The semiotic turn in digital archives and libraries.

Peter Stockinger

The *semiotic turn* in digital archives and libraries

Prof. Dr. Peter Stockinger
Institut National des Langues et Civilisations Orientales (INALCO)
Filière Communication Interculturelle (CFI)

Equipe Sémiotique Cognitive et Nouveaux Médias (ESCoM)
Programme « Archives Audiovisuelles de la Recherche » (AAR)

1/ the meaning of data

To-day, we witness the massive digitization of any kind of media objects, the storing and diffusion of it in form of digital (multimedia) libraries, archives or any other form of digital "(multimedia) information spaces". The general policy is to make available the enormous quantities of media data in form of relevant (critical) *knowledge* or *knowledge resources* for a *target public* – a person, a social group, an institution (this policy underlies the shift from an "information society to a knowledge society" as proclaimed, for instance, in the Lisbon 2001 Declaration or again in the actual Europe 2020 program). However if we want to progress in this historically and culturally certainly highly exciting and innovative direction, a series of serious – technical, social and scientific – problems has to be solved.

One of the most complex ones is without any doubt the question of how to process the *symbolic* or the *meaning* of (digital) media data with respect to a given – analogically speaking - *market place of meaning production, sharing and consumption* (potential users of meaning, cultural expectations of meaning, needs and desires of meaning, uses and exploitations of meaning, added value of specific forms of meaning …). Indeed:

> … a (*digital) media data* (a still image, a video, a sound record, an oral record, a printed document, …) is not in itself already a *genuine cognitive resource* for a given "reader" or "community of readers", that will say a *relevant* "means" for an agent to solve a problem, to answer a question or again to satisfy a (personal or collective) goal or a need.

A digital media data is, in other terms, only a *potential* cognitive resource. It has to "undergo" more or less significant qualitative transformations in order to become a user or a user community relevant one. These qualitative transformations are performed through series of concrete operations such as the constitution and classification of relevant corpora of digital records, the description and indexing of records, the processing of digital data (segmentation, tagging, linking, montage, …), the (cultural, linguistic) versioning (commenting, translating, …) of given source records or again the (re-)publishing of digital records.

These and other operations constitute what we call the *semiotic processing* of (digital) media objects, corpora of (digital) media objects or again entire archives and libraries. They demonstrate practically and theoretically the well-known "*from data to meta-data*" or the "*from (simple) information to (relevant) knowledge*" problem – problem that obviously determines the *effective* use and also the *future* of digital knowledge archives.

In short, the central question here is that of the *semiotic structure* of (digital) media data, i.e. the *structural organization* of (digital) media data, their *status* and *function(s)* for a given user or community of users and of how to deal with them, how to work concretely with them in order to achieve not only theoretical but also practical - educational, economical or other - objectives.

2/ some major challenges exemplified through the French ARA program

With regard to this general challenge we call the *semiotic processing* of digital media objects, we would like to introduce and present here briefly the French programme Audio-visual Research Archives (ARA)[1] we have started in 2001 in Paris. This archive actually recovers more than 6000 hours of online streamed videos, accessible for everybody, covering a large diversity of disciplines in human and social sciences. The records composing the ARA corpus contain individual interviews with researchers, conferences, lectures, research seminars, workshops, small reportages about the daily life in research labs, scientific expositions, documentaries, travelogs and road movies, ethnographical films and field work recordings. The offer is composed of almost 270 research seminars, symposia and workshops as well as of more than 380 interviews with researchers and experts in social and human sciences. More than 2800 researchers, scholars and professionals from 80 countries have contributed to the cultural and scientific heritage stored and diffused through the ARA portal. Even if half of all videos are in French language, there also exist large corpora of videos in English, Italian, Spanish, German, Chinese or Russian. The average growth per month of this audio-visual archive is of about 30 hours on line, this means of about 60 hours of video taking and digitizing.

The principal user communities of the ARA cultural and scientific heritage are the (French speaking and international) research community itself, the educational community and specialized professional communities (science journalists, stakeholders especially in non-profit organization, governmental agencies …). The proposed content composes indeed a highly specific "market niche" in the digital (audio/video) content production and communication. All available indicators show that this archive has a potentially high impact for teachers in formal contexts but also for educators and learners in informal settings (lifelong learning, etc.), for professionals working, for instance, in the sector of specialized and highly specialized information media or in organizations where an expertise based on knowledge produced by the social and human sciences is indispensable (NGOs, community and territorial structures, political actors, …).

But such as, this whole heritage has however to undergo a complex process of digital repurposing, of digital *re-writing* of streamed videos in order to fit more precisely with specific user profiles and user contexts.

The reason for this is that the specific *auctorial profile*, the "*auctorial identity*" of the digital media objects composing the fonds of the ARA program does not necessarily fit with the expectations, needs or desires of an individual user or a target user community. Another reason is that this auctorial profile simply isn't perceived by a target user community. It is, so to speak, hidden either from a strictly cognitive point of view (the content remains *not understandable* for a target community) or from a physical point of view (the content is *not attainable*). In a nutshell, there are (among other) the following three serious challenges that have to be faced in a digital repurposing (re-writing) process of digital media objects or corpora of objects:

[1] URL of web ARA web portal : http://www.archivesaudiovisuelles.fr/EN/

1/ the *language limitation:* the diffusion of digital content has to be "improved" in opening and making it available to an *intrinsically multilingual knowledge market* by the means of an extensive use of hints and aids for an at least basic linguistic understanding of a content performed in a given source language;

2/ the *hidden information:* the quality of existing digital content has to be improved in eliciting, systematizing and classifying the *hidden information* in large audio-visual databases by the means of *context-sensitive* description and indexing of digital media corpora using the *same analytical (meta-linguistic or, more broadly speaking, meta-semiotic) resources* (verbal, iconic or other kinds of thesauri, ontologies, description models…) or again analytical resources which are *interoperable* (which, in some way, are able to *communicate* between them);

3/ the *adaptation to specific purposes:* the quality of existing digital content has to be improved for *specific contexts of use* (especially: formal and informal education) and an intrinsically *multicultural market* (characterized by diverging knowledge and value references, by a diversity of expectations and beliefs, interests and needs, …) with the help of *context-sensitive* re-authoring and re-publishing models and tools.

The scientific, technical and practical work on these three limitations has constituted (and still constitute) the principal motivation of a series of European and French R&D projects which we have coordinated or in which we have participated as a consortium partner since 1989[2].

In the following chapters we would like to present globally the general "philosophy", the general assumptions that underlie our research agenda since the last twenty years and which we summarize under the label of the *semiotic turn in digital archives*.

