Data Management at the French National Audiovisual Institute (Ina)

Eleonore ALQUIER
Project manager, Collections division, Ina
e-mail: ealquier@ina.fr

Data Management at the French National Audiovisual Institute (INA)

ABSTRACT

The French National Audiovisual Institute has been responsible since 1974 for the preservation of the audiovisual heritage produced by national broadcasting corporation (or "Office de radio et television française": ORTF, for French radio and television corporation). The massive digitalization of these collections in the 1990s, the native digital capture of 120 channels since 2001, the opening of a "general public" website in 2006, are some of the steps taken by the Institute to progressively take into account the digital technologies to benefit the audiovisual preservation. This proposal of presentation would provide an update on the evolution of our processing, concerning most specifically a multi-year project which aims, linked to a new big data policy, to harmonize descriptive metadata according to common thesaurus and to streamline production processes as well as to promote new uses of these contents within the Institute (partial automation of documentary processing by automatic detecting of quoted or represented entities (faces, names, …), automatic articulation of documentary and legal metadata, …), but also outside of the Institute (online access to open data, access to media by technical data mining, …).

Key words: The French National Audiovisual Institute, digitalization, processing, descriptive metadata, technical data mining

Il trattamento dati all'Istituto audiovisivo nazionale francese (INA)

SINTESI

L'Istituto audiovisivo nazionale francese (INA) è responsabile dal 1974 della conservazione del patrimonio audiovisivo prodotto dalla società televisiva nazionale (o "Ufficio della radio e televisione francese", ORTF). La digitalizzazione massiva di queste collezioni dagli anni '90, la registrazione in nativo digitale di 120 canali dal 2001, l'apertura di un sito web nel 2006 sono alcuni dei passaggi compiuti dall'Istituto per prendere progressivamente in considerazione le tecnologie digitali in modo da favorire la conservazione audiovisiva. Questa proposta di presentazione vorrebbe fornire un aggiornamento sull'evoluzione dei nostri procedimenti, concernenti più specificatamente un progetto pluriennale che ha lo scopo, in collegamento con un nuova grande politica dei dati, di armonizzare i metadati descrittivi con un comune thesaurus e di ottimizzare i processi produttivi così come di promuovere nuovi utilizzi di questi contesti all'interno dell'Istituto (parziale automazione dei processi documentali tramite un rilevamento automatico delle entità citate o rappresentate (facce, nomi…), un'articolazione automatica dei metadata documentali e legali, anche al di fuori dell'Istituto (accesso online agli open data, accesso ai media tramite estrazione tecnica dei dati…).

Parole chiave: The French National Audiovisual Institute, digitalizzazione, elaborazione, metadata descrittivi, estrazione di dati tecnici

Upravljanje s podatki na Francoskem nacionalnem avdiovizualnem inštitutu (INA)

IZVLEČEK

Francoski Nacionalni avdiovizualni inštitut je od leta 1974 odgovoren za ohranjanje avdiovizualne dediščine, ki nastaja na nacionalni radio televiziji (Office de radio et television française": ORTF – Francoska radio televizija). Množična digitalizacija teh zbirk v devetdesetih letih prejšnjega stoletja, digitalni zajem lastnih 120 kanalov od leta 2001, odprtje "jave" spletne strani v letu 2006, so le nekateri od ukrepov, ki jih je inštitut izvedel in pri tem progresivno upošteval digitalne tehnologije, ki lahko prispevajo k ohranjanju avdiovizualnega gradiva. Prispevek govori o razvoju obdelave, ki se, v povezavi z novo politiko velikih baz podatkov, nanaša posebej na

večletni projekt, katerega cilj je uskladitev popisnih metapodatkov v skladu s skupnim tezavrom in racionalizacija produkcijskih procesov, kot tudi za spodbujanje novih načinov uporabe teh vsebin v okviru inštituta (delna avtomatizacija dokumentacijske obdelave z avtomatskim prepoznavanjem citiranih ali zastopanih entitet (obrazi, imena, ...), avtomatska artikulacija dokumentacijskih in pravnih metapodatkov, ...), in tudi zunaj zavoda (spletni dostop do javnih podatkov, dostop do medijev s pomočjo tehničnega rudarjenja podatkov, ...).

