

Compte rendu : 2^e séance Séminaire archives audiovisuelles : 3 mai 2016

Présents : *Alain Bertho* (MSH Paris Nord), *Serge Wolikow* (Arcmc), *Aurelia Vasile* (Arcmc), *Céline Alazard* (Arcmc), *Marianne Herard* (MSH Paris Nord), *Bernard Gannes* (projet Canson, CNRS Lyon), *Mélanie Karli* (BDIC), *Rosa Olmos* (BDIC), *Emmanuelle Papinot* (FMSH), *Nicolas Bontemps* (LESC), *Eléonore Alquier* (INA), *Malika Mekdad* (INA), *Hélène Allioui* (INA), *Cécile Kattnig* (BNF), *François Lemaître* (FMSH), *Damien Poïvet* (FMSH), *Anne-Claire Delaunay* (FMSH).

Présentation indexation « FMSH Production » : Anne-Claire Delaunay

Le service dont Anne Claire Delaunay a la responsabilité a été créé il y a 4-5 mois.

Contexte : « FMSH Production » : entité de la FMSH qui a comme but de produire des contenus vidéo. Le but :

- réaliser des captations vidéo des séminaires, conférences, etc qui émanent notamment du **Collège d'études mondiales** porté par la FMSH.
- Produire des contenus de communications : interview, films institutionnels
- Répondre à des besoins extérieurs (demandes ministérielles pour des MOOC, interview pour une remise de prix, etc)

Il existe maintenant des formats, des contenus multiples ce qui pose le problème de l'indexation. Une grosse partie est indexée via Canal U. Pour cela, il est nécessaire de créer un compte et de remplir des champs déjà identifiés. (Le rendu visuel pour le lecteur se présente comme dans l'image 1). Les données sont conservées sur les serveurs de la FMSH dans un data center à Vitry.

Particularités de l'indexation sur Canal U (éléments définis lors des discussions) : chaque contributeur remplit les fiches. Il ne faut pas que cette procédure soit trop lourde et que les champs soient pas trop variés pour ne pas décourager les contributeurs qui peuvent aller sur You tube.

Critères pour être contributeur : appartenir à des institutions d'enseignement supérieur et de recherche, institutions patrimoniales, etc ; y compris MSH, bien que le nombre de MSH soit actuellement réduit.

Difficultés :

- Le problème apparaît lorsque les vidéos n'ont pas vocation à se retrouver sur Canal U. Ces vidéos doivent être indexées d'une autre manière. Il y a un tableau pour faciliter l'indexation avec quelques champs prédéfinis, mais les champs ne sont pas

automatisés, c'est trop manuel.

- Sur Canal U, les champs sont prédéfinis et donc également. On peut rajouter des onglets, mais pas des champs. Dans cet onglet on peut préciser des détails, des descriptions, mais il n'y a pas des champs en plus à la disposition des contributeurs. (Cela peut être étudié sur demande et effectué, par l'équipe technique du Canal U).
- La FMSH a récupéré les fonds du CERIMES (60 ans de vidéos scientifiques) et des AAR (Archives Audiovisuelles de la Recherche) qui contient environ 7000h de vidéo. Comment intégrer ces fonds et l'indexation correspondante de manière à l'harmoniser avec l'indexation de Canal U. Les AAR seront migrées automatiquement sur Canal U (avec les données d'indexation) grâce à un programme développé spécialement.

Questions : Est-ce que le système d'indexation de Canal U est suffisant pour qu'un internaute s'y retrouve ? (dans un premier temps, oui, mais plus il y a des vidéos, plus cela devient complexe de retrouver la vidéo qui intéresse). Existe-t-il un standard d'indexation, une indexation idéale/normée ? Y a-t-il des logiciels pour cela ?

