://www.inist.fr/wp-content/uploads/donnees/co/module_Donnees_recherche_7.html (consulté le 5 août 2023).

11. Citons par exemple le projet arXiv qui concerne une archive ouverte de prépublications dans les domaines des mathématiques, physique, informatique, économie, et le projet *Software Heritage*, archive ouverte qui collecte, préserve et partage les codes sources de tous les logiciels publiquement disponibles.

12. Voir l'article de synthèse sur le sujet par Kembellec et Bottini (2017).

13. Le « père » du Web.

14. Stylo est un éditeur de textes scientifiques, conçu par l'équipe de la chaire de recherche du Canada sur les écritures numériques avec le soutien d'Érudit et Huma-Num. Les articles produits dans Stylo sont structurés, peuvent être annotés collaborativement et exportés en différents formats selon les plateformes ou processus de publication visés. Pour en savoir plus : https://apropos.erudit.org/stylo-un-outil-dedition-numerique-innovant-adapte-a-la-publication-savante/

15. Les structures de classement et l'indexation des corpus en ligne sont des transpositions conceptuelles de l'archivistique « physique » : plans de classement, règles de nommage des documents, normes d'encodage, accès pérennes, organisation des stockages (ici de fichiers), contextualisation et description à l'aide de référentiels.

16. En 1951, Suzanne Briet pose le principe que tout objet ou même être vivant peut devenir document s'il est intégré à un système de connaissance matérialisé par des outils d'inventaire ou de description : « Une étoile est-elle un document ? Un galet roulé par un torrent est-il un document ? Un animal vivant est-il un document ? Non. Mais sont des documents les photographies et les catalogues d'étoiles, les pierres d'un musée de minéralogie, les animaux catalogués et exposés dans un zoo » (Briet, 1951, p. 7).

17. Lettres et sciences humaines et sociales.

18. Centre Roy Rosenzweig pour l'Histoire et les Nouveaux Médias, voir https://rrchnm.org/our-story

19. Voir le projet portail éponyme qui inventorie une partie des textes et livres écrits, traduits, enluminés, collectionnés ou inventoriés de l'Antiquité au XVIII^e^ siècle : https://portail.biblissima.fr

20. Voir le projet *Glossae Scripturae Sacrae-electronicae* qui édite plus de 320 000 sentences exégétiques associées au texte de la Bible, encodées au format XML/TEI pour en permettre le filtrage par différents critères : https://gloss-e.irht.cnrs.fr/

21. Selon Marie-Anne Chabin : « La diplomatique est l'étude de l'authenticité et de la fiabilité des actes écrits au travers de leur processus d'élaboration, de leur forme (support et format, mais aussi structure et mise en page), de leur diffusion » (Marie-Anne Chabin « Diplomatique », blog *Esprit critique et grain de sel*, [s.d.], https://www.marieannechabin.fr/diplomatique/).

22. Voir projet ALPAGE : AnaLyse diachronique de l'espace urbain Parisien : approche GEomatique – Alpage (huma-num.fr), https://alpage.huma-num.fr/

23. Voir le dispositif de consultation lié au projet : https://critiquesdart.univ-paris1.fr/

24. Les éléments méthodologiques présentés dans cette partie sont résumés de manière pragmatique dans une ressource tutorielle interactive du projet Données de la recherche apprentissage numérique (Doranum) : 10.13143/7a03-1j03.

25. *Unified Modeling Language* et Merise sont deux systèmes conventionnels graphiques, utilisés pour représenter les processus et les objets que l'on intégrera à un développement informatique.

26. Le lecteur curieux pourra retrouver en bibliographie les deux exemples de projets présentés dans cette partie, « *Adressbuch der Deutschen in Paris von 1854* » (König *et al.*, 2023) et « Bibliographies de critiques d'art francophones » (Kembellec, 2020).

27. Voir le site du consortium du standard IIIF (https://iiif.io/) et les collections numérisées de la *Yale University Library* (https://collections.library.yale.edu/).

28. Décret n° 2021-1572 du 3 décembre 2021 relatif au respect des exigences de l'intégrité scientifique par les établissements publics contribuant au service public de la recherche et les fondations reconnues d'utilité publique ayant pour activité principale la recherche publique.

29. Texte complet de l'article 6 : « Les établissements publics et fondations reconnues d'utilité publique mentionnés au troisième alinéa de l'article L. 211-2 du code de la recherche définissent une politique de conservation, de communication et de réutilisation des résultats bruts des travaux scientifiques menés en son sein. À cet effet, ils veillent à la mise en œuvre par leur personnel de plans de gestion de données et contribuent aux infrastructures qui permettent la conservation, la communication et la réutilisation des données et des codes sources. » Décret nº 2021-1572 du 3 décembre 2021 relatif au respect des exigences de l'intégrité scientifique par les établissements publics contribuant au service public de la recherche et les fondations reconnues d'utilité publique ayant pour activité principale la recherche publique.

30. https://anr.fr/fr/lanr/engagements/faq-pgd/ consulté le 6 août 2023.

31. https://doranum.fr/

32. https://dmp.opidor.fr/

33. On peut consulter dans cet ouvrage le texte d'Annick Boissel et Véronique Ginouvès qui retrace les étapes d'élaboration d'un PGD autour du fonds de l'anthropologue Jean-Pierre Olivier de Sardan à la Maison méditerranéenne des sciences de l'homme.

34. https://datapartage.inrae.fr/Partager-Publier

35. Voici quelques exemples d'entrepôts de recherche gratuits ou *freemium* disponibles au moment de l'écriture de ce texte (consultés le 16 novembre 2023) :

– Mendeley Data (Elsevier), https://data.mendeley.com/, 10 GB par dataset (tous les fichiers) ;

– Harvard Dataverse, https://dataverse.harvard.edu/, 2 GB par fichier ;

– Open Science Framework, https://osf.io/, 5 GB par fichier ;

– Zenodo (CERN, OpenAIRE), https://zenodo.org/, 50 GB par fichier ;

– Science Data Bank, https://www.scidb.cn/, 8 GB par fichier.

36. Plateforme OUverte Numérique Transdisciplinaire [POUT] : https://pount.unistra.fr/

37. https://journalofdigitalhistory.org

38. Les notions de *close* et *distant reading* sont empruntées à Franco Moretti pour distinguer, sans les opposer, la lecture attentive de l'humain et l'analyse computationnelle qui peut faire émerger de nouvelles hypothèses que le spécialiste humain sera à même d'interpréter. Ces notions sont donc complémentaires.

39. Voici quelques exemples de revues internationales publiant des *data papers : Data in Brief,* revue en libre accès, coéditée par ScienceDirect et Elsevier, spécialisée dans les *data papers* dans toutes les disciplines ; *Scientific data,* revue en libre accès éditée par Nature Publishing Group, publie des *data papers* dans toutes les disciplines ; *Harvard Data Science Review,* multidisciplinaire, favorise le dialogue entre les chercheurs, les formateurs et les praticiens des données.

40. La *Revue française des sciences de l'information et de la communication* a consacré en 2022 un dossier spécial à la méthode des *data paper* (Kembellec et Le Deuff, 2022).

41. Présentation du second Plan national pour la science ouverte : https://www.enseignementsup-recherche.gouv.fr/fr/le-plan-national-pour-la-science-ouverte-2021-2024-vers-une-generalisation-de-la-science-ouverte-en-48525 (consulté le 20 novembre 2023).

AUTEURS

GÉRALD KEMBELLEC

Maître de conférences en sciences de l'information, chercheur au Dicen-IDF Cnam/Paris (EA 7339)

CLAIRE SCOPSI

Maître de conférences en sciences de l'information, chercheuse au Dicen-IDF, Cnam/Paris (EA 7339)

Infrastructures, architectures et outils des données de la recherche

Nicolas Sauret, Jean-Luc Minel, Mélanie Bunel et Stéphane Pouyllau

1 Conceptualisés sous l'acronyme FAIR[1], les principes de la Science ouverte et de l'accès libre se sont focalisés ces dernières années sur les données de la recherche. Ce focus s'inscrit dans la continuité du mouvement d'ouverture des publications de la recherche à la fin des années 1990.

2 Comme le soulignent Cousijn, Kenall et Ganley, les scientifiques considèrent les données sur lesquelles ils étayent leurs raisonnements et leurs résultats comme des objets de recherche à part entière : « *Data citation is based on the idea that the data underlying scientific findings or assertions should be treated as first-class research objects* » (2018, p. 4).

3 Sans reprendre ici l'ensemble des principes FAIR abondamment décrits dans la littérature scientifique (Wilkinson *et al.*, 2016 ; F. W. Group, 2020 ; Wittenburg et Jong, 2020), il est utile pour la suite de ce texte de rappeler l'importance des principes de citation, de découvrabilité et d'accessibilité des données de la recherche : pouvoir citer une donnée, donner les moyens de la trouver, de la découvrir et d'y accéder. Les recommandations concernant la citation des données de la recherche dans les publications académiques ont été publiées en 2014 par la communauté FORCE11 dans la *Joint Declaration of Data Citation Principles* [JDDCP] soutenue par une centaine de sociétés savantes, d'éditeurs de revues scientifiques et d'agences de financement (D.C.S. Group, 2014). FORCE11 regroupe des chercheurs, des bibliothécaires, des archivistes, des éditeurs et des bailleurs de fonds de la recherche avec l'objectif de faciliter la transition vers une meilleure création et un meilleur partage des connaissances.

4 Si l'accès ouvert et l'ouverture des données sont parfois perçus comme des injonctions pouvant heurter les pratiques des disciplines académiques, cela fait presque deux décennies que les infrastructures de recherche françaises dédiées aux disciplines des SHS œuvrent à une traduction adaptée de ces principes dans les pratiques de ces disciplines. C'est dans ce contexte que sont nées, au tournant de 2010, les grandes plateformes de diffusion, d'archivage, d'accès et de découverte des publications et des

données numérisées, puis numériques. Combinant les pratiques anciennes de construction d'instruments de recherche par les scientifiques et la conception de dispositifs d'accès à la documentation scientifique et technique, ces plateformes ont su par ailleurs tirer profit des principes du Web et du Web sémantique (Berners-Lee et Fischetti, 1999) développés dès la fin des années 1990. C'est ainsi que le Web, cet ensemble de dispositifs sociotechniques, est devenu le socle des instruments développés par les infrastructures de recherche françaises, redéfinissant les principes de la circulation des données et des publications, puis des informations (Bermès, Isaac et Poupeau, 2013).

5 Ce chapitre[2] propose un parcours dans les infrastructures de la recherche dédiées à la conservation et à la découverte des données de la recherche en prenant en considération les pratiques de dépôt et de recherche qui se sont développées autour et avec ces infrastructures. Dans un premier temps, nous tentons de saisir ces pratiques récentes à partir d'une revue de littérature présentant quelques études de cas. Dans un second temps, nous présentons les initiatives DANS, CLARIN et ISIDORE-NAKALA, trois infrastructures représentatives de la place primordiale prise par les infrastructures de la recherche dans la gestion et la réutilisation des données de la recherche. Nous développerons plus précisément les spécificités du couple ISIDORE et NAKALA pour lequel vient d'être mené un chantier opérationnel de *fairisation* de leurs données. Enfin, en guise de conclusion, nous livrons le résultat d'un travail prospectif projetant les usages avancés que pourraient offrir des infrastructures exploitant les grands volumes de données qui y sont déposées.

Les pratiques de recherche et de dépôt des données de la recherche

6 Quelles relations entretiennent les chercheurs et les chercheuses avec les infrastructures dédiées à la gestion de leurs données ? Que ce soit pour le dépôt de leurs données ou pour la recherche de données déposées, les usagers des entrepôts ont développé des pratiques spécifiques que plusieurs études ont cherché à décrire. À la lecture de ces études, nous mettrons en exergue les principales recommandations proposées par leurs auteurs. L'examen des relations entre déposants de données, consultants des données et gestionnaires d'infrastructures de dépôt de données de la recherche nous permettra d'insister sur l'importance d'une politique de médiation, de sa définition et de sa mise en œuvre par les gestionnaires des infrastructures.

Les pratiques de dépôt pour le partage des données de la recherche

7 Les études sur les pratiques de dépôt de recherche sont relativement peu nombreuses, en raison de leur coût (temps d'analyse) et du faible taux de réponses des utilisateurs aux questionnaires ou aux demandes d'entretien. Néanmoins, quelques études de cas permettent de comprendre le comportement des chercheurs et des chercheuses vis-à-vis du partage de leurs données de recherche.

8 L'étude menée en 2018 par DANS[3] (Borgman, Darch et Golshan, 2018) présente les résultats de 9 entretiens réalisés avec des déposants. Ces échanges ont permis d'obtenir

quelques réponses sur les motivations des déposants interviewés pour partager leurs données de recherche :

- pour préserver les données dans un temps long, c'est-à-dire au-delà de la carrière professionnelle du déposant ;
- pour répondre à l'exigence de l'agence ou de l'institution qui finance le projet du déposant ;
- pour permettre à d'autres chercheurs d'exploiter leurs données.

9 L'étude met aussi en lumière les pratiques parfois inconsistantes des déposants. Par exemple, alors que la plateforme EASY gérée par DANS assigne un DOI[4] aux données déposées, les déposants ne mentionnent pas systématiquement cet identifiant dans leurs publications.

10 L'étude menée en 2017 par l'université Rennes 2 (Serres *et al.*, 2017) porte sur l'analyse des pratiques, des besoins et des attentes des chercheurs et des chercheuses des unités de recherche en sciences humaines et sociales [SHS] de l'université Rennes 2 en termes de stockage, de partage et de diffusion des données de la recherche. Cette étude approfondie, fondée sur des analyses quantitatives et qualitatives, met en exergue de nombreux points intéressants qui incitent à une réflexion globale sur les systèmes de dépôt et de partage de données, non seulement d'un point de vue technique, mais aussi épistémologique et sociologique. Contrairement à l'étude de cas précédente, les répondants n'évoquent pas comme motivation l'exigence du financeur dans le partage des données, mais mettent en avant l'indépendance des chercheurs. Par ailleurs, si l'idée de l'exploitation de leurs données par d'autres chercheurs leur paraît séduisante et altruiste sur un plan philosophique, en accord avec les principes de l'accès libre, des barrières psychologiques viennent cependant neutraliser cette motivation. En effet, l'hypothèse que les données produites dans un certain contexte et dans un but bien précis peuvent être réutilisables dans d'autres situations de recherche, voire d'autres disciplines, n'apparaît pas comme une évidence pour les chercheurs. Ce scepticisme influe énormément sur leur capacité ou leur volonté de partager leurs données. Le partage de la donnée doit alors se faire dans un contexte sécurisé, avec un contrôle sur les modalités du partage (stockage et archivage maîtrisé) afin d'assurer une réutilisation tenant compte, d'une part, de la propriété intellectuelle et, d'autre part, du contexte de production de cette donnée, en particulier l'aspect disciplinaire et le type de données concernées (qualitatives ou quantitatives). Finalement, la reconnaissance et la valorisation scientifique ne semblent pas être des facteurs de motivation dans le partage de la donnée, dans la mesure où ils sont plus efficacement couverts par la publication scientifique (articles, monographies).

11 La troisième étude de cas, réalisée en 2016 (Kim et Stanton, 2016), étudie les facteurs institutionnels et individuels qui influencent les comportements des scientifiques en matière de partage de données dans différentes disciplines scientifiques. Les auteurs s'appuient, d'une part, sur le modèle de la théorie néo-institutionnelle (Scott, 2001) et, d'autre part, sur la théorie de l'action planifiée (Ajzen, 1991). Le modèle de la théorie néo-institutionnelle identifie trois types d'influence : la contrainte régulatrice, la pression normative et la pression cognitivo-culturelle. La théorie de l'action planifiée explique le comportement d'un individu en fonction de ses intentions comportementales qui sont à leur tour influencées par son attitude à l'égard de la perception des normes subjectives. Sur la base de ces deux théories, Kim et Stanton proposent un modèle de recherche pour expliquer et prédire les comportements des scientifiques en matière de partage de données. Ils identifient deux groupes de facteurs

d'influence des comportements : les facteurs institutionnels et les facteurs individuels. Le modèle et les hypothèses développés ont été validés empiriquement en utilisant des données d'enquêtes recueillies auprès d'un panel de scientifiques appartenant à 53 disciplines (l'échantillon final comprenait 1 317 scientifiques). Les auteurs concluent leur étude par une série de recommandations :

- mettre en œuvre des politiques strictes de partage des données par les agences de financement et les revues ;
- promouvoir des normes communautaires de partage des données en s'appuyant sur les associations professionnelles ;
- développer un système d'incitation pour fournir des crédits pour le partage des données ;
- réduire les efforts nécessités par la mise en œuvre du partage des données en standardisant les protocoles de dépôts ;
- faciliter l'altruisme scientifique individuel des scientifiques en promouvant une culture altruiste de partage des données dans la communauté scientifique.

Les pratiques de recherche des données de la recherche

12 L'étude de Gregory *et al.* (2019) vise à identifier les points communs dans la façon dont les utilisateurs issus de cinq communautés de recherche (astronomie, sciences de la terre et de l'environnement, biomédecine, fouilles archéologiques, sciences sociales) recherchent et évaluent les données de recherche. Les auteurs ont collecté puis analysé la littérature sur la recherche de littérature scientifique et de données de la recherche. Cette littérature ne provenant que de la base Scopus[5], seules certaines disciplines des sciences humaines et sociales sont représentées[6]. Néanmoins, l'analyse des 400 articles de recherche collectés apporte des résultats pertinents pour notre réflexion.

13 Les auteurs notent tout d'abord que la recherche d'information et de données se fonde sur un processus identique en trois étapes : 1°) besoins utilisateurs, 2°) actions de l'utilisateur et 3°) évaluation. Ils insistent cependant sur les différences de pratiques, toutes disciplines confondues, entre recherche de publications (*Information Retrieval*) et recherche de données de la recherche.

14 Dans une seconde étude, Gregory, Cousijn et Groth (2019) articulent une analyse bibliométrique de la littérature scientifique consacrée aux pratiques de recherche avec des entretiens d'utilisateurs de la plateforme DataSearch (22 participants installés dans 12 pays) développée par Elsevier. Les auteurs présentent plusieurs résultats importants. En premier lieu, le rapport insiste sur l'importance des interactions entre un utilisateur et la communauté scientifique avec laquelle il entretient des liens. En second lieu, l'étude insiste sur l'importance d'un moteur de recherche offrant des fonctionnalités combinant différents filtres de recherche de publications et l'accès aux données de recherche associées ou citées par ces publications. En troisième lieu, le processus de recherche des données de la recherche est considéré comme un puissant levier pour mettre en œuvre des collaborations interdisciplinaires qui exploiteront au mieux les données partagées.

15 S'appuyant sur ces résultats, le rapport propose plusieurs recommandations concernant les fonctionnalités importantes que devraient offrir ces dispositifs sociotechniques :

- standardiser les métadonnées qui décrivent les données de la recherche ;
- incorporer des techniques d'enrichissement des métadonnées ;

- développer des fonctionnalités qui permettent de stimuler des collaborations autour des données ;
- développer des API[7] qui permettent d'automatiser certaines recherches dans le dépôt de données ;
- développer des outils d'interface de représentation visuelle qui permettent d'appréhender les dépôts de données selon différents points de vue (Börner et Record, 2017 ; Scharnhorst, 2015).

16 Nous ajoutons à ces recommandations l'idée de favoriser ces intersections sociotechniques en conservant et en explicitant le contexte de création et de dépôt des données, mais aussi en inférant des passerelles pertinentes entre jeux de données.

Les relations entre déposants, consultants et gestionnaires des plateformes

17 Dans leur étude, Borgman, Darch et Golshan (2018) examinent les rôles et les relations entre des déposants de données, des consultants des données et des archivistes de la plateforme DANS/EASY. Les auteurs insistent sur l'importance du rôle de médiation scientifique et technique joué par les archivistes. Ces derniers se font en effet médiateurs du libre accès aux données de plusieurs manières. L'une d'elles consiste à fournir l'infrastructure – humaine, technique et institutionnelle – facilitant le dépôt, la récupération et la gestion des données. Ils régissent les règles d'échanges entre les déposants et les consultants. Par exemple, alors que le dépôt avec des licences *Creative Commons* réduirait au minimum la médiation requise, ce modèle limiterait la capacité du DANS à acquérir des données auprès de chercheurs et chercheuses universitaires. Cette communauté a en effet exprimé qu'elle soumettrait plus volontiers des données si elle pouvait garder le contrôle sur les personnes qui y auront accès. Le verrouillage des ensembles de données oblige les consultants potentiels à s'inscrire auprès du DANS, à renseigner leur nom et à contacter directement les déposants pour demander l'accès. Le processus de demande d'accès crée un canal secondaire permettant aux déposants et aux consultants de négocier l'accès aux ensembles de données. Dans le meilleur des cas, une conversation fructueuse conduit à un partage sélectif des ensembles de données appropriés et potentiellement à une collaboration. Les données pouvant être difficiles à interpréter en dehors de leur contexte d'origine, ces relations personnelles entre déposants et consultants peuvent s'avérer essentielles à la réutilisation des données.

18 Les auteurs notent que les modèles d'utilisation des plateformes présentent les mêmes caractéristiques que les distributions de type « longue traîne » identifiées dans d'autres études sur le comportement des utilisateurs dans la recherche d'informations (Case, 2006), c'est-à-dire avec un nombre limité de grands consultants ou déposants et de nombreux utilisateurs occasionnels. Cette distribution des pratiques a plusieurs conséquences.

19 Les déposants, qui soumettent un ensemble de données une ou deux fois par an, ou peut-être une fois dans leur carrière, ont besoin d'aide au moment du dépôt pour structurer et documenter leurs données. Les documentalistes chargés des plateformes doivent vérifier les métadonnées, la documentation et l'intégrité des données pour s'assurer que les données déposées répondent aux normes minimales. Sans cette assistance par des professionnels, les données se révèlent inutilisables. Néanmoins, si les normes et la classification des métadonnées peuvent assurer un certain niveau de

découverte de base, les auteurs estiment qu'il est pratiquement impossible de normaliser les formats et les vocabulaires dans une plateforme polyvalente qui couvre plusieurs disciplines. Ils en concluent que des investissements plus importants dans les métadonnées, la documentation et les outils de recherche permettraient d'améliorer la découverte, mais des compromis sont nécessaires dans ces investissements à forte intensité de main-d'œuvre.

Infrastructures de dépôt de données de la recherche

20 Un entrepôt (ou dépôt) de données[8] est « une collection de données thématiques, intégrées, non volatiles et historisées, organisées pour le support d'un processus d'aide à la décision » (Espinasse, 2021)[9]. Il existe un très grand nombre de dépôts de données de la recherche. Le catalogue Re3Data[10] référence 2 774 dépôts[11]. Il fournit un moteur de recherche et une API qui permettent de filtrer ou de parcourir le catalogue par discipline, pays, licence, etc.[12] (Buddenbohm *et al.*, 2021). De même, FAIRsharing.org[13] dénombre une liste de 1 797 dépôts[14] qu'il est possible de parcourir suivant différents critères. Chacun de ces entrepôts propose des caractéristiques spécifiques selon qu'il est géré par des institutions académiques (Harvard Dataverse, Dimonea de l'EHESS), par des institutions privées (Figshare), par des institutions disciplinaires (Pangea, CLARIN) ou multidisciplinaires (Dryad, Figshare, Mendeley, Zenodo), ou encore dédié à un seul projet (CERN Open Data Portal). Le thésaurus « Science ouverte » de l'INIST-CNRS[15] définit 7 types d'entrepôts de données ouvertes : archive ouverte, dépôt d'archive OAI[16], entrepôt agrégateur, entrepôt certifié, entrepôt disciplinaire, entrepôt institutionnel et entrepôt recommandé.

21 Il est important de souligner, à l'instar de nombreux auteurs (Borgman *et al.*, 2016 ; Karasti et Blomberg, 2017), que ces infrastructures techniques doivent être considérées comme des maillons de ce que Borgman, Darch et Golshan (2018) appellent des « infrastructures de connaissances » (*Knowledge Infrastructures)*, c'est-à-dire des « *robust networks of people, artifacts, and institutions that generate, share, and maintain specific knowledge about the human and natural worlds*[17] » (Edwards *et al.* 2006, p. 13), et non comme des boîtes noires dans lesquelles des données sont déposées puis recherchées.

22 Dans la suite du chapitre, nous présentons trois infrastructures de connaissances décrites selon l'angle institutionnel et selon leur attachement à une institution locale, nationale ou européenne. Les descriptions s'appuient sur leurs sites Web, sur des rapports ou des articles publiés dans des revues scientifiques et sur le rapport du COSO[18] de 2020 intitulé « Étude comparative des services nationaux de données de recherche Facteurs de réussite » (Hugo, 2020).

Une infrastructure nationale : Digital Archiving and Networked Services [DANS]

23 Fondé en 2005 par l'Académie royale néerlandaise des arts et des sciences [KNAW] et l'Organisation néerlandaise pour la recherche scientifique [NOW], DANS[19] est chargé de la gestion des données de recherche numériques en sciences humaines et sociales provenant des organisations qui l'ont précédé. DANS offre plusieurs services :

1. Electronic Archiving System (EASY)[20] pour l'archivage à long terme ;

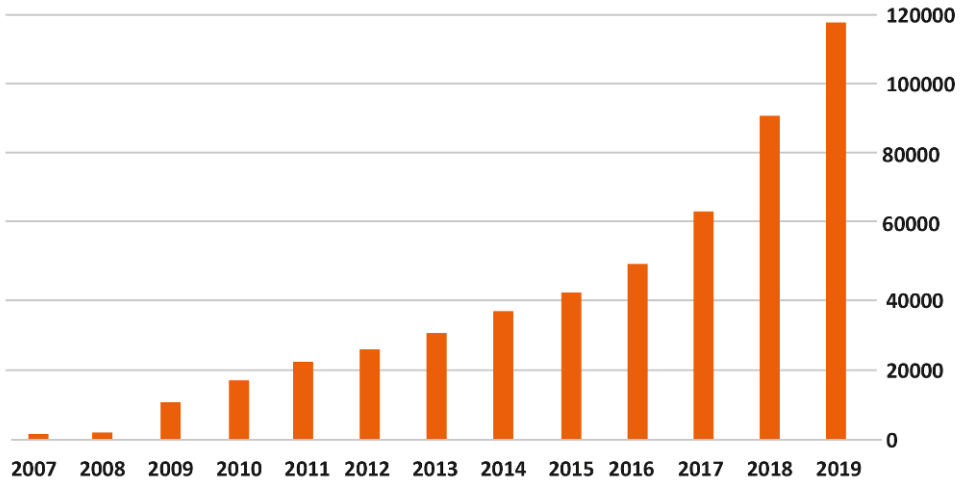
2. DataverseNL[21], un service de dépôt pour les universités, les instituts de recherche et l'enseignement supérieur ;
3. NARCIS[22], le portail national d'information sur la recherche (Doorn, 2020). DANS-EASY a reçu une certification CoreTrustSeal[23] garantissant la fiabilité des référentiels selon seize exigences. Ces exigences portent sur l'infrastructure organisationnelle, la gestion des objets numériques et la technologie sur laquelle repose DANS-EASY.

24 En 2014, DANS a créé *Research Data Netherlands*[24], une alliance visant à promouvoir les meilleures pratiques en matière de gestion et de préservation des données en partenariat avec d'autres fournisseurs néerlandais d'archives de données et d'infrastructures de recherche. DANS est par ailleurs impliqué dans de nombreux réseaux nationaux et internationaux tels que l'*European data infrastructure for scientific research* (EUDAT)[25], l'*Advanced Research Infrastructure for Archaeological Dataset Networking* (ARIADNE)[26], l'*European Open Science Cloud* (EOSC)[27] et l'*European Holocaust Research Infrastructure* (EHRI)[28].

25 DANS est organisé en trois services : « Projets et politique », « Archives et Support », « Recherche et Innovation », qui regroupent 58 personnes dont les activités sont coordonnées par un directeur. La gouvernance de DANS s'appuie sur un comité de pilotage, un comité consultatif (*advisory board*), un comité consultatif spécifique à NARCIS et un comité consultatif spécifique à DataverseNL. Le comité de pilotage de DANS supervise la gestion et le fonctionnement du réseau et des politiques menées par le directeur, ainsi que les résultats obtenus. Le comité consultatif fait part de ses recommandations en matière de stratégie et de politique générale auprès du comité de pilotage. Le comité consultatif propre à NARCIS oriente les choix de la direction de DANS au sujet du développement et du fonctionnement de NARCIS. Il est composé de représentants de sept universités et de la Bibliothèque nationale néerlandaise. Le comité consultatif spécifique à DataverseNL a pour objet de conseiller la direction de DANS sur les axes stratégiques de développement. Les 13 institutions partenaires y sont représentées.

26 En 2020, Peter K. Doorn, directeur de DANS, a publié une étude sur la montée en puissance de la plateforme EASY entre 2007 et 2019 (Doorn, 2020), dont nous reprenons ci-dessous quelques éléments.

Figure 1. Croissance du nombre de datasets dans DANS EASY, 2007-2019.



Extrait de P. K. Doorn (2020).

27 La figure 1 illustre la progression du nombre de *datasets*[29] déposés dans EASY. Après la phase de démarrage, à partir de 2012, le nombre de dépôts croît d'environ 15 à 20 % par an, puis connaît une brusque accélération avec 30 à 40 % de croissance, à partir de 2017. Doorn (2020) explique cette croissance par les conventions passées avec les universités et les institutions de recherche pour que le dépôt EASY soit utilisé comme second dépôt par ces organisations. Ces dépôts sont réalisés automatiquement sous la forme de paquets (*bulk*) échangés entre le dépôt de l'organisation et le dépôt EASY. Par ailleurs, Doorn montre que les sciences sociales représentent 30 % des dépôts et les humanités (hors archéologie) un peu moins de 10 %.

28 De ce fait, DANS répond à deux finalités :

1. la prise en charge du dépôt de données pérennes indépendamment d'une réutilisation de ces données à court ou moyen terme ;
2. la prise en charge du dépôt de données de la recherche pour répondre à des besoins de réutilisation de ces données dans l'optique de la science ouverte.

29 Doorn (2020) note une évolution importante du choix des déposants sur le type d'accès aux dépôts. Ainsi en 2012, 50 % des dépôts étaient en accès ouvert contre 70 % en 2016. L'auteur interprète cette augmentation comme un signe de la prise de conscience de l'importance d'une science ouverte.

30 Depuis 2016, DANS n'exige plus des utilisateurs un enregistrement préalable pour télécharger des données déposées en accès ouvert dans EASY. En termes d'usage, le nombre annuel d'utilisateurs enregistrés est d'environ 4 000 pour 312 472 *datasets* téléchargés entre 2007 et 2019. Ces téléchargements sont différemment répartis suivant les disciplines. Ainsi, les *datasets* en sciences sociales et dans les humanités représentent environ 4 000 téléchargements par an.

31 Dans le cadre du projet européen *Fostering Fair Data in Practices in Europe* (FAIRsFAIR)[30], le réseau DANS a collaboré avec le *Digital Curation Center* (DCC)[31] et la *Middlesex University* à la production de l'outil FAIR-Aware[32], outil d'auto-évaluation aux principes FAIR développé par DANS NL, le *Digital Curation Center* et l'université de Brême, et qui vise à sensibiliser les chercheurs et les gestionnaires de données à l'importance des principes FAIR (Hugo, 2020).

32 Les données déposées dans DANS-EASY sont accessibles *via* le moteur de recherche NARCIS. L'objectif est d'offrir aux utilisateurs les outils pour lier l'ensemble de leurs productions de manière à développer les principes d'une « *research in context* » : liens entre données, publications, chercheurs, financement et organisation.

33 DANS illustre parfaitement la notion d'infrastructure de connaissances. Cette institution interagit avec de nombreux acteurs publics et privés dans le monde entier, que ce soit des fournisseurs de moteurs de recherche, des réseaux de bibliothèques, des portails du patrimoine culturel et des sites Web qui recueillent et exploitent les données du DANS. L'infrastructure nationale se révèle ainsi être un nœud entre les infrastructures de connaissances dont elle acquiert les données et les communautés qui les consomment. Ainsi l'infrastructure déploie la technologie, définit les politiques, rédige les contrats et gère les ensembles de données qui lui sont confiés. Elle crée également des communautés en sollicitant des ensembles de données, des formations et des actions de sensibilisation. Comme ces communautés évoluent sur de longues périodes, les archives de données numériques assurent une continuité de fait articulant les différentes générations d'utilisateurs. Néanmoins, ces infrastructures sont coûteuses, nécessitent une grande quantité de temps de travail et leurs gains sont difficiles à mesurer. Leur temps de conception et de réalisation se compte en dizaines d'années et nécessite un investissement continu pour limiter une dégradation rapide (Borgman, Darch et Golshan, 2018).

Une infrastructure européenne : CLARIN

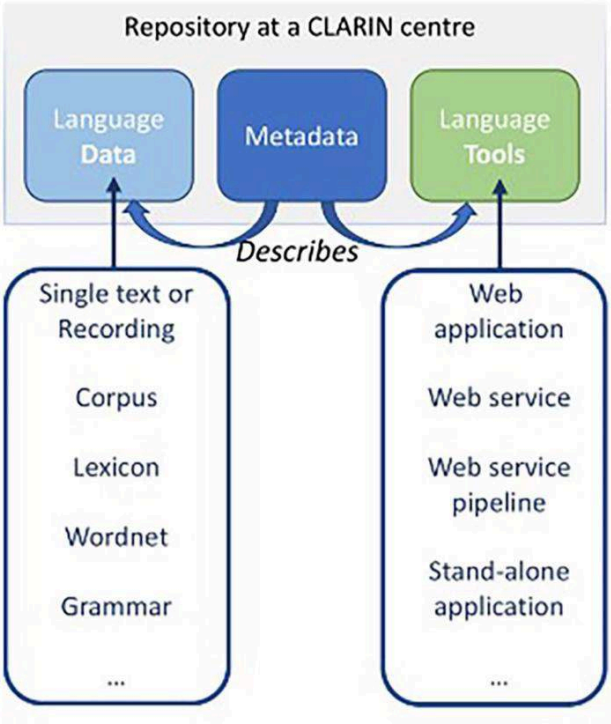
34 *Common Language Resources and Technology Infrastructure* [CLARIN][33] est un *European Research Infrastructure Consortium* [ERIC][34] créé en 2012 avec neuf membres[35] fondateurs. La tâche principale du consortium est de construire, d'exploiter, de coordonner et d'entretenir l'infrastructure de CLARIN. Il ne mène ni ne finance d'activités de recherche. CLARIN est l'une des infrastructures de recherche qui ont été sélectionnées pour la feuille de route européenne sur les infrastructures de recherche par l'ESFRI et le Forum stratégique européen sur les infrastructures de recherche. CLARIN a été créé avec le soutien financier de la Commission européenne par le biais du projet de la phase préparatoire de CLARIN (2008-2011), mais est maintenant entièrement financé par les pays participants. En 2016, CLARIN a reçu le statut de « Landmark » sur la nouvelle feuille de route. En 2017, le consortium CLARIN comprend dix-neuf pays membres et deux observateurs (dont la France) et a passé une convention avec l'université Carnegie Mellon (États-Unis). CLARIN vise une collaboration interinstitutionnelle et intersectorielle, notamment avec le secteur GLAM[36] et avec l'industrie.

35 Actuellement, CLARIN fournit un accès aux données linguistiques numériques (sous forme écrite, parlée ou multimodale) pour les chercheurs en sciences humaines et sociales. CLARIN offre également des outils avancés pour découvrir, explorer, exploiter, annoter, analyser ou combiner ces ensembles de données, où qu'ils se trouvent. Une des particularités de CLARIN est de s'appuyer sur une fédération de centres en réseau qui se distinguent selon trois types de services : des centres de dépôts de données linguistiques, des centres de services et des centres de connaissances. Les outils et les données des différents centres sont interopérables, de sorte que les collections de données peuvent être combinées et que les outils de différentes sources peuvent être articulés pour effectuer des opérations complexes (CLARIN, 2020).

36 Les centres de connaissances, au nombre de neuf en 2020[37], constituent un réseau (*Knowledge Sharing Infrastructure* [KSI]) dont une des missions est de réaliser la médiation (*the glue*) entre les infrastructures techniques et les utilisateurs. Ces centres peuvent se spécialiser dans certaines langues ou dans certaines technologies ou données. Les *workshops,* organisés par les centres et financés par le consortium CLARIN, sont considérés comme un instrument clef pour le partage des connaissances et pour le développement de nouvelles idées.

37 Les centres de dépôts peuvent se spécialiser dans une langue, une modalité (écrite ou orale), un type de données (lexicale, syntaxique, etc.) ou un type de traitement et s'engagent à être interopérables au sens où ils doivent maintenir le protocole OAI-PMH[38] pour l'échange des données. Les centres de dépôt s'engagent à respecter les principes FAIR. Le standard commun utilisé pour décrire les métadonnées est le *Component Metadata Infrastructure* [CMDI] et l'identifiant pérenne *(persistent identifier)* est un *handle* (Fig. 2).

Figure 2. Fonctionnalités générales d'un centre de dépôt CLARIN.



Extrait de Jong *et al*., 2020.

38 Les données des centres de dépôts sont moissonnées et accessibles *via* le moteur de recherche *Virtual Language Observatory* [VLO] [39] qui offre des fonctionnalités de recherche textuelles et des facettes de sélection. Néanmoins, dans leur article, les auteurs (Jong *et al*., 2020) précisent que la recherche dans un entrepôt de plus d'un million de ressources constitue un défi. Ils détaillent aussi plusieurs limitations qui tiennent aux principes mêmes de l'organisation de CLARIN. De nombreux corpus ayant été ajoutés aux dépôts nationaux ne peuvent toujours pas être identifiés dans le VLO à cause de l'absence de mots-clés ou de champs de description, ou du fait de choix idiosyncrasiques ou vernaculaires utilisés pour les dénominations. De même, des

informations sur les périodes temporelles, les annotations linguistiques et les licences d'utilisation sont absentes. L'hétérogénéité dans la granularité des *datasets* dont la taille peut varier d'un simple fichier à une archive contenant des milliers de fichiers soulève aussi des problèmes.

