Le portail des archives sonores culturelles

Journées d'étude sur la valorisation des archives sonores :
le documentaliste, le juriste et le chercheur
Maison méditerranéenne des sciences de l'homme
Aix-en-Provence, 22 novembre 2005

Michel Fingerhut

Introduction

L'État, par l'intermédiaire des actions menées par le Ministère de la culture et de la communication avait inscrit, depuis 1999, dans le cadre du Plan national de numérisation les archives sonores dans ses appels à projets mettant ainsi les technologies de l'information et de la communication au service des archives sonores. Près de 3000 heures de documents sonores ont ainsi été numérisés depuis 1999 par l'association Les Musiques de la Boulangère dirigée par le compositeur Nicolas Frize.

Dans ce cadre, la Mission de la recherche et de la technologie (MRT) avait missionné l'Ircam pour la réalisation d'un dispositif de mise en ligne et de valorisation de ces fonds, déposés auprès de quatre organismes :

- Le Centre historique des archives nationales (CHAN), pour des fonds de témoignages oraux sur la déportation.
- La Maison des civilisations de l'Europe et de la Méditerranée[1].
- La Maison méditerranéenne des sciences de l'homme (MMSH).
- Le Museum national d'histoire naturelle.

Cette communication dresse le bilan de ce chantier qui a compris quatre axes principaux[2] :

- L'aspect juridique. La numérisation et la mise en ligne de contenus sonores sont sujettes à la loi sur la propriété intellectuelle (droits de reproduction, droits de représentation, droits voisins). La nature des fonds concernés par ce projet et le cadre organisationnel dans lequel ils ont été collectés étant particulièrement complexes, une « approche découplée » a été élaborée pour la gestion des droits sur le plan contractuel et s'est manifestée dans la réalisation du dispositif technique. Ce point étant abordé par Ludovic Le Draoullec dans le cadre de sa communication, nous ne nous y arrêteront pas ici.
- L'aspect technologique. Dans l'optique de permettre à chaque organisme de garder son autonomie totale par rapport à la gestion de ses contenus et, en au

[1] Ex Musée des arts et traditions populaires.
[2] Auxquels ont participé, à diverses étapes, Ludovic Gaillard, Samuel Goldszmidt, Ludovic Le Draoullec et Xavier Sirven.

premier chef, de les utiliser en interne pour ses besoins scientifiques, une approche découplée a été mise en place pour la réalisation du dispositif :

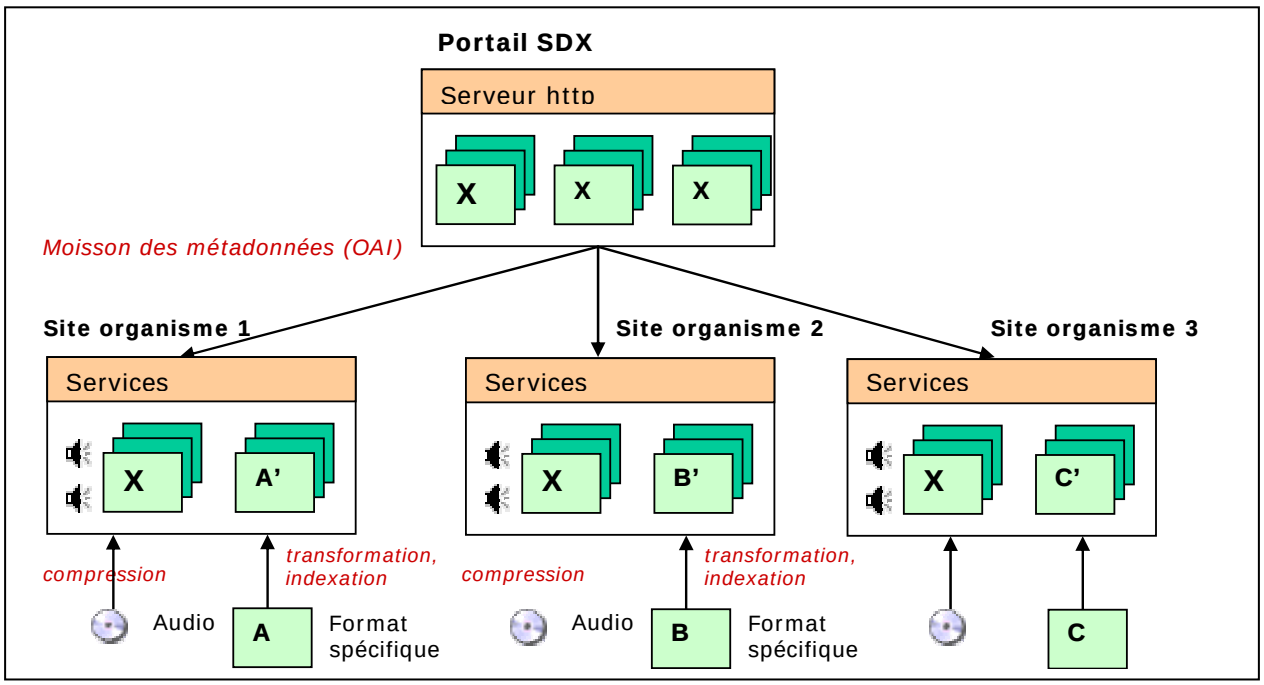
  - D'une part : quatre sites indépendants ont été réalisés pour héberger les fonds de chacun des quatre organismes, sites qui ont vocation à être intégrés dans les dispositifs existants ou futurs de ces organismes. Chacun de ces sites est directement consultable sur l'internet et fournit l'accès aux contenus que l'organisme aura décidé d'autoriser à l'écoute.
  - D'autre part, un « portail agrégatif » recueille (tel un catalogue commun) toutes les métadonnées des sites indépendants, offrant ainsi un point d'accès commun (qui se rajoute aux accès directs).