Ključne besede: The French National Audiovisual Institute, digitalizacija, obdelava, popisni metapodatki, tehnično rudarjenje podatkov

1 The French National Audiovisual Institute: one establishment, many assignments…

a. … archives...

As a public establishment created when the Office de radiodiffusion-télévision française (ORTF - French Radio and Television Broadcasting Authority) was wound up in 1974, the Institut national de l'audiovisuel (Ina - French National Audiovisual Institute) was set up to collect and conserve French audiovisual resources, initially seen basically for production purposes of national broadcasters. However, this perimeter of action was gradually extended as the French audiovisual landscape changed: former public-service channels were privatised, an increasing number of private channels appeared and cable and satellite coverage grew, offering consumers an alternative to the content of the 'traditional' (or terrestrial) channels. This led Ina to develop its service offer as a third-party archivist of audiovisual resources, whatever their origin or status. The original statutes of the Ina were compatible with this new positioning: as a 'public establishment of an industrial and commercial nature', it could offer a catalogue of services to private as well as public partners.

Aside from its basically archive-based tasks, Ina was also positioned as a training, research and creative production organisation when it was founded - however, such operations will be not be explored here.

On the basis of this first duty of statutory collection (laws of 1974, 1982, 1986 and 2000), concentrating on what are termed 'professional' audiovisual archives - in other words, content produced by public television channels and radio stations - Ina has acquired an initial collection of more than 1.5 million hours, including a million digitised hours that can be accessed online via a site devoted to professional production services. This collection continues to grow year on year as Ina fulfils the terms of contracts it has signed with audiovisual partners (public and also private radio stations and television channels) or organisations in other sectors (cultural establishments, private companies, etc.). The types of service associated with these resources not only require Ina's archiving skills, but also legal analysis of intellectual property rights, so as to enable marketing and royalty payments.

However, another source of content and new archiving regulations that appeared at the start of the 1990s radically changed Ina's positioning. Adding to the legal deposit requirement applied from the 16th century on and related to different forms of printed material (monographic works, periodicals, posters and later music scores, commercial sound and video media, etc.), the legislature extended it to radio and television broadcasts, treated as forms of publication. All programmes produced or co-produced by French national broadcasters were now added to Ina's archives under the terms of the legal deposit rules for French radio and television, i.e. terrestrial radio stations and television channels, from 1995 on (law of the 20th June 1992). Initially, collection was based on physical recordings by broadcasters on media storing programmes covered by legal-deposit regulations, but the system changed in 2001 when Ina began to record direct digital streams from stations and channels with a legal deposit obligation. This change in positioning ended Ina's dependency on broadcasters in terms of content supply, but also changed the very principle of collection for deposit - from selection at source based on statutory criteria to the comprehensive recording of broadcast content 24 hours a day. The perimeter of collection was gradually extended beyond 'traditional' broadcasting in 1995 and today it takes in 100 television channels and 20 (soon 64) radio stations that are recorded continuously, generating 1 million hours of new content every year.

In 2006 (with the DAVDSI law of the 1st August 2006), Ina's collection perimeter was enlarged for the last time when the legal deposit requirement was extended to media websites. Ina and Bibliothèque nationale de France (BnF - National Library of France) share responsibility for this collection. Consequently, Ina manages the collection and conservation of the content of all broadcaster websites (9,000 in 2013), web radio and web TV, as well as content published by French audiovisual operators and sites dedicated to their content (programmes, series, fan sites, etc.). Today, collection also covers social media, with the regular monitoring of YouTube and DailyMotion accounts, along with Twitter hashtags.

b. … and consumer populations

To distribute content whose context and statutory collection system may have evolved since 1974, Ina has developed different systems of distribution suited to the diversity of the target populations.

