



HAL
open science

Les apports de la sémiotique dans la valorisation des données numériques.

Peter Stockinger

► **To cite this version:**

Peter Stockinger. Les apports de la sémiotique dans la valorisation des données numériques. : L'exemple des archives audiovisuelles. Semiotique et sciences humaines et sociales: la semiotique face aux defis societaux du XXIème siècle., Université de Limoges - CeReS (Centre de Recherches Sémiotiques), Nov 2015, Limoges, France. hal-01227616

HAL Id: hal-01227616

<https://hal.science/hal-01227616>

Submitted on 19 Nov 2015

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution - NonCommercial - ShareAlike 4.0 International License

Les apports de la sémiotique dans la valorisation des données numériques.

L'exemple des archives audiovisuelles¹.

Peter Stockinger

Institut National des Langues et Civilisations Orientales (INALCO)
Filière Communication Interculturelle (FCI)

Résumé

Dans cette brève contribution, nous souhaitons montrer le rôle, à notre avis central, de la sémiotique structurale pour tout projet scientifique visant la collecte, l'analyse, la diffusion et le partage de corpus de données textuelles *lato sensu* documentant un patrimoine culturel. En nous basant sur nos activités de recherche et de développement que nous menons depuis 2001 sur les *archives audiovisuelles*, nous nous proposons de traiter plus en avant les points suivants :

- a) la problématique des « données » au sens d'*objets textuels* potentiellement signifiants ;
- b) la place des archives numériques (ici : audiovisuelles) dans ce monde des données ;
- c) l'apport théorique de la sémiotique dans des projets de valorisation de données numériques archivées ;
- d) une approche méthodologique se basant essentiellement sur la *sémiotique structurale du texte/discours* pour travailler avec les archives numériques.

1 – La problématique des « données »

De très grandes quantités de données captées, obtenues, collectées par tous les moyens techniques et technologiques se trouvent aujourd'hui *en principe* à la disposition d'acteurs sociaux les plus différents qui peuvent en faire – de nouveau : *en principe* - des usages (très) variés : usages politiques, économiques et commerciaux, usages techniques et scientifiques et, enfin, usages pédagogiques.

Les données, elles, sont de provenance diverse. Une *première catégorie* de données – souvent mais pas exclusivement de nature quantitative - réunit toutes celles qui sont générées par les calculs scientifiques, les données d'observation de la terre, du climat, du cosmos, etc.

Une *deuxième catégorie* de données, fortement convoitées aujourd'hui, concerne ce que l'on appelle l'*identité* ou le *profil numérique* de tout un chacun (son, comme on dit en anglais, « quantified self »). Ces données – captées et enregistrées via des dispositifs particuliers (comme, par exemple, des *interfaces mobiles* ou des « *wearable interfaces* » tels que objets interconnectés, bracelets numériques, lunettes augmentées, capteurs haptiques, enregistreurs audiovisuels, ...) – sont extraites de nos actions et interactions quotidiennes, de nos comportements de consommation, documentent nos réseaux personnels et professionnels, nos goûts et attitudes, nos préférences et valeurs.

¹ Support d'une conférence donnée dans le cadre du colloque « *Sémiotique et sciences humaines et sociales : la sémiotique face aux défis sociétaux du XXIème siècle* » organisé par le Centre de Recherches Sémiotiques de l'Université de Limoges du 25 au 27 novembre 2015 à Limoges.

Une *troisième catégorie* de données – à forte valeur économique ajoutée – est collectée d’une manière infatigable par les directions de communication des annonceurs, les agences de communication-conseil, les instituts de sondage travaillant seul ou en réseaux internationaux. Il s’agit de données (quantitatives ou qualitatives) relatives aux *marchés* de portfolios de produits et de marques, aux segments de *consommateurs* et aux *habitudes* et *pratiques* de consommation.

