THE IMPACT OF SEMANTIC TECHNOLOGIES ON THE ARCHIVES & THE ARCHIVAL

Guy Maréchal, TITAN, Belgium

At the 2009 IASA Conference in Athens, the Cataloguing and Documentation Committee was re-named Organising Knowledge (OK). The change in name reflects the intention to adopt a new perspective on how we create, enhance, manage, link, and share metadata about our collections and, crucially, to understand and harness the possibilities of the semantic web.

During the following IASA Conferences (Philadelphia, Frankfurt, New-Delhi, and Vilnius) the IASA-OK task force has taken various initiatives focused on "awareness."

Beyond inspirational papers and a name change, what is next? Those responsible for cataloguing and documentation in established institutions face major challenges in this area. One is to understand better the landscape of resource discovery, navigation, knowledge as brand and product, user behaviours, and how contextualising our metadata as knowledge can promote discovery.

The existing IASA recommendations on "Cataloguing & Documenting" have been designed for human understanding. The intention of semantic technologies is to represent things in a way that computers could 'understand' them and present them to people in a form that they can exploit. The IASA recommendations remain valid! Semantic technologies allow the empowerment of recommendations for better visibility and optimal linking of data from archives with current data on the Web. Many archiving organisations have implemented retrieval and access to their catalogue items with the same intentions but without using semantic formalism and standards. This means that, for most organisations, the situation could stand as it is. But, interoperability between archive websites remains unrealized. A possible migration to representations compliant with semantic standards could occur in parallel. This is the reason IASA-OK has given priority to establishing guidance for a graceful migration of existing records associated with a 'catalogue item' to interoperable semantic resources.

Semantic technologies offer extra capabilities by constructing typed relationships between things (such as "A" is a part of "B"; "Q" is an instance of a "Person"; or "Q" is the author of "W"), by expressing the internal structure of a work and the links between works and associated documents. Further, native semantic contents will have to be archived with their semantic power.

Semantic technologies are also used for empowering the responsibilities of archival management, in particular: cataloguing, documenting, enriching; managing and clearing rights; retrieving and accessing, and persistent preserving.

1. Introduction

The current method of managing archival catalog data is called "**flat modeling,**" which refers to the fact that each archived asset can be seen as an island in an archipelago (one catalogue item in a collection).

Although "no archives is an island,"[18] the flat modelling approach is sufficient in most current situations.

The "**networked modelling**" offered by semantic technologies is particularly fitting when the archived material is, by nature, interconnected or more easily retrieved by semantic navigations. Networked modelling fits well for archiving news, sports, political, ethnological, architectural, thesaurus, and multilingual information.

18 The focus of the IASA-2008 conference in Sydney, Australia.

1.1. What is the Semantic Web?[19]

The Semantic Web is a web that is able to describe things in a way that computers can understand. Sentences such as "The Beatles was a popular band from Liverpool", "John Lennon was a member of the Beatles", and "Let It Be was recorded by the Beatles" can be understood by people. But how can they be understood by computers? Statements are built with syntax rules. The syntax of a language defines the rules for building the language statements. But how can syntax become semantic? This is what the Semantic Web is all about, describing things in a way that computer applications can understand.

One can argue the fact that the semantic technologies will not emerge by pinpointing their current limitations. The article of Wikipedia in reference to the semantic web (Link 6 in the useful links section at the end of this paper) and Stefano Cavaglieri's tutorial (included in the bibliography at the end of this paper) include excellent commentary on the subject. However, the dynamic cannot be surrounded! The analysis of the semantic approach and of its potentialities remains worthwhile.

2. Illustration of the change by a simple example

The usual way of cataloguing and documenting media assets is to fill-in a metadata template for each of the assets and then to store that data in a database. The list of metadata elements depends on the nature of the asset (e.g., a book or a sound recording), of its cultural domain, and on other classification and sector rules. For example, three typical metadata elements are very general: the name of the asset, the name of the contributor, and the hyperlink to the file representing the asset.