39 Un module de validation des métadonnées (*Curation Module*) a été développé (Ostojic, Sugimoto et Durco, 2017). Cette application contrôle un large éventail de critères (validité du schéma, présence de champs comme la langue et la disponibilité, etc.). Sur la base de ces contrôles, un score de qualité globale est calculé. Une partie très spécifique de la validation des métadonnées consiste à vérifier la validité des liens. Ces liens peuvent prendre la forme d'un identifiant de ressource de forme unique ou d'un identifiant persistant (par exemple un *handle* ou un DOI). La pratique a montré qu'environ 10 % des 5,2 millions de liens ne pouvaient être résolus par le moteur de recherche VLO. C'est pourquoi le module de curation comprend un composant spécifique qui explore régulièrement tous les liens rencontrés et stocke le résultat de l'accès à ces liens dans une base de données. Les auteurs remarquent que ce problème est générique et que plusieurs institutions développent le même type d'applications (DataCite, Europeana). Ils proposent de fédérer ces différentes bases de données dans le cadre de l'EOSC.

40 Afin de mesurer la qualité de l'infrastructure, douze indicateurs de performance *Key Performance Indicator* [KPI] ont été définis. Ces indicateurs forment un sous-ensemble des vingt indicateurs de performance définis par le groupe de travail de l'ESFRI (Report 2019).

Infrastructures FAIR : l'exemple d'ISIDORE et NAKALA

41 Le début des années 2000 a vu la production d'un très grand nombre d'études portant sur l'analyse des pratiques d'accès des chercheurs aux sources d'information et aux documents bibliographiques (publications, données de série, etc.). Dans leur étude, Ihadjadene et Chaudiron (2008) identifiaient plus d'une centaine de travaux de ce type. Dans leur conclusion, ils insistaient sur un point qui semble toujours d'actualité :

> Un moteur de recherche n'est plus simplement le « lieu » où s'apparient différentes structures cognitives dans le cadre d'interactions, mais il est considéré comme un système plus global dans lequel entrent en jeu de multiples variables : l'espace cognitif des acteurs, les caractéristiques contextuelles psychologiques, sociales et organisationnelles, ainsi que le changement des besoins d'information. Il est important d'appréhender l'usager en situation de recherche d'information de manière beaucoup plus globale que dans les modèles cognitifs et, *a fortiori*, dans l'approche système qui sous-tend encore souvent les études d'usage actuelles des moteurs de recherche (p. 29).

42 Au cours des dix dernières années, après l'arrivée de Google Scholar[40], le nombre de plateformes de recherche de documents, de données, de publications et donc d'informations a fortement augmenté (Gusenbauer 2019). L'étude menée par Know-Center (Breitfuss, Barreiros *et al.*, 2020) recense plus de 47 plateformes, incluant les plateformes privées (ResearchGate, Academia, My Science Work, etc.) ou non gouvernementales (Semantic Scholar, Dimensions, etc.). Cette étude corrobore les résultats publiés en 2016 par Lopez-Cozar, Orduna-Malea et Martin (2018), à savoir que Google Scholar est le moteur de recherche utilisé par 89 % des utilisateurs.

43 Dès la fin de la décennie 2000, la multiplication des moteurs de recherche académiques s'est doublée d'une croissance très importante de la mise à disposition de plateformes de découverte au sein des bibliothèques universitaires (Simonnot, 2012). Les *discovery tools* se sont d'ailleurs très souvent hybridés avec les moteurs de recherche académiques et les moteurs de recherche du Web (Bermès, Isaac et Poupeau, 2013 ; Gandon, Faron-Zucker et Corby, 2012). À l'échelle européenne, le développement et l'évolution de portails tels que NARCIS aux Pays-Bas, *Cultura Italia* en Italie, ou plus récemment REDIB en Espagne et en Amérique latine, ont permis aux chercheurs et chercheuses d'avoir accès de façon complémentaire et coordonnée à la littérature scientifique et aux sources de données pour les recherches dans leurs disciplines, que ce soit les publications en libre accès ou les documents sous droits, rendus accessibles *via* des API dédiées ou plus largement *via* les proxys intégrés aux portails des bibliothèques universitaires pour gérer les abonnements payants. En Europe, le développement de Driver (2006-2009) puis, depuis 2008, de la plateforme OpenAIRE[41], qui regroupe « 50 partenaires, de tous les pays de l'UE et au-delà », a offert, au cours de la décennie 2010-2019, un ensemble de « briques » pouvant être utilisées par des portails ou dispositifs de recherche de données, qu'ils soient thématiques ou disciplinaires (Manghi, Bardi et Schirrwagen, 2018).

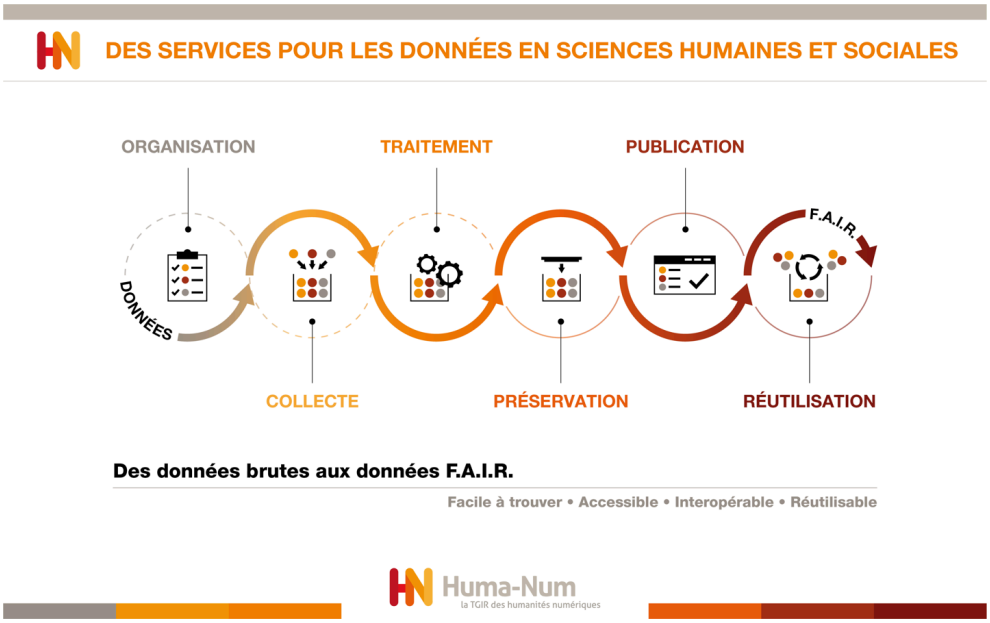
44 Dans le même temps, le développement du libre accès aux données de la recherche et aux publications (mouvement de l'*Open Access* puis de la Science ouverte) et, dans une moindre mesure, le Web sémantique (Bermès, Isaac et Poupeau, 2013), ont libéré des masses très importantes de métadonnées et de documents qui ont été intégrés à la plupart des outils de découverte sous la forme de base de données satellites. Ces dernières permettent de développer des portails associant moteurs de recherche (fondés sur l'indexation des métadonnées et du texte intégral) et outils de rebond ou d'extension par recherche fédérée vers différentes bases de données accessibles sous la forme de multiples API grâce à la proxyfication des dispositifs (Pouyllau *et al.*, 2012).

45 C'est dans ce contexte, et dès 2010, qu'ISIDORE a été imaginé et mis en œuvre (Maignien, 2011 ; Pouyllau, 2011). Au-delà de la dimension « moteur de recherche », ISIDORE s'est orienté dès le début vers l'enrichissement sémantique et la publication de métadonnées en *Linked Open Data* (Poupeau, 2016) à l'aide de référentiels scientifiques élaborés par les communautés de recherche et des bibliothèques. Et ce, dans une dimension nationale jusqu'en 2015, puis internationale avec le passage aux enrichissements multilingues à partir de 2015. L'ajout de fonctionnalités de réseau social académique dans ISIDORE est venu compléter, en 2018, un dispositif sociotechnique largement centré sur la mise en relation des savoirs avec les travaux des productions de classification et d'organisation proposées par les communautés de recherche.

46 Si NAKALA a été imaginé dans un premier temps comme un réservoir de documents et de métadonnées dans le *Linked Open Data*, sans interface Web de consultation des métadonnées et des données, ce n'est que lors de sa refonte, en 2020, que son positionnement en complémentarité de contenus avec ISIDORE a été proposé. Les deux services s'inscrivent depuis dans une interopérabilité de services (Maignien, 2011) au sein de l'écosystème d'Huma-Num. À l'image des dispositifs intégrés tels que OpenAIRE, Europeana, etc., ISIDORE et NAKALA poursuivent cette tendance en s'appuyant sur des « briques communicantes » pour faciliter leurs usages par les publics cibles. Ces briques constituent le cœur du dispositif cohérent des services mis en place par Huma-Num

pour faciliter l'accès, le signalement, la conservation et l'archivage à long terme des données de la recherche en SHS (Fig. 3).

Figure 3. Les services pour les données en SHS de Huma-Num.



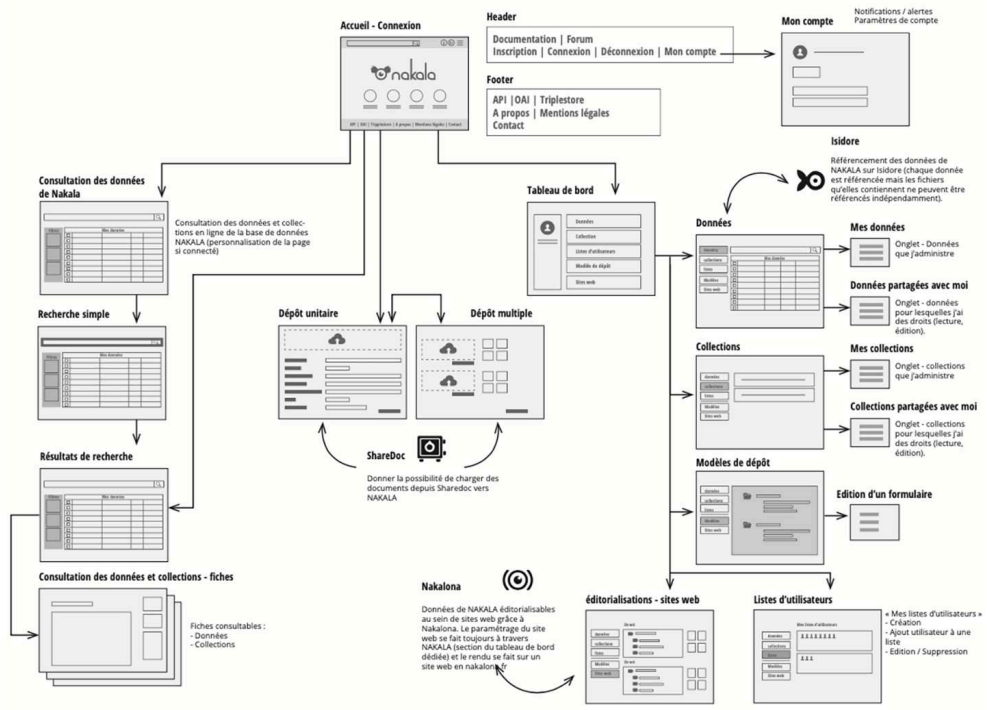
NAKALA

47 NAKALA (Fig. 4) est un service d'Huma-Num[42] permettant à des chercheurs, enseignants-chercheurs ou équipes de recherche, de partager, publier et valoriser tous types de données numériques documentées (fichiers textes, sons, images, vidéos, objets 3D, etc.) dans un entrepôt sécurisé (Fig. 5) afin de les publier en accord avec les principes FAIR et plus largement ceux de la science ouverte (accès ouvert, immédiat et réutilisable des données publiques)[43].

Figure 4. Page d'accueil du service NAKALA.

48 L'entrepôt Nakala assure à la fois l'accessibilité aux données et aux métadonnées ainsi que leur « citabilité » dans le temps à l'aide d'identifiants stables fournis par Huma-Num et fondés sur des identifiants de type Handle (jusqu'en 2021) et DOI[44] (depuis le 19 décembre 2020). Il s'inscrit dans la politique du Web des données qui permet notamment de rendre interopérables les métadonnées, c'est-à-dire la possibilité de les connecter à d'autres entrepôts existants suivant ainsi la logique des données ouvertes et liées. Par ailleurs, il propose un dispositif d'exposition des métadonnées qui permet leur référencement par des moteurs de recherche spécialisés comme ISIDORE. La description riche, précise et harmonisée des données avec NAKALA permet d'assurer leur pérennité, de garantir leur traçabilité sur le long terme et d'encadrer leur réutilisation. L'utilisation de NAKALA a pour finalité de cibler des projets visant à publier en ligne un ensemble de données associées à des métadonnées descriptives ayant une cohérence scientifique, comme des corpus, des collections, des reportages, etc. L'objectif de NAKALA est ainsi de viser la publication de jeux de données ou d'ensemble de données ayant une valeur scientifique ou culturelle importante[45].

Figure 5. Architecture du service NAKALA.



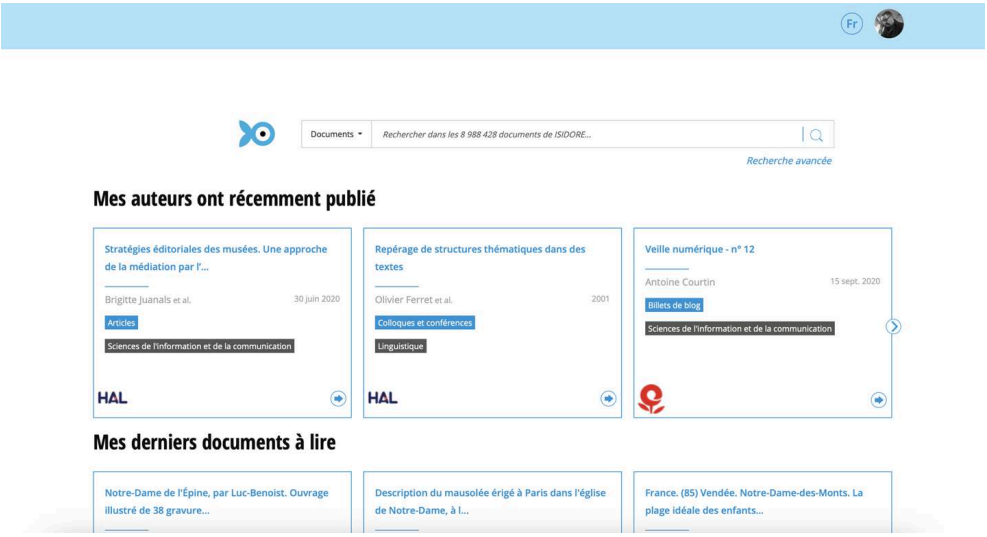
Source : L'Atelier Universel, 2021.

ISIDORE

49 Lancé le 8 décembre 2010, ISIDORE est le fruit de la collaboration du très grand équipement Adonis (Maignien, 2011) du CNRS (2007-2013), du Centre pour la communication scientifique directe [CCSD] et des sociétés Antidot, Mondéca et Sword (Pouyllau *et al.*, 2012). Il est actuellement développé et exploité par l'IR*[46] Huma-Num[47].

50 ISIDORE[48] est un moteur et assistant de recherche permettant de trouver des publications, des données numériques et profils de chercheurs et chercheuses en sciences humaines et sociales venant du monde entier (Fig. 6). Il permet de rechercher dans plusieurs millions de documents (articles, thèses et mémoires, rapports, jeux de données, pages Web, notices de bases de données, description de fonds d'archives, etc.) et des signalements d'événements (séminaires, colloques, etc.). Il propose aussi des fonctionnalités de réseau social scientifique (profil personnel, suivi d'auteurs, partage de collections bibliographiques, etc.). Il offre aussi de nombreuses fonctionnalités pour organiser sa veille scientifique (collections bibliographiques, alerte sur des requêtes, etc.).

Figure 6. Page d'accueil du service ISIDORE (anglais, espagnol et français).



51 Plus qu'un simple moteur de recherche, ISIDORE constitue une plateforme de traitement et d'enrichissement des données avec pour objectifs :

- d'offrir aux chercheurs un point d'accès unifié aux différentes ressources structurées produites dans le domaine des SHS en France ;
- d'exposer, selon les principes du *Linked Data,* les données bibliographiques structurées de la recherche en sciences humaines et sociales en France ;
- selon la logique d'une boucle de rétroaction, d'offrir les moyens aux producteurs de récupérer l'enrichissement automatique effectué par le moteur sur les données indexées.

52 ISIDORE et NAKALA constituent d'ores et déjà le « couple » central et moteur de l'écosystème Huma-Num. En 2021, Huma-Num a lancé dans le cadre du programme HNSO[49] un chantier pour renforcer ce couple et améliorer la FAIRisation des deux plateformes en travaillant sur les trois dimensions suivantes : 1°) un processus d'authentification unique, 2°) la convergence des référentiels, 3°) la procédure de moissonnage des données NAKALA par ISIDORE.

53 1. ISIDORE et NAKALA exploitent aujourd'hui le même dispositif d'authentification HumanID. Ce « hub » d'authentification offre la possibilité d'accéder en un clic à l'offre de services et d'applications de l'IR* Huma-Num en termes de stockage, de traitement, de diffusion ou encore d'exposition des données scientifiques[50]. HumanID est compatible avec l'ensemble de l'écosystème de l'enseignement supérieur et de la recherche internationale (*via* EduGAIN ou ORCID en particulier), mais aussi avec les outils les plus courants pour se connecter facilement à des services numériques. Développé avec le logiciel *open source* LemonLDAP, HumanID permet aux utilisateurs d'Huma-Num de faire des demandes d'accès aux services de l'écosystème Huma-Num et de visualiser les services connectés au Web dont ils sont ou peuvent être utilisateurs. Une évolution concrète pour l'amélioration de l'interconnexion entre ISIDORE et NAKALA consisterait à profiter de l'authentification partagée pour faciliter l'indexation des données NAKALA à partir du profil ISIDORE de l'utilisateur. Par exemple, une personne connectée et disposant de contenus « revendiqués » dans ISIDORE bénéficierait de propositions automatiques et profilées par discipline et/ou par thématique lors du dépôt de ses données dans NAKALA. Plus largement il s'agirait d'exploiter dans la base ISIDORE les informations sur les chercheurs venant des

plateformes de publications en SHS, afin de nourrir en métadonnées et de favoriser ainsi l'indexation de qualité dans NAKALA. Inversement, il serait possible d'utiliser les données déposées par un chercheur dans NAKALA pour suggérer dans ISIDORE des lectures relatives (disciplines, mots-clés, mots du titre) [Pouyllau, 2020].

54 2. ISIDORE et NAKALA exploitent les mêmes référentiels scientifico-documentaires et les mêmes auteurs. Ces référentiels communs permettent aujourd'hui de proposer à un déposant de données des labels de mots-clés (Fig. 7) et des « formes auteurs » (Fig. 8) par complétion automatique lorsque le déposant saisit ces mots-clés dans l'interface de saisie des métadonnées de NAKALA. Ce dispositif assure également une cohérence conceptuelle et une meilleure précision dans les processus de recherche d'information dans ISIDORE. Le chantier de FAIRisation propose d'améliorer le dispositif en exploitant également les URIs des labels afin d'assurer le suivi des modifications des concepts dans les référentiels et la prise en compte du multilinguisme[51].

Figure 7. Complétion automatique des labels de mots-clés présents dans les référentiels ISIDORE depuis l'interface de NAKALA.

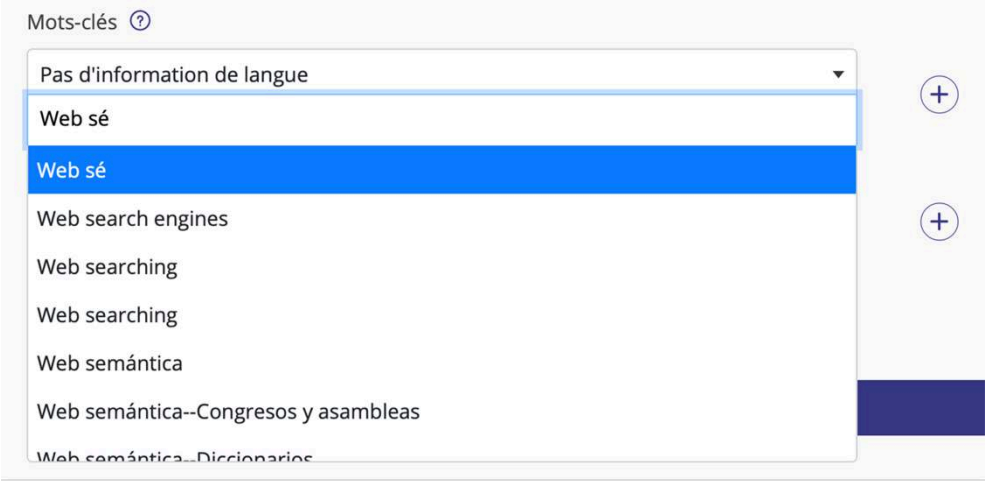
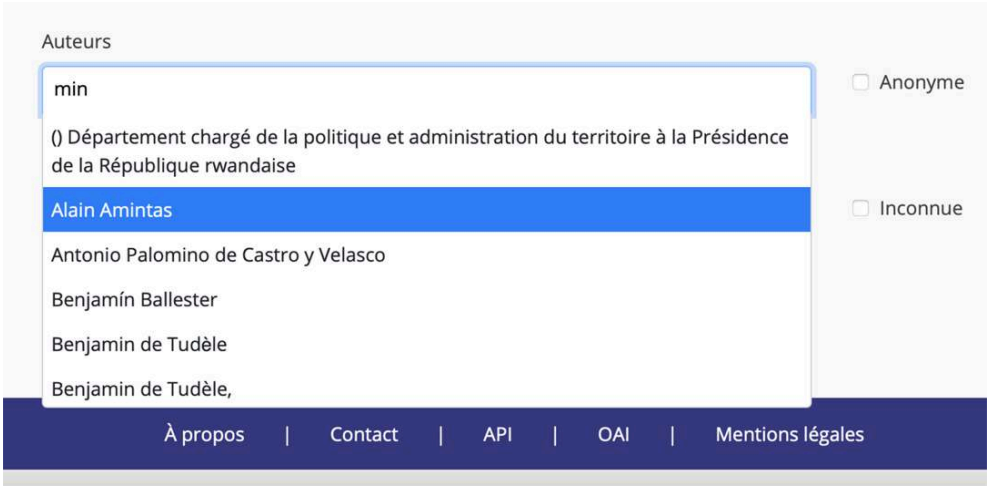


Figure 8. Complétion automatique des formes auteurs présentes dans ISIDORE depuis l'interface de NAKALA.



55 3. La création d'une collection dans NAKALA par un déposant entraîne aujourd'hui la création automatique d'un *set* dans l'exposition selon la norme OAI-PMH. Le chantier de FAIRisation prévoit qu'avec l'accord du déposant et à l'aide d'un menu accessible depuis l'interface Web de NAKALA, cette collection soit signalée à ISIDORE et automatiquement moissonnée. Alors que l'ajout de nouvelles sources ISIDORE résulte aujourd'hui d'une procédure manuelle, un ajout automatisé de nouvelles sources ISIDORE à la demande du déposant devrait très nettement améliorer la visibilité des données NAKALA.

Conclusion

56 Ce chapitre a présenté différentes infrastructures de recherche nationales et européennes dédiées à la conservation et à la découverte des données de la recherche. Nous avons mis à jour l'articulation entre infrastructures de la recherche et pratiques de la communauté académique en termes de dépôt de leurs données et de recherche au sein de ces entrepôts. Ces dernières années ont été marquées par une forte augmentation à la fois de la production de données et à la fois des usages de ces infrastructures. Usages et infrastructures sont ainsi amenés à évoluer rapidement, dans un contexte sociotechnique qui se transforme constamment comme le montre la montée en puissance des IA dans tous les domaines.

57 Sans pouvoir préjuger des orientations que prendront les différentes institutions face aux développements des pratiques, ni même celles d'Huma-Num vis-à-vis des instruments que sont ISIDORE et NAKALA, il nous semble intéressant de noter l'introduction récente des méthodes d'apprentissage machine et d'apprentissage profond qui ouvrent de nouvelles perspectives pour la gestion et le traitement des données de la recherche pour le classement, le requêtage ou la génération automatique de synthèse. Les infrastructures de recherche ont un enjeu particulier à produire des chaînes de traitement adaptées à des jeux de données à la fois très spécialisés et fortement enrichis. C'est une plus-value qui doit les différencier de la simple mise à disposition d'outils, y compris sémantiques, opération qui relève désormais davantage des DSI que des « infrastructures de recherche » au sens d'« infrastructures pour faire de la recherche ».

58 Les initiatives actuelles tentent d'exploiter le potentiel d'association des données fondée sur les principes du LOD[52], susceptible de créer des relations sémantiques entre des objets informationnels de natures différentes et provenant de plusieurs agrégateurs de données européens. Au-delà de l'Europe, le programme canadien LINCS[53] est l'un des premiers projets à mettre en œuvre une telle approche. Avec cette dernière, il devient envisageable de dépasser le portail Web de recherche documentaire pour proposer une découverte et une navigation entre concepts scientifiques, communautés de recherche et experts. L'un des enjeux de tels dispositifs sera de fournir des points d'entrées éditoriaux dans la masse des données, par exemple à partir d'une sélection de concepts disciplinaires ou en lien avec l'actualité scientifique.

BIBLIOGRAPHIE

Icek AJZEN, « The Theory of Planned Behavior », *Organizational Behavior and Human Decision Process*, 52, 2, 1991, p. 179-211.

Madeleine AKRICH, *The De-Scription of Technical Objects*, Cambridge, MIT Press, 1992.

Emmanuelle BERMÈS, Antoine ISAAC et Gauthier POUPEAU, *Le Web sémantique en bibliothèque*. Bibliothèques [Ressource électronique], Paris, Cercle de la librairie, 2013.

Tim BERNERS-LEE et Mark FISCHETTI, *Weaving the Web: The original design and ultimate destiny of the World Wide Web by its inventor*, Harper Business, 1999.

Christine L. BORGMAN, Herbert VAN DE SOMPEL, Andrea SCHARNHORST et Henk VAN DEN BERG, « Who uses the digital data archive? An exploratory study of DANS », *Proceedings of the Association for Information Science and Technology*, 1-4, 2015, https://doi.org/10.1002/pra2.2015.145052010096.

Christine L. BORGMAN, Peter T. DARCH, Ashley E. SANDS et Milena S. GOLSHAN, « The durability and fragility of knowledge infrastructures: Lessons learned from astronomy », *Proceedings of the Association for Information Science and Technology,* 53: 1-10, 2016, http://dx.doi.org/10.1002/pra2.2016.14505301057.

Christine L. BORGMAN, Peter T. DARCH et Milena. S. GOLSHAN, « Digital Data Archives as Knowledge Infrastructures: Mediating Data Sharing and Reuse », *JASIST*, 1-31, 2018, http://arxiv.org/abs/1802.02689.

Kathy BÖRNER et Elizabeth RECORD, « Macroscopes for making sense of science ». *Proceedings of the practice and experience in advanced research computing on sustainability, success and impact*, 2017, p. 64-74.

Gert BREITFUSS, Carla BARREIROS *et al.*, *Report on Stakeholder and Opportunity Analysis TRIPLE Project*, 2020, https://doi.org/10.5281/zenodo.3925662.

Stefan BUDDENBOHM, Maaike DE JONG, Jean-Luc MINEL et Yoann MORANVILLE, « Find Research Data Repositories for the Humanities – The Data Deposit Recommendation Service », *International Journal of Digital Humanities*, 2021, https://doi.org/10.1007/s42803-021-00030-7.

Daniel Owen CASE, *Looking for Information: A Survey of Research on Information Seeking, Needs, and Behavior* (2nd ed.), San Diego, Academic Press, 2006.

CLARIN, *CLARIN in a nutshell*, 2020, https://www.clarin.eu/content/clarin-nutshell-0.

Helena COUSIJN, Amye KENALl et Emma GANLEy, « A data citation roadmap for scientific publishers », *Scientific Data* 5, 2018, p. 180-259.

Data Citation Synthesis Group, *Joint declaration of data citation principles, San Diego CA, FORCE11,* 2014, https://doi.org/10.25490/a97f-egyk.

Peter K. DOORN, P. K., « Archiving and Managing Research Data : data services to the domains of the humanities and social sciences and beyond : DANS in the Netherlands », *Der Archivar* 73(01), 2020, p. 44-50, https://pure.knaw.nl/portal/en/publications/archiving-and-managing-research-data-data-services-to-the-domains.

Paul N. EDWARDS, Steven J. JACKSON, Melissa K. CHALMERS *et al.*, *Knowledge infrastructures : Intellectual frameworks and research challenges*, Ann Arbor, Deep Blue, 2006, http://hdl.handle.net/2027.42/97552.

ESFRI Report Working Group, *Monitoring of Research Infrastructures Performance*, 2019, https://www.esfri.eu/sites/default/files/ESFRI_WG_Monitoring_Report.pdf.

Bernard ESPINASSE, *Introduction aux entrepôts de données,* 2021, https://pageperso.lis-lab.fr/bernard.espinasse/wp-content/uploads/2021/12/2-Intro-Entrepots-4p.pdf

European Commission, Directorate-General for Research and Innovation, *Six Recommendations for implementation of FAIR practice by the FAIR in practice task force of the European open science cloud FAIR working group*, Publications Office, 2020, https://data.europa.eu/doi/10.2777/986252.

Fabien GANDON, Catherine FARON-ZUCKER et Olivier CORBY, *Le Web sémantique : comment lier les données et les schémas sur le Web ?,* coll. « InfoPro. Management des systèmes d'information » , Paris, Dunod, 2012.

Kathleen GREGORY, « A dataset describing data discovery and reuse practices in research », *Scientific Data* 7, 2020, https://doi.org/10.1038/s41597-020-0569-5.

Kathleen GREGORY, Helena COUSIJN, Paul GROTH *et al.*, « Understanding data search as a socio-technical practice », *Journal of Information Science*, 2019, https://doi.org/10.1177/0165551519837182.

Kathleen GREGORY, Helena COUSIJN, Paul GROTH *et al.*, « Searching data: A review of observational data retrieval practices in selected disciplines », *Journal of the Association for Information Science and Technology,* 2019, https://doi.org/10.1002/asi.24165.

Michael GUSENBAUER, « Google Scholar to overshadow them all ? Comparing the sizes of 12 academic search engines and bibliographic databases », *Scientometrics,* 18, 1, 2019, p. 177-214. https://doi.org/10.1007/s11192-018-2958-5.

Catherine HUGO, *Étude comparative des services nationaux de données de recherche Facteurs de réussite, rapport du COSO*, 2020.

Madjid IHADJADENE et Stéphane CHAUDIRON, « Quelles analyses de l'usage des moteurs de recherche », *Questions de communication,* 14, 2008, p. 17-32, https://doi.org/10.4000/questionsdecommunication.604.

Franciska DE JONG, Bente MAEGAARD, Darja FIŠER, Dieter VAN UYTVANCK et Andreas WITT, « Interoperability in an Infrastructure Enabling Multidisciplinary Research: The case of CLARIN », *Proceedings LREC 2020*, p. 3406-3413, https://www.aclweb.org/anthology/2020.lrec-1.417.

Helena KARASTI et Jeanette BLOMBERG, « Studying Infrastructuring Ethnographically », *Computer Supported Cooperative Work (CSCW),* 2017, https://doi.org/10.1007/s10606-017-9296-7.

Youngseek KIM et Jeffrey M. STANTON, « Institutional and individual factors affecting scientists' data-sharing behaviors: a multilevel analysis », *Journal of the Association for Information Science and Technology,* 67, 2016, p. 776-799, http://dx.doi.org/10.1002/asi.23424.

Emilio DELGADO LOPEZ-COZAR, Enrique ORDUNA-MALEA et Alberto MARTIN, « Google Scholar as a data source for research assessment », *Computer Science*, 2018, https://doi.org/10.31235/osf.io/pqr53.

Yannick MAIGNIEN, *ISIDORE, de l'interconnexion de données à l'intégration de services*, 2011, https://archivesic.ccsd.cnrs.fr/sic_00593320v2/document.

Paolo MANGHI, Alessia BARDI et Jochen SCHIRRWAGEN, *D7.2 – Interoperability with EOSCpilot services*, 2018, https://doi.org/10.5281/zenodo.3701434.

Davor OSTOJIC, Go SUGIMOTO et Matej DURCO, « The Curation Module and Statistical Analysis on VLO Meta-data Quality », *Selected papers from the CLARIN Annual Conference 2016*, 2017, p. 90-101.

Gautier POUPEAU, « Bilan de 15 ans de réflexion sur la gestion des données numériques », *Les petites cases (blog),* 12 octobre 2016, http://www.lespetitescases.net/bilan-reflexion-sur-la-gestion-des-donnees-numeriques.

Stéphane POUYLLAU, *ISIDORE : une plateforme de recherche de documents et d'information pour les Sciences Humaines et Sociales*, 2011, https://archivesic.ccsd.cnrs.fr/sic_00605642.

Stéphane POUYLLAU, *Classifieur de titres utilisant les données du moteur de recherche ISIDORE et l'API Keras*, 2020, https://doi.org/10.5281/zenodo.3991994.

Stéphane POUYLLAU, Jean-Luc MINEL, Shadia KILOUCHI et Laurent CAPELLI, *Bilan 2011 de la plateforme ISIDORE et perspectives 2012-2015. Comité de pilotage du TGE Adonis,* 2012, https://hal.archives-ouvertes.fr/sic_00690558.

Stéphane POUYLLAU, Mélanie BUNEL, Jean-Luc MINEL et Laurent CAPELLI, *"We": a Proposal for the TRIPLE platform*, 2020, https://doi.org/10.5281/zenodo.4032622.

Nicolas SAURET, « Design de la conversation scientifique : naissance d'un format éditorial », *Sciences du Design,* n° 8, 2, 2018, p. 57-66, https://www.cairn.info/revue-sciences-du-design-2018-2-page-57.htm.

Andrea SCHARNHORST, « Walking through a library remotely. Why we need maps for collections and how KnoweScape can help us to make them », *Les cahiers du numérique,* 11, 2015, p. 103-27.

Richard W. SCOTT, *Institutions and Organizations*, Thousand Oaks, CA, Sage Publication, 2001 [1995].

Alexandre SERRES, Marie-Laure MALINGRE, Morgane MIGNON, Cécile PIERRE et Didier COLLET, *Données de la recherche en SHS. Pratiques, représentations et attentes des chercheurs : une enquête à l'Université Rennes 2 : Rapport ; Annexe 1 : Résultats de l'enquête statistique ; Annexe 2 : Croisements statistiques ; Annexe 3 : Extraits des entretiens ; Synthèse des résultats*, 2017, https://hal.archives-ouvertes.fr/hal-01635186v2.

Brigitte SIMONNOT, *L'accès à l'information en ligne. Moteurs, dispositifs et médiations*, coll. « Systèmes d'information et organisations documentaires », Hermès Lavoisier, 2012, https://archivesic.ccsd.cnrs.fr/sic_00804286.

Niels STERN, Jean-Claude GUÉDON et Thomas WIBEN JENSEN, « Crystals of Knowledge Production. An Intercontinental Conversation about Open Science and the Humanities », *Nordic Perspectives on Open Science,* 1, 2015, p. 1-24, https://doi.org/10.7557/11.3619.

Mark D. WILKINSON, Michel DUMONTIER, Ijsbrand AALBERSBERG *et al.*, « The FAIR Guiding Principles for scientific data management and stewardship », *Scientific Data* 3, 160018, 2016, https://doi.org/10.1038/sdata.2016.18.

Peter WITTENBURG et Franciska DE JONG, « State of FAIRness in ESFRI Projects », *Data Intelligence,* 2, 1-2, 2020, p. 230-237.

NOTES

1. FAIR est l'acronyme de *Findable, Accessible, Interoperable, Reusable* et se traduit en français par « Faciles à (re)trouver, Accessibles, Interopérables, Réutilisables ».

2. Ce texte reprend des parties des chapitres de l'ouvrage en édition continue « Propositions méthodologiques pour ISIDORE et NAKALA » réalisé dans le cadre du programme Huma-Num Science Ouverte [HNSO] financé par le Fonds national pour la Science ouverte.

3. Institut Digital Archiving and Networked Services (Pays-Bas).

4. *Digital Object Identifier* : identifiant pérenne et unique d'un fichier en ligne.

5. Scopus est une base de données transdisciplinaire de résumés et de citations de publications scientifiques lancée par l'éditeur scientifique Elsevier en 2004.

6. En janvier 2020, 30,4 % des titres de Scopus sont issus des sciences de la santé, 15,4 % des sciences de la vie, 28 % des sciences physiques et 26,2 % des sciences sociales. Scopus dispose d'un processus d'examen étendu et bien défini pour l'inclusion des revues ; 10 % des quelque 25 000 sources indexées dans Scopus sont publiées par Elsevier (Gregory 2020).

7. Une API (interface de programmation d'application) est une interface logicielle qui permet de « connecter » un logiciel ou un service à un autre logiciel ou service afin d'échanger des données.

8. En anglais : *Datawarehouse.*

9. D'après la définition d'Inmon (1992). Dans sa présentation, Espinasse présente les différents types de données comme suit :

- thématiques ou orientées sujet : un entrepôt de données rassemble et organise des données associées aux différentes structures fonctionnelles de l'entreprise, pertinentes pour un sujet ou thème et nécessaires aux besoins d'analyse ;
- intégrées : les données résultent de l'intégration de données provenant de différentes sources pouvant être hétérogènes ;
- historisées : les données d'un entrepôt de données représentent l'activité d'une entreprise durant une certaine période (plusieurs années) permettant d'analyser les variations d'une donnée dans le temps ;
- non volatiles : les données de l'entrepôt de données sont essentiellement utilisées en interrogation (consultation) et ne peuvent pas être modifiées (sauf certains cas de rafraîchissement).

10. *Registry of Research Data Repositories*. Registre mondial des entrepôts de données de recherche. Disponible à l'url : https://www.re3data.org.

11. Au 2 décembre 2021.

12. En appliquant les filtres « Humanities and Social Sciences », « non profit institution » et « FAIR », le catalogue recense un total de 45 infrastructures.

13. Service lié à l'université d'Oxford. Disponible à l'url : https://fairsharing.org/databases/.

14. Au 2 décembre 2021.

15. Disponible à l'url : https://www.loterre.fr/skosmos/TSO/fr/.

16. *Open Archives Initiative* (initiative pour des archives ouvertes).