Une *quatrième catégorie* de données concernent plus particulièrement (mais pas exclusivement) l’e-gouvernement, l’e-citoyenneté et l’e-démocratie. Il s’agit de données générées par les administrations de l’Etat et des collectivités territoriales, par les divers bureaux d’études et de prospectives. Ces données documentent, par exemple, l’usage et la gestion de nos richesses économiques et écologiques, le suivi du train de vie des élus, le processus participatif dans l’élaboration des lois et, plus généralement, des règles de coexistence citoyenne. Citons ici, pour la France, le très officiel site des données [data.gouv.fr](https://www.data.gouv.fr/)² qui invite tout acteur social à partager et à utiliser (réutiliser) cette catégorie de données appelée *données publiques*.

Et, enfin, grâce notamment à la banalisation et à la « massification » de capteurs et d’enregistreurs visuels, sonores et audiovisuels, on assiste, depuis une bonne dizaine d’années à la production massive voire ultra-massive de données dont la première finalité est d’exprimer, de documenter des croyances et de connaissances, des valeurs, des visions, des savoir-faire, des traditions, etc. – bref, tout ce qui renvoie – pêle-mêle – à la dimension « immatérielle » de la ou d’une *culture* (à sa noosphère ou encore aux *mentifacts* de Julian Huxley). En premier lieu, il s’agit ici de données documentant les cultures personnelles et/ou les cultures de communautés d’un genre assez particulier – virtuelles, distribuées dans l’espace, à participants anonymes – autour d’un « leader », d’une marque, d’un thème, d’une préférence, d’un comportement, d’un style de vie, etc. Tout le monde connaît, par exemple, les grandes plateformes de production, de diffusion et d’échange de données textuelles, visuelles ou audiovisuelles qui forment en quelque sorte la colonne vertébrale du *web des réseaux sociaux* : FB, Twitter, Instagram, Pinterest, YouTube, Daily Motion, Google+, Google Livres, Google Actualités, Google Panoramio, ...

Les acteurs publics, aussi, ont développé des programmes très ambitieux de numérisation et de diffusion de très grandes masses de données textuelles et audiovisuelles documentant la culture, le patrimoine culturel dans toute sa diversité. Citons à titre d’exemple Gallica³, la bibliothèque numérique de la BnF, la médiathèque en ligne [ina.fr](http://www.ina.fr/)⁴ ou encore, au niveau européen le projet Europeana⁵ dont l’objectif est de constituer un point d’accès unique à l’ensemble du patrimoine public en Europe. Enfin, on rencontre, aujourd’hui, dans le monde de la recherche toute une diversité de programmes et projets consacrés à la collecte, au stockage, à la diffusion et à la réutilisation de données sous forme de collections de corpus linguistiques et textuels numériques, d’archives audiovisuelles en ligne, et ainsi de suite. Qu’on nous permette de citer ici, comme exemple, le programme des *Archives Audiovisuelles de la Recherche* que nous avons conçu au début des années 2000 et qui forme, aujourd’hui, une plateforme de production, d’analyse et de diffusion de contenus audiovisuels⁶.

Ces différentes catégories de données posent un ensemble de problèmes qui leurs sont communs et qui sont assez bien identifiés :

- 1) La valeur ou le « sens » de ces données (problème auquel s’intéresse surtout les recherches théoriques et appliquées sur les « *smart data* » et les « *linked data* », i.e. le *web des données*) ;
- 2) Les stratégies d’appropriation/de réappropriation, de publication/republication et, enfin, d’utilisation/réutilisation de ces données par des *acteurs* et dans des *contextes sociaux* les plus divers)⁷ ;
- 3) Le cycle de production, gestion, analyse, diffusion, réutilisation et conservation des données (= question des nouveaux écosystèmes sémiotiques et culturels de communication *lato sensu*) ;
- 4) Le statut des données au sens de *données ouvertes*.

Ces quatre problèmes renvoient tous à la *même question centrale* qui est celle de savoir « quoi faire avec (ces données) », i.e. comment *appréhender leur sens*, comment les *rendre signifiants*, comment en faire – pour parler avec Greimas – des *objets de valeur*.