Il existe des multiples usages de l'indexation en fonction de la nature du public : chercheur, opérateur vidéo/technicien (intéressé par des informations concernant l'autorisation de filmer le public, le nom du cadreur, de l'opérateur). L'indexation pour la recherche comprend des champs plus détaillés avec plus de description (comme l'avait fait Francis Lemaître pour studio ASA avec un système fin et complexe d'indexation qui demande une expertise pour remplir chaque champ)

Objectif : trouver une méthode pour réaliser une indexation harmonisée/standardisée de manière à faciliter la migration vers d'autres sites et plateformes.

Présentation indexation INA (en 3 temps : Eleonore Alquier, Malika Mekdad, Hélène Allioui)

Eleonore Alquier : introduction

L'INA a des collections qui ont été constituées en deux temps autour de statuts juridiques bien distincts :

- la récupération et l'héritage des fonds de l'ORTF (avec les fonds dévolus, car l'INA est l'archiviste « naturel », par convention, des diffuseurs nationaux, soit essentiellement France télévisions et Radio France). C'est une pratique que l'INA a étendu sous la forme de prestations auprès d'interlocuteurs qui sont soit des diffuseurs comme par exemple TF1, soit des structures dont la diffusion n'est pas le métier mais qui produisent des archives audiovisuelles (ex. des banques, des instituts culturels,

etc.) ; toute structure peut solliciter les services de l'INA pour déléguer l'archivage a- v. Cela représente plutôt la position d'usage sur des « volumes maîtrisés » et avec une finalité de gestion de droit de la propriété intellectuelle, puisque la prestation que l'INA propose comprend non seulement l'archivage, mais aussi la gestion de la vie des archives y compris dans la réutilisation qui peuvent en être faites, les ventes, etc.

- au titre du dépôt légal qui existe depuis plus de 20 ans pour la télévision et la radio, et depuis 10 ans pour internet.

Le premier statut de collecte représente aujourd'hui environ 1, 5 millions heures, alors que le dépôt légal représente actuellement plus de 13 million d'heures depuis 20 ans et de manière croissante (environ 1 million d'heures collectées chaque année), puisque le périmètre est plus large (plus de chaînes de diffusion)

L'INA développe un projet de refonte de son système documentaire, qui est aujourd'hui bicéphale puisqu'il est constitué par rapport à ces deux statuts : deux environnements de travail co-existent, deux environnements d'hébergement, deux bases qui dialoguent épisodiquement.

Ce nouveau projet, estimé à une durée de 5 ans, consiste en la refonte des données, dans une démarche de *big data*, et la refonte de l'applicatif. Les questions posées sont similaires à celles de la FMSH : qu'est-ce qu'on peut voir/ utiliser comme fonctionnalité pour proposer une indexation et d'autres formes de traitement documentaire.

Pour les chiffres et historique des projets INA depuis les années 1960, voir pages 2, 3, 4, 5 du PDF).

Malika Mekdad : Les fonds professionnels (FP)

L'Etat a obligé le diffuseur public (ORTF puis France Télévisions et Radio France), par une convention d'archivage, à déposer son fonds à l'INA : l'INA recevait donc à sa création la dévolution des droits des productions des diffuseurs publics télévisuels ou radiophonique ; l'INA récupérait les *droits producteurs*. Suite à la loi d'août 2000, c'est France télévision qui récupère les droits producteurs à partir de 1997.

C'est par le biais de cette procédure de récupération des fonds que l'INA a constitué les fonds dits « professionnels », sur la base de ce qui était produit et coproduit par le diffuseur public.

Chronologie : (voir page 6 du pdf). Début par les fonds radio (y compris les radios régionales). Tout n'est pas conservé. Il y a un comité de sélection qui sélectionne les programmes de radio qui vont être conservés au titre de fonds « professionnels ».

Un fonds intéressant, sollicité par le public et les chercheurs : « Les actualités françaises »

(1800 heures) : très riche pour la période de la 2GM : support films-cinéma de 35 mm, très bonne qualité.