17. « Des réseaux solides de personnes, de dispositifs et d'institutions qui génèrent, partagent et maintiennent des connaissances spécifiques sur les mondes humain et naturel » (Traduction des auteurs).

18. COmité pour la Science Ouverte.

19. Cette description s'appuie sur les articles de Borgman *et al.* (2015) et Doorn (2020). Disponible à l'url : https://dans.knaw.nl/nl/.

20. Disponible à l'url : https://easy.dans.knaw.nl/ui/home.

21. Disponible à l'url : https://dataverse.nl/.

22. Disponible à l'url : https://www.narcis.nl/.

23. Voir à l'url : https://www.coretrustseal.org/.

24. Disponible à l'url : https://researchdata.nl/.
25. Disponible à l'url : https://eudat.eu/european-data-initiative.
26. Disponible à l'url : https://ariadne-infrastructure.eu/.
27. Disponible à l'url : https://eosc-portal.eu/.
28. Disponible à l'url : https://www.ehri-project.eu/.
29. Un ensemble de données (*data set*) EASY est l'équivalent d'une « collection » dans la terminologie de la *Dublin Core Metadata Initiative*. Les ensembles de données sont étiquetés avec un ou plusieurs codes de classification disciplinaire.
30. Disponible à l'url : https://www.fairsfair.eu/.
31. Disponible à l'url : https://www.dcc.ac.uk/.
32. Disponible à l'url : https://fairaware.dans.knaw.nl/.
33. Cette description s'appuie sur CLARIN 2020 ; Jong *et al.*, 2020 ; Wittenburg et Jong, 2020.
34. Un ERIC est une entité juridique internationale, créée par la Commission européenne en 2009.
35. Les membres de CLARIN sont des gouvernements ou des organisations intergouvernementales.
36. Acronyme de Galleries, Libraries, Archives, Museums qui désigne le secteur des galeries, bibliothèques, archives et musées.
37. En France, le consortium CORLI est un centre de connaissance https://corli.huma-num.fr/fr/.
38. *Open Archives Initiative* Protocol *for Metadata Harvesting.* Ce protocole, développé par l'Open Archives Initiative, a pour objectif d'échanger des métadonnées entre institutions afin de multiplier les accès possibles aux documents numériques concernés.
39. Disponible à l'url : http://vlo.ChapitreInfrastructures-latest-img5.eu
40. Google Scholar a été lancé fin 2004.
41. Voir à l'url : https://www.openaire.eu/openaire-history
42. Disponible à l'url : https://www.nakala.fr.
43. Voir documentation sur NAKALA disponible en ligne : https://documentation.huma-num.fr/nakala/.
44. Disponible à l'url : https://www.doi.org/.
45. Par exemple, un fichier vidéo déposé dans NAKALA peut être inséré dans des pages Web, comme dans le cas d'un carnet de recherche *Hypothèses* (disponible à l'url : https://fr.hypotheses.org/) ou dans un webdocumentaire.
46. Appelées jusqu'en 2021 « Très Grandes Infrastructures de Recherche » [TGIR], les IR* sont les infrastructures relevant d'une politique nationale et d'un budget du ministère de l'Enseignement supérieur, de la Recherche et de l'Innovation, contrairement aux IR qui sont sous la responsabilité des opérateurs de recherche.
47. Sur l'historique de l'outil, nous renvoyons le lecteur à la documentation sur ISIDORE disponible sur : https://documentation.huma-num.fr/isidore.
48. Disponible à l'url : https://isidore.science/.
49. Voir https://hnlab.huma-num.fr/blog/projets/hnso/.
50. Voir https://humanum.hypotheses.org/5754.
51. La description et l'administration des référentiels communs à ISIDORE et NAKALA sont décrits dans les chapitres Référentiels, Concepts, Définitions et Administration, Les référentiels utilisés par ISIDORE et NAKALA et Administration des référentiels utilisés dans ISIDORE de l'ouvrage HNSO « Propositions méthodologiques pour ISIDORE et NAKALA », en ligne https://hnlab.huma-num.fr/blog/2022/03/15/ouvrage-HNSO/.
52. *Linked Open Data* : « données ouvertes et liées ».
53. Linked Infrastructure for Networked Cultural Scholarship, https://lincsproject.ca

AUTEURS

NICOLAS SAURET

Maître de conférences, chercheur au Laboratoire Paragraphe, université Paris 8.

JEAN-LUC MINEL

Professeur émérite, université Paris Nanterre. Président du conseil scientifique de l'IR* Huma-Num.

MÉLANIE BUNEL

Ingénieure d'études spécialisée en ingénierie documentaire, projet Huma-Num Science Ouverte [HNSO].

STÉPHANE POUYLLAU

Ingénieur de recherche au CNRS, responsable du HN Lab.

Vie des archives et plan de gestion de données rétrospectif : récit d'une expérience à partir du fonds de l'anthropologue Jean-Pierre Olivier de Sardan

Annick Boissel et Véronique Ginouvès

1 Lorsqu'un projet obtient un financement de l'Agence nationale de la recherche ou de l'Union européenne[1], la rédaction d'un plan de gestion de données [PGD] ou *data management plan [DMP]* est désormais obligatoire. Pour justifier de cette obligation, les instances de financement et d'évaluation s'appuient sur le développement de la politique de la science ouverte, corrélée aux principes FAIR[2] qui régissent les bonnes pratiques en matière de données. En effet, l'objectif du PGD est de consigner, dès la conception du projet, l'ensemble des données, qui seront recueillies avec leurs caractéristiques, et de préciser leurs conditions d'utilisation et leur diffusion pendant et après le projet. Pour faciliter le travail des porteurs de projet, plusieurs institutions proposent des guides pour la rédaction des PGD. Pour aider la communauté académique française, l'Institut français de l'information scientifique et technique [INIST] a développé la plateforme de formation *DoRANum*[3] qui regroupe des modules de formation et des outils, dont le site OPIDoR[4] et sa déclinaison dédiée à l'élaboration des plans de gestion des données de la recherche : DMP-OPIDoR.

2 En réalité, les archivistes ont toujours explicité les informations qui régissent la rédaction des PGD. La norme de description archivistique de l'*Encoded Archival Description*[5] *[EAD]* engage le catalogueur à préciser le contexte de production des données traitées <origination>, la date de leur production <unitdate>, leur(s) langue(s) <langmaterial>, leur description physique <physdesc> et leurs typologies <physfacet> ou encore les droits afférant à leur diffusion <accesrestrict>[6]. Les principes FAIR sont venus conforter, eux aussi, les pratiques archivistiques. Les normes et les standards utilisés par la profession permettent effectivement de rendre les données plus faciles à

trouver, accessibles, interopérables et réutilisables. Au sein du secteur Archives de la recherche de la médiathèque & Phonothèque de la MMSH, les archivistes se sont emparés du PGD comme une forme de récit de ses fonds. Au-delà d'un projet, il s'agit de décrire le processus d'intégration des données d'un fonds, de leur collecte à leur conservation sur le long terme, dans ce que nous appelons, « un plan de gestion de données rétrospectif ».

3 Cet article, rédigé par deux archivistes, est issu d'un travail commun de longue haleine qui a débuté par une réflexion sur la rédaction d'un plan de gestion de données rétrospectif à la suite du don de l'anthropologue Jean-Pierre Olivier de Sardan. Nous nous sommes rendu compte de la complexité d'une telle rédaction lorsque le don du chercheur était discontinu, parfois circonstanciel, avec des objectifs évolutifs en fonction de sa propre recherche. Il fallait à la fois revenir en arrière pour reconstituer l'histoire du fonds et son contexte de production, comme une forme d'analepse, et aller de l'avant en enchâssant le passé, le présent et les années à venir. Rédiger ensemble ce texte a été pour nous l'occasion de proposer des éléments réflexifs sur notre expérience, d'en évaluer l'intérêt et de revenir très concrètement sur l'agencement de cette rédaction afin de partager les chemins qui nous restent à parcourir.

Pourquoi concevoir un plan de gestion de données rétrospectif ?

4 Avant d'explorer les raisons qui nous ont poussées à rédiger des plans de gestion de données rétrospectifs, il nous semble nécessaire de décrire le contexte de notre activité et de préciser nos missions.

5 Les fonds pour lesquels nous rédigeons ce type de plan sont ceux du secteur Archives de la recherche de la médiathèque SHS & Phonothèque de la Maison méditerranéenne des sciences de l'homme [MMSH]. La MMSH est un campus de recherche, unité d'accompagnement de la recherche qui associe l'université Aix-Marseille et le CNRS autour des études aréales méditerranéennes et africaines relevant des sciences humaines et sociales. Jusqu'en septembre 2021, sa médiathèque était composée d'une bibliothèque, d'une phonothèque et d'une iconothèque. C'est la phonothèque qui a réuni les deux auteures : Véronique Ginouvès en est responsable depuis la création de la MMSH et, dans ce cadre, elle a reçu en stage Annick Boissel alors en formation à l'Intd/Cnam[7]. Le traitement et l'archivage des données sonores et audiovisuelles mis en œuvre à la phonothèque ont été décrits dans plusieurs articles[8], mais une réorganisation interne consistant à intégrer tous les types d'archives dans un seul secteur a engagé l'équipe[9] dans une description rigoureuse du processus de la vie des données. Un autre événement, beaucoup plus brutal, a imposé cette nécessité. Le 27 avril 2020, un virus rançonneur a fait disparaître toutes les données des serveurs de la MMSH[10]. L'activité numérique de la MMSH est restée perturbée plus d'une année après la cyberattaque, mais cet événement a été instructif : décrire l'ensemble des processus d'archivage et de sauvegarde est nécessaire pour que les données puissent être interprétées avec le contexte technologique du moment. Un article prophétique[11] de David Zeitlyn débute par une image terrifiante : celle du musée de Rio de Janeiro en flammes la nuit des 2 et 3 septembre 2018. Des milliers d'artefacts de la société amérindienne ont disparu, mais David Zeitlyn s'appuie sur cet événement pour nous démontrer que la sauvegarde matérielle des objets n'est pas le seul enjeu. Il est

nécessaire que ces objets soient décrits et numérisés, que les questions juridiques et éthiques aient été réglées afin que leur image numérique puisse être réutilisée simplement. Depuis la catastrophe de Rio de Janeiro, d'autres drames sont venus conforter son point de vue. Le 18 avril 2021, c'est la bibliothèque de l'université de la ville du Cap [University of Cape Town] qui est dévastée par les flammes. La section des Études africaines de l'UCT abritait une collection d'archives constituée dès 1953 et conservait des films concernant toute l'Afrique ; elle a été en grande partie détruite[12]. Des archivistes du monde entier ont entrepris de reconstituer cette collection[13], mais restaurer les films ou retrouver des copies prendra des années.

6 À plus petite échelle, la catastrophe informatique de la MMSH a en quelque sorte agi comme une déflagration. Certes, la phonothèque n'a pas perdu de données parce qu'elle bénéficiait dès 2007 des outils de sauvegarde mis en place par une infrastructure de recherche créée par le CNRS[14]. Mais le message a été compris. Face au chaos qui peut survenir à tout moment, il est capital non seulement de s'engager dans une conservation sur le long terme, mais aussi d'expliciter précisément le contexte et le processus de traitement de nos fonds. Il ne suffit pas de décrire les modalités de leur sauvegarde ou de préciser finement les questions juridiques et éthiques, il faut aussi informer de notre environnement numérique et humain afin que, dans les années à venir, ces archives puissent être réutilisées avec la meilleure compréhension possible du contexte de leur production comme de leur traitement.

7 Ainsi, le PGD rétrospectif est pour nous plus qu'un document administratif. Il est un outil pour décrire et penser le traitement des données, un instrument collaboratif utile pour les producteurs, les informaticiens, les archivistes et l'ensemble des réutilisateurs des données, présents et à venir. Au fil de sa rédaction et de ses versions successives, il identifie les actions à réaliser pour poursuivre le traitement du fonds et, à terme, facilite la compréhension du fonds dans sa globalité. Il est aussi le résultat d'une enquête menée méticuleusement pour reconstituer la production, la collecte et le traitement des données. La décision de se lancer dans la rédaction d'un tel PGD est finalement assez circonstancielle. À la différence du PGD de projet, elle peut se décider à n'importe quel moment, mais ce plan est systématiquement réalisé lorsque nous terminons le traitement du fonds et qu'il est prêt à être déposé sur les serveurs du CINES[15] pour une sauvegarde sur le long terme. Le PGD rétrospectif est une forme de récit qui raconte l'histoire d'un fonds et de ses collections en prenant en compte toutes celles et ceux qui ont participé à sa production et à son analyse ainsi que tous les outils qui leur ont été utiles.

Retour d'expérience sur la rédaction d'un plan de gestion de données rétrospectif : le fonds Jean-Pierre Olivier de Sardan

Figure 1. Supports originaux du fonds Jean-Pierre Olivier de Sardan.

8 La rédaction du plan de gestion de données, appliqué au fonds de l'anthropologue franco-nigérien Jean-Pierre Olivier de Sardan[16], est pour nous l'occasion d'illustrer les particularités archivistiques d'un plan de gestion de données rétrospectif. En 2017, le chercheur a fait don d'une partie de ses matériaux de terrain à la phonothèque de la Maison méditerranéenne des sciences de l'homme. Il était particulièrement curieux de l'existence d'une phonothèque et très intéressé à l'idée de pouvoir partager ses données avec le Laboratoire d'études et de recherche sur les dynamiques sociales et le développement local[17] [LASDEL] à Niamey (Niger), dont il a été membre fondateur en 2001. Les données audiovisuelles, sonores et textuelles liées à ses recherches ont été collectées entre 1965 et 2008 sur différents supports physiques tels que cassettes audio, bandes magnétiques sur bobine, cassettes VHS, cahiers de terrain et autres documents papier. Les principaux thèmes représentés dans les archives du chercheur sont liés à l'aire géographique nigérienne. Dans les années 1970 et 1980, l'anthropologue réalise des entretiens en langue zarma dans le cadre de ses recherches sur la culture et la civilisation songhay-zarma. En 1986, il recueille des témoignages sur la pratique et l'évolution des cultes de possession chez des émigrés nigériens au Ghana, au Bénin et en Côte d'Ivoire. Puis, dans les années 2000, il étudie le fonctionnement des services administratifs, techniques, sanitaires, éducatifs et la crise alimentaire au Niger, dans le cadre de son activité de chercheur au sein du LASDEL. En dehors de ce terrain nigérien, il revient en France au début des années 1980, où il réalise une étude sur l'identité et l'avenir d'habitants de la région de la Margeride en Lozère. Jean-Pierre Olivier de Sardan a également réalisé plusieurs films ethnographiques entre 1967 et 1992, liés pour la majorité d'entre eux aux thèmes cités ci-dessus et pour lesquels il a déposé, d'une part, des copies ou enregistrements originaux, et, d'autre part, des rushes sonores. L'ensemble des documents analogiques ont été numérisés progressivement, réunissant sous forme de données numériques entretiens, rushes sonores, films édités,

vidéos montées, notes de terrain, transcriptions d'entretiens, notes sur la langue zarma et lexiques, correspondance, schémas, articles, préparations de cours ou coupures de presse. En 2020, à l'occasion d'échanges sur le traitement archivistique, le chercheur a communiqué à la phonothèque d'autres données textuelles complémentaires nativement numériques cette fois (fiches de lecture, transcriptions d'entretien d'anciens combattants nigériens et une bibliographie du chercheur).

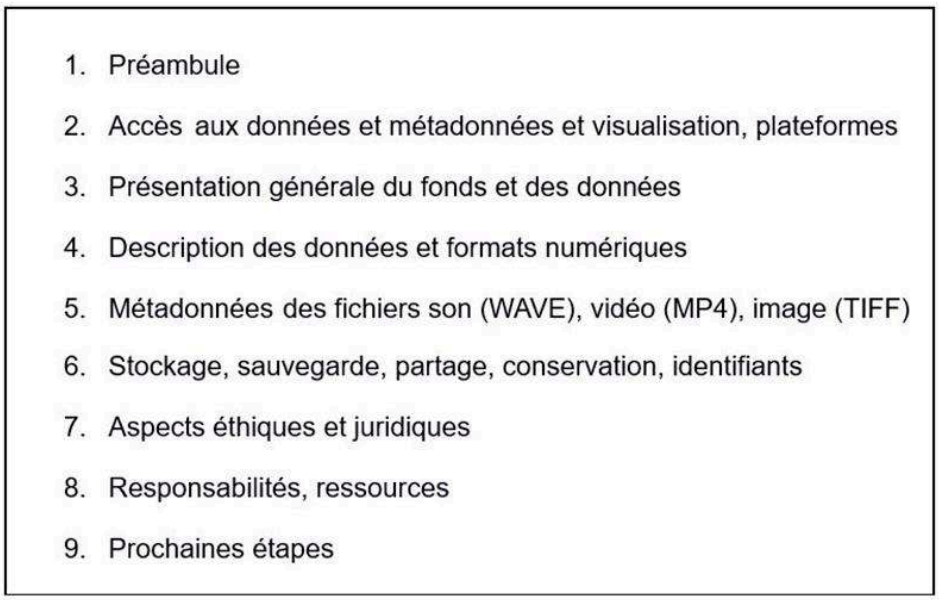
9 À ces données de recherche et données du chercheur s'ajoutent les métadonnées, créées par les archivistes, qui sont la documentation archivistique des données[18] et qui constituent un enrichissement indispensable de la donnée pour son processus de FAIRisation. S'ajoutent également les documents juridiques tels que les contrats signés entre le chercheur et les personnes interrogées lors des entretiens et les contrats de dépôt, d'utilisation et de diffusion signés avec l'institution, les contrats de collaboration/partenariats avec d'autres institutions comme le LASDEL, ainsi que les nouveaux documents créés par les archivistes : instruments de recherche (tableaux d'inventaires, catalogues), entretiens réalisés avec le chercheur, mémoires d'étudiants ou billets de blog (à répertorier), courriers et courriels ayant une valeur juridique, ainsi que le plan de gestion de données.

10 C'est cet ensemble qui constitue le fonds et qui a fait l'objet de la rédaction d'un plan de gestion de données. Ce plan, bien décrit par les services de l'Institut de l'information scientifique et technique [INIST], vise à décrire et à rendre compte du traitement et de l'organisation des données et métadonnées du fonds, au cours de leur cycle de vie : création ou collecte, traitement, analyse, conservation, accès, réutilisation. Son objectif, au début d'un projet, est de préparer le partage, la réutilisation, le stockage et la conservation sur le long terme, en précisant quel type de communication des données est envisagé, le cas échéant, les raisons de leur fermeture.

11 La rédaction d'un premier PGD nécessite d'avoir recours à la fois à des réflexions théoriques, pour comprendre l'agencement et les raisons du plan, et à des ressources pratiques représentées par les divers PGD que l'on peut consulter sur Internet ainsi que ceux déjà créés le cas échéant par sa propre institution et par ses pairs. Outre le réseau professionnel signalé précédemment de l'INIST-CNRS et sa plateforme de formation *DoRANum*, il existe sur le Web de nombreux sites qui permettent de rédiger ce type de plan. Les pays anglo-saxons en particulier les organisent de façon très pragmatique, à l'exemple du MIT[19] de l'université de Stanford[20] ou de celle d'Oxford[21]. On voit que les outils et les modèles sont nombreux, mais il faut comprendre quel est celui qui correspondra le mieux au fonds que l'archiviste est en train de traiter. Pour le fonds Jean-Pierre Olivier de Sardan, la première trame de rédaction s'est appuyée sur le modèle proposé par l'Agence nationale de la recherche, pour ses caractéristiques standard qui pourraient être amendées au fur et à mesure des interactions réflexives des différents acteurs du secteur Archives de la recherche de la médiathèque & Phonothèque de la MMSH.

12 À partir des principaux éléments[22] que doit comporter un PGD, celui du fonds Jean-Pierre Olivier de Sardan[23] s'organise aujourd'hui, dans sa quatrième version, en neuf grandes rubriques, présentées et commentées ci-dessous, en mettant l'accent sur la prise en compte de la spécificité archivistique et rétrospective. Ce PGD est mis à jour lorsque de nouveaux éléments nécessitent une modification.

Figure 2. Plan de gestion de données du fonds Jean-Pierre Olivier de Sardan : plan général.



Préambule

13 Ce paragraphe a pour objet de présenter le document, d'apporter des informations administratives, ainsi que des informations générales sur les auteurs, les versions successives et la méthodologie de rédaction.

Accès aux données et métadonnées et visualisation, plateformes

14 Cette rubrique, placée dès le début du PGD, a été ajoutée pour répondre au plus tôt à la question primordiale de l'accès au fonds, aux données, métadonnées et documents archivistiques décrits précédemment. Les liens vers les données et métadonnées sont indiqués. Ici, pour le fonds Jean-Pierre Olivier de Sardan, on apprendra que les données sont cataloguées sur le Catalogue en ligne des archives et des manuscrits de l'enseignement supérieur (Calames)[24], qu'elles sont accessibles sur les plateformes Nakala[25], Ganoub[26], Calames ainsi que sur le site physique de la MMSH, qu'elles sont conservées sur les serveurs de la TGIR HumaNum et qu'elles seront, à terme, déposées au CINES pour une conservation sur la longue durée. Ces plateformes et infrastructures, ainsi que d'autres[27] liées à l'activité du fonds et du secteur Archives de la recherche de la médiathèque & Phonothèque de la MMSH, sont décrites dans cette partie, ainsi que le service lui-même et son histoire. Notons que la description de ces plateformes pourra être semblable pour tous les PGD du secteur Archives de la recherche de la médiathèque & Phonothèque de la MMSH.

Présentation générale du fonds et des données

15 Il s'agit tout d'abord de comprendre rapidement le fonds, dans sa globalité, en décrivant succinctement les sujets de recherche de l'anthropologue. Sont cités également les autres fonds présents dans le service, qui sont reliés par une aire

géographique commune ou un sujet de recherche commun. Cinq autres fonds le sont effectivement. De même, les mémoires d'étudiants et billets universitaires en rapport avec le fonds sont également cités et cliquables. Ensuite, les informations plus techniques et organisationnelles sont apportées : volume des données, plan de classement, plan de nommage. Les acronymes utilisés dans les noms de fichier sont explicités. Puis une classification générale des données sous forme de tableau est présentée, réunissant toutes les données ou produits de recherche (dénomination selon la plateforme DMP Opidor). Deux catégories apparaissent : les données du chercheur d'une part et les données archivistiques d'autre part, elles-mêmes subdivisées en sous-catégories. Le tableau présente également la nature des supports physiques originaux utilisés lors de la collecte des données, leur type de contenu et les formats numériques des données, après numérisation le cas échéant.

Tableau 1. Classification générale des données. Fonds Jean-Pierre Olivier de Sardan.

Catégories de données / documents		Support original de la donnée	Type de contenu	Formats numériques*	Natif ou non
Données du chercheur	Documents sonores	- cassettes audio - bandes magnétiques sur bobines	- entretiens - rushes sonores	wave (44,6khz/16bits) mp3 80 Go	Fichiers issus de la numérisation des supports physiques
	Documents audiovisuels	- cassettes VHS - DVD	- films édités - vidéos montées	mp4 16,3 Go	
	Documents papiers	- cahiers - feuilles	- notes de terrain, transcriptions d'entretiens, notes sur la langue zarma et lexiques, correspondance, schémas, articles, préparations de cours ou coupures de presse	tiff jpg pdf 11 Go	
	Documents numériques	- fichiers texte	- transcriptions d'entretiens - fiches de lecture - bibliographie	doc pdf 0,12 Go	Natif num.
Données archivistiques	Instruments de recherche	- fichiers tabulés - fichiers textes	- tableaux inventaires - catalogues	odt ods doc pdf	Natif num.
	Documents juridiques	- feuilles	- contrats	inf à 1 Go pdf inf à 1 Go	Mixte
	Documents d'activité des archivistes	- courriels et courriers - fichiers tabulés - fichiers textes - liens vers sites/pages web - fichiers son - divers	-correspondance - documentation - documents de travail - mémoires d'étudiants - billets de blog - URL - entretiens avec le chercheur - plan de gestion de données	odt csv ods doc xls pdf url wav mp3 4 Go	Mixte

* formats de conservation : wave, tiff, pdf/A ; formats de diffusion : mp3, jpeg, pdf

Description des données et formats numériques

16 Après une note générale sur les formats de conservation et de consultation arrive la description détaillée, sous forme de tableau, des fichiers sonores, audiovisuels et textuels issus de la numérisation des supports analogiques ou nativement numériques et des documents archivistiques que sont les instruments de recherche (inventaires, catalogues), les documents juridiques (contrats, coordonnées, courriers, courriels), entretiens avec le déposant et autres documents de travail des archivistes. Une colonne « commentaire » permet d'ajouter une note sur le traitement restant à appliquer à la donnée, dans la poursuite du processus de traitement archivistique du fonds et de sa FAIRisation. Les problématiques et les actions correctrices sont ainsi mises en évidence.

Le tableau en exemple ci-dessous décrit les instruments de recherche, en termes de support, nature, format et nom de fichier, quantité.

Tableau 2. Description des instruments de recherche. Fonds Jean-Pierre Olivier de Sardan.

Instruments de recherche (inventaires, catalogues)					
Support et nombre	Nature	Type fichier	Nb	Fichiers	Commentaire
1 fichier	Inventaire général	ods	1	MMSH_ODS_INV_GENERAL_V20211111	Convertir en csv Dépôt Nakala
2 fichiers	Catalogue à destination des membres du LASDEL	docx pdf/A	1 1	MMSH_ODS_CAT_LASDEL	Dépôt Nakala
2 fichiers	Catalogue à destination de JPODS (corpus Margeride)	docx pdf/A	1 1	MMSH_ODS_CAT_Margeride	Dépôt Nakala

Métadonnées des fichiers son (WAVE), vidéo (MP4), image (TIFF)

17 La structuration des métadonnées est décrite selon les standards archivistiques généraux et propres au traitement documentaire des archives sonores[28]. Les données sonores et audiovisuelles sont décrites dans Calames, selon la logique de description hiérarchique, telles que représentées dans les normes ISAD(G)[29] et EAD. Dans la base de données Ganoub, les métadonnées sont structurées selon le standard international Dublin Core[30]. Dans l'entrepôt de données Nakala, les métadonnées sont structurées selon le standard Dublin Core et Dublin Core qualifié.

Stockage, sauvegarde, partage, conservation, identifiants

18 Les espaces de stockage utilisés en fonction du cycle de vie des données sont cités, de l'étape de la collecte qui correspond à la réception du fonds (serveurs locaux, Huma-Num Box) à celle du traitement et de l'analyse (serveurs locaux et espaces partagés), jusqu'à l'archivage au CINES. La question de la consommation excessive d'énergie pour la conservation des données est soulevée, orientant les choix vers une économie de la donnée, limitant les duplications inutiles, notamment des fichiers de consultation. Le lieu et les conditions de conservation des supports physiques sont indiqués. D'autre part, FAIRiser les données, c'est les rendre partageables, dans le respect des principes éthiques et juridiques. Il est donc précisé qu'à ce titre, la plupart des entretiens sont consultables uniquement sur le site de la MMSH. Les entretiens disponibles en ligne sont cités, ainsi que la majeure partie des papiers du chercheur[31]. Ensuite, il est indiqué que toutes les métadonnées produites par la phonothèque du secteur Archives de la recherche sont accessibles. Elles sont placées dans le domaine public anticipé des Creatives Commons CC0 (pas de droits réservés) et peuvent être agrégées et publiées par n'importe quelle plateforme sous les mêmes termes et utilisées par tout le monde dans n'importe quel projet et sans aucune restriction. Enfin, un projet quasiment abouti de partage des données et métadonnées du fonds Olivier de Sardan est en cours avec le LASDEL, à Niamey, car les données ont été produites dans le cadre de recherches conduites au sein de ce laboratoire.

Aspects éthiques et juridiques

19 Cette rubrique indique le nom du fichier dans lequel sont consignées les coordonnées du déposant, puis sont cités les divers contrats en lien avec le fonds. Une note informe sur le traitement des données personnelles et l'embargo appliqué à certains papiers du chercheur, ainsi que sur les règles de réutilisation des données. La réutilisation est autorisée à des fins de recherche et dans le cadre de l'enseignement, dans la limite juridique liée aux droits des personnes concernées dans les entretiens, après présentation et acceptation du projet.

Responsabilités, ressources

20 Les différentes personnes ayant contribué au traitement du fonds sont toutes citées, quel que soit leur statut, et les éventuelles spécificités du traitement sont relevées. Enfin, les prestataires de numérisation sont identifiés et l'on indique si des contrats spécifiques ont permis de financer ce travail. Ici, le catalogage rétrospectif du fonds a été réalisé avec le soutien d'un financement de l'Agence bibliographique de l'enseignement supérieur et de la recherche.

Prochaines étapes

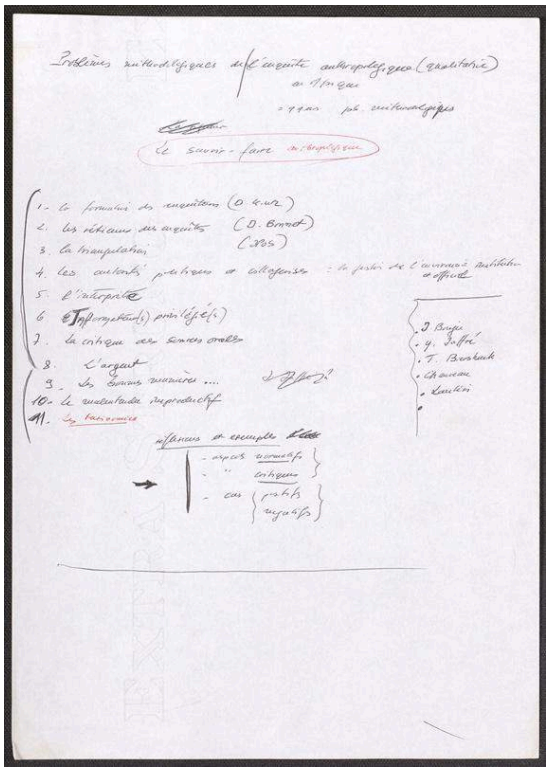
21 Cette dernière partie permet de synthétiser les différentes actions à apporter pour améliorer la FAIRisation des données : par exemple, réaliser le catalogage dans Calames, finaliser les documents juridiques, convertir les fichiers en formats archivables sur le long terme ou établir la correspondance entre les paquets de fichiers versés au CINES et les métadonnées qui l'accompagnent.

22 Puisque le plan de gestion de données respecte les principes FAIR, nous n'omettons pas de préciser le régime juridique sous lequel il est lui-même placé. Au sein du secteur Archives de la recherche & Phonothèque de la Médiathèque de la MMSH, les plans de gestion de données mentionnent que les auteurs et tous ceux qui ont participé à sa rédaction acceptent que tout ou partie du texte puisse être réutilisé et personnalisé pour un autre PGD.

23 Le temps d'élaboration et de rédaction du PGD du fonds Olivier de Sardan a été une période d'échanges constructifs qui ont permis à l'équipe non seulement de vérifier le travail de traitement et son processus, mais aussi de proposer des actions correctrices, transposables à d'autres types de fonds en SHS. Ce travail collectif nous a offert l'occasion de construire des briques communes pour envisager la rédaction de documents génériques. Enfin, son achèvement a permis de réunir les données du chercheur et celles des archivistes dans ce document final, pour produire une nouvelle entité unique nommée « fonds ».

En quoi la rédaction des plans de gestion de données fait-elle évoluer nos façons de travailler ?

Figure 3. Savoir-faire anthropologique/savoir-faire archivistique.



Note de travail de l'anthropologue Jean-Pierre Olivier de Sardan, Documentation diverse, chemise 3. MMSH_ODS_B3_C-07_029[32].

24 Depuis 2018, l'équipe de la phonothèque rédige des PGD rétrospectifs et les publie en ligne[33] depuis 2021 pour accompagner les dépôts de ses archives au CINES, mais aussi pour partager le récit des traitements des fonds. Les équipes ont été fortement invitées à le faire par la TGIR Huma-Num, dans le cadre du consortium Archives des ethnologues clos en décembre 2021[34]. Cet article nous donne l'occasion de revenir sur le travail collectif au sein de ce consortium : il a été inspirant et constructif, car durant les dix années de son existence (2011-2021) ses membres n'ont eu de cesse de travailler à améliorer les bonnes pratiques archivistiques. Ce consortium regroupait neuf centres détenteurs d'archives de chercheurs en sciences humaines et sociales [SHS]. Le consortium était créé avec un triple objectif : soutenir les partenaires pour assurer la pérennisation des données sociales, culturelles et historiques, numériques ou numérisées des chercheurs ; faire en sorte que les partenaires assurent la contextualisation et la documentation des matériaux conservés ; mettre à disposition les données traitées en suivant des standards et formats internationaux dans le respect des règles éthiques et juridiques.

25 Entre 2017 et 2021, Véronique Ginouvès a encadré ce programme avec Fabrice Melka (IMAF). Cette expérience commune et ce partage de compétences ont éclairé notre réflexion et facilité la rédaction des plans de gestion de données. Il ne nous semble pas qu'il existe hors de France un soutien et des financements comparables pour encourager un travail réflexif commun sur les données de la recherche. Or ces ressources sont nécessaires, car s'engager sur un plan de gestion de données rétrospectif n'est pas chose simple. Ce plan traverse tout le processus du cycle de vie de la donnée de façon diachronique, synchronique et proactive. Il implique l'explicitation

de la démarche de production et du traitement de la donnée en prenant en compte les évolutions politiques, nationales et locales et les changements de personnels au fil du temps. Il inclut les transformations technologiques des supports, des machines, des logiciels et des modalités de sauvegarde ainsi que des plateformes et des bases de données, et leurs conséquences sur le traitement et la conservation des archives.

26 L'exemple du fonds Jean-Pierre Olivier de Sardan est intéressant, car, bien qu'il ait été traité sur un laps de temps assez court malgré sa relative ampleur (2,7 mètres linéaires), des changements notables ont interféré dans sa rédaction, entre la date du dépôt du fonds en 2017 et la quatrième version du plan. L'USR 3125 et la phonothèque ont disparu. Cette dernière a intégré, le 1er septembre 2021, le secteur Archives de la recherche de la médiathèque & Phonothèque de la MMSH, et l'unité de service et de recherche (USR) est devenue une unité d'appui à la recherche (UAR) le 1er janvier 2022. Pour complexifier la situation, en décembre 2021, GB Concept, l'éditeur et l'intégrateur du logiciel Alexandrie qui gérait la base Ganoub, a été acheté par la branche Alfeo-division *Knowledge Management* de la société Archimed. Ce rachat a rendu impérative la rétroconversion de l'ensemble des données saisies sur Alexandrie, soit quelque 10 000 notices.

27 L'étape de numérisation intervenue au début du travail sur ce fonds nous a également donné l'occasion de mettre en pratique une nouvelle plateforme technique pour la numérisation de l'audiovisuel à la phonothèque[35]. Une bonne pratique consiste à numériser les documents analogiques sans les lire au préalable, car chaque passage dans l'appareil de lecture abîme la bande. Les cassettes VHS[36] de Jean-Pierre Olivier de Sardan étant peu documentées, nous avons numérisé une dizaine de films qui ont dû être ensuite éliminés. Voilà un enjeu de taille que les rédacteurs d'un PGD ne doivent pas omettre : donner quelques éléments du contexte historique et technologique lorsque cela peut avoir des conséquences sur les supports traités. Au moment où ces films ont été transférés sur VHS, les fictions ou les émissions diffusées à la télévision ou empruntées dans les Vidéos Club étaient copiées sans répit sur ces cassettes. Dans les années 1980, comme le rappelle Françoise Taliano-Des Garets dans son ouvrage *Un siècle d'histoire culturelle en France*[37], les magnétoscopes ont fait leur entrée en masse dans les foyers et la pratique de la copie de film s'est fortement répandue lorsque la redevance sur les magnétoscopes a été abandonnée. Jean-Pierre Olivier de Sardan copiait des films de toutes sortes pour son intérêt personnel. Nous avons donc dû faire une sélection et nous avons recherché les supports d'origine ou d'éventuelles copies de meilleure qualité. C'est ainsi qu'en repérant certains films disponibles sur la plateforme Web de l'Institut agronomique méditerranéen de Montpellier nous avons pu entrer en contact avec le responsable de l'atelier multimédia de cet établissement universitaire. Pierre Arragon, qui est également coréalisateur de trois films de Jean-Pierre Olivier de Sardan, a pu nous fournir une copie en haute définition de ces films. Nous avons également engagé une collaboration scientifique avec le CNRS Images, détenteur des films et des rushes réalisés par Jean-Pierre Olivier de Sardan, pour éventuellement partager les missions de conservation et de documentation de ces documents.

28 Le dernier exemple est précieux pour notre équipe, car il s'agit d'une formidable expérience : les données du chercheur sont partagées avec un autre laboratoire de recherche situé sur un autre continent, là où une partie des données a été produite. Dans une conférence donnée dans le cadre du séminaire « Archives de chercheurs » de la MMSH[38]. Jean-Pierre Olivier de Sardan insiste sur le lien naturel qui lie les données

de la recherche et leur lieu de production. Mais ce lien se rompt lorsque l'anthropologue doit rendre compte de ses travaux dans son propre pays. Le retour de ces données est l'occasion de revenir à ce que Jean-Pierre Olivier de Sardan nomme « réel de référence[39] » et partager ces données avec le Niger, terrain où elles ont été produites, peut contribuer à l'histoire du pays. Une convention de collaboration scientifique a donc été signée au cours de l'été 2022 avec le directeur du LASDEL et le président d'Aix-Marseille Université. Le PGD devra rendre compte de l'enrichissement des données, de leurs modifications, des éventuelles traductions, de l'état des recherches, des droits d'utilisation, etc. qu'occasionnera cette collaboration.

