Introduction

We discussed the issue of *analyzing* digital audiovisual corpora in [STO 11a] using a variety of concrete examples. We also examined in this book a working environment which enables us to analyze these corpora that are specifically adapted to their users' professional or personal needs.

Analyzing audiovisual corpora involves identifying the stages in work processes which define every digitizing project of knowledge heritages (i.e. scientific, cultural, etc.). This follows the (audiovisual) data collection stage which documents a "research terrain" and precedes the publication and dissemination of (analyzed, i.e. described, classified, annotated, interpreted etc.) data. Along with the technical processing ("cleaning", improving visual creation, etc.) and auctorial (i.e. mounting) stages of data collection and creation of field corpora, this is an unavoidable movement from transforming the status of virtually relevant digital data into a *potential resource* that is a potential "*advantage*" for the audience and its expectations, needs, and curiosities for various contexts of usage. In [STO 11a], various approaches are discussed which might define a concrete analysis and approaches such as:

– *stricto sensu* textual analysis, which consists of locating (identifying) only those passages in an audiovisual text that are genuinely relevant for an analysis project, presuming that for a given

Introduction written by Peter STOCKINGER.

analysis not the whole audiovisual text (no matter if it is a documentary or a simply “raw” recording of a cultural or scientific manifestation) is necessarily relevant. Once the passage(s) is/are identified, the *analyst* (a role given to an individual or group undertaking an analysis project) segments the audiovisual text, that is (virtually) extracts the identified passage(s), provides them with a provisory title, and records the temporal values corresponding to their beginning and end of the audiovisual text’s linear progression. The analyst can always return to this analysis to modify, for example, the beginning and end of the segmented passage, to suppress segmented passages, or redefine identified passages, and so on;

– *meta-description* of which the object is not so much the audiovisual text or the analyzed passage of an audiovisual text. The aim of meta-description, rather, is to explain the aim of the analysis itself: its authors, aims, the *area of expertise* (i.e. the area of knowledge concerned) or even the *genre* that it represents (basic versus detailed analysis, an excerpt versus the whole audiovisual text, overall content versus visual and acoustic patterns, analysis as linguistic adaptation of the audiovisual text or part of it). Meta-description also defines its authors rights as well as the rights for its use and exploitation by an interested public (in a new analysis project, for example);

– *paratextual* description initially aims to specify the formal identity of the audiovisual text and/or each identified and virtually extracted excerpt. This involves identifying indicative information related to the author(s), directors, producers, and so on, of the audiovisual text (or an excerpt of it). This description enables us to identify the genre(s) which the audiovisual text represents and also identify the important times and places in the “life” of the text in question, for example the time and date of production, publication and dissemination, latest update, and so on. Another important point of the paratextual description is that of the explicit rights that govern ownership of the audiovisual text (or a particular passage) and of its various uses by the interested public;

– *audiovisual* description specifically focuses on analyzing the audiovisual patterns which are composed of the text (or a specific passage which was identified and virtually extracted). The focus of audiovisual description, for example, ranges from recording of different frames in a filmed event to the sounds accompanying it or synchronization between "acoustic" and "visual" patterns;

– *thematic* description involves explaining the content conveyed by an audiovisual text which is being analyzed (or an excerpt of this) including thematized events or situations (i.e. the domain to which the text refers), discursive thematization strategies (i.e. the perspective in which an event or situation is being examined, the progressive development of a theme in the text etc.);

– *pragmatic* description – in the wider sense, has three objectives: 1) highlighting the potential interest, the potential value of a text (or a specific passage) for an audience and a context of use, 2) enriching the text (or excerpt of it) (via commentaries, bibliographies (including Web) suggestions etc.) in order to adapt the text to the needs and interests of a targeted audience and the specific constraints of a specified context of use, 3) producing, if necessary, an appropriate linguistic version of the audiovisual text (or an excerpt of it) for an audience that cannot understand, or fully understand, the text's language of production (in general, this involves text translation (not necessarily a literal or "faithful" translation) or producing a *linguistic version* of the text which is better adapted to the sociolinguistic register of the audience's language. Note that the pragmatic description ends where technical and authorial processing of an audiovisual text or corpus begins: pragmatic description, specifically analysis in general, does not focus on the audiovisual text's "materiality". Modifying the audiovisual text's "materiality" involves processing audiovisual corpora from a purely technical perspective ("trimming" collected audiovisual data, cleaning files, improving acoustic or visual quality etc.) or from an authorial perspective (displaying audiovisual data according to a specific scenario, with additional music and voice overs (production technique where voice is not part of the narrative), post synchronization of visual and sound patterns etc.) .