- L'aspect documentaire. Une attention particulière a été prêtée à l'organisation des contenus, qui s'est manifestée par :
  - L'organisation inter-documentaire : la réalisation d'un modèle commun de métadonnées qui puisse décrire de façon uniforme, notamment dans le portail, les fonds des organismes en question et qui permette d'y effectuer des recherches transversales et d'y « naviguer ».
  - La navigation intra-documentaire : la réalisation d'un outil de navigation *dans* des enregistrements sonores d'une longue durée : les fonds du CHAN, contrairement aux trois autres, contiennent des documents d'une durée pouvant excéder une heure et pour lesquels il faut donc fournir des moyens de navigation internes, à l'instar de la navigation dans les fonds de documents.
- La valorisation. Dans le cadre du projet, plusieurs réalisations mettant en valeur une partie d'un fonds ou d'un autre ont été créées et mise en ligne dans le cadre de ce dispositif.

Le dispositif technique

La conception et le développement du dispositif ont été effectués en deux étapes. Une première modélisation consistant en un site avec quelque 600 enregistrements de trois des quatre organismes, réalisée (en moins de trois semaines) à l'occasion de la Première Fête de l'internet scientifique, qui s'est tenue du 29 mars au 2 avril 2004, a permis de valider certaines hypothèses techniques et de réaliser un schéma préliminaire du modèle commun des métadonnées. En particulier :

- Organisation hiérarchique des métadonnées :
  Organisme → Fonds → Collection → Sous-collection → Item
  les niveaux intermédiaires étant facultatifs, et leur appellation modifiable.



- Utilisation de standards (en l'occurrence : Dublin Core, vCard, cld, XML) pour la structuration et le codage des métadonnées.
- Utilisation de SDX[3] pour la gestion des métadonnées.

L'étape suivante a permis d'aborder la question du dispositif réparti[4], dans lequel les métadonnées (et les données associées) sont stockées dans les sites des organismes que le portail va interroger périodiquement pour en aspirer les métadonnées mises au format commun et offrir ainsi un accès transversal à tous les fonds présents dans ce dispositif. Une interrogation effectuée dans le portail renverra les métadonnées correspondantes. Dans le cas où des données (en l'occurrence : des enregistrements sonores) sont attachées à ces métadonnées, l'accès se fait par l'entremise du site de l'organisme qui les diffuse et sous son contrôle.

Le portail « aspire » les métadonnées de chacun des sites en utilisant le protocole de collecte de métadonnées OAI[5], qui est inclus dans SDX. Contrairement à la façon de procéder avec le protocole Z39.50[6], la récolte des métadonnées par le portail ne s'effectue pas au moment où une interrogation est lancée sur le serveur, mais de façon

[3] Logiciel libre comprenant un moteur de recherche et un environnement de publication libre pour documents en XML, développé à l'instigation de la MRT. Cf. <http://adnx.org/sdx/>.
[4] Disponible à l'adresse <http://www.archison-culture.fr.eu.org/>.
[5] Plus précisément OAI-PMH (*Open Archive Initiative Protocol for Metadata Harvesting*), cf. <http://www.openarchives.org/>. Pour lire une introduction claire (en français) de ce protocole, cf. <http://www.bnf.fr/PAGES/infopro/journeespro/pdf/AFNOR2005/OAI.pdf>.
[6] Cf. l'article de Martin Sévigny (l'auteur du logiciel SDX qui inclut OAI…) sur « La norme Z39.50 : un outil essentiel pour l'uniformisation de la recherche d'information », disponible à l'adresse <http://www.ebsi.umontreal.ca/cursus/vol1no1/sevigny.html>.

asynchrone et programmée (par exemple : une fois par semaine). Les réponses que fournit le portail peuvent ainsi être quelque peu moins actuelles que celles fournies par chacun des sites, mais elles ont l'avantage d'être immédiates[7].