29 Nous avons parlé des pratiques que nous considérons comme majeures : le travail collectif, la précision des contextes politiques, technologiques et culturels, le partage avec celles et ceux qui ont participé à la création de ces archives. Il nous reste à aborder l'enjeu de la transmission. En 2022, l'équipe a été sollicitée par le groupe *Scripto* du réseau national des maisons des sciences de l'homme [RNMSH] pour animer des ateliers sur le thème « Rédiger des PGD rétrospectifs : méthodes, enjeux, perspectives »[40]. Très concrets, ces ateliers suivent le plan de travail d'un PGD. Le premier portait sur la rédaction de son préambule et du contexte de production (4 février 2022). Le deuxième, organisé en collaboration avec l'équipe du programme *ERC HornEast*[41], dressait l'état des lieux de données produites, les typologies et les formats (12 avril 2022). Le troisième était consacré aux questions juridiques et éthiques (7 juillet 2022). Le quatrième, organisé en collaboration avec l'équipe de l'ANR LiPoL[42], a porté sur le nommage et l'organisation des fichiers (9 septembre 2022).

30 Ces ateliers sont l'occasion d'exposer et de confronter nos méthodes et nos pratiques. Chacun de ces échanges, toutes les réflexions de nos collègues et de ceux qui consultent nos archives nous sont nécessaires pour améliorer la rédaction de nos PGD et ses modalités d'écriture. Dans 200 ou 300 ans, nos archives seront consultées dans un tout autre contexte de production et d'usages des documents et nous présumons que la lecture du PGD sera d'une grande aide pour ces futurs lecteurs. Mais, dans l'immédiat, le plan de gestion de données est un outil qui incite à explorer plus avant les étapes du cycle de vie des données et des métadonnées. Contraint à approfondir son enquête pour décrire les contextes de production, l'archiviste est conduit à explorer et à s'approprier les solutions juridiques et techniques qu'elles soient éprouvées ou innovantes.

BIBLIOGRAPHIE

BERT, Jean-François, *Qu'est-ce qu'une archive de chercheur ?*, OpenEdition Press, 2014. http://books.openedition.org/oep/438

COLOMBIÉ, Hélène, « Numériser des archives audiovisuelles à la phonothèque de la MMSH », *Les carnets de la phonothèque,* 9 juin 2017. https://phonotheque.hypotheses.org/21896.

FAMDT. Commission documentation, *Patrimoine culturel immatériel. Traitement documentaire des archives sonores inédites. Guide des bonnes pratiques*, sous la coordination de Claire Marcadé, 2014, 82 p. (3e mise à jour en langue française). https://halshs.archives-ouvertes.fr/halshs-01065125

GINOUVÈS, Véronique, et GRAS, Isabelle (dir), *La diffusion numérique des données en SHS-Guide des bonnes pratiques éthiques et juridiques*, Presses universitaires de Provence, 2018, https://hal-amu.archives-ouvertes.fr/hal-01903040.

MONAGHAN, Peter, "How Will an Important Archive Ever Recover?", *Moving Image Archive News*, 10 juin 2021. https://web.archive.org/web/20220903065726/https://www.movingimagearchivenews.org/how-will-an-important-archive-ever-recover

OLIVIER DE SARDAN, Jean-Pierre, « Archives de la recherche : Pourquoi et pour qui un chercheur dépose-t-il ses archives ? », *Carnet Hypothèses Archives de la recherche & phonothèque*, 4 mai 2021. https://phonotheque.hypotheses.org/33401

OLIVIER DE SARDAN, Jean-Pierre, *La rigueur du qualitatif. Les contraintes empiriques de l'interprétation socio-anthropologique*, Louvain-la-Neuve, Academia-Bruylant, 2008, 372 p.

STEYN, Daniel, "South Africa: Restoring Cape Town University's burnt archives will take years", *ZAM*, 19 juillet 2022. https://web.archive.org/web/20220907075047/ https://www.zammagazine.com/politics-opinion/1506-south-africa-restoring-cape-town-university-s-burnt-archives-will-take-years

TALIANO-DES GARETS, Françoise, « Chapitre 7. Les années Mitterrand, culture et communication », dans *Un siècle d'histoire culturelle en France*, éd. par Françoise TALIANO-DES GARETS, Armand Colin, 2019, p. 225-262.

ZEYTLIN, David, "Archiving ethnography? The impossibility and the necessity", *Ateliers d'anthropologie*, n° 51, 2022, DOI : https://doi.org/10.4000/ateliers.16318.

NOTES

1. Le PGD est obligatoire pour tous les programmes d'Horizon Europe 2021-2027 ; il l'était dès 2014 pour les programmes H2020.
2. *Findable Accessible Interoperable Reusable* [Facile à trouver, Accessible, Interopérable, Réutilisable].
3. Données de la Recherche : Apprentissage Numérique. https://doranum.fr/
4. Optimiser le Partage et l'Interopérabilité des Données de la Recherche. https://opidor.fr/
5. L'*Encoded Archival Description [EAD]* est un standard d'encodage des instruments de recherche archivistiques fondé sur le langage de balisage XML.
6. Site officiel de la norme EAD : https://www.loc.gov/ead.
7. En stage de septembre 2020 à avril 2021, Annick Boissel a réalisé un mémoire de fin d'études sur *Les enjeux d'un plan de gestion de données réalisé a posteriori sur les archives de chercheurs en sciences sociales, à partir du fonds Jean-Pierre Olivier de Sardan*. Titulaire du diplôme de licence professionnelle d'Archiviste audiovisuel du Conservatoire national des arts et métiers, elle exerce aujourd'hui en tant qu'archiviste indépendante.
8. Les articles de V. Ginouvès sont en ligne sur HALSHS : https://cv.archives-ouvertes.fr/vginouves.
9. À partir de septembre 2021, Émilie Groshens, archiviste au CNRS, a rejoint Véronique Ginouvès pour la création de ce service.

10. La *WayBack Machine* en garde trace : https://web.archive.org/web/2022/hhtp://www.mmsh.fr.

11. David Zeytlin, 2002.

12. Peter Monaghan, 10 juin 2021.

13. Daniel Steyn, 19 juillet 2022.

14. Il s'agit de l'infrastructure Adonis devenue ensuite la TGIR Huma-Num [Très grande infrastructure de recherche des Humanités numériques], puis l'IR* Huma-Num. Elle a pour mission de construire, avec les communautés des SHS, une infrastructure numérique de niveau international leur permettant de développer, de réaliser et de préserver sur le long terme les données et outils de leurs programmes de recherche, dans un contexte de science ouverte et de partage des données : https://www.huma-num.fr/. Voir à ce propos, dans cet ouvrage, l'article de Stéphane Pouyllau et Nicolas Sauret.

15. Le CINES [Centre informatique national de l'enseignement supérieur] est un établissement public à caractère administratif national (EPA), situé à Montpellier et placé sous la tutelle du ministère de l'Enseignement supérieur, de la Recherche et de l'Innovation [MESRI] : https://www.cines.fr/.

16. Pour mieux connaitre le parcours et les travaux de Jean-Pierre Olivier de Sardan, voir http://www.calames.abes.fr/pub/ms/FileId-3805.

17. Le LASDEL est un laboratoire nigérien et béninois de recherche en sciences sociales, menant des travaux qualitatifs à base empirique (enquêtes de terrain de type socio-anthropologique), fondé en 2001 à Niamey (Niger), puis étendu en 2004 à Parakou (Bénin). Il regroupe, en 2022, 33 chercheurs africains dont onze sont titulaires d'un doctorat et quatre sont habilités à diriger des thèses ; site : https://web.archive.org/web/sitemap/https://www.lasdel.net.

18. Les métadonnées de l'enregistrement sonore d'un entretien sont constituées par des données (informations) réunies par l'archiviste, qui sont celles énoncées dans le guide de bonnes pratiques *Patrimoine culturel immatériel. Traitement documentaire des archives sonores inédites*, 2014 (3e mise à jour en langue française), mis à jour en 2021 en langue espagnole.

19. Massachusetts Institute of Technology : https://libraries.mit.edu/data-management/plan/why.

20. https://library.stanford.edu/research/data-management-services/data-management-plans.

21. https://researchdata.ox.ac.uk/home/managing-your-data-at-oxford/data-management-planning.

22. Les éléments principaux qui doivent figurer dans un PGD sont : les informations administratives ; la description des données ; la documentation, les métadonnées, les standards ; le stockage et le partage ; la conservation et l'archivage ; la sécurité des données ; les aspects éthiques, les responsabilités, les coûts. Source : https://doranum.fr/plan-gestion-donnees-dmp/plan-de-gestion-des-donnees-fiche-synthetique_10_13143_cgv4-0k53/.

23. Le PGD du fonds Jean-Pierre Olivier de Sardan est librement accessible sur https://phonotheque.hypotheses.org/files/2021/12/PGD-ODS-MMSH-V4a.pdf.

24. Lien vers le catalogue du fonds Jean-Pierre Olivier de Sardan : http://www.calames.abes.fr/pub/#details?id=FileId-3805.

25. Collection des données textuelles du fonds Jean-Pierre Olivier de Sardan : https://nakala.fr/collection/10.34847/nkl.c6e9i4j8.

26. Lien vers les données sonores et audiovisuelles du fonds Jean-Pierre Olivier de Sardan, documentées sur la plateforme Ganoub à partir de la *Wayback Machine* puisque la plateforme est destinée à disparaître : https://web.archive.org/web/20220907080447/http://phonotheque.mmsh.huma-num.fr/dyn/portal/index.xhtml?page=alo&aloId=13239.

27. Carnet de recherche *Archives de la recherche & Phonothèque* : https://phonotheque.hypotheses.org/, plateforme *Transcrire :* https://transcrire.huma-num.fr/, Consortium *Archives des ethnologues :* https://ethnologia.hypotheses.org/.

28. *Patrimoine culturel immatériel. Traitement documentaire des archives sonores inédites. Guide des bonnes pratiques*, 2014, 82 p. (3e mise à jour en langue française) : https://halshs.archives-ouvertes.fr/halshs-01065125 Mis à jour en 2021 en langue espagnole : https://halshs.archives-ouvertes.fr/halshs-03475977.
29. Norme générale et internationale de description archivistique ISAD(G) [International Standard Archival Description – General) : https://www.ica.org/fr/isadg-norme-generale-et-internationale-de-description-archivistique-deuxieme-edition.
30. Standard international Dublin Core : https://www.dublincore.org/specifications/dublin-core/.
31. La démarche et les réflexions de la phonothèque de la MMSH sur ces questions éthiques et juridique est précisée dans l'ouvrage collectif, guide de bonnes pratiques sur ces questions (2018) et dans le carnet de recherche *Questions d'éthique et de droit* : https://ethiquedroit.hypotheses.org.
32. ID Nakala : 10.34847/nkl.bbab1otw/4c064921d977847893d449829c42bdc785a2098a.
33. Le premier plan rédigé en 2018 était celui de l'ANR Colostrum ; il a été publié en 2021. À partir de cette date, le secteur Archives de la recherche de la médiathèque & Phonothèque a commencé à publier régulièrement des PGD au fur et à mesure du traitement des fonds : https://phonotheque.hypotheses.org/33813.
34. Soit, entre 2017 et 2020, le CEH, Centre d'études himalayennes ; l'IMAF, Institut des mondes africains ; le CRBC, Centre de recherche bretonne et celtique ; le LESC, Laboratoire d'ethnologie et de sociologie comparatives ; la MSHS [Maison des sciences de l'homme et de la société] de Poitiers (USR 3565) ; la phonothèque de la MMSH et, jusqu'en 2019, les laboratoires du LAS, CREDO et IRASIA.
35. Hélène Colombié, 9 juin 2017.
36. Video Home System. Format de cassette vidéo destiné à un usage par le grand public.
37. Françoise Taliano-Des Garets, 2019.
38. Jean-Pierre Olivier de Sardan, 4 mai 2021.
39. Jean-Pierre Olivier de Sardan, 2008.
40. Les diaporamas des ateliers sont en ligne sur : https://phonoteque.hypotheses.org/3911.
41. Pour cet atelier, l'archiviste Maryasha Barbé s'est jointe à nous ; le plan de gestion de données qu'elle a publié est accessible en ligne : https://horneast.hypotheses.org/2195.
42. https://anr.fr/Projet-ANR-19-CE27-0024.

AUTEURS

ANNICK BOISSEL

Archiviste indépendante

VÉRONIQUE GINOUVÈS

Responsable du secteur Archives de la recherche, médiathèque de la Maison méditerranéenne des sciences de l'homme (AMU-CNRS)

Partie 4 - Les contenus des réseaux sociaux

Le temps des plateformes : enjeux, différences et complémentarité de l'archivage des médias sociaux numériques à la Bibliothèque nationale de France et à l'Institut national de l'audiovisuel

Alexandre Fay, Jérôme Thièvre et Valérie Schafer

Introduction

1 La mainmise d'Elon Musk sur Twitter à l'automne 2022[1] et le mouvement qui s'ensuit de départ d'une partie des usagers vers le fédiverse[2], notamment vers Mastodon, ainsi que leur demande d'effacement de leurs *tweets*[3] et données, témoignent, s'il en était besoin, de l'éphémérité des données numériques et, plus largement, des réseaux sociaux numériques [RSN]. Il y a des antécédents de disparition de RSN ou de leurs contenus, par exemple celui de Geocities (Milligan, 2017), très populaire dans les années 1990, ou encore ceux de MySpace (Gomez-Mejia, 2018) et Vine[4]. Les RSN, qui se sont imposés depuis la décennie 2000 comme des plateformes importantes d'information et de communication, sont aussi des témoins de leur temps, notamment lors de crises, comme les printemps arabes (Tufekci, 2019), le mouvement *Black Lives Matter* ou la pandémie de la Covid-19. Ils y jouent un rôle important, par exemple lors du mouvement #metoo, ou de la coordination des Gilets jaunes en France.

2 Inégalement préservés, les RSN entrent dans les enjeux d'archivage du web de manière plus ou moins précoce. La Bibliothèque nationale de France [BnF] a archivé Dailymotion entre 2007 et 2013, Twitter depuis 2012 et YouTube à partir de 2017. L'Institut national de l'audiovisuel [Ina] archive YouTube et Vimeo depuis 2009-2010 et Twitter depuis 2014. Si toutes les plateformes ne bénéficient pas encore d'une activité de captures

régulières, de nouveaux projets de collecte améliorent la couverture. Ainsi la BnF gère des collectes Instagram depuis 2020 et TikTok depuis 2022.

3 Quels sont les enjeux de conservation, mais aussi d'exploitation de ce patrimoine nativement numérique ? Quels sont les tendances, complémentarités, défis partagés, mais aussi spécificités propres à ces fonds ? Après avoir dressé un panorama rapide du développement international de l'archivage des RSN et des défis qu'ils posent, les approches en France de la BnF et de l'Ina sont mises en regard et en perspective. Elles permettent de s'intéresser dans une troisième partie à des collections plus particulières, à l'instar de fonds dédiés aux attentats, aux mouvements sociaux français ou encore à la crise de la Covid-19, pour mettre en valeur des cas d'usages et des recherches en cours sur ces collections inédites, diverses, mais complémentaires, qui invitent à penser ensemble les nouveaux paradigmes de préservation et de recherche.

Les données des réseaux sociaux numériques, un patrimoine nativement numérique en constitution

4 L'archivage du web n'a plus à démontrer son importance (Brügger, 2018 ; Musiani *et al.*, 2019). Depuis les initiatives, au milieu des années 1990, de la fondation Internet Archive ou encore de la Bibliothèque nationale d'Australie (Hegarty, 2022), il a peu à peu gagné bien des pays et été entrepris par des institutions souvent liées au monde des bibliothèques et parfois encadré par des lois sur le dépôt légal, comme en France en 2006 ou en Grande-Bretagne en 2013. L'archivage du web a connu des changements, mais il est aujourd'hui plutôt stabilisé et bien organisé, tout en laissant encore de la place à des évolutions concernant par exemple les interfaces, les modes de recherche et d'accès aux archives du web, les métadonnées, etc. Les RSN ont évidemment aussi attiré l'attention des archivistes du web sans que l'on puisse réellement parler ici de stabilisation des pratiques. Elles sont en constante adaptation, tandis que les RSN se développent et créent de nouveaux espaces à considérer, avec, par exemple, la montée des usages de TikTok depuis son lancement en 2016. Chacune de ces plateformes a ses caractéristiques, des approches plus textuelles ou visuelles, des marqueurs propres (*retweet*, *like*, etc.), des utilisateurs différents (ceux de Facebook, Twitter et TikTok ne se recoupent pas forcément), et des usages spécifiques de l'hyperlien, des renvois, des fils de discussion, des partages. Leurs conditions d'usage ou d'accès peuvent régulièrement évoluer, ce qui implique de reconfigurer les modes de collecte : les API[5], c'est-à-dire les modalités d'accès aux interfaces de programmation permettant de récupérer des données des RSN, peuvent changer dans le temps et, avec une tendance générale à la limitation du nombre de données récupérables, induisent une baisse des volumes collectés. Au niveau des infrastructures, la durée des collectes – en particulier pour la vidéo – et les besoins de stockage restreignent la production. La collecte de YouTube par la BnF illustre bien cette double tendance : les limitations de l'API youtube-dl ne permettent pas de récupérer l'ensemble des métadonnées associées aux vidéos pour les chaînes les plus volumineuses et la durée ainsi que le poids des vidéos augmentent les besoins matériels en termes de stockage et de nombre de robots employés.

5 Surtout, ces réseaux sont « bavards » et produisent des masses de données qui nous renvoient, comme le web l'avait fait précédemment, l'impossibilité de l'exhaustivité. La Bibliothèque du Congrès aux États-Unis en a fait l'expérience quand elle a passé un accord avec Twitter en 2010 pour récupérer ses archives depuis 2006, puis a décidé en

2017 de revenir à une politique plus sélective, face à des difficultés techniques de stockage et d'accès, qui ne sont toujours pas résolues (LoC, 2017). Si l'exhaustivité est impossible, comprendre certains mouvements sans prendre en considération ces espaces d'expression l'est tout autant. C'est pourquoi les archivistes du web recherchent, à défaut d'exhaustivité, une forme de représentativité permettant de saisir les vibrations (Boullier, 2015) parfois intenses des RSN. Bien sûr, l'archivage des RSN n'est pas propre aux crises et à l'affût des seuls signaux forts. Les campagnes électorales, qui font l'objet de collectes précoces à la BnF[6], invitent à suivre l'expression politique sur les RSN en complément de l'archivage des sites des candidats. Cependant, lors des attentats de 2015[7], face à un événement soudain dont l'impact affecte l'ensemble de la société, l'adaptation passe par la mise en place d'une collecte en urgence et la prise en compte de la diversité des voix qui s'expriment. Ce qui ressort alors, c'est le besoin de collecter, de manière parfois un peu idéalisée, la diversité des expressions personnelles. Par exemple, l'Ina suit, dès 2014, des comptes Twitter de journalistes ou de chaînes dans le cadre de son périmètre dédié à l'audiovisuel, mais, avec les attaques terroristes, c'est bien un besoin d'élargir la collecte aux voix ordinaires et aux réactions « par le bas » qui se manifeste.

6 Les RSN permettent de dessiner des temporalités très fines : on pense par exemple à la qualification des attentats du Bataclan d'abord mentionnés comme #tirs puis comme #attentats dans les *hashtags*[8] (Schafer *et al.,* 2019), ou encore à des enjeux mémoriels transnationaux comme ceux analysés lors des commémorations de la Grande Guerre par Frédéric Clavert (2018). Les collectes des #deconfinementJ1, J2, etc. par l'Ina procèdent aussi de ces temporalités très précises au sein d'un mouvement plus large, celui de la crise de la Covid-19 (Schafer *et al.*, 2020). Reste une réalité : les données des RSN ne peuvent pas toutes être conservées et, lors des attentats de 2015 par exemple, Périscope[9], qui capte des images des événements en direct, n'est pas préservé. De même, lors de la crise de la Covid, par ailleurs particulière par sa durée, des choix sont faits. Si la Bibliothèque royale du Danemark se lance dans la collecte de traces de TikTok ou Reddit (Schostag, 2020), ce sont essentiellement Facebook ou encore Twitter qui concentrent les efforts des institutions d'archivage européennes. Il y a plusieurs explications à ces choix nécessaires : il y a bien sûr des enjeux techniques, que l'on passe par un robot spécifique ou par les API des RSN, plus ou moins ouvertes. Ensuite, les logiques de flux sont complexes à capter. Les robots ne sont pas toujours adaptés, les paradigmes d'archivage parfois difficiles à cerner et ils doivent combiner plusieurs approches (par *hashtag*, par compte, etc.). Il faut aussi choisir entre capturer essentiellement le contenu et des données (approche de l'Ina, cf. partie suivante) ou tout un environnement numérique (approche de la BnF). Il faut aussi prendre en compte la question des données privées et publiques, qui soulève certes des enjeux éthiques, mais laisse aussi dans l'ombre une partie des traces numériques des comptes privés. Et sur les RSN qui entrent dans le cadre des collectes, des contenus peuvent être omis, car les choix de capture à un instant T sont extrêmement complexes : il faut anticiper des tendances, mener une veille en temps réel.

7 Il y a également des enjeux institutionnels : des enjeux de stockage, des enjeux de budget, des enjeux de disponibilité des équipes, car l'archivage repose sur des choix humains et un intense investissement des équipes. Archiver les RSN, et d'autant plus quand il faut le faire en urgence, impose de définir une politique d'archivage, des choix de *hashtags*, de comptes, etc., et aussi de collecte, cette dernière pouvant se faire au détriment d'une autre. En outre, l'énorme investissement demandé n'est pas toujours

rentable, au regard de la qualité du résultat. À moyen terme, l'expérimentation de nouveaux outils, moins automatisés comme Webrecorder/Browsertrix[10], pourrait permettre de contourner certains blocages techniques que relève Ben Els, archiviste à la Bibliothèque nationale du Luxembourg : « Facebook actively tries to block crawler robots and as a result you have to collect much more data to obtain reliable results. So, for the cost of one capture on Facebook you can archive a regular website many times and the capture of Facebook may even be unusable, with videos not working, for example » (Schafer et Els, 2020)[11].

8 Il n'en reste pas moins que les RSN sont aujourd'hui au centre des préoccupations de bien des institutions, mais aussi d'initiatives mêlant une pluralité d'acteurs, dont des activistes et des chercheurs, que ce soit dans le cas de *Documenting the Now*[12] qui suit le mouvement *Black Lives Matter* ou encore de SUCHO[13] [*Saving Ukrainian Cultural Heritage Online]*, initiative lancée au début du conflit contre l'Ukraine pour préserver son patrimoine culturel nativement numérique. Dans l'urgence, il s'agit de collecter. Quand l'urgence est moindre, les institutions et les parties prenantes peuvent aussi se concerter, mener un travail prospectif comme dans le cas du projet belge BESOCIAL (Vlassenroot *et al.*, 2022) qui a fait un tour d'horizon des pratiques européennes de préservation des traces des RSN. Celui-ci n'a pas manqué de s'intéresser aux pratiques déjà en place, et notamment aux approches de la BnF et de l'Ina, qui représentent des voies différentes et complémentaires d'approche de l'archivage des RSN.

Les approches de la BnF et de l'Ina

9 À l'Ina, les premières collectes sont engagées dès la fin des années 2000, ce qui est précoce au regard des autres institutions européennes qui ont préféré archiver les médias et les plateformes de partage ou d'hébergement de médias, avant d'archiver les RSN proprement dits.

L'archivage des médias des plateformes de partage ou d'hébergement de médias à l'Ina depuis la fin des années 2000

10 Le périmètre du dépôt légal du web de l'Ina étant par définition centré autour de la radio et de la télévision, il est apparu essentiel, dès le début de sa mise en œuvre, d'être en mesure de collecter les médias publiés sur le web. Or, en simplifiant la publication de contenus audiovisuels et en rendant gratuit leur hébergement, l'apparition des plateformes de partage de vidéos a permis leur massification spectaculaire tout comme celle de leurs usages. Cette concentration des vidéos sur quelques plateformes aurait pu conduire à un risque important, pour les archivistes du web, de voir ces contenus leur échapper dans le cas où les plateformes empêcheraient leur collecte. C'est plutôt l'inverse qui s'est produit, car les contenus sont restés, jusqu'à aujourd'hui, accessibles et le nombre restreint de plateformes a permis de rationaliser les procédés de collectes des médias sur le web.

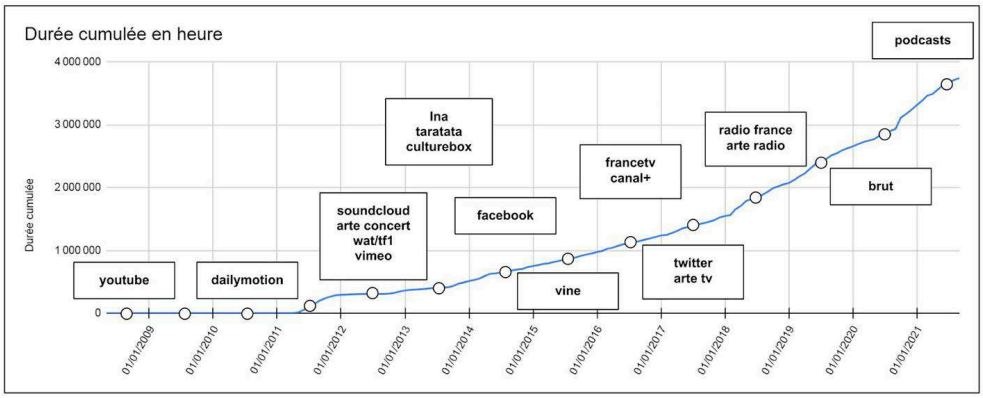
11 Face à la montée en puissance des plateformes de partage de médias, telle que YouTube, à la fin des années 2000, puis des plateformes de *replay* et de *podcasts* quelques années plus tard, l'Ina s'est donc investi dans la création d'outils de collecte adaptés à chacun de ces modes de publication. Ces collectes spécifiques sont venues compléter la collecte des sites web de l'audiovisuel français réalisés par les robots de moissonnage

des pages web. En effet, les robots de collecte de sites web ou *web crawlers* s'étaient rapidement révélés peu adaptés à la collecte des plateformes de partage de médias, et ce pour plusieurs raisons. Les ressources audio et vidéo sont rendues disponibles dans différents formats et leurs URLs[14] sont le plus souvent construites dynamiquement ou nichées dans des strates de codes que les *web crawlers* ne peuvent pas interpréter. Les modes de publication de médias ont évolué à plusieurs reprises afin de s'adapter aux usages, notamment nomades avec, par exemple, un découpage des pistes audio et vidéo en de multiples segments afin d'adapter, durant la lecture, la qualité des vidéos à la bande passante fluctuante des réseaux mobiles. Ces évolutions constituent des obstacles qu'un *web crawler* générique ne peut pas franchir.

12 Les volumes de vidéos disponibles et le poids de ces objets ont aussi des implications sur le fonctionnement des robots collecteurs. Dans les rares cas où ils seraient capables de découvrir les ressources vidéo, comment s'assurer de ne collecter et stocker qu'une seule fois ces ressources coûteuses ? Un mécanisme de déduplication assure l'unicité des contenus lors de l'archivage, mais ne permet pas de s'assurer d'un téléchargement unique de ces ressources et remet donc en question une utilisation raisonnable de la bande passante des plateformes archivées.

13 En plus de ces barrières inhérentes au fonctionnement des *crawlers*, peuvent venir s'ajouter des verrous mis en place par les plateformes elles-mêmes, afin de gêner ou de freiner la collecte des médias (c'est par exemple le cas sur YouTube depuis 2016). Aussi, l'Ina a fait le choix de développer des modes de collectes propres à chacune des plateformes dans un robot de collecte de médias du web, ou *media crawler*. Cela permet d'assurer pour chaque plateforme une collecte unique de chaque média dans un format explicitement identifié et de récupérer également les métadonnées textuelles telles que le titre, la description, l'auteur, la date de publication, etc. Ce choix impose, en conséquence, de suivre les évolutions techniques régulières de chacune des plateformes afin d'adapter le code de ce robot et le maintenir fonctionnel.

Figure 1. Mise en place de collectes de médias par plateforme.



14 Sur les plateformes de partage YouTube, Dailymotion et Vimeo, près 10 000 comptes ou chaînes ont été identifiés par les documentalistes de l'Ina, selon différentes modalités :

- le média a été découvert dans une ressource web archivée (page ou *tweet* par exemple) ;
- le média est publié sur une plateforme de partage par un compte ou une chaîne qui ont été sélectionnés pour archivage ;
- le média est publié sur une plateforme dont tout le catalogue est collecté.

15 Ces comptes regroupent des contenus directement liés à l'audiovisuel français, les chaînes de radio et de télévision nationales, bien sûr, mais aussi les chaînes locales qui se trouvent hors périmètre du dépôt légal de la radio et de la télévision. Cette sélection comprend aussi des comptes de vidéastes web ou « youtubeurs » français qui rencontrent un succès populaire ou critique auprès du public français. Les nouvelles publications de ces comptes sont identifiées et collectées quotidiennement et vingt millions de vidéos ont été collectées sur ces trois plateformes depuis 2009.

16 Les réseaux sociaux comme Twitter et Facebook sont aussi hébergeurs d'une partie des vidéos partagées par leurs utilisateurs. Les collectes médias sur ces deux plateformes ont permis d'archiver près de treize millions de contenus, identifiés à partir de pages web ou de *tweets* archivés.

17 Le succès des plateformes de partage a profondément modifié les modes de consommation des contenus audiovisuels et conduit les diffuseurs historiques que sont la télévision et la radio à publier leurs contenus sur le web. Autour de 2016, les sites web des chaînes de télévision ont évolué, passant d'une vitrine des programmes vers un catalogue de contenus majoritairement de *replay*, mais aussi de création exclusive au numérique. Le *media crawler* est aussi utilisé pour collecter les médias de certaines de ces plateformes. Ainsi, France.tv et Arte.tv font l'objet d'une collecte intégrale et près d'un million de contenus ont été collectés pour ces deux plateformes.

18 On observe le même phénomène pour les radios qui proposent de plus en plus d'écoutes en *podcast* de leurs émissions et s'investissent aussi dans la création de contenus exclusivement web. Les *podcasts* de Radio France et Arte Radio sont collectés depuis 2018, ce qui représente aujourd'hui 1,6 million d'émissions ou épisodes. En 2021 une approche différente a été entreprise afin de couvrir un périmètre de collecte des *podcasts* plus large, en utilisant les flux de syndication RSS comme sources de collecte. En effet, les *podcasts* reposent encore en grande partie sur ce standard web qui a l'avantage d'être bien spécifié et ouvert et qui permet de découvrir les nouveaux épisodes de programmes audio avec leur flux média et leurs métadonnées. Près de 5 000 *podcasts* (et plus de 800 000 épisodes) ont été ainsi sélectionnés, car ils sont soit en lien avec l'audiovisuel français, soit populaires en France.

L'archivage des réseaux sociaux numériques à l'Ina depuis les skyblogs

19 Lié à la radio française Skyrock, le site skyblog.com a été un précurseur des réseaux sociaux en France en proposant une plateforme de blogs à partir de 2002. Celle-ci évolue au cours de la décennie et propose des fonctionnalités caractéristiques des réseaux sociaux, telles que les listes d'amis, la messagerie et la gestion de profils. En 2008, ce site devient le septième réseau social le plus populaire dans le monde, avant de tomber progressivement dans l'oubli dans les années 2010. Les archives de ce site représentent une source précieuse pour qui s'intéresse par exemple aux usages numériques de la jeunesse des années 2000.

20 À partir des années 2010, les RSN supplantent progressivement les sites personnels et les blogs. Facebook compte aujourd'hui encore plus de 30 millions d'inscrits en France et Twitter est une plateforme très prisée par les politiques, les médias et la presse. Ces deux plateformes sont aussi particulièrement utilisées pour échanger en temps réel

autour des programmes télévisés, ce qui apporte un éclairage important sur leur réception par les téléspectateurs. L'Ina s'est intéressé à leur archivage en 2012 avec, comme objectif, de pouvoir collecter les publications en rapport avec les chaînes de télévision et de radio, leurs programmes et personnalités.

21 Plusieurs difficultés ont conduit à opter pour une approche « sur mesure », similaire à celle des plateformes de partage précédemment évoquée. Tout d'abord, le regroupement de multiples publications dans une même page pose problème lors de la collecte et de la consultation en rendant impossible une recherche plein-texte précise. Le nombre de publications et leur volatilité peuvent être source de collectes redondantes pour les comptes publiant peu, mais aussi de collectes partielles pour les comptes publiant beaucoup. L'utilisation de technologies mal supportées par les *web crawlers*, comme le chargement dynamique des publications ou la nécessité de s'authentifier, peuvent rendre inopérantes les collectes. Enfin, la mise en place progressive par ces plateformes de barrières anti-robot, afin notamment d'éviter les aspirations massives de publications, empêche aussi les collectes légitimes des institutions d'archivage du web.

22 L'alternative a été de se tourner vers les API publiques que mettent à disposition les RSN. Ces API présentent l'avantage de pouvoir identifier, requêter et collecter les publications selon des critères pertinents et en adéquation avec les besoins de préservation :

- publications d'un compte,
- publications contenant certains mots-clés ou *hashtags*.

23 Elles résolvent en grande partie les problématiques d'efficience, avec la possibilité de récupérer en grand nombre et rapidement des publications. Celles-ci sont de plus accessibles dans un format structuré, dont le modèle est relativement stable dans le temps.

24 Les premières collectes sur Twitter ont été menées en 2014. Le périmètre de collecte sur Twitter peut être défini comme :

- toute publication émise par un compte en lien avec l'audiovisuel français, notamment les chaînes, leurs émissions et les personnalités qui produisent, animent ces émissions ou y ou participent,
- toute publication qui fait référence à une chaîne ou à une émission.

25 Les *tweets* sont collectés par un robot spécifique qui utilise les API publiques de Twitter dans leur version 1.1. L'API Timeline permet de récupérer les dernières publications émises par des comptes, le robot de collecte interroge quotidiennement cette API pour chacun des 15 000 comptes identifiés afin d'archiver leurs nouvelles publications. L'API Filtered Stream donne accès à un flux en temps réel de publications correspondant à une requête textuelle. Cette API est utilisée pour collecter les *tweets* contenant au moins un des 1 500 *hashtags* suivis quotidiennement. Plus de 3 000 *hashtags* ont été suivis depuis le début des collectes, dont 1 800 en lien direct avec l'audiovisuel et 1 100 autour d'événements médiatiques à fort retentissement sur les réseaux sociaux. Enfin l'API Search, qui permet de rechercher des *tweets* par requêtes textuelles sur les sept derniers jours, est utilisée pour collecter les *tweets* contenant la mention d'un des 15 000 comptes identifiés, mais aussi de pouvoir rattraper un retard de nomination d'un *hashtag*.

26 Toutefois, ces API n'offrent pas un accès illimité aux données de Twitter. Afin de pouvoir y accéder, il est nécessaire de posséder un compte Twitter et d'obtenir une autorisation, dont les modalités d'obtention ont beaucoup évolué ces dernières années, notamment à la suite du scandale Cambridge Analytica[15] en 2016, qui a imposé aux plateformes de réseaux sociaux un plus grand contrôle sur les données de leurs utilisateurs. L'accès aux publications peut être limité en nombre et dans le temps. Ainsi, l'API Timeline permet d'accéder aux 3 200 dernières publications d'un compte, l'API Search est limitée à une profondeur temporelle d'une semaine et l'API Filtered Stream ne permet de suivre que 400 *hashtags* par compte et est limitée à un volume de 50 *tweets* par seconde. Afin de pouvoir assurer la collecte du périmètre, nous devons utiliser aujourd'hui plusieurs comptes.

27 Le périmètre des collectes sur Twitter a été élargi, à partir de 2015, à de grands événements médiatiques. Une première collecte de ce type a été déployée en urgence durant l'attentat dans les locaux de *Charlie Hebdo*[16]. Ce type de collecte a été maintenu puis étendu à d'autres événements ayant un fort impact sur les réseaux sociaux et dans les médias, comme le mouvement #MeToo et #BalanceTonPorc, les grands événements sportifs, les élections présidentielles, le mouvement de #GiletsJaunes ou bien la crise de la Covid-19.

28 Des collectes spécifiques sont aussi menées autour de partenariats ou de projets de recherche. Cela a été le cas avec le projet RwandaMap2020 autour de la mémoire du génocide des Tutsis au Rwanda pour lequel une collecte a été réalisée en 2019, ou pour le projet BodyCapital avec une collecte des *tweets* relatifs au procès du Mediator[17].

De la collecte des sites web à celles des médias sociaux à la BnF

29 Le dépôt légal du web à la BnF couvre l'ensemble des sites web et des contenus des plateformes ne relevant pas du périmètre de l'Ina. Pour produire une archive représentative de l'Internet français et des sites hébergés en France, la BnF réalise chaque année une collecte large qui a atteint le chiffre de 5,9 millions de sites en 2022. L'intérêt pour le web social, d'abord les blogs, puis les nouveaux médias sociaux, s'est développé au cours des années 2010 avec la mise en place de collectes ciblées consacrées à un événement ou un thème. La BnF a commencé à collecter spécifiquement les réseaux sociaux à l'occasion des élections présidentielles de 2012 et la plateforme de diffusion de vidéos Dailymotion dès 2007. Cependant, contrairement à l'Ina, la BnF s'est toujours appuyée sur le robot de collecte Heritrix[18].

30 La sélection des médias sociaux s'effectue en continuité avec les autres collectes ciblées. Elle est réalisée par les correspondants de la BnF, répartis au sein des différents départements de la bibliothèque et des bibliothèques partenaires. Cette organisation assure une large couverture thématique et géographique. L'application BCweb, développée par la BnF, est utilisée par les correspondants pour saisir et documenter les sélections. Elle offre une vue documentaire cohérente par collecte. La sélection des médias sociaux n'est pas séparée de celles des sites. Cela facilite la prise en compte de l'ensemble des médias composant l'espace numérique d'une personnalité ou d'une institution : sites web, chaîne YouTube, comptes Twitter, Instagram ou TikTok. Les périodes électorales sont souvent l'occasion d'élargir le périmètre des plateformes couvertes pour suivre l'élargissement des pratiques numériques des candidats et leur stratégie multicanale. Des chercheurs participent également aux sélections dans le

cadre de partenariats. Dans le cadre de son DataLab, la BnF propose un service de collecte à la demande dit « Corpus de recherche ». Ce dernier a donné lieu en particulier à deux collectes spécifiques d'un corpus de chaînes YouTube dans le cadre de projets de recherche[19].

