L'histoire du Cédric : penser un dispositif archivistique en histoire des sciences

Gérald Kembellec, Raphaël Fournier-S'Niehotta, Pierre Cubaud

L'histoire du Cédric : penser un dispositif archivistique en histoire des sciences

Mise en œuvre d'une « fusée documentaire à trois étages »[1]

Gérald Kembellec
Laboratoire Dicen-Idf, Cnam.

Raphaël Fournier-S'niehotta
Laboratoire Cédric, Cnam.

Pierre Cubaud
Laboratoire Cédric, Cnam.

Résumé

Cet article présente la synthèse du travail de collaboration interdisciplinaire réalisé au Cnam entre des historiens des sciences et des techniques, des informaticiens et des chercheurs en sciences de l'information et de la communication. Autour de la genèse du laboratoire d'informatique du Cnam, une plateforme d'archivage numérique des documents historiques est modélisée et développée. L'enjeu principal est de rendre accessible les documents pour les chercheurs des disciplines concernées dont les pratiques diffèrent, aussi bien au cours du temps pour une même discipline, que de manière interdisciplinaire. Après un état de l'art des pratiques du domaine, l'article présente la modélisation retenue pour le projet, le développement de celui-ci, avec une attention particulière sur la valorisation des corpus.

Mots-clés : archives numériques, histoire, système d'information, classification multi-point-de-vue, recherche d'information.

1 Ce sous-titre — inspiré par une formule humoristique inventée par les auteurs de cette étude — symbolise les trois aspects de cette structure documentaire : le stockage physique, le modèle archivistique et les affichages.

Introduction

En 2018, le Centre d'Études et De Recherche en Informatique et Communications (laboratoire Cédric du Cnam, EA 4629) fêtera les trente ans de sa création officielle. Pour mieux comprendre la genèse de ce laboratoire, un groupe de recherche interdisciplinaire s'est formé dans le cadre d'un projet au sein du Laboratoire d'Excellence HASTEC (Histoire et Anthropologie des Savoirs, des Techniques et des Croyances). Des chercheurs et chercheuses en histoire des sciences et des techniques, en information-communication et en informatique en sont les moteurs. Le point de départ de cette investigation réside dans le constat selon lequel le Cédric ne s'est pas formé comme une entité homogène spontanée en 1988. Durant les deux décennies qui précèdent cette date, de nombreux événements ont eu lieu et leur compréhension est cruciale pour appréhender la création du Cédric. Un travail d'examen des documents produits par les acteurs de l'époque est en cours. Pour le mener à bien, une plateforme d'archives numériques a été élaborée au sein du projet. Nous décrivons dans cet article les motivations du projet et l'avancée de la mise en œuvre de ce dispositif archivistique.

Contexte et objectifs

Les membres du Cédric sont enseignants-chercheurs en Informatique, en Mathématiques ou en Électronique, et situés au Cnam à Paris. Cette assertion, dans notre contexte actuel semble aller de soi, aussi bien pour le statut des acteurs que pour les disciplines énoncées. Pourtant, si l'on revient cinquante ans en arrière, avec le cadre scientifique encore flou d'une discipline naissante, les choses n'étaient pas aussi simples.

Pour comprendre la complexité de cette genèse, ce projet est conduit en collaboration étroite avec des historiens, ce qui l'ancre résolument dans l'historiographie des sciences et des techniques. Pour mieux saisir les éléments factuels d'époque, il faut se recentrer sur les écrits, les acteurs et leurs témoignages. Intégrer ces derniers dans un système d'information permet de les archiver, mais aussi de les croiser et d'en faciliter l'analyse *a posteriori*. C'est la raison pour laquelle ce travail a engendré une réflexion en sciences de l'information et de la communication, plus particulièrement en documentation.

Notre objectif est de proposer un dispositif d'accès à de l'information historique reposant sur un ensemble d'archives numérisées, de notices bibliographiques ou catalographiques. Après avoir considéré la possibilité de mettre en œuvre un système d'information documentaire existant (spécialisé en gestion de l'information scientifique historique), nous avons envisagé la réalisation *ad hoc* d'une hybridation de systèmes d'archivage scientifique ayant la capacité d'intégrer un modèle liant les acteurs et les objets scientifiques. Un tel modèle, décidé en concertation avec nos collègues historiens, offre une flexibilité

d'analyse importante en évitant de cloisonner *a priori* les informations, ce qui devrait permettre de faire émerger des liens peu apparents dans d'autres modèles d'analyse. Notre modèle propose aussi la liaison des acteurs et des objets scientifiques avec les entités institutionnelles, c'est-à-dire les laboratoires de recherche et structures administratives ou pédagogiques au sein desquelles les futurs membres fondateurs du Cédric ont pu évoluer au cours des deux décennies précédant sa création. Tous ces acteurs et entités ont produit une masse importante de documents dont nous pouvons commencer à définir le périmètre. Une partie d'entre eux relève de la production scientifique (articles de recherche, thèses, mémoires), pour lesquelles la notion de co-signature est fondamentale en vue d'analyser les dynamiques de groupes entre acteurs. Les autres documents seront de nature très diverse, mais une partie importante d'entre eux témoigne de l'activité para-scientifique des acteurs : comptes rendus de conseils de laboratoires, annuaires de structure, correspondance, etc. L'intégration d'entretiens d'acteurs de la période est aussi envisagée (matériel audio voire audiovisuel).

Réflexions sur des travaux précédents

Pour appréhender la problématique des archives numériques, Stockinger *& al.* (2015) présentent une dualité de sens entre (1) la banque de données ouvertes, dans un objectif de ré-exploitation indéterminé, mais libre, basée sur une plateforme et (2) un ensemble patrimonial de ressources au sens d'un corpus dont l'objectif est de garantir l'accès aux documents et de les structurer dans un but précis et forcément contraint sans être dépendant de l'évolution des technologies. Comme dans le cadre que nous présentons ce corpus devra servir principalement de ressource à une activité scientifique, nous prendrons plutôt en compte le second sens donné aux archives numériques.

Ainsi, de ce point de vue, « *les archives existantes ou à venir servent d'objet de recherche théorique et appliquée en vue, par exemple, d'améliorer les accès aux données archivées, d'avancer sur le terrain de l'interconnexion des données archivées* […] *de rendre possible la préservation de l'information à travers différents formats, d'améliorer les conditions de stockage des données et des métadonnées, de mieux connaître les procédés et les "façons de faire" qui déterminent les pratiques courantes d'archivage* » (Stockinger & *al.,* 2015, p. 12). Pour continuer dans cette logique, il faut penser la « *réutilisation active* » d'un ensemble de données archivées dans le cadre de l'histoire des sciences et des techniques en général, de l'informatique en particulier, ce qui implique une transformation qualitative desdites données pour les rendre exploitables dans ce(s) contexte(s) par les différentes populations d'usagers.

