Department of Informatics
University of Fribourg, Switzerland

A FRAMEWORK FOR STRUCTURING MULTIMEDIA ARCHIVES AND FOR BROWSING EFFICIENTLY THROUGH MULTIMODAL LINKS

THESIS

Submitted to the Faculty of Science, University of Fribourg (Switzerland)
to obtain the degree of Doctor scientiarum informaticarum

Maurizio Rigamonti
from
Lugano, Switzerland

Thesis N° 1587
Imprimerie Saint-Paul, Fribourg
2008

Accepted by the Faculty of Science of the University of Fribourg (Switzerland), on the recommendation of:

- Prof. Ulrich Ultes-Nitsche, University of Fribourg, Switzerland (Jury president)
- Prof. Rolf Ingold, University of Fribourg, Switzerland (PhD director)
- Prof. Jean-Daniel Fekete, University of Paris-Sud, France (External expert)
- Prof. Jean-Marc Ogier, University of La Rochelle, France (External expert)
- Dr Denis Lalanne, University of Fribourg, Switzerland (Expert)

Fribourg, January 15th, 2008

PhD Director	Faculty Dean
Prof. Rolf Ingold	Prof. Titus Jenny

Acknowledgments

Cette thèse est la conclusion d'un grand voyage, qui m'a permis de rencontrer beaucoup de personnes. Celles-ci ont influencé d'une manière ou d'une autre ma vie pendant ces années de thèse. À elles s'adressent donc mes remerciements.

Un primo pensiero va ai miei genitori, a mia nonna, a mio fratello e a tutta la mia famiglia. Mi hanno sempre sostenuto, anche quando avevano soltanto una vaga idea di cosa stessi facendo.

Merci à mes superviseurs : à Rolf Ingold, qui a su canaliser mes excès d'énergie et de caractère, et à Denis Lalanne, qui m'a appris beaucoup et qui est devenu un ami.

Merci au jury de thèse. À Ulrich Ultes-Nitsche pour l'avoir présidé. À Jean-Daniel Fekete et Jean-Marc Ogier pour avoir accepté d'expertiser ce travail. Sans oublier la brillante discussion qu'ils ont animée après ma défense privée, pendant le souper chez Rolf.

Merci aux responsables du projets IM2 (http://www.im2.ch), et donc au Fonds National Suisse (http://www.snf.ch), pour avoir soutenu cette thèse.

I would like to thank Patrick Chiu, Mike Christel, Berna Erol, Marti Hearst, Paul Janecek, Ka-Ping Yee, Agnès Lisowska, Stefan Rüger, and Jacqueline Hansson of IEEE for permitting me to reproduce the figures in Chapter 2.

Merci à tous les anciens et nouveaux collègues de la bande DIVA. À Oliver Hitz, qui m'a accompagné à Paris pour un premier voyage dans le monde de la recherche. À Jean-Luc Bloechle, avec qui j'ai créé Xed. Et fait les 400 coups. À Florian Evéquoz, qui a travaillé efficacement au développement d'Inquisitor. À Catherine Pugin, qui m'a encouragé à porter un costume pour ma défense privée. Et à tous les autres, qui mériteraient d'être cités ici pour plusieurs raisons.

Merci à mes amies et amis, avec qui j'ai partagé la vie de la nuit, faite de lumières évanescentes, de fantasmes éphémères et de folies bien trop concrètes. Énumérer toutes ces personnes ici est vraiment impossible.

E infine grazie a mio nonno, che sarebbe fiero di leggere queste righe.

Abstract

This thesis proposes a method for indexing and browsing archives of multimedia documents, and in particular meeting recordings, using printable documents and links. Existing systems for indexing and browsing multimedia data have four main limits. First, the indexing requires high-level abstractions extracted from multimedia documents, which is still an unsolved problem for rich media such as images or videos. Second, existing systems do not take into account the correlations between multimedia documents, but manage them as isolated documents. Third, users are only weakly involved within the indexing process. Fourth, users can search specific documents, but rarely can browse an archive.

In this thesis, we propose a methodology acting at three levels, as solution to these limits. At the first level, printable documents are annotated with high-level information. The user can either supervise an automatic analysis process or add supplemental information. At the second level, groups of correlated multimedia documents, e.g. belonging to a meeting, are aligned in order to elicit their relationships. The latter can be represented with temporal links, thematic links, etc. The users can validate the links created automatically. At the third level, all multimedia documents are aligned. At this stage the information is indexed thanks to annotations, whereas calculated links are stored. Consequently, this solution enables users to search and browse all types of multimedia documents archives.

This methodology has been integrated in a system. Its architecture is flexible and can be easily extended. At the first level, a novel analysis tool automatically annotates PDF documents with their structural information. The annotations can be validated and extended using a graphical user interface. At the second level, we apply multimedia alignments and lexical analysis for creating links between all the multimedia documents of a meeting. A relational indexing system has also been integrated: it structures meeting archives, by indexing textual information and by storing links. Finally, two browsers have been developed for respectively replaying a meeting or browsing an entire collection of multimedia documents.

The integration of an indexing and browsing system for meeting archives validates our model. Furthermore, various user evaluations have assessed the usability of links and the usefulness of printable documents for browsing a meeting. For this purpose, two evaluation methods have been set up, for evaluating respectively the individual modalities and the components of a meeting browser.

The main contribution of this thesis is a model for structuring archives of multimedia documents, validated with meeting collections. Our model is centered on the transfer of information from printable documents to other multimedia documents and integrates the user at each level. Various tools have also been integrated and evaluated, validating our technology for indexing and browsing multimedia archives. In the future, the model presented in this thesis could be applied to other collections of multimedia data, such as personal information or television news.

Résumé

Cette thèse propose une méthode pour indexer et explorer des archives de documents multimédias, en particulier des enregistrements de réunions, grâce à des documents imprimables et à des liens. Aujourd'hui, les systèmes d'indexation et d'exploration de données multimédias ont quatre limitations majeures. 1) L'indexation requiert des abstractions de haut niveau extraites des documents multimédias. Le problème demeure irrésolu pour des médias riches comme l'image et la vidéo ; 2) les systèmes actuels ne considèrent pas les corrélations entre documents multimédias, qui sont plutôt traités comme des documents isolés ; 3) l'utilisateur est souvent peu impliqué dans le processus d'indexation ; 4) les utilisateurs peuvent chercher des documents spécifiques, mais peuvent rarement explorer une archive sans requête précise.

Dans cette thèse, nous proposons une méthodologie sur trois niveaux comme solution à ces limitations. Au premier niveau, les documents imprimables sont annotés avec des informations de haut niveau, tels que le contenu et leurs structures. L'utilisateur peut soit superviser le processus automatique d'analyse, soit éditer de l'information supplémentaire. Au deuxième niveau, des groupes de documents multimédias corrélés, par exemple appartenant à une réunion, sont alignés pour découvrir leurs relations, qui peuvent être représentées avec des liens thématiques, temporels, etc. Les utilisateurs peuvent valider ces liens créés automatiquement. Au troisième niveau, l'information est indexée grâce aux annotations et les liens sont sauvegardés.

Cette méthodologie a été intégrée dans un système à l'architecture flexible et extensible. Au premier niveau, un outil d'analyse automatique annote des documents PDF avec leurs informations structurelles. Les annotations peuvent être validées et étendues avec une interface graphique. Au deuxième niveau, des alignements multimédias et lexicaux créent les liens entre tous les types de documents multimédias dans une réunion. Puis, un système d'indexation structure les archives de réunions, en indexant l'information textuelle et en sauvegardant les liens. Enfin, deux navigateurs permettent respectivement de rejouer une réunion et d'explorer une collection entière de documents multimédias.

L'intégration d'un système d'indexation et d'exploration pour des archives de réunions a permis de valider notre modèle. De plus, plusieurs évaluations utilisateur ont validé l'utilité des liens et des documents imprimables pour explorer les enregistrements d'une réunion. Deux méthodes d'évaluation ont été ainsi définies, pour évaluer respectivement les modalités individuelles et les composantes d'un navigateur de réunions.

La contribution principale de cette thèse est un modèle pour structurer des archives de documents multimédias, validé avec des collections de réunions. Notre modèle est basé sur la liaison automatique entre des documents imprimables et d'autres médias et il intègre l'utilisateur à tous les niveaux. Plusieurs outils ont été intégrés et testés, en validant notre technologie d'indexation et d'exploration d'archives multimédias. Dans le future, le modèle présenté tout au long de cette thèse pourrait être appliqué à d'autres collections de données multimédias, comme par exemple des informations personnelles ou des journaux télévisés.

Contents

Chapter 1

Introduction

During the last decades, text was the standard form for exchanging digital information between people, either in websites or in electronic documents. At best, textual documents were enriched with graphics and images, e.g. charts, tables, and photos. Recently, a new type of communication raised up: the production and dissemination of multimedia information, in which text, graphics and images can be coupled with different media streams, such as video and audio. The word "multimedia" is composed of Latin suffix "multi" and of the plural of noun "medium". A medium is a vector of information, i.e. a document such as videos, audio, images, or textual documents. Thus, the word "multimedia" indicates a group of heterogeneous and coexisting media. The term "multimedia" is widely used in different contexts, e.g. information technologies, art, advertisement, etc., but its definition is ambiguous. For instance, if the term indicates the coexistence of different streams of information, a newspaper containing both text and photos could be either a single medium or a multimedia document. In this thesis, we consider multimedia documents as data made of streams that might be both recorded and reproduced by different physical devices. Therefore, according to this assumption, a newspaper with photos is a unique medium.

Multimedia data are characterized by several peculiarities and advantages: they are rich in content, time-based, and today can be produced with low-cost devices. These characteristics are only a part of the benefits that have encouraged the rapid diffusion of multimedia information in several aspects of everyday life.

A recent trend consists in recording lectures, conferences, and meetings. Preparing and recording these events produce a huge amount of multimedia information: textual documents and slideshows presented by speakers, audio recordings of the dialogues, videos of participants, websites, and so on. In fact, multimedia documents preserve information that is hard to represent or consult in the classical printable documents. For instance, a lecture contains more than the script distributed to the students: it is also composed of professor's speech, notes, and his gestures. Similarly, the spoken dialogues, persons' emotions, and the discussed textual documents are part of meetings and conferences' information.

The production and dissemination of multimedia information involve the collaboration of several processes, such as the recording, the processing of multimedia content, and the access to the structured information. These processes are defined in Section 1.1, which sketches the requirements of a system for managing multimedia collections. Defining such a system is a complex task, because it must coordinate diverse technologies developed in various fields of computer science, e.g. document or speech recognition, information retrieval, information visualization, and so on. Furthermore, these technologies must be considered in an integrated process, rather than independently. Their limits are detailed in Section 1.2.

In this thesis, we propose a novel solution for structuring and accessing archives of multimedia documents. Our solution take into account the limits of existing technologies, which are organized in order to mutually compensate their respective lacks. More precisely, we propose a complete system for organizing multimedia collections of meetings and various interactive visualizations for reproducing them.

1.1 Requirements for Organizing Multimedia Data Collections

The main processes involved in multimedia information management are recording, mining, indexing, retrieval, and replay.

The *recording* of multimedia data consists of capturing and storing the information using camcorders, microphones, scanners, electronic ink, and so on. The documents synthesized for an event, for instance textual documents, slideshows, and images, also belong to the recorded data.

Structuring this recorded information is a non-trivial task that needs the understanding of documents' content, the classification of data in respect of the message they contain, and a mechanism for accessing the arranged information. These requirements are focused in three disciplines of the computer science: the data mining, the indexing and the retrieval. The *Data mining* consists in extracting useful and relevant information from the original document, composed of one or more media. Mined information can be high-level abstractions such as semantic and temporal information, summaries, structures, etc. Extracted information is further involved with *indexing*, which associates one or more indexes to the multimedia document, in order to classify it according to its content. The *retrieval* aims to match the information of the indexed archive with a query submitted by a user.

Finally, when a user retrieves multimedia information, he can *replay* it: recorded and structured multimedia data can be reproduced on physical devices such as screens and loudspeakers. The replay takes into account how to present the data to the users and proposes functionalities for interacting with the displayed multimedia documents.

Nowadays, search engines such as Google widely use techniques of mining, indexing, and retrieving to organize huge collection of multimedia information. Similarly, these techniques are required for structuring and accessing large collections of lectures, conferences and meetings recordings.

1.2 Limits of the Technologies for Multimedia Data Organization

The amount of produced multimedia information is nowadays considerable, but the necessary technologies for its organization and redistribution are not efficient. For instance, Google, the most powerful existing search engine, only permits to search either textual documents or labelled images. Moreover, users can not search these media at the same time. This lack is provoked by the limitations of technologies for mining, indexing, retrieving, and replaying that are still in their infancy for some categories of media, such as video and audio recordings. Most of these technologies have been developed in monomodal[1] context and are affected by the following drawbacks:

Unequal Quality of High-level Abstractions

The monomodal techniques for mining information extract high-level abstractions from documents, describing more or less precisely the meaning and the message of the medium. These differences can be critical when indexing and retrieving multimedia documents.

Systems analyzing videos and images are tuned for recognizing precise thematic contents, i.e. persons, locations, objects, etc. Recognized information can be associated to text, permitting to index those media. Unfortunately, since systems are specialized for only detecting well-defined information, the content of videos and images is not fully indexable. Moreover, in the case of meetings, most videos are focused on the speakers during the whole recording and, thus, the document is hard to segment in thematic scenes. This lack of structures implies that the indexing will be less precise.

On the other hand, still in the context of meetings, the recordings of dialogues can be transcribed through automatic analysis techniques. This high-level information is rich in thematic contents, but at same time it can be redundant and weakly structured.

To wrap up, when thematic information does not describe the full content of multimedia documents, the amount of indexes is limited. Consequently, the documents are difficult or impossible to retrieve by users. Similarly, the lack of structures does not allow to identify the exact parts of document containing the searched information. High-level information, which evenly describe thematic and structural contents for each category of multimedia documents, improve both the indexing and the retrieval mechanism.

[1] The "monomodal" term indicates the use of a unique modality, such as the voice, the vision and so on. The coexistence of modalities is expressed by "multimodal" word. The "monomodal analysis" aims to extract from a document a well-defined modality of human communication, i.e. the gestures or gazes from videos, the speech from audio, etc.

Correlations between Multimedia Documents Ignored

In general, analysis methods do not take into account the correlations between multimedia documents. For instance, mining of newspapers and newscasts containing both audio and video is accomplished by methods that take into account a unique medium a time. Thematic, temporal, and geographic correlations between these documents are frequently ignored. Eliciting the relationships between several media can improve the efficiency of the individual mining techniques, by correcting and mutually completing the high-level abstractions extracted from each medium. Furthermore these relationships can be used as indexes for facilitating the organization of multimedia data collections.

Limited Role of Users

The technologies involved with multimedia organization either ignore the users or do not sufficiently explore the opportunities of interaction between them and the automatic systems. Users can be implied in the organization of multimedia data collections, by validating the results of mining techniques and by creating high-level information useful for the indexing. From the retrieval point of view, the current trend consists in improving the performances of automatic methods for organizing and accessing data. For instance, indexing engines try to detect the most relevant documents and to define rankings of documents matching a query. Although these functionalities are useful and interesting, we believe that retrieval mechanisms shall more and more involve the users, because persons are able to interpret information better than computers. For instance, a good visualization of the data allows to present hundreds of documents to the users, which are able to find the expected information. Thus, the future retrieval systems must offer to the users the opportunity of exploring either the whole collection of documents or a consistent part of it, instead of automatically proposing the documents that are supposed to be the most interesting.

1.3 Goals of This Work

This PhD thesis proposes a novel approach for organizing and reproducing multimedia data recordings. The approach is an attempt at overcoming the limits enumerated in the previous section. In particular, it tackles three main challenges:

Textual Documents as Structuring and Thematic Vectors

Although textual documents are less rich of contents than media such as video and audio, most of their mining techniques are reliable and provide structured and thematic information. The latter is well-suited for indexing mechanisms. In this thesis, we propose to use the high-level information extracted from textual documents to complete that of other media, more difficult to mine. Completing the high-level information enriching media such as video, audio and images will enhance their indexing.

Cross-media Alignments

The relationships between multimedia documents can be thematic, as well as temporal, geographic, artificially defined by users in respect of their perception, and so on. In this thesis we propose to elicit and extract these correlations, in order to use them as indexes for structuring a meetings collection. Furthermore, we propose to develop interactive visualizations that exploit the relationships between documents as main artifact for navigation.

Information Visualization to Complement Mining

Finally, we propose to represent the meetings collection using techniques of information visualization. These methods allow to display a huge amount of multimedia documents and to elicit the structure of the collection. Moreover, the interactive visualizations allow to integrate new searching and browsing mechanisms, which are not used in existing search engines such as Google. These mechanisms also imply the development of new techniques for information retrieval.

1.4 Structure of This Thesis

In addition to this introductive chapter, this manuscript is composed of six other chapters.

Chapter 2 is a state of the art that presents the limits and peculiarities of recent related works, which deal with the organization and navigation in multimedia archives. Four fields of research are involved in this thesis and are presented in four respective sections: multimedia data management, events indexing and browsing, methods for analyzing textual documents and visualization techniques. A fifth section illustrates the scientific contribution of this thesis relative to the presented domains.

Chapter 3 proposes a generic model for multimedia data representation. The general approach of our method is explained and all its actors are described in details. We further present an instance of the model applied to meetings collections.

Chapter 4 presents the whole analysis chain for a collection of events, e.g. of meetings. The first step consists of automatically analyzing a single medium. In our approach, we focus on the analysis of printable documents in PDF format. After analysis, a specific tool allows the users to correct these results. Next, the chapter explains how the results of individual media analysis are involved in the analysis of an event. In particular, this analysis elicits the hidden correlations between multimedia documents belonging to an event. Again, users should validate automatically calculated results through specific interactive tools.

Chapter 5 is dedicated to the indexing of events collections. The indexing follows and takes full benefit of the results obtained by events analysis. Firstly, this chapter introduces the full architecture of the indexing system and then it details the different components. In particular, it explains how to import events recordings, to remove multiple instances of identical documents, to index multimedia documents and to align them. Multimodal alignment is a technique that discover the thematic and temporal relationships between multimedia documents. Moreover, this chapter presents techniques for automatically extracting keywords from multimedia documents, for creating a ranking and for clustering similar documents. The querying and retrieval mechanisms are as well explained. Then, the chapter discusses users' opportunities for manipulating and updating the data indexed by our system. Finally, this chapter summarizes the performances of the system, obtained while automatically importing two collections of meetings.

Chapter 6 illustrates the browsers that are proposed to the users to search and browse into the events collections. The first browser is *JFriDoc*, which has been developed to browse through a single event and which explores new visualization and interaction paradigms for managing synchronized multimedia documents. The second browser is *FaericWorld* that allows to visualize an entire collection of events, to browse within, and to validate its automatically calculated indexes. The navigation task is insured thanks to a new browsing paradigm, which takes benefit of the elicited relationships between multimedia documents. Chapter 6 also presents the dataflow and architecture of the system for indexing and browsing meeting collections. Finally, the chapter lasts with three JFriDoc user evaluations, which target at measuring the efficiency of alignments for searching information, of textual document modality for navigating in a meeting, and of the interactive components composing the browser.

Finally, Chapter 7 concludes this thesis. After a summary of the thesis, the chapter analyzes the limits of our techniques and proposes future extensions and potential improvements.

Chapter 2

State of the Art in Multimedia Indexing and Management Technologies

Multimedia data production has drastically augmented in the last decade. On one hand, low-cost devices such as web-cameras and multimedia mobile phones are more and more powerful and accessible to everybody. On the other hand, storing device capacity as well as Internet's speed have significantly augmented and people are being stimulated to produce and redistribute multimedia information.

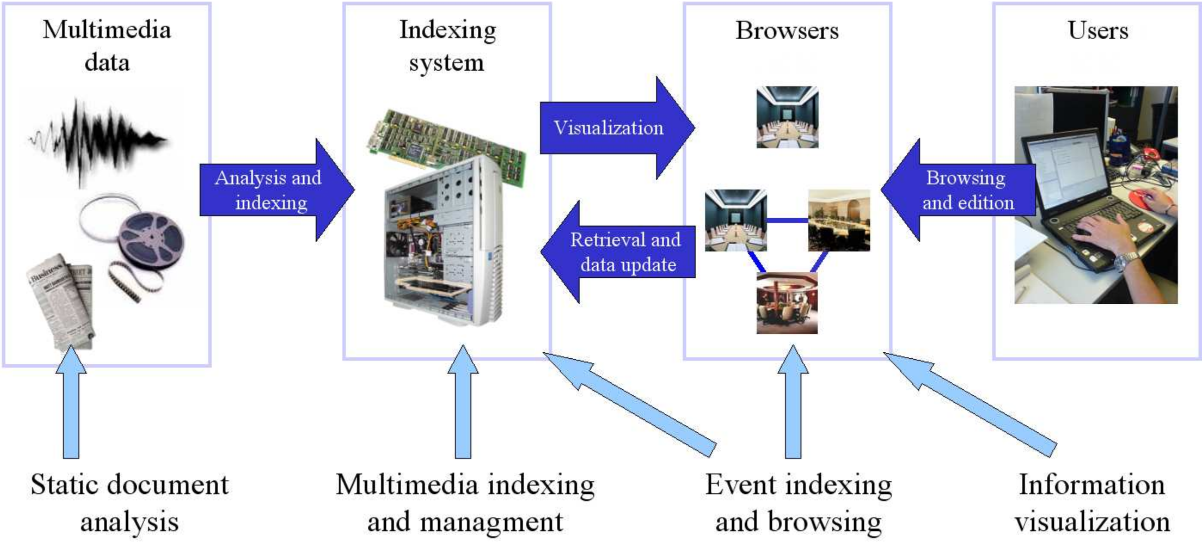


Figure 2.1: The schema synthesizes how the different fields of research presented in this chapter influence multimedia indexing and management technologies.