Cette organisation permet de rajouter à ce dispositif des organismes qui auraient déjà leur propre site de gestion de leurs contenus : il leur faut uniquement rajouter à leur site un point d'accès par OAI (bien plus facile à intégrer que Z39.50), qui fournisse, lorsqu'il est interrogé par le portail, les métadonnées au format commun, codées en XML.

La consultation des contenus se fait de l'une des deux façons suivantes :

- par l'entremise d'un lecteur audio (celui qui est installé sur le poste de la personne qui écoute), pour tous les enregistrements ; ceux-ci sont installés sur les sites des organismes, et disponibles par serveur de flux (*streaming*, en anglais) au format MP3 en général ;
- par l'entremise d'un *plugin* pour Flash, permettant de naviguer dans les enregistrements sonores parvenus du CHAN ; ceux-ci ont été fournis avec des métadonnées décrivant la structure de chacun d'eux au format EAD ; un des développements effectués dans le cadre du projet génère automatiquement une présentation permettant d'y naviguer par chapitrage ou par mots-clé, de voir le résumé des sections, etc.

Enfin, le dispositif assure la mise en œuvre de la politique d'accès que chaque organisme choisit de mettre en place pour ses propres contenus, *via* des contrôles explicites (ou implicites) qu'il spécifie dans les métadonnées correspondantes et qui déterminent trois périmètres concentriques : sont-ils consultables uniquement dans l'organisme qui les détient, ou dans les organismes partenaires du dispositif, ou partout sur l'internet ?

Le modèle des métadonnées

La conception de la structuration des métadonnées du dispositif a été entreprise en consultation avec les organismes, en prenant en compte d'une part les métadonnées qu'ils possédaient déjà[8] et, d'autre part, les préconisations de l'AFAS[9] et de Minerva[10].

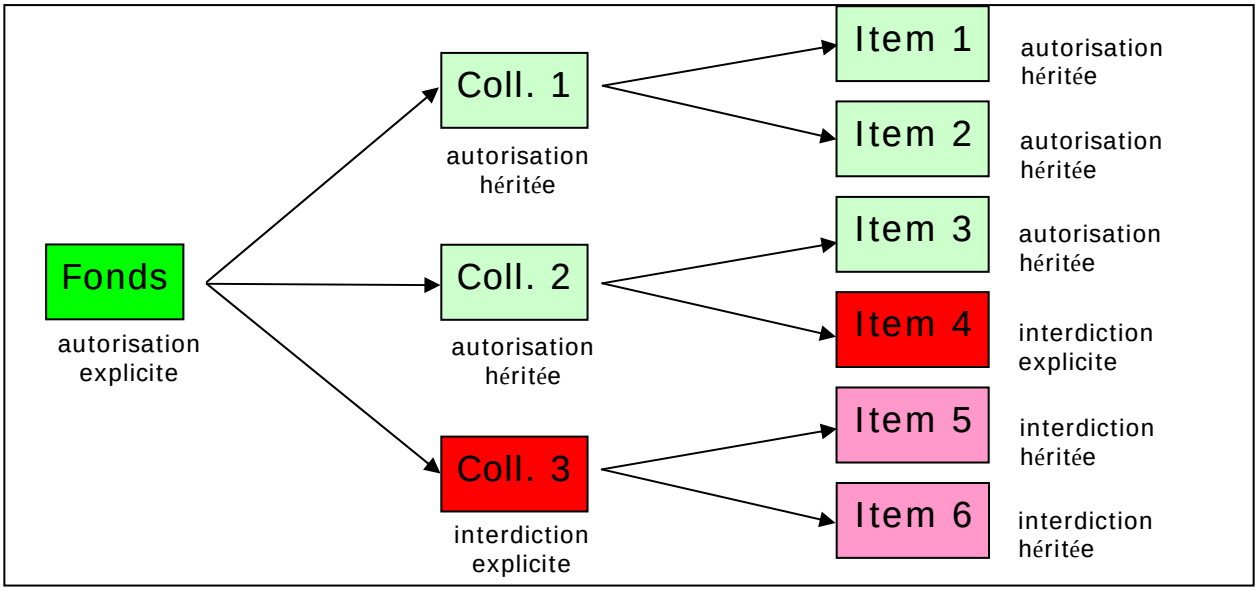
[7] Une interrogation par Z39.50 est communiquée à chacun des sites, le portail récoltant les réponses au fur et à mesure de leur arrivée. Certains sites peuvent ne pas répondre, d'autres le faire lentement, ce qui rend cette méthode parfois aléatoire.
[8] Structurées de façon différente, et codées dans des systèmes distincts (Word, FileMaker, 4D, EAD…).
[9] AFAS-FAMDT : *Guide d'analyse documentaire du son inédit pour la mise en place de banques de données*, 2001. Voir < http://afas.mmsh.univ-aix.fr/CATALOGAGE.htm>.
[10] Le réseau ministériel européen pour la valorisation des activités de numérisation, cf. <http://www.minervaeurope.org/>.