Classer et indexer les documents d'archives

Dans un précédent numéro de la revue des *Cahiers d'histoire du Cnam,* Quantin *& al.* (2016) analysent la situation actuelle de mise en exploitation de l'histoire des sciences et des techniques. Ils situent ladite mise en exploitation des corpus grâce à (1) l'usage de dispositif de numérisation, (2) de compilation, (3) de visualisation et (4) de valorisation des archives. Leur analyse est en substance fondée sur deux observations distinctes :

- grâce aux (relativement) nouvelles orientations en termes de politique publique des sciences, la mise à disposition de sources à destination de la recherche est actuellement un phénomène acquis ce qui renvoie aux points (1) et (2) ;

- cependant, la possible valorisation de ces sources semble être une impasse pour la connaissance scientifique, qualifiée de « *branche morte* » (*sic*) et qui interroge les points (3) et (4).

- La difficulté de la problématique exposée n'est donc pas liée à la possibilité d'identifier, numériser, stocker et d'indexer les contenus et documents, mais bien de les valoriser d'un point de vue de l'usager final.

Même si Michel Cotte (2007), cité par Quantin *& al.* (2016, p. 110), place l'innovation du numérique – notamment en termes de patrimonialisation – au cœur de l'évolution des pratiques des historiens, il met l'accent sur l'importance de l'alignement de ces données avec les connaissances et les besoins (réels, et non supposés) des chercheurs et les témoignages récoltés. Ce dernier aspect recentre le dispositif autour des terrains étudiés, de l'épistémologie en histoire des sciences et des techniques et, bien sûr, des usagers. Ce point de vue oblige à une réflexion, faisant sortir l'artefact de consultation patrimoniale du simple paradigme documentaire traditionnel issu des sciences de l'information et de la communication pour glisser vers des besoins spécifiques en histoire des sciences et des techniques et se plonger dans trois mondes historiquement quasiment disjoints dans le cadre de la consultation des archives scientifiques et des bibliographies d'acteurs :

a. celui des historiens, observant au prisme micro-historique de la genèse d'un laboratoire la discipline scientifique informatique en train de se construire en France – puisque c'est le projet qui nous concerne. Ce public n'est pas forcément au fait des règles de classification en informatique technique ou scientifique, mais comprend très bien les enjeux politiques et structurels historiques dans lesquelles la création du laboratoire Cédric a eu lieu ;

b. celui des chercheurs en informatique, agents administratifs, ingénieurs et enseignants en activité sur la période étudiée, pour la plupart retraités à

ce jour. Ces acteurs maîtrisent les classifications des grandes instances scientifiques et techniques de l'informatique au moment de leur période d'activité [2]. De plus, les méthodes d'indexation documentaire appliquée en archivistique peuvent ne pas aller de soi pour ces usagers ;

c. les usagers lambda, au sein desquels nous incluons les chercheurs d'autres disciplines comme les Sciences de l'information et de la communication (SIC). Ces utilisateurs peuvent être partiellement au fait du cadre historique et peuvent également avoir des bases en utilisation de dispositifs numériques d'archivage.

La pluralité des temporalités, typologies d'acteurs et usagers étudiés pose un certain nombre de défis à relever : comment faire cohabiter dans une archive patrimoniale étendue sur plus de vingt ans et consultée par des usagers scientifiques de plusieurs disciplines un système de classification documentaire efficace tant à l'indexation qu'à l'archivage ?

Pour éclairer le questionnement, il a fallu se poser la question du classement dans un même système des versions numériques de boîtes d'archives administratives d'un laboratoire en construction contenant à la fois : (1) des articles et livres scientifiques, (2) des rapports de recherche, mais aussi (3) des documents pédagogiques (procès-verbaux de soutenances d'élèves ingénieurs), ou encore (4) des documents administratifs tels des correspondances ou des procès-verbaux des instances administratives. Qui plus est, pour complexifier encore le propos, en dehors de la simple question des publics de cultures scientifique et administrative hétérogènes, les temporalités interviennent comme facteur de complexité au sein même des catégories homogènes. Un chercheur en informatique qui travaille par exemple sur le stockage de la donnée en informatique n'utilisera pas le même vocabulaire entre le début et la fin de la période observée tout simplement parce que les concepts et/ou leurs implémentations ont changé ou ont disparu.

La gageure de rendre notre plateforme adaptée à différents publics de chercheurs (historiens, sociologues, etc.) reste ainsi secondaire devant la question de la mise à disposition efficace de l'information. À titre préliminaire, il est donc envisagé que les utilisateurs finaux de la plateforme seront avant tout les collègues chercheurs du laboratoire HT2S (voir catégorie (a) ci-dessus). L'adaptation aux catégories (b) et (c) d'utilisateurs étant prévue pour un travail ultérieur. Initialement, la plateforme et son interface sont donc développées et testées en tenant compte des besoins et retours d'utilisation des historiens. D'autre part, remarquons que cet article constitue un point d'étape dans le travail d'élaboration de cette plateforme, permettant de présenter à la communauté un retour d'expé-

2 Cependant ces indexations évoluent dans le temps en même temps que la science et les techniques comme nous le verrons en détail plus tard.

© Pierre Cubaud

Figure 1 : Les archives papier « administratives », « pédagogiques » et « techniques » du département d'informatique sur la période concernée dans leur état actuel de classification et de stockage

rience sur la méthodologie de conception d'une telle plateforme de travail. Nous proposons donc un banc d'essai de certaines solutions logicielles d'indexation et de stockage de documents, puisque cela a fait partie intégrante du développement de la plateforme. En revanche, nous ne présenterons pas les détails complets de la synthèse réalisée, peu pertinents pour la cohérence de cet article.

Valoriser les documents d'archives par la visualisation

Dans un cadre similaire, Quantin & *al.* (2016, p. 95), relèvent deux défis simultanés pour réaliser un outil d'archivage patrimonial : le modèle des données stockées et leur visualisation. « *Du côté de la production, le numérique ne doit pas être un carcan pour l'histoire. Les notions d'incertitudes, temporelles et spatiales ; ainsi que le contexte immatériel de l'objet ("conditions de travail" par exemple)*

devront être pris en compte et accessibles à la valorisation. » En effet, les règles traditionnelles d'archivistiques et de systèmes d'information documentaires (plus ou moins rigides et péremptoires) doivent s'harmoniser pour que la classification, le nommage et l'indexation des documents soient transparents pour l'utilisateur final du système tout en demeurant suffisamment cohérents pour garantir l'accès à l'information. Les questions du point de vue et de l'appropriation sont ici cruciales pour garantir une utilisabilité maximale. Quantin & *al.* (2016, p. 95) poursuivent : « *Du côté de la valorisation, l'accès devra être adapté à l'utilisateur et à la nature des connaissances historiques. Les mêmes contenus ne peuvent pas être présentés au visiteur de passage et à l'expert du domaine*. » Dans un strict cadre documentaire, de ce point de vue, il faut donc penser l'outil d'accès au corpus comme accessible aux néophytes comme aux historiens des sciences et des techniques ou encore aux spécialistes de la documentation et aux simples visiteurs curieux du sujet. Cette description des besoins heurte la traditionnelle vision classificatoire, dite du « *Web sémantique* » pour glisser vers le « *Web cognitivement-sémantique* » (Caussanel & *al.*, 2002) qui hybride la classification et les besoins communicationnels sociaux, naturellement plus individualisés tels que décrits par (Zhang & Marchionini 2004 ; Zhang 2007) avec l'*exploratory search*[3]. Pour aller plus loin dans le paradigme d'accès informationnel socio-sémantique, Zacklad & *al.,* dans une logique d'appropriation du dispositif par l'usager, proposent un système d'information (SI) documentaire disposant d'une interface aux accès pluriels, *a minima* avec des facettes, voire personnalisée avec les propres tags des usagers (Zacklad & *al.*, 2011). Dans notre vision du système, à défaut d'une indexation réellement sociale, complexe à mettre en œuvre et peu utilisable par l'usager de passage, nous avons envisagé un système d'accès par points de vue comme exposé *infra*, ce qui suppose de modéliser des méthodes de classement et d'indexation adaptés à des méthodes de visualisations distinctes.