Unfortunately, existing technologies for organizing and disseminating multimedia data are often primitive or not adapted for reacting to this data production growth. This chapter is dedicated to the state of the art in the domain of multimedia corpora indexing and browsing and presents the lacks this thesis tries to tackle. In fact, multimedia management systems must mix knowledge coming from different fields of computer engineering, which are discussed in four different sections. Figure 2.1 shows how these fields relate to each other. Section 2.1 presents current techniques dedicated to the creation of annotations on multimedia information and emphasizes their strengths and weaknesses. Annotations can extend an original document with semantic information, in order to improve the efficiency of multimedia indexing and retrieval systems. Section 2.2 presents systems developed for events recording and browsing. Events such as conferences or meetings. Each system is based on a specific type of media for accessing the whole event's information. Section 2.3 presents processing techniques for extracting annotations from static documents, which are the cornerstone of our multimedia management system. Section 2.4 presents browsing and visualization techniques useful for representing the correlations existing within multimedia documents collections. Finally, the Section 2.5 announces the challenges tackled in this thesis, highlighting its contributions.

2.1 Multimedia Indexing and Management

All existing systems that manage documents need to represent them in a concise manner, in order to efficiently retrieve the information required by the users. Documents classification and retrieval are performed thanks to indexes. Typically, documents are indexed with text strings, i.e. represented by few and discriminating words (or stems) that summarize their content. The indexing systems calculate these indexes in an automatic manner, using different analysis techniques. For instance, it can create an inverted file containing documents' relevant words [103]. Nowadays, these techniques obtain satisfying results with textual documents (see Section 2.3). At opposite, indexing multimedia information is more difficult: although media such as video and audio can be replayed by computers, their thematic content can not be directly accessed by indexing systems. The solution for overcoming this problem consists in annotating these media with supplemental data, which indexing systems can interpret and use. Currently, we distinguish four methodologies for creating annotations useful for indexing multimedia documents: manual, automatic, combined manual and automatic, and, finally, alignment based annotations.

Users perform *manual annotation*, without any support from analysis systems. Persons try to interpret images, video and audio, in order to add the labels that summarize their content. At a glance, this methodology seems to produce good indexes, but in fact it has two big drawbacks. Firstly, people interpret media's content using personal perception, experience and knowledge and, consequently, the quality and universality of manual annotation is not guaranteed. YouTube [109] is a typical example of a system, in which users upload videos and then add labels for indexing their documents. Frequently, these annotations correspond only marginally with the real content and, moreover, they are vague and do not take into account the different sequences of the video. The second drawback of manual annotations is the cost. Professional labelling, which is de facto a solution to the previous problem, is very expensive in time and need experts who are friendly with indexing systems: choosing a good label means not only that the semantic is correct, but also that the chosen words are useful for retrieving the document. For instance, Janecek [53] and Yee [108] use corpora annotated by experts in their works.

Automatic annotation follows another approach, which tries to overcome manual annotation drawbacks. In this case, analysis systems process the documents in order to extract high level information. Such systems are rather complex and tuned by experts of specific analysis domains. In [92], Swain presents various works that integrate automatic methods for indexing audio, video, and images. In general, the integrated analysis techniques extrapolate low-level features that are specialized for describing a well-defined subset of documents. For instance, the features used for analyzing images can be based on distribution of colors, textures, etc. Videos analysis techniques are derived from images analysis methods and furthermore take into account temporal features, as well as spatial and motion activities. All these low-level features are poor of semantic information and need to be associated to specific themes and concepts. Audio analysis methods use both low-level characteristics such as phonetic units and high-level features such as dictionaries. In the most cases, existing analysis systems take into account only one particular category of media. Recently, Smith et al. proposed an automatic method [91], which uses *model vectors* for indexing multimedia documents. Each vector is correlated with a specialized semantic concept detector and is mapped in the model. Although obtained results are interesting, the set of concepts is restricted.

The *hybrid approach* combines both manual and automatic annotations. This method targets at automatically producing valid annotations from low-level features, in order to reduce and simplify the user's involvement in repetitive tasks. M4Note [37] is a multimodal system that allows to record and to annotate videos with basic features such as time, format, etc. Further, users use electronic ink and speech for easily annotating video frames with textual content. The drawback of these systems consists in the calculated low-level features, i.e. video format, frames per second, etc., that are often too basic, not interesting for the final user, and not useful for the indexing system. For creating annotations, other works have explored more evolved methods based on document analysis. In [70], Moënne-Loccoz et al. present a framework that allows to create systems for indexing large collection of videos. The framework combines methods for the automatic analysis of video with manual annotation defined within an ontology. FutureViewer [19] is a framework that provides a graphical user interface for visualizing and annotating low-level features extracted automatically from videos. The visualization elicits videos' semantics and lets users define and tag clusters of shots. In general, these systems generate sophisticated information in an automatic way, but require also that users are confident with the domain, in order to correctly interpret the results and to produce coherent annotations. To wrap up, the hybrid approach is promising, but, in general, automatic annotations are either simplistic or poor of semantics. Consequently, user intervention is still too important for defining useful annotations. Moreover, in general existing methods take into account only one document at a time, without creating annotations that describe the implicit relationships within the entire document collection.

Alignment based annotation is a recent trend that studies the various similarities between different multimedia documents, with the purpose of transferring information from a document to another and, thus, of indexing the data. Originally, alignment techniques were developed in order to correlate multilingual documents.

For instance, the MUMIS system [60] indexes videos of football matches thanks to alignment of related textual documents. These documents correspond to news, commentaries transcripts, etc. in three different languages. All the analysis results are combined in respect of well-defined rules, in order to align the different sources, to complete the missing content, and to improve their indexing. Behera [10] extracts from videos the slides presented during a meeting or a conference. The detected images are then aligned with the original slideshows, augmenting both media: presented documents are enriched with timestamps, whereas videos inherit the content from the slide. Likewise, Lalanne and Mekhaldi's technique of multimodal alignment is based on textual content similarity [68] and takes mainly into account the static documents presented in a meeting and the transcript of speakers' audio recordings. The alignment augments both media, adding in particular timestamps to static documents and structuring the transcribed dialogues thanks to static documents. In these works, document alignments have produced reliable results and have demonstrated how to automatically create annotations and indexes on multimedia documents difficult to mine with monomodal techniques. For this reason, we believe that document alignment can be used to align static documents, from which semantic information can be easily extracted, with other kinds of media such as audio, video, and so on.

2.2 Events Indexing and Browsers

A recent trend consists in recording some events of interest such as lectures, conferences and meetings. We refer to them globally as "events". They are containers of synchronized and heterogeneous multimedia documents that coexist at the same moment and that are thematically linked. Information associated to events include:

- Printable documents such as papers, news, agendas, newspapers, reports, accounting tables, etc.;
- Speakers' audio and video recordings;
- Slideshows;
- Websites;

- Participants' notes through electronic ink;
- Whiteboard marks with pen strokes;
- Metadata specifying recording attributes, i.e. date, hour, participants' identities, place, etc.

Although all the events are composed of similar types of multimedia data, their global structures are different. Rogina suggests that meetings are more difficult to index than lectures, because they are less planned [86] and also because they involve multiple speakers. In particular, three differences are highlighted: 1) the quality of speech acoustic is better in lectures, when the speaker wears the microphone; 2) the availability of presented documents is sometimes lacking in meetings; and 3) the speaking style is more controlled when presenting a lecture. We consider that talks at conferences are more similar to lectures than meetings, because their structure is rather rigid and strictly related to the structure of a scientific paper. Despite of these structural differences, the works and the researches based on events use analogous indexing techniques and in general share the same targeted user-tasks, i.e. playing the synchronized multimedia data, browsing and searching for specific content and summarizing the event. Firstly, this section illustrates the most relevant works in the domain of events recording, indexing and browsing, and secondly presents a general discussion about them, in order to highlight the unsolved problems in this research field.

2.2.1 Lectures and Conferences Recording

The Cornell Lecture browser [71] aims at automatically producing a multimedia document from recordings of seminars, talks, or class lectures. The authors focused their work on passive recording, which targets at preserving speaker's behavior, i.e. the system does not require invasive and explicit actions for facilitating multimedia indexing. For instance, the speaker is not forced to use specific devices such as whiteboards or to prepare its talk in a specific manner defined by the recording system. The media involved in this project are videos and slides, which are automatically synchronized and aligned. Users can play the event, by viewing the slides and one video composed of alternated shots taken by two cameras. The navigation is either slide- or time-based, i.e. the user can browse the meeting by clicking a discussed slide or through a timeline. Further, the Lecture Browser has been extended by Berkeley Multimedia Research Center with functionalities that allow searching in audio, adding bookmarks, and interacting with a white board [87, 15].

The eClass [14] project (formerly, Classroom 2000 [2]) targets at capturing the classroom experience in a collaborative environment, where the students share their notes. The system requires a pre-production step, in which the teacher prepares the slideshows and the web pages to be presented. During the lecture, eClass records audio and video streams, as well as the text written by teachers on the electronic whiteboard and students' handwritten notes. These notes are time stamped and used as the main artifact for indexing video and audio streams. After the lecture, students can play the entire lecture or only part of it, accessing at all the recorded resources. Moreover, eClass displays a time line enriched with significant moments, occurred when the teacher changed the focused slide or presented a website. Students can browse in the lecture by clicking slides, web links, and handwritten notes, in order to select a specific topic. Finally, the system offers features for searching keywords in the lecture and editing collaborative web pages.

Authoring on the Fly (AoF) [69, 49] is a formal framework that defines strict guidelines for creating multimedia documents from recordings of classroom experiences. Mueller et al. distinguish three different phases involved in lecture recording. In the pre-processing phase, slideshows, images, simulations, etc. that will be presented in the lecture are packaged. During the presentation, this material is explained and annotated by teachers using a whiteboard, while the system records automatically streams of audio, graphics, video, and application commands. Additionally, remote students are able to participate to the classroom via telepresentation. In the post-production, the teacher can eliminate imperfections and speech pauses from recorded data. In the same phase, the recorded streams are synchronized and linked. AoF does not only offer capabilities for replaying and browsing the lecture using slides, but also explores new paradigms for accessing recorded data. In particular, Huerst studied the indexing of audio transcript and slides, taking into account criteria such as text color, size, etc., in order to retrieve pertinent textual content [47]. Moreover, he proposed elastic panning approach [48], which allows students to use an elastic slider to set replaying speed, with the purpose of faster retrieving or overviewing recorded information.

The SMAC project [96] aims at recording, analyzing, indexing, and browsing multimedia conferences archives. The system stores scientific papers and captures the projected slides, speakers' talks, and related videos. Audio and video streams are directly synchronized at recording stage. SMAC matches the slideshows video with the original electronic presentation [10], in order to create indexes for navigating. Moreover, it extracts the electronic content from papers using Xed (see Subsection 4.2.3) and thematically aligns the results with the slideshow presentation [68]. The project provides two interfaces for accessing stored data. The first one is based on SMIL technology [90] and visualizes the image of current slide as well as the videos of speakers, conference room, and slideshow presentations. Either a timeline or a bar containing the images of projected slides allow to navigate in the synchronized conference. The second interface is web-based and proposes the same functionalities described above. Furthermore, it visualizes the pages of the presented paper segmented in textual blocs, which can be used as indexes for navigating in the conference by themes. Recorded conferences are managed thanks to Indico system [52], which permits to prepare the different phases involved in the organization of the event, generating at the same time metadata useful for indexing. SMAC system is currently used at CERN [1] institute.

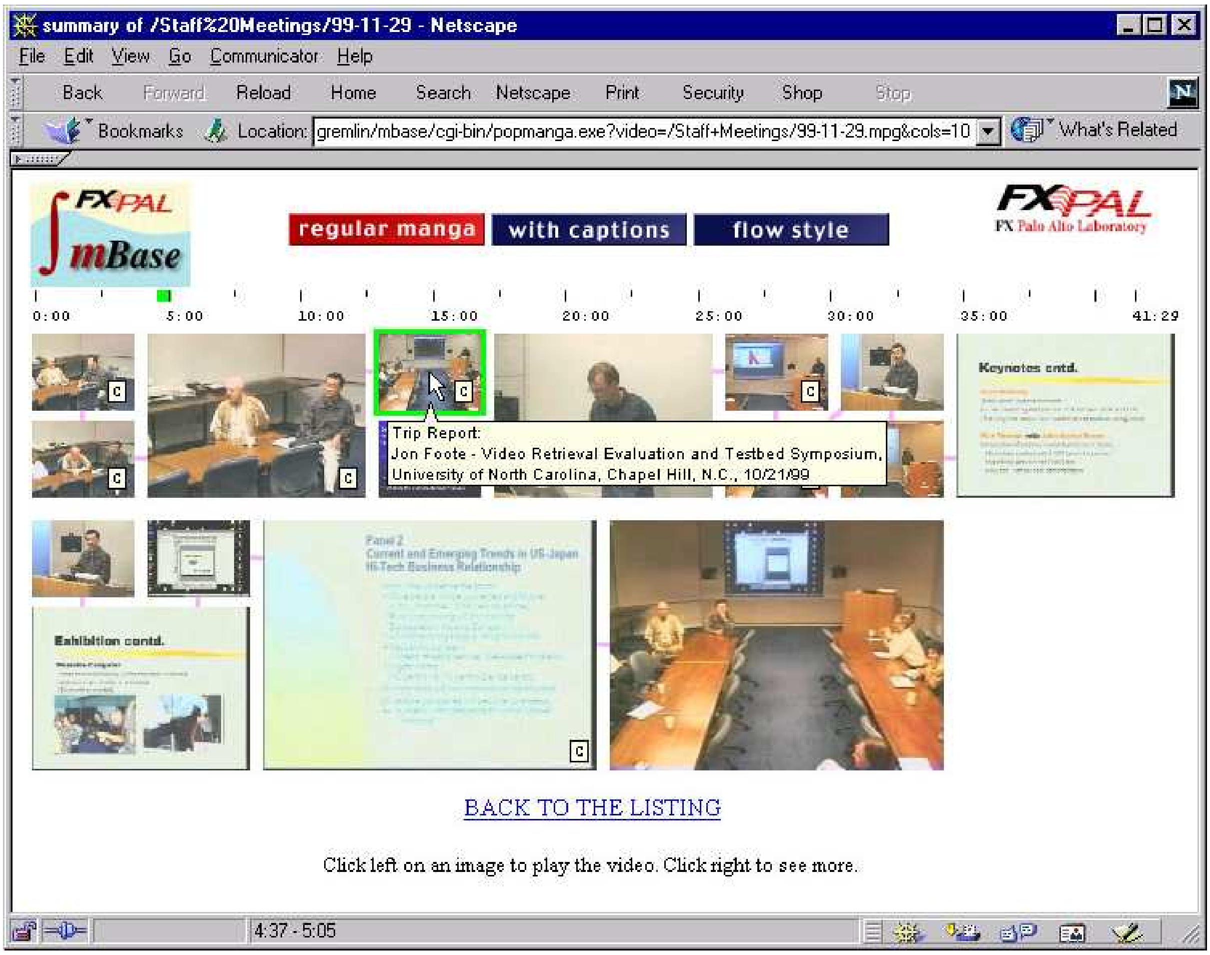
2.2.2 Meeting Recordings

Distributed Meetings [30] is a system for meeting and teleconferencing management, which deals with audio, video, whiteboard marks, and presented documents. An interesting feature of DM is the portable hardware: a unique ring camera equipped with 5 directional microphones captures all the participants at the same time.The system segments the speech in respect of participant intervention, calculated thanks to the multimodal fusion of audio source localization and of speakers tracking in video. Moreover, it creates indexes from the whiteboard content, taking into account the timestamps of pen strokes. DM provides a browser that offers both a basic view of the meeting for teleconferencing and a postproduction view.The latter is the most complete and allows to visualize the current speaker, the panoramic view calculated with the different video streams, the whiteboard key frames, and a timeline with dialogues segmentation. Users can jump in the synchronized meeting via whiteboard marks or through the timeline. An exciting browsing mechanism permits to remove pauses in speech or to speed up the meeting replay without changing audio pitch.

SMaRT [98] is a system dealing with meetings that aims at enhancing human-human interaction and at integrating computers in a non-intrusive manner. In fact, a consistent part of the project studies how to interact with the meeting room in real time, using people and voice recognition in order to change slides, to control devices, to view summaries, etc. Recording of audio and video streams are supported for post-meeting activities thanks to the browser developed at Interactive Systems Laboratories of the Carnegie Mellon University [97]. The browser allows to access a collection of meetings, which are enriched with speakers' identities and attitudes, summaries obtained through dialogue analysis, etc. More precisely, the browser allows to consult the speech transcript, summaries, and participants' videos, which are time-aligned. Finally, this browser offers capabilities for navigation, edition, and retrieval of audio and textual content.

LiteMinutes [24] is an Internet-based system developed at FX Palo Alto Laboratory that is focalized on meeting minutes in HTML format. Its main purpose is the use of emails for enabling collaborative activities on the minutes, which are composed of notes taken by participants during the meeting. The system records video and audio streams, the time stamped images of presented slides, and all the notes taken by participants on their personal computers. LiteMinutes permits to already consult these notes during the meeting: the new annotations are immediately accessible to each person, segmented into time stamped items and linked with multimedia data such as images and video. Multimedia data that have been prepared in pre-production phase, but that have not been presented, are ignored by the system. Users can browse and replay the meetings thanks to notes that are linked to existing multimedia resources. Furthermore, the browser visualizes video key frames,calculated in an automatic manner, in both a timeline and

[1] European Council for Nuclear Research


summary of /Staff%20Meetings/99-11-29 - Netscape
File Edit View Go Communicator Help
Back Forward Reload Home Search Netscape Print Security Shop Stop
Bookmarks
What's Related
FXPAL mBase
regular manga
with captions
flow style
FXPAL FX Palo Alto Laboratory
0:00
5:00
10:00
15:00
20:00
25:00
30:00
35:00
41:29
Keynotes cntd.
Exhibition contd.
Panel 2 Current and Emerging Trends in US-Japan Hi-Tech Business Relationship

Location: gremlin/mbase/cgi-bin/popmanga.exe?video=/Staff+Meetings/99-11-29.mpg&cols=10

Trip Report:
Jon Foote - Video Retrieval Evaluation and Testbed Symposium,
University of North Carolina, Chapel Hill, N.C., 10/21/99

BACK TO THE LISTING

Click left on an image to play the video. Click right to see more.

4:37 - 5:05

Figure 2.2: In LiteMinutes, the Manga view attributes a major visual relevance to significant shots. *Image used with permission from FXPAL [24].*

a Manga view (confer Figure 2.2). The latter contains significant frames, annotated with notes produced at the same time. Finally, the edition task allows to correct or to extend meeting summary: after each modification, the minute is sent via email to other participants and to the system.