² url : <https://www.data.gouv.fr/fr/> ; au niveau européen, cf. le site portail des données ouvertes de l’Union Européenne : <https://open-data.europa.eu/fr/data>

³ url : <http://gallica.bnf.fr/>

⁴ url : <http://www.ina.fr/>

⁵ url : <http://www.europeana.eu/portal/>

⁶ pour en savoir plus : Stockinger (2003) ; Stockinger (2011) ; Stockinger (2012)

⁷ cf., par exemple, les travaux et recherches du réseau international DDI (Document, Discover and Interoperate) : <http://www.ddialliance.org/>

Sous le sigle de la « sémantisation des données » (i.e. des « smart datas »), se pose trois questions tout à fait complémentaires :

- 1) Comment identifier, localiser – dans la masse des données numériques – les données qui possèdent une pertinence, une *valeur* pour un acteur social donné, qui forment, autrement dit, des *signes potentiels* pour un acteur social en quête d'informations ou de connaissances sur un domaine, un objet documenté par ces données ?
- 2) Comment décrire, analyser la *structure sémiotique* – le *contenu* et son *expression* – de ces données pour pouvoir expliciter leur *valeur (potentielle)* pour un acteur social en quête d'informations ou de connaissances sur un domaine, un objet particulier ?
- 3) Comment expliciter, enfin, le *positionnement* des données les unes par rapport aux autres, i.e. comment expliciter les *rapports intertextuels* entre un ensemble de données qui constituent (hypothétiquement) les éléments signifiants – la *semiosphère* - du paysage textuel – du *textscape* – d'un domaine ou d'un objet dont un acteur social est en quête de connaissance.

Ces trois questions renvoient à la problématique plus générale de la *valorisation* de données ou encore à celle de la création de *chaînes de valeurs* (pas obligatoirement restreintes aux organisations économiques – bien loin de là !) à partir de ou sur la base d'un fonds de données stockées, par exemple, dans des entrepôts ou archives numériques. Technologiquement parlant, il s'agit ici de la conception, implémentation et utilisation de *plateformes d'outils*, de *méthodes* et de *bonnes pratiques* pour :

- 1) la *localisation* de données dans un fonds (fermé ou ouvert) ;
- 2) la *description*, la *classification*, l'*analyse*, l'*interprétation*, ... de données localisées
- 3) l'*éditorialisation* (publication/replication, visualisation, ...) et l'utilisation de données préalablement localisées et analysées.

Si nous prenons l'exemple des fonds de données composant les archives audiovisuelles, un des problèmes les plus ardues est celui des *modèles* qu'il faut utiliser pour interpréter – classer, indexer, commenter, comparer, ... – des corpus choisis de ces données afin d'expliciter leur *valeur* (i.e. leur sens) pour un acteur social donné. Ces modèles doivent à la fois tenir compte des *contraintes structurales* des données textuelles et des usages pour lesquels ces données sont prévues. C'est ici où la sémiotique – et je pense ici plus particulièrement à la sémiotique structurale (pour ne pas dire greimassienne) – peut jouer un rôle pivot.

La question des modèles de description de corpus de données prend place dans le cadre des recherches théoriques et appliquées consacrées aux *ontologies* au sens informatique du terme, i.e. aux *structures conceptuelles* qui sous-tendent un modèle ou une bibliothèque de modèles de description. Les ontologies sont des *vocabulaires*, des *terminologies* qui expriment ou désignent des *notions*, des *concepts* relatifs à un *domaine de connaissances* et selon la *culture (l'expertise)* de leurs auteurs et les *contextes d'usage* auxquels ils sont destinés. Les activités de conception, développement, validation et enrichissement d'ontologies font partie du domaine de l'*analyse et de la représentation de connaissances* et s'apparentent à maints égards à celles de l'analyste de corpus de textes *lato sensu* et aussi à celles du lexicologue-terminologue.

Aujourd'hui, on peut constater une tendance de prolifération incontrôlée de telles structures conceptuelles, ce qui pose toute une série de questions scientifiques et techniques extrêmement épineuses telles que :

- la portée et la valeur empirique d'une ontologie ;
- les principes de construction d'une ontologie ;
- la structure canonique qui « fonde » une ontologie ;
- le degré de généricité d'une ontologie ;
- la « portabilité » d'une ontologie (i.e. la possibilité de l'utiliser en dehors du domaine pour lequel elle a été conçue) ;
- l'identité et le statut des *catégories conceptuelles de base* d'une ontologie, voire de toutes les ontologies ;
- la distance conceptuelle entre deux ou plusieurs ontologies et la traductibilité d'une ontologie à une autre ontologie ;
- l'intégration, dans une ontologie spécialisée, de parties d'un langage « populaire » (exemple : l'intégration de certaines parties pertinentes de RAMEAU de la BnF⁸ dans une ontologie spécialisée sur la diversité culturelle) ;