Valoriser les documents d'archives par l'ouverture et l'exposition des métadonnées

Les récents travaux sur les aspects documentaires du Web portent sur la manière de systématiser la valorisation les données et métadonnées contenues dans les archives documentaires (Stockinger & *al.,* 2015). Ces données et métadonnées se doivent d'être accessibles non seulement pour les humains avec les aspects d'ergonomie, mais aussi pour les systèmes logiciels avec le Web de données. Ce principe fut présenté par Tim Berners-Lee et

3 Nous avions repris ce concept largement décrit par White & Roth (2009) dans leur ouvrage de synthèse, dans le prototype OntologyNavigator, avec l'heuristique de recherche QBQ-S qui combine le traditionnel moteur de recherche et l'accès par graphe avec possibilité d'enrichissement sémantique (Kembellec, 2013).

implémenté au moyen de diverses technologies[4] largement présentées en France par l'INRIA et normalisées par le W3C. Ainsi, en répondant à ces besoins dans le cadre d'infrastructures documentaires en SHS, Stéphane Pouyllau a proposé un principe d'ouverture des données liées (publications, archives, référentiels), ce qui a permis la création d'outils de recherche nouveaux. Ces derniers permettent à la fois de lier des portails documentaires et des applications pouvant être embarquées dans des sites Web. Un exemple de réutilisation possible est la recommandation de contenus thématiquement liés sur les blogs scientifiques vers les documents indexés par la plateforme Isidore (Pouyllau, 2016). Cet aspect semble particulièrement important pour permettre un usage optimisé et sans contrainte de la plateforme aux usagers et à leurs outils de collecte.

Penser la modélisation du dispositif

Comme évoqué *supra*, la production documentaire sur l'époque étudiée a bien sûr pour auteurs des individus, mais ceux-ci s'inscrivent temporellement dans des entités que l'on peut qualifier de majeures des points de vue respectifs du laboratoire Cédric et des historiens des sciences et des techniques. Ces entités majeures du système sont déterminées par les acteurs et sont, par exemple, des groupes informels de chercheurs, des équipes, des structures institutionnelles, des revues, etc. Il s'agit de mettre au jour et d'analyser les dynamiques collectives du travail de ces chercheurs : qui publie avec qui ? Quels sont les groupes formels et informels qui se créent ? Qui sont les acteurs centraux de ces groupes de travail ? Compte tenu de la période considérée, ces chercheurs ont contribué à faire émerger une discipline scientifique et leur travail témoigne des aléas de ce processus de légitimation aux côtés de disciplines existantes. L'analyse des archives résultant de ces périodes d'activité est donc prépondérante pour le travail historiographique. Un travail préliminaire des chercheurs du HT2S a montré que la complexité des dynamiques collectives était importante, les chercheurs en informatique ayant travaillé à la fois en groupes informels pour cosigner des articles, mais aussi des groupes hérités de structures institutionnelles (départements, laboratoires, équipes en émergence). Un collectif d'auteurs prend parfois même un nom *ad hoc*, par exemple Cornafion (1981), pour la publication de l'ouvrage *Systèmes répartis*, non sans rappeler le groupe Bourbaki.

Une collaboration des équipes historienne, infodocumentaire, et informaticienne a permis d'identifier les relations que les entités collectives entretiennent entre elles. Le résultat de cette collaboration a été synthétisé par les chercheurs du HT2S sous la forme de cartes heuristiques, comme celle illustrées dans les figures 2 et 3 pour les groupes et les publications.

4 Soit directement dans les pages soit au moyen d'API ou encore de SParQL endPoint [URL : https://www.w3.org/wiki/SparqlEndpoints]. Voir la synthèse de B. Menon (Menon, 2016)

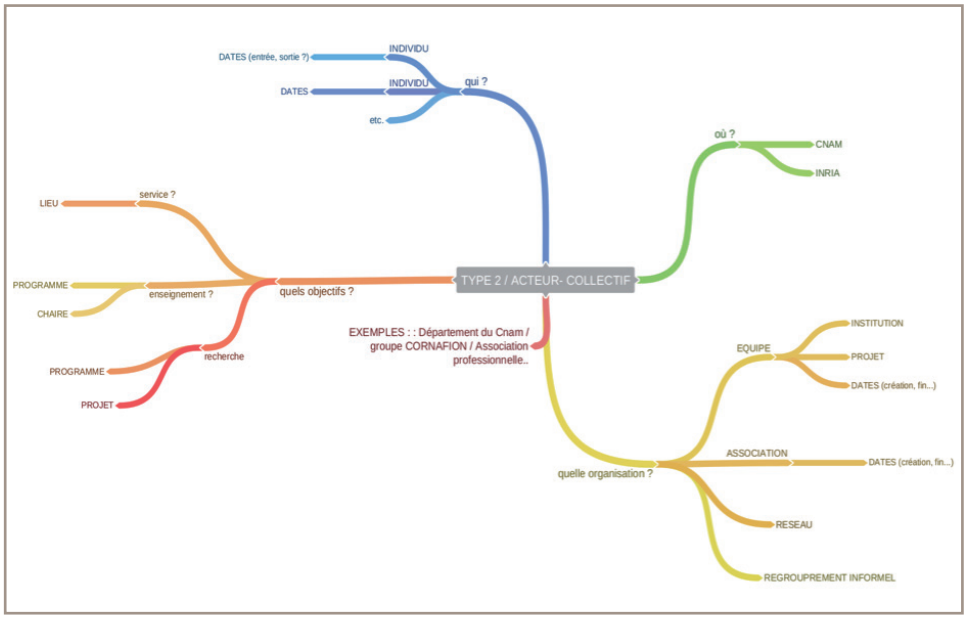
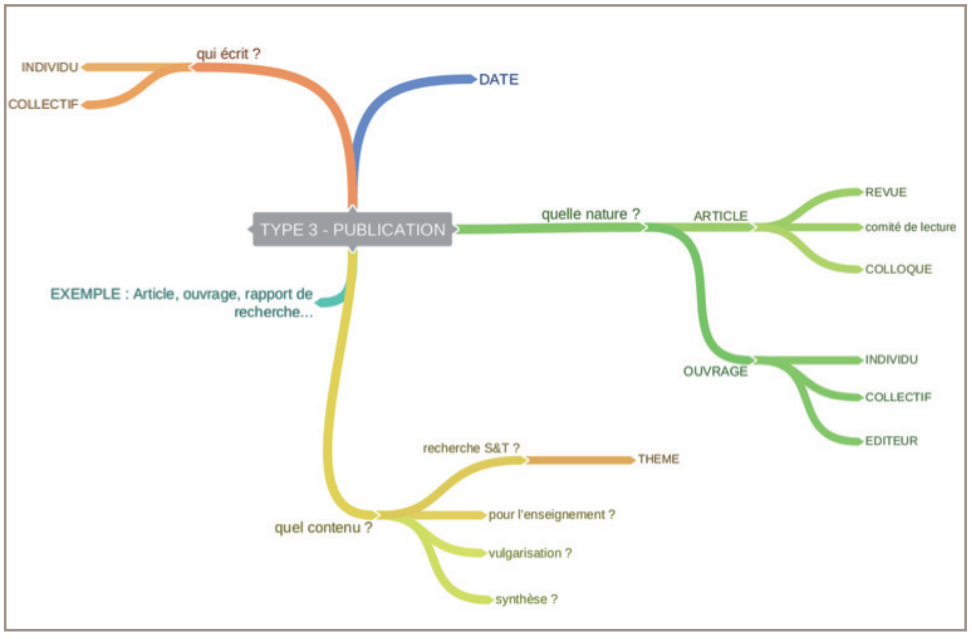