Archives collected under a fixed agreement or contract based on commercial objectives negotiated by Ina with the initial intellectual property rights holder are made available to a professional population, which requires a high level of content quality and accessibility. Today, that accessibility is guaranteed by a dedicated online access interface known as 'InaMediapro' (https://www.inamediapro.com/), dedicated to authorised users. This online previewing, selection and purchase platform enables content searches using metadata entered by the establishment's archivists indexing team, and provides online low-definition content viewing. Viewing is itself enabled by Ina's 'digital preservation plan', which has been implemented for the last 15 years to conserve historical archives by digitising storage that is oldest, most vulnerable to obsolescence and most fragile, and facilitating access to it, like any digitisation policy.

Since 2006, aside from this 'professional' distribution based on a legal and commercial system of rights analysis and the invoicing of content sold, Ina has developed a 'general public' access interface called 'ina.fr' (http://www.ina.fr/), whose purpose is to promote awareness among 'uninformed' Internet users of the diversity of its collected content, particularly in historical archives. Several tens of thousands of hours of content are distributed in this way (i.e. a small proportion of the collected archives). Selection is based on two criteria. Firstly, the aim is to identify content in the public domain, which can therefore be freely distributed online without breaching intellectual property rights. Secondly, the free distribution of extracts or full content is governed by an editorial selection policy. Content distributed in this way must be subject to an optimisation strategy related to a news item (death of a personality, anniversary of a historic event, etc.) and / or an approach that will increase the dissemination of sets and collections rather than an individual resource. In this way, the Internet user will discover a 'rebound' effect, taking them from one piece of content to another and facilitating their perception of the audiovisual archive as a component of wider groups or even collections - a prime example is collections of television advertisements (http://www.ina.fr/pages-carrefours/publicite/).

The legal deposit collection system most recently deployed by Ina has at last enabled a third type of consumer to access audiovisual archives: an 'academic' or 'researcher' population. Since legal deposit status requires access to be limited to the premises of the depositary establishment, in 1995, Ina opened a research reception area at the Bibliothèque nationale de France (National Library of France). Aside from access to collected content on VHS, DVDs and finally servers (according to the level of technology), access to the collections is based on the availability of descriptive media metadata produced by Ina's teams, which enable three search modes:

- browsing and advanced searches in databases using customisable search tables that can be adapted according to the researcher's needs;
- the preparation of sets of metadata than can then be exported from databases, with the option of adding new levels of indexing to the sets so produced, in such a way as to provide new angles of inquiry related to the researcher's spectrum of studies;
- the indexing and sequencing of media content in parallel with viewing, using viewing tools that also enable the capture of thumbnails that are themselves timecoded and indexed for both quantitative and qualitative research.

These 'expert' consulting modes take into account the ban on removing viewed content from the

premises of the establishment. However, the metadata themselves - at least in terms of primary cataloguing data - are accessible on inquiry and for display online, enabling a first stage of research to be carried out and a remote working programme to be established.

Finally, new methods of 'intermediate' consultation (midway between general public use and specialist research) are being explored, especially using a network of dedicated terminals deployed on the premises of partner organisations (film, media and university libraries, etc.). These access terminals are remotely controlled by Ina and only enable 'passive' viewing of archived content, but they provide an additional way of accessing audiovisual archives, also using search tables that can be adapted according to the level of detail of the research conducted.

2 A sophisticated but ageing information system

The variety of media collected by Ina (the oldest content dates back to the appearance of radio broadcasts in the 1930s) and the many ways in which they can be used have naturally led to the production of metadata that are adapted to both the content and the methods of access to it.

a. History of Ina's documentary databases

Originally developed to handle two different types of archive - 'professional' and 'legal deposit' - over the last 20 years, Ina's collections have naturally been managed by two parallel systems designed to operate independently. Each one plays a specific role: for the first, the distribution of a relatively limited volume of content (on the scale of the establishment) for the commercial and professional purpose of selling individual extracts; for the second, the management of an ever-growing volume of content for uses that notably enable long-term research into trends applied to thousands or even tens of thousands of data sets. However, links have gradually been established between these two systems of collection and two types of use, if only because the content stored in these two different archives could fall into both categories - for instance, public television news, covered by both legal deposit and the terms of the service agreement signed between Ina and national public broadcasters.