3/ digital media object repurposing

One of the major challenges (if not the major challenge) for the constitution of a genuine knowledge community within an intrinsic multilingual and multicultural world, is the *context-sensitive* exploitation of existing digital resources such as videos, sound tracks, electronic versions of printed texts, etc. produced by the concerned users themselves or

---

[2] The full list of the R&D projects in which we have been implied since 1989 can be found here: http://semioweb.msh-paris.fr/escom/ressources_enligne/p_stockinger/2010/CVAnglais_2010.pdf. We would like to stress here more particularly the importance of the ANR founded French project ASA-SHS (Audio-visual semiotic workshop for processing and describing video corpora in social and human sciences) which has offered us the possibility to design and to develop an environment for researchers working in the humanities and aiming at the production of "personal" audio-visual research archives. This environment is composed of a whole *meta-language* (ontology, thesaurus and description models) adapted to the specificities of research in the humanities as well as of the necessary tools for adapting the generic version of the meat-language to domain specific requirements, for segmenting and tagging video files, for describing and indexing video files and, finally, for republishing video-files online. Information about the ASA-SHS project which has started in 2009 and finished in 2012 can be found here: http://asashs.hypotheses.org/. The recently started ANR founded project – Campus AAR – continues to develop this environment. The main objectives are the following ones: 1) integration, in the existing environment, of the processing of other audiovisual media objects: still images, sound records and printed material; 2) interoperability of the used meta-language with major standards and thesauri; 3) multilingual version of the used meta-language; 4) user-friendly and user-adapted versioning of the existing environment. More information of the Campus AAR project which has started in January 2014 for three years can be found here: http://campusaar.hypotheses.org/.

other authoring instances and stored in central or distributed digital libraries. This challenge is identified under the heading of *(re-)purposing* of (digital) media data.

(Re-) purposing of digital media objects is a more or less complex (individual or collective) process by the means of which a digital media object or a corpus of digital media objects (such as a corpus of video clips, a corpus of printed or handwritten texts, a corpus of still images, a corpus of sound recordings, etc.) are adapted, attuned to a *specific context* of use and, this, following a sort of *authoring* or *publishing scenario*. There exist many different related notions that capture the one or the other specific aspect of the process of repurposing of digital media objects – notions such as:

- the segmenting or re-segmenting of video or audio-files,
- the delimitation or re-delimitation of 2-D regions in still images,
- the classification or re-classification of media objects (or parts of a media object),
- the description or re-description of the content of media objects (or parts of a media object),
- the description or re-description of the audiovisual expression, of the formal and the the physical organization of a media object (or a part of a media object),
- the "subjective", theory-bounded, … interpretation and annotation of media objects (or parts of it),
- the (thematic, rhetorical, …) positioning or re-positioning of a media object (or a part of it) within a field (a context) of related media objects,
- the processing/re-processing of specific features of a media object (for instance, a video or a still image) itself: the blurring of a person's face, …,
- the publishing/republishing of (virtual) parts of one or more audiovisual media objects (in form of, for instance, mash-ups, web documentaries, thematic, pedagogical or bilingual folders, etc.),
- the collaborative and personalized archive building and diffusion for a given domain of discourse;
- the channelization of digital media assets with respect to a user's interests or preferences,
- etc.

In considering more precisely the (re-)processing process (figure 1) itself, it has to be explained with respect to:

1. The *type and genre of (re-)purposing:* content selection in a given corpus of existing digital objects, explanation & completion of selected content; existing digital object versioning; content translation; digital media object interlinking; creation of new content parts; (visual, sound, …) content expression modification; etc.

2. The specific profile of the source media: *type* and *genre* of digital objects: technical and scientific texts; audio-visual ethnographic documentaries; cultural heritage images; etc.

3. The *goals* of a (re-)purposing process: *contexts of use* (learning, teaching, science popularisation …); *destines* (pupils, students, any person, specific social group …); *publishing genre* (courses, thematic folders, glossaries, info flashes, educative games …); *forms of distribution*; etc.

4. The *resources*, *means* and *tools* of the (re-)purposing process: *human resources* (authors, domain specialists, translators; publishers; …); *conceptual resources* (ontologies; thesauri; publishing models; aids, hints and fully developed explanations and methodologies; …); *technical resources* (media processing tools,

indexing tools; annotation tools; translation tools; publishing tools; …); *economic resources* (budget …).

5. The *process* of (re-)purposing *itself*: the phases and tasks composing a *(re-)authoring chain* (such as "publishing genre selection and preparation"; "corpus constitution"; "description and indexing"; "translation"; etc.).

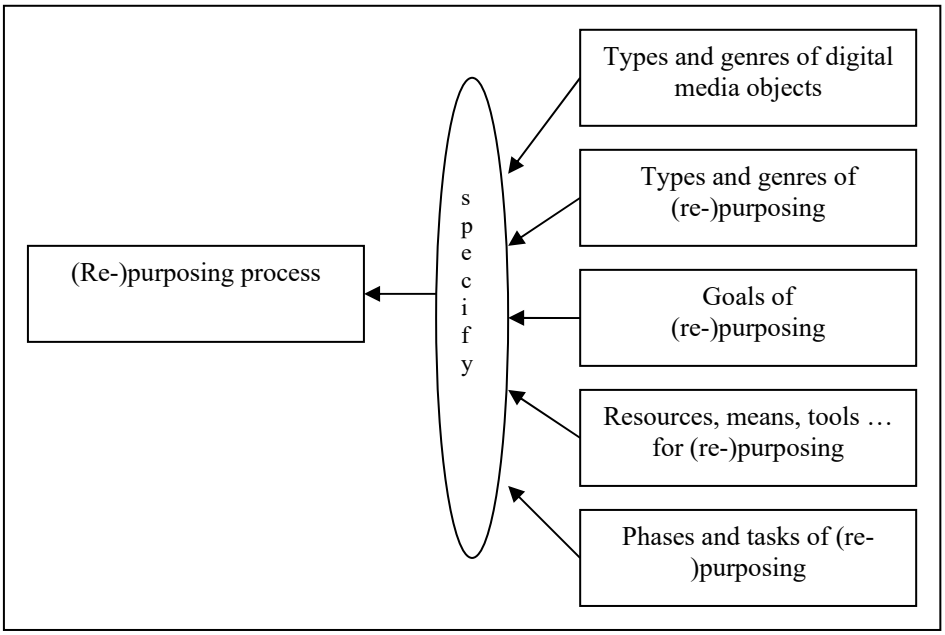

Figure 1: *Central components specifying the (re-)authoring process*

The (re-)authoring process, can be understood as a – in a broad sense – *cultural translation* process. Digital (media) objects are meaningful entities with respect to context(s) of uses, destines, social practices, etc. for which they have been designed. In other words, they belong and refer to a *culture* embodied in the life-world of social actors, viz. user communities.