31 Techniquement, la collecte des réseaux sociaux ne diffère pas fondamentalement de la collecte des sites web. Le robot collecte les contenus publiés sous la forme de pages web et produit des fichiers au format standardisé WARC[20]. Le résultat des collectes est consultable à partir de l'application Archives de l'Internet comme les autres archives web produites, permettant ainsi une navigation fluide entre les publications et les pages citées dans les publications. La BnF s'appuie sur des outils partagés et maintenus dans le cadre de la communauté internationale des archivistes du web, notamment *via* le Consortium international pour la préservation d'Internet [IIPC], lui-même à l'origine du format WARC.

32 La collecte par robot requiert une organisation et une planification des flux. La configuration technique détermine la profondeur et la fréquence de collecte. Dans le cas des réseaux sociaux, la réactivité est un élément essentiel, notamment lorsqu'il s'agit de capturer au plus près les réactions à un événement. Ainsi pour Twitter, et précédemment pour Facebook, deux collectes sont lancées quotidiennement pour une profondeur de récupération limitée à 20 *tweets* par capture. Pour Instagram, TikTok et YouTube, les collectes s'effectuent sur un rythme ponctuel à des moments précis de l'année.

33 Outre les configurations techniques, chaque plateforme demande un travail d'instruction, notamment pour guider au mieux le robot vers les contenus ciblés. Cette phase donne lieu à des tests pour définir des filtres positifs et négatifs, qui constituent une base de connaissance tenue à jour en fonction de l'évolution technologique des plateformes. Pour autant, la principale difficulté rencontrée reste avant tout le risque de blocage des robots. Pour le contourner, il est nécessaire d'obtenir directement un accord de la plateforme (c'est le cas pour YouTube) ou de mettre en place des solutions de contournement en collectant par exemple une version mobile du site (cela a été le cas à plusieurs reprises pour Facebook) ou une application tierce (la collecte Instagram s'effectue *via* un agrégateur de contenu). Néanmoins, face à des blocages répétés et en l'absence de solution technique, il peut être décidé de ne plus collecter certaines plateformes. C'est le cas de Facebook, dont la collecte a cessé en juillet 2020 du fait d'un taux d'échec très élevé lié à l'impossibilité de dépasser la page de test captcha. Finalement, l'utilisation d'un seul et même outil de collecte n'empêche pas de devoir continuellement adapter le processus d'archivage pour chacune des plateformes.

34 Malgré ces difficultés, la BnF collecte régulièrement quatre plateformes : Twitter depuis 2012, YouTube depuis 2017, Instagram et TikTok depuis 2020 et 2022.

Les plateformes collectées par la BnF : Twitter, YouTube, Instagram et TikTok

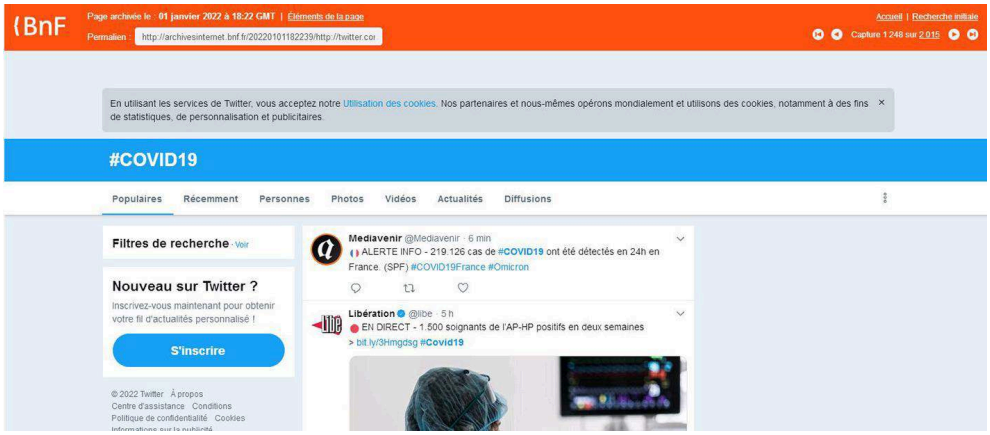
35 **Twitter** est le média social le plus collecté par la BnF. La plateforme de micro-blogging occupe une place centrale dans les réseaux d'information. Elle a été largement investie par les médias presse et les journalistes. Elle entre dans plusieurs collectes de la BnF et, en premier lieu, dans celle de l'Actualité. Cette collecte lancée quotidiennement couvre

les sites de presse et les portails d'information (MSN, Orange, Yahoo...), ainsi qu'une centaine de comptes Twitter qui leur sont associés.

36 L'activité de sélection des correspondants a suivi la popularité de la plateforme. Elle est particulièrement soutenue dans le cadre des collectes électorales et de la collecte de l'Actualité éphémère, mise en place en septembre 2018 pour suivre les événements en cours et archiver les réactions avec une plus grande réactivité. Twitter et Facebook représentaient ainsi 15 % des sélections faites pour la collecte électorale de 2012, 40 % pour celle de 2017 ou encore près de 25 % des sélections liées à la crise sanitaire de la Covid-19. Ce taux monte à 75 % pour les attentats de novembre 2015 et le mouvement des Gilets jaunes[21].

37 L'utilisation d'un robot de collecte permet de sauvegarder l'interface web de la plateforme et le contexte des publications. L'objectif est de permettre la rejouabilité de l'archive et la restitution de l'expérience de l'utilisateur. Les contenus populaires au moment de la collecte et les algorithmes de la plateforme impactent ainsi le résultat de la collecte. L'archive produite constitue ainsi une sorte d'artefact numérique de la page Twitter telle qu'elle se présente aux internautes. Pour autant, la collecte par robot présente plusieurs défauts déjà soulignés : la récupération des contenus est partielle, du fait des limites de profondeur fixées, et lacunaire, car le robot ne peut accéder aux liens profonds et ne parvient pas systématiquement à collecter toutes les images et les vidéos. La rejouabilité reste donc relative. Néanmoins, l'archive produite présente l'avantage de donner une idée du fonctionnement de la plateforme et permet de retracer son évolution historique, de même qu'elle éclaire également la circulation des contenus sur un mode plus qualitatif.

Figure 2. Page Twitter #COVID19 du 1er janvier 2022.



Collection des archives du web de la BnF.

38 La première collecte **YouTube** réalisée à la BnF a eu lieu en mai 2018. Elle a nécessité une importante supervision technique sur plusieurs semaines. Elle est lancée en moyenne deux fois par an pour un poids cible qui varie entre 3 et 5 To. La collecte s'effectue au niveau de la chaîne prise dans son intégralité et non de la vidéo. Elle s'appuie sur le logiciel open source youtube-dl qui sert notamment à la récupération des métadonnées (informations techniques et documentaires sur les vidéos telles que le titre, la date, la durée). Ces données permettent d'évaluer le poids des vidéos de chaque chaîne et, dans certains cas, de réévaluer les choix documentaires puisque le poids de certaines chaînes peut dépasser 1 To. Ce sont en tout plus de 3 000 chaînes qui ont été

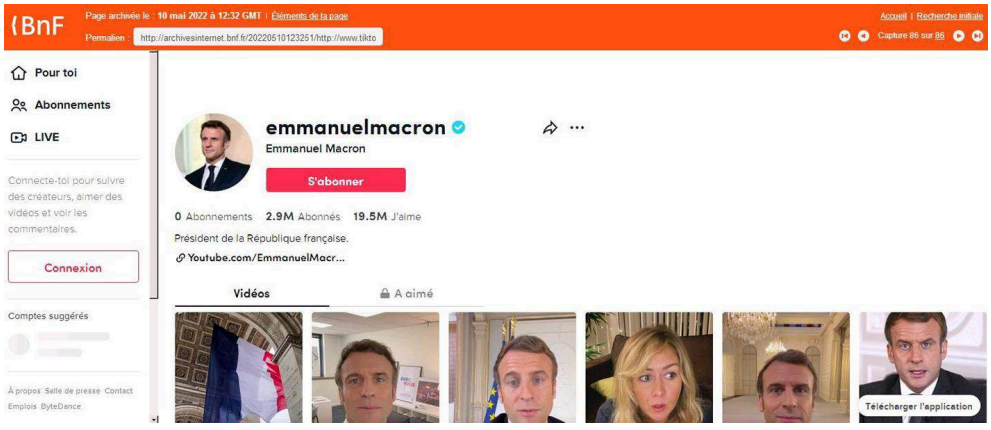
collectées entre 2018 et 2022, dont près de 600 dans le cadre des collectes électorales (20 %) et 110 dans le cadre de la collecte Covid-19 (3,5 %).

39 La collecte **Instagram** a été lancée en juin 2020. Elle se rapproche de la collecte YouTube dans son mode d'organisation. Il s'agit d'une collecte planifiée, jusqu'à quatre fois par an, qui requiert une préparation et une supervision technique. Comme dans le cas de Facebook, la plateforme bloque rapidement les robots en cas de requêtes répétées. Pour contourner ce blocage, la BnF s'appuie sur un agrégateur de contenu qui met les publications, images et vidéos à disposition des internautes sans qu'il soit nécessaire de s'authentifier.

40 La sélection documentaire privilégie la continuité avec les collectes thématiques et est moins reliée à l'actualité. La collecte porte sur un maximum de 200 comptes et *hashtags* et les captures sont limitées à 30 publications correspondant à la taille d'une page-écran. Ces limitations prudentes, tout comme les règles de politesse adoptées par le robot, visent à constituer une archive de bonne qualité. La première archive constituée en 2020 est d'un poids relativement réduit, avec 6 Go de données produites pour 170 comptes.

41 **TikTok** est la plateforme la plus récemment ajoutée au périmètre de nos collectes de réseaux sociaux numériques : elle est archivée depuis l'élection présidentielle de 2022. La collecte est effectuée directement à partir du site tiktok.com avec le robot Heritrix qui récupère les 16 dernières vidéos publiées sur un compte ou en lien avec un *hashtag*. L'interface web de la plateforme est également collectée ainsi que les premiers commentaires des utilisateurs. Les vidéos archivées peuvent être rejouées à partir des pages web. Pour la première collecte TikTok, une centaine de comptes et *hashtags* ont été sélectionnés par les correspondants. Il s'agit essentiellement des comptes des candidats et de leurs soutiens, de comptes de médias traditionnels et de *pure players*[22], ainsi que des principaux *hashtags* (une cinquantaine) permettant de recueillir des réactions militantes ou de simples particuliers. Comme pour les collectes Instagram et YouTube, la collecte TikTok devrait être lancée plusieurs fois par an selon une fréquence qui reste à déterminer.

Figure 3. Page TikTok d'Emmanuel Macron, capture du 10 mai 2022.



Collection des archives du web de la BnF.

42 Ces modalités de collecte ont des implications sur la recherche et la visualisation des contenus archivés. Pour consulter l'ensemble de ses archives qui représentent

aujourd'hui 49 milliards d'URL et 1,8 pétaoctet de données, la BnF propose une application appelée « Archives de l'internet ». Elle permet de retrouver l'ensemble des captures d'un site web, une page, un compte de réseau social, une vidéo, etc., à partir de son adresse URL et de naviguer de page en page et de site en site.

43 L'ensemble des sites sélectionnés dans le cadre des collectes ciblées réalisées par la BnF, dont les comptes et *hashtags* des plateformes de réseaux sociaux, sont disponibles sur le site API et Jeux de données (https://api.bnf.fr). Ces sélections prennent la forme de fichiers au format CSV contenant adresses URL, informations techniques sur la fréquence et la profondeur de collecte et informations documentaires (thèmes, mots-clés, description). Ces listes représentent un outil essentiel pour connaître la collection et retrouver les médias sociaux archivés.

44 Les listes de métadonnées sont d'autant plus essentielles que la masse des données produites d'année en année ne permet pas d'envisager pour le moment une indexation en plein texte de la totalité des archives. Une recherche par URL permet d'interroger l'ensemble des collections, mais il n'est pas toujours évident de connaître l'adresse d'un site disparu aussi bien pour les débuts du web que pour des périodes plus récentes. Pour compléter cette recherche par URL, les équipes de la BnF ont développé des fonctionnalités de recherche textuelle sur une partie des collections qui ont fait l'objet de demandes particulières, notamment dans le cadre de projets de recherche. Sont ainsi disponibles en plein texte les collections « Presse et actualités », « Le web des années 1990 », « Les attentats parisiens de 2015 », « La première vague de l'épidémie de Covid-19 » et prochainement « Le web littéraire francophone ». Ces fonctionnalités permettent de retrouver facilement des contenus issus de Facebook et Twitter. Attention, ces résultats portent sur le contenu intégral des pages collectées et non sur des publications unitaires (texte d'un *tweet* ou d'un post) et ne permettent pas de réaliser une réelle analyse quantitative des résultats.

45 Un parcours guidé « Des vidéos à la chaîne » a été spécialement créé pour faciliter l'accès aux chaînes et vidéos YouTube et Dailymotion. Les parcours guidés constituent une forme de médiation des archives qui vient compléter les outils de recherche. Ils présentent une série de captures autour d'un grand thème. Chaque capture est brièvement décrite et contextualisée. Le parcours guidé décrit et donne accès à l'ensemble des chaînes et vidéos collectées. Il permet un accès par thème (art, jeu vidéo, militantisme, etc.), par événement (élections, Covid-19) ou simplement par titre de chaîne. Pour chaque chaîne, il est également possible de faire une recherche sur les titres des vidéos.

Figure 4. Parcours guidé « Des vidéos à la chaîne, un état des lieux » proposé par la BnF.

46 Malgré des biais et des limites propres à certaines plateformes, l'archivage des médias sociaux par la BnF enrichit de manière de plus en plus notable les collections et suscite un intérêt croissant des chercheurs et des partenaires. Pour l'avenir, les enjeux d'amélioration concernent autant les modalités de collecte de ces plateformes que celles de consultation des contenus une fois archivés.

47 En premier lieu, les plateformes de réseaux sociaux évoluant en permanence, les outils de collecte doivent sans cesse être adaptés pour tenter de réaliser des collectes de la meilleure qualité possible. La BnF et, plus largement, la communauté internationale s'intéressent aujourd'hui à des outils de collecte intégrant un navigateur « sans tête », c'est-à-dire sans interface graphique. L'objectif est d'utiliser les capacités du navigateur à interpréter le code complexe des pages web pour améliorer la récupération des contenus et des médias. Face à ce défi, plusieurs bibliothèques nationales ont fait le choix, à l'instar de l'Ina, de s'orienter vers l'utilisation des API et d'abandonner la collecte web de ces plateformes. C'est le cas par exemple de la Bibliothèque royale du Danemark [KB].

48 Le second enjeu de collecte concerne l'arrivée et l'essor de nouvelles plateformes, qui ont acquis une réelle popularité souvent au détriment de Twitter. Mastodon et Telegram vont ainsi donner lieu à des tests de collecte. Néanmoins, même lorsqu'une partie des développements est partagée par la communauté des archivistes web, la multiplication des collectes génère *de facto* de nouvelles tâches et des besoins de supervision pouvant atteindre les limites de capacité d'action des équipes techniques.

49 Du côté de l'accès et pour permettre l'exploitation de ces contenus à des fins de recherche, la BnF souhaite continuer à travailler avec les utilisateurs pour définir avec eux les développements les plus utiles. Il importe en effet de proposer de nouveaux modes d'entrée et d'appropriation des collectes. L'exploitation plus systématique des métadonnées, comme les listes de sélections des correspondants, ouvrirait des perspectives et répondrait aussi à des besoins exprimés par les chercheurs. L'objectif est ici autant d'offrir de nouvelles possibilités de recherche que de donner à voir la collection notamment *via* la production de datavisualisations sur une thématique ou sur une plateforme particulière.

50 Plusieurs outils ont pu être expérimentés avec des laboratoires de recherche partenaires en 2022. L'outil SolrWayback, outil *open source* de recherche plein texte, de fouille et de visualisation de données, a ainsi été testé sur une collecte dédiée au thème

de l'alimentation coproduite avec l'université de Strasbourg (laboratoire SAGE, ERC Bodycapital). L'outil Hyphe, développé par le Medialab de Sciences Po pour constituer et cartographier des corpus web, a été adapté pour être utilisé sur les archives de l'Internet. Si ces outils ne sont pas spécifiques à l'exploitation des archives de médias sociaux, ils présentent un fort potentiel notamment en termes d'analyse computationnelle. Ils répondent à des intérêts convergents dans la manière de travailler sur les corpus web, que ceux-ci proviennent du web vivant ou du web archivé.

Diversité et complémentarité

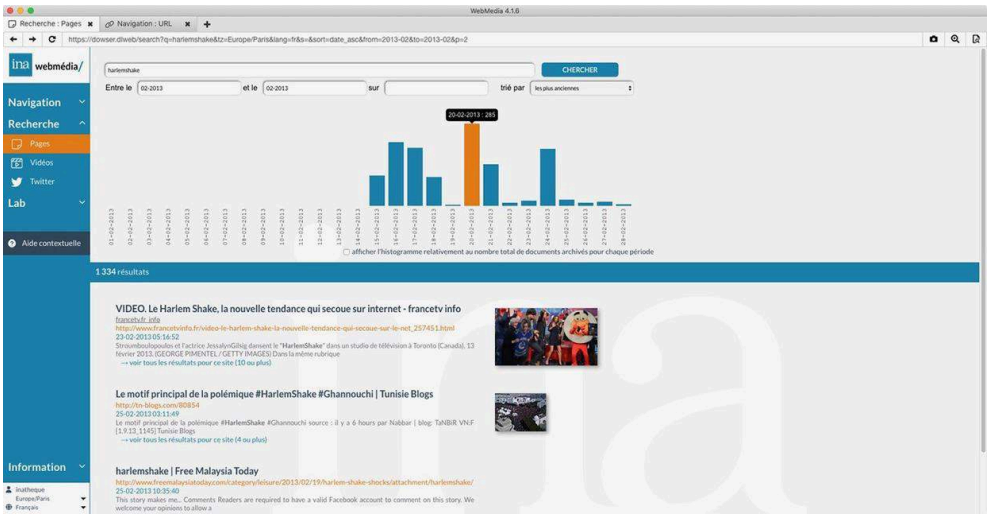
51 Appliquée à la recherche, cette diversité des modes de collecte et de consultation complexifie la démarche du chercheur, qui doit prendre connaissance du paysage institutionnel et des périmètres de collecte de chaque organisme dépositaire, de leurs méthodes et de leurs outils, avant de pouvoir étudier les archives du web. Le manque d'interopérabilité entre les données oblige les chercheurs à composer avec des formats différents, fichiers JSON[23] ou WARC, comme l'a souligné le projet WARCnet[24], en portant un regard sur les collectes du web mises en place au niveau international lors de la pandémie de la Covid-19 (Aasman *et al.*, 2021). Or cette échelle est par exemple pertinente pour analyser les phénomènes viraux en ligne (« mèmes[25] » qui circulent, *hashtags* comme #metoo, etc.) et l'impact d'événements globaux comme la crise sanitaire.

52 Après avoir pris connaissance de cette diversité intrinsèque des archives, les chercheurs peuvent néanmoins s'appuyer sur les complémentarités des collections et des outils, et notamment de ceux proposés par l'Ina et la BnF.

53 Une première illustration de ces complémentarités se lit dans l'étude de la viralité en ligne menée dans le cadre du projet HIVI[26]. L'exemple du *Harlem Shake*, forme de danse collective filmée qui connaît une popularité aussi soudaine qu'éphémère en 2013, offre un point de comparaison sur la manière dont les collections de l'Ina et de la BnF témoignent de ce phénomène. Les vidéos se révèlent nombreuses dans les collections de l'Ina grâce à la collecte des plateformes Dailymotion et surtout YouTube. Elles peuvent être mises en relation avec les productions télévisuelles, notamment le premier résultat issu de francetv.fr qui utilise les termes « secouer » et « nouvelle tendance » pour parler du phénomène. À l'inverse, les collections de la BnF proposent essentiellement des traces laissées par les vidéos YouTube embarquées sur les pages des sites de presse. Pour autant, ces traces définissent un corpus analysable grâce aux données de l'API YouTube et permettent de suivre une forme de diffusion géographique de la viralité notamment *via* la presse régionale. La « nouvelle tendance », terme repris dans ces articles, se retrouve dans une grande diversité de rubriques (« insolite », « sport », etc.), qui rappellent les termes employés sur la plateforme YouTube, mais aussi dans le monde audiovisuel. Cette complémentarité des sources et des données permet de confronter plusieurs résultats de recherche. L'existence d'une continuité documentaire entre les médias sociaux et les autres sources – contenus télévisuels pour l'Ina, sites web et presse en ligne pour la BnF – facilite la compréhension des circulations à l'heure de l'hybridation des médias (Esprit, Médias hybrides, 2022). Dans l'exemple du *Harlem Shake*, la comparaison des différentes sources montre ainsi clairement que le monde de l'audiovisuel et celui de la presse jouent des mêmes codes performatifs en mobilisant les mêmes termes. L'analyse d'un autre phénomène viral

antérieur, le *lip dub* (2007-2009), montre cette fois que les archives de la BnF conservent de fait un corpus vidéo représentatif de la diffusion du phénomène en France, et ce grâce à la collecte large et aux collectes de la plateforme Dailymotion. On notera donc que l'exploitation des archives du web, et en particulier celles des médias sociaux, doit s'appuyer sur une connaissance de l'histoire des plateformes, mais aussi de l'historique des collectes.

Figure 5. Recherche *Harlem Shake* dans les collections vidéo de l'Ina.

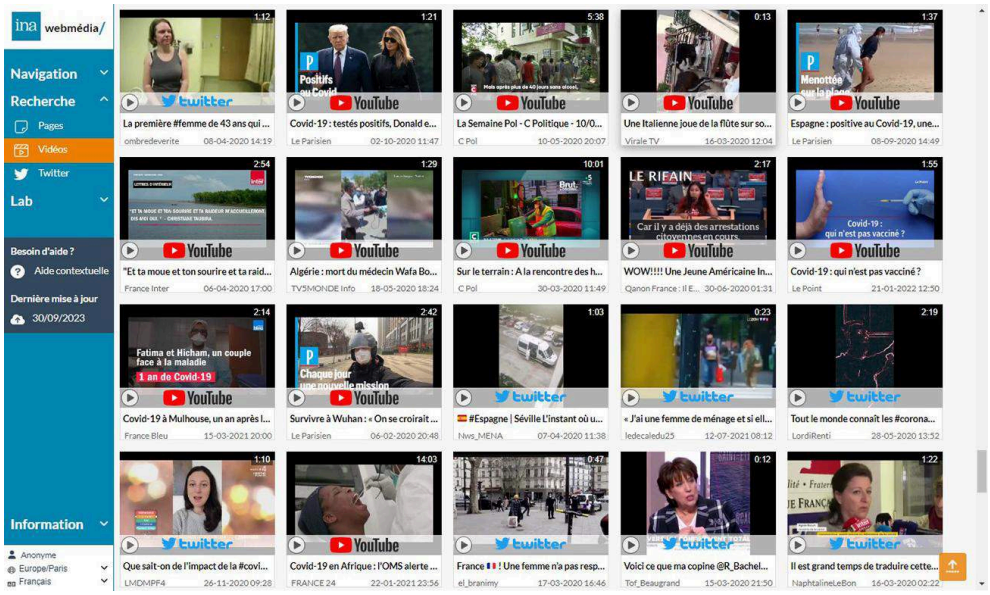


54 Les collectes Attentats 2015 et Covid-19 éclairent, elles aussi, ces complémentarités. Les deux événements historiques n'ont pas eu la même durée et, dans le cas des attentats, la réactivité des équipes fut essentielle à l'archivage des réseaux sociaux, qu'il s'agisse de sélectionner rapidement les sources (comptes et *hashtags*) à suivre ou de planifier une nouvelle collecte dans le cas de la BnF. Alors que l'Ina lançait une collecte massive de *tweets* au moment des attentats de janvier et novembre 2015, la BnF choisissait une approche plus sélective notamment pour la collecte de novembre, comme le note Annick Le Follic : « nous avions déjà plusieurs chantiers en cours [...] et étions donc moins disponibles, en termes de moyens techniques et humains. Ensuite, comme pour *Charlie*, nous avions collecté beaucoup de matériaux, qui n'étaient pas tous forcément pertinents, nous avons été plus sélectifs en novembre. Sur Twitter, nous avons collecté les comptes et *hashtags* qui remontaient le plus (préfecture de Paris, #attentats, etc.) » (Borelli et Schafer, 2016). Au contraire, l'Ina a choisi de poursuivre sa collecte Twitter de manière plus durable dans le temps, capturant les récurrences des *hashtags* liés à *Charlie Hebdo* au moment des attentats du Bataclan ou de Nice.

55 À la suite de la collecte Attentats, la BnF a réorganisé et complété sa méthode de collectes d'urgence en en dédiant une à l'Actualité éphémère. Ce nouveau dispositif est pleinement utilisé lors de la crise de la Covid-19 permettant de maintenir une collecte dynamique et collaborative sur une période de près de trois ans (Faye, 2022). Une attention particulière est portée au monde amateur et aux expressions personnelles. Ainsi la collecte YouTube lancée en juillet 2020 a cherché à atteindre la plus grande diversité possible en archivant aussi bien des chaînes d'institutions (Inserm, CHU) que des chaînes créées *ad hoc* lors du confinement et proposant des tutoriels pour créer des masques ou des vidéos d'humour. Le chercheur retrouve par ailleurs à l'Ina, en complément de ressources audiovisuelles en ligne liées au monde de l'audiovisuel et

des chaînes d'information par exemple, des vidéos Twitter, permettant, en croisant les fonds de la BnF et de l'Ina, d'avoir une vision large et complémentaire des usages et contenus vidéo.

Figure 6. Recherche spécifique dédiée à Femmes et Covid sur les collections web et audiovisuelles de l'Ina dans le cadre du projet WARCnet.

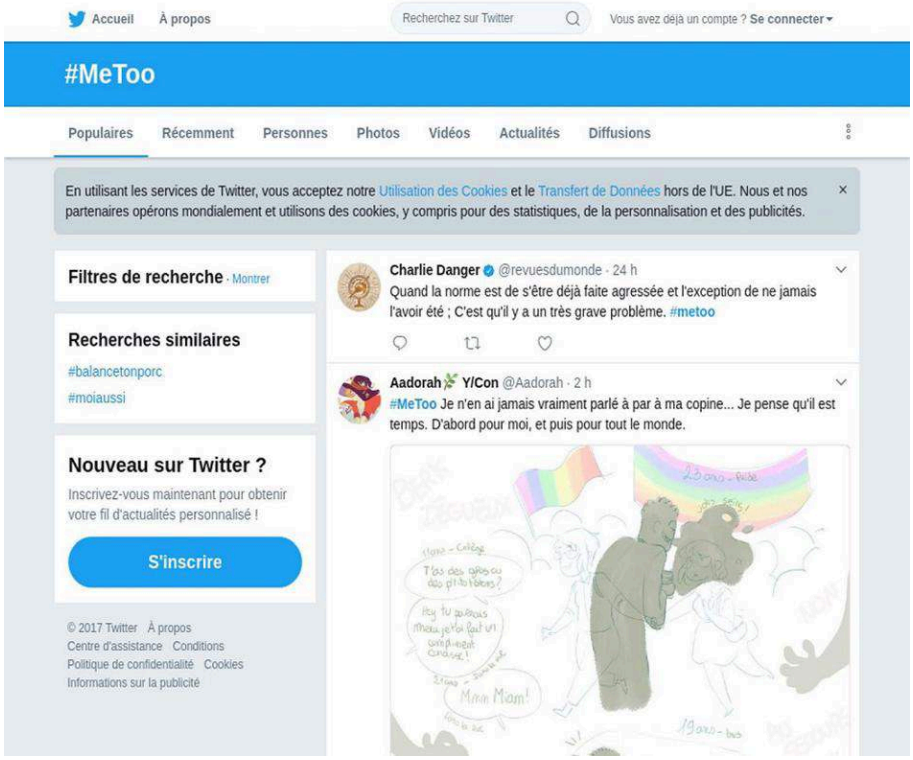


56 Les expériences originales de créations collectives ont aussi été recherchées par la BnF en vue de leur archivage, qu'elles aient pris la forme de sites web – on citera par exemple le projet d'écriture collaboratif Pandémonium proposé par le master Écopoétique et création de l'université d'Aix-Marseille autour de l'idée de « contamination ou maladie textuellement transmissible[27] » – ou bien de défis créatifs lancés sur les réseaux sociaux comme avec le *hashtag* #ArtEnQuarantaine, qui incitait les twittos à reproduire des œuvres de musée en se mettant en scène. Cet archivage des pratiques d'amateurs relève d'une pratique bien établie à la BnF depuis 2007 et la première collecte de blogs personnels en partenariat avec l'Association pour l'autobiographie [APA]. Les communautés de blogueurs ont ainsi représenté, comme le relève Christine Genin, « des collectifs, qui se commentaient entre eux en une sociabilité [qui précède] les réseaux socio-numériques » (Schafer, 2022). Cette continuité dans la collecte du web participatif, qui est aussi tangible à l'Ina, depuis la collecte du site skyblogs.com à l'Ina pour ses traces plus « anciennes », donne la possibilité aux chercheurs de travailler sur un temps long et de saisir l'évolution des identités et sensibilités numériques.

57 De la même manière, l'étude des mouvements sociaux peut s'appuyer sur l'historicité des collections, permettant ainsi de reconstruire des généalogies. Pour la BnF, la collecte du web militant fut lancée dès 2008 en partenariat avec des chercheurs, la BDIC (devenue La Contemporaine), le Centre d'histoire sociale de Paris 1 et le CERI-Sciences Po et a donné lieu à un parcours guidé. La collection de l'Ina, grâce à un archivage massif, montre à quel point les usages militants des réseaux sociaux et des vidéos se sont renforcés et influencent en retour les pratiques de collecte. 81 *tweets* et 95 vidéos répondent à une requête « contrat première embauche » renvoyant aux mobilisations anti-CPE de 2006. Dix ans plus tard, les résultats pour Nuit Debout (2016),

à savoir 785 000 pages web, 57 500 *tweets* et 2 300 vidéos, montrent l'essor des usages de Twitter et de sa collecte, ainsi que le développement des vidéos en ligne. Ils témoignent aussi des usages renforcés des RSN par les citoyens et les mouvements sociaux et de l'évolution des cultures numériques et de leur collecte. Quant à la collecte spéciale de l'Ina consacrée au mouvement des Gilets jaunes, elle conserve deux millions de pages web, 26 millions de *tweets* et 111 500 vidéos. Cette évolution démontre également la capacité des institutions à mettre en place des collectes spéciales et à mobiliser les moyens nécessaires pour cela.

Figure 7. Page Twitter #MeToo *Tweet* du 18 octobre 2017.



Collection des archives du web de la BnF (contenu sous droit, avec l'aimable autorisation d'Aadorah).

58 Pour autant, la promesse d'une archive plus ouverte aux expressions individuelles reste critiquable dans son effectivité. Elle correspond à un nouveau paradigme, qui n'est pas propre à l'archivage du web et qui se propose de réaliser des collectes plus sensibles et inclusives en veillant à ne pas omettre les « voix ordinaires » dans la construction d'une mémoire citoyenne, comme en témoignent l'inclusion des archives familiales dans la Grande Collecte du centenaire de la Grande Guerre, la sauvegarde des mémoriaux éphémères qui témoignaient de la solidarité des citoyens avec les victimes des attentats de 2015 ou encore l'ouverture de collectes participatives, lors du confinement de mars 2020, dont l'objectif était, le plus souvent, la conservation de témoignages individuels. Les chercheuses Sarah Gensburger et Marta Severo ont cependant montré que « la nature participative des collectes liées à la mémoire de la Covid renforce l'uniformité sociale des participants » (Gensburger et Severo, 2021). Les archives du web ne sont pas non plus sans biais sociologique, mais leurs méthodes de constitution, qui combinent souvent sélections et masse des données, pourraient aboutir *in fine* à la conservation

d'une plus grande diversité de voix – question à laquelle s'attache le projet web mémoires porté par les deux chercheuses.

59 Enfin, les complémentarités entre la BnF et l'Ina s'expriment au niveau des outils et des fonctionnalités de recherche proposées. Dans le cas de l'Ina, l'exploitation de métadonnées structurées permet des recherches particulièrement fines et efficaces. Il est par exemple possible d'établir la chronologie des réactions sur Twitter, *via* l'usage des *hashtags*, images ou émoticônes. Pour des études *crossmedia*, les chercheurs pourront également faire défiler une émission audiovisuelle et le fil Twitter associé. Ces possibilités ouvrent la voie à de nouvelles analyses fondées sur l'exploitation des métadonnées et la datavisualisation. C'est la raison d'être des récents labs de la BnF et de l'Ina, qui proposent d'accompagner la communauté des chercheurs vers le développement de nouvelles méthodologies d'analyse computationnelles (Carlin et Laborderie, 2021). Des synergies entre les deux institutions sont envisageables aussi bien au niveau des outils que de l'animation d'une communauté de chercheurs et de professionnels autour des archives du web.

Conclusion

60 La collecte des RSN induit de nouveaux paradigmes tant d'archivage que de recherche.

61 Dans le premier cas, il s'agit de renoncer à l'exhaustivité pour privilégier la représentativité, de créer des archives vivantes (Rollason-Cass et Reed, 2015), en temps réel et qui continuent de s'étendre au fil du temps et de leurs rappels. Il convient de traiter et rendre accessible des contenus, mais aussi, de plus en plus, des données et des métadonnées, et de prendre en compte des logiques de flux. Ces questions ne sont pas forcément nouvelles, mais se trouvent amplifiées par la préservation des RSN.

62 Médiation entre les outils développés par les institutions et les usages des chercheurs, l'interface d'accès et ses fonctionnalités induisent aussi de nouveaux paradigmes en termes d'analyse et favorisent la lecture scalable, oscillant entre lecture proche et distante (Armaselu et Fickers, 2022), étude fine du contenu et traitement computationnel de masses de données difficiles à saisir par le seul regard du chercheur.

63 Au-delà des méthodes, ce sont aussi les contenus qui changent et se tournent davantage vers les messages ordinaires des usagers, voix individuelles qui, mises ensemble, créent une forme de *big data* à même de renouveler l'histoire culturelle ou politique, ou encore les études mémorielles. Patrimoine nativement numérique largement inclusif, les traces des RSN posent aussi la question de leur appropriation par les publics. Actuellement essentiellement utilisées par les chercheurs, elles pourraient aussi conduire à de nouvelles formes de coconstruction pour développer des projets d'histoire publique. On observe d'ailleurs l'arrivée de nouveaux acteurs, notamment des musées, dans le domaine de l'archivage des réseaux sociaux (Canelli, 2022). Ils opèrent avec une approche différente de celle développée par les bibliothèques nationales, puisque leur objectif est souvent de collecter très peu, au niveau de l'objet, avec une qualité maximum dans un but d'exposition publique. L'élargissement des publics nécessiterait sans doute de développer la curation de ces contenus en partenariat avec d'autres institutions de mémoire, des musées ou des centres d'archives, mais aussi avec des associations et des collectifs qui souhaitent voir représentée leur histoire (ou mémoire) numérique.

BIBLIOGRAPHIE

Susan AASMAN *et al.*, “Chicken and Egg. Reporting from a datathon exploring datasets of the Covid-19 special collections”, WARCnet Papers, Aarhus, 2021. https://cc.au.dk/fileadmin/dac/Projekter/WARCnet/Aasman_et_al_Chicken_and_Egg.pdf

Florentina ARMASELU et Andreas FICKERS (dir.), *Zoomland. Exploring Scale in Digital History and Humanities,* Berlin, De Gruyter, 2023, en cours de publication.

Marguerite BORELLI et Valérie SCHAFER, « Entretien autour des collectes d’urgence au moment des attentats de janvier et novembre 2015 avec Annick Le Follic, chargée de collections numériques au département de dépôt légal de la BnF », *ASAP (carnet de recherche Hypothèses)*, 21 mars 2016, https://asap.hypotheses.org/168

Dominique BOULLIER, « Les sciences sociales face aux traces du big data : société, opinion ou vibrations ? », *Revue française de science politique*, n° 5, vol. 65, 2015, p. 805-828.

Niels BRÜGGER, *The Archived Web. Doing History in the Digital Age*, Cambridge, MA, The MIT Press, 2018.

Beatrice CANELLI, “Mapping social media archiving initiatives: state of the art, trends, and future perspectives”, *IIPC blog*, 30 novembre 2022 [En ligne] https://netpreserveblog.wordpress.com/2022/11/30/mapping-social-media-archiving-initiatives-state-of-the-art-trends-and-future-perspectives/.

Marie CARLIN et Arnaud LABORDERIE, « Le BnF DataLab, un service aux chercheurs en humanités numériques », *Humanités numériques* [En ligne], 4, 2021. http://journals.openedition.org/revuehn/2684.

Frédéric CLAVERT, « Face au passé : la Grande Guerre sur Twitter », *Le Temps des médias*, vol. 31, 2, 2018, p. 173-186.

Alexandre FAYE, « Archiver le web durant la Covid-19 et le premier confinement : organisation, bilan et perspectives », *Revue COSSI*, 11, 2022. https://revue-cossi.numerev.com/articles/revue-11/2749-archiver-le-web-durant-la-covid-19-et-le-premier-confinement-organisation-bilan-et-perspectives.

Sarah GENSBURGER et Marta SEVERO, « L’espace public du confinement. Archives, participation et inclusion sociale », *Revue d’histoire culturelle* [En ligne], 2021, http://revues.mshparisnord.fr/rhc/index.php?id=662.

Gustavo GOMEZ-MEJIA, « La fabrique de la désuétude. Regards diachroniques sur Geocities et Myspace », dans V. Schafer (dir.), *Temps et temporalités du web*, Nanterre, Presses universitaires de Paris Nanterre, 2018.

Kieran HEGARTY, “The invention of the archived web: tracing the influence of library frameworks on web archiving infrastructure”, *Internet Histories*, 2022, DOI: 10.1080/24701475.2022.2103988.

Library of Congress [LoC], “Update on the Twitter Archive at the Library of Congress”, 2017, https://blogs.loc.gov/loc/files/2017/12/2017dec_twitter_white-paper.pdf.

Aymerick MANSOUX et Roel ROSCAM ABBING, “Seven Theses on the Fediverse and the becoming of FLOSS”, dans K. Gansing et I. Luchs (dir.), *The Eternal Network. The Ends and Becomings of Network Culture*, Institute of Network Cultures, Amsterdam, 2020, p. 124-140.