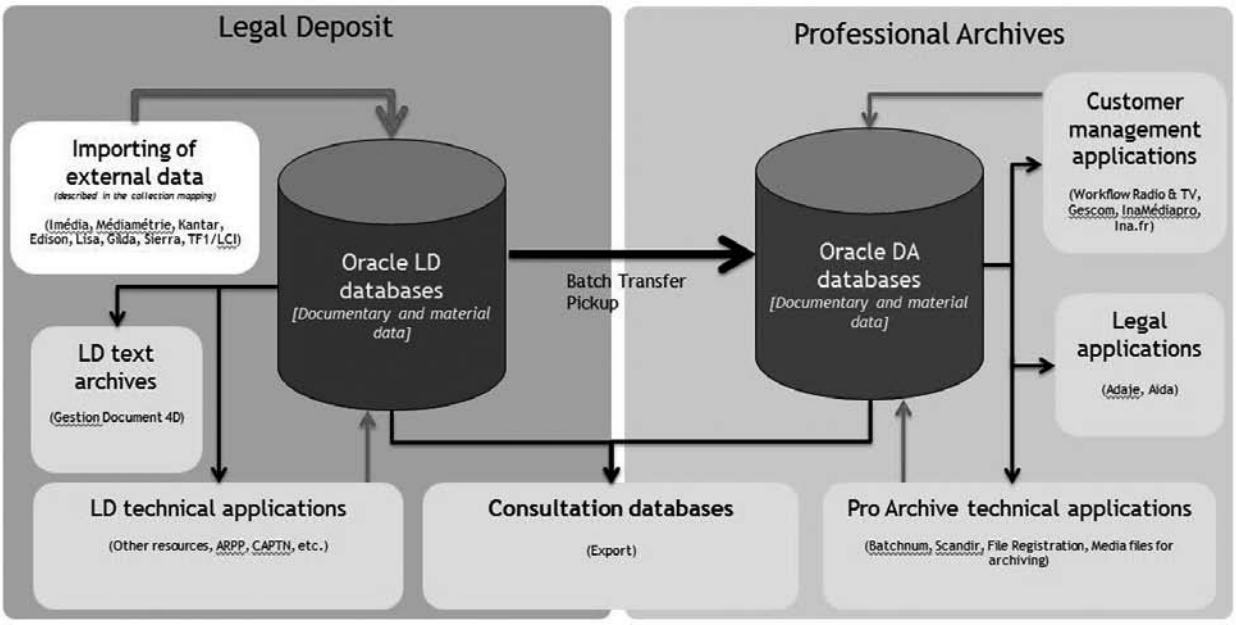

Figure 1 - Architecture of Ina documentary systems: co-existence of two metadata production environments

So the silo rationale that led to the coexistence of two distinct information systems tends now to play a diminishing role. Firstly because of the observed redundancy of some of the data present in both

databases; secondly, because developments in data processing - in terms of both technological possibilities and the regulatory framework - have understandably led to efforts to counter silo phenomena and prioritise interaction between data sets with shared characteristics.

b. From the reconciliation to the merging of related databases: big data at INA

Based on the observation that different databases existed within the organisation to supply different services - which, while distinct, employed information that was fundamentally similar in nature (identification and description of audiovisual content according to criteria such as programme genre, names listed in the credits, year of production, etc.) - an initial rationalisation initiative implemented over the last few years has consisted of designing a documentary processing chain in which the different silos align themselves with each other. Documentary data - rarely created from scratch, but usually imported from a third-party source - enters the system via a single, systematic gateway to one of the two databases (in this case, the one dedicated to mass processing since it was set up). After a first stage of processing (alignment of the imported terms with internal references, cataloguing control and even detailed documentary processing with indexing and the production of a summary), the data is then transferred to a second environment. This one is devoted to professional distribution services, which may require specific editorial processing before publication - an editorial processing that is nevertheless facilitated by the first documentary stages completed when the data entered the system. This processing pattern (shown in figure 1) does not prevent data redundancy between the databases (since each one is connected with its own processing applications and dedicated consultation interfaces) or the risk of divergent changes in an originally shared piece of information since the two databases are not systematically co-synchronised. Even so, this model has the advantage of establishing communication between the databases and throwing light on the shared information they contain, a necessary preparation and even source of awareness prior to the redesign of these applications in a single shared environment.