Étant communes à tous les organismes, elles comprennent essentiellement les principaux champs communs à tous, et sont donc souvent moins détaillées que celles de chaque organisme, auxquelles elles ne cherchent pas à se substituer, mais à faciliter la localisation.

Ce modèle est organisé en plusieurs classes :

- Les personnes physiques ou morales (leur rôle – interprète, collecteur… – est

Notice (Fonds, Collection, Sous Collection, Item)								
Attribut	**Propriété RDF**	**Equivalent Minerva**	**Lien**	**Définition**	**Ob**	**Op**	**Rep**	**REF**
Attributs généraux								
Identifiant	dc:identifier	Identifier		Numéro de classement attribué par l'organisme	X			
Niveau	(integer)			Niveau de la notice (Fonds=10, Collection=20, Sous Collection=30, Item=40)	X			
Titre	dc:title	Title		Le nom de l'objet	X			
Résumé	dc:description	Description		Une description de l'objet		X		
Langue	dc:language	Language		La (les) langue(s) utilisée(s) dans le document	X		X	
Type	dc:type	Digital Document Type		Le type de contenu (son inédit, vidéo,...)	X			
Consultation	cld:accessControl (sous-propriété de dc:rights)	Access Control		Conditions de consultation	X			
Date	dc:date			Date de création de l'objet		X		
Sujet								
Description	dc:subject							
Genre	*(string)*			Genre de l'objet (Chanson, conte, récit, poésie,...)		X		
Expression	*(string)*			Form et expression musicale (nb d'interprètes, instruments,...)		X		
Nature Enregistrement	*(string)*			Contexte de l'enregistrement		X		
Lieu Enregistrement	vcard:adr:pobox vcard:adr:street vcard:locality vcard:region vcard:pcode vcard:country	Address		Lieu de l'enregistrement		X		
Contexte Ethno	*(string)*			Contexte ethno/sociologique.		X		
Analyse	*(string)*			Description du contenu (Oeuvres, Personnes, Lieux cités)		X	X	
Couverture Temporel	dcq:temporel (sous-propriété de dc:coverage)	Spatial coverage		La couverture spatiale de l'objet (classement à 3 niveaux)		X	X	
Couverture Spatial	dcq:spatial (sous-propriété de dc:coverage)	Temporal coverage		La couverture temporelle (périodes ou dates) de l'objet		X	X	
Note	cld:note			Informations non saisies		X		
Agents associées								
Contributeur	dc:contributor		Personne	Personne morale ou physique	X		X	X
Role	*(référence)*			Le rôle de l'agent	X		X	X
Ayant Droit	cld:owner	Owner	Personne	Ayant droit de l'objet	X		X	X
Structuration								
Id Parent	dcq:isPartOf (sous-propriété de dc:relation)		Notice	Lien vers une notice parent (objet Super Collection de CLD)		X		X
Référence Bibliographique	cld:seeAlso			Lien vers une bibliographie		X	X	
Ressource Complémentaire	cld:hasAssociation (sous-propriété de dc:relation)			Lien vers une ressource complémentaire		X	X	X
Localisation Numérique	cld:hasLocation (sous-propriété de dc:relation)		Manifestation	Référence vers les ressources numériques correspondantes	X		X	
Saisie								
Documentaliste	dc:creator		Personne	Auteur de la fiche		X		X
Date saisie	dc:date			Date de création de la fiche		X		



indiqué dans les notices documentaires)..