To enable everyone to carry out "their" analysis projects with this variety and richness of approaches, a small group of researchers and engineers working at ESCoM (*Equipe Sémiotique Cognitive et Nouveaux Médias* [Cognitive Semiotics and New Medias Lab]) at

FMSH (Fondation Maison des Sciences de l'Homme)[1] in Paris have developed a sophisticated digital work environment called the *ASW Studio* in reference to the ASW-HSS (*Audiovisual Semiotics Workshop for analyzing corpora in Humanities and Social Sciences*)[2]. The ASW Studio is composed of several specialized workshops: the *Segmentation Workshop* (virtual) for audiovisual data, the *Description Workshop*, the *Publishing Workshop*, and the *Modeling Workshop* for the metalinguistic resources necessary for carrying out the analysis/description of the audiovisual data. The analysis of audiovisual corpora is carried out in the segmentation and description workshops presented in further detail in [STO 11a].

This book will develop the analysis of audiovisual corpora in further detail in terms of the *analysis projects*. Like any *project*, an analysis project of audiovisual corpora follows precise objectives, stages, and so on, (e.g. defining the project, analyzing the needs and project related audiovisual information that already exists, carrying out the project etc.) and is led by a person or team working in a given framework (social, community, institutional etc.).

We will examine the three types of analysis for audiovisual corpora which, in one way or another, explore new aspects and perspectives of production, dissemination, sharing, and enriching knowledge in an entirely digital context.

The first type of work relies on the creative reuse of audiovisual corpora which have often (but not necessarily always) been already analyzed and published (on a Web portal for example) to submit them to a new cycle of analysis and publication with a view to create new products or services from it (see [STO 99]) for specific audiences or fixed usage. This is carried out in what is presently known as

[1] "House of the Sciences of Man", (http://www.msh-paris.fr/en/foundation/missions).

[2] ASW-HSS (*Audiovisual Semiotics Workshop for analyzing corpora in Humanities and Social Sciences*) is a research project by ESCoM/FMSH financed by the ANR (Agence Nationale de la Recherche [National Research Agency]) in France. Its reference number is: *ANR-08-BLAN-0102-01*. ASW-HSS began in January 2009 and will officially end in December 2011. For further details, visit the official Website of the ASW-HSS project: http://www.ASW-HSS.fr/

repurposing or even *document reengineering*. Chapter 1 presents an analysis project focusing on a corpus which has already been published online based on *A Thousand and One Nights*[3]. Chapter 2 is dedicated to analyzing and re-analyzing corpora which have been partially published online focusing on the traditional production of bread in France and Portugal[4]. Chapter 3 examines different types of publications/republications which have proven particularly useful for experiments carried over a number of years at ESCoM in its Research and Development "Archives Audiovisuelles de Recherche" (Audiovisual Research Archives) program[5] (see [STO 11a] for further details), with support/help from a series of European and French research and development projects[6].

The second type of research focuses on using the *ASW Studio* for creation of digital audiovisual and specialized geographically thematized corpora, as well as for analyzing and publishing (or republishing) data composing these archives. Chapter 4 describes an area of experimentation based on constructing and using Andean Quechuaphone communities. This area of experimentation is part of the European research and development program *Convergence*[7]

[3] The audiovisual corpora are sourced from the AAR program ("Archives Audiovisuelles de la Recherche") [Audiovisual Research Archives] from the ESCoM Research and Development program, presented in further detail in [STO 11a]; see also the program's official Website: http://www.archivesaudiovisuelles.fr/EN/

[4] Parts of this corpus are also published on the AAR portal.