For instance, there exists a corpus of audio-visual files in the already quoted audio-visual research archive for human and social sciences (i.e. the ARA program[3]) which deals extensively with the topic of linguistic diversity from a *socio-linguistic* point of view and of which the main target public are linguists, experts in this field. However, this corpus can be (re-)purposed (re-authored) for a high diversity of potentially relevant contexts of use and potentially interested destines. It can, for instance, be repurposed as a semantically restricted audio-visual library dedicated exclusively to this domain of knowledge[4]; it also can be re-authored by the means of different publishing formats: as an *introductory seminar* in philological studies, as a *complementary course* in anthropology or sociology, as a *mash-up* for a wider interested public, as a set of *info flashes* for people only with local interests in this domain, as a *discovery game* for children.

[3] http://www.archivesaudiovisuelles.fr/EN/

[4] Thanks to a 7th FP European founded R&D project (LOGOS, 2006 – 2009), we have had the possibility to repurpose a small corpus of interviews with (socio-)linguists and to republish (parts of) them in form of a *semantically restricted virtual video-library* called « video-lexicon ». This specific video-lexicon is dedicated to the diglossic situations all over the world. The URL of this small semantically defined video-library is: http://www.archivesaudiovisuelles.fr/FR/_Encyclo_Situation_dlc.html.

4/ digital media repurposing examples from the ARA program

In order to make more concrete our explanations, we would like to quote here briefly four concrete examples taken from the ARA program. Indeed, thanks to the already quoted European and French R&D projects, we have had the possibility to constitute a research team aiming at the design, the implementation and the practical experimentation of different repurposing (or republishing) approaches in form of:

1) the *instrumentation* of typical repurposing tasks and activities (= working environment and methodologies);
2) a library of *dynamic publishing templates*.

We will come back again to the definition and general architecture of a working environment for digital media repurposing activities (cf. next chapter). Concerning the second one – the definition and implementation of a library of dynamic publishing templates -, we have experimented mainly with following genres:

1) **specialized web portals** for *diffusing* and *sharing* thematically or otherwise circumscribed events and videos;
2) **semantic video libraries** for *accessing* videos or segments (sequences) of videos with respect to their topics, their rhetorical specificity, their audiovisual specificity, etc.;
3) **thematic video folders** dedicated to the editing of a selection of video segments that are relevant for a given topic;
4) **bilingual folders** dedicated to the opening of monolingual source videos to a multilingual public;
5) **dynamic video corpora constitution** "through the time" with respect to a specific theme, a period, a place, a personality, etc.;
6) **dynamic video-books** based on the re-editing of a given source video in taking as a cultural reference the traditional "book"-genre.

Figure 2: *Repurposing of a source video in different publishing formats*

Figure 2 shows us the republishing of a given source video with the help of different publishing or republishing formats. The source video is a documentary about the small Chilean commune *Alto Bio Bio* (8th Region) and the Pehuenches population threaten by the construction of a giant dam in the 90ies. The documentary has been produced by a group of anthropologists under the direction of José Bengoa[5] of the Universidad Academia de Humanismo Cristiano in Santiago de Chile. A first republication of this documentary has been undertaken on the ARA web portal in form of a digitized, streamed video[6]. This first republication is a kind of a "simple" digital copy of the original source documentary. The goal of this republication has been three-folded: first to diffuse more broadly the content of this exceptional document showing the disintegration of traditional social structures opposed to important industrial and financial interests; second the long-term preservation of this document; third the possibility to share and to reuse this document for research, educational and also (broadly speaking) political aims. A second and third republication of this documentary has been in form of a selection of (virtually segmented) sequences of the whole documentary in form of "chapters" and the commenting, annotating of each chapter in order to produce a sort of a hypermedia, interactive web book.

Figure 3: *Repurposing of a digital video corpus in form of a domain restricted web portal*

Figure 3 shows a second example of republishing digital audiovisual media objects in form of a *semantically restricted video-library* and a *dedicated web portal*. In this concrete case, a corpus of more than 350 hours of videos, originally diffused through the ARA web portal, have been selected, partially re-segmented ("de-linearized"), re-categorized, re-described and re-indexed by the means of a domain-specific ontology and library of description models and republished in form of a web portal specifically dedicated to Latin America's history and culture[7]. One obvious objective of this semantically restricted video-

[5] http://antropologia.academia.cl/iciis-investigadores/jose-bengoa-2
[6] http://www.archivesaudiovisuelles.fr/EN/Event.asp?id=1129&url=/1129/presentation.asp
[7] http://www.amsur.msh-paris.fr/

library and dedicated web portal is to provide (pedagogically…) relevant material about this world region. A second – important and – following our point of view – innovative objective is to build up a *common digital (local, regional or global) meaning production, sharing and exploitation place* where individual and institutional actors can cooperate as knowledge producers and/or users.

It's the upper menu bar which suggests us the principal features of the repurposed video corpus: on the one hand there are so-called "automatic" publications of this corpus based on a previous description and indexing process and on the other hand there are re-authored versions of this corpus (parts of this corpus) in form of collections of video-folders. These collections of video folders group together sequences of different videos which refer to a *common knowledge objet*. There are, for instance, a series of video-folders dedicated to Amerindian languages and civilizations, other video-folders are dedicated to the environmental and ecological questions in Latin America, a third series is dedicated to the historical and cultural presentation of Latin-American countries; and so on. Each video folder is thoroughly analyzed, explained, enriched with relevant resources belonging to exterior web sites.

The organization of the virtual video-library dedicated to the history and culture of Latin America, recovers, as shown by figure 3 and 4, a series of *headings* offering a diversity of accesses to and exploration paths of the whole repurposed video corpus: one access takes into account the *main topics* of a video or of a part of a video; another access uses a domain specific thesaurus; a third access is based on the global narrative or rhetorical structure of a video distinguishing, for instance, between scientific exposés, historical presentations, documentaries, and so on.

Figure 4: *The "Subject-area" of the semantic virtual video-library of the AMSUR portal*

As figurer 4 shows us, the access "Topics" is organized in several main collections of topics related to circumscribed, more or less well identified knowledge domains such as "Countries, regions and localities", "Cultural references", "History", etc. Each one of these collections is composed of one or a list of more specialized subjects. For instance, the topic referring to the broad knowledge domain "Cultural diversity" can be actually explored through three more specific subjects referring to: "Religious culture and popular believes", Figures of veneration", "Religious practices".