Ian MILLIGAN, “Welcome to the web: the online community of GeoCities during the early years of the World Wide Web”, dans N. Brügger et R. Schroeder (dir.), *The Web as History: Using Web Archives to Understand the Past and the Present*, Londres, UCL Press, 2017, p. 137-158.

Francesca MUSIANI, Camille PALOQUE-BERGÉS, Valérie SCHAFER et Benjamin THIERRY, *Qu’est-ce qu’une archive du web ?*. Marseille, OpenEdition Press, 2019.

Sylvie ROLLASON-CASS et Scott REED, “Living Movements, Living Archives: Selecting and Archiving Web Content During Times of Social Unrest”, *New Review of Information Networking*, n 2, vol. 2, 2015, p. 241-247.

Valérie SCHAFER, Gérôme TRUC, Romain BADOUARD, Lucien CASTEX et Francesca MUSIANI, “Paris and Nice terrorist attacks: Exploring Twitter and web archives”, *Media, War & Conflict*, vol. 12, 2, 2019, p. 153-170.

Valérie SCHAFER, Jérôme THIÈVRE et Boris BLANCKEMANE, “Exploring special web archives collections related to COVID-19: The case of Ina”, *WARCnet Paper*, Aarhus, 2020. https://cc.au.dk/fileadmin/user_upload/WARCnet/Schafer_et_al_Exploring_special_web_archives.pdf

Valérie SCHAFER et Ben ELS, “Exploring special web archives collections related to COVID-19: The case of the BnL”, *WARCnet Paper*, Aarhus, 2020. https://cc.au.dk/fileadmin/user_upload/WARCnet/Schafer_et_al_COVID-19_BnL.pdf.

Valérie SCHAFER, « Préserve-moi ! Des journaux intimes à ceux de confinement dans les archives du web », *Le Temps des médias*, n° 38, vol. 1, 2022, p. 175-194.

Sabine SCHOSTAG, “The Danish coronavirus web collection – coronavirus on the curators’ mind”, *International Internet Preservation Consortium Blog*, 2020. https://netpreserveblog.wordpress.com/2020/07/29/the-danish-coronavirus-web-collection/

Zeynep TUFEKCI, *Twitter et les gaz lacrymogènes : Forces et fragilités de la contestation connectée*, Caen, C&F Éditions, 2019.

Eveline VLASSENROOT, Sally CHAMBERS, Sven LIEBER, Alejandra MICHEL *et al.*, “Web-archiving and social media: an exploratory analysis”, *International journal of digital humanities*, 2022, p. 107-128.

NOTES

1. Après des mois de négociations et de tergiversations, Elon Musk, milliardaire américain, patron de Tesla et cofondateur d’Ebay, a racheté Twitter au prix de 44 milliards de dollars en octobre 2022. Il en a licencié le PDG, a dissous le conseil d’administration, renvoyé de multiples employés ou encore réintégré des milliers de comptes bannis pour avoir violé les règles de la plateforme.
2. Réseau social qui repose sur une fédération de serveurs interconnectés. Pour en savoir plus voir Mansoux et Abbing (2020) ou la traduction en français de leur article sur Framablog : https://framablog.org/2021/01/26/le-fediverse-et-lavenir-des-reseaux-decentralises/.
3. Messages postés par les utilisateurs sur le réseau social Twitter.
4. Geocities, service gratuit de création de pages personnelles sur Internet, a fermé le lundi 26 octobre 2009, entraînant la disparition de plusieurs milliers de pages web. Le site d’hébergement de blogs Myspace, créé en 2003, existe toujours, mais, en mars 2019, des millions

de contenus créés avant 2016 sont perdus, apparemment à la suite d'une erreur technique. Créée en 2012 par Twitter inc., Vine était une application mobile de partage de courtes vidéos très populaire. L'application est abandonnée en octobre 2015.

5. API : « *application programming interface* » ou « interface de programmation d'application ».

6. Dès 2002.

7. Attentats terroristes très meurtriers perpétrés le 13 novembre 2015 à Paris (notamment contre les spectateurs de la salle de concert du Bataclan) et à Saint-Denis.

8. Mot-clé précédé du signe #, permettant le référencement des posts sur les sites de microbloging tels que Twitter, Instagram ou TikTok.

9. Application de diffusion vidéo en direct disponible sur Android et iOS, acquise par Twitter en 2015. Le service a été arrêté en 2021.

10. Webrecorder/Browsertrix se fonde sur les fonctionnalités des navigateurs et leur capacité d'interprétation du code pour améliorer la récupération des éléments de la page web. L'objectif est de produire des archives de haute qualité, ciblées et guidées manuellement pour des sites et plateformes trop complexes pour les robots habituellement utilisés par la communauté des archivistes.

11. « Facebook cherche activement à bloquer les robots de collecte et, par conséquent, vous devez collecter beaucoup plus de données pour obtenir un résultat intéressant. Ainsi, pour le coût d'une simple capture de Facebook, vous pouvez archiver régulièrement un site web, à cela s'ajoute le fait que la capture de Facebook peut même s'avérer inutilisable, par exemple avec des vidéos impossibles à lancer » (traduction par Alexandre Faye).

12. https://www.docnow.io/.

13. https://www.sucho.org/.

14. « *Uniform Resource Locator* » ou « localisateur uniforme de ressource » : adresse pointant vers une ressource unique sur le web.

15. Cambridge Analytica est une entreprise britannique spécialisée dans le conseil en communication et l'analyse de données, qui a été utilisée par Donald Trump au cours de sa campagne pour la présidentielle de 2016 et qui a analysé à leur insu les données de dizaines de millions d'utilisateurs de Facebook.

16. Ces agressions terroristes meurtrières dirigées notamment contre l'équipe de rédaction du journal satirique *Charlie Hebdo* et un supermarché se sont produites du 7 au 9 janvier 2015.

17. Le procès est lié au scandale sanitaire concernant les personnes victimes de la prise de benfluorex, commercialisé sous le nom de Mediator par les laboratoires Servier de 1976 à 2009. Le procès en appel du Médiator a débuté le 9 janvier 2023 devant la cour d'appel de Paris, presque deux ans après la première condamnation des laboratoires Servier.

18. Heritrix est un robot d'indexation sous licence libre développé par Internet Archive à partir de 2003.

19. Plus précisément : 33 chaînes sur le thème de l'alimentation, collectées dans le cadre du projet Bodycapital en partenariat avec le laboratoire SAGE (Université de Strasbourg) et 16 chaînes animées par des youtubeurs en histoire, pour les besoins du doctorant Arthur de Forges de Parny accueilli dans le cadre du dispositif d'accueil chercheur associé de la BnF.

20. Web ARChive.

21. De novembre 2018 à juin 2019.

22. Entreprises qui exercent leur activité professionnelle uniquement en ligne.

23. JavaScript Object Notation.

24. https://cc.au.dk/en/warcnet.

25. Élément (texte, vidéo, audio, image, etc.) repris, adapté et partagé massivement en ligne.

26. Projet de recherche soutenu par le Fonds national de la recherche [FNR] au Luxembourg (C20/SC/14758148) et portant sur l'histoire de la viralité en ligne. hivi.uni.lu

27. Le site éphémère Pandémonium a été collecté entre juin 2020 et mars 2021 à l'adresse pandemik.org.

AUTEURS

ALEXANDRE FAY

BnF

JÉRÔME THIÈVRE

Ina

VALÉRIE SCHAFER

C²DH, Université du Luxembourg

Le traitement des données de masse au sein du dépôt légal du Web de l'Institut national de l'audiovisuel

Boris Blanckemane

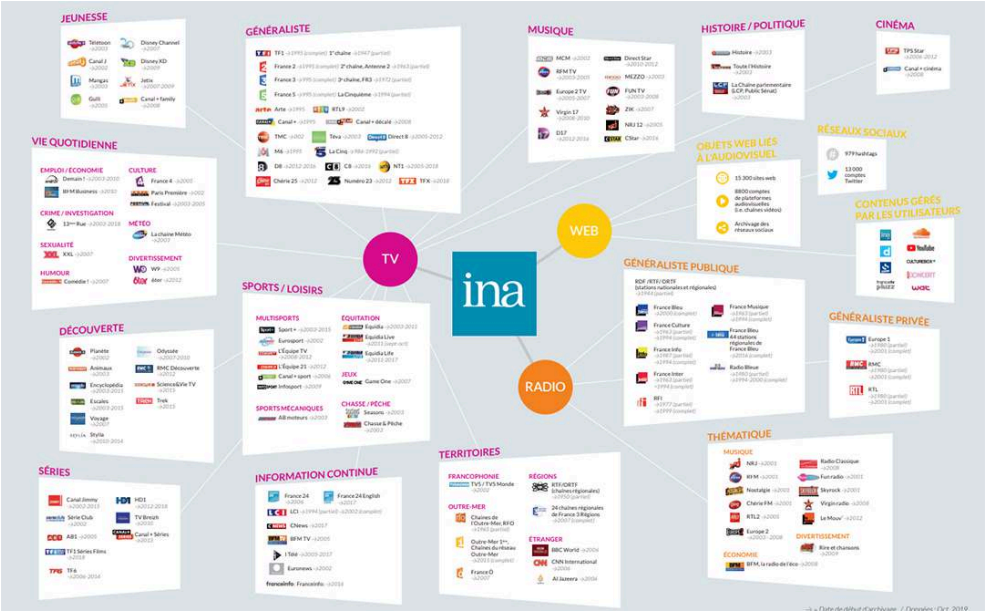
1 Qu'elles soient matérielles (bobines de film, cassettes vidéos, DVD, documents diffuseurs au format papier) ou immatérielles (fichiers vidéos, fichiers audios, fichiers textes), les données que l'Ina, depuis sa création en 1974 jusqu'à nos jours, a été amené à collecter, gérer, archiver et valoriser, présentent des volumes de plus en plus importants. Ces données s'accompagnent généralement de métadonnées (données décrivant les données) qui sont à considérer avec autant, sinon plus, de soin que les données. Dans la suite de cet article, on regroupera les notions de données et de métadonnées sous le terme d'« omnidonnées[1] ». La gestion des omnidonnées produites par les acteurs de l'audiovisuel français est donc l'une des composantes clés de la mission d'archivage qui incombe à l'Ina. L'objet de cet article est de décrire quantitativement et qualitativement ces omnidonnées massives, les outils mis en place pour les collecter, ceux mis en œuvre pour les archiver et ceux créés pour en assurer la valorisation et la mise à disposition auprès du public. Plus particulièrement, il se concentrera sur la gestion des omnidonnées massives telle qu'elle est opérée par et au sein du service du dépôt légal du Web depuis la sélection jusqu'à leur valorisation en passant par le stockage, l'indexation et la documentation.

Contexte et mission de l'Ina

2 Créé à la suite de l'éclatement de l'ORTF [Office de radiodiffusion-télévision française], l'Institut national de l'audiovisuel a pour mission l'archivage des chaînes publiques de télévision et de radio. En 1992, le dépôt légal français est étendu à toutes les chaînes du paysage audiovisuel français. Cette loi étend alors le périmètre de collecte à 120 chaînes de télévision et de radio. Cette mission menée à bien *via* la captation de ces chaînes en format numérique, 24 heures sur 24, a conduit à la collecte de 23 millions d'heures de programmes télédiffusés et radiodiffusés avec une croissance d'un million d'heures par

an. La loi DADVSI et son décret d'application du 19 décembre 2011[2] ont étendu le périmètre du dépôt légal aux sites web. Cette mission est partagée entre l'Ina, pour les sites liés à l'audiovisuel français, et la Bibliothèque nationale de France pour les autres sites du domaine français.

Figure 1. Quelques exemples de médias collectés par l'Ina.



Source : http://www.inatheque.fr/medias/Brochure_2020_INAtheque_CnC.pdf, consulté en octobre 2022.

Périmètre de collecte du dépôt légal du Web de l'Ina

3 Le périmètre de collecte du dépôt légal du Web de l'Ina s'articule sur quatre types d'objets web dits « sources » :

- les sites web :
  - les sites officiels des diffuseurs nationaux (par exemple https://www.france.tv, https://www.radiofrance.fr/),
  - les sites officiels des émissions produites par les diffuseurs nationaux (par exemple https://www.6play.fr/le-meilleur-patissier-p_1807),
  - les sites contenant des informations relatives à l'audiovisuel français : sites de communautés de fans de programmes télévisés, forums d'échange, sites non officiels (par exemple https://www.plusbellelavie.org/) ;
- les comptes de réseaux sociaux :
  - les comptes liés aux médias : comptes officiels des chaînes (par exemple @francetv, @CNEWS), comptes officiels des programmes (par exemple @cashinvestigati, @TPMP),
  - les comptes liés aux acteurs des médias : comptes de journalistes, de présentateurs et de présentatrices, de directeurs et de directrices de programmes, de podcasteurs, etc. ;
- les comptes de plateformes vidéos (ou *UGC* pour *User-Generated Content)* francophones :
  - les chaînes YouTube françaises,
  - les chaînes Dailymotion françaises,
  - les chaînes Vimeo françaises,
  - les chaînes Twitch françaises,

    - les chaînes Soundcloud françaises,
    - les chaînes Mixcloud françaises,
    - les chaînes Vine françaises,
    - les chaînes Wat françaises,
    - les *podcasts* français ;
  - les tweets associés :
    - aux *hashtags* liés aux médias français (par exemple #FranceInter, #EnvoyeSpecial),
    - aux *hashtags* liés aux sujets d'actualités qui résonnent fortement dans les médias français (par exemple #presidentielle2022).

4 En pratique, la collecte consiste à archiver :

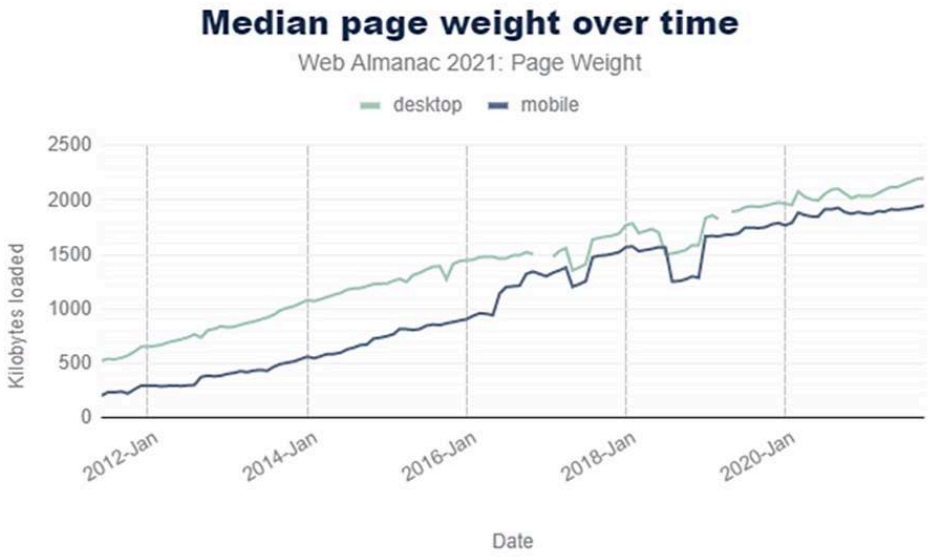
- pour un site web donné : toutes les pages et les ressources (textes, médias, CSS, scripts) permettant de reproduire pleinement la navigation,
- pour un *hashtag* : tous les tweets contenant le *hashtag* et qui sont fournis par l'API Twitter,
- pour un UGC : toutes les vidéos publiques de la chaîne,
- pour les comptes de réseaux sociaux : toutes les publications publiques des comptes,
- pour les *podcasts* : tous les épisodes.

Omnidonnées sur le Web

5 À un objet collecté correspond un nombre important d'omnidonnées associées. Par exemple, une page web contient en moyenne 2 Mo de données (images, textes, polices), un tweet contient une trentaine de métadonnées (date de publication, texte, lien), une vidéo YouTube, une quarantaine (description, durée, titre).

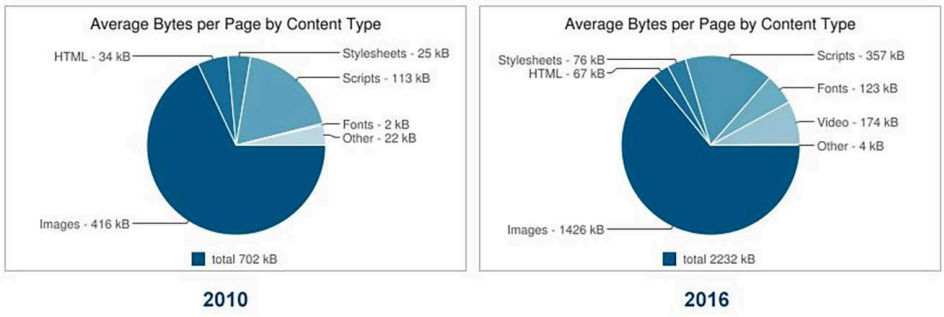
6 En outre, en raison de la démocratisation des ordinateurs personnels et de leurs usages, et de la diminution du coût du stockage grand public, ces métriques tendent à augmenter avec les années comme le montrent les illustrations ci-dessous.

Figure 2. Évolution du volume moyen de données contenues dans une page web.



Source : https://almanac.httparchive.org/en/2021/page-weight, consulté en octobre 2022.

Figure 3. Taille d'une page web et répartition par type de données.



Source : https://www.nosyweb.fr/google-analytics-web-analytique/la-croissance-de-la-taille-des-pages-web.html, consultation octobre 2022.

Processus de collecte des omnidonnées

Sélection des sources à collecter

7 La sélection des quatre types de sources est effectuée par l'équipe documentation du dépôt légal du Web de l'Ina. Cette équipe est constituée, côté métier, d'une chargée de mission qui établit et formalise la méthodologie de sélection et de documentation des sources, d'un ou plusieurs documentalistes en session chargés de sélectionner, de documenter et d'ajouter ces sources à la collecte et, côté maîtrise d'œuvre, d'un responsable de projet chargé de construire et mettre à disposition les outils informatiques permettant ce travail de sélection et de documentation.

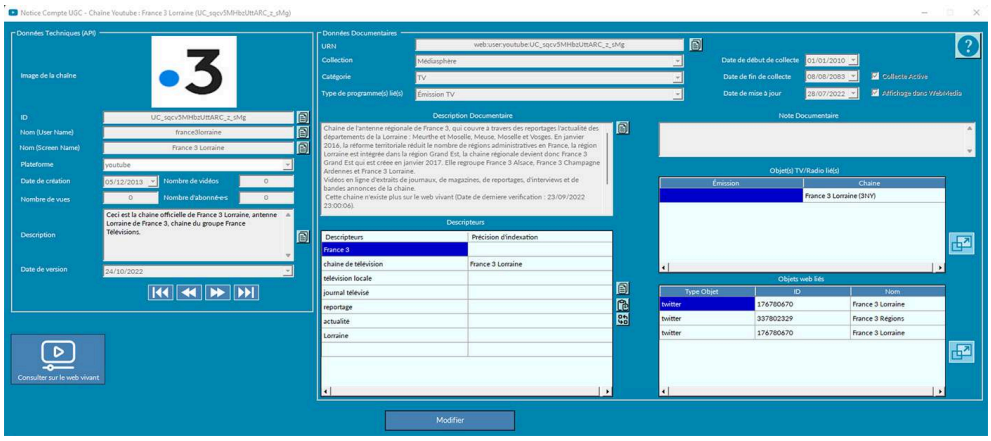
8 Les sources éligibles peuvent être détectées *via* des outils de veille automatisés (Talkwalker, Trends24), des inscriptions à des newsletters à la thématique audiovisuelle ou des agrégateurs de flux (Hootsuite) fondés sur des mots-clés relatifs à l'audiovisuel. Les sources peuvent aussi être issues des habitudes de consommation et des connaissances personnelles des documentalistes en session. L'ajout d'une source peut aussi s'effectuer dans le cadre d'un projet de recherche ou faire suite à la demande d'un organisme tiers qui nous communique une liste de sources à collecter.

Ajout des sources à collecter

9 L'ajout des sources s'effectue *via* un outil dédié et développé par le dépôt légal du Web : DUMBOW[3]. Ce client lourd développé en VB/.NET permet de consulter, d'ajouter et de modifier la base de données SQL/MariaDB contenant les sources à collecter.

10 Chaque source dispose d'une notice regroupant à la fois les informations de collecte (date de début de collecte, date de fin de collecte) et les informations documentaires (description, descripteurs, catégorie).

Figure 4. Notice documentaire de la chaîne YouTube de France 3 Lorraine.



Volumétrie des sources et des objets collectés

11 Le tableau ci-dessous présente les métriques des sources et des objets collectés par le dépôt légal du Web de l'Ina en septembre 2022.

Tableau 1. Volumétrie des sources de collectes et des objets collectés par le dépôt légal du Web.

Type de sources	**Sous-type**	**Nombre de sources**	**Type d'objets collectés**	**Nombre d'objets collectés**
Sites web	N/A	16 537	Pages	124 300 000 000
Hashtags	N/A	3 091	Tweets	2 891 830 636
Comptes de réseaux sociaux	Twitter	16 217		
Chaînes de plateformes vidéos	YouTube	6 822	Fichiers vidéos/ audios	11 841 101
	Dailymotion	1 561		6 954 592
	Soundcloud	720		734 047
	Vimeo	1 090		322 257
	Mixcloud	209		N/A
	Wat TV	116		1 326 454
	Twitch	63		N/A
	Vine	464		19 405
Podcasts	N/A	5 374	Épisodes	2 254 971

12 Les volumétries de ces sources de collecte évoluent de manière croissante comme le montrent les graphiques ci-dessous.

Figure 5. Évolution du nombre de *hashtags* collectés en fonction du temps.

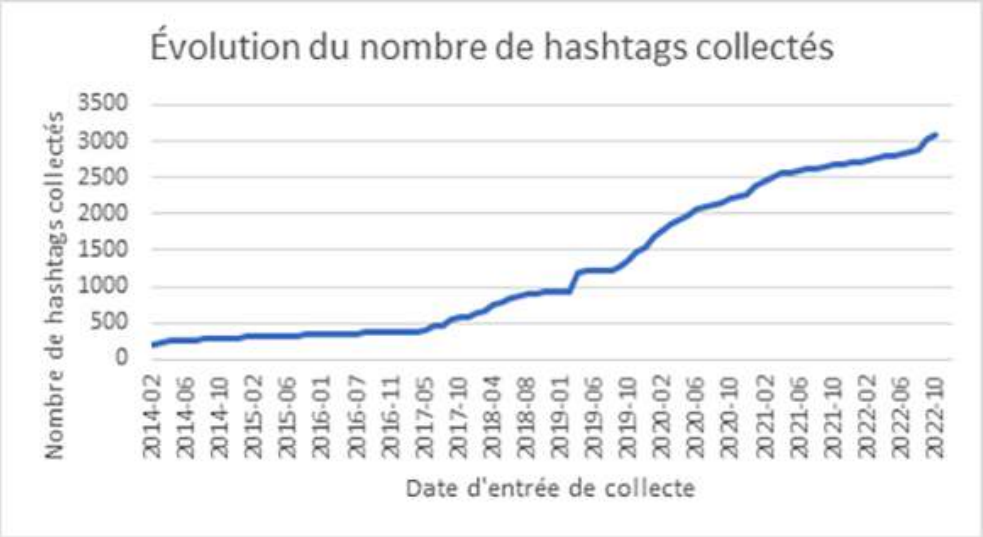


Figure 6. Évolution du nombre d'UGC collectés en fonction du temps.

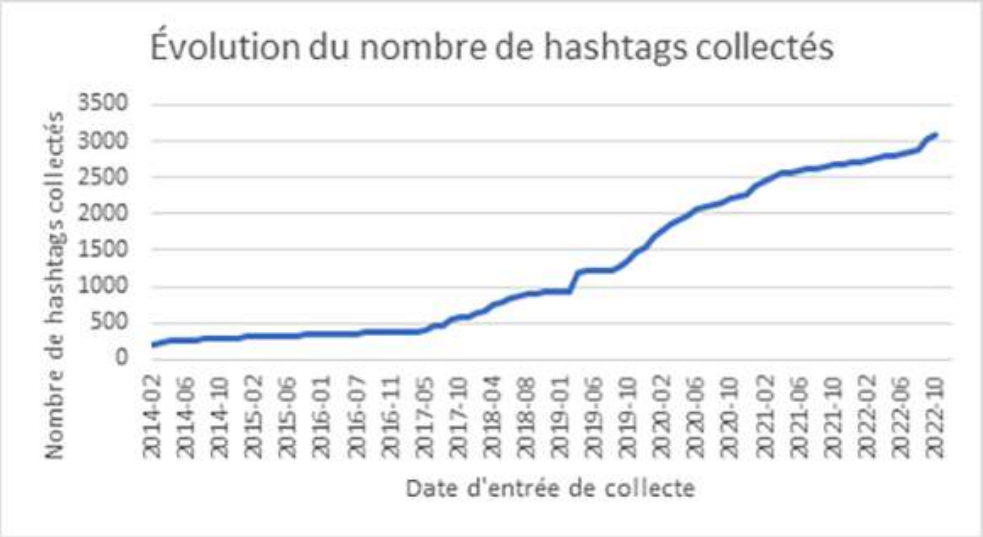


Figure 7. Évolution du nombre de comptes de réseau social collectés en fonction du temps.

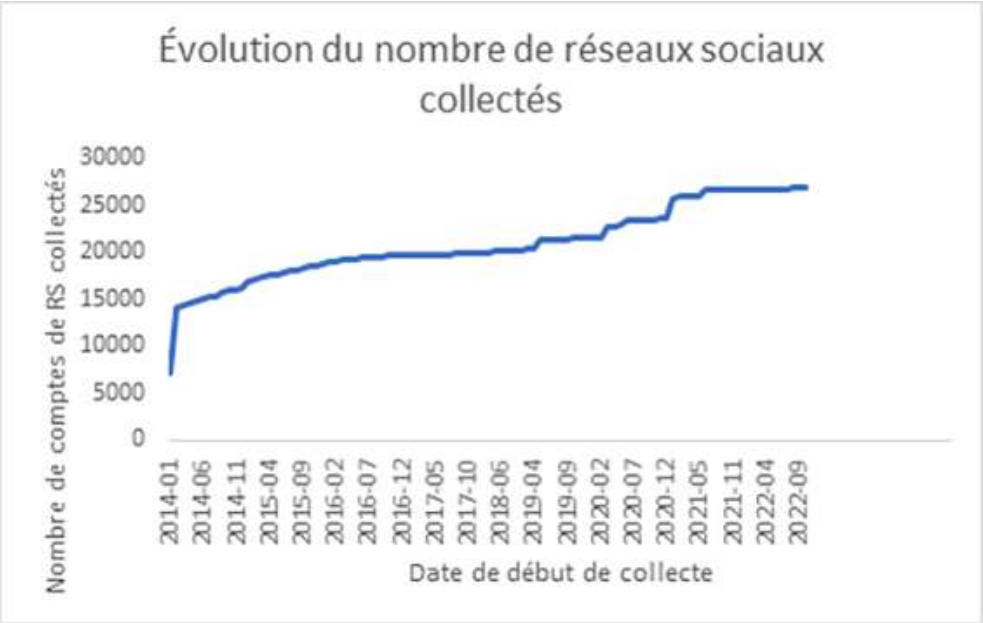
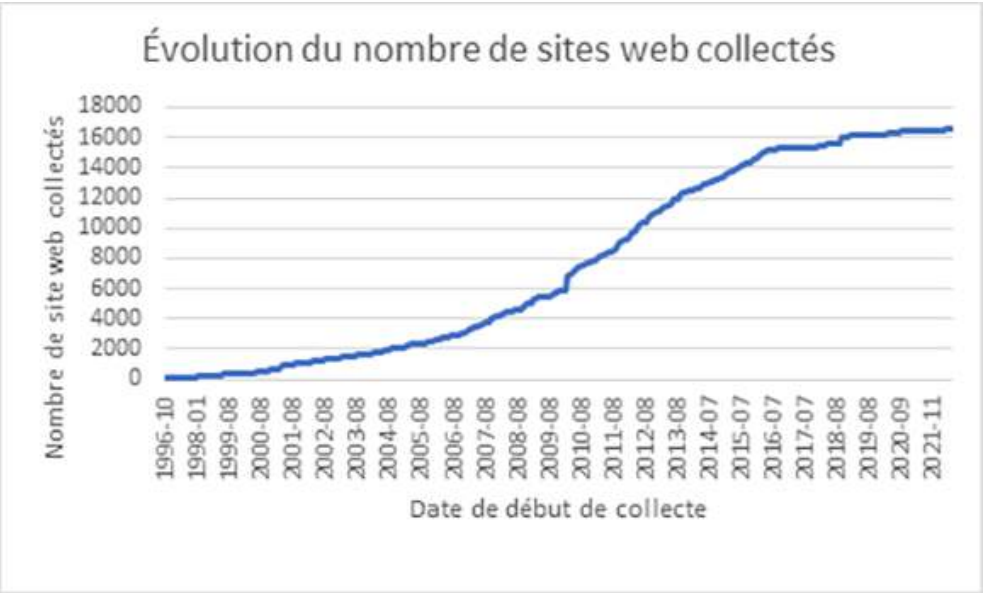


Figure 8. Évolution du nombre de sites web collectés en fonction du temps.



13 Les sources du périmètre de la collecte évoluant de façon croissante, le volume d'objets collectés augmente logiquement selon la même tendance comme l'illustrent les figures 9, 10 et 11. Cependant, pour une source ajoutée à la collecte, ce sont plusieurs dizaines d'objets correspondants (publications, vidéos, pages) qui seront collectés. De même, il est à noter que, si le dépôt légal du Web maîtrise complètement le nombre de sources collectées, il ne maîtrise pas le nombre d'objets à extraire de ces sources.

Figure . Évolution du nombre d'objets web cumulés en fonction du temps.

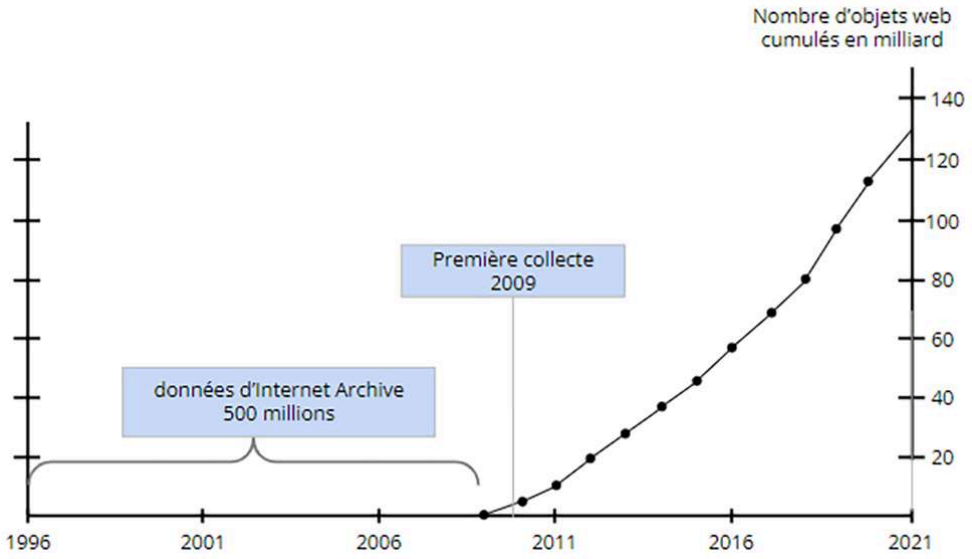


Figure 9. Évolution du nombre d'heures de vidéos collectées en fonction du temps.

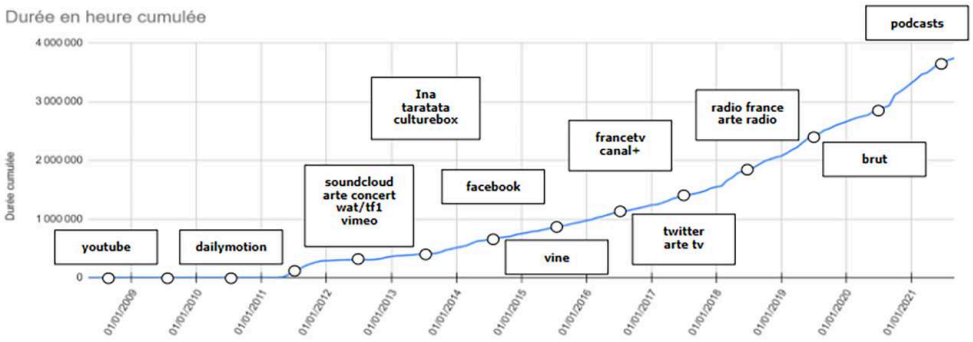
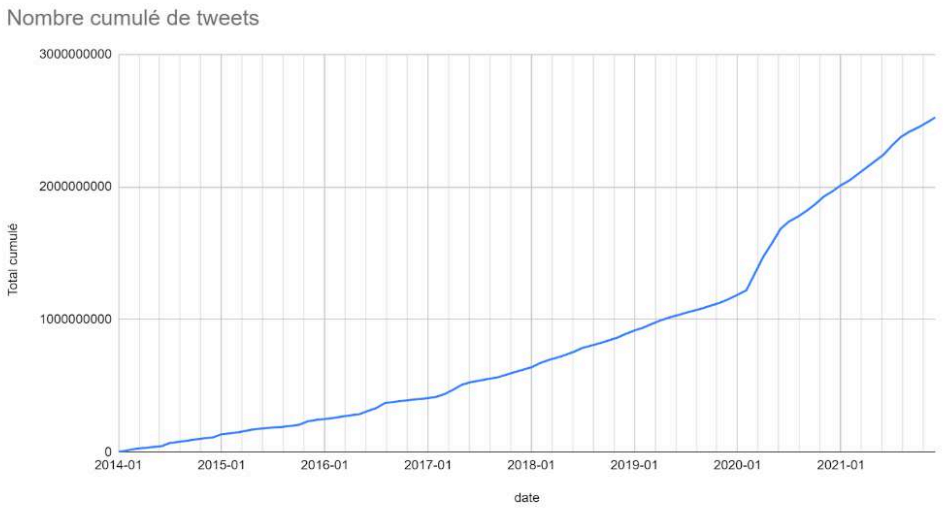


Figure 10. Évolution du nombre de tweets collectés en fonction du temps.



14 Les volumétries des sources de collecte, dont l'ordre de grandeur est le millier, ainsi que les volumétries des objets collectés, dont l'ordre de grandeur est le million, attestent du caractère massif des omnidonnées issues du fonds du dépôt légal du Web.

15 Ce caractère massif va être déterminant dans le choix des outils de sélection, de collecte et d'exploitation de ces omnidonnées.

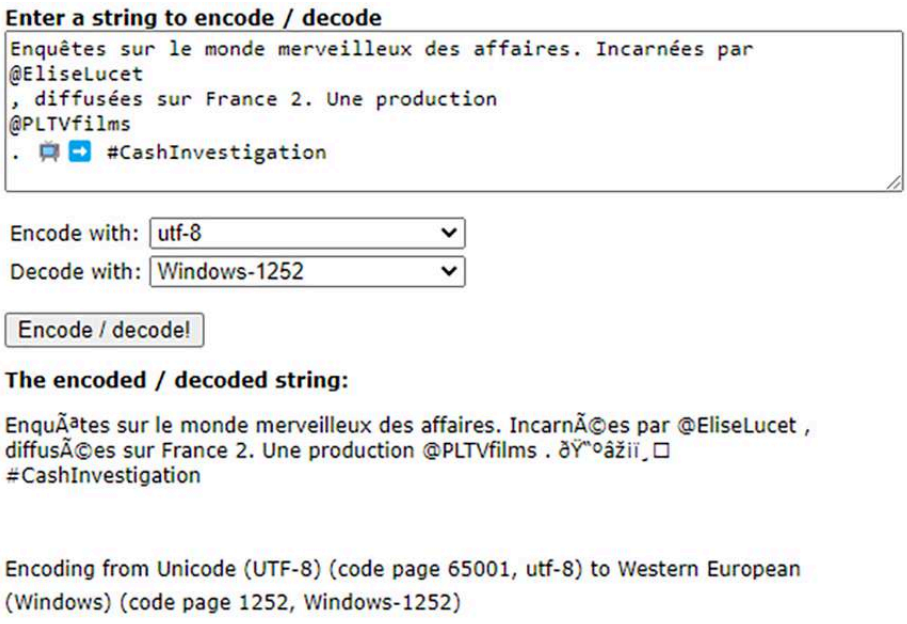
Exploitabilité des omnidonnées massives

16 Aussi importantes que les données, les métadonnées permettent d'indexer les données et *a fortiori* d'exploiter et de valoriser le fonds archivé. Ainsi, l'exploitabilité des données dépend directement de la qualité de l'indexation de ces métadonnées.

17 Au-delà de la question du volume, la première problématique rencontrée à l'Ina, et plus particulièrement au sein du service du dépôt légal du Web, réside dans l'hétérogénéité des données (images, fichiers vidéos, fichiers audios, données textuelles) et dans l'hétérogénéité des métadonnées (codecs, résolution, bitrate[4]).

18 On observe aussi une hétérogénéité dans les formats des métadonnées. Ainsi, une métadonnée contenant une date peut se présenter sous divers formats. Par exemple, la date de création est stockée sur Dailymotion au format EPOCH[5] et au format ISO 8601[6] sur YouTube. L'encodage des chaînes de caractères dans les métadonnées retournées par les API est aussi un point d'attention. En effet, la description d'un compte Twitter peut contenir des émojis. Si cette description est stockée dans ou est utilisée par un champ technique n'ayant pas le même encodage, la donnée archivée ne sera pas parfaitement identique à la donnée originale.

Figure 11. Exemple d'une conversion d'un encodage prenant en charge les émojis (UTF-8) vers un encodage ne les prenant pas en charge (Windows 1252).



19 Un travail préalable d'étude et de normalisation des formats des métadonnées est donc nécessaire.

20 Dans un deuxième temps, il faut déterminer quelles sont les omnidonnées pertinentes et ne conserver que ces dernières. Cette démarche préalable permet d'économiser du temps machine (traitement, stockage de l'omnidonnées) et du temps humain (dans la prise en charge intellectuelle et logicielle de l'omnidonnée) aussi bien lors de l'étape de collecte que lors de l'étape d'utilisation, de valorisation, de mise en œuvre des omnidonnées.