The Portable Meeting Recorder [64] is a compact system developed at Ricoh Innovations, which is composed of a omni-directional camera and 4 microphones (cf. Figure 2.3b). The camera captures a panoramic view of the meeting, whereas the sound sources are redirected to a multi-channel sound card for processing the signal in real time. The system locates the talking person thanks to analysis of sound's direction, angle, and elevation, jointly with speakers' face detection. This localization permits to create a virtual video of the meeting, composed with the best shots of talking speakers, i.e. when their faces are not occluded and they are facing the camera. Moreover, the Portable Meeting Recorder analyzes the sound in order to detect the

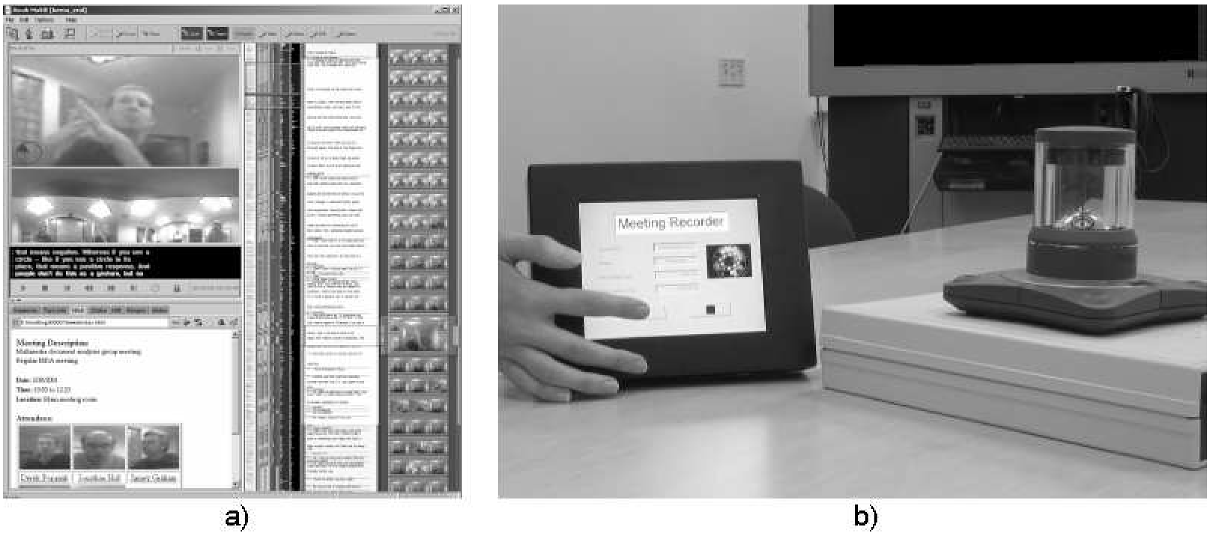
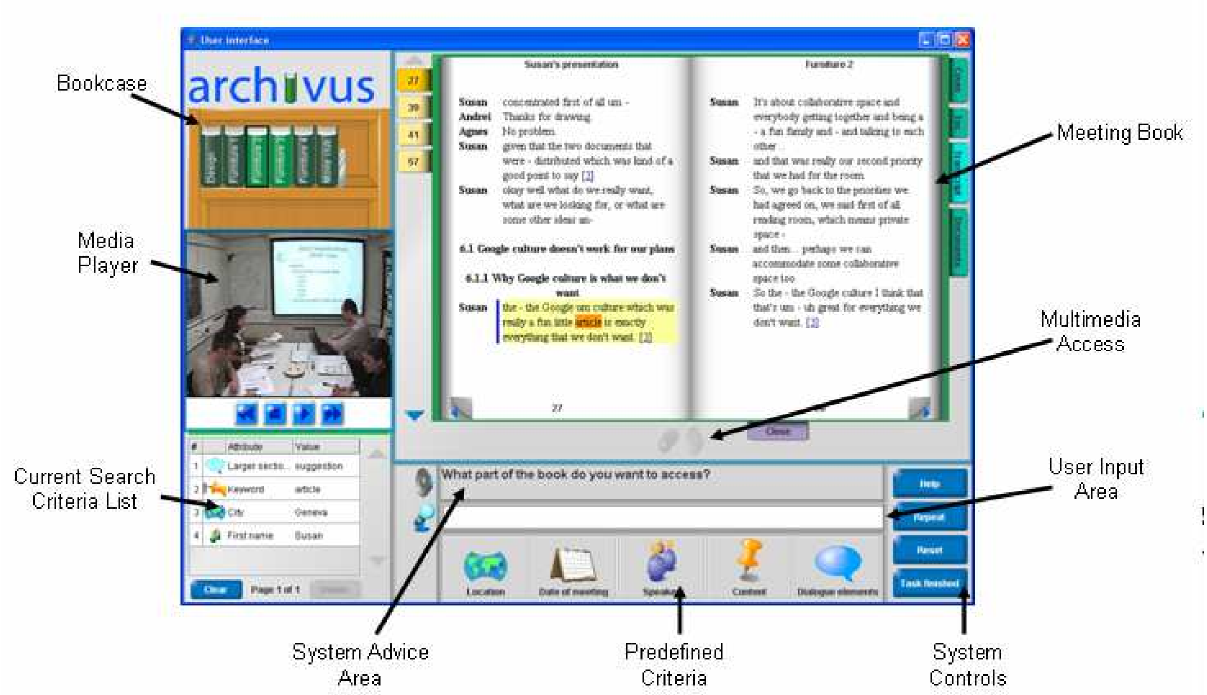


Figure 2.3: The MuVie browser (a) and the hardware of meeting browser recorder (b). *Figures are courtesy of Ricoh Innovations, California Research Center [64].*

intensity of conversation and the speed of interactions between participants. Furthermore, video analysis aims at recognizing both the location of the meeting and participant's identity and motion. MuVIE is the proposed client interface (see Figure 2.3a) that offers several capabilities for editing, searching textual content, etc. In particular, in this project MuVIE visualizes the panoramic view of the meeting, the virtual movie with talking persons, speaker transitions, audio and visual activities, the meeting transcript, and the key frames for navigating in videos. Users can interact with MuVIE in order to browse, filter speakers, search keywords in transcript, and so on.

Archivus [66, 5] is at contemporary a meeting browser and a retrieval system developed for studying multimodal interactions. It displays an archive of meetings using the library metaphor (cf. Figure 2.4), where each meeting is a book, its subject is the title, the participants are the authors, and so on. The metaphor has been designed in order to facilitate the dialogue between inexpert users and the system. In fact, Archivus proposes different modalities of communication, classified as inputs active (voice, keyboard, mouse or pen), input passive (emotions) and outputs (graphics, sound and text). Each input modality is interchangeable, in order to enable users to access the meetings using their preferred modalities. Tests on multimodal dialogue structure have been accomplished using a Wizard Of Oz methodology, in which a human operator, the Wizard, simulates the system and reacts to the inputs of unaware users. Users can search information, thanks to a satisfaction constraints mechanism or to analysis of natural language, and browse into the archive or into one meeting.

The Ferret meeting browser [101] is a modular system developed at IDIAP that allows to manage an archive of meetings. Firstly, the user selects a meeting from the archive, consult-



Susan's presentation

Susan concentrated first of all um -
Andrei Thanks for drawing.
Agnes No problem.
Susan given that the two documents that were - distributed which was kind of a good point to say [2]
Susan okay well what do we really want, what are we looking for, or what are some other ideas an-

6.1 Google culture doesn't work for our plans

6.1.1 Why Google culture is what we don't want

Susan the - the Google um culture which was really a fun little article is exactly everything that we don't want. [2]

27

Furniture 2

Susan It's about collaborative space and everybody getting together and being a - a fun family and - and talking to each other
Susan and that was really our second priority that we had for the room.
Susan So, we go back to the priorities we had agreed on, we said first of all reading room, which means private space -
Susan and then... perhaps we can accommodate some collaborative space too.
Susan So the - the Google culture I think that that's um - uh great for everything we don't want. [2]

What part of the book do you want to access?

#	Attribute	Value
1	Larger sectio...	suggestion
2	Keyword	article
3	City	Geneva
4	First name	Susan

Figure 2.4: The Archivus' interface, based on library metaphor, and its functionalities. *Image reproduced by permission of A. Lisowska.*

ing summary information and participants' pictures. Then, the browser proposes all disposable resources and asks the user to select the data streams of interest, i.e. videos, transcript, documents, etc. So, the browser is configured with the plug-ins necessary for replaying or navigating in the interesting data. All the media are synchronized and the user can interact with them using a VCR-like control or a timeline. Moreover, while browsing the users can dynamically change the visualized resources and load new data streams, replacing or removing the old ones. In the recent past, the authors proposed the new JFerret framework, useful for fast development of meeting browsers and used by JFriDoc (see Section 6.2).

2.2.3 A Discussion About Event Recordings

Various papers have proposed a classification of projects involved in indexing and browsing of multimedia events. Erol presented an exhaustive state of the art in [31], which summarizes several works classified by their capture systems, content analysis, and retrieval methods, as well as by content delivery and evaluations. In a different way, Tucker and Whittaker classified meeting browsers by type of control stream (video, audio, and artifacts) [94], whereas Lalanne and Sire considered the types of streams: control streams, derived streams, and annotations [63]. Unfortunately, all these classifications are not completely satisfying and, sometimes, incomplete or vague for defining a reasonable taxonomy or for integrating all existing projects in an objective manner. In fact, this difficulty can be justified with a few observations:

1. The earlier works take into account the full organization of events, from recording to browsing, whereas more recent researches are focusing mainly on precise phases. Thus these works are not always directly comparable with the same metrics.

2. Frequently, the existing projects take into account the same media streams. Obviously, such a classification implies that most of the projects belong to the same group, with few exceptions.

3. In most cases, research in events management is focused on improvement of different analysis techniques in order to resolve an identical problem, for instance to detect the speaker or to transcribe the speech. Unfortunately, such a classification is not interesting in the context of this thesis.

Even though the classification of related works remains a difficult task, it is possible to elicit some topic of discussions and to highlight lacks of existing projects, from the point of view of printable documents' role and of events relationships.

The Role of Printable Documents

All the projects presented in Subsections 2.2.1 and 2.2.2 take into account and analyze media such as video and audio, but very few investigate printable/printed documents, which are part of the information available during an event and in our daily lives. Moreover, printed documents can be analyzed with robust techniques that produce reliable and useful results. In [14, 24, 71, 69], slides are used as indexes for navigating in lectures, but their thematic content is ignored. Consequently, users could interpret its message in order to browse, but the content is not used for indexing purposes.

Full text search is the first integrated technique using the thematic content of documents for accessing events recordings [64, 98]. Huerst [47] proposes to enhance this functionality with the detection of relevant textual content, thanks to typographic properties of text. Although full text research is an interesting feature, document structures are still ignored. Structures can be used for improving event indexing, for providing useful information in order to analyze other media, and for structuring unplanned events such as meeting. These concepts are explained with more details in the Chapter 4. Document structures have only been integrated in systems developed in our research group [68, 96] or envisaged in works [66] using the results of Xed, our technology for recovering the physical structures of PDF files (see Subsection 4.2.3). We believe that the structures of printable documents are necessary for improving indexing techniques and that this medium can provide a natural and thematic mean for browsing and searching through large multimedia repositories.

Events Relationships

Tucker and Whittaker observe that the existing projects only consider individual and isolated events, without looking at their relationships and context [95]. In fact, few approaches manage the whole archive of events and as access points they mainly propose a simple list that provides basic information [97, 101]. In the best case, the systems group events together by categories and offer functionalities permitting to search full text in the entire archive [96]. This is a top-down approach, because the user firstly selects one event and then browse within it. When the user desires to browse into a new event, it is necessary to go back to the entire list once more.

Archivus [66] is the only browser that proposes some advanced features for browsing and searching at the archive level. The bookshelf shows groups of related meetings and uses colors and shades for representing events that match the user's queries. Moreover, the bookshelf is always visible while browsing and users can select a new meeting without changing the browsing level. Although this browser is a first attempt at representing the archive and its relationships, cross-event management lasts primitive and several problems are not taken into account. For instance, how do event relate to each other? How do the topics evolve in time into correlated events? Which documents have been presented in more than one event? Which topics do recur among uncorrelated events? Actually, these answers are still opened, because existing browsers do not consider correlations in events archive, neither for indexing nor for browsing.

The next Section 2.3 presents the analysis techniques used in order to automatically annotate static documents - in particular PDF documents - and to prepare them for efficient alignment and indexing. In fact, we think that finding the relationships between printable documents permits to further correlate their different events.

2.3 State of the Art in Static Documents Analysis

The word "*document*" derives from the Latin verb "docere" (to teach, show): thus, the purpose of documents is to transmit information that is apprehensible by humans. In particular, *static documents* are documents that can be printed, such as newspapers, books, etc. They have the characteristic to be a vector for thematic and structured information, because their appearance and structures define the rules for reading and interpreting the content. For instance, a scientific paper is composed of several text blocks with specific logical functions: keywords, abstract, chapter, etc. These functions communicate to the reader the relevance of the text they contain and allow to decide the reading depth. In fact, a person could examine only the paper's abstract and be advised of the whole content and, therefore, be conscious that it is not necessary to read further.

Recent works try to automatically extrapolate this high level information, in order to exploit them in various contexts, such as in indexing and retrieval systems. This discipline is called "*document analysis and recognition*". Subsection 2.3.1 is a general overview of the document image analysis techniques, which describes the origins, the vocabulary, and the methods belonging to this discipline. At opposite, Subsection 2.3.2 focuses on the specific domain of PDF document analysis.

2.3.1 Document Image Analysis

Figure 2.5 represents the phases of document production and the recognition processing chain. At production, the writer types in the content and defines logical entities, describing thus the semantic of the document. Then, the editor formats the document, selecting the layout of text, the fonts, and the colors. These properties define the physical structure of the document. At this stage, the document is ready for being rendered as image, which can be printed or viewed on the screen.

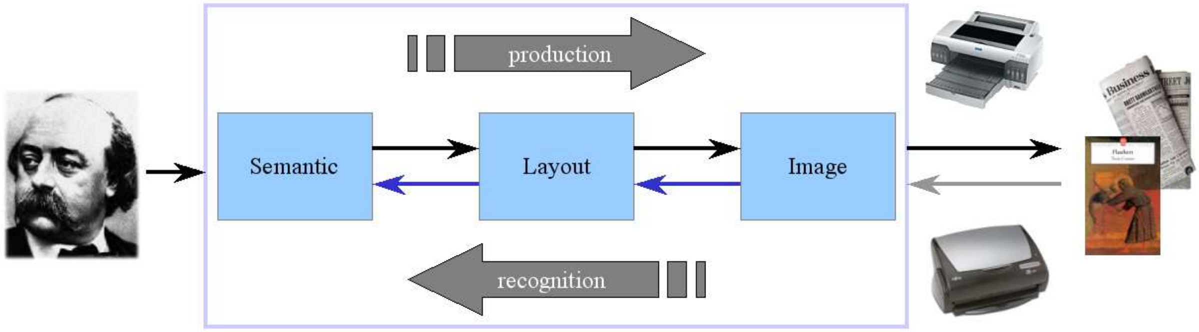


Figure 2.5: Production and recognition chains are reducible to three levels. The semantic or logical level encapsulates author's message; the layout describes the physical appearance; finally, the image is the final reading form.

Document recognition applies a reverse engineering process on printed documents. Although the steps are symmetric to those involved in the production, the recognition requires many analysis and sub-processing. Firstly, the printed document is digitized using a scanner or synthesized. The scanned image is in general afflicted by geometric deformations and noise that must be eliminated or attenuated by pre-processing algorithms.

The filtered or digitized image is further analyzed in order to recognize the *physical structure*, which groups the characteristics of the document's visual form. Various challenges are involved with physical structure recovering: *page decomposition* tackles the problem of separating text blocks, images, frames, and threads. *Optical character recognition* (OCR) targets at restituting the text; the *optical font recognition* (OFR) analyzes the used fonts. In general, physical structures are independent from any specific class of document and their methods are valid for any kind of printable documents, for scientific papers as well as for newspapers.

The next phase is the *logical structure* recognition, which deals with document's logic and interprets the results of the physical structure recognition. For instance, different methods are specialized for labelling physical text blocks and for establishing their logical hierarchy. These methods take into account documents with complex layout, such as newspapers, technical reports, etc. At opposite to physical structure recognition, the methods for recovering the logical structures depend on specific document classes (for example, newspaper) and sub-classes (i.e. *"Times"* and *"Le Monde"*). Moreover, the same class of documents can be interpreted in different manners, according to the targeted application: for instance, one model is adapted for describing a thesis as a sequence of chapters, composed of titles, section and text bodies, while another model considers the titles such as the items of the table of content.

All the techniques discussed in this subsection are exhaustively described in various synthesis books, for instance [17, 23], and conferences proceedings, such as [67, 16].

A recent trend consists in analyzing documents in their electronic form, instead of using their image. Three reasons justify this approach. Firstly, nowadays 90% of the documents are available in their electronic form, whereas 10% remains available in their only printed form. Secondly, eliminating the errors provoked by scanning allows to work essentially on the methods of structures recognition. Thirdly, documents management systems need to profit of the knowledge that could be extracted from electronic documents for purposes of indexing, searching, etc. The methods for analyzing electronic documents, and more precisely PDF files, are discussed in the next Subsection 2.3.2.

2.3.2 PDF Documents Analysis

Nowadays, Portable Document Format [51] has became the de facto standard format for exchanging documents through the Internet. PDF is a robust format and allows to represent all types of document, but it has a major drawback: it is usually generated by automatic tools. These PDF generators privilege the final rendering of the document, regardless of the physical and logical structures, and add a multitude of inconsistencies in the content, such as badly segmented words, unexpected blanks, etc. These artifacts will be analyzed in more details in the Subsection 4.2.2.

Moreover, PDF is very rich and redundant. The same document can be represented in various manners: for instance, one page could be stored in its raster form or detailed though primitives such as text, graphics and images. Consequently, accessing PDF documents' content can be a cumbersome task and an analysis process is necessary.

During the last years, various works and researches tried to extract the content of PDF in a form accessible by both human and machines and, in some case, to discover physical and logical structures. The categorization of existing methods is detailed in the next paragraphs and summarized in Table 2.1.

The methodology based on *document image analysis* targets at recovering the content and the original structures from document's image (see Subsection 2.3.1). Applying these techniques to PDF documents presents the double advantage of using mature methods and, in the major-

<table>
<tr><td>Document image analysis</td><td colspan="3">Electronic content analysis</td></tr>
<tr><td>+ mature techniques
+ independent from document</td><td colspan="3">+ accurate results
+ access to document hidden information</td></tr>
<tr><td></td><td>Extending methods</td><td colspan="2">Restructuring methods</td></tr>
<tr><td></td><td>+ document preserved</td><td colspan="2">+ information easily accessible</td></tr>
<tr><td></td><td></td><td>Conversion</td><td>Reverse engineering</td></tr>
<tr><td></td><td></td><td></td><td>+ content and structures strictly related</td></tr>
</table>

Table 2.1: Taxonomy of methods for PDF document analysis.

ity of situations, of analyzing ideal images, i.e. at very high resolution and noise free [39, 40]. Moreover, image analysis is independent from the inner representation of PDF file, because it acts always on rendered images. In fact, the raster image is the explicit and universal representation of both the document's content and structures. Document image analysis has only two disadvantages. Firstly, the automatic recognition of document is always perturbed by errors, which are cumulated after each analysis step. The error rate is acceptable when equivalent to that of humans accomplishing the same task [9]. Secondly, document images techniques do not take into account electronic information contained in the file, such as PDF textual primitives, which usually is reliable information.

Although *electronic content analysis* [74] is theoretically the complementary methodology to image analysis, it inherits part of its know-how. Its techniques directly access the electronic content of the document that ideally is already segmented in textual and graphical primitives. The electronic content analysis eliminates all the pre-processing steps applied to the image from the scanning of the printed form until the OCR, reducing errors in successive recognition phases. Unfortunately, accessing the content remains difficult because of the format [80] and requires new analysis methods. Moreover, content analysis can not be applied to any types of documents: for instance, when PDF pages are stored as raster, the textual content does not exist and OCR is required for recovering it. Thus, in [41, 81] we argued that merging both methodologies can improve the analysis of PDF documents: respective lacks can be counterbalanced and the results cross-validated.

Table 2.1 shows two families of techniques belonging to *electronic content analysis*: *extending* and *restructuring methods*. *Extending methods* augments the original PDF documents with new annotations. Structures analysis produces results that are simply included in the original document as tags. Extending methods have been successfully used in [8, 42], but they have two main disadvantages: they conserve the PDF format, which is hard to access and manipulate, and often require specific plug-ins for reusing analysis results.

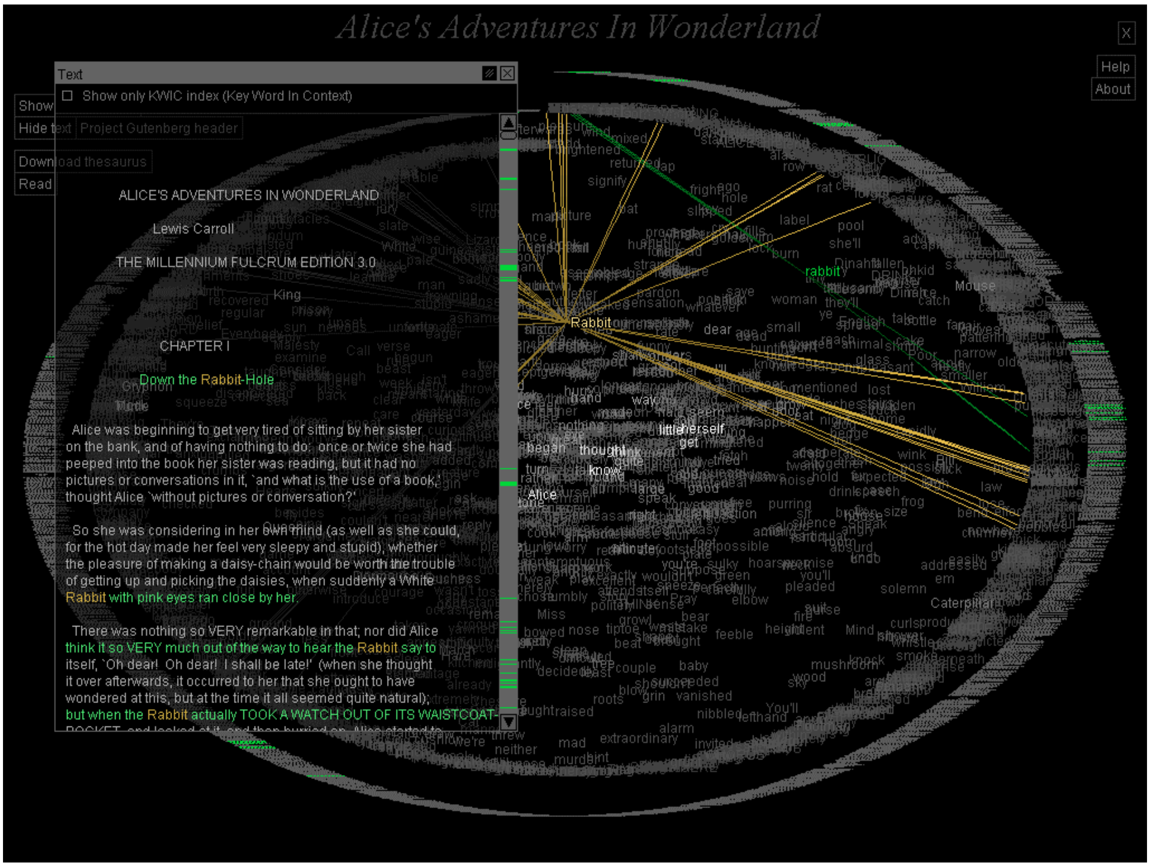
Restructuring techniques use another format than PDF for representing the document after analysis. The output format of predilection is XML [105], because it is structured, human readable and very simple to manage for machines. The most evolved technique of restructuring is *reverse engineering*, which does not only recover the PDF document's content, but tries also to clean it and to extract its implicit structures. Nowadays, several researches [6, 21, 22, 35, 79] and tools [56] perform PDF restructuring with more or less accurate results. The *conversion* is similar to reverse engineering because it allows to extract PDF documents in a more user-friendly format, but no analysis is applied over the content. Products such as [1, 76, 29, 77, 106] are only a part of the existing tools for the conversion of PDF documents. Although they make the document's content accessible, they do not extract its physical and logical structures.