Figure 2 : Les acteurs et groupes d'acteurs au sein du modèle historique décrit

Figure 3 : Modèle « heuristique » documentaire prévisionnel du système d'information pour les documents



Ces cartes ont servi à la modélisation du système info-documentaire qui peut servir à la rédaction du cahier des charges fonctionnel tant pour la création d'une potentielle base de données relationnelles qu'à la validation du choix d'un logiciel ou d'une plateforme existante comme système d'information. En vue de faciliter le travail de collecte des documents et de création de savoirs, nous avons, en lien avec l'équipe des historiens, modélisé un système d'information documentaire et en avons initié le développement. Celui-ci repose sur une architecture comportant trois « couches » (au sens informatique du terme) :

1. Le stockage des documents et notices sous forme de fichiers (couche basse) ;
2. L'organisation des documents (couche intermédiaire) ;
3. La visualisation et la manipulation des documents (couche haute).

État de l'art technique

Lors de l'étude de l'état de l'art, trois systèmes ont été envisagés et ont inspiré notre approche en accord avec les historiens.

Le *Dictionnaire Prosopographique des Inventeurs en France*[5] nous a intéressés en particulier ses différents onglets permettant des « vues » différentes de l'information associée à une notice biographique. L'approche strictement biographique, cependant, est écartée pour notre projet, car c'est le collectif davantage que l'individu qui doit être mis en valeur. En outre, il s'agit là d'un exemple de réalisation d'implémentation *ad hoc*, non d'une solution logicielle réutilisable directement en l'état.

SyMoGih[6] est un système modulaire de gestion de l'information historique (Beretta & Vernus 2012). D'un point de vue méthodologique, son approche par objets (acteurs – dont collectifs ; lieux, objets abstraits ; caractères sociaux, formes concrètes) est la plus pertinente pour notre projet, car le modèle de la base de données associé correspond aux besoins exprimés par l'équipe d'historiens. Cependant, des contraintes techniques d'hébergement du système et des données nous ont amenés à finalement l'écarter.

Omeka[7] est un système documentaire multimédia orienté Web spécialisé dans la publication de collections savantes, muséales et bibliothécaires. Il est développé par des spécialistes de l'histoire au *Center for History and New Media* de l'université états-unienne George Mason. Sa large diffusion dans le monde académique en fait un outil de travail quasi standard. Sa plasticité permet d'y implémenter un système de

5 [URL : http://dpif.cnam.fr/].

6 [URL : http://symogih.org/].

7 [URL : https://omeka.org/].

visualisation et de manipulation personnalisé selon nos méthodologies, de plus il intègre nativement les principes d'exposition des métadonnées. Plusieurs exemples d'implémentation sont disponibles pour présenter les usages possibles de cette plateforme. Un écosystème d'extensions est activement développé par la communauté open source.

Adaptation

Sur la couche la plus basse, celle la plus proche du stockage, se déroule le dépôt des archives numérisées. Afin d'organiser la matière documentaire, un plan de classement a été élaboré sur la base des standards en vigueur, les règles de nommage, au niveau le plus bas du système (stockage des fichiers). Les fichiers sont ainsi nommés avant même d'être placés dans un espace commun, et donc procèdent d'une organisation *a priori*. Cela permet d'éviter la dépendance à la couche intermédiaire et favorise donc la séparation complète des couches, ce qui facilite une migration éventuelle vers d'autres outils.

La couche intermédiaire est un outil de gestion documentaire qui définit le statut documentaire de ces fichiers et organise la liaison avec les notices associées (sources primaires, secondaires, types et genres de documents…). On peut définir des frontières entre des fonds ou autres ensembles documentaires. C'est ici que s'organise la datation, l'authentification et la mise en relation des documents. Pour ce niveau archivistique, nous avons décidé d'utiliser le système Omeka que nous avons adapté et personnalisé.

Sur la couche la plus haute, celle de la visualisation, s'interprètent les données, métadonnées et documents. C'est l'interface utilisateur ou IHM. On y modélise la manière dont les documents sont catégorisés, mis en relations, et visualisés ; cette modélisation est pensée en lien direct avec la méthodologie d'analyse, en l'occurrence une prosopographie qui est aussi une analyse relationnelle (acteurs-réseau). Il est prévu notamment un système de catégorisation souple, par *tags* (étiquettes) qui permettent d'associer les documents à des mots-clefs de manière personnalisable, dite « sociale », en complément de vocabulaires plus contrôlés (comme des thésaurus techno-scientifiques liés à la discipline). Omeka propose déjà un outil de recherche et de visualisation, mais il est également possible de penser une surcouche sous forme de *plug-in* ou d'interface autonome interrogeant la base de données d'Omeka et formalisant les flux extraits pour en faire des « vues » répondant à des besoins spécifiques en histoire des sciences et des techniques.

Réalisation du projet infodocumentaire

La constitution et l'intégration du corpus

Dans le cas du matériau traité, bibliographies académiques et littérature grise pour commencer, les règles classificatoires des documents et les vocabulaires qui y sont associés évoluent aussi au fil du temps, au rythme des avancées technologiques et de leur intégration scientifique comme sujet d'étude. La question du consensus pour le choix d'une classification est pertinente, surtout dans le cadre de l'étude d'une science « en train de se faire », comme c'était le cas au milieu des années 1960, moment où les premières classifications sont proposées.

Le choix d'une structure documentaire pour les notices

Plusieurs grandes associations professionnelles et sociétés savantes ont ainsi proposé des modèles à visée documentaire pour classer, diffuser aisément la production technoscientifique en informatique, que nous avons déjà examinée dans de précédents travaux (Kembellec, 2012). Parmi les plus anciennes et les plus reconnues, nous pouvons mettre en avant celles de l'Association for Computing Machinery (ACM) et de l'Institute of Electrical and Electronics Engineers (IEEE). Ces deux associations possèdent leurs propres bibliothèques en ligne qui autorisent toutes deux des recherches avancées classiques, mais aussi un accès aux corpus par des facettes telles que les entités de recherche (laboratoires, écoles, universités…), les auteurs, les spécificités techniques, la notoriété des écrits et des acteurs[8], autorisant ainsi l'analyse quantitative des réseaux d'auteurs à l'aide de la statistique de leur production.