That is indeed the objective of Ina's big data project, planned in the form of a single data-storage structure known as the 'data lake'. In fact, this 'lake' is made up of different types of database that can be more or less structured according to the nature of the data they contain, but are brought together in a single system, so enabling the reconciliation of data from different production environments. Indeed, aside from the replacement of the dual documentary processing environment (described above) by this new primary hosting infrastructure, the aim is also to recover (as a secondary copy this time) data generated by the other business sectors of the establishment, particularly those related to the legal analysis of audiovisual content and its commercial use.

This adoption of a big data policy co-implemented by a data-storage infrastructure (built and maintained by the IT divisions) and the redesign of the systems producing the data (led by the operational divisions), should lead to the development of new services and new forms of use for the data - without, of course, undermining the collection processes that ensure their quality, whether they come from outside sources or internal activity.

Given the volume of audiovisual data collected (mainly for legal deposit), the specific nature of Ina's metadata indeed relies on the very many imports that generate them and which can come from suppliers such as press agencies supplying television programme guides or public partners working with Ina because of its statutory role in the conservation of national audiovisual heritage. After various stages of processing to ensure compliance with Ina's prevailing data models, these imports are stored in the databases. Depending on the import, the data can be left untouched or enhanced by internal documentation processing. With a total of around a million documentary files produced each year, there can be no question of allowing a change in database infrastructure to undermine this economy of controlled and reformatted imports, sometimes (but more rarely) internally modified by hand. That is why Ina's big data project faces the dual challenge of being based on a new model of documentary data that complies with standards (FRBR) but is ultimately specific given the uniqueness of Ina's duties in the heritage domain, and enables the merging of uses that are currently embodied by two distinct production systems; while at the same time ensuring that no changes are required in the data-import principles that have formed the foundation of database input policy for the last twenty years.