To wrap up, we consider that PDF is an interesting format, robust, complete, and well-accepted by persons. However, we believe that existing PDF documents must be reverse engineered, because their content is difficult to access and their lack of structures.

2.4 Information Visualization of Multimedia

Information visualization (alias *infovis*) is a field of computer science that studies how to present data to users, in order to improve their understandability. Moreover, visualizations can be coupled with interaction and animations, in order to permit the users to focus on interesting information. We believe that information visualization can be used for representing archives of multimedia data, in particular in the case of meetings, in order to complement indexing and to enhance browsing and information retrieval. Moreover, hidden information and relationships can be discovered thanks to infovis techniques. This section introduces recent works and researches that are focused on visualization of digital libraries and then present a brief discussion about them.

TextArc [75] reveals the frequency and distribution of the words in a textual document such as books, mails and news (cf. Figure 2.6). Firstly, the main visualization displays lines of text on an ellipse, in the order they appear in the document. Special lines such as headings, poetries, etc. are highlighted with landmarks in order to preserve document structures. Secondly, the text is redrawn in a second inner ellipse word by word: when a word occurs several times, its location is the mean of all the positions it should occupy. To be precise, the most recurrent words are visualized in the proximity of ellipse's center, whereas single occurrences are located on the perimeter. Moreover, words frequency influences text brightness, which is lighter for recurrent words. Finally, when users point a word with the mouse, TextArc draws lines connecting it to its locations in the document. We consider that TextArc has two main interests. Firstly, drawing lines between words and locations efficiently indicates their correlation. This artifact is also well-suited for connecting similar multimedia documents. Secondly, TextArc calculates the position of a recurrent word as the mean of all its occurrences' locations. This technique allows to eliminate multiple occurrences of the same object and to locate it relative to the general context.



□ Show only KWIC index (Key Word In Context)

ALICE'S ADVENTURES IN WONDERLAND

Lewis Carroll

THE MILLENNIUM FULCRUM EDITION 3.0

CHAPTER I

Down the Rabbit-Hole

Alice was beginning to get very tired of sitting by her sister on the bank, and of having nothing to do: once or twice she had peeped into the book her sister was reading, but it had no pictures or conversations in it, `and what is the use of a book,' thought Alice `without pictures or conversation?'

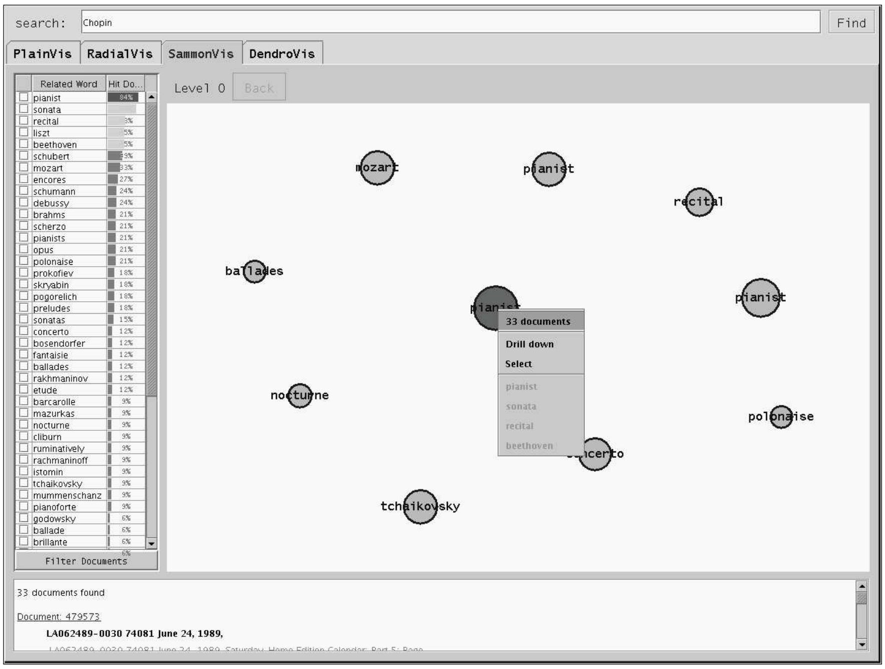
So she was considering in her own mind (as well as she could, for the hot day made her feel very sleepy and stupid), whether the pleasure of making a daisy-chain would be worth the trouble of getting up and picking the daisies, when suddenly a White Rabbit with pink eyes ran close by her.

There was nothing so VERY remarkable in that; nor did Alice think it so VERY much out of the way to hear the Rabbit say to itself, `Oh dear! Oh dear! I shall be late!' (when she thought it over afterwards, it occurred to her that she ought to have wondered at this, but at the time it all seemed quite natural); but when the Rabbit actually TOOK A WATCH OUT OF ITS WAISTCOAT-

Figure 2.6: TextArc visualizes Alice's adventure in wonderland. The word "rabbit" and the lines containing it are highlighted. *W. Bradford Paley [75], ©2002 IEEE.*

Info Navigator [20] is a system dealing with textual documents and visualizing search results in three different views. In the Sammon cluster view (see Figure 2.7), clustered documents are represented with circles in a plane: their distances correspond to their similarity. Each circle is labelled with its most frequent keyword and its size is proportional to the amount of contained documents. Clicking a circle allows to access the whole cluster's keywords and related documents in a scrollable list.

The Dendro map (cf. Figure 2.8) visualization represents the clusters as binary trees. Their depth is limited and the leafs are either documents or sub-clusters, which are labelled with the most representative keyword.



	Related Word	Hit Do...
□	pianist	84%
□	sonata	
□	recital	3%
□	liszt	5%
□	beethoven	5%
□	schubert	9%
□	mozart	33%
□	encores	27%
□	schumann	24%
□	debussy	24%
□	brahms	21%
□	scherzo	21%
□	pianists	21%
□	opus	21%
□	polonaise	21%
□	prokofiev	18%
□	skryabin	18%
□	pogorelich	18%
□	preludes	18%
□	sonatas	15%
□	concerto	12%
□	bosendorfer	12%
□	fantaisie	12%
□	ballades	12%
□	rakhmaninov	12%
□	etude	12%
□	barcarolle	9%
□	mazurkas	9%
□	nocturne	9%
□	cliburn	9%
□	ruminatively	9%
□	rachmaninoff	9%
□	istomin	9%
□	tchaikovsky	9%
□	mummenschanz	9%
□	pianoforte	9%
□	godowsky	6%
□	ballade	6%
□	brillante	6%
		6%

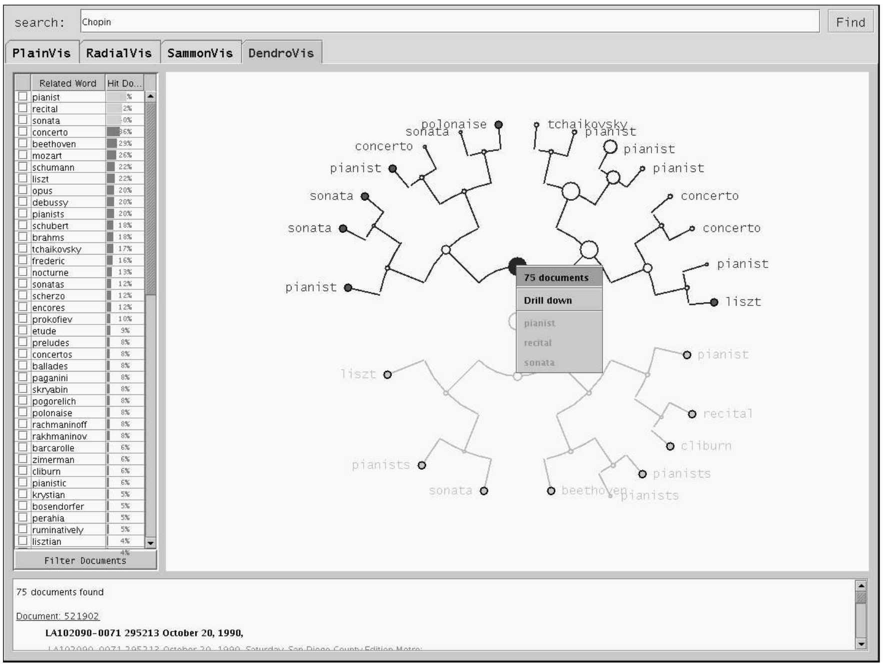
33 documents found

Document: 479573

LA062489-0030 74081 June 24, 1989,

LA062489-0030 74081 June 24, 1989, Saturday, Home Edition Calendar; Part 5; Page

Figure 2.7: The Sammon cluster view displays clusters of documents. The distance between clusters corresponds to their similarity. *Image reproduced by permission of S. Rüger [20].*



	Related Word	Hit Do...
□	pianist	%
□	recital	2%
□	sonata	0%
□	concerto	36%
□	beethoven	29%
□	mozart	26%
□	schumann	22%
□	liszt	22%
□	opus	20%
□	debussy	20%
□	pianists	20%
□	schubert	18%
□	brahms	18%
□	tchaikovsky	17%
□	frederic	16%
□	nocturne	13%
□	sonatas	12%
□	scherzo	12%
□	encores	12%
□	prokofiev	10%
□	etude	9%
□	preludes	8%
□	concertos	8%
□	ballades	8%
□	paganini	8%
□	skryabin	8%
□	pogorelich	8%
□	polonaise	8%
□	rachmaninoff	8%
□	rakhmaninov	8%
□	barcarolle	6%
□	zimerman	6%
□	cliburn	6%
□	pianistic	6%
□	krystian	5%
□	bosendorfer	5%
□	perahia	5%
□	ruminatively	5%
□	lisztian	4%
		4%

75 documents found

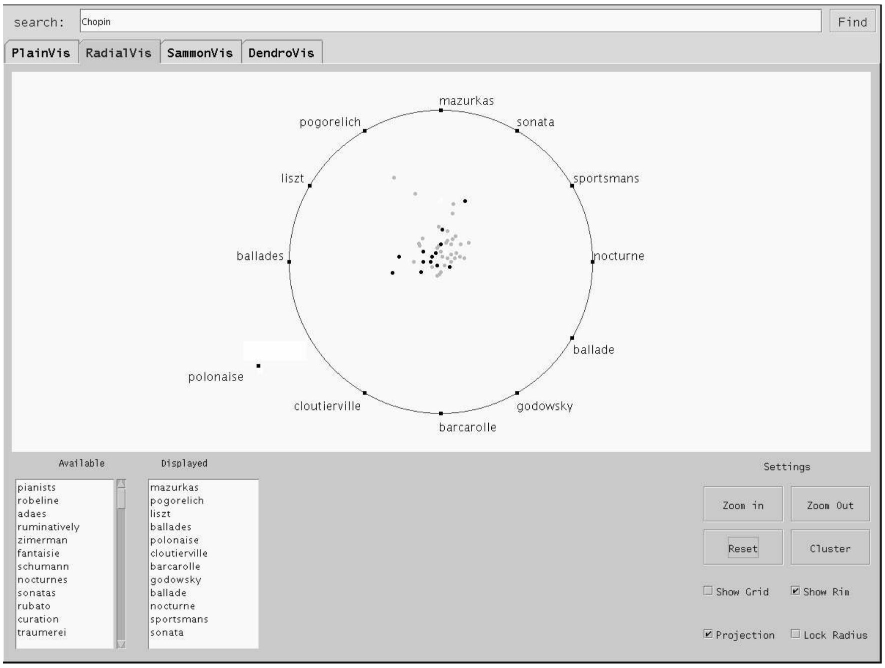
Document: 521902

LA102090-0071 295213 October 20, 1990,

LA102090-0071 295213 October 20, 1990, Saturday, San Diego County Edition Metro;

Figure 2.8: The Dendro Map represents clustered document in a hierarchy. *Image reproduced by permission of S. Rüger [20].*

The Radial view is based on the RadViz visualization [46] (see Figure 2.9), which disposes keywords around a circle, as anchors attracting related documents. When the documents contain a very influencing keyword, they are located in its proximity. At opposite, if the documents contain all the keywords of the query, they are rather situated in the center of the visualization. RadViz is explained with more details in Section 6.3. All the visualizations of Info Navigator can be switched at any moment, in order to browse in the same result set.


search: Chopin

Available: pianists, robeline, adaes, ruminatively, zimerman, fantaisie, schumann, nocturnes, sonatas, rubato, curation, traumerei

Displayed: mazurkas, pogorelich, liszt, ballades, polonaise, cloutierville, barcarolle, godowsky, ballade, nocturne, sportsmans, sonata

□ Show Grid ☑ Show Rim

☑ Projection □ Lock Radius

Figure 2.9: The Radial view visualizes the most relevant documents in the center of the circle. *Image reproduced by permission of S. Rüger [20].*

The application of RadViz to an archive of documents has been principally retained in our work. In fact, RadViz disposes the most relevant documents in the center of the visualization, i.e. where the users are mainly concentrated, and indicates the strength of query's keywords.

ThemeRiver [43] is a visualization that elicits thematic variations in a large collection of textual documents over time. Themes are represented as rivers that change of size relative to themes strengths at a precise moment, allowing users to discover pattern and to validate hypothesis about the collection (cf. Figure 2.10). The river is composed of several colored flows, which correspond to the individual themes. The authors suggest that ThemeRiver could be improved with a feature that assigns colors of the same family to correlated themes. From our point of view, this temporal visualization is very intuitive to interpret and has two advantages. Firstly, the strength of flows and river clearly indicates the importance of query keywords at a given moment. Secondly, ThemeRiver represents the topics as continuous flows, which are more readable and significant for users than plot points or histogram bars.

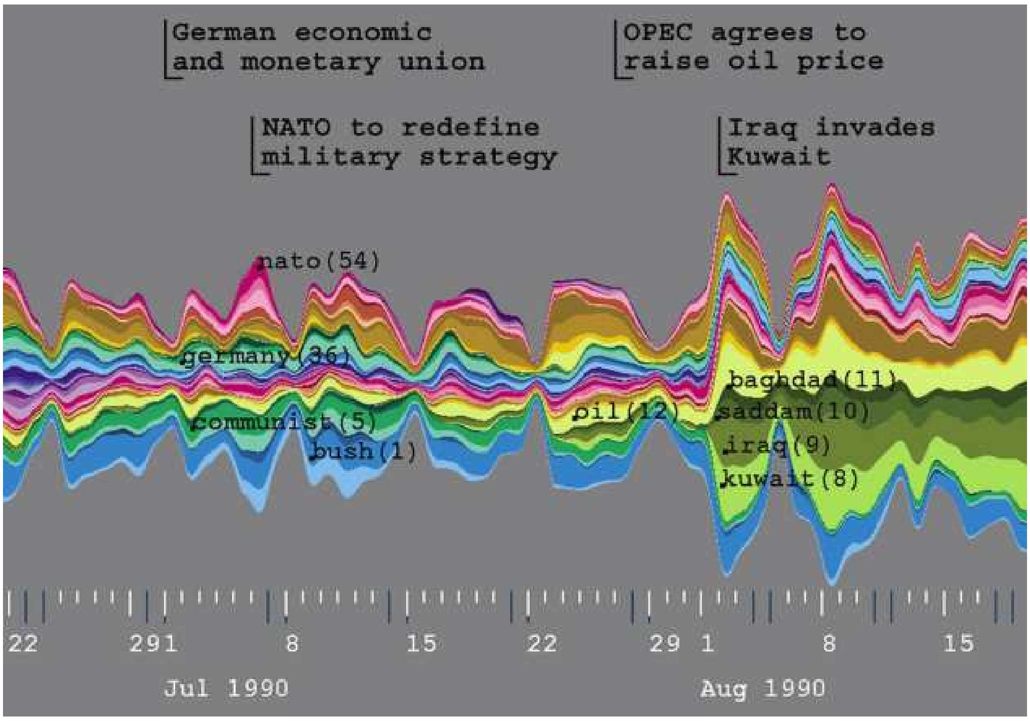


Figure 2.10: The ThemeRiver visualization shows the predominance of topics such as oil, Saddam, etc. during the month of August, after Iraqi invasion of Kuwait. *S. Havre [43], ©2002 IEEE.*

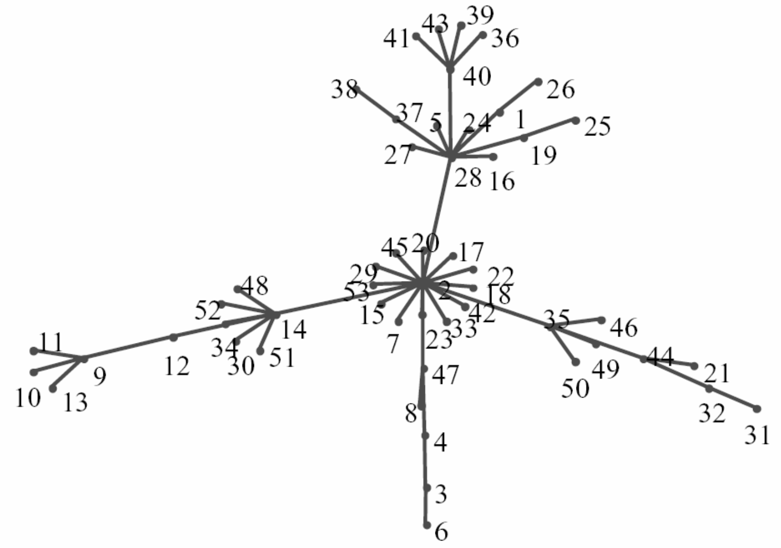


Figure 2.11: The tree shows co-citation count for a group of papers. Probably, number 2 is the precursor within the domain. *S. Noel [75], ©2002 IEEE.*

Noel and al. [73] proposed a visualization based on minimum spanning trees, in order to explicitly represent links between related scientific documents. These relationships are given by the *citation count*, i.e. a measure of similarity indicating how many times a pair of papers has been co-cited. Citation count allows to construct the graph that represents the direct influences existing in a collection of papers. The root of the tree and its closer nodes are drawn in the visualization's center and represent the most influencing literature, whereas the leafs are new research domains (confer Figure 2.11). This visualization is interesting, because the explicit representation of directed links between documents creates thematic spaces, useful for browsing an archive.

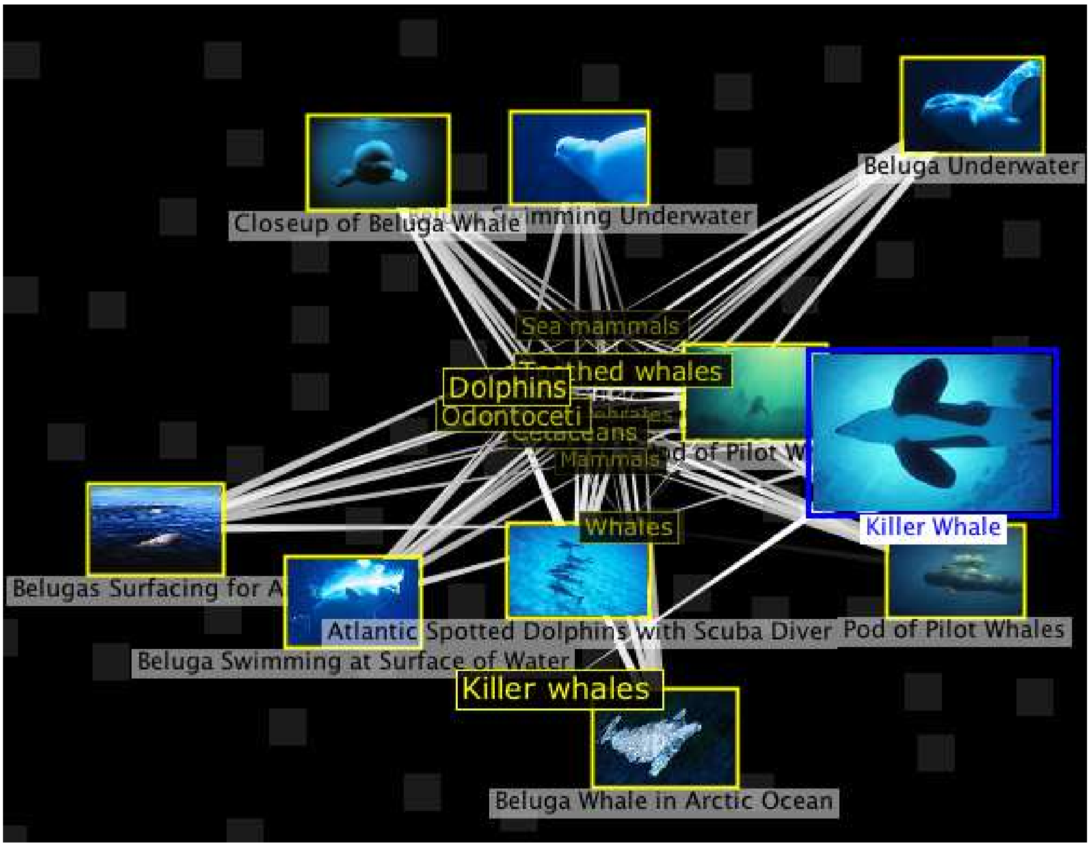
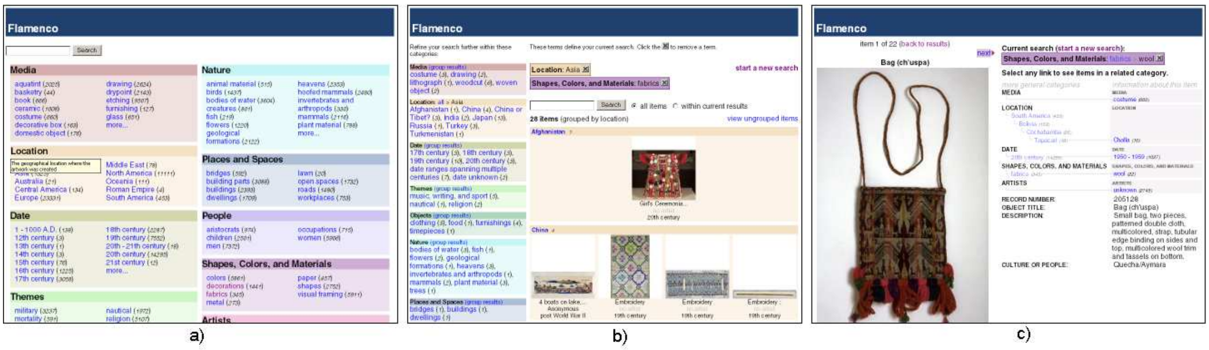


Figure 2.12: The system shows the similarity between an image and related keywords. Moreover, the most similar images are visually preponderant. *Image reproduced by permission of P. Janecek.*

Janecek [53] proposed a system based on fisheye views [34] for opportunistic searching and browsing in a collection of images, which have been manually labelled by professionals with textual metadata. Similarities between images correspond to the distance of the relative labels in WordNet [104] hierarchy. The images are rendered into a spring layout that determines their visual weight in respect of their relevance to the query (cf. Figure 2.12). Spring layout [33] is a dynamic visualization that represents the objects as nodes and their relationships as springs, which repulse or attract the nodes until a stable configuration is achieved. In this project, the springs are syntactic and semantic similarities between images, relative to the browsing context. This work proposes an elegant application of fisheye views, used for efficiently indicating documents relevance in respect of a query.