- *Quelle évidence historique du choix d'une classification ?*

Dans le cadre de la modélisation d'un outil de recherche sur un sujet aussi spécifique que celui envisagé, le choix d'une classification réputée et historiquement documentée était donc un enjeu majeur. Le choix s'est ainsi porté sur la taxonomie ACM[9], dont les différentes versions font consensus, admettent une traçabilité de version et une rétrocompatibilité de classement des documents. De plus, cette classification thésaurise deux vocabulaires contrôlés spécifiques à l'informatique : un pour les entités nommées et un second pour les termes descripteurs.

- *Les problèmes d'indexation liés à l'évolution technologique et scientifique*

Entre 1964 et 2012, l'ACM CCS a connu quatre étapes de classification correspondant à la fois aux évolutions

8 IEEE XPlore [URL : http://ieeexplore.ieee.org/Xplore/home.jsp] avec plus de 4 millions de documents indexés et ACM digital library [URL : http://dl.acm.org] et ses 2,8 millions de documents.

9 ACM Computer Classification System : [URL : http://www.acm.org/about/class].

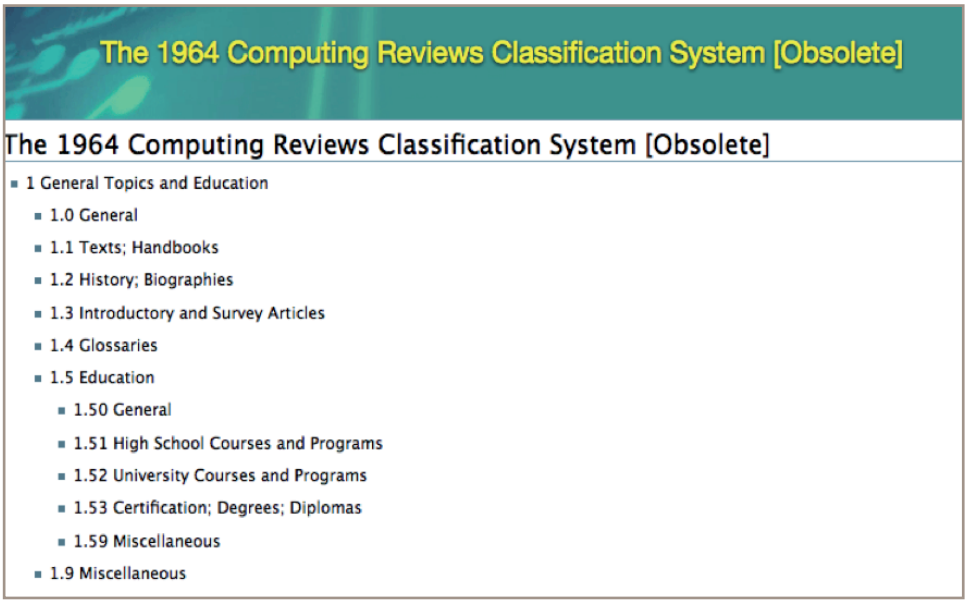


Figure 4 : Extrait de la classification ACM CCS de 1964

Capture d'écran de la page [URL : http://www.acm.org/about/class/cr64]

technologiques liées au sujet technoscientifique et à celles liées aux méthodologies documentaires, mais aussi au besoin de spécification induit par la masse de littérature à traiter.

Le système original décimal à trois niveaux strictement hiérarchiques de 1964 a été utilisé jusqu'en 1991 par les chercheurs auteurs d'articles et les organisateurs de conférences ACM pour l'indexation des articles, ouvrages et actes de conférences (voir figure 4). Sur la période historique qui nous intéresse dans le cadre de ce projet, il est très probable qu'une partie de la production scientifique et de la littérature grise du corpus concerné ait été déjà classée à l'aide de ce système.

L'année 1991 a vu une refonte totale de la classification documentaire de l'ACM avec la mise en œuvre d'un système alpha décimal à quatre niveaux de profondeur prenant bien sûr en compte des mises à jour technologiques et témoignant de l'arrivée du réseau planétaire ainsi que de l'informatique grand public (figure 5).

Des termes descripteurs (*Subject Descriptors*) et un vocabulaire d'entités nommées (*Implicit Subject Descriptors*) ont été également ajoutés pour spécifier les classes trop génériques, implémentant un thésaurus à la classification. Il est à noter qu'avec cette deuxième mouture sont apparues des passerelles entre classes, exprimant la possibilité qu'un concept puisse être lié à plusieurs contextes techno-

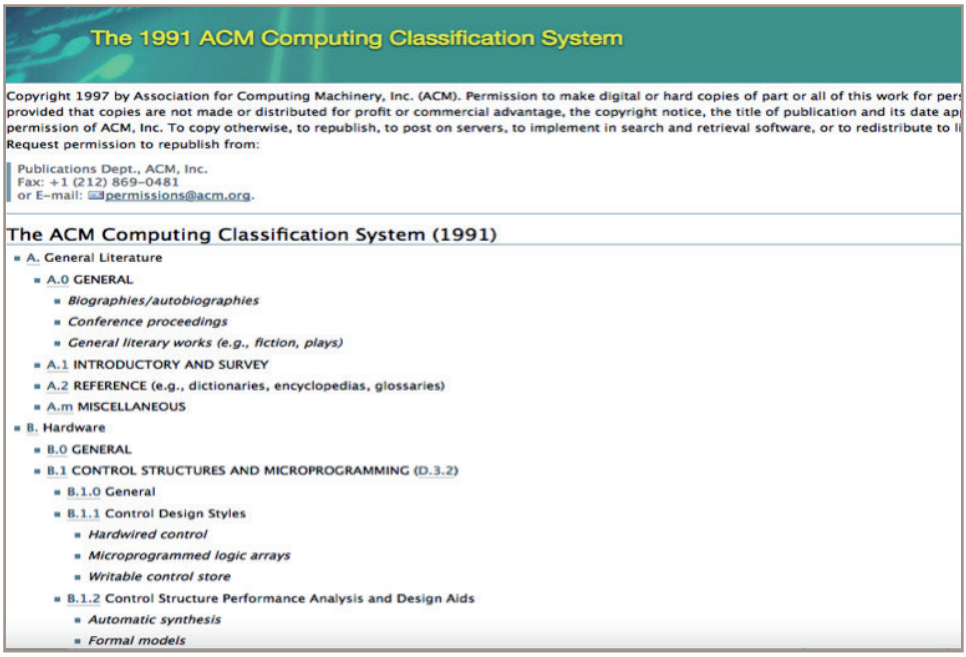


Figure 5 : Extrait de la classification ACM CCS de 1964

Capture d'écran de la page [URL : http://www.acm.org/about/class/ccs91-html]