Entrée de nouveau fonds : des mandats patrimoniaux, donations, producteur privés, institutions, etc. (ex. TF1 a de nouveau recours aux prestations d'archivage de l'INA, même en étant privé ; fédération sportives, etc). + les images amateurs + Fonds photo (illustration des programmes TV + illustrations des émissions radio, etc).

Les outils de traitement ont été conçus sur la distinction des deux types de média, la radio et télé. Cette structuration par média a structuré les bases et les applications qu'utilisent aujourd'hui l'INA, mais aussi les principes d'organisation et les organigrammes. Le cas des émissions issues de la radio filmée illustre le dépassement des frontières par média auxquels les traitements et les outils doivent aujourd'hui s'adapter.

Valorisation voir page 7 et 8 du pdf

- Traitement documentaire des fonds « professionnels » : ne suit pas une logique de flux mais une logique de collection.
- Traitement documentaire dans le cadre de la constitution de corpus multimédia ou éphémérides
- Préparation éditoriale
- L'intégrité des bases de données

A voir les fresques multimédia sur ina.fr

Présentation de la base de données Totem (page 9-12) Outil de gestion intégré documentaire, développé pour l'INA. Il fonctionne avec deux moteurs : une base Oracle pour toute la partie de production documentaire et le moteur Exalead pour la recherche.

Hélène Allioui : Le dépôt légal (p. 14)

Instauré par le décret de 31 décembre 1993 et mis en place en 1995, il stipule que tous les producteurs et les diffuseurs de radiotélévision doivent verser tous les documents a-v et écrits de production ou co-production française de première diffusion et de diffusion nationale auprès de l'INA. L'INA doit conserver ces documents à des fins patrimoniales et doit les mettre à disposition du public à des fins de recherche.

Actuellement, le dépôt légal couvre 100 chaînes de télévision et plus de 50 chaînes de radio.

Le traitement documentaire : doit refléter le flux de la programmation quotidienne et la structuration de chaque émission.

A chaque document av est associée une notice documentaire. Ex : pour les 7 chaînes Hertiennes en 2015 ont été créés 289000 notices documentaires, pour les 58 chaînes câbles

satellites, plus d'un million de notices. Seulement une partie de la captation du flux fait l'objet d'un traitement documentaire, selon des critères mettant en valeur le propre de la diffusion françaises (journaux télévisés, magazines, ...).

Commentaire d'une notice documentaire de travail : page 17

[les outils de création et de traitement sont les mêmes depuis 1995, ce sont des outils sur mesure ; c'est le même environnement de travail, les mêmes bases depuis la création du dépôt légal de la radio et de la télévision]

Une grande partie du travail de description des fonds s'appuie sur les imports de données, soit achetées auprès de sociétés privées (Imedia, Médiamétrie), soit issues des bases documentaires des partenaires de l'Ina (France Télévisions, Radio France)..

Conclusions :

Outils anciens. Objectifs : refaire ces applicatifs, sortir des silos qui peuvent exister entre les différents statuts de collecte qui produisent parfois des doublons. Besoin d'assainir les bases, dans une démarche de big data ; de mettre en communication les outils documentaires et les données d'ordre juridique, la direction de marketing et de vente : il s'agit donc d'un projet de restructuration de la donnée.

En parallèle, par le biais du département de la recherche, la réflexion porte sur les innovations qui ont pu être développées ces dernières années et qui correspondraient à un réel cas d'usage en interne, afin de faciliter le travail et d'améliorer la productivité.

Un exemple de démarche innovante à l'INA : la reconnaissance automatique d'images et de sons, sur la base d'un dictionnaire de voix ou d'images représentatives (ex. présentateurs, visages, marques, etc)

Un autre objectif du projet de refonte : établir une interopérabilité plus efficace, car l'INA a conçu ses propres outils et référentiels, mais cela peut empêcher interopérabilité. L'objectif serait l'ouverture des données et l'homogénéisation.