Refine your search further within these categories:

Media (group results)
costume (3), drawing (2), lithograph (1), woodcut (6), woven object (2)
Location: all > Asia
Afghanistan (1), China (4), China or Tibet? (3), India (2), Japan (13), Russia (1), Turkey (3), Turkmenistan (1)
Date (group results)
17th century (3), 18th century (3), 19th century (10), 20th century (3), date ranges spanning multiple centuries (7), date unknown (2)
Themes (group results)
music, writing, and sport (5), nautical (1), religion (2)
Objects (group results)
clothing (5), food (1), furnishings (4), timepieces (1)
Nature (group results)
bodies of water (3), fish (1), flowers (2), geological formations (1), heavens (3), invertebrates and arthropods (1), mammals (2), plant material (3), trees (1)
Places and Spaces (group results)
bridges (1), buildings (1), dwellings (1)

These terms define your current search. Click the ☒ to remove a term.

Location: Asia ☒

Shapes, Colors, and Materials: fabrics ☒

Current search (start a new search):

Shapes, Colors, and Materials: fabrics › wool ☒

Select any link to see items in a related category.

RECORD NUMBER: 205128

OBJECT TITLE: Bag (ch'uspa)

DESCRIPTION: Small bag, two pieces, patterned double cloth, multicolored, strap, tubular edge binding on sides and top, multicolored wool trim and tassels on bottom.

CULTURE OR PEOPLE: Quecha/Aymara

Figure 2.13: The sequence shows the browsing mechanism through facets. The system proposes different categories (a), which can be selected in order to reduce the result set (b) until a satisfying document is found (c). *Image reproduced by permission of M.Hearst [108].*

Flamenco [108] allows to navigate in a collection of labelled images through faceted metadata. The system automatically associates the images with various relevant categories, thanks to information contained in their metadata. The categories are organized in facets that contain related subcategories (see Figure 2.13a). The users interact with facets in order to browse: after each selection, the system shows new related subcategories and updates the visualized results, filtering images that do not correspond to the query (cf. Figure 2.13b). The current result set is one or more images, grouped per category (cf. Figure 2.13c). The solution proposed in this project is very interesting, but we believe that it limits the browsing possibilities. In fact, navigation through facets is a top-down approach that helps users to clearly identify the searched information, but that does not allow to change of thematic space while browsing. Moreover, the facets hide the relationships between documents, which are rather represented in various categories at the same time.

Video collage [25] is an interface for browsing collections of broadcasted news. Collages are presentations of video and text that summarize multiple video sources. Videos are decomposed in stories and shots and are enriched with metadata relative to shots of news anchorpersons, face detection results, and transcripts extracted thanks to video-closing captioning and speech recognition. Users can drill down in the collage and visualize new data at different levels of detail. Moreover, the interface proposes a map visualization that represents the distribution of geographic regions within one or more stories (confer Figure 2.14a). Further, the map can act as a filter for reducing the result set to shots concerning specific countries. A time line has been

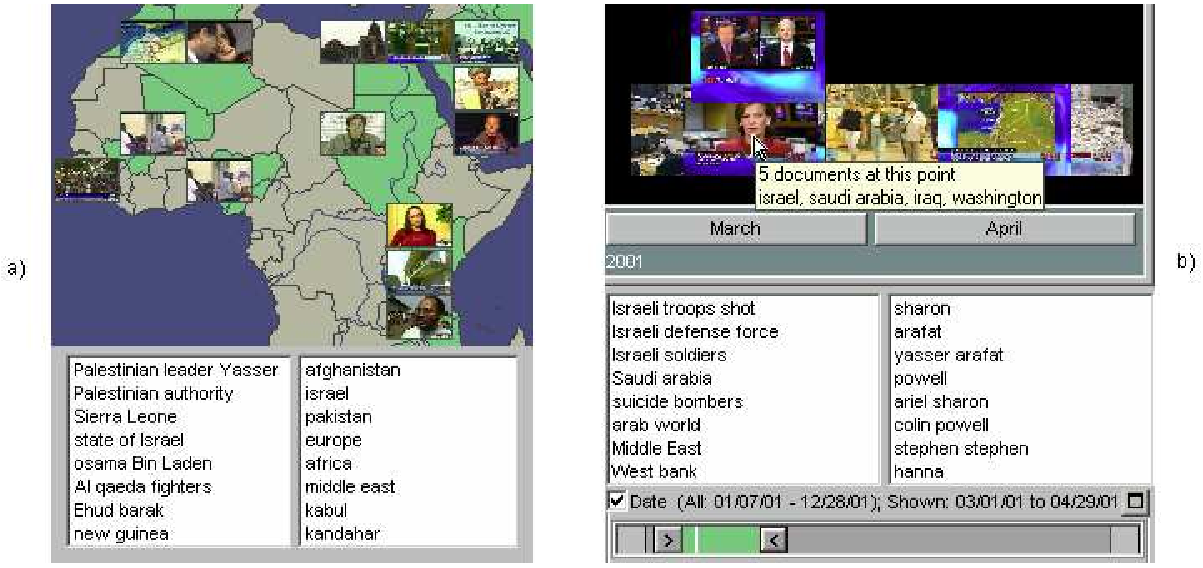
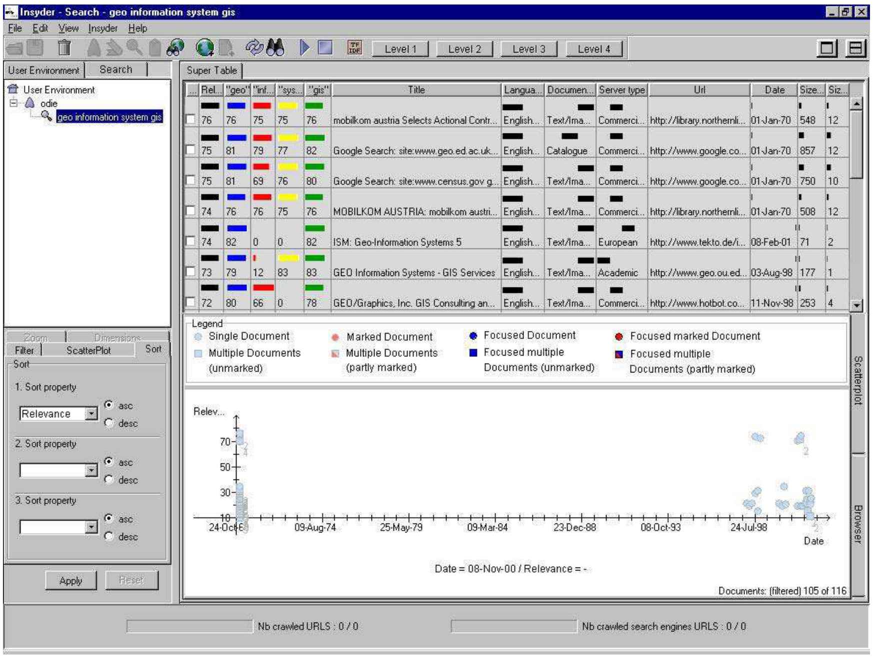


Figure 2.14: In (a), the map visualizes the video collages associated to geographic regions, whereas in (b) the collages are organized by time. *M. Christel [25], (c) 2002 ACM, Inc. Included here by permission. Thumbnails extracted from video provided for educational and research purposes only via Cable News Network, (c) 2001 CNN.*

integrated in order to define a temporal range for the collages (see Figure 2.14b). Finally, text lists allows to select a topic and to consult related videos. The approach is interesting because it combines in the same view different modalities, i.e. textual documents and videos. Moreover, the collage and the map are appropriate for browsing the archive. At opposite, the necessity of switching between two interfaces for browsing is a drawback that should be avoided.

INSYDER [58] is a *visual information retrieval* system that targets at presenting documents retrieved by queries and their metadata. The system provides an interface combining two linked visualizations, a ScatterPlot and a SuperTable (see Figure 2.15). The ScatterPlot allows users to overview query results and to manage them thanks to operations such as filtering, zooming, selecting, and so on. The SuperTable represents documents' characteristics such as date, title, size, content, etc. The main novelty of this representation is the access to results set at different levels of granularity, focusing one, more, or all documents at contemporary and setting the richness of displayed metadata (cf. Figure 2.16).

For instance, at the first level, users consult the whole result set, represented as bars indicating the global relevance of the document and the amount of text, whereas at higher level they can view labels with title, author, etc. and TileBars [44] with terms distribution. Finally, a supplemental browser can substitute the ScatterPlot, in order to view a document and its keywords. From our point of view, the interest of this project consists in the presentation of the archive at different levels of details. One or more documents can be zoomed, while the global



...	Rel...	"geo"	"inf...	"sys...	"gis"	Title	Langua...	Documen...	Server type	Url	Date	Size...	Siz...
□	76	76	75	75	76	mobilkom austria Selects Actional Contr...	English...	Text/Ima...	Commerci...	http://library.northernli...	01-Jan-70	548	12
□	75	81	79	77	82	Google Search: site:www.geo.ed.ac.uk...	English...	Catalogue	Commerci...	http://www.google.co...	01-Jan-70	857	12
□	75	81	69	76	80	Google Search: site:www.census.gov g...	English...	Text/Ima...	Commerci...	http://www.google.co...	01-Jan-70	750	10
□	74	76	76	75	76	MOBILKOM AUSTRIA: mobilkom austri...	English...	Text/Ima...	Commerci...	http://library.northernli...	01-Jan-70	508	12
□	74	82	0	0	82	ISM: Geo-Information Systems 5	English...	Text/Ima...	European	http://www.tekto.de/i...	08-Feb-01	71	2
□	73	79	12	83	83	GEO Information Systems - GIS Services	English...	Text/Ima...	Academic	http://www.geo.ou.ed...	03-Aug-98	177	1
□	72	80	66	0	78	GEO/Graphics, Inc. GIS Consulting an...	English...	Text/Ima...	Commerci...	http://www.hotbot.co...	11-Nov-98	253	4

Legend
- ● Single Document
- ● Marked Document
- ● Focused Document
- ● Focused marked Document
- ■ Multiple Documents (unmarked)
- ▨ Multiple Documents (partly marked)
- ■ Focused multiple Documents (unmarked)
- ▨ Focused multiple Documents (partly marked)

Sort

1. Sort property: Relevance ◉ asc ○ desc

2. Sort property: ◉ asc ○ desc

3. Sort property: ◉ asc ○ desc

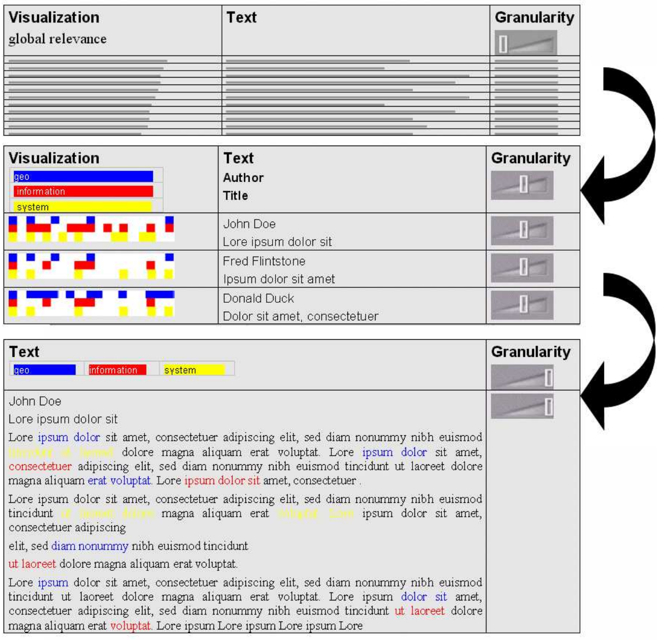
Date = 08-Nov-00 / Relevance = -

Documents: (filtered) 105 of 116

Nb crawled URLS : 0 / 0

Nb crawled search engines URLS : 0 / 0

Figure 2.15: The screenshot is an overview of SuperTable + ScatterPlot interface. *P. Klein [58], ©2002 IEEE.*



Text	
John Doe Lore ipsum dolor sit	
Fred Flintstone Ipsum dolor sit amet	
Donald Duck Dolor sit amet, consectetuer	

John Doe

Lore ipsum dolor sit

Lore ipsum dolor sit amet, consectetuer adipiscing elit, sed diam nonummy nibh euismod tincidunt ut laoreet dolore magna aliquam erat voluptat. Lore ipsum dolor sit amet, consectetuer adipiscing elit, sed diam nonummy nibh euismod tincidunt ut laoreet dolore magna aliquam erat voluptat. Lore ipsum dolor sit amet, consectetuer .

Lore ipsum dolor sit amet, consectetuer adipiscing elit, sed diam nonummy nibh euismod tincidunt ut laoreet dolore magna aliquam erat voluptat. Lore ipsum dolor sit amet, consectetuer adipiscing

elit, sed diam nonummy nibh euismod tincidunt

ut laoreet dolore magna aliquam erat voluptat.

Lore ipsum dolor sit amet, consectetuer adipiscing elit, sed diam nonummy nibh euismod tincidunt ut laoreet dolore magna aliquam erat voluptat. Lore ipsum dolor sit amet, consectetuer adipiscing elit, sed diam nonummy nibh euismod tincidunt ut laoreet dolore magna aliquam erat voluptat. Lore ipsum Lore ipsum Lore ipsum Lore

Figure 2.16: The sequence shows the different levels of zoom into results, from the entire set until a specific document preview. *P. Klein [58], ©2002 IEEE.*

context is preserved at different levels of granularity. For instance, this mechanism is interesting for zooming into one or more meetings, then into documents, and finally into the annotations of a specific document.

All the projects presented in this section use visualization in order to correlate information in large archives of documents. In fact, a visual overview of data can be interpreted very fast by humans and, then, manipulated in order to retrieve precise information on a specific document. This principle is summarized in Shneiderman's mantra *Overview first, zoom and filter, then details-on-demand* [88]. Presented projects are based on different principle for correlating information: evolution of data in time [25, 43], similarity of semantic content [20, 108, 53, 58, 73, 108], inner structures of the document [75] and so on. From an analytical and general point of view, it is inappropriate to talk about lacks in them: visualizations are difficult to compare and their efficiency depends on both the task to be performed and users' perception. However, two considerations motivate the exploitation or the development of infovis techniques in this thesis: 1) existing projects dealing with large datasets either do not take into account multimedia information or separate media in different views [25], and 2) until now no research tackled the problem of visualizing events archives for eliciting their hidden relationships.

2.5 Major Contributions of This Thesis

The major contributions of this thesis are the development of new techniques in order to 1) structure multimedia archives of events and 2) support users in accessing and browsing within this data. The achievement of this task implies that several existing methods presented in this chapter have been used or extended and that new technologies have been developed. For each presented field the contributions of this thesis are listed below:

- **Multimedia Indexing and Management**. The first unavoidable contribution is a model for representing multimedia documents, their metadata, annotations and cross-relationships (Chapter 3). Further, we propose a complete system that uses multimodal alignment techniques in order to automatically annotate and index all categories of multimedia documents (Chapters 4 and 5). The architecture of the system takes into account various aspects such as flexibility, extensibility and modularity (Section 4.1).

- **Event Indexing and Browsing**. The model proposed in Chapter 3 has been applied in order to structure archives of events. Our technique for indexing events is centered on static documents, which are generally ignored or marginally integrated, and takes full benefits from document's structures (Subsection 4.2.1). Further, the originality of our work consist in considering meeting as correlated events, either from thematic or temporal points of view, instead of isolated entities (Chapter 5). Finally, we propose a new document-centric browser in order to navigate into one event (Section 6.2).

- **Static Document Analysis**. In all the systems developed all along this thesis, static documents play a major role. Consequently, a new tool able to extract textual content as well as physical structures from PDF files has been developed. Its main purpose is the restructuring of an original PDF document in a XML format, which allows to annotate the original content with structural information, easily accessible by users and computers (Section 4.2).

- **Information Visualization for Multimedia Data**. We believe that a complete and efficient system for multimedia data retrieval requires the collaboration of both users and algorithms. Users are the only ones that know which information satisfies their needs. Thus, another contribution of this thesis is the use of information visualization techniques in order to represents events and their correlations (Chapter 6). More precisely, the proposed visualizations are interactive and allow users to browse by links into an archive of multimedia documents. Moreover, our interactive visualizations fulfill a lack in the domain of multimedia digital libraries.

Chapter 3

A General Model for Multimedia Archives

The definition of models for representing multimedia data is a delicate task, which has to take into account formal criteria such as clarity, completeness and so on. This chapter presents the solution we propose in order to model and organize multimedia data. In Section 3.1, we present our novel approach for structuring and browsing multimedia archives, which is the cornerstone behind the model. This model is described in Section 3.2 that illustrates data representation. Finally, Section 3.3 applies the model to the use case of meetings archives.

3.1 A New Approach for Structuring and Browsing in Multimedia Archives

The extraction of semantic knowledge from multimedia documents and the successive creation of indexes depend on individual media types. Unfortunately, the quality and the typology of results provided by automatic analysis techniques are not homogenous between the different media. For instance, adding annotations on static documents is generally easier than automatically extracting high level abstractions from video and audio documents. At opposite, textual documents do not provide the temporal information that is contained in video and audio documents. Thus, we believe that structuring multimedia archives is not limited to the application of monomodal analysis, but it must also take into account the explicit and implicit relationships between the different media. Multimedia archives are not composed of individual documents, but of groups of related documents that coexist and share thematic content, geographic properties and so on. For instance, all the documents presented or recorded during an event are explicitly correlated by time. Similarly, all the events talking about the same topics are related by themes. For this reason, we propose to apply multimodal alignment techniques [68] over the whole collection of events, in order to enable the transfer of indexes between multimedia documents and to elicit their correlations. In particular, we propose to use information extracted from static documents to enrich the high-level abstractions of other categories of media. Static documents are naturally and implicitly structured, rich of semantics, and more planned than audio transcripts and other textual information. Moreover, existing techniques for annotating and indexing static documents produce results equivalent to humans' manipulations.

Such manner of structuring multimedia archives is further well suited for enhancing browsing mechanism. Furnas [57] defines browsing as the task that consists in "looking to see what is available in the world". This definition implies that users do not have the exact knowledge of expected documents, but a more or less vague idea to clarify while navigating. The opposite task is searching, in which users look for a "known target". In general, existing engines such as Google [36] only support searching functionality and users retrieve interesting information after various refinements of their queries. Our model for structuring multimedia archives, which explicitly describes the relationships between documents using links, is well adapted for proposing a new approach of navigation that permits both searching and browsing tasks. The links allow to guide the user in the browsing task, because they represent similarity of documents in the archive. For instance, users could consult a document, choose an interesting paragraph and then follow its links for discovering similar documents from a thematic or temporal point of view. This process iteratively continues until the user retrieves the desired information. Moreover, the links facilitate the access to media that are poor of textual indexes and difficult to retrieve by searching. For example, when a user searches for videos but does not knows their labels, she could first find static documents that are correlated and secondly access all videos linked to them. Section 3.2 describes the model that matches this approach and that allows to explicitly represent the links existing in multimedia archives.

3.2 A Relational and Extensible Multimedia Aggregatio

This section presents our model describing in details the representation of the documents in a multimedia collection. An "*archive*" is a simple container of information such as events or, from a more general point of view, multimedia documents, i.e. static documents, video and audio streams, etc. When these data are enriched with additional metadata and annotations, the dataset is called a "*corpus*". In our model, we introduce the links, which are a new category of storable information, and we define the entire collection as being an "*aggregatio*" (pl. "aggregationi"). Thus, in the rest of this thesis the term "aggregatio" is referring to multimedia collections of "*aggregati*" (sg. "aggregatus"), i.e. documents enriched with metadata, annotations and links. Figure 3.1 is an overview of the model.

The structure of an aggregatus is presented in Subsection 3.2.1, whereas the rest of the section describes in details the documents included in an aggregatio (Subsection 3.2.2), the annotations (Subsection 3.2.3) and the links (Subsection 3.2.4).