⁸ cf. à ce propos le site indispensable databnf.fr : <http://data.bnf.fr/semanticweb>

- la prise en compte de standards dans une ontologie (par exemple, la prise en compte du formalisme OWL⁹ dans l'élaboration d'une ontologie spécifique, la prise en compte de standards communs tels que DCMI¹⁰ ou EAD¹¹, etc.) ;
- la motivation – linguistique, cognitive, pragmatique, formelle, ... - d'une ontologie (i.e. l'organisation d'une ontologie est-elle « simplement » justifiée du point de vue de son utilisation, reflète-t-elle des visions culturelles ou scientifiques particulières ; est-elle motivées par une catégorisation lexico-sémantique proche d'une langue naturelle ; ...) ;
- la dimension de l'évolution des concepts d'une ontologie.

Enfin, la question de la valorisation, de la création de chaînes de valeurs autour d'un ensemble de données potentiellement signifiantes pour un acteur social, est étroitement liée à celle des *données ouvertes* (« open data »). Celle-ci renvoie à différents problèmes parmi lesquels on trouve d'une manière récurrente :

- le problème de l'accessibilité des données (numériques) au sens *technique* du terme ;
- le problème de *rendre public* les *données publiques*, c'est-à-dire les données générées par l'administration de l'Etat, des collectivités territoriales et locales, de la recherche publique, etc. ;
- le problème de la conservation à long terme des données d'intérêt général (par exemple, des données scientifiques, des données du patrimoine culturel, ...) ;
- le problème de l'enrichissement des données, de leur utilisation ou réutilisation dans des contextes et cadres d'activités spécifiques.

Ce sont des enjeux importants pour le domaine de l'*e-gouvernement* et de la *démocratie participative*. Citons ici comme exemple le portail du gouvernement français dédié à la *diffusion des données ouvertes*¹² ; le portail OpenData France¹³ des collectivités territoriales et locales ou encore le très intéressant portail de l'association *Parlement&Citoyens* qui utilise l'approche « open data » pour influencer l'élaboration des lois ou des projets de lois¹⁴. La question des données ouvertes est centrale dans le cadre de l'émergence d'une nouvelle économie participative. Il faut mentionner ici le rôle important de la *BPI (Banque Publique d'Investissement) France*¹⁵ avec son plan d'investissement pour soutenir des projets de marché participatif. Elle est, enfin, essentielle dans le cadre de la recherche et de l'enseignement (citons ici, à titre d'exemple, le portail de la *Open Knowledge Foundation*¹⁶ avec ses différents projets et plateformes autour des données ouvertes).

2) Les archives numériques

La notion « archives numériques » revêt deux problématiques différentes mais qui sont indissociablement liées :

1. les archives au sens d'une *banque de données* (ouvertes) ;
2. les archives au sens d'une *ressource ou d'un ensemble de ressources* (d'un centre de ressources).

Les archives au sens d'une *banque de données* (ouvertes, d'accès libre, ...) font partie d'un *projet d'archivage* spécifique, se présentent sous forme de collections de données numériques (composant le fonds – fermé ou ouvert – d'une archive) et peuvent servir *potentiellement* à toute sorte d'exploitations pratiques. Dans le second sens, les archives numériques servent d'objet de recherche théorique et appliquée en vue, par exemple, d'améliorer les *accès aux données archivées*, d'avancer sur les *terrains de la sémantisation et de l'interconnexion* des données archivées (cf. ci-dessus) ou encore de mieux contrôler le processus de *l'éditorialisation* et de la *visualisation (lato sensu)* des données et des métadonnées.