If somebody is interested in specific topics related to the domain "Figures of veneration", he/she has the possibility to access a small dynamic corpus composed of "whole videos" or segments of whole videos dealing with a series of such specialized topics as the *Blessed Virgin of Petorquita* or *La Tirana* (two small localities in Chile), the *Fiesta of San Pedro in Quilama* in Chile, the figure of the *jaguar* in the *Kuna culture* of Panama, and so on.

This small corpus is open in the sense that it can be enriched with other video resources documenting the domain "Figures of veneration". Actually, all the videos are stored in and diffused originally via the ARA web portal but, in principal, the corpus can be composed of video resources located elsewhere (relevant – academic – content providers in France are, for instance, Canal U[8], UOH[9] or HAL Video[10]).

A last example of repurposing digital media objects is shown by figure 5: an interview in French with the researcher Sabine Trebinjac from the CNRS on the *muqam* genre in the Chinese Turkestan. This interview has been *re-versioned* in several target languages (such as English, Chinese, Russian, Turkish or Spanish). The obvious objective of this repurposing activity is to open this interview to a non-French speaking public.

Figure 5: *Re-versioning of an original video in different target languages*

Each one of the five above discussed examples shows (re-)purposing as a process of – so to speak – *cultural adaptation* of the content, form and function of a given source media object with the aim to bring it into line with the (supposed) interests, needs but also abilities of the culture of target publics.

The (re-)purposing process of a digital media object can definitively be understood as the opening of the cultural specificity of a concrete digital media object with respect to a given diversity of target cultures and the attempt to encourage the circulation of a digital media object or artefact within an intrinsically *multi-cultural (knowledge) space*. Linguistic translation, in this sense, is only a very *specific case* of cultural translation.

[8] http://www.canal-u.tv/
[9] http://www.uoh.fr/front
[10] http://www.ccsd.cnrs.fr/

5/ the Studio ASA

The re-purposing is the central activity of the *intentional manipulation of the source profile*, *the source identity of digital media objects* or of corpora of digital media objects in order to attune them to the expected profile of a user community and/or contexts of uses. For example, the use of available digital objects such as videos or images for educational purposes, presuppose in general a whole series of such repurposing tasks which may more or less drastically change the content of these objects, their "look", their purposes and goals.

Thanks to the already quoted French and European R&D projects – and especially thanks to the ANR founded *ASA-SHS* project (2009 – 2012)[11] – we have designed and implemented with our research team in Paris[12] an environment called – in French – Studio ASA (*Atelier de Sémiotique Audiovisuelle*, i.e. *Audio-visual Semiotics Workplace*) with the help of which we realize our different audio-visual archive projects as well as all of our video corpus repurposing activities. Figure 6 shows us in a nutshell the principal components of the Studio ASA.

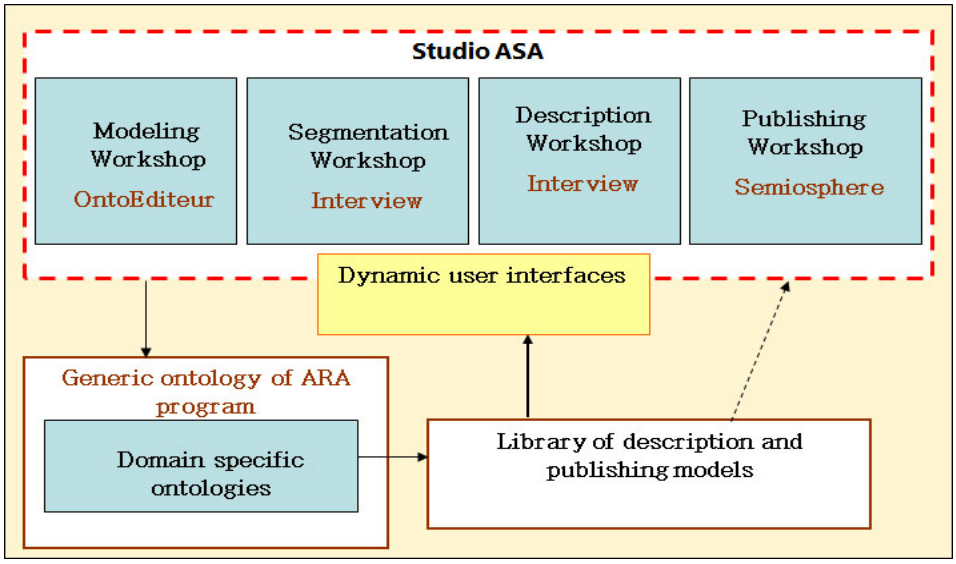

Figure 6: The *Studio ASA*

The Studio ASA is composed of four more specialized "workshops" (i.e. software tools together with user guides and examples of best practices):

1. a video processing workshop,
2. a video description, indexing and annotation workshop,
3. a video publishing workshop
4. and an archive discourse modeling workshop.

Each one of these workshops have been extensively discussed and presented in [STO 11a] and [STO 11b]):

---

[11] http://asashs.hypotheses.org/

[12] Cf. the web site of the ESCoM-AAR research team belonging to the French Fondation Maison des Sciences de l'Homme (FMSH) : http://www.semionet.fr

The *video processing workshop* provides the analyst (i.e. the person or group of persons working with digital media objects) with the technical means he/she needs namely for extracting from a given source video the relevant *segments* (or sequences) which he/she wants to analyze further and to publish/republish.

The *video description workshop* provides the analyst with a whole library of dynamic interactive formularies he/she needs in order to describe, index, enrich, and interlink a video or a part (a segment) of a video.

The *video publishing workshop* enables the user to publish or republish a video, parts of a video or a corpus of several videos previously indexed.

Finally, the *archive discourse modeling workshop* provides the competent user (the knowledge designer, the semiotician, …) with the technical and conceptual resources for designing and developing an archive specific *domain ontology* (a conceptual vocabulary), a domain specific thesaurus of predefined values or descriptors, as well as a library of *description models* which are indispensable for the elicitation of the universe of discourse of the corpus of videos composing the fonds of the archive.

The (software) tools composing these four workshops are provided with *dynamic formularies* based on a common meta-language called the *ASA meta-language*. This ASA meta-language is composed of:

1) an *ontology* (a vocabulary of concepts),
2) a *thesaurus* (predefined values for a subset of concepts)
3) and a *library of description models* (i.e. interrelated concepts or conceptual maps).

Some of the concepts (of the ASA ontology), values (of the ASA thesaurus) and description models are *shared* by all archive or archive projects designed and realized with the Studio ASA. These elements compose the *generic dimension* of the ASA meta-language. For instance, all audiovisual archive projects share a set of models for the description of the audiovisual expression of content in videos; all audiovisual archive projects share a set of models for specifying copyrights and other rights and duties with respect to the use/reuse of videos; and so on.