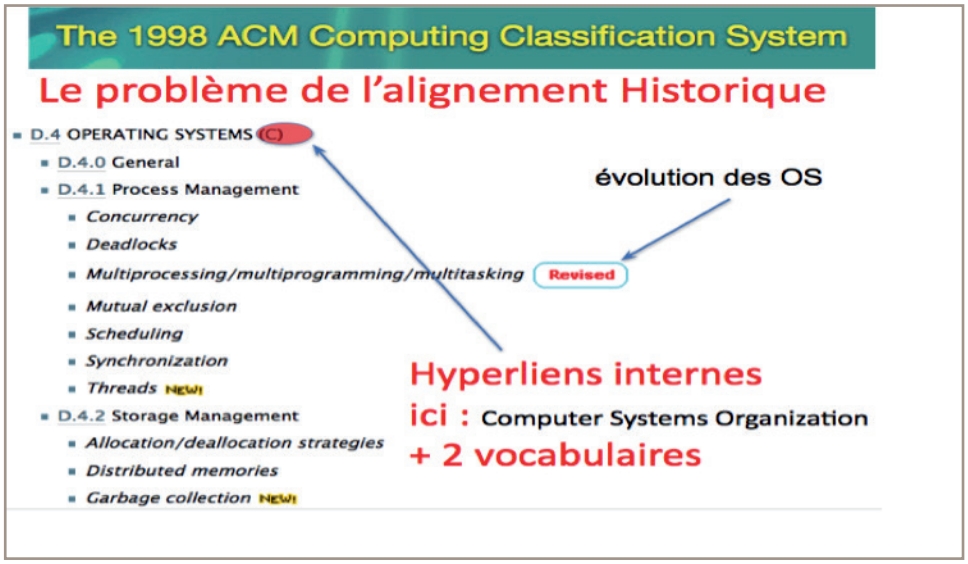


Figure 6 : Extrait de la classification ACM en 1998

Capture d'écran de la page [URL : http://www.acm.org/about/class/ccs98-html]

scientifiques. Cette version a été utilisée de manière internationale dès les débuts du laboratoire Cédric par des chercheurs toujours publiants.

En 1998, une version mise à jour du précédent système alpha décimal est proposée avec des concepts nouveaux et d'autres, révisés, cette version est accompagnée de deux vocabulaires (figure 6).

Les *Implicit Subject Descriptors* (Noms propres ou entités nommées) sont des noms de produits, systèmes, langages et personnes éminentes dans le domaine de l'informatique. L'exemple donné sur le site d'ACM est « C++ » qui est classé sous la catégorie « *D.3.2 Language Classifications* ».

Cette liste est dynamique avec de nombreuses mises à jour. Il était même encouragé de proposer de nouvelles entrées (en motivant sa demande). Cette version de la classification est donc assez proche d'un thésaurus et d'une taxonomie.

En 2012, l'ACM a choisi une nouvelle méthode de classement des documents liés à l'informatique avec un système non codé (ni alphabétique, ni numérique) au sein d'une ontologie poly-hiérarchique : l'accès aux documents se fait intuitivement par spécification selon le modèle directement inspiré du paradigme BQ (*Browsing-Query*) présenté par Zhang (2007). La première étape de la recherche par navigation consiste à descendre dans l'arborescence jusqu'au nœud le plus représentatif du concept recherché. Notons que les documents peuvent être trouvés de manières différentes grâce à la poly-hiérarchie.

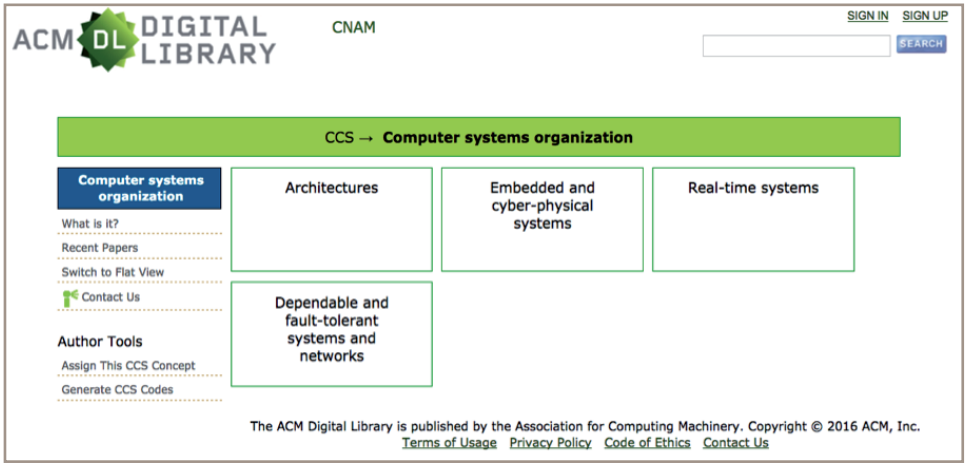


Figure 7 : Extrait de la classification ACM CCS de 2012

Capture d'écran de la page [URL : http://www.acm.org/publications/class-2012]

- *La solution hybride de classement et d'indexation des notices*

Le système de classement techno-scientifique basé sur la classification ACM est pertinent pour indexer la littérature scientifique et partiellement la littérature grise. Bien sûr, pour l'utiliser, il faudrait aligner les différentes versions historiques afin que celles-ci soient compatibles entre elles et que différentes générations d'usagers puissent l'utiliser avec la version qu'ils maîtrisent. Cependant, ce modèle d'indexation n'est pas adapté à tous les publics. En effet, en histoire des sciences et des techniques, cette classification ne sera pas forcément maîtrisée, ni même porteuse de sens. Il est donc possible d'utiliser une version ACM pour indexer dans une approche techno-scientifique et surtout des règles typologiques proposées dans les figures 2 et 3 afin d'offrir une possibilité de recherche plus proche des problématiques d'histoire des sciences. De plus, croiser l'indexation des documents selon deux méthodes avec les groupes et individus peut produire en soi un matériau de recherche prosopographique très intéressant.

Le choix d'une structure documentaire pour les archives

Pour indexer les archives de documents d'activité tels que les comptes rendus de réunions ou les rapports, qu'ils soient administratifs ou pédagogiques, même si le sujet reste l'informatique, l'ACM CCS est insuffisante, voire hors sujet. C'est pourquoi il faut réfléchir à une seconde méthode de classification dédiée aux archives. Cette seconde structure de termes et d'entités nommées devra intégrer les informations issues également du modèle *ad hoc* ci-dessous pour mieux repérer le lien des documents pédagogiques, administratifs ou d'activités avec les acteurs individuels et collectifs.

La collecte des données et documents

Les tâches d'implémentation et de collecte ont été menées sous la direction conjointe des historiens des sciences, des spécialistes de la documentation et des chercheurs en informatique. Le processus alliant sélection et nettoyage des données (curation) a ensuite été effectué en deux passes :

- l'une formaliste et stylistique pour la cohésion des notices a été effectuée par l'équipe documentaire ;

- la seconde, visant à la cohésion des notices avec l'histoire des sciences, portait sur la qualité et la complétude des données : dates, auteurs, sujet et était réservée aux historiens.

La recherche et la formalisation de la bibliographie ont été réalisées en amont de l'implémentation du dispositif documentaire. L'objectif de ce travail était triple :

- aider à penser le système archivistique pour la collecte et le traitement

des archives du Cnam en lien avec l'historique du Laboratoire Cédric ;

- rechercher et enregistrer les notices de la production scientifique sur la période traitée ;

- rechercher et enregistrer de manière normalisée les notices des mémoires d'ingénieurs rédigés sur la période.

La première étape d'un projet de collecte et d'archivage documentaire, quel qu'en soit le cadre, est de fixer une typologie des documents à recenser pour archivage conservation ou indexation s'ils sont déjà présents dans un fonds numérique accessible.