Les archives au sens d'un *entrepôt de ressources épistémiques* mobilisables dans des contextes d'usages les plus variés renvoient à la problématique centrale de la *réutilisation « active »* d'une donnée ou d'un ensemble de données archivées ce qui implique le plus souvent une *transformation qualitative* des données concernées

⁹ Cf. Lalande, Steffen ; Beloued, Abdelkrim et Stockinger, Peter : Modélisation et formalisation RDFS/OWL d'une ontologie de description audiovisuelle ; in/ Les Cahiers du Numérique 11/3, 2015, pp. 39 - 70

¹⁰ Dublin Core Metadata Innovation : <http://dublincore.org/>

¹¹ Encoded Archival Description : http://www.bnf.fr/fr/professionnels/formats_catalogage/a.f_ead.html

¹² Site : <https://www.data.gouv.fr/fr/>

¹³ Site : <http://www.opendatafrance.net/>

¹⁴ Site : <https://www.parlement-et-citoyens.fr/>

¹⁵ Site : <http://tousnosprojets.bpifrance.fr/>

¹⁶ Site : <https://okfn.org/>

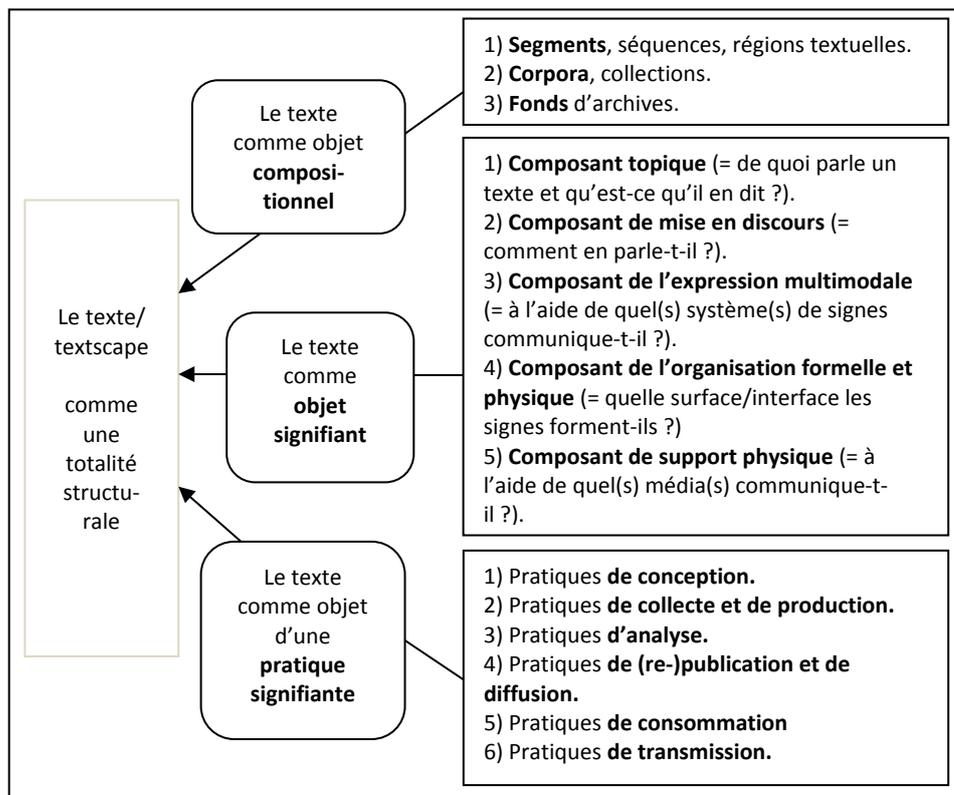
(Stockinger, 2011, 2012, 2015). Voici quelques activités tout à fait récurrentes qui illustrent cette notion de *transformation qualitative* d'une donnée (d'un ensemble de données) en une *donnée signifiante* (un *corpus de données signifiantes*) :

- la *reclassification* (de données archivées et déjà classées dans une archive) selon un plan de classification propre à l'utilisateur ou un groupe d'utilisateurs ;
- la *redescription* et *réindexation* (de données archivées déjà indexées) à l'aide de modèles de description appropriés aux besoins de l'utilisateur ;
- la *recontextualisation* (de données archivées et déjà mises en relation avec d'autres données) ;
- le *redécoupage* virtuel (de données audiovisuelles, visuelles, textuelles... archivées) pour ne retenir que les passages intéressant l'utilisateur ;
- la *réécriture* partielle (*i.e.* la modification même de la textualité d'origine d'une donnée archivée) sous forme, par exemple, d'un montage virtuel de la donnée archivée avec différents éléments textuels (segments audiovisuels, ...)
- la *rééditorialisation* et la republication d'une ou d'un corpus de données archivées.