However, the content strictly speaking (the *subjects* or *themes*) vary from one archive to another. Hence, each archive project possesses also its *own, specific* conceptual vocabulary, thesaurus and library of description models. These elements form the *domain (or archive) specific dimension* of the ASA meta-language.

6/ a text- or discourse based vision of digital archives

The Studio ASA has been designed and developed with respect of what we call a *text- or discourse based vision* in digital archive production, management and exploitation.

Figure 7 summarizes this vision that constitutes the central investigation domain of the actually ongoing ANR funded Campus AAR[13] project (2014 _ 2017) of which a central aim is to provide any actor actively implied in the production and exploitation of digital audio-visual archives with a new and enhanced version of the Studio ASA.

[13] http://campusaar.hypotheses.org/

The upper case in figure 7 shows the typical activities that we summarize under the general label *text/discourse-based processing* (of digital media data, corpora of media data or "whole" archives). Text/discourse-based processing covers all *interactions* between a (human or artificial) *actor* and a digital media data, in our case, a digital audiovisual *source data* in order to transform it in something which possesses a relevancy (an interest, an added value, an "attractiveness") for the concerned actor.

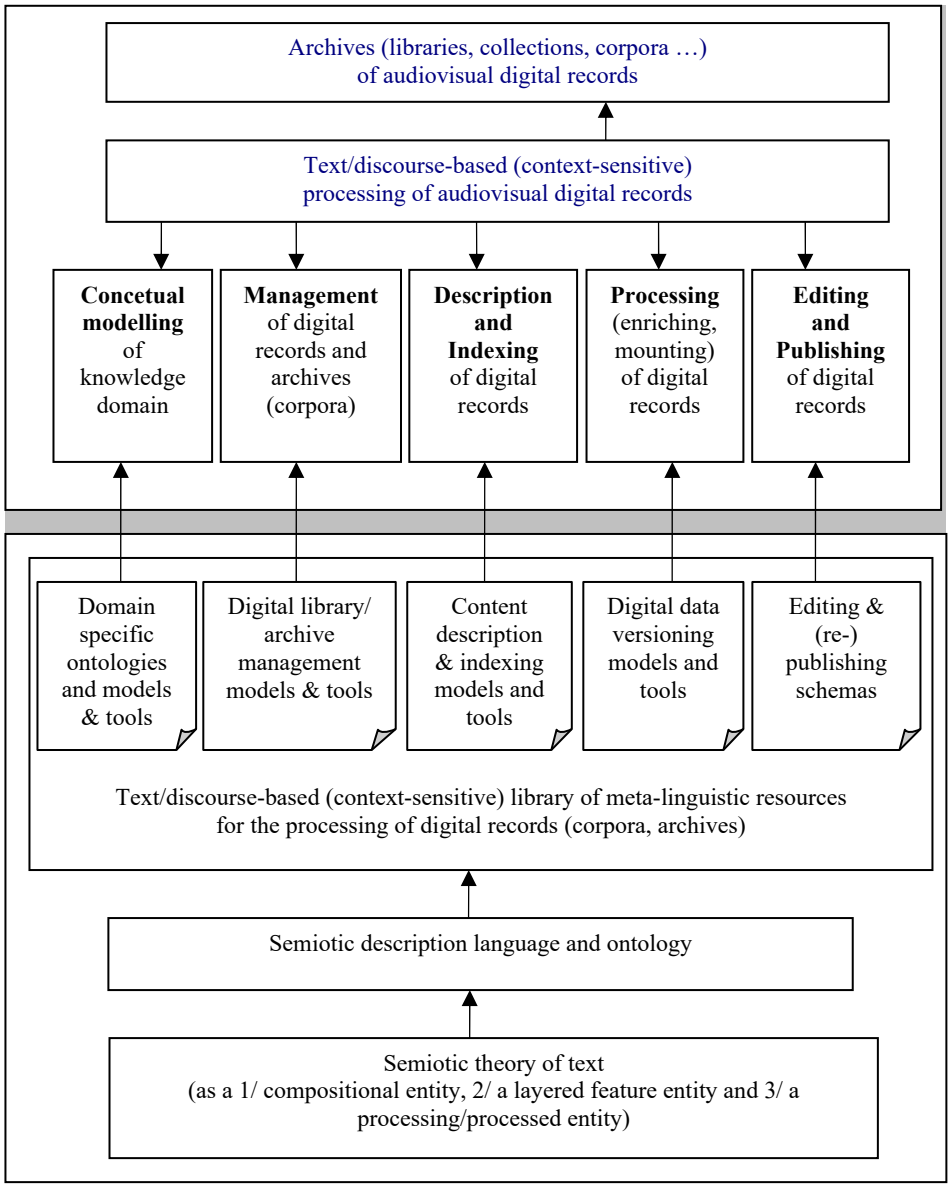

Figure 7: The domain of investigation of the Campus AAR project)

Indeed, we distinguish six main types of activities composing the realm of text/discourse-based processing of digital (audiovisual) records:

1. The *conceptual modelling activity*: its principal objective is to elaborate media processing, domain-appropriate description and indexing as well as (re-)publishing models.

2. The *corpus and/or archive management activity* is concerned with the creation/production and management either of corpora of digital media resources or of (more or less important) archives or libraries[14].

3. The *description* of the *media content* is concerned with the context-sensitive interpretation and indexing of digital media data or corpora of digital media data. The notion "media content" is a complex one and covers, among other features, a *topic feature* (a digital record "speaks" about what?) ; a *discursive function feature* (what is the place of the topic in an argumentation, a narration, a didactic exposé, etc.); a *rhetorical devise feature* (the topic, is it defined, exemplified, literally/metaphorically developed, …?); a *media expression feature* (what are the – visual, acoustic, … – medias used for the expression of the topic and what are the used expression methods ?); a *formal and physical organization feature* (where do you find a topic in a digital record?); etc. All these features determine the specific meaning structure of a digital media record.

4. The *processing activity* covers on the one hand the "*enriching*" of a given digital source record and on the other hand the *qualitative modification of the textual structure* itself of a source record (for instance, through a virtual montage adding a sound track, new filmic elements, … to an existing record). Together with the description/indexing activities, they contribute to the *versioning of a digital record*, i.e. to the *re-purposing*, the *adaptation* of existing media data with respect to the cognitive, cultural and linguistic particularities of a user or a community of users.

5. The editing and publishing activities, finally, intend to make accessible ((re-)described, (re-)commented, (re-)processed…) digital media data for a target public, via specific publishing genres (such as virtual narrative paths through collections of segments, video-dictionaries, bilingual folders …).