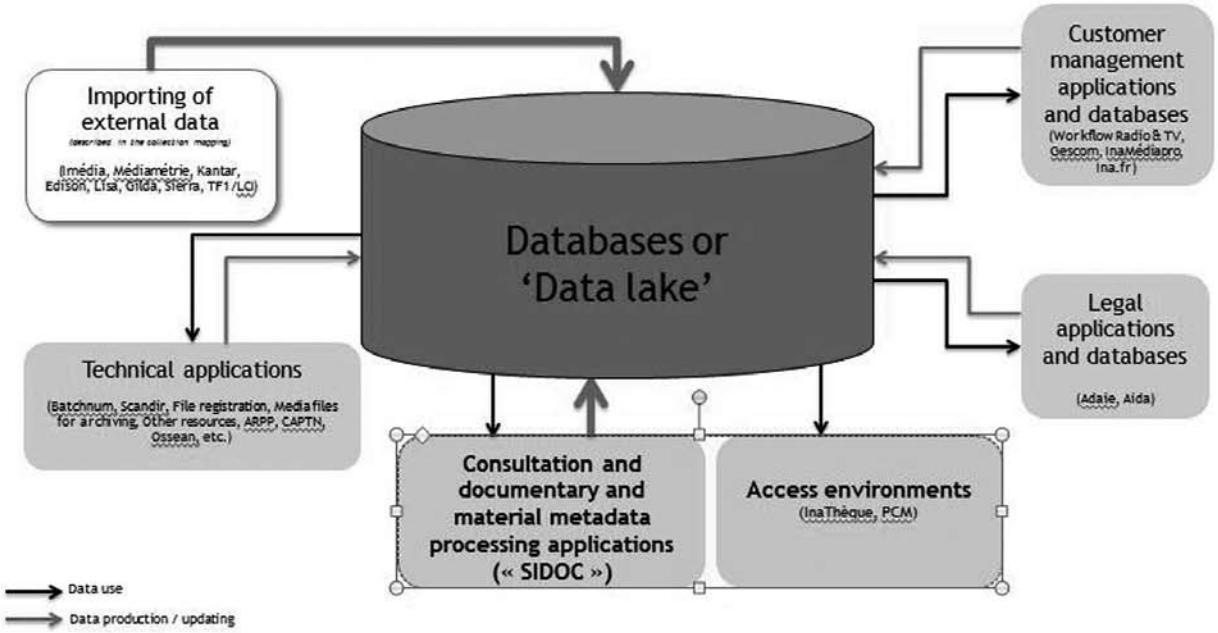

Figure 2 -Ina data lake: eliminating silos without undermining the data-processing rationale

All these challenges must be met over the next five years, also taking into account technological and regulatory developments in progress and yet to come.

3 Assisting and facilitating new services

a. Open data: constraint or opportunity?

Changes in the French regulatory framework in 2015 and 2016 have especially tended to generalise open data policies, which public bodies (public establishments, ministries, etc.) were formerly only encouraged to apply, but that have now become statutory obligations. As a public establishment, Ina is subject to these obligations and will have to implement them over the coming years, at least in relation to metadata whose recognised producer (in the archival sense of the term) and intellectual-property rights holder it is. This concerns only the opening of metadata: the data themselves, i.e. the media archived by the establishment, are protected by intellectual property laws and can't be object of an open data policy. Given the import policies presented above, compliance with these new requirements implies the preliminary detailed mapping in our databases, so as to distinguish (for each data set) between the share of data totally sourced from an outside supplier and the share produced by Ina, either by significantly modifying the initial data or by producing information from scratch. Those are the principles governing the documentary processing of television news programmes with one benchmark edition per day and per channel, managed 100% by Ina's archivists.

Beyond the obligations now framed by law, the implementation of an open-data policy can be part of a voluntary approach, aiming to promote awareness of our data and especially its potential secondary uses, since that is indeed the initial objective of open data (public or not). Opening up our metadata - not just for viewing, but also for reuse - could extend our consumer population base from 'academic' researchers specialising in audiovisual issues to a larger cohort who, by adopting our data, will approach their information from the angle or spectrum of other entry data - for instance, information from library catalogues or archive inventories. So open data will take on all its meaning when based on the principle of linked data, which, aside from increased data access, facilitates the correlation of data from separate bodies (by the use of shared references, for example). Once Ina's internal silos have been eliminated, open data could link audiovisual adaptations of *Les Misérables* (the 19th century novel by Victor Hugo) described in our databases to the original manuscript kept at the Bibliothèque nationale de France (French National Library).

Conversely, these connections to be constructed between databases when they are opened up to each other mean that we will be able to consider new ways of automatically acquiring strictly 'audiovisual' information produced or at least managed under the responsibility of Ina. So acquiring autobiographical entries from Wikipedia will add to our benchmark catalogue of persons, subject to the prior establishment of the appropriate connections between databases, and as long as there is a minimum level of interoperability between our descriptive models.

b. From the evolution of tools to the evolution of occupations

Prefigured by open data and systematic approaches to the correlation of information with external sources, progress in data use ultimately raises the question of the development of the production, management and verification of the data themselves. While Ina has already been studying these questions for nearly 10 years in relation to its growing policy of data imports supplied by third parties, the growing possibilities presented not only by the semantic web practices mentioned above, but also automatic content analysis technology, logically lead us to anticipate changes in information management occupations.

Today, media interpretation and analysis tools provide a glimpse of many approaches to the production of new metadata, such as automatic transcriptions that could be supplied by appropriate tools that are 'educated' according to the nature of the resources to be processed. In these days of full text and Google searches, providing the public with access to these complete 'real-size' metadata directly sourced from the media would certainly lead to 'documentary noise' problems but, here too, technology could reduce their impact - for instance, by limiting the terms of the requested transcription to a list of 'distinguishing' terms or items that already exist in toolkits.