- Les notices documentaires, décrivant autant les items (objets sonores individuels) que leurs regroupements (fonds, collection, sous collection), la seule différence étant manifestée par une indication de niveau ; chaque enregistrement dans cette classe appartient à un niveau plus élevé. Une fonctionnalité importante a été déterminée pour certains champs, celle de l'héritage : pour ceux-ci, il suffit de les renseigner à un niveau pour qu'ils se « propagent » au niveau sous-jacent s'ils n'y sont pas mentionnés. Ainsi, on peut renseigner le collecteur de tous les items d'une collection dans la notice de collection, et cette information s'appliquera à tous les items de la collection, sauf mention spécifique (c'est-à-dire sauf si on renseigne explicitement le contributeur pour un ou plusieurs items particuliers).
- Les supports, décrivant les médias physiques d'origine (souvent analogiques) et les éventuels médias dérivés (tels le CD audio et le CD-Rom produits dans le cadre du plan de numérisation).
- Les ressources numériques, décrivant les contenus numérisés ou dérivés (présentations multimédia de valorisation).
- Un ou plusieurs thésaurus instrumentaux, reflétant ceux utilisés par les organismes.

Les métadonnées sont codées au format XML et utilisent la norme Dublin Core[11], augmentée de standards tels que vCard[12] ou cld[13].

[11] Voir <http://www.dublincore.org/>.
[12] « Carte virtuelle », norme de description d'informations personnelles.
[13] Norme de description de niveaux de collections (***c**ollection **l**evel **d**escription*). Voir, par exemple, <http://www.en.eun.org/eun.org2/eun/en/Celebrate_LearningObjects/content.cfm?lang=fr&ov=3829>.

La définition des restrictions d'accès aux métadonnées est située dans un champ présent dans chacune d'elles, et prend quatre valeurs possibles :

- 0 – visible nulle part
- 10 – visible dans l'intranet de l'organisme
- 20 – visible sur le réseau de partenaires
- 30 – visible sur tout l'internet.

Il n'est pas nécessaire de spécifier explicitement les droits pour *chaque* métadonnée : les droits spécifiés à un niveau (par exemple : une collection) passent automatiquement, selon le principe d'héritage, au niveau inférieur (par exemple : à tous les items qui en font partie), sauf mention explicite du contraire, qui ne peut être que plus restrictive ; ainsi, si une collection est autorisée à la consultation uniquement dans le réseau de partenaires, tous ses items le seront aussi ; on ne pourra en autoriser la consultation hors de ce réseau, mais par contre, on pourra en limiter l'accès encore plus, de façon à ce qu'ils ne soient accessibles que dans l'enceinte de l'organisme dépositaire.

Ces restrictions permettent, par exemple, de ne pas donner accès à des informations personnelles (noms de contributeurs, par exemple).

La valorisation

Plusieurs présentations, disponibles sur les sites, mettent en valeur certaines collections. Ayant été réalisées à des périodes différentes, elles utilisent des technologies distinctes.

Les perspectives

Le projet est actuellement suspendu en attente de financements qui permettraient d'en poursuivre le développement dans les axes suivants :

- La propagation des droits d'accès aux données : actuellement, le dispositif de contrôle d'accès ne concerne que les métadonnées, et ne s'applique pas automatiquement aux données (en l'état actuel, ce contrôle est rajouté manuellement dans le serveur audio).
- La durée d'exploitation : dans l'éventualité où les contrats limiteraient la mise à disposition dans le temps, le dispositif devrait pouvoir en empêcher l'accès une fois cette période échue.
- L'arrière-boutique (*back-office*) : il s'agit des outils de gestion des sites (destinés à l'import des métadonnées et des données et à leur mise en correspondance[14], à la spécification des droits d'accès, au rajout des ressources externes - produits de la

[14] Problème particulièrement épineux : les données ne sont pas fournies avec les métadonnées, et la correspondance entre elles est souvent impossible à établir entièrement automatiquement.

valorisation –, etc.) et du portail (configurations diverses, dont celle du réseau de partenaire).

- Le rajout de partenaires, qui nécessite certains développements communs (le rajout de l'accès OAI à leur site ou l'installation d'un site générique et le développement d'outils d'import correspondants).

L'architecture de « portail agrégatif » – et l'utilisation de SDX pour le réaliser – ont été depuis réutilisées pour la réalisation (en cours) d'un point d'accès commun à toutes les bases documentaires (et événementielles) de l'Ircam, et a vocation à s'ouvrir à d'autres organismes.