The well-known "*Eine kleine Nachtmusik*" was composed by Mozart. According to XML, Dublin Core, and METS syntaxes, these metadata elements could be expressed in the following manner:

```
<dc:name>Eine kleine Nachtmusik</dc:name>
<dc:contributor>Mozart</dc:contributor>
<mets:file ID="FILE_W002" ADMID="TMD_W002"
MIMETYPE="audio/wav" GROUPID="GW003" SIZE="1"
CHECKSUMTYPE="MD5" CHECKSUM="the_md5_file_check-
sum here">
  <mets:FLocat LOCTYPE="URL" xlink:href="file://root/path/
subdir/S_2069-B-01-W3.ogg" />
 </mets:file>
```

Anyone who has a minimum knowledge of music history will understand that it is meant that the composer is Wolfgang Amadeus Mozart (1756–1791) and that the music involved is the usual name of the serenade identified K.525. Everyone should also forget about the hyperlink and simply assume that a file coded in the "ogg" format is available representing the audio recording.

For Information Technology (IT) it is precisely the reverse: "Mozart" is simply and not more than a string of characters and "*Eine kleine Nachtmusik*" another one. However, through the complex hyperlink, IT has what is required for presenting you in evidence the beautiful sound of Mozart's music!

The fundamental intention of semantic technologies is to ensure, by construction, the **interoperability** of applications and navigations through the expression of the relations existing between representations of concepts and their instances with their characteristics. Such representation is usually expressed according to a combination of standards languages (using the XML syntax) of the W3C, in particular, RDF (Resource Description Framework) and OWL (Ontology Web Language).

19 See the Tutorial of Stefano Cavaglieri (included in the bibliography at the end of this paper).

Semantic technologies support keeping current representations according to the usual cataloguing and documentation rules expressed using the well known DC, MARC, or MODS models, among others (collectively referred to as "**Flat**" models). Semantic models (collectively referred to as "**Rich**" models) can hook into and integrate the metadata from "Flat" models.

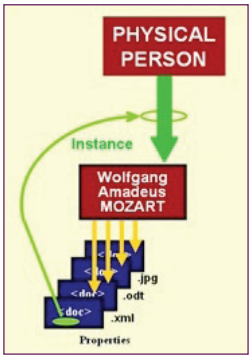

Figure 1: "Is an instance of"

In the previous example, for semantic technologies, the representation of Mozart is a resource being an instance of the class of things called "Physical person." Figure 1 illustrates the approach. The rectangles represent "Resources" being identified. The upper red rectangle has the class "Physical person." The relation "**is an instance of**" is expressed by the green arrow.

The middle red rectangle represents the resource carrying the representations and properties of Mister Wolfgang Amadeus Mozart as an instance of the class "Physical person." It 'owns' the lower rectangle representing the existence of the resource and its associated properties, including the relation "is an instance of," which links it to the class "Physical person." The instance inherits all the characteristics of "Physical person."

The lower blue rectangles represent the digital files representing Mozart. In the example, the .xml file could carry the classical "Flat" model according to the Dublin Core of the structural representation of his life (e.g., date of birth; marriage; or date of death); the .odt file could carry a bibliography; and the .jpg file could carry the scan of a painting representing him.

Any of the relations could, through the Web, link data present in distinct databases: this is what is called "Linked Open Data" (LOD). The OWL definition of the class "Physical person" is in one semantic database (FOAF for example) while its instances, including you and "Wolfgang Amadeus Mozart" could be described in a variety of independent semantic databases linked by LOD and alias. This constructs a **network of related data**.

3. The graceful migration of "Flat models" to "Networked models"

The existing IASA recommendations on "Cataloguing & Documenting" are intended, meant, and dedicated to be applied by human, expert cataloguers. The expression of metadata and identifiers is assumed to be read by humans.

Figure 2 is an excerpt of these recommendations (Section 0.B.2). It represents an example of what will be the subject of the 150 pages of the recommendations.