• *Collecter les notices*

Ensuite, la collecte des notices d'articles, livres, actes de conférences produites par les chercheurs du Cnam a été menée sur les bibliothèques numériques en accès par abonnement et accessibles grâce au service de documentation du Cnam. L'équation de recherche, adaptable à chaque plateforme portait sur les noms, institutions ou groupe et la période. Les principales bases consultées, par ordre de priorité sont dl.acm, IEEE Xplore, Taylor & Francis et Springer, mais aussi les archives numériques ouvertes comme l'américain ArXiV et le français HAL. Une troisième passe de recherche a été effectuée sur les moteurs de recherche scientifique comme Google scholar, avec vérification des notices. Les pages personnelles des acteurs et le site du laboratoire ont été autant de sources de complémentaires d'information et de points de comparaison avec les résultats trouvés. Un total de 246 notices a été retrouvé.

La collecte, la gestion et le partage de ces notices ont été effectués au sein du groupe de travail, grâce au Logiciel de Gestion de Ressources Bibliographiques (LGRB) Zotero[10], développé comme Omeka par des spécialistes de l'histoire au Center for History and New Media de l'université George Mason. Cet outil permet de détecter les notices présentes dans les pages Web et de les enregistrer sur un ordinateur individuel et éventuellement dans un espace partagé. Cette dernière solution a été retenue pour permettre le partage en lecture-écriture à tous les membres du projet. Ce mode de travail collaboratif a permis de travailler sur le dédoublonnage des notices alors même que la collecte n'était pas terminée et de soulever des questions intéressantes. Par exemple, une mystérieuse thèse est apparue qui semblait avoir été soutenue deux fois avec le même titre et des auteurs distincts (voir sur la figure 8, les doublons détectés par le logiciel, mais aussi le problème de cohérence cité précédemment avec les annexes communes).

La réponse à cette énigme, après une discussion avec les chercheurs historiques du Cédric, a été que les deux auteurs – Boulenger et Kronental – ont soutenu leurs thèses respectivement,

10 [URL : https://www.zotero.org/]

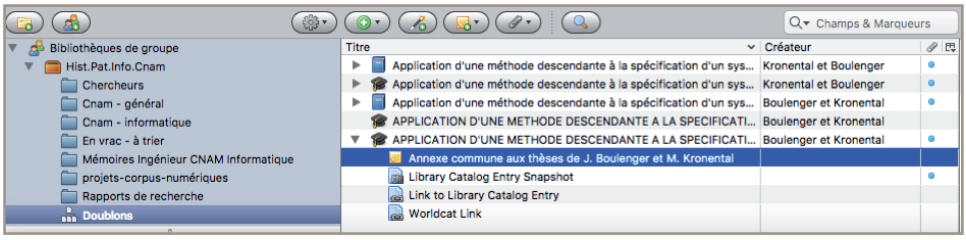


Figure 8 : Capture d'écran de Zotero : fonction de gestion des doublons

mais de manière complémentaire sur des aspects distincts d'une même problématique, des pratiques que l'on retrouve par ailleurs[11].

Ces questions ont mis en valeur la qualité et la fiabilité des données documentaires et des métadonnées bibliographiques sur Internet. Selon les sources la qualité des notices exposées à la capture par un logiciel de gestion de références bibliographiques peut varier comme le montre le tableau 1, ci-dessous :

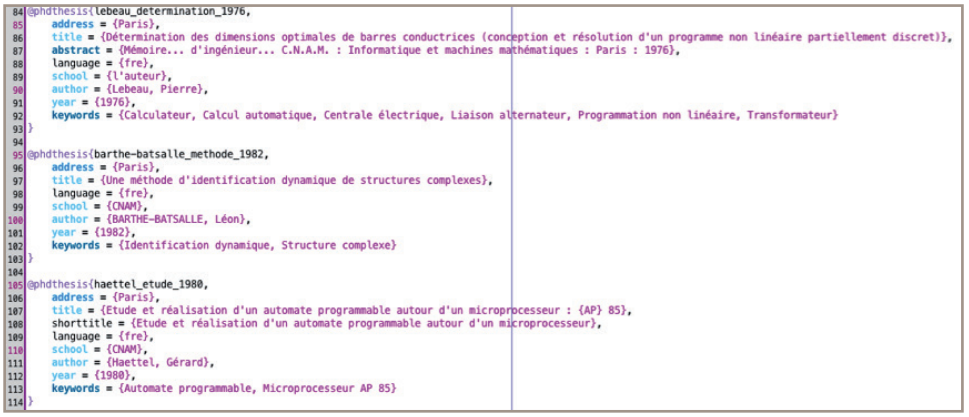
Implémentation du système documentaire

Une recension des plateformes existantes dans le domaine a été effectuée, en vue d'en déterminer les caractéristiques principales et de choisir celle qui pourrait, le cas échéant, servir de base à la plateforme réalisée pour le projet. Sa synthèse a conduit au choix d'Omeka, comme indiqué *supra*. C'est une plateforme open source répandue dans la communauté, et d'administration assez aisée. Une version

Excellente qualité de notices	IEEE, ACM, Springer, HAL, Techniques de l'Ingénieur...
Notices de qualités variables	Google Scholar, base Orion, certains CMS universitaires
Notices non compatibles	Sites d'enseignant-chercheur, certains CMS universitaires.

Tableau 1 : Qualité constatée du formalisme des notices exposées dans les pages Web, et leur compatibilité avec les logiciels de gestion de références bibliographiques (LGRB)

11 Voir par exemple les travaux de S. Natkin et G. Florin sur les réseaux de Petri, développés dans des thèses respectivement de 3e cycle et d'État soutenues en 1985 (*cf.* article de Paloque-Berges et Petitgirard dans ce même volume).



```
@phdthesis{lebeau_determination_1976,
    address = {Paris},
    title = {Détermination des dimensions optimales de barres conductrices (conception et résolution d'un programme non linéaire partiellement discret)},
    abstract = {Mémoire... d'ingénieur... C.N.A.M. : Informatique et machines mathématiques : Paris : 1976},
    language = {fre},
    school = {l'auteur},
    author = {Lebeau, Pierre},
    year = {1976},
    keywords = {Calculateur, Calcul automatique, Centrale électrique, Liaison alternateur, Programmation non linéaire, Transformateur}
}

@phdthesis{barthe-batsalle_methode_1982,
    address = {Paris},
    title = {Une méthode d'identification dynamique de structures complexes},
    language = {fre},
    school = {CNAM},
    author = {BARTHE-BATSALLE, Léon},
    year = {1982},
    keywords = {Identification dynamique, Structure complexe}
}

@phdthesis{haettel_etude_1980,
    address = {Paris},
    title = {Etude et réalisation d'un automate programmable autour d'un microprocesseur : {AP} 85},
    shorttitle = {Etude et réalisation d'un automate programmable autour d'un microprocesseur},
    language = {fre},
    school = {CNAM},
    author = {Haettel, Gérard},
    year = {1980},
    keywords = {Automate programmable, Microprocesseur AP 85}
}
```


Figure 9 : Portion de l'extraction erronée au format BibTeX de la liste des mémoires d'ingénieurs

de la plateforme est accessible en interne du Cnam[12]. Des comptes utilisateurs ont été créés, pour permettre les premiers tests. Une réflexion sur les droits d'accès (« lecture seule » ou « lecture et écriture » des contenus) de chacun est en cours, afin de faciliter le travail et éviter les maladresses. Une tâche de développement courte a été nécessaire pour normaliser les notices bibliographiques des mémoires d'ingénieurs collectées par les chercheurs en histoire. En effet, le passage par Zotero ne permettait pas de gérer aussi finement que nécessaire les notices dans la version par défaut. Le principal problème étant que Zotero ne permet pas nativement de gérer les notices des mémoires d'ingénieurs, le fichier d'export de la bibliographie au format BibTeX contenait donc des erreurs formelles récurrentes : les 596 mémoires d'ingénieurs étaient formalisés comme des thèses de doctorat (voir figure 9), ce qui en empêchait la distinction avec les réels doctorats soutenus sur cette période. De plus, quelques erreurs de fond et de forme avaient échappé à la vigilance des curateurs, ici par exemple l'établissement de rattachement contenait une information erronée dans la première notice (*idem*).