21 La pertinence d'une omnidonnée se juge au regard de l'usage métier qui en sera fait et de son exploitabilité technique. Par exemple, sur une collecte de tweets uniquement francophones, stocker la métadonnée du tweet qui contient le code langue n'est pas un choix pertinent. Pour faciliter ce travail, il peut être intéressant de distinguer les métadonnées selon deux catégories :

- les métadonnées documentaires qui contiennent des informations de nature à documenter la donnée. Par exemple, la date de création d'une chaîne YouTube ;
- les métadonnées techniques qui contiennent des informations de nature à qualifier techniquement la donnée. Par exemple, le format du conteneur d'une vidéo.

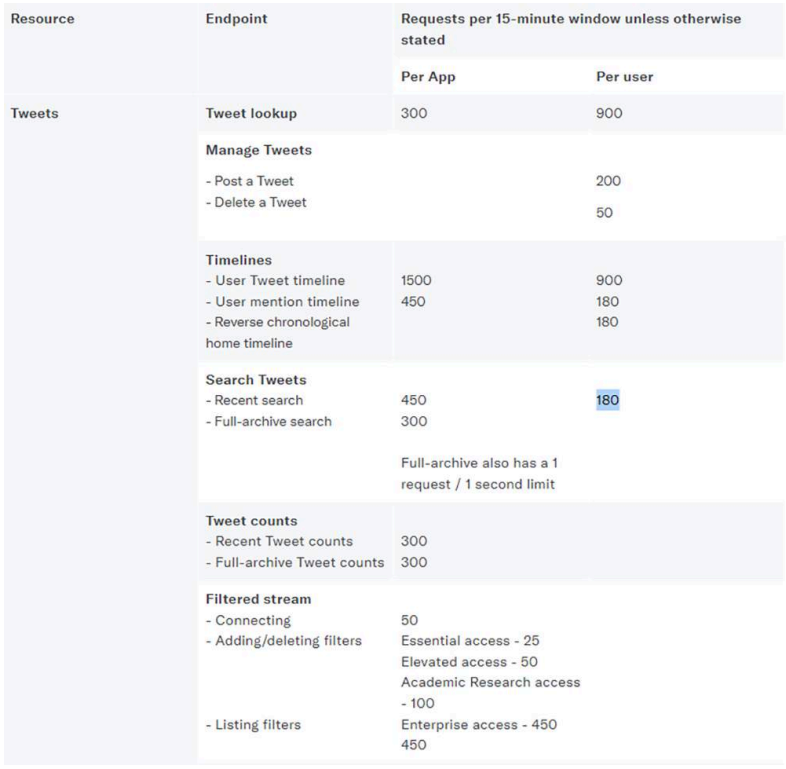
22 Le choix des omnidonnées nécessite donc un travail *ad hoc*, à effectuer préalablement à la collecte et en prenant en compte les usages métiers. Il peut être facilité par le fait que les omnidonnées sont généralement structurées et décrites de façon exhaustive dans les documentations des API des plateformes.

23 Cependant, les API présentent des inconvénients :

- Une évolutivité/versatilité forte. Les plateformes (YouTube, Twitter) sont souvent pionnières en termes d'usage et implémentent/déprécient régulièrement des fonctionnalités.
- Une limitation d'usage. Souvent gratuites, les API limitent les accès sous forme de quotas (cf. Fig. 13) qui sont à prendre rigoureusement en compte dans le développement des outils. Ainsi, le dépôt légal du Web de l'Ina a fait le choix de ne pas archiver l'évolution des engagements d'un tweet (nombre de *likes*, de retweets et de citations). En effet, archiver ces métriques de façon régulière nécessiterait de réinterroger l'API pour les 2,9 milliards de tweets.
- Aucun engagement contractuel n'existe entre l'utilisateur accédant gratuitement à l'API et la plateforme éditrice, il n'y a aucun engagement de qualité et de continuité de service de la part de celle-ci.

24 L'utilisation d'API suppose de fait une adaptabilité forte et une réactivité suffisante de la part des équipes de développement comme des équipes maîtrise d'ouvrage et métier qui doivent aussi se tenir informées des évolutions des API afin d'adapter les méthodologies et les stratégies d'usage.

Figure 12. Quelques limitations de l'API Twitter V2.



Resource	Endpoint	Requests per 15-minute window unless otherwise stated	
		Per App	Per user
Tweets	Tweet lookup	300	900
	Manage Tweets		
	- Post a Tweet		200
	- Delete a Tweet		50
	Timelines		
	- User Tweet timeline	1500	900
	- User mention timeline	450	180
	- Reverse chronological home timeline		180
	Search Tweets		
	- Recent search	450	180
	- Full-archive search	300	
		Full-archive also has a 1 request / 1 second limit	
	Tweet counts		
	- Recent Tweet counts	300	
	- Full-archive Tweet counts	300	
	Filtered stream		
	- Connecting	50	
	- Adding/deleting filters	Essential access - 25 Elevated access - 50 Academic Research access - 100	
	- Listing filters	Enterprise access - 450 450	

Source : https://developer.twitter.com/en/docs/twitter-api/rate-limits, consulté le 19 octobre 2022.

Collecte des données

25 La collecte des données par le dépôt légal du Web de l'Ina s'effectue *via* un ensemble de briques logicielles appelé « Robot de Collecte ».

26 Développés en interne par l'équipe Recherche & Développement (4 personnes), ces robots se fondent sur un ensemble de langages (PERL, JavaScript, Python) et de librairies tierces pour collecter les différents objets au périmètre et stocker le résultat de cette collecte.

27 La volumétrie importante des omnidonnées impose d'adapter la stratégie de collecte. Afin de pouvoir collecter suffisamment et de façon efficiente, cette stratégie doit privilégier :

- le parallélisme, lancement de plusieurs collectes de façon simultanée et maîtrisée, *via* des ordonnanceurs paramétrables et monitorés par une équipe dédiée (au dépôt légal du Web, cette équipe dite « Exploitation » est constituée de deux personnes à plein temps),
- la dédication. Il est plus judicieux de disposer d'une stratégie de collecte adaptée à la nature de l'objet collecté plutôt que d'une collecte polyvalente qui sera plus compliquée à mettre en œuvre et à maintenir.

28 Ainsi, on n'applique pas la même stratégie de collecte à un compte Twitter qui publie plusieurs dizaines de tweets par jour et à une chaîne YouTube qui publie quelques dizaines de vidéos par mois.

Stockage et traitement des omnidonnées

Stockage et indexation des omnidonnées

29 Le stockage et l'indexation des omnidonnées revêt deux aspects concomitants :

- Un aspect technique. Quels formats de stockage utiliser, quelles solutions logicielles développer, comment choisir et dimensionner les infrastructures matérielles,
- Un aspect structurel. Comment hiérarchiser, organiser, exploiter et assurer l'intégrité des omnidonnées.

30 Le stockage des omnidonnées se fait en général sur des disques de type HDD [*Hard Disk Drive*] qui présentent un rapport qualité-prix-capacité-rapidité-robustesse intéressant.

31 Cependant, le stockage a tendance à être de plus en plus fait sur des disques de type SSD [*Solid State Drive*]. Ces disques n'utilisant pas une technologie mécanique, ils sont moins sujets aux pannes et présentent des temps d'accès plus rapides. Bien que leurs prix soient plus élevés, ils se rapprochent de plus en plus de ceux des disques HDD (3 à 5 fois plus chers aujourd'hui contre 10 fois plus chers il y a une dizaine d'années[7]).

32 Au-delà du format du disque, l'interface matérielle du dispositif de stockage joue aussi un rôle primordial dans les performances. Ainsi, un disque SSD utilisant le protocole SATA [*Serial Advanced Technology Attachment*] est moins rapide qu'un disque SSD utilisant le protocole plus récent NVMe [*Non-Volatile Memory Express*].

33 Le choix de la solution de stockage doit aussi prendre en compte la nécessité d'assurer une réplication des données. En fonction de la politique de réplication des données du projet et/ou de la politique de réplication de l'entreprise, les besoins en stockage peuvent ainsi être multipliés par un facteur 2 ou 3.

34 La solution de stockage dépend aussi des objectifs d'accessibilité souhaités. Le temps de mise à jour des index, le temps d'accès aux index (qui conditionne le temps de réponse des moteurs de recherche), le nombre d'utilisateurs potentiels sont des facteurs déterminants qui dépendent de la puissance de calcul du serveur (processeur du serveur ou CPU), de la quantité de mémoire vive (RAM du serveur), mais aussi de la qualité du réseau par lequel transitent les informations.

35 Enfin, au-delà des considérations matérielles, il faut aussi choisir le format d'archivage. Au dépôt légal du Web, le format utilisé pour archiver les données collectées et leurs métadonnées est le DAFF[8], un format créé en interne, autodécrit, lisible (par un humain), « agnostique » au protocole (http, ftp), prenant en charge la déduplication complète et permettant un contrôle d'intégrité fiable.

36 Les omnidonnées des sources à collecter ont une volumétrie de quelques centaines de milliers d'éléments[9]. Ces données sont utilisées :

- par les documentalistes en session au sein du dépôt légal du Web,
- par les usagers pour consulter la liste des sources collectées,
- par les robots de collecte.

37 L'usage est donc modéré, limité à quelques dizaines d'utilisateurs.

38 De fait, la solution retenue a été le stockage de ces sources dans une base de données relationnelles MariaDB utilisant le langage de requêtage SQL. Très orientée vers un usage métier, cette base de données a été conçue de façon à être la plus lisible et la plus compacte possible. Chaque type d'objet (*hashtag*, compte de réseaux sociaux, UGC, sites

web) est stocké dans une table unique et dédiée qui est pleinement exploitable quand elle est exportée en l'état (export dit « à plat »).

39 De même, les contraintes d'intégrité des données et de leurs relations à implémenter au niveau de la structure de la base lors de sa conception ont été limitées au maximum. Il nous a semblé plus judicieux d'utiliser des contraintes métiers à la place. Elles sont implémentées dans l'outil de saisie qui nourrit cette base de données, la base de données reste ainsi plus adaptative afin d'être en capacité de gérer des objets qui seront collectés ultérieurement (comptes Facebook, Instagram) sans nécessiter de modifications structurelles fortes. Par exemple, dans la table des UGC, il n'aurait pas été judicieux de rendre obligatoire la date de création de la chaîne vidéo, cette dernière n'étant pas toujours communiquée par les plateformes.

40 En outre, un schéma relationnel réduit à son expression la plus simple permet d'effectuer plus facilement des modifications directement dans la base de données (*via* des requêtes SQL basiques les moins verbeuses possibles).

41 Les omnidonnées des objets collectés représentent quant à elles 129 milliards d'enregistrements (116 milliards d'enregistrements de métadonnées et 13 milliards d'enregistrements de données), soient 12 pétabytes de données réduites à 2.5 pétabytes après déduplication et compression.

42 À titre d'exemple, les omnidonnées résultant de la collecte des sites Web se constituent de :

- 4 000 000 000 d'enregistrements contenant les données des sites collectés (représentant un volume de 60 térabytes),
- 6 000 000 000 d'enregistrements contenant les métadonnées des sites collectés (représentant un volume de 1 térabyte).

43 La solution choisie pour répondre à ces contraintes a été de mettre en place des *frameworks*[10] distribués sur plusieurs serveurs :

- Hadoop pour le traitement de données, le travail de jointure des métadonnées et des données,
- Elastic Search pour le moteur de recherche plein texte.

44 Ces solutions distribuées présentent l'avantage de pouvoir être redimensionnées à la volée et en fonction des besoins (*scalability*).

45 Le temps d'indexation de ces contenus a été fixé à 5 jours et le temps de réponse acceptable des moteurs de recherche a été fixé à 5 secondes. L'atteinte de ces objectifs repose sur deux clusters :

- un cluster d'indexation constitué de 10 serveurs contenant les *frameworks* Hadoop et Elastic Search contenant chacun 12 disques HDD de 4 To, un processeur puissant, 64 Go de RAM,
- un cluster de recherche composé de 12 serveurs Elastic Search contenant 4 disques SSD de 1 To et 96 Go de RAM.

46 Dans les collections web, des outils de datamining permettent de détecter (et donc de collecter) de manière automatique les URL, les vidéos, les tweets présents au sein d'une page web.

Manipulation des données

47 La manipulation des omnidonnées sera facilitée par un certain nombre de compétences techniques. Entre autres, il convient de se familiariser avec les spécifications des formats suivants :

- CSV [*Comma Separated Values*] : format texte ouvert normé (RFC-4180) présentant des données tabulaires sous forme de valeurs séparées par des virgules,
- JSON [*JavaScript Object Notation*] : format ouvert normé (RFC-8259) dérivé de la notation des objets du langage JavaScript, structuré sous forme de couple (clé, valeurs),
- XML [*Extensible Markup Language*] : format ouvert normé (W3C XML Schema) structuré sous forme de couple (clé, valeurs) reposant sur la syntaxe balisée du langage HTML.

48 Afin de procéder aux opérations de tris, nettoyages, formatages, normalisations, la manipulation des jeux de données peut s'effectuer via plusieurs types d'outils :

- un éditeur de texte amélioré (NotePad++, SublimeText) : il possède des fonctionnalités d'édition en colonne, des fonctions « recherche/remplace » puissantes, gère différents encodages, prend en charge de nombreux formats, propose la coloration syntaxique[11] et offre la possibilité d'utiliser des extensions de manipulations de données développées par des communautés souvent très actives,
- un tableur (Microsoft Excel, Google Sheets, Libre Office Calc) : il possède des fonctionnalités de tri, de filtre, prend en charge des formules de manipulations de chaînes de caractères et l'utilisateur peut y utiliser un langage de programmation (VBA, JavaScript) pour créer des traitements automatiques,
- un logiciel de nettoyage de données (OpenRefine, Data Wrangler) : dédié complètement à cet usage, il permet de manipuler les données en utilisant les mêmes paradigmes que les outils de requêtages dans une base de données relationnelle (chaque donnée correspond à l'intersection d'une ligne et d'une colonne).

49 Enfin, idéalement, la connaissance d'un langage orienté « manipulation de données » peut s'avérer être un gain de temps considérable lors d'opérations de manipulations et de valorisation des omnidonnées. R et Python sont deux langages largement utilisés à cet effet, très bien documentés et bénéficiant d'une communauté en ligne active. En complément, nous utilisons le langage d'expression régulière (Regex) qui permet d'effectuer des opérations puissantes sur des chaînes de caractères ayant des propriétés communes pour les détecter et leur appliquer un traitement automatisé (ajout, remplacement, modification, suppression).

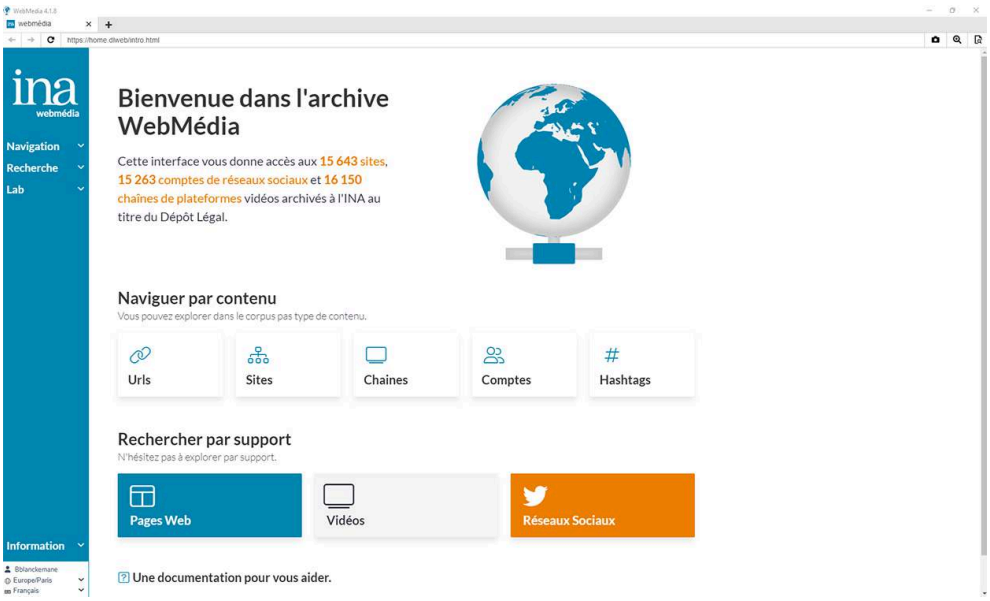
Valorisation et usages des données

Mise à disposition des données

50 Les données collectées par le dépôt légal du Web sont mises à disposition des usagers au sein des médiathèques partenaires et des emprises de l'Ina *via* un poste informatique dédié (le PCM, poste de consultation multimédia) sur lequel est installé l'outil de consultation de l'archive web de l'Ina : Webmédia.

51 Webmédia est un client lourd se présentant comme un navigateur web qui permet de naviguer dans l'archive.

Figure 13. Page d'accueil de l'outil de consultation de l'archive web de l'Ina.



52 Cet outil offre plusieurs parcours de navigation à l'usager qui peut :

- sélectionner des objets web à partir de filtres textuels appliqués à leurs métadonnées (Fig. 15),
- effectuer une recherche plein texte dans l'un des trois moteurs de recherche[12] (Fig. 16).

Figure 14. Recherche dans Webmédia des comptes Twitter contenant le terme « journaliste » dans la description documentaire.

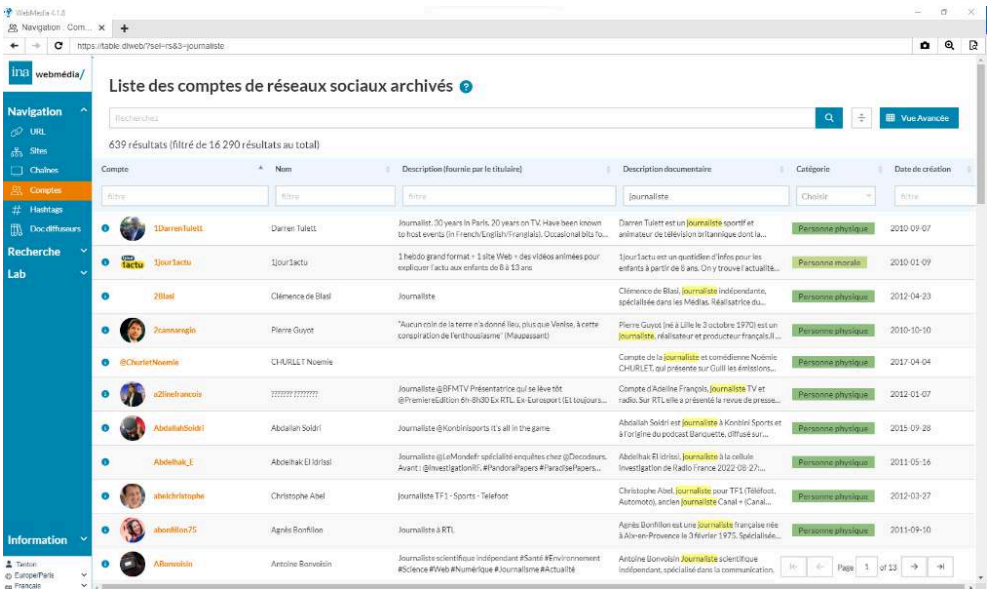
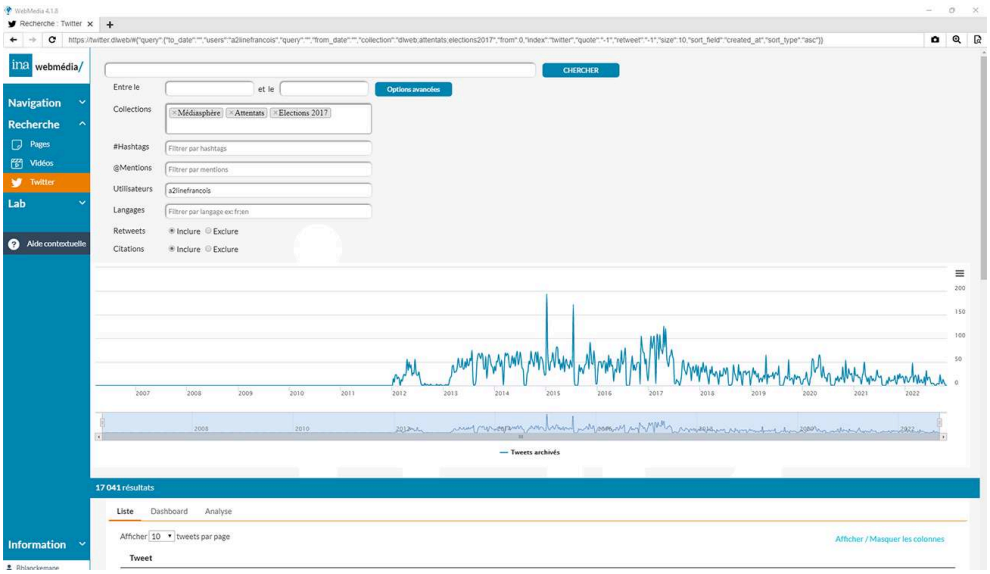


Figure 15. Recherche des tweets publiés par le compte @a2linefrançois via le moteur de recherche de l'archive Twitter.



Valorisation des données

53 Bien que la mission première de l'Ina, et par extension du dépôt légal du Web, soit l'archivage des omnidonnées sur son périmètre, c'est la valorisation de ces données qui permet de mettre en avant la richesse et l'intérêt de l'archive.

54 Cette valorisation peut être portée par des travaux, des projets de recherche s'appuyant sur nos collections comme :

- « Les cérémonies Miss France, de la télévision à Twitter. Une ritualisation des commentaires (2015-2019) », article de revue. Ce travail mené par Laurence Leveneur-Martel (Université Toulouse 1) a été réalisé en partie à partir d'exports de l'archive Twitter constituée à l'Ina.
- *JPICH-Cultural Heritage, Identities & Perspectives* : projet européen autour des archives du passé colonial. Sophie Gebeil (Université Aix-Marseille), en tant que chercheuse associée au projet, utilise les archives du web de l'Ina pour étudier les enjeux, les acteurs et les temporalités des débats liés à l'héritage colonial dans les années 2000[13].
- *ASAP*[14]*-Archives & Sauvegarde Attentats Paris*. Menée par le CNRS et l'Institut des sciences de la communication, cette étude se fonde sur la collecte des tweets relatifs aux attentats terroristes et publiés depuis janvier 2015.
- *Body Capital*[15] : bénéficiant d'un financement ERC [*European Research Council*] et menée par l'université de Strasbourg, cette étude se fonde sur la collecte des tweets relatifs au procès du Mediator et publiés depuis juillet 2019.

55 Le service du dépôt légal du Web mène aussi des projets de valorisation en interne. C'est le cas de l'étude *Épidémie de COVID-19 & Collecte Twitter*[16]. Elle concerne les tweets contenant au moins un *hashtag* lié à la thématique de la pandémie (parmi une sélection) et qui émanent d'une sélection de comptes liés aux médias audiovisuels ou les mentionnent. Ces tweets ont été extraits et analysés *via* des outils de *data-mining* et de *text-mining* pour générer des visualisations sous forme de graphiques et nuages de mots-clés. Cette étude, qui, par le prisme de Twitter, rend compte du traitement de la

COVID-19 par les chaînes TV/Radio d'actualité, est un exemple de travail dans lequel on transforme la valeur quantitative d'un jeu de données en valeur qualitative.

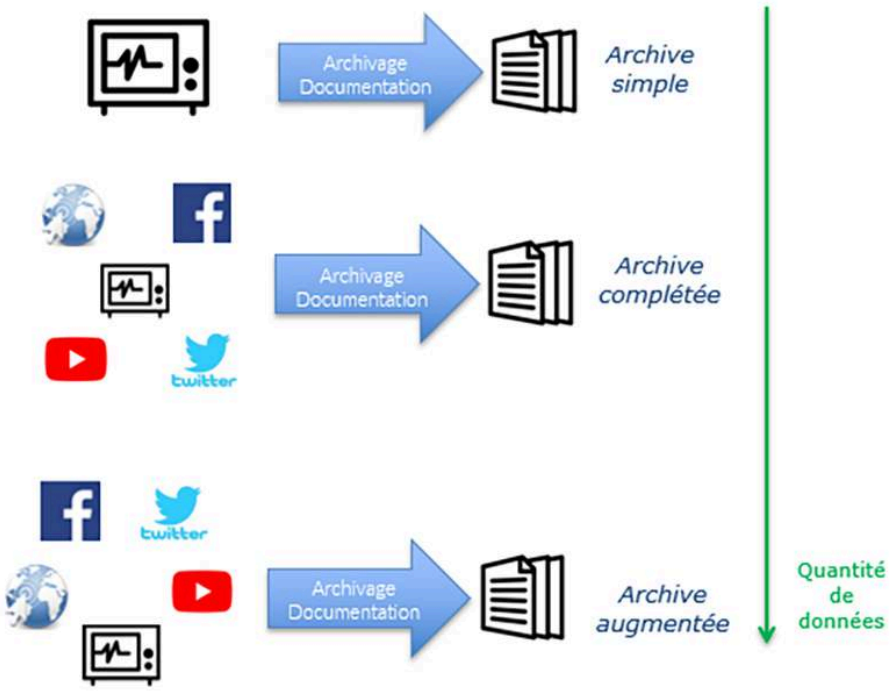
Valorisation et archive augmentée

56 Avant la création d'Internet, les diffuseurs historiques TV/Radio n'utilisaient que le réseau hertzien pour la diffusion de leurs programmes. Ceux-ci pouvaient faire l'objet d'une documentation d'accompagnement (conducteurs d'émission, résumés) généralement disponible au format papier. L'archive était donc mixte, constituée d'un fonds audiovisuel matériel (DVD, cassette) et d'un fonds d'accompagnement papier.

57 L'arrivée d'Internet et sa démocratisation au début des années 2000 ont progressivement mué cette diffusion monocanale hertzienne en diffusion multicanale hertzienne et web. Les diffuseurs ont commencé à diffuser ou rediffuser les émissions en ligne, sur des portails dédiés ou des plateformes tierces, à proposer sur le Web des contenus additionnels aux programmes, et ont progressivement mis en ligne la documentation d'accompagnement. Le programme TV/Radio reste l'objet central qui dicte les stratégies d'éditorialisation, mais il est complété d'objets tiers qui s'adjoignent à l'archive, laquelle devient une **archive complétée**.

58 Aujourd'hui, s'adaptant aux nouvelles pratiques de consommation, la diffusion hertzienne n'est plus LE canal de diffusion, mais UN canal de diffusion. La majorité des programmes est disponible en *replay* sur les applications des chaînes ou *via* leur site web et un programme se décline aussi en ligne *via* un site web ou un compte de réseau social. La documentation d'accompagnement est, quant à elle, intégralement numérique. Le programme n'est donc plus au centre de la stratégie éditoriale des diffuseurs et se présente comme un objet multimédia, multivecteur et multiformat. L'archive TV/Radio originelle devient une **archive augmentée.**

Figure 16. De l'archive simple à l'archive augmentée.

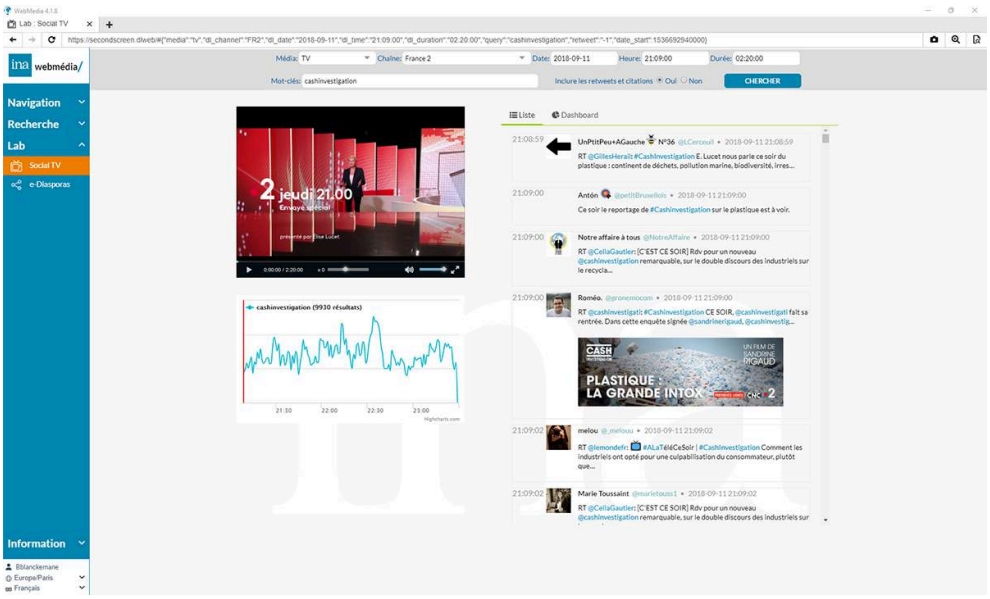


59 Afin de proposer cette archive augmentée à la consultation, la valorisation des données issues de la collecte web consiste à mettre en regard l'archive web et l'archive TV/Radio.

60 C'est l'objet du projet Social TV, développé en interne et disponible dans l'outil de consultation, qui permet de visualiser de manière synchronisée un programme télévisé et les tweets envoyés pendant sa diffusion linéaire, en filtrant sur un *hashtag*, un mot-clé ou un compte.

61 Dans la figure ci-dessous, l'outil présente les échanges qui ont lieu sur Twitter pendant la diffusion d'un numéro de l'émission *Cash Investigation*. La temporalité des tweets affichés correspond à la temporalité de l'émission.

Figure 17. Social TV, un outil permettant de créer des passerelles entre l'archive TV/Radio et l'archive web.



Influence de la gestion des métadonnées sur les métiers de l'archive

62 L'aspect massif des omnidonnées entraîne nécessairement une redéfinition des métiers de l'archive. De plus en plus centrés sur la donnée, ceux-ci ne peuvent faire l'économie de la maîtrise, ou tout au moins de la connaissance, des concepts et des outils qui permettent d'appréhender au mieux ces omnidonnées.

63 Pour mener à bien ses missions, le professionnel de l'archive massive doit donc connaître *a minima* la structure et le contenu des omnidonnées des objets archivés. Sans en maîtriser parfaitement tous les aspects, il doit aussi connaître les offres technologiques afin de pouvoir communiquer aisément avec les équipes techniques (équipe Recherche & Développement, Maîtrise d'œuvre, *Data Scientists*, *Data Managers*), participer à la conception des modèles de données et choisir les outils qui lui permettront de manipuler, d'ordonner et de normaliser ces métadonnées pour être capable de la valoriser de façon pertinente et efficace auprès des usagers.

64 À l'Ina, cette importance des omnidonnées est prise en compte depuis plusieurs années. La récente réorganisation des services a été pensée et structurée afin d'intégrer au mieux les omnidonnées dans les différents aspects de sa mission patrimoniale et audiovisuelle. En outre, afin d'accompagner les documentalistes vers une approche de plus en plus data-centrée de leurs métiers, l'Ina leur propose des formations (à la demande) sur les outils et méthodologies (langages de programmation, langage de requêtage de base de données) et les utilisations (datavisualisation, *text-mining*) possibles des omnidonnées liées à l'audiovisuel.

65 Enfin, sont organisés pour les documentalistes des séminaires d'acculturation et des ateliers de réflexions. Ces derniers permettent non seulement de proposer aux documentalistes des pistes de réflexion pour redéfinir avec eux et au mieux les nouvelles prérogatives de leurs métiers, mais aussi de prendre en compte leurs idées,

leurs suggestions, leurs éventuels points de blocage afin de faciliter et d'œuvrer à leur côté pendant cette transition de leurs métiers.

NOTES

1. Du latin *omnis* (tout), l'omnidonnée représente aussi bien la donnée, la métadonnée et le couple (donnée, métadonnée(s) liée(s)).

2. Décret n° 2011-1904 du 19 décembre 2011 relatif au dépôt légal, *Journal officiel*, n° 0295 du 21 décembre 2011

3. Documentation unifiée et management de la base des objets web.

4. Débit binaire : mesure de la quantité de données numériques transmises par unité de temps. Plus la valeur de bitrate est élevée, meilleure est la qualité de la vidéo.

5. Format de date dans lequel la date est représentée en nombre de secondes par rapport à une date initiale (en général le 1er janvier 1970). Exemple : 1181739828 correspond au 13 juin 2007 15:03:48.

6. Format de date fondé le calendrier grégorien et le système horaire de 24 heures. Exemple : 2011-01-04T00:45:42Z.

7. Source : https://generationcloud.fr/post/quel-type-de-disque-dur-choisir-ssd-vs-hdd-vs-sshd-et-2-5-vs-3-5-cas-pratiques, consulté le 15 octobre 2022.

8. *Digital Archive File Format*.

9. Pour chaque source on compte 10 à 15 éléments. Par exemple pour un compte Twitter, il s'agit du nom du compte, du nombre d'abonnés, de la date de création, etc.

10. Ensemble de composants logiciels et d'outils de développement permettant de mettre en place une architecture logicielle complète et évolutive.

11. Proposée par certains éditeurs de texte, la coloration syntaxique est un outil visuel qui permet de formater automatiquement chacun des éléments du texte affiché en utilisant une couleur caractéristique de son type. Par exemple : la coloration syntaxique d'un document HTML permet de distinguer les balises HTML.

12. Il s'agit de trois moteurs Elastic Search. Chacun dispose de son propre index et de sa propre « machinerie » (serveurs, stockage, etc.).

13. https://madi.hypotheses.org/883.

14. https://asap.hypotheses.org/tag/bataclan.

15. https://bodycapital.unistra.fr/.

16. https://inatheque.hypotheses.org/files/2020/10/INADlweb-Etude-Twitter-Coronavirus.pdf.

AUTEUR

BORIS BLANCKEMANE

Chef de projet au sein du service du dépôt légal du Web (direction des collections, Institut national de l'audiovisuel/Ina)

« Toucher » le public : nouveaux modes d'interaction par le numérique

Rosine Lheureux et Julia Moro

1 Objets polysémiques par excellence, vecteurs d'information d'ordre administratif comme historique ou mémoriel, les archives impliquent la participation active de qui s'y intéresse, par obligation pour faire valoir un droit, ou par choix. Par cette implication même, conjuguée à la médiation indispensable des archivistes, sur site comme à distance grâce au numérique, les archives imposent échange et dialogue, se prêtant idéalement au partage de découvertes, d'expériences et de connaissances.

2 Pour les Archives départementales, progressivement mais inexorablement désertées par leur public traditionnel (lecteurs en salle, visiteurs d'exposition), repenser le rôle d'équipement culturel de proximité suppose de réfléchir à leur mission première de démocratisation des connaissances ainsi qu'aux enjeux de citoyenneté que représente l'inscription dans un territoire précis comme dans le temps chahuté de ce début de siècle, et de songer à ce qui peut faire lien, pour et par la mémoire, si par mémoire on entend la part subjective de l'Histoire qui, mise en commun, favorise l'appropriation de celle-ci par d'autres.

3 Démultipliant les voies de médiation, les nouveaux modes d'interaction offerts par le numérique permettent de « toucher » le public en temps réel avec le matériau protéiforme des archives. Grâce aux archives, parce que celles-ci font écho en chacun, – et il appartient aux archivistes de faire en sorte qu'elles fassent écho auprès du plus grand nombre –, la mobilisation de l'émotion, loin d'être asservie à l'immédiateté de la réaction, favorise le souvenir, la réflexion et la perception du temps long de l'Histoire dans un espace donné.

4 Dès leur création, consécutive à la naissance des départements de la petite couronne parisienne, les Archives départementales du Val-de-Marne ont mobilisé les techniques les plus récentes afin d'accomplir leur mission de réunir et enrichir les archives de ce territoire administrativement neuf et en quête d'une identité commune pour les

habitants de son ressort. En suscitant et conservant d'abondantes ressources audiovisuelles, films et témoignages des bouleversements profonds de la quasi-totalité du département durant la seconde moitié du XX^e^ siècle, elles ont bâti ce patrimoine local du temps présent, pendant plus accessible au grand public que les archives administratives traditionnelles.

5 Exploitant à présent les potentialités du Web, elles cherchent à améliorer la diffusion de leurs ressources par le biais de l'interactivité avec le public. Deux expériences nées des contraintes de la crise sanitaire se prêtent à un bilan : l'une, l'opération Mémoire de confinement, fut une expérience unique, l'autre, l'animation des réseaux sociaux, démultipliée durant ce même confinement, s'inscrit dans le quotidien du service.

L'opération Mémoire de confinement

6 Le 17 mars 2020, la France fut brutalement mise à l'arrêt en raison de la pandémie de Covid-19. À l'exception des personnes obligées de travailler à l'extérieur dans une grande insécurité sanitaire, chacun fut confiné, soumis au respect de règles de vie très restrictives excluant quasiment les sorties, renvoyé à la sphère de l'intime mais en contact permanent, grâce aux médias et aux réseaux, avec l'extérieur plongé dans le même désarroi. Cet événement traumatique majeur, d'une durée initiale indéterminée, se prolongea jusqu'au 11 mai, date de la levée du confinement et d'une reprise de l'activité, dans un contexte qui resta dégradé plus d'une année.

7 La collecte de témoignages s'imposa immédiatement au directeur des Archives départementales des Vosges qui lança un appel sur Twitter dès les premiers jours du confinement et fut suivi rapidement par quelques autres services d'Archives départementales ou municipales. Sa nécessité ne se fit jour dans le Val-de-Marne qu'au terme de quelques semaines, quand il apparut, alors que le confinement se prolongeait, que les gens avaient passé l'étape de la sidération et s'étaient organisés pour vivre le moins mal possible ce temps suspendu. Il fallait également que le service des Archives, qui avait basculé dans le travail à distance sans aucune expérience préalable du télétravail, fût capable de prendre en charge une telle opération.

8 Lancé le 17 avril, au terme du premier mois de confinement, l'appel fit l'objet d'un communiqué de presse du président du département et fut relayé sur le site de la collectivité et celui des Archives. Unique sous cette forme en Île-de-France, il retint également l'attention des radios et de la presse[1].

Pourquoi ?