Joyride [sound recording] / Roxette. - Solna : EMI, p 1991. - 1 sound disc (ca. 49 min.) : analogue, stereo, 33 rpm ; - 30 cm

Words and music: Per Gessle (unless otherwise stated)
Lyrics on inner sleeve

Contents: Joyride -- Hotblooded / P. Gessle, M. Fredriksson -- Fading like a flower (every time you leave) –Knockin' on every door -- Spending my time / P. Gessle, M. Persson -- Watercolours in the rain / P. Gessle, M. Fredriksson, C. Öfwerman -- The big L. -- (Do you get) Excited? / P. Gessle, M. Persson -- Small talk -- Physical fascination -- Things will never be the same -- Perfect day / P. Gessle, M. Persson

EMI: 7960481

Figure 2: Catalogue metadata

The rules for elaborating these catalogue items are detailed in the recommendations. The expertise of the cataloguers ensures the correctness and fitness of the data. A trained reader will understand the record without ambiguity. The casual reader will have some difficulties to fully understand and exploit the data. But, in most archiving organisations, access is ensured through Web pages created from queries of records in a database. Hence, for casual users, the intuitive presentation of Web pages facilitates access and understanding.

Everybody knows **Wikipedia**, the open encyclopaedia. A few years ago, volunteers coded its structures and coding rules according to semantic standards. They also developed a program for an automatic migration of Wikipedia records into a semantic database called **DBpedia**—a crowd-sourced community effort that extracts structured information from Wikipedia and makes this information available on the Web according to the "Linked Open Data" protocol. DBpedia allows you to ask sophisticated queries against Wikipedia. It links the different data sets on the Web to Wikipedia data. It provides easier access to the huge amount of information in Wikipedia and allows this information to be used in new and interesting ways. Furthermore, it might inspire new mechanisms for navigating, linking, and improving the encyclopaedia itself. Wikipedia and DBpedia are both available in parallel. More than 10,000 websites are now accessible through the Linked Open Data protocol, which constitutes a huge unique distributed database.

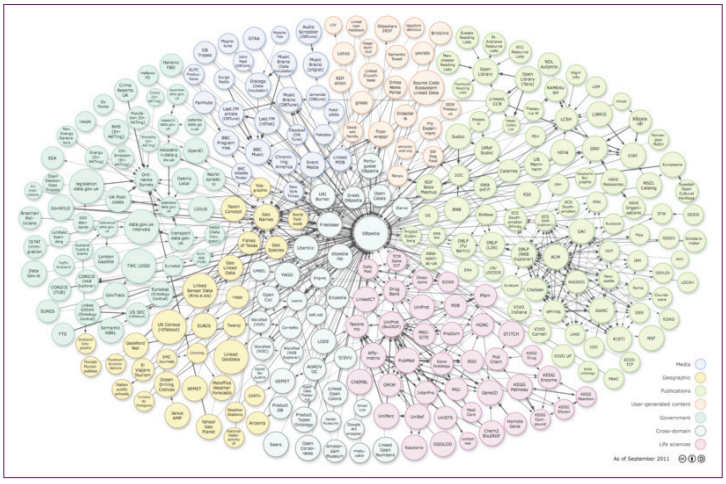

Figure 3: The semantic Web (as September 2011)

Figure 3 represents the main Web-3 sites as of September 2011. The current scenario can no longer be presented as one figure. The prominent place of DBpedia is clearly visible in the figure. In the media sector (in blue) the BBC appear to be very active.

This same type of initiative could easily be implemented for the migration of the records associated with archives that are documented according to IASA recommendations. It could become a concrete initiative and will be analysed during the IASA-2014 conference in Cape Town, South Africa. The process will be to represent the IASA guidelines and recommendations according to semantic standards. This step will initiate the issue of what is called an "**IASA-Knowledge Base**." It constitutes a formalisation "understandable" by computers of the existence of and of the meaning of the red additions presented in Figure 4.