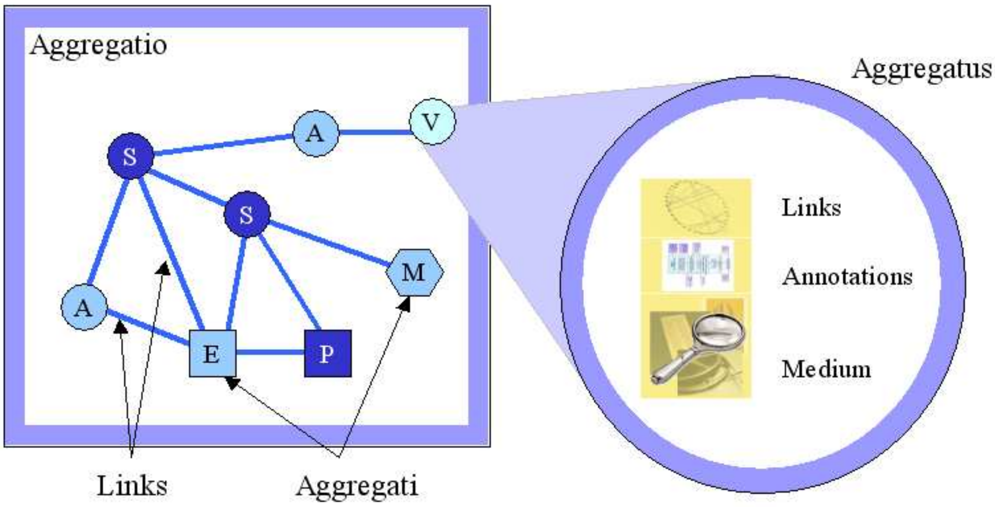


Figure 3.1: The model represents a multimedia archive in the form of an aggregatio. The multimedia documents are encapsulated within aggregati, which contain their links and annotations. The labels "A", "S", "V", "M", "P", and "E" on the shapes respectively stand for "audio", "static document", "video", "multimedia document", "person recording", and "event".

3.2.1 An Aggregatus: Document, Annotations and Links

An aggregatus belonging to the aggregatio is a triptych of document, annotations and links. A *document* is a medium, such as a static document, an audio recording, an image, or a video, and contains exclusively the original message to be transmitted to the users in its raw form. Multimedia documents such as websites and slideshows are composed of multiple data sources. The *annotations* are supplemental information augmenting the document, either added manually by users or computed automatically thanks to analysis techniques. Each medium can contain one or more annotations, which allow to access data in a structured manner and, consequently, to create more precise indexes [18]. For instance, static documents may be annotated with their logical structures: two terms are thematically more significant when belonging to the same document's title rather than to the entire stream of words. Moreover, annotations can be used for augmenting documents that are poor in textual descriptions, such as videos and audio files. Finally, the *links* are calculated thanks to annotations and bring out the relationships between a document and another one contained into the aggregatio. Each aggregatus possesses an undefined amount of links toward one or more aggregati.

3.2.2 Documents Categories

The model takes into account three categories of document: media, multimedia documents and pseudo-documents.

- *Media* are instances of static documents (as defined in Section 2.3, e.g. newspapers, articles, emails, etc.), images, and raw stream of information corresponding to video and audio recordings.

- *Multimedia documents* are composed of two or more categories of media. For instance, slideshows and websites are multimedia documents if containing textual content mixed with video or audio streams.

- *Pseudo-documents* contain metadata and descriptors instead of data streams. On one hand, pseudo-documents regroup media and multimedia in events such as meetings, conferences and lectures. In this case, the pseudo-document implies the coexistence of documents and, thus, it defines their temporal relationships. On the other hand, a pseudo-document describes identities and personal information of groups and persons. For instance, people are filmed in videos, participate to events as speakers, or produce and edit the documents in the aggregatio.

3.2.3 Annotations

Although in this thesis the term "annotation" is used for indicating any kind of additional information augmenting the documents, we distinguish properly said annotations from metadata. Both categories have not yet been differentiated with universally accepted definitions. For instance, Popescu-Belis proposes that metadata are static information about an event or a document, whereas annotations are time-dependent information related to input media [78]. In the DIVA group at University of Fribourg, we make a more general distinction:

- Metadata are extra-information that is not present or directly deductible from the content of documents and that belongs to another level of abstraction (for instance, the author, a date, etc.);

- Annotations are directly extractible from the content of the document and augment its original information.

We consider that both metadata and annotations are either manually added by users or automatically calculated. In our model, we take into account different types of annotations and metadata for each category of document presented in the previous subsection. *Static documents* are mainly annotated with their physical and logical structures. For instance, physical structures describe the layout with textual blocs and line, while logical structures explain their hierarchy and logical functions. Both annotations segment the static documents in thematic fragments. *Images* are augmented with annotations describing either structural information, i.e. colors and patterns, or semantics, e.g. content, symbols, etc. *Videos* are annotated with information about the related audio track, recording time, length, etc. Moreover, video's content is annotated with thematic information. A video could even be segmented into time-stamped scenes and shots.

Audio recordings are enriched with temporally segmented textual transcripts. These annotations have different levels of granularity and can represent words, utterances, thematic episodes, and so on. In general, transcription parts are characterized by a start and an end timestamp, the name (or the id) of the speakers, and the textual content. Annotations on audio are rich of temporal and thematic information.

Multimedia documents are obviously enriched with all the annotations and metadata that can be extracted from individual media. *Events* are annotated with metadata containing the location, the day and the hour of the recording. The descriptors are often completed with an event's title, type (i.e. belonging or not to a scenario), and participants. Last but not least, annotations contain references towards all the documents presented or recorded throughout the events. Finally, *persons* are registered and entirely defined by metadata, which are personal and descriptive information such as first and last name, email, and telephone number. Moreover, persons are annotated with information about produced and edited documents, navigation profiles, etc.

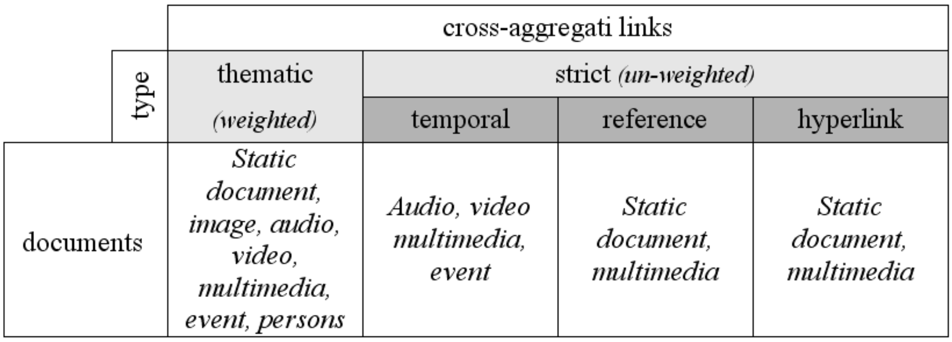
3.2.4 Links Between Aggregati

Links are a natural way for representing relationships between aggregati. Recently, they have been employed successfully in systems such as Citeseer [13] and LinkedIn [65]. In Citeseer, links cross-connect scientific publications that share citations, authors, etc. Similarly, LinkedIn creates communities profiting of social relationships between subscribers. Thus, we believe that links could be useful for structuring multimedia data and for enhancing browsing mechanism.

In our model, we distinguish two families of links: *intra-aggregatus* and *cross-aggregati* links, which interconnect two annotations belonging respectively to the same aggregratus or to two different aggregati. For instance, in static documents, intra-aggregatus links connect logical articles with physical blocks. At opposite, a cross-aggregati link represents for instance the thematic correspondence between an article and video shots. In our model, we define a unique type of intra-aggregatus links, but we distinguish different types of cross-aggregati links [80, 84]. These types are summarized in Table 3.1, as a function of documents categories producing them.

Thematic links are typically calculated using multimodal alignment techniques [68]: each aggregatus is compared with other ones into the aggregatio, by accessing the structured content of documents through annotations. For instance, each physical bloc of a static document is compared with each utterance of an audio recording's transcript. If the comparison detects a similarity, the documents are aligned with a thematic link.

Temporal links are generated for time-based documents. For instance, an event contains several documents discussed, viewed, or created by participants [63]: all the documents coexist at the same moment and, consequently, are temporally related. Thus, non time-based data such as static documents and user records acquire a temporal dimension thanks to links. Moreover, temporal links can be combined with thematic links to achieve a finer granularity: for instance,



		cross-aggregati links			
	type	thematic	strict *(un-weighted)*		
		(weighted)	temporal	reference	hyperlink
documents		*Static document, image, audio, video, multimedia, event, persons*	*Audio, video multimedia, event*	*Static document, multimedia*	*Static document, multimedia*

Table 3.1: Each category of documents generates specific links.

when a paragraph in a static document is thematically linked with a speech's utterance, the former is also synchronized with the timestamp of the latter.

References are links to other documents such as bibliographical entries, citations, and so on. In this thesis, we consider that *hyperlinks* are artifacts for navigation created by persons. Both references and hyperlinks are extracted from textual documents and web pages by lexical analysis.

The aggregati model presented in this chapter can be easily extended with new categories of documents, annotations and links. Section 3.3 presents the application of the multimedia model to the meetings use case.

3.3 Use Case: Events Recordings

The model presented in this chapter has been applied to events and, more precisely, to meetings recordings. Meetings are an interesting type of events, which contains documents belonging to all the categories presented in Subsection 3.2.2 and sharing several relationships among them. During a single meeting, presented topics can dynamically change and evolve. Moreover, various meetings can be strictly related, for example by being subsequent, involving the same participants, or by focusing on similar topics.

The following users' tasks rise up the importance of structuring and indexing meetings corpora:

- *Replaying.* Persons that did not participate to the meeting replay the whole recorded information, which is consulted following the order in which documents were presented. While users replay a meeting, the recorded data must be synchronously reproduced. Moreover, if a user focuses on a document presented at a precise moment, all meeting resources are expected to shift at the same time;

- *Information retrieval.* Users search and browse in order to retrieve a specific information, without knowing a priori which document contains it. Searching and browsing tasks require that multimedia data have been annotated with textual information, linked, and indexed;

- *Minutes creation.* Persons consult the recordings of a meeting in order to prepare its multimedia summary. It can be composed of partial copies of existing documents, hyperlinks toward existing data, or of new information created by users;

- *Edition.* Participants edit the data presented during the meeting or correct the annotations and metadata produced after the event. These manipulations should be allowed when users need to finalize incomplete documents, when automatic systems make errors during the recording or the extraction of high-level abstractions, and so on;

- *Survey.* Users may study how topics of interest evolve throughout time into several meetings. Such task implies that users can access more meetings at contemporary. Specialized views of the meetings are necessary for facilitating this task;

- *Meeting preparation.* When users participate to various related sessions, they consult old meetings recordings, in order to create or update the documents to present in the successive sessions. This task can be considered a mix of minutes creation, edition, and survey.

3.4 Conclusion

In this chapter, we have presented the general model for representing multimedia documents and the relationships existing between them. In particular, we have introduced the concepts of "aggregatus" and "aggregatio".

An aggregatus contains a document, its annotations, and the links. A document can be a medium (e.g. a static document, a video, etc.), a multimedia document, or a pseudo-document, such as a meeting or person recordings. The documents are enriched with additional information: annotations or metadata. This supplemental information is used for enriching the document, accessing its content, and for creating links. The links belong to two main categories: intra-aggregatus and cross-aggregati. Intra-aggregatus links represent the relationships between annotations of a same document, whereas cross-aggregati links connect together two aggregati, in order to elicit their temporal, thematic, etc. similarity.

The aggregatio is a collection of aggregati, which are connected through cross-aggregati links.

In the next three chapters, we present respectively how to analyze the data in order to create aggregati from meeting recordings (Chapter 4), the indexing of the meetings aggregatio (Chapter 5) and how support users in the task described in the use cases above (Chapter 6).

Chapter 4

Analysis of Document-centric Meetings

The analysis of multimedia data is the process that produces both the annotations over documents and the links between them. In this chapter, we present the technologies that act at different levels of granularity, from analysis of static documents until the whole meetings collection, in order to create aggregati as presented in Chapter 3 and to prepare multimedia data for indexing.

The analysis of multimedia data has to deal with two major challenges:

- The heterogeneity of documents implies that several technologies must be integrated and combined for working together;
- Although analysis methods can be performing and reliable, a full automatic system does never obtain perfect results. Moreover, applying sequential techniques provokes errors that are cumulated from one analysis step to the following one. Consequently, tools for supervising the analysis processes and for preventing error propagation are necessary.

These two aspects are tackled in the different sections of this chapter. Firstly, Section 4.1 gives an overview of the different technologies involved in our analysis system. Then, Section 4.2 focuses on the analysis of static documents and, in particular, of PDF files. Finally, Section 4.3 investigates events analysis.

4.1 Functionalities and Architecture

The analysis of meeting collections allows to create the aggregationi as described in Section 3.2. The analysis acts at three levels: intra-document level, intra-event level and cross-event level. The successive steps involved in the analysis of a meetings collection are illustrated in Figure 4.1.

At *intra-document level*, mining techniques are monomodal and target at annotating the documents with high-level information. In this thesis, we mainly focus on the analysis of static documents, which are annotated with their physical and logical structures. Since logical structures

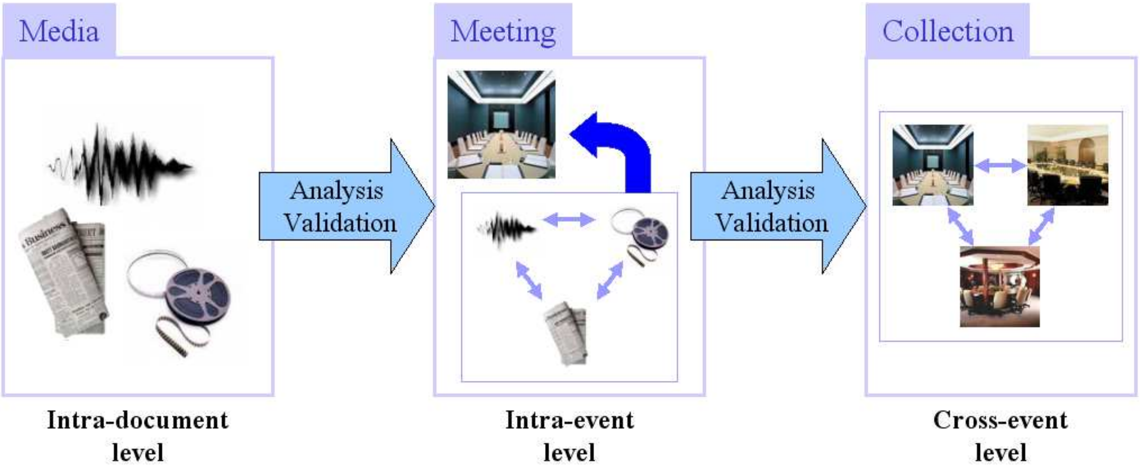


Figure 4.1: The analysis is performed at different levels of granularity. First, mining techniques are applied to the media. Then, their results are involved in the analysis of a meeting. Finally, the whole collection of meeting is analyzed.

rely on physical structures, e.g. an article is composed of multiple text blocks, intra-aggregatus links are created to connect those annotations. The final step at intra-document level consists in validating or correcting annotations and links with interactive tools, in order to avoid the propagation of errors to the successive levels of analysis.

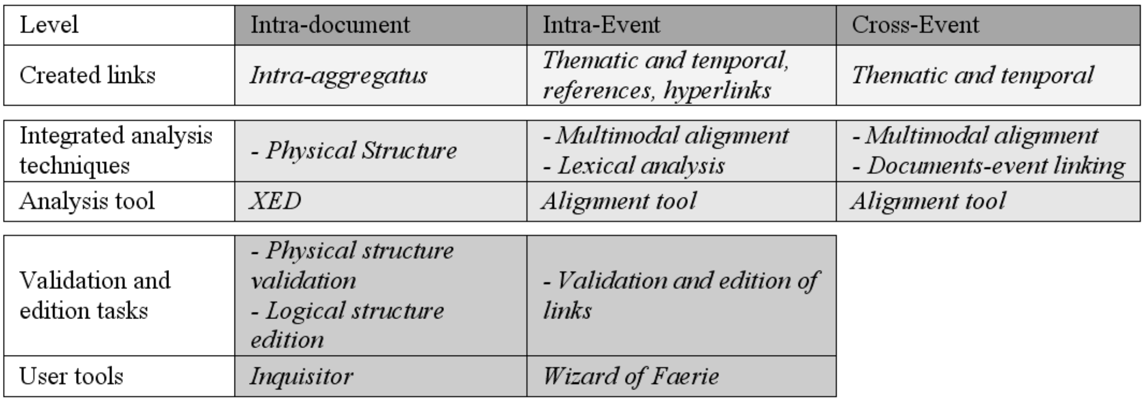
At *intra-event level*, the results of monomodal analysis techniques are used for structuring an event. Firstly, the physical and logical structures of static documents are compared with audio recordings' transcript, in order to create temporal and thematic links. The comparison is accomplished through multimodal alignment technique: each textual block is compared with each dialog part and, if they are similar, they are linked together. Multimodal alignment produces both thematic and temporal links, which represent respectively thematic similarity and media synchronization.

Secondly, lexical analysis is applied to the annotations of video and audio recordings, in order to extract participants' identities and to thematically link videos and audio recordings. Furthermore, lexical analysis allows to extract references and hyperlinks from audio transcript and static documents.

Thirdly, the users can correct or delete the calculated links using interactive tools. New links can also be created when analysis techniques did not discover relationships between the multimedia documents belonging to the event.

Finally, at *cross-event level*, the multimodal alignment is applied over the whole collection in order to create thematic links between events. Temporal links are also created between each meeting and the multimedia documents presented or recorded during this event.

Table 4.1 summarizes the analysis chain presented above and introduces the systems im-



Level	Intra-document	Intra-Event	Cross-Event
Created links	*Intra-aggregatus*	*Thematic and temporal, references, hyperlinks*	*Thematic and temporal*
Integrated analysis techniques	*- Physical Structure*	*- Multimodal alignment* *- Lexical analysis*	*- Multimodal alignment* *- Documents-event linking*
Analysis tool	*XED*	*Alignment tool*	*Alignment tool*
Validation and edition tasks	*- Physical structure validation* *- Logical structure edition*	*- Validation and edition of links*	
User tools	*Inquisitor*	*Wizard of Faerie*	

Table 4.1: The table summarizes the types of links created during the analysis of meetings collections, the integrated analysis techniques and the involved supervision task. Developed tools and systems are also enumerated.

plementing it. These systems are presented in details in the following sections. In particular, Section 4.2 presents static document analysis, the first step of the whole process.

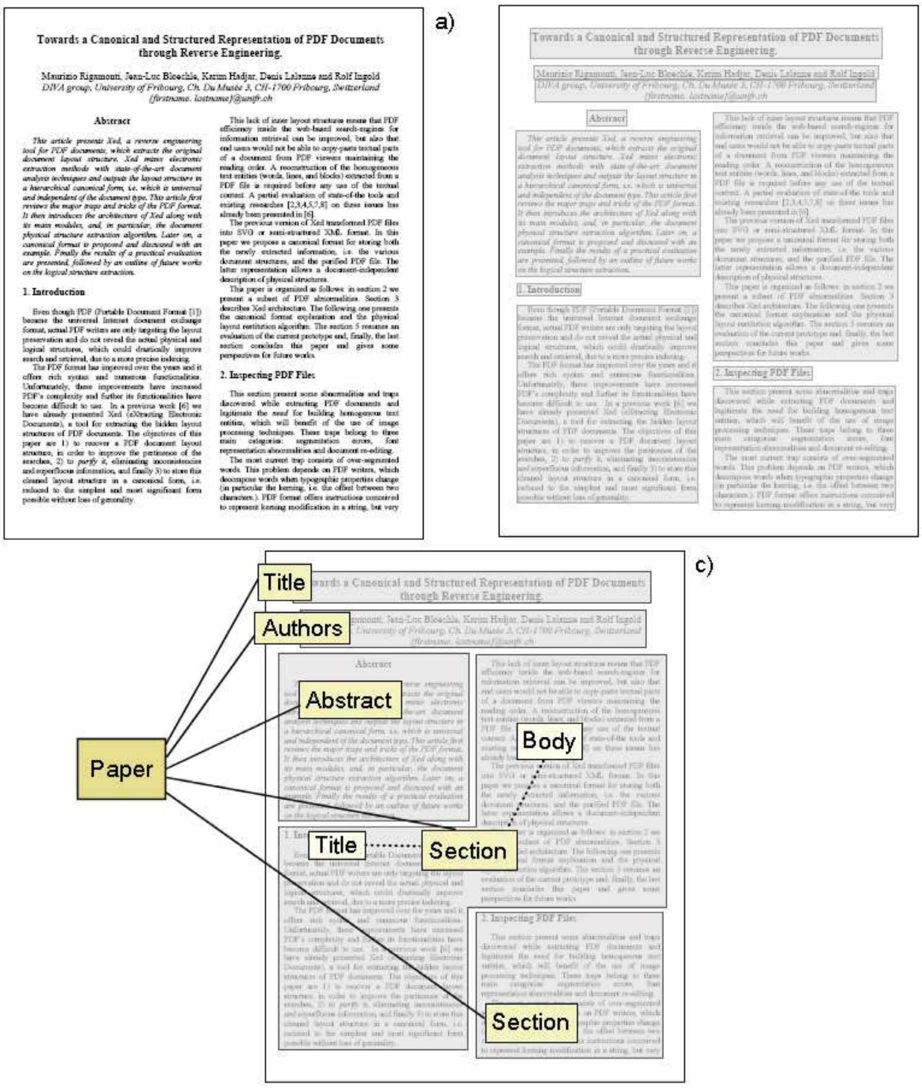
4.2 Static Documents Analysis

This section discusses the analysis at intra-document level and is dedicated in particular to the analysis of static documents, as defined in Section 2.3. In Subsection 4.2.1, we advocate the importance of annotating static documents with their structures. Subsection 4.2.2 illustrates the analysis of PDF files, a specific and standardized type of static documents. In Subsection 4.2.3, we describe Xed, our tool for annotating PDF files. Finally, Subsection 4.2.4 presents Inquisitor, an interactive tool for validating and correcting analysis results.