Figure 7 also shows us that no activity can be performed without the recourse to a specific type of intellectual or cognitive resources we summarize under the general label of *meta-linguistic* (or more generally speaking *meta-semiotic) resource,* this means of a language – a *meta-language* – we use for dealing with the object *text* broadly speaking. Such meta-linguistic (meta-semiotic) resources are for instance *ontologies* and (verbal, iconic, acoustic …) *thesauri*.

The lower case in figure 7 represents schematically the dependencies between semiotic processing activities of digital records and a library of functionally diversified meta-linguistic resources such as content description and indexing models, versioning models, publishing models, etc.

7/ some major traditions and trends in dealing with (digital) media data

In order to show the specific intellectual and scientific context in which we place our research activities on digital (audiovisual) archives in the sense of an integrated set of text- or discourse based activities, let us discuss very briefly some major traditions in the field of (digital) media object storing, processing, indexing and publishing.

[14] Cf. our explanations in [STO 11a] as well as my online paper Digital audio-visual archives, semiotics and digital humanities : http://www.academia.edu/5877963/Digital_audiovisual_archives_semiotics_and_digital_humanities

There have been, in the last (three) decades, many valuable contributions for organizing and explaining the semantic structure of (digital or non-digital) media data mainly in form of *controlled vocabularies*, *thesauri*, *classifications*, *ontologies*, *folksonomies*, *conceptual schemas*, *conceptual graphs* or again other *topic* or *cognitive maps* that are supposed to represent the cognitive structure of a knowledge domain. We may distinguish, roughly, between at least five main trends in this domain of research and development:

1. A first important tradition is "materialized" in a wide range of "meta-linguistic resources" such as (mono- or multilingual, domain specific…) thesauri, controlled vocabularies, terminologies, etc. which are largely used in the context of digital libraries and archives. All these resources represent kinds of *lexicalized visions* of knowledge domains. Well-known examples are, among many others, the *WebDewey* forming the "heart" of the *WordCat* – the worldwide network of libraries of the *Online Computer Library Centre* of Ohio, the UNESCO's *thesaurus*[15], Getty's *AAT – Art & Architecture Thesaurus*, the *RAMEAU* indexing language of the Bibliothèque Nationale de France[16], the Library of Congress Subject Headings (LCSH)[17], the *Social Science Thesaurus* of *the* Leibniz-Institut für Sozialwissenschaften *GESIS*[18], the *Humanities and Social Science Electronic Thesaurus* (HASSET)[19] of the UK Data Archive. But the structural and cognitive organization of a (digital) media data doesn't play any particular role in these and similar meta-linguistic resources.

2. A second trend has to do with the accelerated production of norms and standards in ICT – from basic ones such as the *Dublin Core Metadata Initiative*[20] or *MARC standards*[21] to specialized ones aiming at the "management" of high level cognitive activities in the work with and exploitation of digital data. Three important examples here are the PAIMAS and OAIS standards for archive projects[22], the e-learning standard *LOM* (for France, more particularly, LOMFR[23]) for dealing with pedagogically relevant data, and the *TEI* (*Text Encoding Initiative*)[24] dedicated to the representation (composition, formal organization, …) of electronic texts.

3. A third trend of research is dedicated to the automation processes of cognitive "high level activities". Important issues here are that of the *automatic segmentation* of audiovisual records (in smaller segments or "scenes"), the *automatic recognition* of visual or acoustic forms and figures in (audiovisual) scenes, the (for audiovisual archives) extremely important *speech-to-text conversion* (i.e. the conversion of spoken discourse into text), the *automatic semantic or conceptual indexing* based on statistical and/or linguistic procedures with the aim to map textual features ("words") to concepts or conceptual graphs representing the meaning of a given domain, the *automatic* or semi-automatic *production of domain specific terminologies* or *ontologies* based on textual corpora, and the *automatic translation* between *pairs of languages*.

4. A fourth and for our own research activities important trend covers (cognitive, linguistic, formal and applied) researches in knowledge description and representation. These researches have started in the 70ies and 80ies (cf. for instance the researches on

---

[15] http://databases.unesco.org/thesaurus/
[16] http://rameau.bnf.fr/
[17] http://www.loc.gov/aba/cataloging/subject/weeklylists/
[18] http://www.gesis.org/en/services/research/thesauri-und-klassifikationen/social-science-thesaurus/
[19] http://www.data-archive.ac.uk/find/hasset-thesaurus
[20] http://dublincore.org/
[21] http://www.loc.gov/marc/marc.html
[22] Cf. the explanations of Kari R. Smith on the blog "Digital Archives on the MIT Libraries": https://libraries.mit.edu/digital-archives/visualizing-paimas-and-oais/
[23] http://www.lom-fr.fr/
[24] http://www.tei-c.org/index.xml

semantic networks, topic maps or again *conceptual graphs* based on J. Sowa's[25] groundbreaking work) and are "materialized" to-day in the development of generic and domain-specific ontologies, tools and environments (cf., for instance, the Univ. of Stanford's Protégé environment).

5. The fifth trend of researches is next to our own theoretical assumptions and research objectives. It covers all those activities in the last 30 years that have aimed at a *systematic and operational understanding of the semiotic structure of texts broadly speaking* (i.e. the organization of content and its expression in form of mono-media, multimedia or again in form of cross- and trans-media objects). We think here 1) on structural and functional semiotics (i.e. mainly on the work of A.J. Greimas, R. Barthes and M.A.K. Halliday); 2) on linguistic, pragmatic and cognitive text and discourse approaches (among many different examples, we would like to mention here the researches in text linguistics of W. Dressler or T.A. van Dijk, B. Mann's and S. Thompson's persuasive *RST* (*Rhetorical Structure Theory*)[26] and 3) on applied researches in text and discourse comprehension and generation (cf., for instance, K. Mc Keown's seminal work on text generation based on rhetorical devises[27]).

8/ the structural organization of a data

Let us have now a closer look on the semiotic structure of a digital (media) data. As already stressed, the conception of the ASA Studio (figure 6) is based on structural (text and discourse) semiotics and *other related approaches* such as the rhetorical structure theory, discourse analysis, thematic and cognitive analysis of texts, knowledge representation … The development of the Studio ASA has followed, as precisely as possible the "instructions" formulated from the side of human sciences (in other words, we have tried to adopt the holistic vision called "humanistic computing" which stands for a human/social science based design of technical or technological systems[28]).

In order to understand and to deal adequately with *the semiotic structure of a (digital) media data* stored *in* an archive or *as* an archive (cf. figure 7), it is obviously *not enough* to deal only with the cognitive structure of a *given knowledge domain*: we have to take into account the structure, the organization of the data *in itself* given the self-evident fact that the knowledge of a domain is necessarily mediatized by the data that "contain" a knowledge of a domain, that produce it, that communicate it, that conserve it. And these data are (written or oral) *data*, *photos*, *drawings*, *maps, films,* and so on. And, as obvious too, all these textual "artefacts" (called simply *text* in the specialized literature) possess a highly specific and constraining structure.