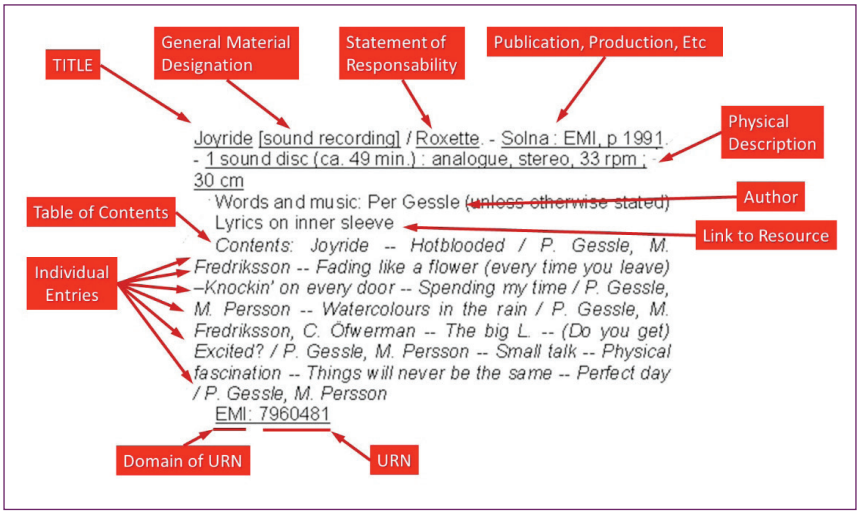

Figure 4: The initialisation of the IASA-Knowledge Base

For each of the "red entries," semantic formalism will express not only the typing and properties, but also the coding rules for the representations dedicated to people. Figure 4 is the same as Figure 2 but simply the rules for "understanding" the intentions and their expressions are now expressed in a form "understandable" by computers and by humans (independently of the selected access device).

Similar to what has been done for harvesting Wikipedia to create DBpedia, each archive database could be used to create its semantic database mirror. This could be done under the direction of the IASA-KB (a tool to support mapping IASA-conformant catalogue records to valid RDF instances). Such extractors are usually designed for systematic updates to ensure that the semantic database remains a mirror of the flat database. It is a synchronisation process allowing the presence of both representations standing in parallel. The inherent multilingual capability of semantic technologies implies that contextual translations could be made or verified.

The resulting semantic database (SDB) could be used only within the local organisation but could also be accessible and referenced on the Semantic Web. In this second case, your website would become a new node in the diagram in Figure 3. In both cases the SDB will be accessible according to the LOD protocol. Obviously, here "open" does not mean that any free access is allowed! It means that, if access is allowed, the data will be "understandable" in an open, standardized way. It means the opposite of "proprietary" coding.

4. Enhancing the Knowledge Bases and the contents of Semantic Databases

The representation of knowledge has to be seen in independent levels, distinctly from the point of view of humans and from that of ICTs (Information and Communication Technologies). Each of the levels could be empowered by the next higher level. The movement between levels occurs for humans through higher cultural and social education and for ICTs by training, trials, validations, or error corrections, under the control of educated persons.

Each of the ways of representing knowledge could be added to the Knowledge Bases and serve the Semantic Databases of the instances.

4.1. Textual representation

The textual expression of knowledge is very powerful for producing knowledge and for accessing it by humans. In this level, interoperability is ensured between individuals sharing the same culture, the same language, and having common social repositories and locators. The complexity and the richness of grammar rules, of syntaxes, of poetry; the voluntary multiple, evocative, and ambiguous meanings; and the games between sound and sense all open doors to utterances that are above knowledge. These expressions can be stored in a persistent manner with no loss of information, but they are not normalised and have poor precision and recall capabilities. This is the level of the Web-1. When using semantic technologies, the structure of sentences and other logical elements could be extracted in a similar way. The same with identifying that some of the words in the text are names of typed items (e.g., persons, organisations, places, dates, moments, periods, or concepts). Each of the typed items could be an object of a specific LOD entry. Each applicable occurrence of the word in any text could be linked with the article or to a simple tag. Semantic technology allows for resolving ambiguities in free-text representation: the word Paris usually refers to the capital of France; but it could also designate a small city in Texas or the surname of a person. The empowerment of texts by semantic technologies becomes also an empowerment of Knowledge Bases; and vice-versa, the use of existing Knowledge Bases could be made in order to empower texts and local Knowledge Bases. The main empowerments of KBs take the form of Tags, Taxonomies, and Thesauri.

4.2. Tagging

Tagging has a low threshold. For most cases it is sufficient for a single human user. At this level, interoperability is ensured between individuals and machines through simple standards. Tagging offers moderate precision on large databases, although there remains poor precision with regard to the meaning attached to the tags. The control of consistency is also limited at this level. This is the level of the Web-2. Tags could be associated to structural or typed classes and be parts of KB or linked in instances.