4.2.1 The Relevance of Document Structuring

Annotating static documents with physical and logical structures allows to access their content in a structured manner. In fact, these structures allow to recover the original semantics of the homogenous text blocks. For instance, the logical structures define the hierarchy of text in an article, in which title's words are more significant than those in the body. Figure 4.2 shows an example of document (a), with its physical (b) and logical (c) structures superposed.

The physical structures represent words regrouped in lines, which similarly compose homogenous text blocks. Each word is labelled with its syntactic function, e.g. "term", "number", "punctuation" and so on. These structures are not dependent on document class, for instance "newspaper" or "scientific paper", and they can represent each document in a unique manner. Today, methods for automatic extraction of physical structures are performing well and their efficiency is comparable to that of humans.



Towards a Canonical and Structured Representation of PDF Documents through Reverse Engineering.

Maurizio Rigamonti, Jean-Luc Bloechle, Karim Hadjar, Denis Lalanne and Rolf Ingold
DIVA group, University of Fribourg, Ch. Du Musée 3, CH-1700 Fribourg, Switzerland
{firstname. lastname}@unifr.ch

Abstract

This article presents Xed, a reverse engineering tool for PDF documents, which extracts the original document layout structure. Xed mixes electronic extraction methods with state-of-the-art document analysis techniques and outputs the layout structure in a hierarchical canonical form, i.e. which is universal and independent of the document type. This article first reviews the major traps and tricks of the PDF format. It then introduces the architecture of Xed along with its main modules, and, in particular, the document physical structure extraction algorithm. Later on, a canonical format is proposed and discussed with an example. Finally the results of a practical evaluation are presented, followed by an outline of future works on the logical structure extraction.

1. Introduction

Even though PDF (Portable Document Format [1]) became the universal Internet document exchange format, actual PDF writers are only targeting the layout preservation and do not reveal the actual physical and logical structures, which could drastically improve search and retrieval, due to a more precise indexing.

The PDF format has improved over the years and it offers rich syntax and numerous functionalities. Unfortunately, these improvements have increased PDF's complexity and further its functionalities have become difficult to use. In a previous work [6] we have already presented Xed (eXtracting Electronic Documents), a tool for extracting the hidden layout structures of PDF documents. The objectives of this paper are 1) to recover a PDF document layout structure, in order to improve the pertinence of the searches, 2) to *purify it*, eliminating inconsistencies and superfluous information, and finally 3) to store this cleaned layout structure in a canonical form, i.e. reduced to the simplest and most significant form possible without loss of generality.

This lack of inner layout structures means that PDF efficiency inside the web-based search-engines for information retrieval can be improved, but also that end users would not be able to copy-paste textual parts of a document from PDF viewers maintaining the reading order. A reconstruction of the homogeneous text entities (words, lines, and blocks) extracted from a PDF file is required before any use of the textual content. A partial evaluation of state-of-the tools and existing researches [2,3,4,5,7,8] on these issues has already been presented in [6].

The previous version of Xed transformed PDF files into SVG or semi-structured XML format. In this paper we propose a canonical format for storing both the newly extracted information, i.e. the various document structures, and the purified PDF file. The latter representation allows a document-independent description of physical structures.

This paper is organized as follows: in section 2 we present a subset of PDF abnormalities. Section 3 describes Xed architecture. The following one presents the canonical format explanation and the physical layout restitution algorithm. The section 5 resumes an evaluation of the current prototype and, finally, the last section concludes this paper and gives some perspectives for future works.

2. Inspecting PDF Files

This section present some abnormalities and traps discovered while extracting PDF documents and legitimate the *need* for building homogeneous text entities, which will benefit of the use of image processing techniques. These traps belong to three main categories: segmentation errors, font representation abnormalities and document re-editing.

The most current trap consists of over-segmented words. This problem depends on PDF writers, which decompose words when typographic properties change (in particular the kerning, i.e. the offset between two characters.). PDF format offers instructions conceived to represent kerning modification in a string, but very

Figure 4.2: A scientific paper (a) is enriched with physical (b) and logical (c) structures.

Logical structures are based on physical structures, of which they describe the logical functions. We consider that extracting logical structures is a two steps process: firstly, the physical entities are labelled with their functions and, secondly, they are organized in a hierarchy. For instance, physical text blocks extracted from a newspaper can be labelled as "title", "author", or "body" and then grouped in articles. The logical structures are described by *models*, strictly related to a document class or sub-class. For example, models for newspapers do not correspond to those for scientific papers.

A static document can be structured in several manners, using different logical models. For instance, a user can consider the titles of a book as being either titles of chapters or items of a table of contents. Consequently, the automatic systems for the recognition of logical structures are configured with one or more specific models.

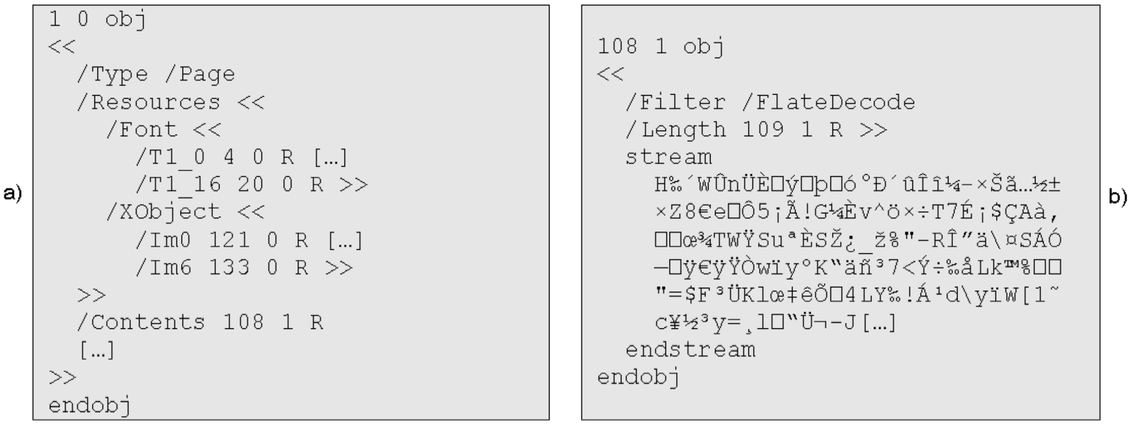
In this thesis, we mainly take into account automatic extraction of physical structures, in order to analyze each type of documents, despite of its class. These physical structures are useful to fragment the static documents [18], i.e. to decompose them in tiles corresponding to homogenous text blocks. The fragments permit a more precise indexing than the whole document, because they convey the proximity of words into text blocks.

4.2.2 The PDF Format

In our work [11, 12, 41, 81, 80], we have analyzed static documents represented in PDF format, because nowadays most of static documents exist in this form. *Portable Document Format* [51] has become the de facto standard for exchanging electronic documents through the Internet and between persons. In the last years, this format has been constantly improved and nowadays it is so powerful to represent each class of documents. Unfortunately, PDF has three main drawbacks.

Firstly, the format has been specialized for the visual representation of the documents, but it does not preserve any information about physical or logical structures. In recent past, some works [8] tried to solve this problem, by adding specific operators for including annotations in PDF files, but usually tools and users producing PDF ignore these options.

Secondly, the extension of the format has increased the amount of operators, sometimes redundant, which actually can be only managed by automatic generators and interpreters of PDF files. The mix of ASCII and binary information is too complex for allowing end users to directly manipulate PDF documents (cf. Figure 4.3).



a)

```
1 0 obj
<<
  /Type /Page
  /Resources <<
    /Font <<
      /T1_0 4 0 R […]
      /T1_16 20 0 R >>
    /XObject <<
      /Im0 121 0 R […]
      /Im6 133 0 R >>
  >>
  /Contents 108 1 R
  […]
>>
endobj
```

b)

```
108 1 obj
<<
  /Filter /FlateDecode
  /Length 109 1 R >>
  stream
    H‰´WÛnÜÈ□ý□þ□ó°Ð´ûÎî¼–×Šã…½±
    ×Z8€e□Ô5¡Ã!G¼Èv^ö×÷T7É¡$ÇAà,
    □□œ¾TWŸSuªÈSŽ¿_ž%"-RÎ”ä\¤SÁÓ
    —□ÿ€ÿŸÒwïy°K“äñ³7<Ý÷‰åLk™%□□
    "=$F³ÜKlœ‡êÕ□4LY‰!Á¹d\yïW[1˜
    c¥½³y=¸l□“Ü¬-J[…]
  endstream
endobj
```

Figure 4.3: The PDF streams in ASCII (a) and primitives in binary format (b).

Thirdly, PDF has been thought and optimized for transmitting information through the Internet. This optimization involves an over-segmentation and a reorganization of files' content, for broadcasting purposes.

The consequence of these drawbacks is the difficulty of accessing document's content, because PDF syntax is complex to interpret and because generators provoke several troubles. Consequently, search engines for information retrieval can not exploit the full content of PDF documents and frequently end users are not able to copy-paste textual and graphical primitives. The main troubles in PDF files are detailed in the first part of this subsection, whereas in the second part we propose a solution for efficiently managing these documents using restructuring techniques.

PDF Troubles

We classify the main troubles encountered while analyzing PDF files in three categories: segmentation errors, abnormalities in font representation and re-editing incoherencies. The most recurrent trouble consists in over-segmentation of words. In fact, PDF generators target at only preserving the visual appearance of documents and split the words when typographic properties change. For instance, a term is segmented when the kerning, i.e. the distance between two terms or characters, is modified. Moreover, sometimes PDF producers add supplemental blank characters, which do not exist in the original text.

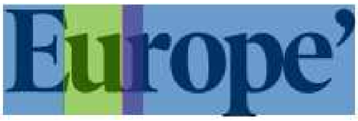


Figure 4.4: The word *Europe'* has been over-segmented and a blank space added.

Figure 4.4 is an example of word extracted from the "International Herald Tribune" newspaper (February 21, 2005) that illustrates both these abnormalities. The word is over-segmented in three tokens, i.e. "E", "u", and "rope'", and a blank space has been added between the characters. In Figure 4.4, the dark rectangle overlapped with "r" corresponds to the bounding box of this supplemental blank.

Another frequent inconsistency is the re-mapping of fonts. The PDF format is conceived for encapsulating proprietary fonts in the raw document file. This feature permits to represent typographic information and characters descriptions, which are composed of numeric codes (e.g. 100), glyphs' names (in this case, "Dsmall") and charstrings (i.e. the graphical representation of the printed character). In several documents, the generators substitute the numeric codes defined by standard encodings with custom values. For instance, the code 100 that usually represents the "d" character is associated to "e". Consequently, the original content is not more accessible, although the character is correctly visualized and printed. This inconsistency frequently occurs in "Le Monde" newspapers, in which advertising's original text is hidden.

People that add unmeaning information provoke the last category of troubles. For example, users of WYSIWYG [1] editors finalize the layout of their documents adding blank spaces, instead of defining or modifying the paragraph's properties. The generators of PDF files do not filter this information, but they store it as meaningful content. Consequently, the blanks are not visible at screen, but they complicates the extraction of content while indexing or copy-pasting the text of PDF files.

Other minor troubles, encountered while dealing with PDF format, consist in some inconsistent implementations relative to the official reference [51], in non-meaning and invisible objects contained in the documents such as empty frames, and so on. All these troubles demonstrate that accessing the content and the structures of PDF documents is a complex task. Thus, we propose to apply restructuring techniques over PDF files, in order to 1) correct their inconsistencies and 2) annotate their original content with high-level information, well-suited for indexing process.

From PDF to Structured XML

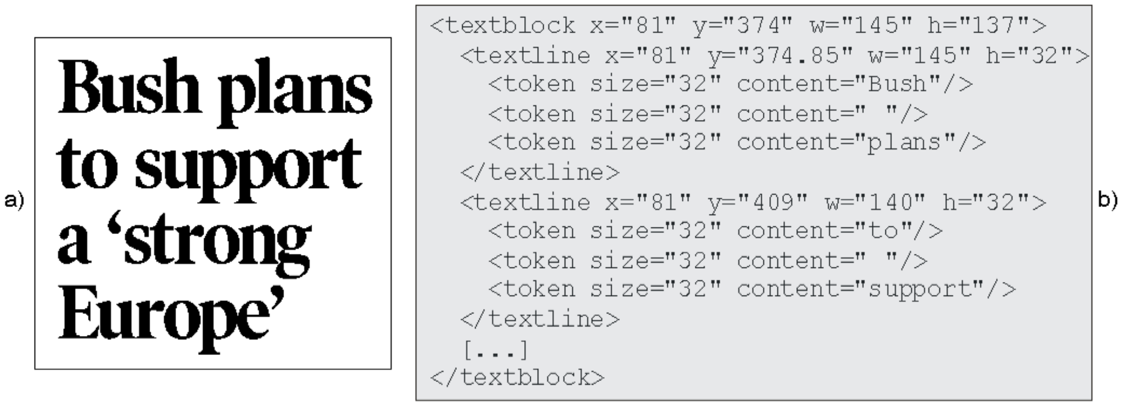
Restructuring techniques, as described in Subsection 2.3.2, resolve the problems enumerated above. We propose to apply three processes for restructuring PDF files: the cleaning, the structuring and the conversion. Firstly, we propose to access the electronic content of PDF documents and to *clean* it from inconsistencies. This process allows to eliminate superfluous information and to correct over- and under-segmentation errors.

Secondly, we suggest annotating the proper document with its *physical structures*, calculated using techniques of document image analysis. These structures facilitate the access to the documents and enable efficient copy-past operations. For instance, when copy-pasting multi-column text from European newspapers using Acrobat Reader [50], the reading order is not respected: the lines of each column are horizontally merged. The analysis of physical structures reestablishes the reading order and text lines can be vertically copy-pasted.

Finally, we propose to *convert* the analyzed document in an XML format [105], which benefits of several advantages relative to the PDF. Firstly, PDF proposes a linear representation of document's content and annotations. At opposite, XML is a meta-language based on a tree structure that is well suited for representing documents in a structured manner. Then, the richness of PDF format complicates the access to document's content, whereas XML is very simple to parse. Further, PDF files are readable by computers, but they are not understandable by people (cf. Figure 4.3). XML format has been designed for being readable by both machines and humans. Finally, developing a PDF converter is a cumbersome task, while XML could be transformed very fast in another format. To wrap up, we propose to convert the PDF files into XML, in order to represent the document in a structured manner and to simplify the access to its content for successive manipulations.

An efficient XML representation for static documents has been proposed by Bloechle [11, 12].

[1] Acronym of *What you see is what you get.*



Bush plans
to support
a ‘strong
Europe’

```
<textblock x="81" y="374" w="145" h="137">
  <textline x="81" y="374.85" w="145" h="32">
    <token size="32" content="Bush"/>
    <token size="32" content=" "/>
    <token size="32" content="plans"/>
  </textline>
  <textline x="81" y="409" w="140" h="32">
    <token size="32" content="to"/>
    <token size="32" content=" "/>
    <token size="32" content="support"/>
  </textline>
  [...]
</textblock>
```

Figure 4.5: An article's title is extracted in XCDF. This representation preserves the document's physical structures.

The format is called XCDF (*XML canonical document format*) and it has been developed for representing both document's content and physical structures. Figure 4.5 shows an example of a newspaper's title and its XCDF representation. This format is based on four principles and offers multiple advantages:

- Simplicity: at opposite of PDF, XCDF format is easy to read and manipulate for both computers and people.
- Universality: as well as in PDF format, each static document can be represented in XCDF, in despite of its class.
- Completeness: although the set of operators is restrained relative to PDF, no loss of information occurs during the restructuring from PDF to XCDF.
- Uniqueness: text, graphics and images are represented in a unique and non-ambiguous manner. Thus, the XCDF format provides a unique representation for two or more PDF documents composed of different primitives, but identical at rendering and printing.

In Subsection 4.2.3, we present Xed, our tool for restructuring PDF files into the XCDF format, and its analysis techniques.

4.2.3 Xed: a Tool for Annotating Static Documents with Their Physical Structures

Xed (acronym of *Extracting Electronic Documents*) is the Java tool we propose in order to extract PDF documents in the XCDF format. It targets at purifying PDF files and at annotating them with their original structures, using classical methods of document image analysis, such as filtering, clustering of homogenous zones, etc. as presented in Section 2.3. In this subsection, we present firstly tool's functionalities and secondly the results of its preliminary evaluation.

Cleaning and Structuring PDF Documents with XED

Figure 4.6 illustrates Xed's architecture, which mainly consists of extracting primitives such as text, graphics, and images, cleaning document's content, and recovering the layout. The conversion into XCDF format is well-suited for further manipulations and analysis of the document.

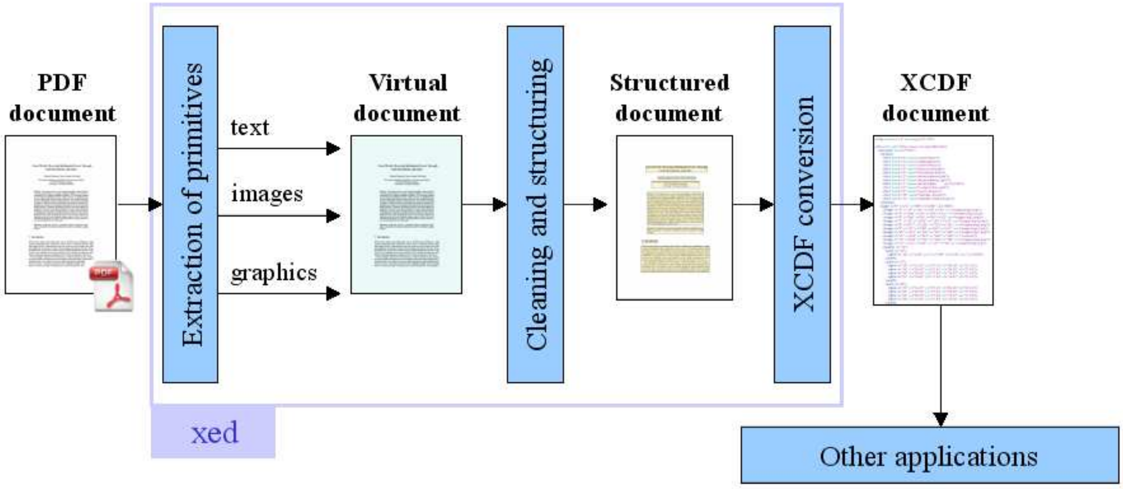


Figure 4.6: Xed's architecture includes two stages for respectively extracting the primitives and the layout. After analysis, the document is converted into XCDF format, ideal for supplemental analysis tools.

The *extraction of raw primitives* is based on two separated phases: the parsing and the creation of the virtual document. The *parsing* acts at low-level: it aims at reading PDF files and at converting their data into a tree, containing the operators and operands that will define document's appearance. The main issue of this phase consists in providing a high-level interface for accessing files' content. In fact, the PDF format allows various representations of primitives (text, graphics and images) and uses encoders for representing data into a compact form. For instance, a textual string can be a sequence either of hexadecimal values or characters. Similarly, the string can be stored in its standard representation or firstly encoded. The parsing homogenizes the representation of primitives, decodes the entire information, and decrypts encrypted files. The *creation of the virtual document* consists in interpreting the tree constructed during the parsing, in order to recreate the document. This phase implements the same functionalities of any PDF viewer, but the documents are rendered into an inner representation, called the "virtual document", rather than at screen. The virtual document is created page by page. For each page, Xed load resources such as fonts, images, color spaces, and extended graphic states as specified in [51]. When the resources are ready, Xed creates and draws the primitives into the page. During the whole creation of the virtual document, Xed updates a graphic state that describes the properties of the next object to create. These properties contain colors, size and appearance of the stroke, coordinates, etc. In fact, the PDF objects do not contain any information about their position and appearance, which depend on the configuration of the graphical state. For instance, a graphical object describes a shape, whereas position, color and stroke are only defined in the current graphical state. When the virtual document has been created, all the primitives have also been enriched with their planar and rendering attributes.

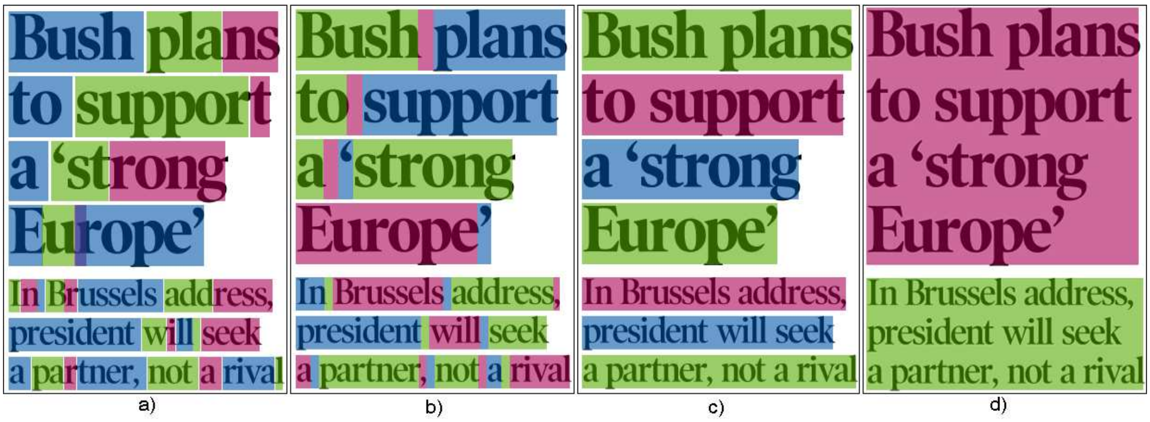


Figure 4.7: Xed analyzes the PDF data (a) and extracts words (b), lines (c) and homogenous text blocks (d). In each image, the filled rectangles correspond to the bounding boxes of text elements.