Therefore, if we want to progress in our (theoretical as well as operational) understanding of the qualitative transformation of (digital) media data as the result or consequence of user-centric activities of adapting or repurposing "original" (or better given "source") data in order to transform them into genuine added value *cognitive* (or more generally, *epistemic*) *resources* for a target user community, the central problem is obviously the appropriate understanding of the *structural organisation of the digital data itself*.

From a structural or semiotic point of view, a digital media object (a digital text, film, image) can be characterised, described with respect to a set of *central features* such as the

[25] http://www.jfsowa.com/
[26] cf. http://www.sfu.ca/rst/
[27] http://www1.cs.columbia.edu/~kathy/
[28] Cf. for instance the programme of the International Journal of Social and Humanistic Computing : http://www.inderscience.com/jhome.php?jcode=ijshc

themes or topics encapsulated in a media data, the (linear or non-linear) development of a topic provided by a media data, the forms of (verbal and/or non-verbal) expression of a topic, and so on.

Like the grammar of a natural language, these features form together the specific identity or profile of a media data or of a corpus of media data. Such a profile or identity is not arbitrary but belongs to a tradition represented by *genres* or *cultural models* of production and sharing of meaning to which people refer and which people use in their own activities of writing, reading, exchanging information and knowledge.

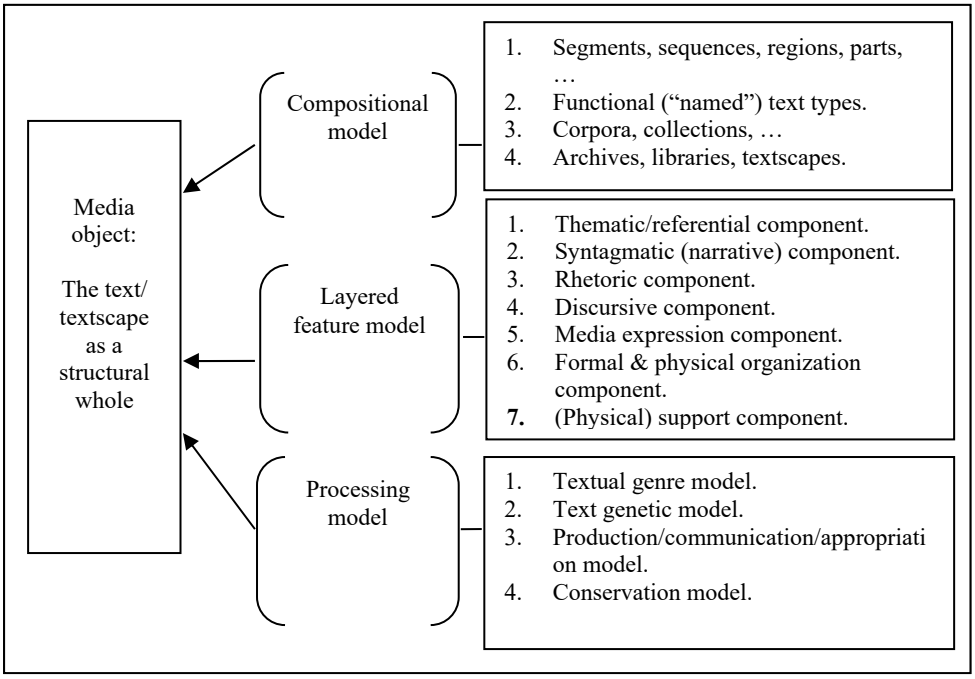

Figure 8: *The general structural model of a digital resource in a semiotic perspective*

Figure 8 shows the general structural picture of the (digital) media data understood as a *meaning loaded text* or *network of texts* ([STO 99], [STO 01], [STO 12]) which becomes manifest, basically, through three complementary aspects:

1. The *text as a compositional entity*: in principal, a text always can be decomposed in *smaller parts* (sequences, scenes, statements…) and it belongs always to *inter-textual fields* – archives, libraries, collections, etc. - forming more or less institutionalized or un-formal and ephemeral *textscapes* or *textspheres*.

2. The text as a *layered feature entity*: a text possesses a number of layers assuring the specific *information-to-meaning* processing: a *topical layer* ("what is the content of a text?"), a *syntagmatic layer* ("how information is merged in a text?"), a *media expression layer* ("what are the medias expressing an information?"), etc.

3. The *text as a processed entity"*: A text is an *historical entity* (it refers to traditions in form of genres), a *genetic*[29] *entity* (it evolves, for instance, from a simple draft or plan to a definitive product) and a *functional entity* (a text belongs to specific social practices and entities).

[29] "genetic" in the sense of the – philologically inspired - "genetic text analysis" (cf. for instance the works of the French research laboratory ITEM: http://www.item.ens.fr/

The structural organization of a text becomes manifest through its compositional nature (a text as a *stand-alone entity,* as a *whole composed of parts entity* or as a *part of a whole entity*), its nature as a layered feature entity (corresponding to a more precise formulation of the traditional Saussurian or Hjelmslevian distinction between *content* and *expression of content*) and as a processed/processing entity (as a *historical entity*, as a *text-genetic entity* evolving from a first sketch to some steady state version, as a *functional entity* integrated in a social practice).

9/ semiotics and digital archives

In order to conclude our article, we want to briefly systematize the role (or, more precisely, the potential role of semiotics in the field of digital archives.

Very generally speaking, semiotics can be defined as a theoretical framework and methodology for describing and analyzing ("expertizing"):

1. the production, exchange, sharing, conservation, reuse, … of **messages**
2. by the means of one or more **medias** and in form of **texts** (broadly speaking)
3. in using a **language** (the "semiosphere" peculiar to the culture of a social actor)
4. attuned to the specific the **context of use** ("social structure").