Discussions et échanges (1h05)

Normalisation : il existe une norme d'indexation des documents audiovisuels, AFNOR¹ : qui a instauré une liste de champs qui peuvent être adaptés en fonctions des besoins.

Serge Wolikow soulève 2 questions sur les **archives audio-visuelles du web** : comment l'indexation s'articule avec l'indexation de base ? [*Réponse INA* : depuis que le dépôt légal

¹ AFNOR FD Z 44-065 Septembre 1998 : Documentation - Catalogage des vidéogrammes - Rédaction de la description bibliographique
AFNOR FD Z 44-066 (1988) : idem pour les enregistrements sonores

existe, l'INA a construit une machine dédiée partant du principe qu'une page internet est un élément composite (photo, texte, vidéo) et on n'a pas les moyens humains pour faire une indexation, mais la page internet en elle-même représente l'articulation entre les éléments qui la composent qui sont archivés distinctement. Cela amène à la reconstitution de tous les éléments, tels qu'ils existaient au moment de leur création, pour la consultation. Ce qui est indexé au niveau web, c'est le site achevé (ligne éditoriale, politique, globalement, indexation macro). En revanche, l'indexation fine du contenu qui existe dans les pages, ce n'est que de l'indexation automatisée, sur la base des techniques de fouille de données, de datamining qui consiste à récupérer des contenus et à le rendre interrogeables. A savoir : l'indexation des sites ne s'appuie pas sur les mêmes référentiels que l'indexation du flux av. D'où les projets de fusion des bases et d'homogénéisation des référentiels]

Archives audiovisuelles de la recherche, produites par des chercheurs (documents hétérogènes : films de chercheurs, discours de chercheurs) est un objet qui intéresse particulièrement le consortium Arcmc. Comment les outils maisons de l'INA et l'indexation proposée peuvent-ils répondre à des besoins de recherche et d'indexation d'autres institutions comme les AN, BDIC, les universités, MSH, etc. ? Il pourrait être intéressant de développer des coopérations communes entre l'INA et ces autres institutions, par le biais de la TGIR. [Réponse INA : Les archives de la recherche sont à envisager sous l'angle du statut d'archives publiques, en lien avec les Archives nationales pour l'archivage définitif. L'INA réfléchit à une cartographie des fonds audiovisuels, dans une démarche de collecte au-delà du seul flux du dépôt légal, afin d'identifier les domaines/patrimoines audiovisuels où elle peut apporter une plus-value, dont la recherche peut faire partie.]

Perspectives institutionnelles et objets de coopération :

Serge Wolikow : compte tenu des mandats et de l'historique de l'INA il semble normal qu'il y ait des coopérations avec les universités, les MSH, les grandes bibliothèques publiques dans le domaine de l'av. Il n'est peut-être pas nécessaire que toutes les structures réinventent des choses qui existent déjà, il faut du partage, notamment par le biais des consortiums et des grandes infrastructures de recherche.

CK (BNF) : Le premier niveau minimal de partage devrait être l'utilisation des référentiels uniques qui puissent favoriser la circulation et la mutualisation des données. (Utilisation de Dublin Core minimum). ACD (Canal U) suggère même la mise en place d'un logiciel unique pour tous les MSH pour réaliser l'indexation.

SW: considère que l'objectif fondamental de ce séminaire est de réfléchir, dans la perspective du lancement d'une nouvelle phase de labellisation du consortium, à l'élaboration d'un

ensemble de propositions et préconisations concernant le domaine de l'audiovisuel. La position du consortium à l'intersection d'un ensemble de MSH, laboratoires, archives, bibliothèques. L'objectif serait de faire travailler ensemble toutes ces institutions ensemble ; c'est une démarche originale dans le champ de la recherche en France, car jusqu'à maintenant la coopération a été bilatérale, mais presque jamais multilatérale. Ce séminaire peut fonder les bases de réflexion pour des collaborations futures ; on met en place des synergies qui doivent nous permettre d'avancer vers des bonnes pratiques, des documents de référence, des personnes ressources etc. Cela doit se faire dans le respect de l'identité de chacun, dans le respect des tutelles de chacun.