9 Pour obtenir l'autorisation de cette collecte de traces immédiates du confinement, il a fallu convaincre de sa spécificité par rapport aux projets ludiques portés par les villes ou les offices de tourisme[2] et de son intérêt majeur.

10 Sur le moment, il s'agissait tout d'abord de faire œuvre de mémoire commune dans un territoire défini, celui d'un département francilien très peuplé, contrasté géographiquement et socialement et dont la population est d'ordinaire extrêmement mobile, les actifs travaillant majoritairement hors de leur commune d'origine. Les premiers mots de l'appel affichent clairement cette intention : « Vous êtes chez vous ou bien mobilisés pour répondre aux nécessités d'urgence : partagez votre quotidien avec

nous, il deviendra la mémoire de demain ! Participez à l'écriture de notre histoire, enrichissez vos Archives ! Nous collectons vos témoignages pour garder trace de la vie en Val-de-Marne durant cette période unique. »

11 Tout autant, cette collecte avait pour but d'ouvrir un espace empathique d'expression dans ce moment de repli forcé sur la sphère privée pour la plupart mais paradoxalement exposé aux bruits contradictoires et anxiogènes du monde entier. En cela les Archives départementales accomplissent leur mission de proximité, celle-ci ayant eu pour but, en temps de crise, de créer du lien et de permettre à certains participants de rompre leur isolement. L'appel se poursuivait ainsi :

> [...] Durant cette période de confinement et d'incertitude, nous avons tous adapté notre quotidien, modifié nos habitudes, déployé de nouveaux modes de fonctionnement, fait preuve de solidarité et d'entraide, mais aussi d'innovation, d'inventivité et d'originalité ! Cet événement collectif vécu pour chacun dans l'intimité revêt un caractère unique. Parce que nous traversons tous ces moments dans des conditions bien différentes et, parfois, dans la peine, il est important d'en préserver les traces individuelles. Elles sont un pan de l'histoire que nous vivons ensemble. Aussi, contribuez par vos récits, vos créations, témoignez de votre vie, des objets, des activités qui vous aident, que vous soyez seuls ou entourés, et invitez celles et ceux, petits et grands, qui vous accompagnent dans ce confinement, à distance ou auprès de vous, à apporter aussi leur pierre à notre édifice commun, notre mémoire partagée.

12 Pour la suite, ces archives, une fois mises à disposition, devaient permettre à chacun de revenir sur ce temps particulier et servir de matériau pour les chercheurs, se prêtant par exemple à la comparaison avec celles recueillies ailleurs en France.

Quels contenus ?

13 Le délai de réflexion d'un mois avant le lancement de l'opération a permis de s'inspirer des modalités de collecte mises en place ailleurs. Le choix a été fait de ne pas recourir à des formulaires, mais de laisser les donateurs totalement libres de leur envoi, en leur demandant simplement de signaler leur commune d'origine, ce qu'ils ne firent pas toujours, et en leur garantissant l'anonymat lors d'une éventuelle diffusion, ainsi que l'absence de toute sélection. Un seul envoi fut refusé, car il contenait des propos racistes et diffamatoires, et l'envoyeur fut prévenu de ce rejet.

14 Plus de 200 documents ont été reçus d'une centaine de personnes et de plusieurs établissements scolaires, dont la diversité reflète la population valdemarnaise et un peu au-delà, de Parisiens ou d'habitants de départements limitrophes s'étant tournés par défaut vers le seul service départemental francilien à avoir organisé une collecte : élèves, travailleurs « de l'extérieur », parents jonglant entre intendance familiale, télétravail et apprentissage du métier d'enseignant, étudiants, jeunes actifs comme retraités privés subitement de vie sociale, militants associatifs au contact de populations, comme les Roms, invisibilisées lors de la crise.

15 Quelques participants ont rédigé leur témoignage à l'intention des Archives, tel celui, lapidaire, d'une infirmière évoquant sa tristesse d'avoir perdu deux proches du Covid et de devoir sortir pour travailler tous les jours. Mais la grande majorité, parfois lancée dans une sorte de déconfinement intellectuel ou dans une boulimie d'activités domestiques, a envoyé généreusement ses productions, de toute nature et sur tout support numérique, du fichier PDF à la webradio et au padlet[3] : carnets de bord,

journaux intimes ou écrits divers principalement, mais aussi recettes de cuisine, canevas et masques cousus main, détournement d'objets, chansons, playlist, dessins poétiques ou humoristiques, bandes dessinées, photomontages, photographies, vidéos, blogs, contes audio, etc. Ces réalisations livrent un quotidien parfois joyeux lorsque les familles d'ordinaire séparées en semaine profitent de l'occasion d'être ensemble, mais devenant de plus en plus monotone. Elles témoignent de la promiscuité des appartements trop petits ou à l'inverse de la solitude, des pensées mêlées d'anxiété (le décompte jour après jour des morts), de la curiosité, des défis (faire du sport dans son couloir, apprendre le Tamoul), de la colère, de l'incompréhension ou de la résignation, de l'ennui… L'ennui est tangible dans les audioblogs, poèmes et récits très libres d'élèves envoyés par des professeurs de huit collèges et lycées du Val-de-Marne, dans lesquels un jeune adolescent évoque « la boucle infinie du confinement », ou bien quand se disent les difficiles conditions de l'enseignement improvisé à distance dans la webradio d'un lycée professionnel : « Les professeurs, quand vous envoyez vos cours sur Pronote, mettez des détails, parce qu'on n'a pas tous des ordinateurs, moi je suis sur mon téléphone, la plupart on est sur notre téléphone [...]. » En creux, dans l'appel de l'assistante sociale du lycée à tous les élèves majeurs isolés à prendre contact avec elle, se devinent des dangers plus grands que le décrochage scolaire[4].

16 Les modalités de la collecte ont occasionné divers échanges avec les donateurs : leur envoi très libre était suivi d'un nouveau contact pour formaliser le don, les autorisations de diffusion et régler parfois des soucis techniques liés au format d'envoi, en vue de l'archivage. Quelques donateurs ont créé des routines : ainsi, une retraitée du Val-de-Marne, qui avait commencé la rédaction d'un journal sur un carnet pour sa petite-fille qu'elle ne pouvait plus voir, l'a recopié pour l'envoyer en plusieurs livraisons durant tout le confinement. D'autres ont demandé très rapidement la mise en ligne et le partage des témoignages. Pour des personnes peu sollicitées et peu habituées à s'exprimer, cette collecte a comblé un besoin de reconnaissance en leur offrant un moyen simple de faire connaître le bouleversement de leurs vies et, bien plus, de participer à la construction d'un matériau pérenne pour l'Histoire, portée par un service public que très peu connaissaient initialement.

17 La facilité de l'envoi des contributions aux Archives a cependant induit un « biais d'immédiateté » : si les premiers jours les donateurs ont été recontactés dans la journée ou le lendemain pour la formalisation du don, le succès de l'appel a rapidement rallongé les délais et certains, quelques jours après, n'ont plus répondu, comme ces jeunes contributeurs ayant envoyé leur rap du Covid que l'on n'a pas pu retrouver. Un autre frein à la collecte fut le tout numérique. Les incitations à envoyer également des productions par courrier sont restées vaines, seuls deux envois sont parvenus aux Archives, dont l'un était un carnet de croquis déjà reçu sous forme de fichier durant le confinement.

Les aspects techniques de l'archivage d'une collecte « à portée de clic »

18 Opérer la mise en archives a consisté à obtenir confirmation des dons auprès de chaque donateur, s'assurer de la pérennité des supports numériques reçus, les référencer, en envisager la valorisation.

19 La chaîne opératoire mise en place à distance a reposé :

- pour la collecte même, sur l'archiviste en charge d'ordinaire de la salle de lecture désignée comme responsable (réception et inscription dans le tableau des entrées, visionnage, relations avec les donateurs) et la webmestre (médiation de la collecte puis valorisation) ;
- pour l'archivage, sur le chef de projet archives électroniques, aidé de la régie audiovisuelle pour les formats spécifiques et des archivistes en charge des fonds privés et de la collecte des archives des établissements scolaires pour l'indexation.

20 Ce projet commun fut bénéfique pour la cohésion des équipes alors que chacun travaillait à domicile. Le travail de reprise des données en vue de leur conservation, qui s'est prolongé bien après la collecte, a permis de tester l'archivage de formats parfois peu habituels, alors que le service envisageait de se doter d'un système d'archivage électronique, acquis en 2022.

21 Les témoignages ont été archivés par donateur, accompagnés de leurs messages d'envoi. La consigne donnée dans l'appel aux dons, envoi de fichiers PDF ou JPEG (1 Mo maximum pour les photographies et 10 Mo pour les vidéos) n'a guère été suivie.

22 Quelques schémas permettent de comprendre le processus d'archivage[5] :

Figure 1. Réception et copie.

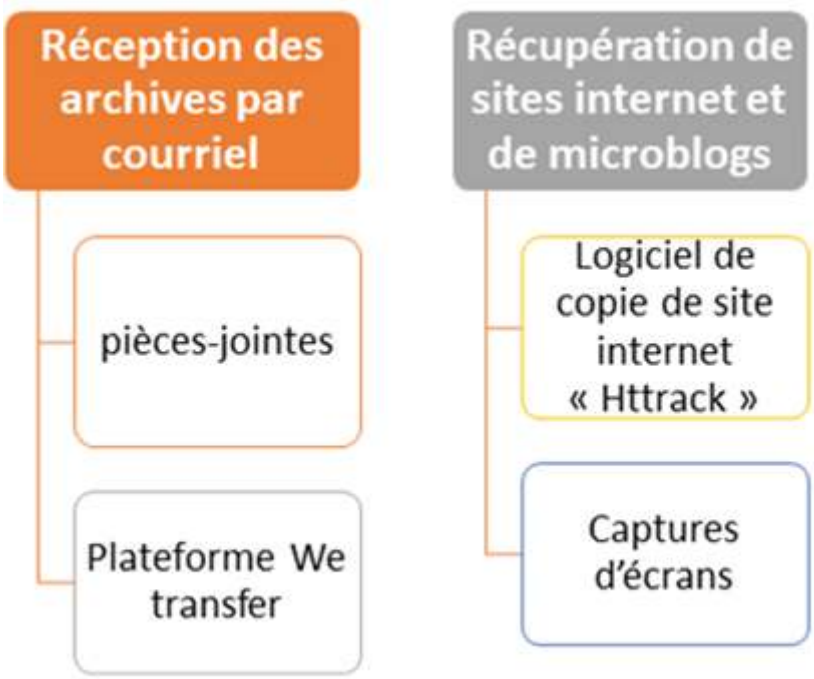


Figure 2. Typologie technique des archives collectées : une majorité de fichiers numériques bureautiques et audiovisuels isolés.

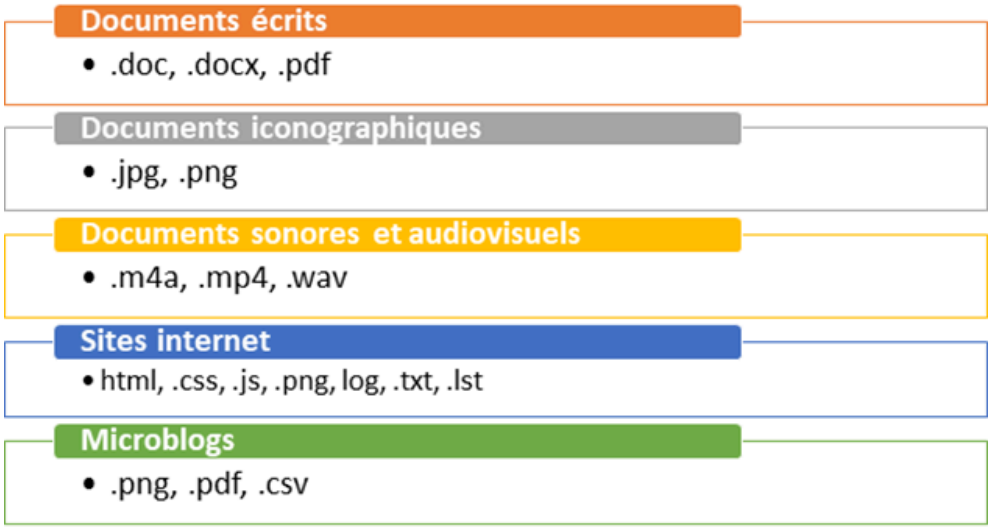


Figure 3. La conservation : migration et stockage.

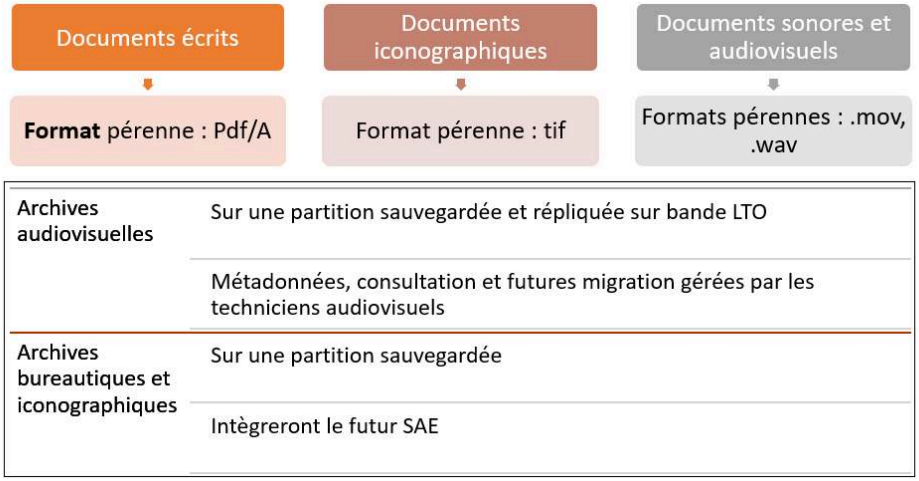
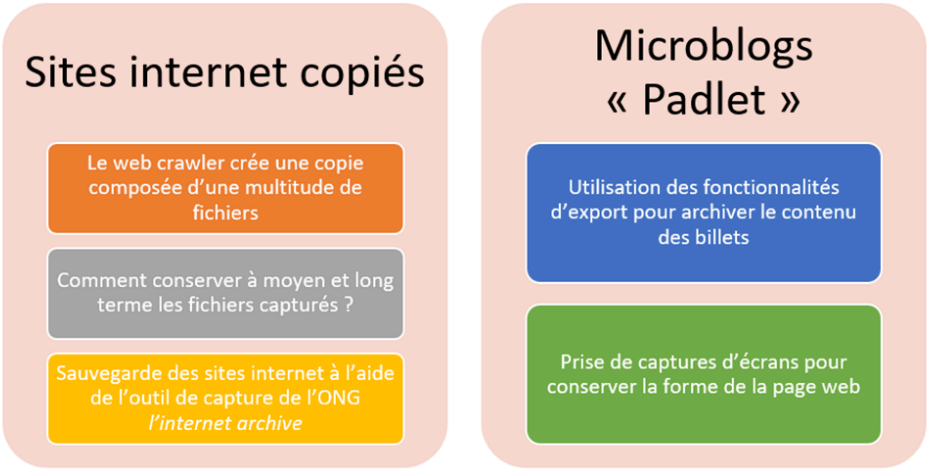


Figure 4. L'archivage du Web.



23 L'archivage des sites Internet constitue un défi pour les Archives départementales qui ne sont pas outillées pour assurer leur conservation pérenne et leur communication. La copie des sites hors connexion semble fonctionner actuellement, mais pour combien de temps ? La copie de sauvegarde créée *via* la *wayback machine* de l'ONG *Internet archive* ne constitue pas une solution viable pour un service public d'archives. Plus complexe encore est l'archivage des microblogs Padlet : la plateforme Padlet propose des modalités d'exports peu avancées permettant de récupérer séparément les images, les vidéos ou le contenu écrit, mais il est impossible de restituer la forme et l'interactivité du blog et il faut compenser cette perte à l'aide de captures d'écrans. Ce mode d'archivage dégradé trouve ses limites en transformant un objet d'archives cohérent en plusieurs sous-objets à réassembler.

Le résultat : les archives de la mémoire du confinement dans le Val-de-Marne

24 Le produit de la collecte est évalué à 200 témoignages, mais ce nombre est sans doute supérieur, car certains documents, comme le Padlet d'un collège, en comportent plusieurs dizaines.

25 144 fichiers informatiques différents (4,31 Go), collectés auprès de plus d'une centaine de contributeurs particuliers, ont donné lieu à 39 entrées d'archives privées, complétés par 7 versements d'archives publiques provenant d'une école élémentaire, de 4 collèges et de 2 lycées (2,60 Go). 19 fichiers vidéos et 30 fichiers audios ont été traités, ainsi que 2 documents sur support papier.

26 En décembre 2021, il restait encore 12 articles non traités faute d'avoir pu renouer le contact avec les donateurs pour la formalisation du don. Dans un souci d'exhaustivité du fonds, il a été décidé de les archiver également et d'en permettre la communication, puisque l'appel à projets précisait qu'on s'en réservait le droit à condition de les anonymiser, mais seulement aux Archives, sans diffusion en ligne.

27 Aucun terme officiel n'a été mis à la collecte et quelques rares témoignages sont encore entrés après juillet 2020. Il a été décidé de ne pas relancer la démarche lors des confinements ultérieurs qui n'avaient plus la rigueur et le caractère d'exceptionnalité du premier. Une restitution a eu lieu dès l'automne 2020, sous la forme d'un portail

présentant un échantillon des documents reçus réunis par thème[6] : s'occuper ; apprendre et enseigner ; situations insolites ; réfléchir et se questionner. Le dernier thème, retour sur le déconfinement, a été ajouté à la date anniversaire du début du confinement, le 16 mars 2021.

Bilan

28 La question de savoir si les documents collectés sont ou non des archives au sens classiquement admis par les professionnels a été posée, fait encore débat, mais n'enlève rien au fait que ces traces mémorielles du confinement, adressées aux Archives départementales et « mises en archives » par celles-ci selon les procédés archivistiques en vigueur, font en tout cas Histoire, au même titre que celles collectées ou à collecter auprès des services, organismes, associations ayant fait face à la crise.

29 Cette collecte ne masque pas l'impossibilité de collecter autre chose, ailleurs et plus tard. Bien des archives de l'année 2020 se trouvent encore dans les services versants pour quelques années. Un regret persiste cependant, celui de n'être pas parvenu à intéresser les nombreuses associations du Val-de-Marne venues en aide à la population, bien qu'un appel ait été lancé à leur intention durant l'été 2020, appel peut-être trop précoce. En revanche, à l'occasion du changement de mandature en 2021, sont entrées des archives d'élus et de directeurs en poste au printemps 2020 et relatives à la gestion de la crise sanitaire.

30 À propos des opérations de collecte lancées en 2020, la comparaison avec la Grande Collecte d'archives de la Première Guerre mondiale en 2014 permet de mesurer quelques changements profonds : en 2014, une collecte nationale longuement préparée d'archives dont le temps avait fait le tri, de traces de souvenirs distanciés apportés par des donateurs séparés des producteurs par plusieurs générations, de documents sur support classique papier ou photographique ; en 2020, des collectes dispersées, immédiates, modestes, de témoignages de première main qui, posent très concrètement la question de ce qui restera de notre société numérique. Elles furent un moyen, certes exploratoire, empirique et imparfait, de fixer l'empreinte volatile de milliers de vies banales percutées par un événement extraordinaire.

Se réinventer pendant la crise : le rôle et l'usage du Web et des réseaux sociaux

31 Dès mars 2020 et durant toute la crise sanitaire, s'adapter, se réinventer et continuer à exister malgré l'absence de relations physiques avec l'usager ont été les maîtres mots des Archives départementales du Val-de-Marne, et plus généralement des institutions culturelles. Le parti retenu pour les Archives a été celui d'investir massivement le Web et les réseaux sociaux. Déjà présent sur la toile grâce à son site Internet, deux réseaux sociaux majeurs (Facebook et Twitter[7]) et deux médias sociaux (Viméo et Soundcloud), l'institution a fait le choix dès l'annonce du premier confinement de réinterroger chacun de ses outils de manière à construire une nouvelle relation aux publics. Au temps suspendu, elle a saisi son double rôle social et culturel en souhaitant offrir aux lecteurs confinés derrière leur écran la possibilité de s'instruire, mais aussi de se

distraire et de communiquer entre eux autour de l'histoire locale, de leurs rapports au territoire et plus généralement à l'archive.

Programmation au Web pour rester visible

32 Au printemps 2020 la programmation culturelle des Archives départementales du Val-de-Marne prévoyait notamment une saison « Les Archives font leur cinéma ». Reposant sur la diffusion dans leur hall d'exposition, transformé provisoirement en salle de cinéma, d'une bibliothèque audiovisuelle de 176 films découpés en plusieurs cycles de projections, l'initiative devait mettre en lumière la diversité des thématiques représentées dans les collections des Archives départementales :

- Regards sur les transports
- Regards sur la ville
- Regards sur la jeunesse
- Regards sur la culture et le patrimoine
- Regards sur les migrations et la mémoire

33 Continuer à être visible, marquer la présence de la vie malgré ce temps d'arrêt brutal dans ce qui fait l'essence même de la société a été le leitmotiv affirmé de la direction. Aussi, n'a-t-on pas réfléchi au report de l'initiative à une date ultérieure et encore bien incertaine mais plutôt à la mise à disposition rapide des contenus sur le Web.

34 Dans une première phase expérimentale, en mars et avril 2020, des extraits de films ont ainsi été présentés sur les réseaux sociaux, manière de donner à voir et à entendre la richesse d'une collection de plus de 18 000 documents, soit plusieurs milliers d'heures de films et d'enregistrements sonores provenant aussi bien des services du conseil départemental, de l'État, des communes, mais aussi d'entreprises, d'associations ou de familles du département[8]. Pour contenter les plus curieux, mais aussi dans l'optique d'inciter davantage le public des réseaux sociaux à approfondir sa connaissance des archives, une page recensant l'intégralité des films proposés au cours du cycle a également été créée sur le site Internet des Archives départementales, vitrine des activités et missions de la direction[9].

35 Dans un second temps, en octobre 2020, après l'annonce d'un nouveau confinement et fort de cette première expérience, le choix s'est porté sur l'organisation de rendez-vous « Facebook première[10] » permettant de programmer des films préenregistrés et de les diffuser en direct sur le réseau social. Cette opération « #ConfinésMaisPasPrivésDeCiné » a permis à l'usager, peu enclin à quitter son réseau social pour basculer sur le site Internet des Archives, de visionner l'intégralité des films disponibles et d'interagir directement sur la page Facebook pendant les diffusions. Avec en moyenne une vingtaine de personnes présentes en simultané lors de ces « direct », et des réactions, commentaires et partages pouvant toucher jusqu'à 2 000 personnes, une forte dynamique est enclenchée. Des records sont notamment enregistrés pour les films du cycle consacré aux communes du Val-de-Marne : les documentaires sur Orly, Villejuif ou encore Ivry-sur-Seine ont attiré des habitants attachés à voir ou revoir leur ville à une époque révolue et à en apprendre davantage sur son histoire. L'outil de communication qu'est Facebook devient alors, au-delà du partage des connaissances, un nouveau vecteur de lien social et d'appropriation de l'histoire locale. Il a cette force, contrairement au site Internet, parfois trop institutionnel, trop aride, de revisiter la relation avec l'usager par un dialogue ouvert. Il renverse le rapport de passivité du

lecteur pensé comme un réceptacle d'information ou de contenu pour lui offrir la possibilité d'exprimer son point de vue ou d'apporter à son tour des éléments de compréhension de l'histoire.

36 Les restrictions sanitaires courant jusqu'au printemps 2021, empêchant notamment la réouverture des établissements culturels aux publics, le problème d'adaptabilité au format numérique s'est également posé pour la sortie de l'exposition *+ 2 °C ? Les Val-de-Marnais, le climat et l'environnement, 1780-1945* prévue en mars 2021. Dans une même logique que le travail réalisé durant l'année 2020, l'objectif de la direction a été de ne pas priver le public des contenus réalisés. L'exposition a donc été adaptée sous une forme immersive avec un format de visite innovant à « 360° [11] », ainsi que sous une forme « virtuelle », véritable déclinaison numérique du catalogue de l'exposition agrémenté de contenus multimédia (présentation d'iconographies à 360°, mises en voix de textes d'archives) et ludique (puzzles, mots mêlés)[12]. Relayés sur les réseaux et invitant toujours davantage le lecteur non initié aux archives à la découverte et au dépassement des idées reçues sur ce qu'est et ce que propose un service d'archives, ces nouvelles formes de déclinaisons numériques ont démultiplié le nombre de visiteurs habituels des expositions. À la fin de l'année 2021, on comptabilise ainsi 926 visiteurs pour l'exposition à 360° et 867 visiteurs pour l'exposition virtuelle, chiffres qui ne cessent encore d'augmenter aujourd'hui grâce à la disponibilité offerte par le format Web malgré la fermeture physique de l'exposition en juillet 2022.

37 Cette nouvelle offre culturelle 100 % numérique, née de la volonté de préserver la valorisation de l'activité, a de multiples effets positifs. Les vues sur le site Internet connaissent une augmentation constante depuis le confinement (+ 32 % de pages vues sur le site Internet en 2020 par rapport à 2019, avec notamment une croissance de 82 % pendant le premier confinement par rapport à 2019 sur la même période). Même chose pour les réseaux sociaux : le nombre d'abonnés a doublé sur Facebook entre mars 2020 et octobre 2022 (2 000 à 4 613 abonnés avec un pic d'augmentation pendant le premier confinement à + 260). Sur Twitter, réseau social plus professionnel, la croissance est un peu plus timorée avec environ 700 abonnés supplémentaires depuis 2020 (135 nouveaux abonnés pendant le confinement), soit 1 144 en octobre 2022. À cela s'ajoutent l'arrivée d'un nouveau public jusqu'à présent peu familier des archives, des demandes croissantes de nouveaux contenus, une personnalisation du rapport à l'archive et à l'archiviste, l'instauration d'un dialogue de plus en plus régulier et profond entre les *followers* et le *community manager* des archives. À l'aune de ces nouvelles pratiques, c'est donc toute la stratégie des Archives départementales qui se trouve réinterrogée. *A posteriori*, il est bien établi que les visiteurs en ligne ont été plus importants que le public physique habituel. L'ouverture d'événements en ligne révèle un public insoupçonné, intéressé et curieux mais peu enclin à se déplacer. Cela nous amène donc à repenser complètement l'organisation des initiatives culturelles non plus dans un cloisonnement des événements pensés comme soit physiques soit virtuels, mais dans une optique de capitalisation des publics en offrant aux usagers les deux possibilités sur des objets similaires.

Conquérir de nouveaux publics, accroître sa présence

38 Jusqu'en mars 2020, les Archives départementales publiaient en moyenne 3 posts par semaine sur les réseaux sociaux Facebook et Twitter. Le confinement a fait ressortir la

nécessité de distraire, d'informer le public habituel, mais aussi d'attirer de nouveaux *followers*. Dans cette optique, le rythme des publications s'est intensifié en passant à 1 post par jour minimum et parfois jusqu'à 3 par jour, avec des contenus ludiques ou informatifs, programmés à l'avance ou réalisés sous le coup de la spontanéité et des demandes. Rapidement un constat s'est fait jour : le confinement a bouleversé les pratiques des *followers* demandant toujours plus de réactivité et d'adaptabilité. Plus le nombre de publications augmentait, plus le besoin d'échanger s'est fait sentir. On vient désormais sur la page Facebook des archives comme si l'on allait au cinéma, au théâtre ou au musée, etc., pour rechercher une forme de plaisir, d'évasion ou d'instruction. L'archive sur les réseaux sociaux devient le livre, les posts défilent comme les pages se tournent, avec toujours le même empressement d'arriver au contenu suivant.

39 Afin de répondre à cette présence accrue sur les réseaux sociaux en période de pandémie, plusieurs initiatives destinées à distraire les internautes tout en leur apportant une connaissance des fonds et des missions d'un service d'archives ont été imaginées par les Archives départementales en adoptant notamment le *hashtag* « #CultureChezNous » lancé par le ministère de la Culture[13], puis par celui imaginé par le service communication de la mairie de Crépy-en-Valois, « #RemèdesContreLaMorosité », repris de manière virale dans la communication publique des collectivités et établissements culturels. Puzzles, mots croisés et mots-mêlés[14] ont ainsi fait leur apparition sur les fils Twitter et Facebook des Archives départementales, aux côtés des jeux déjà expérimentés avant le confinement sur les photographies et cartes postales (exemple : « Saurez-vous reconnaître la ville », etc.). L'insertion dans la « *battle* de cocottes[15] » lancée par les Archives de la ville et de la métropole d'Orléans a également été un temps particulièrement appréciable d'échanges avec le public mais aussi entre la communauté des archivistes qui, pour faire face à la situation, a su faire preuve d'imagination, de cohésion et de dialogue. Ces nouvelles publications, qui ramènent fraîcheur et bonne humeur, ont permis de corriger l'image parfois trop austère des archives. Elles ont montré les différentes utilisations qui peuvent être faites des fonds et ont ouvert également la voie à un usage familial des archives sur les réseaux sociaux (puzzles, cocottes en particulier[16]). Fort de ce nouveau public captif, aux intérêts souvent différents pour l'archive (certains attendent le puzzle, d'autre plutôt une image à découvrir ou à identifier), un travail de fidélisation a été entamé par la création de nouveaux rendez-vous fixes, repris désormais pour certains de manière régulière : #PromenadeDuDimanche, #MercrediPuzzle, #LAfficheDeLaSemaine, #ConfinésMaisPasPrivéDeCiné, #MondayMotivation #Jeu, #QuestionPourUnChampion.

40 L'ensemble de ces publications a mis en évidence le rôle social des archives. La diffusion de fonds iconographiques et audiovisuels est un moment privilégié entre l'archiviste et le *follower*. Ces fonds tissent un dialogue régulier autour de la mémoire et de l'histoire du territoire, souvent empreint de nostalgie et d'anecdotes (évocations des souvenirs, de l'enfance, personnes qui se retrouvent, etc.). Preuve de l'importance de la construction et de l'entretien de ces nouvelles relations, l'inversion sur la connaissance du territoire se fait naturellement. Le *follower* corrige, rectifie et enrichit les descriptions, rappelle des événements marquants sur l'histoire de sa commune ou du département. Ce rapport de réciprocité qui se fait jour entre l'archiviste et l'usager actif s'apprécie d'autant qu'il peut déboucher sur un travail de collecte d'archives papier ou orales.

41 La valorisation des fonds d'archives sur les réseaux a ceci d'intéressant qu'elle fait parler l'archive. L'internaute nous livre « ses » sentiments sur l'évolution de sa ville, de son quartier, de son département, documentant la parole méconnue ou peu connue des anonymes dans l'histoire du quotidien, de l'aménagement, de l'urbanisme, de l'évolution des rapports sociaux, etc. Tous ces témoignages recueillis sous forme de « commentaire », plongée fascinante dans l'histoire des mentalités et des représentations, sont autant de nouvelles sources, de nouveaux matériaux pour l'historien cherchant à approcher une histoire urbaine et socio-culturelle de l'intime et du sensible. Après le temps de l'enquête orale vient ainsi pour l'archiviste le temps de l'enquête écrite (même si elle n'en porte pas le nom) portée par la diffusion de contenus qu'il n'appartient plus seulement à l'historien ou à l'érudit local de commenter. Quand l'archive devient créatrice d'archives se pose alors d'inévitables questions sur lesquelles il convient désormais de se pencher attentivement sans quoi le fil de cette nouvelle relation tomberait dans l'anonymat des méandres du Web : quelles méthodes adopter pour saisir et rendre compte des paroles enfouies dans les fils des réseaux sociaux ? Faut-il repenser nos modes de communication pour mieux appréhender ces paroles offertes comme l'on mènerait une enquête orale ? Faut-il susciter volontairement la parole de l'internaute, au risque de briser la spontanéité de son commentaire et donc de le dénaturer ?

42 Le succès évident de cet investissement sur le Web et les réseaux sociaux a bousculé nos codes, nos manières de travailler et de percevoir nos relations avec des *followers* venus à nous pour leur simple loisir et ne connaissant souvent rien de nos missions. Il nous a semblé pertinent de leur proposer, par de nouveaux contenus, un aperçu quotidien de notre travail. « #EnCoulisses », « #LesExperts », nouveaux rendez-vous réguliers des réseaux sociaux, donnent ainsi à voir l'archiviste en action dans ses différents domaines de compétences et permet de gommer le caractère parfois impersonnel des réseaux sociaux institutionnel.

43 Les relations plus proches et régulières avec le public impliquent un investissement humain accru, un ajustement du temps de travail[17], mais aussi une remise en cause permanente de l'usage des réseaux et de nos pratiques. Connaître son public, s'adapter aux nouvelles modes et aux nouvelles tendances du Web, c'est aussi accepter de multiplier les expériences et les présences avec, en ligne de mire, toujours le même vœu : amener de nouveaux publics à la culture. L'ouverture d'un compte Instagram en septembre 2022 marque une autre étape dans cette réflexion permanente.

Conclusion

44 La démarche de collecte participative de l'opération Mémoire de confinement, comme la démultiplication de l'offre sur les réseaux sociaux Facebook et Twitter, auxquels s'est ajouté récemment Instagram, favorise des modes d'interaction nouveaux avec un public lui aussi nouveau. Reposant sur la facilité et l'immédiateté d'accès pour le *follower*, les réseaux n'en sont pas pour autant superficiels et fondent leur attractivité sur le sérieux de contenus proposés de façon ludique. Le public des réseaux, que peut aussi attirer un appel à témoignage ou à collecte numérique, n'est plus celui des Archives, ni même celui du site Internet de celles-ci, où sont également proposés des contenus de valorisation culturelle et plusieurs chantiers d'indexation collaborative de moins en moins alimentés. Il consomme mais donne également du contenu qu'il faut

qu'à notre tour, nous, les professionnels des archives, apprenions à appréhender avec le recul nécessaire. Ce contenu ne servira peut-être pas tant à la connaissance des archives dont nous avons la charge qu'à documenter directement l'histoire du territoire dans lequel il s'inscrit dans un temps donné.

45 Se déporter sur de nouvelles pratiques, aller vers un autre public, encore largement méconnu, partager avec lui l'apport des Archives sans forcément chercher à l'y amener, pour frustrant que ce puisse sembler, revient à honorer la mission de l'archiviste qui motive le fait de conserver sur des kilomètres linéaires et des serveurs informatiques une part de la mémoire collective, celle d'offrir au plus grand nombre, en se mettant à sa portée, le patrimoine dont il a la charge, sous quelque forme que ce soit et par les moyens les mieux adaptés, et ce quel qu'en soit l'usage qui en est fait.

NOTES

1. Entre avril et juillet 2020, France Bleue locale et nationale, France Inter, *Le Point*, *Le Parisien*, *Citoyens94* et *Sortir Télérama*.

2. Tels que « Prenez une photographie depuis votre fenêtre. »

3. Mur de messages virtuel.

4. Arch. dép. Val-de-Marne, 4487 W 3. Témoignage de confinement, Radio-Lycée Val-de-Bièvre, premier enregistrement, appel d'une lycéenne en 2e année de C.A.P.

5. Cette description est due à Kamel Hamichi, chef de projet archives électroniques des Archives départementales du Val-de-Marne.

6. https://archives.valdemarne.fr/r/328/retour-sur-la-collecte-/

7. Les Archives départementales du Val-de-Marne disposent d'un compte Facebook depuis 2014 et d'un compte Twitter depuis 2019.

8. En 2018, les Archives départementales ont décidé de mettre progressivement en ligne leur collection d'archives audiovisuelles. Actuellement, une partie des séries 3 et 13AV (services du conseil départemental et de l'État) sont consultables en ligne : https://archives.valdemarne.fr/f/FondAudioVisuels/tableau/?&reset_facette=1.

9. https://archives.valdemarne.fr/r/287/les-archives-font-leur-cinema/.

10. Facebook première permet, outre la diffusion en direct, de conserver ensuite le film projeté dans le fil d'actualité de la page du réseau social : le public peut ainsi revoir le film à volonté et continuer à commenter la publication sur Facebook.

11. https://archives.valdemarne.fr/r/357/.

12. https://archives.valdemarne.fr/r/356/l-exposition-virtuelle/.

13. Ce hashtag permettait d'identifier les contenus créés spécifiquement pendant le confinement.

14. https://archives.valdemarne.fr/r/289/a-vous-de-jouer-/.

15. https://archives.valdemarne.fr/r/447/operation-battle-cocotte-/.Il s'agit d'une adaptation du célèbre jeu des cocottes en papier qui permet, dans le cadre des archives, de tester ses connaissances sur l'histoire et la mémoire d'un territoire. Chaque bord permet ainsi de tirer une question à laquelle est bien sûr rattachée la bonne réponse.

16. Fortes de cette expérience, les Archives départementales ont depuis également proposé des coloriages ainsi qu'un jeu des sept différences : https://archives.valdemarne.fr/r/398/coloriages/ ; https://archives.valdemarne.fr/r/400/jeux-des-sept-differences/.
17. La webmestre et *community manager* consacre désormais deux jours par semaine aux réseaux au lieu d'un avant 2020 et poste également hors de son temps de travail (week-end et jours fériés).

AUTEURS

ROSINE LHEUREUX

Directrice des Archives départementales du Val-de-Marne

JULIA MORO

Webmestre et *community manager* aux Archives départementales du Val-de-Marne

Remerciements

1 Ce livre collectif a pu voir le jour grâce au soutien du Laboratoire d'Excellence d'Histoire et anthropologie des savoirs, des techniques et croyances, de l'École pratique des hautes études – Université PSL, partenaire fidèle du séminaire « Les nouveaux paradigmes de l'archive », que nous remercions chaleureusement.

2 Notre gratitude va également à la Mission de la diffusion scientifique aux Archives nationales. Les relectures rigoureuses de Claire Béchu-Bénazet ont été décisives pour la qualité de cette publication.

3 Merci, enfin, au Dicen-IDF, Cnam/Paris pour son investissement dans la direction de l'ouvrage et surtout aux autrices et auteurs qui, à plusieurs reprises, nous ont fait l'honneur de partager leur savoir dans le cadre du séminaire et de son ouvrage.

4 Nous portons une mention particulière à Olivier Poncet, de l'École nationale des chartes, à qui est revenue la tâche délicate d'introduire l'ensemble.