[5] The official site can be found at: http://www.archivesaudiovisuelles.fr/EN/

[6] This consists primarily of three projects; SAPHIR, LOGOS and DIVAS. SAPHIR ("Système d'Assistance à la Publication Hypermédia" [Assistance System for Hypermedia Publication]) is a French research project financed by the French ANR (Agence Nationale de la Recherche [National Research Agency]) and coordinated by the INA Recherche which began in 2006 and finished at the end of 2010. LOGOS is an acronym for *"Knowledge on demand for ubiquitous learning"*, a European research and development project financed in the 6th PCRD which began in January 2006 and finished in February 2009. DIVAS is an acronym for "*Direct Video & Audio Content Search Engine*" which is a European research and development project, which was also financed in the 6th PCRD beginning in January 2007 and ending in February 2009. For further details, see the glossary at the end of this book.

[7] *Convergence* is a European research and development project which will run from June 2010 to February 2013. It is coordinated by the CNIT (Consorzio Interuniversitario per le Telecomunicazioni [Interuniversity Consortium for Telecommunication]) in Rome and financed by the 7th PCRD, No. FP7-257123). The aim of the *Convergence*

developing, among other things, technologies which enable us to trace all uses of digital data (such as a video) over the Internet. Chapter 5 describes an area of research in the ASW-HSS project based on creating an archive dedicated to cultural heritages (in this case, that of Azerbaijan).

The third type of research focuses on analyzing and publishing projects for audiovisual corpora using new possibilities presented by social networks, Web 2.0 or even mobile communications to better circulate, share and enrich previously analyzed and published audiovisual extracts. Chapter 6 examines case studies for publication and sharing of scientific information via *Facebook* and *Twitter.* Chapter 7 examines the importance of using dissemination platforms for digital data such as *YouTube*, *DailyMotion*, or *Vimeo*. Chapter 8 demonstrates how to use, in an analysis/publication project, "aggregators of Web 2.0 content" such as *Netvibes* or *scoop.it* or even research communities such as Louvre.fr. Finally, Chapter 9 explores the importance of "usage tracing" technology for digital data developed in the *Convergence* project by developing in a highly detailed and technical way, the area of experimentation examined in Chapter 4, including the diffusion of sensitive audiovisual content which documents the intangible heritage of Quechuaphone communities in Peru and Bolivia.

To conclude this examination, it should be emphasized that this book is the product of collective and interdisciplinary work converging "fundamental" and applied research, informatics, and human social sciences (specifically those of semiotics and linguistics). This has been led over a period of 10 years by a small team of researchers and

project is to enrich the Internet with a new *publication-subscription* service model focused around *content* and based on a shared container for each type of digital data, including people and real world objects (RWOs). This shared container, called a *Versatile Digital Item* (VDI) is a structured set of digital data and metainformation, identified in isolation (i.e. such as the URL on a Web page) and which includes the concept of a *Digital Item* defined by MPEG-21. The significance of ESCoM and the ANR program in this project relates to the fact that all uses of a video online can be traced via VDI technology. This opens up the possibility of circulating digital content and appropriating those that respect the rights of authors and owners. The official *Convergence* site can be found at: http://www.ict-convergence.eu/

engineers who have also written this book and [STO 11a]. I would like to formally acknowledge all of them and convey my regards to them.

Throughout the past 10 years of research and development, our team has benefited from the support of a number of friends and colleagues from France and elsewhere. In particular, we thank the following people: Patrick Courounet, Steffen Lalande, Abdelkrim Beloued, Bruno Bachimont (INA Recherche); Jocelyne and Marc Nanard (CNRS-Lirmm); Marie-Laure Mugnier, Michel Chein, Alain Gutierrez (CNRS-Lirmm); David Genest (University of Angers-Leria); Danail Dochev, Radoslav Pavlov (Académie des Sciences de Bulgarie); Stavros Christodoulakis, Nektarios Moumoutzis (Université technologique de la Canée);

Finally, we profusely thank Muriel Chemouny (FMSH-ESCoM) for having reviewed the contributions to this book and Elisabeth de Pablo (FMSH-ESCoM) for creating the manuscript.