Piste de réflexion sur la nature de l'indexation et ses différents usages

ACD : Canal U est un programme qui respecte l'identité de chaque contributeur et il est regrettable que les MSH ne sollicitent pas davantage les services de cette plateforme on-line. Un premier pas dans la perspective de collaboration institutionnelle, serait d'inciter les MSH à diffuser leur contenu à travers Canal-U ; cela peut contribuer à l'harmonisation d'indexation.

EA (INA) : considère que la solution n'est pas forcément dans la mise en place d'un logiciel commun, ni même dans la mise en place des référentiels standard, bien que cela soit utile, mais dans le dialogue entre les personnes en charge des différents projets qui permettraient une indexation plus ou moins affinée, en fonction des nécessités de chaque projet.

FL (FMSH ; AAR) : soulève la question de la définition du terme indexation qui peut vouloir avoir de multiples sens et usages. Dans le cas de AAR cela fait référence à la fonction de catalogage, c'est-à-dire des notices standard de description (titre, auteur, durée, thématique principales), issus de la norme Dublin Core. Dans le cas de Canal U, il possède un moteur OAI ce qui lui permet d'être moissonné par Isidore et donc de diffuser ces données à travers d'autre plateforme de requête. Cette indexation de base peut varier en fonction des fonds qui sont indexés et du public auquel il s'adresse (un chercheur peut avoir des besoins plus précis en terme de référentiels par rapport à ce que Dublin Core offre.)

EA (INA) : l'INA propose à l'Inathèque l'outil *Media Corpus*, qui permet la constitution de corpus par les chercheurs même. Ce sont les chercheurs qui créent leurs propres corpus pour la recherche (une sélection subjective et assumée à partir des fonds proposés par l'INA). Ce n'est pas l'INA qui crée les corpus pour les chercheurs, mais les chercheurs eux-mêmes. L'INA propose un formulaire de recherche sur le modèle des légos, c'est-à-dire manipulable en fonction des besoins de l'utilisateur, leurs propres critères.

Le travail fait par le chercheur dans la sélection des fonds et la constitution de son propre

corpus ouvre la question de l'indexation collaborative, car le travail que le chercheur fait pour constituer son corpus pourrait contribuer à l'enrichissement de la base avec des éléments de description plus fins. Cela n'est pas encore possible à l'INA, car l'espace de constitution du corpus appartient uniquement au chercheur. La base de consultation est une base passive. Un angle de réflexion de l'INA serait de fusionner les bases de production et de consultation. Le problème de l'indexation collaborative est plus large, car une fois le principe accepté, il reste la question de l'organisation : pour donner des consignes aux chercheurs/utilisateurs sur les normes d'indexations acceptées, pour mettre en place une table de validation etc. Aujourd'hui les limites sont surtout techniques et l'objectif INA est de se focaliser sur la maintenance de ces bases déjà existantes pour permettre leur continuité dans le temps et offrir aux chercheurs la possibilité de revenir à leur corpus des années après la finalisation de la recherche.

RO (BDIC) : à la BDIC la pratique revient d'utiliser une indexation Rameau. Depuis deux semaines la BDIC a pris la décision de réaliser des chapitrages chronologiques dans la description de Calames pour faciliter l'accès aux documents. La BDIC a l'avantage d'être une institution plus petite que les AN ou l'INA, ce qui lui permet d'avoir une autonomie propre et, donc, de prendre l'initiative et d'avancer plus vite sur de nombreux sujets et dossiers, y compris des questions d'indexation.