L'intégration des données et documents

La poursuite de ce travail peut emprunter une voie assez évidente. Il reste à continuer la mise en adéquation d'Omeka avec le modèle de données fin élaboré en collaboration avec les chercheurs en histoire. Ensuite ou en parallèle, l'automatisation de la jonction entre la couche basse (numérisation des documents) et intermédiaire (Omeka) devra être réalisée et testée. À plus long terme, un travail d'amélioration de l'interface pourra

12 Accessible seulement depuis une machine du Cnam, à l'adresse [URL : http://163.173.230.26/omeka-2.4/]. Une ouverture plus large est prévue ultérieurement, lors de la mise en production.

également avoir lieu : nouveaux regroupements de documents, enrichissement d'annotations, etc.

Le travail de collecte, numérisation et d'organisation des données a porté pour l'année 2016 sur :

- la liste des publications des chercheurs pour la période 1975-1980 et 1988-1990 (début et fin de la période d'incubation du laboratoire de recherche) – références prélevées dans les archives ;

- la liste des mémoires d'ingénieurs soutenus dans cette période – notices prélevées dans les catalogues du Cnam ;

- quelques archives papier (pour un corpus de test).

L'équipe historienne doit encore exploiter les données et documents traités dans le système. Ce sera l'objet de la suite du projet dont les buts additionnels à court terme sont :

- l'intégration de documents hétérogènes (dont des notices d'objets informatiques, en l'occurrence des ordinateurs, matériels et logiciels) et leur mise en relation ;

- l'enrichissement des données documentaires (lien avec des d'autres données sur le Web) ;

- la création d'une interface de visualisation par chronologie interactive.

Conclusion

La qualité du travail qui pourra être réalisé avec cette plateforme repose bien entendu sur la qualité des sources initiales et du respect des règles d'indexation lors de l'archivage. Un soin particulier devra donc être apporté à la sélection des sources, qui devront être validées par des curateurs — les chercheurs du laboratoire HT2S. Les systèmes automatisés comme Zotero ont montré, sur les notices de mémoires d'ingénieurs par exemple, qu'il pouvait y avoir des écarts sensibles avec ce qui est souhaité.

Le succès de ce dispositif réside à n'en pas douter sur la réception que les usagers feront de l'interface et de ses données. La capacité de l'outil à proposer des données de qualité et d'en exposer les métadonnées dans des formats et standards compatibles avec les agents logiciels les plus usités par les chercheurs et usagers participeront également de l'utilisabilité de l'archive. L'interface de visualisation et de création du savoir (relations entre entités, chercheurs, publications, lieux, etc.) sera fondamentale ; elle devrait être déployée fin 2017.

Bibliographie

Beretta F. & Vernus P. (2012). « Le projet SyMoGIH et la modélisation de l'information : une opération scientifique au service de l'histoire ». *Les Carnets du LARHRA,* pp. 81-107.

Caussanel J., Cahier J.-P., Zacklad M. & Charlet J. (2002). « Les Topic Maps sont-ils un bon candidat pour l'ingénierie du Web Sémantique ? ». *Actes des 13e journées francophones d'ingénierie des connaissances* (IC).

Cornafion (1981). *Systèmes informatiques répartis : concepts et techniques.* Paris : Dunod Informatique.

Cotte M. (2007). « La génétique technique a-t-elle un avenir comme méthode de l'histoire des techniques ? » *In* Rey A.-L. (dir.), *Méthode et Histoire, journées d'études de la SFHST.* Lille : publications de la SFHST, pp. 187-201.

Kembellec G. (2012). « Bibliographies scientifiques : de la recherche d'informations à la production de documents normés ». Thèse en sciences de l'information et de la communication de l'Université Paris 8, Saint-Denis.

Kembellec G. (2013). « Recherche exploratoire : proposition d'une méthode basée sur une ontologie de domaine. », *Actes du 9e Colloque ISKO-France : Contextes, langues et cultures dans l'organisation des connaissances,* pp. 281-302

Menon B. (2016). « Comprendre les standards du web de données ». *I2D – Information, données & documents,* 2/53, pp. 32-34.

Pouyllau S. (2016). « Isidore Suggestion, des recommandations de lecture pour les blogs de science ». *I2D – Information, données & documents,* 2/53, p. 44.

Quantin M., Laroche F., & Kerouanton J.-L.. (2016). « Récit historique et objet technique : outil de valorisation mutuelle ». *Cahiers d'histoire du Cnam*, 5, pp. 93-120.

Stockinger P., Lalande S., & Beloued A. (2015). « Le tournant sémiotique dans les archives audiovisuelles. Vision globale et éléments conceptuels de mise en œuvre ». *Les Cahiers du numérique,* 3/11, pp. 11-38.

White R. W., & Roth R. A. (2009). *Exploratory search : beyond the query-response paradigm* (Synthesis lectures on information concepts, retrieval & services). San Rafael (CA.) : Morgan and Claypool Publishers.

Zacklad M., Desfriches-Doria O., Bertin G., Mahe S., Ricard B., Musnik N., Cahier Jean-P., Bénel A., Lewkowicz E. (2011), « MiipaDoc : vers une gestion de l'hétérogénéité des classifications documentaires en entreprise ». *In* I. Saleh, L. Massou, S. Leuleu-Merviel, Y. Jeanneret, N. Bouhai, P. Morelli (dir), *Hypermedia et pratiques numériques – H2PTM'11.* Paris : Hermès Sciences-Lavoisier, pp. 323-333

Zacklad M. (2005). « Vers le Web Socio Sémantique : introduction aux ontologies sémiotiques. ». *Ingénierie des Connaissances.* pp. 1-15, [URL : https://archivesic.ccsd.cnrs.fr/sic_00001347/document], accédé le 14 septembre 2017.

Zhang J., Marchionini G. (2004). « Coupling browse and search in highly interactive user interfaces: a study of the relation browser++ ». *In* H. Chen, M. Christel, E.P. Lim, *Proceedings of the 2004 Joint ACM/IEEE Conference Digital Libraries. Global Reach and Diverse Impact, JCDL 2004*, p. 384.

Zhang J. (2007). « Visualization for information retrieval ». *Springer Science & Business Media,* volume 23, pp. 1-5 [URL : https://people.uwm.edu/jzhang/files/2016/05/p2009_4-1ldij8e.pdf] accédé le 14 septembre 2017.