4.3. Taxonomies and Thesaurus

This level offers very high precision but, by nature is difficult and tedious to maintain and is hardly scalable. Interoperability at this level is ensured as long as no changes occur. Combined with the level of free-text expression, this level could be very powerful. This is the level of Web-2 with data mining enrichment tools. Retrieval services such as Google and Yahoo have demonstrated the power of this level, but simple searches could generate thousands of hits. Taxonomies and Thesauri could be associated with structural or typed classes, used in Tags, and could be parts of KBs or linked in instances. Important thesauri and taxonomies could be combined to construct networks of Multilingual Knowledge Bases (MKB). UNESCO has such a MKB for which plans are under way to present it in LOD.

4.4. Semantic

Semantic expressions of knowledge are very powerful for producing or accessing knowledge by ICTs. But the modes of representation of this knowledge, in a way suitable for human understanding, remain a research area. At this level, precision and scalability are

without limits and interoperability is ensured for all the situations where formal modelling could apply. Recall and retrieval is optimum—the thousands of returns from a query at level 3 are refined to only pertinent and serendipitous hits. In concrete trials in large semantic databases, we often obtain only 30 replies (with 20 or more pertinent); while for the same searches at level 1 (Web-1), millions of replies were frequent. Navigation in semantic databases and LOD is simple. This is the level of Web-3.

4.5. Operational

Semantic formalisms open the door to the capacity of computation, inference, and operations through "intelligent" agents. Such associated technologies are partly available and already in use in targeted domains. Again, this means that the empowerments of SDBs and of MKBs will occur in an automatic way made by ICT processes working in the background.

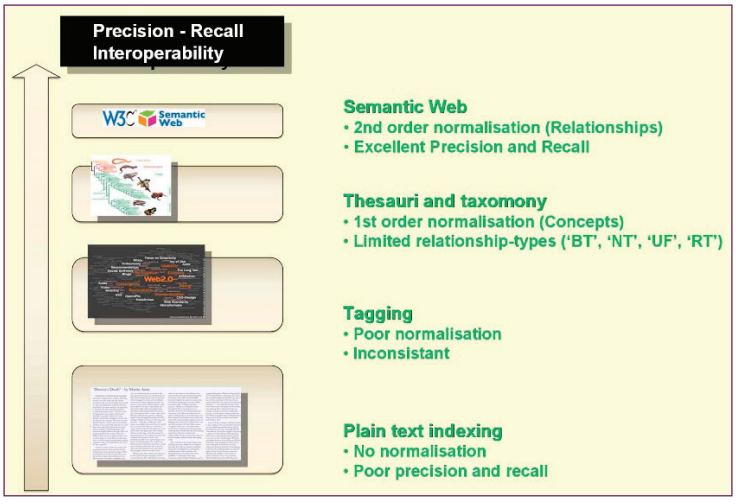

Figure 5: The ways of representing Knowledge to be federated by semantic technologies20

4. Structuring inside the media

Some media have, by nature, an inner structure. Semantic technologies cover the inner structuring of these cases: the "Fragment" standard.

A typical example is the recording of an interview. It implies interviewer(s), interviewee(s) and the subject(s) of the interview. The semantic structuration would typically include:

- presentation of the agents (links to the MKB(s))
- presentation of the subjects (links to the MKB(s))
- transcription of the sentences pronounced by each agent
- translation of theses sentences
- keyword spotting and their linking to the associated tags
- documentation of the event process

Another typical example is audio-visual media associated with a sporting match or with a television news program.

20 Illustration by courtesy of Maarten Verwaest.

For these important cases, specific ontologies have been designed for the modelling of the event processes and of their associated media. In particular, a combined ontology for News and Sports // Events and Media has been developed. In order to ease the exchange of news, the International Press Telecommunication Council (IPTC) has developed the NewsML Architecture (NAR).

It combines the NewsML G2 standard and the EventsML G2 (see Link 9 in the references section of this paper: http://www.eurecom.fr/~troncy/Publications/Troncy-iswc08.pdf).