CK (BNF) : remarque le fait que Rameau est en train de changer, de bouger grâce aux partenariats (avec un projet sur les instruments de musique). Du coup, le langage issu de ce partenariat enrichit la liste d'autorité de Rameau. Cela est un exemple qui montre que des projets phares peuvent contribuer à faire bouger les normes.

SW : soulève la question compliquée de l'indexation de types de documents qui ne sont pas des livres et des revues, mais des enquêtes, des travaux de recherche, de la littérature grise (le web pose aussi cette question, car il réunit une variété de type d'écriture : texte, vidéo, images, audio), de l'audio-vidéo. Faut-il prendre des références de la documentation, des bibliothèques ou des archives ? Dans le cas d'une sélection sur des grands thèmes (Front populaire/2GM) comment faire pour interroger la diversité des sources ?

CF (BNF) : l'accès doit se faire par des « points d'accès » et des référentiels. Mais c'est un processus en train de se faire, ce n'est pas encore fait.

SW : « Il se trouve que le monde de la recherche, le monde la culture, les bibliothèques, les archives ont des histoires compliquées, au moment du numérique on est confronté à cet héritage et en même temps à des possibilités et à l'effort de travailler en commun pour surmonter et rendre un certain nombre de choses plus « fluides ». Il faut essayer de ne pas reproduire dans l'avenir ce type de cloisonnement. Les préoccupations de la recherche dans le

domaine des SHS doivent exister dans tous les dispositifs de l'accès au numérique pour que les étudiants, les citoyens puissent accéder à ces contenus. Cela peut être très compliqué, car les moteurs de recherche « grand public » masquent l'accès aux contenus. D'où les efforts de la TGIR, du CNRS, des MSH pour avoir un certain nombre de portées d'entrée. C'est pourquoi des MSH se mettent parfois à créer leur propre outil (comme les propres chaînes YouTube par exemple) ».

Intervention de Bernard Ganne au sujet du projet « Canson » et de la collaboration avec l'INA pour la sauvegarde et la valorisation du fonds.

Alain Bertho : Conclut sur une proposition pour le prochain séminaire : se focaliser sur l'indexation collective, sur l'indexation d'usage, de la façon dont s'est mis en œuvre par les chercheurs et de la façon dont on peut l'intégrer comme un type d'indexation soit à accès privé soit à accès public. Il y a des gens qui ont des pratiques sur cette question, comme Patrick Déshayes. Une date avant les grandes vacances. Pas les vendredis, ni les mercredis. Le mardi, c'est le jour préféré par les participants.

Il présente également la plateforme du chantier du Condorcet, appelée « Grand équipement documentaire » (?) qui a été pensée au départ non pas en lien direct avec le numérique audiovisuel, mais qui va être obligé de se mettre à jour.... Les représentants de l'INA ont été en contact avec ce GED. AB propose une réunion d'échange thématique sur la question de l'accès à la documentation audio-visuelle à partir des différentes institutions et des espaces qui seront localisés à Condorcet.

Serge Wolikow soutient cette suggestion et propose d'organiser cette rencontre dans le cadre du séminaire audiovisuel du consortium de la TGIR, car c'est une structure « neutre », ni une MSH, ni un labo.

SW : précision du calendrier futur : le consortium doit présenter son bilan en septembre 2016 à la TGIR. Mais en même temps, il s'agit d'une phase de lancement des nouveaux projets dans la perspective d'une nouvelle labellisation. Cela demande la préparation d'un site dans lequel les activités des séminaires av doivent avoir une place pleinement exploitée.

Il faut essayer de tracer sur 2 ans l'esquisse d'un programme raisonnable qui nous semble utile pour la communauté dans son ensemble. Il faut faire vivre cela dans le cadre d'un partenariat élargi : mettre en place une convention avec l'INA. On peut imaginer que d'autres institutions puissent aussi se rapprocher de notre travail, des bibliothèques, ou autres établissements intéressés.