Ces différentes activités montrent toute la pertinence de la *vision sémiotique* (du « semiotic turn », *cf.* Stockinger, 2015 ; *cf.* aussi Craven 2008 qui adopte un point de vue similaire) pour la recherche sur les archives numériques : d'une manière générale, le fonds d'une archive peut être considéré comme un ensemble (quantitativement plus ou moins important) de *données textuelles* au sens large du terme qui à la fois documentent, conservent et transmettent le *discours* d'un *acteur social* (d'une organisation, d'un groupe social, d'une communauté, voire d'une personne) sur un domaine, un objet ou encore une époque¹⁷.

3) Sémiotique des archives

Afin de rendre opérationnel cette vision générale du « semiotic turn », il faut avoir recours à un cadre théorique du texte et du discours dont les principaux éléments sont représentés par la figure 1.



¹⁷. Cette vision fait référence aux travaux antérieurs de Michel Foucault sur la fonction des archives dans la constitution d'une *rationalité* (par exemple *médicale* ou *pénale*, en ce qui concerne les recherches de Foucault), *i.e.* sur la constitution d'un ensemble de savoirs et savoir-faire formant un ensemble de standards (de traditions, normes ou règles) auxquels se réfère un corps social et/ou professionnel (médecins, psychiatres, pénalistes...) pour exercer son métier.

Figure 1. La vision sémiotique structurale sous-tendant le processus de modélisation conceptuelle des procédures d'analyse d'un objet/corpus d'objets média

Brièvement résumé, l'objet « texte » n'est pas du tout restreint à son avatar culturel du texte imprimé (du livre, du journal...) mais recouvre plutôt toute « entité signifiante » (selon Greimas, 1966 et Greimas et Courtés, 1979) pour un acteur, *i.e.* toute entité qui possède un certain sens pour un acteur (figurent ainsi parmi les « données textuelles » notamment les *artefacts*, les *objets naturels* qui possèdent un sens pour un acteur donné ainsi que les *gestes* qui expriment les activités de toutes sortes). Dans le cas des archives audiovisuelles qui nous intéressent ici, ces entités signifiantes sont restreintes à quelques grandes catégories : l'*écrit* et l'*imprimé*, le *visuel statique*, l'*audiovisuel* et le *sonore* (comprenant la parole, la musique et le bruit).

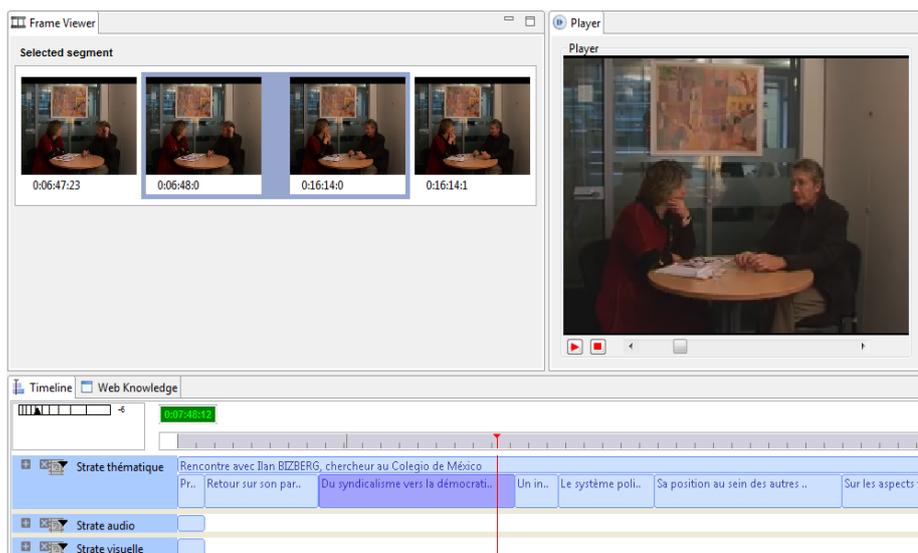


Figure 2. Interface donnant accès au modèle d'identification et de segmentation – thématique, visuelle, sonore... – d'une vidéo)