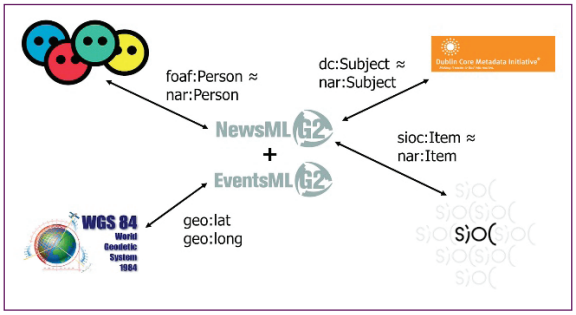

Figure 6: The NAR profile

This technology allows for structuring and labelling as illustrated in Figure 7. Similar allowances are made for interviews and the news (from headers; weather forecasts; sport sequences, up to the final summary).

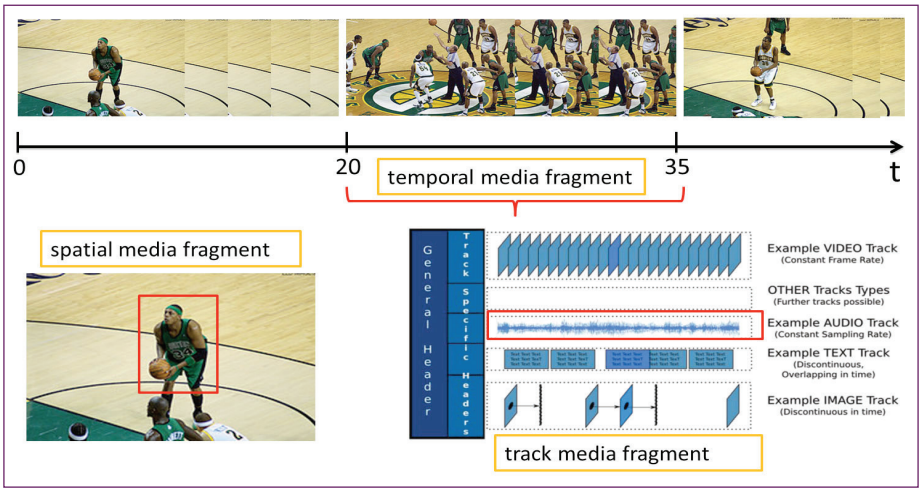

Figure 7: Illustration of an NBA match[21]

5. The power of semantic technologies for archiving processes and management

Semantic technologies are powerful for the empowerment of archives. But they can also be valuable in managing archival processes and duties.

The previous sections have demonstrated the power of semantic technologies for cataloguing, documentation, and enrichment. Their power for managing archives and clearing the associated

21 Image by courtesy of Erik Mannens.

rights is also demonstrated in many projects (in particular the Presto4U project). Semantic formalisms improve navigation, discovering, and access. The power of semantic technologies for "Preservation & Persistence" according to the OAIS standard have also been demonstrated.

6. Conceptual reference models and resolvable identifiers

Semantic technologies are powerful for implementing "Conceptual Reference Models."

The most important models are the FRBR standard (Functional Requirements for Bibliographic Records) and its associated FRBR-OO (Cidoc-CRM). These are already in the semantic trend.