As European authorities are studying the question of authorising text and data mining or not, automatic recognition process for still images (logos, pictorial works, buildings, etc.), moving images (faces) and voices, which use appropriate toolkits based on dictionaries of images and sounds, have already allowed us to evaluate the suitability of these tools for conducting research or surveys. Ultimately, we may be able to anticipate using these solutions to launch the documentary processing of an archive that available human resources have not been able to analyse and index. All those methods of use of identified resources that are of an experimental nature today and may be based on industrialised principles tomorrow can be considered in order to facilitate access to sources that we know to be growing more and more rapidly.

Examples such as the recent appearance of filmed radio, which breaks down the wall between radio, audiovisual and the Internet, or the growing number of web documentaries that are using the Internet as a means to exclusively distribute audiovisual content that was formerly shown on television, can only lead us to consider the nature of the medium and its future. Since the usual compartments between media are tending to disappear, it also seems reasonable to take a fresh look at the role of documentary information professionals. They are soon to be or are already working with semi-automated processes that are essential in the production of a streamed description of mass content, but they still play an essential role. No longer necessarily as the originator of data, but as the controller of the quality and coherency of the information they have validated and made accessible to users. In the same way, consultation interfaces are also being studied, in order to weigh aims of popularisation (facilitating access to complex content for the greatest number of people) against a (semi-)professional goal (to improve the interoperability of our data to enable connection with external databases). These practices do not, however, involve contradictions, but rather complementarities: whatever the type of use or consumer, information professionals always have a duty to pass on, guarantee and manage the information they conserve and to which they provide access.

SUMMARY

The French National Audiovisual Institute has been responsible since 1974 for the preservation of the audiovisual heritage produced by national broadcasting corporation (or "Office de radio et television française": ORTF, for French radio and television corporation). The massive digitalization of these collections in the 1990s, the opening of a "general public" website in 2006, are some of the steps taken by the Institute to progressively take

into account the digital technologies to benefit the audiovisual preservation. In parallel to these original missions, the law of legal deposit applied to television and radio since 1995 gave to Ina the responsibility of collecting and keeping the memory of a growing number of channels, till 100 television channels and 20 (soon 64) radio stations that are recorded continuously, generating 1 million hours of new content every year. These contents are submitted to a very strictly defined frame of consultation and use: they are indeed dedicated to research use only, and can't be proposed for any commercial use. These different contexts of collecting collections are connected to different media proposed to publics for their consultation and eventually reuse: on the web or limited in the physical consultation rooms of the establishment (dedicated to research), three main ways ok making our collections accessible co-exist, adapted to the different publics concerned: audiovisual professionals, researchers, but also citizens. However, because of a historically "fragmented" IT architecture, the Institute has been suffering these last years from an under-utilization of its millions metadata collected in support of audiovisual collections, in particular in the frame of the legal deposit of radio and television. This is why since 2014 the Institute has been remodeling its documentary IT, in close coordination with the broader construction of a "data lake", which will allow merging the metadata from all enterprise tools (documentary, legal, and commercial). The adoption of a big data policy tends to develop new ways of using metadata, without putting into question the process of collecting them, guaranteeing their quality, whether they come from external sources or from internal activities of documentary description. This rationalization of the metadata management should also help to build and systematize their opening policy, according to legal criteria to be determined such as their origin (purchase, gift, exchange). This multi-year project aims to harmonize descriptive metadata according to common thesaurus and to streamline production processes as well as to promote new uses of these contents within the Institute (partial automation of documentary processing by automatic detecting of quoted or represented entities (faces, names, …), automatic articulation of documentary and legal metadata, …), but also outside of the Institute (online access to open data, access to media by technical data mining, …). As a matter of fact, the remapping of our data in the frame of a data lake should simplify our IT architecture and facilitate the integration of external metadata such as the ones produced by the National Library of Wikipedia, provided that our new data model will be generic enough to allow interoperability. These evolutions lead naturally to an interrogation concerning the duties and competences of an establishment like Ina, dedicated to the national audiovisual memory. In which way evolutions of tools contain in themselves the evolution of occupations?

Tipology: 1.02 Review Article
Submitting date: 27.01.2016
Acceptance date: 20.02.2016