Cette vision sous-tend et motive l'ensemble des ressources métalinguistiques (terminologies, modèles de description) que nous avons commencé à concevoir, sous l'acronyme ASA (Atelier de sémiotique audiovisuelle), dès 2006 dans le cadre de plusieurs projets européens et français (cf. Stockinger 2011a ; Stockinger 2012). Une première version de ces ressources métalinguistiques qui forment des classes de *graphes conceptuels* a été concrètement utilisée en 2009 dans le cadre d'un projet ANR du même nom¹⁸. Une deuxième version est actuellement en cours d'élaboration dans le cadre du projet *Campus AAR*¹⁹ qui sera conforme aux exigences formelles propres au web sémantique (RDFS ; OWL). L'objectif du projet Campus AAR est de fournir à la « communauté » académique des SHS (sciences humaines et sociales) un environnement intégré de production, d'analyse, d'édition et de diffusion d'archives audiovisuelles. Cet environnement est composé d'un outil appelé *Studio Campus AAR* qui comprendra un outil d'analyse de corpus d'objets média et un autre outil réservé à l'édition et la publication de corpus audiovisuels. Le « cœur » du studio Campus AAR est, autrement dit, composé de deux classes de modèles conceptuels :

– Le premier type recouvre les *modèles d'analyse* (ou scénarios d'analyse) qui représentent l'*univers du discours d'une archive* appréhendé selon les différentes facettes composant le schéma dans la figure 1 et qui forment *autant de clés* pour l'interprétation ou la réinterprétation du patrimoine d'une archive. L'analyste d'une archive dispose, simplement dit, d'une série de *formulaires dynamiques* (formant l'interface de l'outil d'analyse d'objets média) à l'aide desquels il décrit son corpus d'objets média.

– Le deuxième type recouvre les *modèles de publication* mis à la disposition de la (ou des) personne(s) ayant en charge l'édition et la diffusion des objets média préalablement analysés.

¹⁸ Le projet ASA-SHS de 2009 à 2012), cf. aussi le blog du projet <http://asashs.hypotheses.org/>

¹⁹ Ce projet coordonné par Peter Stockinger, comprend un consortium de 4 partenaires français (Equipe Sémiotique Cognitive et (nouveaux) Médias/FMSH, Institut National de l'Audiovisuel, CNRS – CCSd et société Armadillo) et a débuté en janvier 2014 pour une durée de 36 mois. Plus d'informations sur le blog du projet: <http://campusaar.hypotheses.org/>

Concrètement parlant, l'analyste souhaitant interpréter ou réinterpréter une ressource audiovisuelle, l'enrichir, l'adapter à un usage (par exemple pédagogique ou scientifique) spécifique, dispose d'un outil lui permettant de localiser la ou les ressources, voire telle ou telle séquence spécifique d'une ressource (figure 2). Ce même outil intègre une interface composée de toute une série de modèles à l'aide desquels l'analyste peut ensuite procéder à son travail d'interprétation (figure 3) : analyser les sujets ou thèmes d'un objet média ou d'un corpus d'objets média, analyser la mise en discours d'un sujet ou d'un thème, analyser les stratégies d'expression visuelle et/ou acoustique d'un thème, expliciter le champ intertextuel d'un objet média ou d'un de ses thèmes, etc.