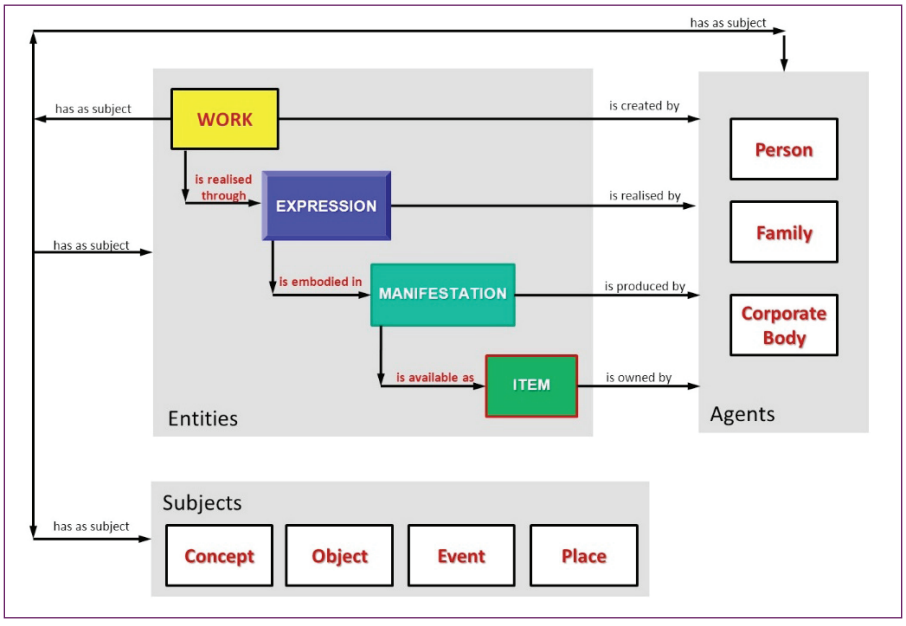

Figure 8: FRBR (Entities, Subjects and Agents)

FRBR is a powerful conceptual tool for bibliographic situations and associated items. It could serve as a reference in many derived domains. However, its mapping to semantic technologies has required adaptations, in particular to cope with the capacity of declaration of the existence of an object, independently of the declaration of the existence of the several independent models of the object; the incorporation of management of "Rights" information; the reuse of works in works; and the explicit declaration of roles, of characters, and of the temporal availability of works (such as acquiring rights to enjoy contents in streaming mode).

In the context of two projects (Memories and MediaMap), the non-profit association, Titan, has migrated the FRBR concepts into a full semantic representation, covering any object and subjects and their documentations and structuration. This effort has been undertaken by collaboration with the UNESCO MoW programme, the Radio France archive department, the Celtic telecommunication sector of the Eureka programme, and a few others. The result has been named FRAR, standing for Functional Requirements for Assets and Rights.