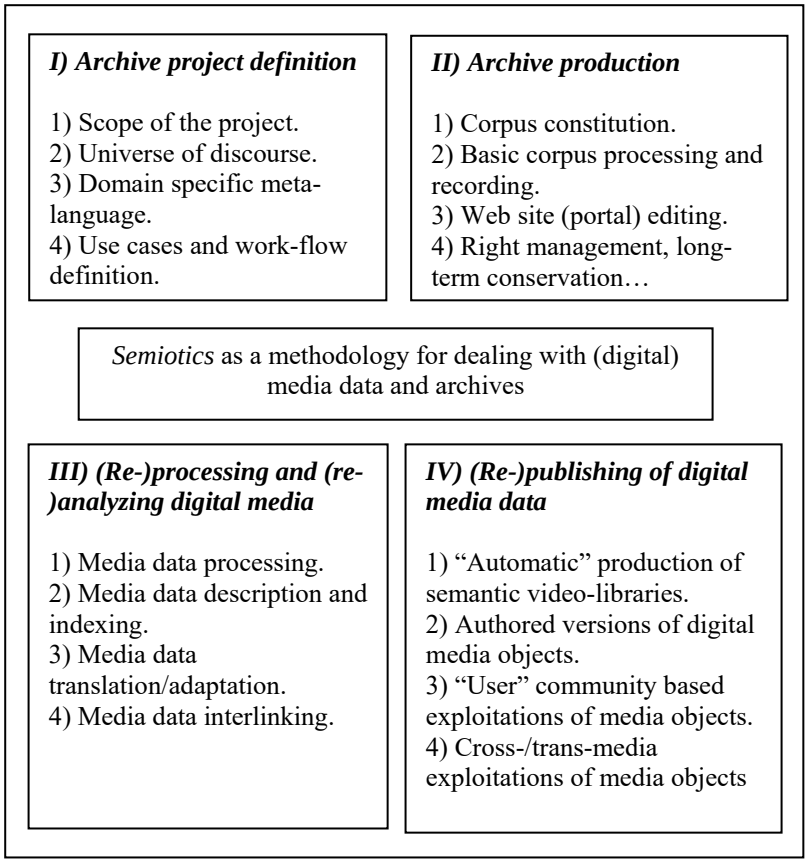

Figure 9: *4 major application domains of semiotics in the field of digital archives*

As well known, there are different traditions in semiotic research. One important tradition is represented by the structuralism in social and human sciences and structural linguistics going back to de Saussure, R. Jakobson, L. Hjelmslev, C. Lévi-Strauss, A.J.

Greimas, R. Barthes; a second one is represented by the functional text-semiotic approach of M.A.K. Halliday; a third tradition is represented by the cultural semiotics approach of J. Lotman and the Tartu school in Estonia; a fourth one is represented by the pragmatic and sign-theoretic approach of Ch. S. Peirce. It is beyond the scope of this small article to investigate and compare these several traditions and to assess their interest for the field of digital (audiovisual) archives. However, personally, I believe that in general semiotics can be used in several ways and for several objectives.

Figure 9 summarizes our vision of the central role of semiotic in the field of digital (audiovisual) archives and libraries that we have tried to implement partially in one of our previous R&D projects – the already quoted French ANR funded *ASA-SHS project*[30] – and that we continue to exploit in the actually ongoing *Campus AAR project*[31].

As suggested in figure 9, *one* important application of semiotics in this domain is its use as a methodology for a *digital archive project*: the definition of the *scope* of an archive project (= its domain and its context of use); the definition of the *universe of discours*e of the intended archive (archive-specific topics, rhetorical genres, auctorial visions, …); the specification of the appropriate *meta-language* (a domain-specific ontology, a domain-specific thesaurus, a domain-specific library of description models and, eventually, archive-specific publishing/republishing templates).

A *second* application domain for semiotics is its role as a methodological tool during the process of the *constitution, the production of a digital archive*: definition and production of an *appropriate corpus* of media data; the definition of basic *recording procedures and templates* of the media objects composing an archive; the design and editorial follow-up of the *web portal* used for diffusing the media data of an archive.

A *third* application of semiotics in the field of digital archives consists in its use as a methodology and guide for the *appropriate use* of the domain specific meta-language (library of description models) during the processing, description, indexing, commenting, versioning/translating, interlinking, … of media data composing the fonds of an archive. This possible role of semiotics includes, for instance, the writing of guides and "best practices" for analysts, the design of courses for future analysts, etc.

A *fourth* and important field of application for semiotics in the field of digital archives consists, finally, in its possible role as a methodological reference for the authoring (re-authoring) of digital media data with the help of (archive-specific) publishing/republishing templates. The role of semiotics here is to provide the users (publishers) with publishing guides and – once more again – "best practices" as well as with pedagogical resources for people intending to become publishers.

[30] http://asashs.hypotheses.org/
[31] http://campusaar.hypotheses.org/

BIBLIOGRAPHY

[CRA 08] CRAVEN, L. (éd.), *What are Archives. Cultural and Theoretical Perspectives : A Reader*, Ashgate, 2008

[ESE 10] EUROPEANA SEMANTIC ELEMENTS SPECIFICATION. EUROPEANA V1.0, EUROPEAN UNION 2010

[GRE 66] GREIMAS, A.J., *Sémantique structurale. Recherche de méthode,* Larousse, 1966

[GRE 79] GREIMAS, A.J., COURTES, J., *Sémiotique. Dictionnaire raisonné de la théorie du langage,* Hachette, 1979

[GRE 83] GREIMAS, A.J., *Du Sens II. Essais Sémiotiques,* Seuil, 1983

[ISA 04] ISAAC, A, et TRONCY, R. (eds.), *Designing and Using an Audio-Viusal Description Core Ontology*, Institut National de l'Audiovisuel (INA); en ligne 2004 (http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.97.9559&rep=rep1&type=pdf)

[SOW 84] SOWA, J., *Conceptual Structures : Information Processing in Mind and Machine*, Addison-Wesley 1984

[STO 99] STOCKINGER, P. : *Les nouveaux produits d'information. Conception et sémiotique du document,* Paris, Hermes Science Publications, 1999.

[STO 01] STOCKINGER, P. : *Traitement et contrôle de l'information,* Paris, Hermes Science Publications, 2001.

[STO 11a] STOCKINGER, P., (éd.): *Les archives audiovisuelles : description, indexation et publication*. Paris – Londres, Editions Hermes Science Publishing 2011 (trad. en anglais aux éditions John Wiley & Sons, NY)

[STO 11b] STOCKINGER, P., (éd.): *Nouveaux usages des archives audiovisuelles numériques*. Paris – Londres, Editions Hermes Science Publishing 2011 (trad. en anglais aux éditions John Wiley & Sons, NY)

[STO 12] STOCKINGER, P.: *Analyse des contenus audiovisuels. Métalangage et modèles de description* ; Paris/Londres, Hermes Science Publishing 2012 (350 pages) – traduction en anglais chez J. Wiley & Sons (NY, 2012)

[TFC 03] TSINARAKI, C., FATOUROU, E. et CHRISTODOULAKIS, S., *An Ontology-Driven Framework for the Management of Semantic Metadata describing Audiovisual Information*, MUSIC-TUC (Technical University of Crete), en ligne 2003 (http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.104.3374&rep=rep1&type=pdf)