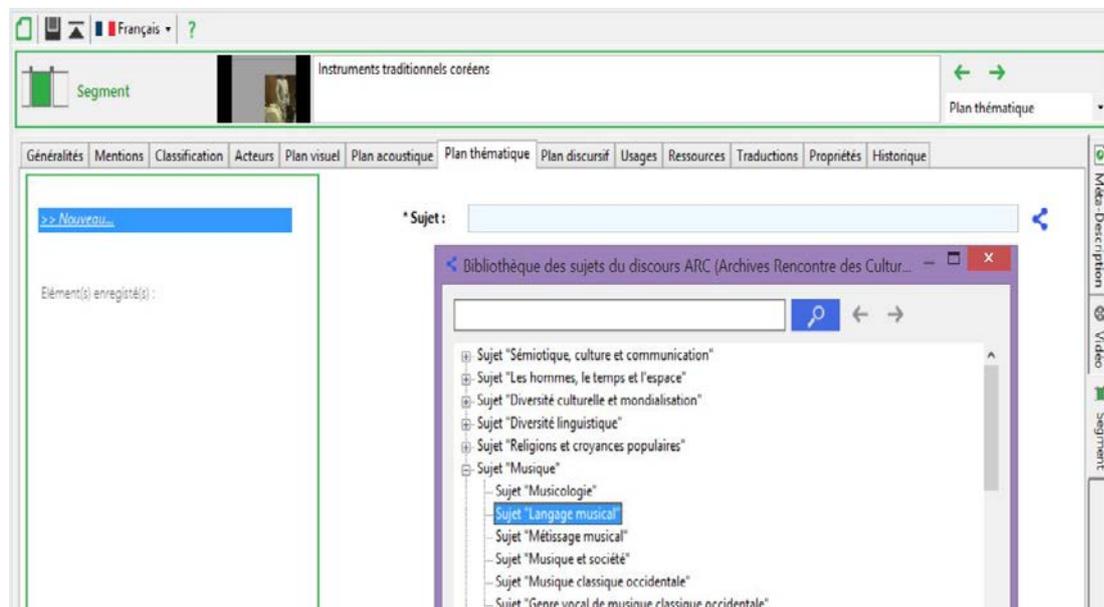


Figure 7. Interface donnant accès à une bibliothèque de modèles d'analyse servant à l'analyse d'un objet média/ d'un corpus d'objets média

Ce travail d'interprétation ou de réinterprétation de ressources audiovisuelles équivaut, dans la terminologie du web 3, à la sémantisation, à l'interconnexion ou encore à l'ouverture auctoriale d'une donnée. Ajoutons encore que la (dans la terminologie du web 3) *visualisation des données* est prise en charge par les modèles de publication ou de republication d'une ou d'un corpus de ressources audiovisuelles. Nous n'avons pas l'espace nécessaire ici pour développer cet aspect évidemment essentiel dans une politique de valorisation de données. Dans le cadre de nos recherches sur la valorisation des archives audiovisuelles, cette partie se divise en une partie appelée « *publication automatique* » et une deuxième appelée « *publication auteur* ». La partie publication automatique utilise toutes les métadonnées générées dans la phase de l'analyse d'une donnée audiovisuelle (ou d'un corpus de données audiovisuelles) pour en produire une *nouvelle forme éditoriale*. La partie « publication auteur » offre à l'analyste de *personnaliser* une nouvelle édition d'une donnée audiovisuelle préalablement analysée. Ces deux processus sont exemplifiés par une série de portails à thème que nous avons conçus et réalisés avec notre équipe ESCoM depuis 2011. Il s'agit des portails AMSUR²⁰ (dédié à la valorisation et diffusion d'un patrimoine audiovisuel sur l'Amérique latine), du portail ARC²¹ (dédié à la valorisation et diffusion d'un patrimoine audiovisuel sur la diversité culturelle), du portail AGORA²² (dédié à la valorisation et diffusion d'un patrimoine audiovisuel en SHS) ou encore du portail AHM²³ (dédié à la valorisation et diffusion d'un patrimoine audiovisuel en histoire des mathématiques).

Voici donc, en quelques lignes, un exemple d'utilisation de la sémiotique (structurale) dans le contexte de projets de production et de valorisation d'archives numériques qui font partie de ce gigantesque monde des données dont l'un des enjeux centraux est de les « *faire parler* », de les *rendre signifiantes*, pour utiliser une expression d' A.J. Greimas.

²⁰ Cf. <http://www.amsur.msh-paris.fr/>

²¹ Cf. <http://www.arc.msh-paris.fr/>

²² Cf. <http://www.agora.msh-paris.fr/>

²³ Cf. <http://www.ahm.msh-paris.fr/ahmfr.aspx>