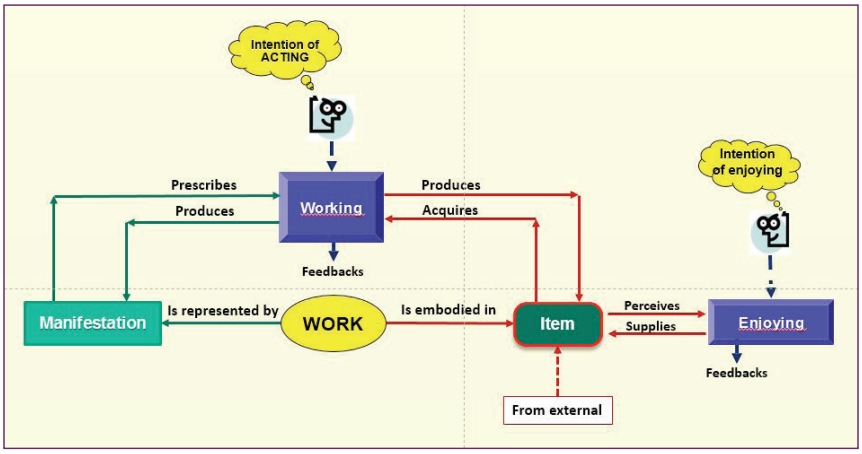

Figure 9: FRAR (Functional Requirements for Assets and Rights) Top view

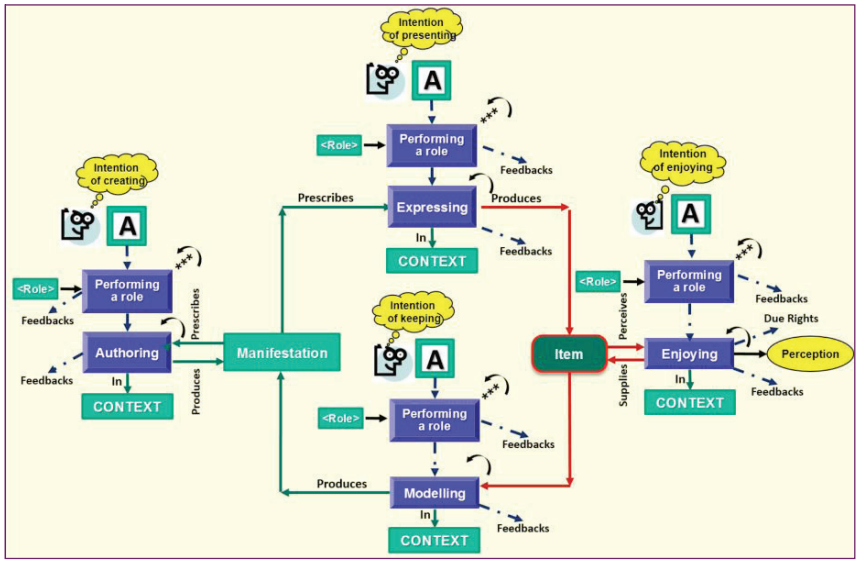

Figure 10: FRAR (Functional Requirements for Assets and Rights) Detailed view

It is out of the scope of this paper to detail the FRAR model. It is simply a way of introducing the need for a reference conceptual model as foundation for collaborative work in the AV archiving sector.

Another important area to be covered is the elaboration of an open identification system that could cope with many identification systems for the objects and for the models of the objects. The systems should be able to resolve URN and URL in both cases.

7. The IASA-OK Task-Force

In IASA, the OK dynamics is, up to now, a transient effort of a few persons. After a few years focused on "**awareness**," I am convinced that IASA should enter into a more concrete limited set of actions organised jointly with other similar associations such as FIAT and AMIA. Links with UNESCO also should be established, in particular with the Memory of the World (MoW) programme and the Multilingual Cultural and Educational thesauri. Such projects could be sponsored by public authorities. These projects should also be built in continuity with existing achievements. The IASA-OK should set the foundation for a possible future preparation of Recommendations and Rules for structuring the archives and managing them.

IASA is entering into a third phase of its focus where IASA-OK can have a leading presence.

Foundation: **Phase 1: AV assets represented according to ANALOGUE models**
Restoration and preservation of carriers; cataloguing; metadata; ethical guidelines

2004: **Phase 2: AV assets represented according to DIGITAL FLAT models**
AV carriers; formats; wrappers; records in databases

2011: **Phase 3: Native semantic & interoperability of Objects, Subjects, and their Relations**
Archiving of complex subjects (series with bonus and associated documents); preservation of the substances; profiles; ontologies; taxonomies; cataloguing; knowledge bases; metadata according to several standards; linked open data

8. Conclusion

The semantic enhancement of existing "Flat model" databases can occur through automatic migration processes. The prerequisites are the construction of Semantic Knowledge Bases specific to the local organisations and their linking to general purpose SKBs. By systematic synchronisation the "Flat model" database and the "Networked model" database could remain active in parallel ensuring a graceful migration to the power of semantic technologies.

9. Acknowledgements

The author would like to thank Stefano Cavaglieri, Jean-Pierre Évain, Erik Mannens, Maarten Verwaest, and Tobias Bürger who offered the authorisation of including here short excerpts of their texts or presentations.
I would like also to thank Roger Roberts (president of the non-profit association TITAN) for its continuous involvement in the research and implementation of concepts suitable for producing contents "native semantic" and for persistent archiving in a semantic context.

10. Bibliography

Cavaglieri, Stefano. "Semantic Web for the IASA (Tutorial)." Tutorial at the IASA-2013 Conference (Vilnius), 2013. (Available on the IASA Web site).
Evain, JP and T. Bürger. "Semantic Web, linked data and broadcasting." In EBU T Review (2011 Q1).
Mannens, Erik. Interoperability of Semantics in News Production. ISBN 978-90-8578-415-9 (2011 Q).

11. Useful links

[Link 0] http://www.iasa-web.org/cataloguing-rules
[Link 1] http://www.ted.com/talks/tim_berners_lee_on_the_next_web.html
[Link 2] http://www.w3.org/DesignIssues/LinkedData.html
[Link 3] http://www.w3.org/2008/WebVideo/Annotations
[Link 4] http://www.w3.org/TR/2010/WD-mediaont-10-20100608/
[Link 5] http://tech.ebu.ch/docs/tech/tech3293v1_2.pdf
[Link 6] http://en.wikipedia.org/wiki/Semantic_Web
[Link 7] http://www.iptc.org/site/News_Exchange_Formats/NewsML-G2/
[Link 8] http://www.w3.org/TR/WebIDL/
[Link 9] http://www.eurecom.fr/~troncy/Publications/Troncy-iswc08.pdf