The *extraction of physical layout* recovers the document's physical structures. The current implementation of the analysis algorithm has been integrated by Jean-Luc Bloechle [11, 12] and recognizes textual blocks, lines and words (cf. Figure 4.7). The algorithm is composed of eight successive steps:

1. Remove all blank spaces from textual content. This step is unavoidable, in order to eliminate superfluous spaces added by PDF producers (confer the word "Europe" in Figures 4.4 and 4.7). Furthermore, the most part of under-segmented tokens is split in words.

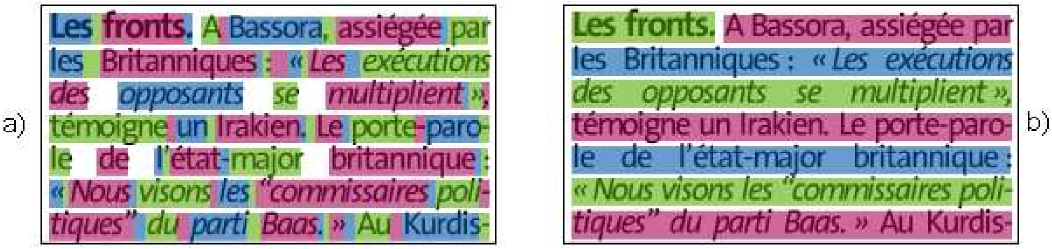
2. Horizontally merge the token in strings, given a threshold. This step eliminates over-segmentation errors.

3. Parse the merged strings, in order to separate punctuations, parenthesis, number, etc. from words.

4. Horizontally merge the tokens into line, given a threshold.

5. Add necessary blank spaces between consecutive tokens belonging to a same line.

6. Vertically merge lines into blocks, using the average distance between lines as threshold.



Les fronts. A Bassora, assiégée par
les Britanniques : « Les exécutions
des opposants se multiplient »,
témoigne un Irakien. Le porte-paro-
le de l'état-major britannique :
« Nous visons les "commissaires poli-
tiques" du parti Baas. » Au Kurdis-

Les fronts. A Bassora, assiégée par
les Britanniques : « Les exécutions
des opposants se multiplient »,
témoigne un Irakien. Le porte-paro-
le de l'état-major britannique :
« Nous visons les "commissaires poli-
tiques" du parti Baas. » Au Kurdis-

Figure 4.8: The retroactive merging detects over-segmented lines (a) and corrects them (b).

7. Retroactively merge. The blocks are reanalyzed in order to correct over-segmented lines. In fact, text justification provokes errors in Step 4, when the distance between tokens is higher than the threshold. Figure 4.8, shows an example of over- segmented lines after Step 4 and the result of the retroactive merging.

8. Parse all the tokens and merge sequential blanks, punctuations, digits, etc. At the same time, label each token with its syntactical category ("word", "number", "sign" and so on).

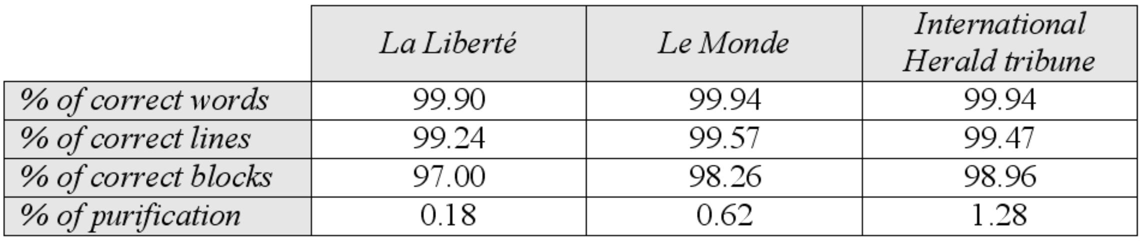
In addition to text, the extraction of physical structures takes into account the analysis of graphical primitives: Xed merges over-segmented objects; it distinguishes threads and frames from other graphics; and it eliminates occluded or invisible primitives.

After the analysis of physical structures, the document is *coverted* into XCDF format, which can be further analyzed or manipulated. For instance, Subsection 4.2.4 presents Inquisitor, a tool for validating and editing the results of document analysis, that has been developed for working with XCDF files.

A Preliminary Evaluation: Extraction of Physical Structures from Static Documents

The performances of Xed have been assessed in a preliminary evaluation [80, 11]. The database of PDF files used for this experiment consists in 30 front pages of three different and representative newspapers, i.e. "La Liberté", "Le Monde", and "International Herald Tribune", which are characterized by different layouts and segmentations of textual content. We chose newspapers for this evaluation, because they have a complex and variable layout, difficult to extract automatically.

Firstly, Xed restructured each PDF front page in the XCDF format, cleaning the document and calculating its physical structures. Then, the results have been visualized in an XCDF viewer and three experts appreciated the quality of the extraction, respecting XCDF specification. The investigated criteria were the percents of words, lines and blocks correctly restructured. Moreover, a purification rate has been measured to indicate if the original documents were under-segmented (rate < 1) or over-segmented (rate > 1). This rate only illustrates the variability of the PDF files for the three classes of newspapers. Table 4.2 summarizes the results of this preliminary evaluation.



	La Liberté	*Le Monde*	*International Herald tribune*
% of correct words	99.90	99.94	99.94
% of correct lines	99.24	99.57	99.47
% of correct blocks	97.00	98.26	98.96
% of purification	0.18	0.62	1.28

Table 4.2: The evaluation results take into account the percent of words, lines and blocks correctly restructured.

Although a formal evaluation would require a larger database, these preliminary results depict that Xed is useful for automatically producing the physical structures required in this thesis. Moreover, the results are reliable and, in our experience, further analyses could be applied without systematically validating Xed's results. However, in Subsection 4.2.4 we propose to validate them and we explain why validation of analysis results is an unavoidable task.

4.2.4 Visualization and Supervised Document Analysis

The CIDRE philosophy [9] promotes the idea that a general document recognition system does not work in a fully automatic way, but cooperates with the user. Its feedbacks through interactive tools improve the efficiency of the system that incrementally learns and adapts itself to document's properties. This philosophy is applied in the works of Robadey [85] and Hadjar [38], where the analysis of newspapers is supervised and the system incrementally learns from users' feedbacks. Similarly, the DocMining project [27, 26, 28, 3] allows to build analysis chains with various technologies, e.g. packages for pattern recognition, layout extraction, etc., and to supervise each individual process. All these works integrate the *xmillum* framework, proposed by Hitz in [82, 45].

xmillum offers a set of tools that allows to visualize the images of documents and the results of analysis accomplished over them. Users interact with xmillum, in order to validate and correct the displayed results, while the analysis system learns and updates itself. xmillum highlights two of the major advantages of the CIDRE philosophy:

- The visualization of results helps users to rapidly retrieve and correct errors;

- Interacting with graphical objects is more intuitive than directly tuning the analysis systems. Consequently, non-experts of the analysis domain are also enabled to tune the system.

In this thesis, the validation and the supervision of analysis results are unavoidable tasks, as discussed in the first part of this subsection. In the second part, we present a new tool called Inquisitor that, in some way, is the successor of xmillum.

Importance of Good Analysis Results

In this thesis, we propose a full system for analyzing corpora of events that requires the collaboration of several and heterogeneous technologies. The analysis of the data is a delicate task because errors at this stage are propagated to the indexing process. Obviously, errors in the indexing of information imply that users will encounter problems in retrieving information.

At-intra document level (as defined in Section 4.1), errors involved in the analysis of textual documents are words segmentation, characters substitutions, and failures in structures recognition. *Words segmentation errors* correspond to truncated or merged words, whereas *character substitutions* consists in replacing the characters of a string with wrong text. In both cases, the original mean of the text is altered or lost. These errors essentially provoke two drawbacks in the successive analysis processes. First, computing the thematic similarity between two documents does not provide reliable results. Second, the indexing process creates wrong indexes, which do no represent the original content of the document. In the worst case, a document can not be retrieved by users, because of malformed indexes.

Errors in structure recognition are less critical, but in any case they compromise the efficiency of the retrieval mechanism. In fact, if static documents' parts have not been correctly detected, submitting a query risks to produce a result set that is either incomplete or overburdened. For instance, a user, who is interested in articles talking about both Bush and Europe, submits a query with these two keywords. Although the archive contains several articles matching the query, their parts have been wrongly extracted and never contain the keywords together. Consequently, the result set is empty and the user cannot find the searched information.

Thus, supervising analysis methods and correcting their results are necessary tasks, for avoiding error propagation to successive analyses and for building an efficient system for managing multimedia data. This requirement of supervising analysis is enforced for mining systems at intra-document level, because they are the basis of the full analysis process.

Inquisitor: a Tool for Validating and Editing Static Documents Analysis Results

Inquisitor[2] [83, 32] is the successor of xmillum [45], a framework for validating, correcting and editing each type of document analysis results. Moreover, its edition functionalities allow to manually create new annotations over the documents.

In the next paragraphs, we presents Inquisitor. Firstly, we summarize xmillum's philosophy and we enumerate the limits of the framework, in order to advocate the development of Inquisitor. Then, we compare the architectures of both systems, highlighting the novelties of Inquisitor. In

[2] The prototype has been implemented by Florian Evéquoz, in the context of his master thesis.

particular, we present its inner XML language, which supports reversible transformations, and its main interactive capabilities.

xmillum and its Limits

The main novelty of xmillum was to be a flexible tool, suitable for surrounding each type of document analysis systems. This flexibility has been achieved using XML [105] and XSLT [107] technologies. In fact, xmillum does not define a strict inner language for representing the documents to visualize and manipulate, but it accepts any XML format as input. The interpretation of the data is done by mean of an XSLT stylesheet, describing the input format and the configuration of xmillum's viewer. In fact, the graphical user interface (GUI) of xmillum is flexible and customizable: the users defines the plug-ins to visualize, their appearance, and their interactive functionalities using the XSLT stylesheet.

Although xmillum is a powerful tool, which has been successfully integrated in several projects [85, 38, 27], we consider that it has some drawbacks:

- The opportunity of defining an XSLT stylesheet for configuring the input format, the graphical tools, and the interactive functionalities is at the same time an advantage and a disadvantage. This solution is elegant and powerful, but the configuration of xmillum is very complex for end users. In fact, a unique stylesheet involves two levels of abstraction, i.e. one for setting up the GUI and one for defining how to interpret the input data, that are confusing for users.

- xmillum is designed for accepting each category of input XML formats, but it does not provide any mechanism for outputting the manipulated document in its original format. Consequently, xmillum can be wrapped with an analysis system, but it not useful as stand-alone application.

- xmillum provides plug-ins for basic visualizations and interactions, whereas we need to display intra-aggregatus links, different annotations at contemporary, etc. xmillum can be extended with new and more complex plug-ins, but their implementation is a cumbersome task. In fact, the flexibility of GUI's configuration imposes several constraints for development: for instance, a unique graphical object is defined by different plug-ins, specifying object's geometry, appearance, and interactive opportunities.

Although Inquisitor shares most of its capabilities with xmillum, it has been developed for overcoming the limits enumerated above.

xmillum Versus Inquisitor

Figure 4.9 shows a comparison between the architectures of xmillum and Inquisitor. Both tools accept documents represented in any XML file format as input, which is transformed in an inner

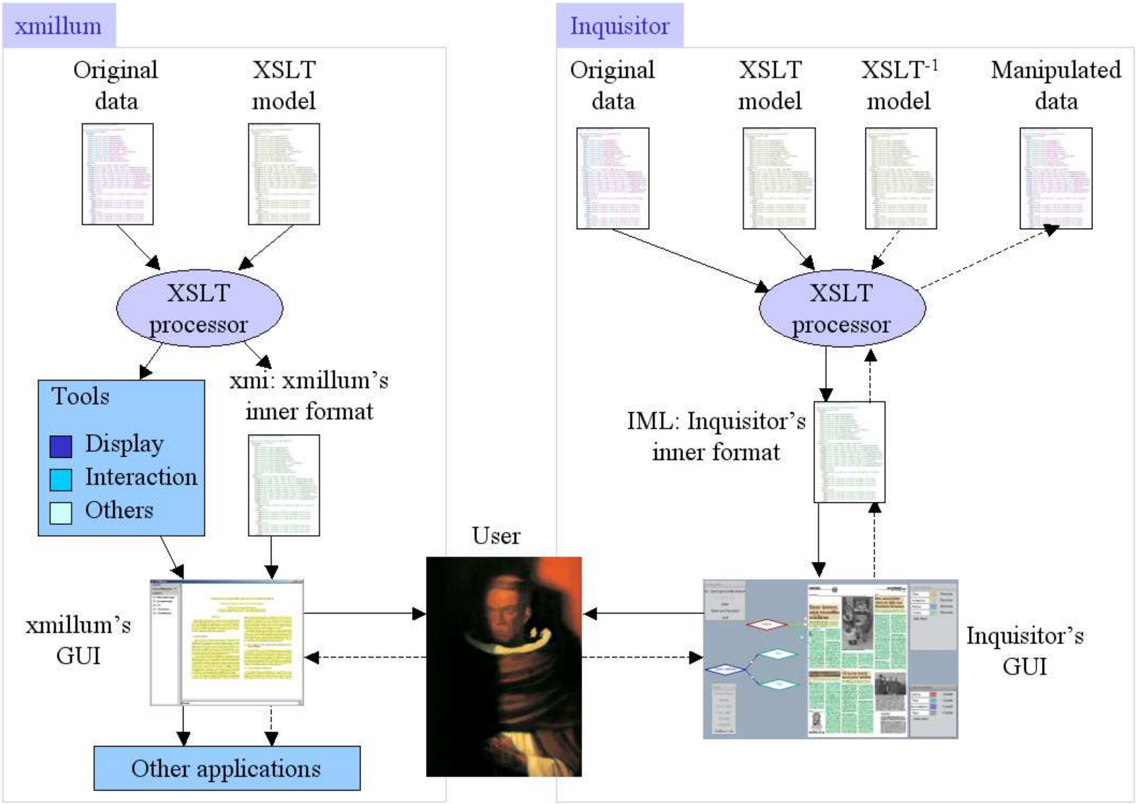


Figure 4.9: The image compares the architecture of xmillum (left) and Inquisitor (right).

format using an XSLT stylesheet [107]. In xmillum, the stylesheet requires to be completed with information about the graphical and interactive tools composing the GUI. As soon as the document is displayed in xmillum or Inquisitor, users can manipulate its analysis results. The main advantage of Inquisitor is that edited documents can be reconverted in the original input format, thanks to an inverse XSLT mechanism. In addition to this extension, we decided to define a graphical inner language for Inquisitor, in order to facilitate the configuration of the GUI. End users only take into account the transformation of the input format in this inner language.

Inquisitor's Inner Language

The *Inquisitor Modelling Language* (IML) is an XML inner format, which proposes a restricted set of primitives, but sufficient for representing most document analysis results. The format is composed of the following elements:

- The ***image*** is mainly used for visualizing the images of printable documents;
- The ***shape*** is a basic primitive that allows to represent bounding boxes and graphics. It

can be completed with meta-information, useful for displaying coordinates, labels, and so on;

- The ***layers*** are containers of the others primitives belonging to IML format. They allow to group together annotations of a specific category such as "text lines", "blocks", etc;
- The ***links*** correspond to the intra-aggregatus links in our model, as described in Subsection 3.2.4. They join related shapes and/or entities;
- The ***entities*** are a special type of primitives for representing high-level abstraction of documents. For instance, the entities are well-suited for representing the nodes of a hierarchal tree associated with a logical model.

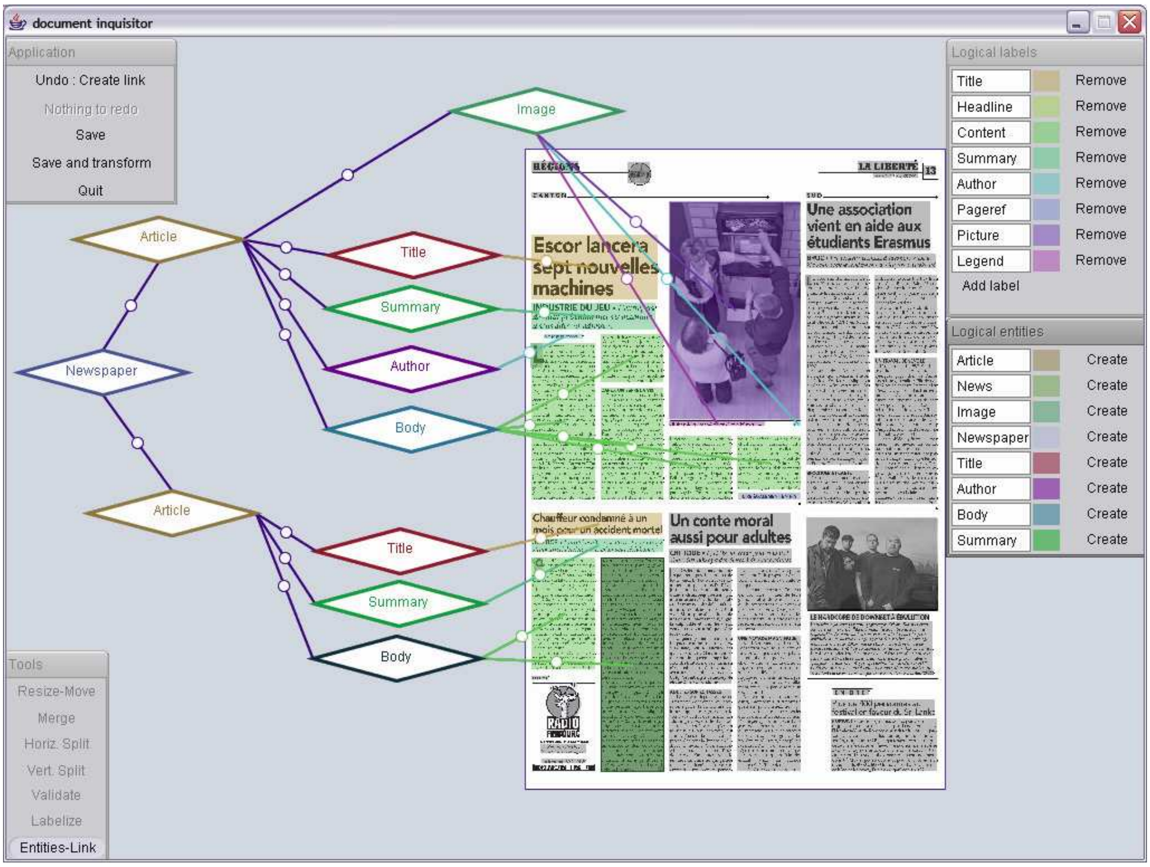
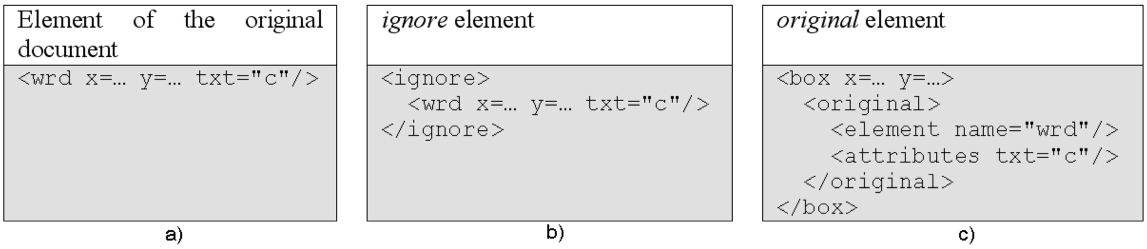


Figure 4.10: Inquisitor loads document's image and draws overlapped rectangles, which correspond to physical blocks. The rhombi are entities associated with logical structures. The links either join physical with logical structures or represent logical hierarchies.

Although IML is a minimal language, it allows to visualize and interact with the XCDF documents extracted by Xed (see Subsection 4.2.3). IML represents with shapes the bounding boxes of XCDF's words and symbols. The physical structures are associated to layers, which are encapsulated with the aim of conserving their hierarchy. For example, the first layer represents text blocks and it contains a second layer for text lines. If a user describes the logical structure of a newspaper, she creates an "article" entity, linked with three other entities, e.g. "title", "body" and "author". Then, she links these entities with the shapes that represent the corresponding text blocks. Figure 4.10 illustrates an example of XCDF file loaded into Inquisitor GUI, together with a first logical structuring.

IML is not only simple and generic enough to represent document analysis results, but it also has been conceived for being reversible. In fact, while converting an XML format into IML, it is necessary to avoid the loss of information that is not visualized in Inquisitor. The solution we propose consists in saving 1) all the elements and attributes of the original XML document and 2) the hierarchy of primitives composing the original document. The first condition is fulfilled including the two elements *ignore* and *original* in the IML format. *Ignore* allows to copy an entire subtree of the original document, without transforming its content. *Original* simply stores an element or an attribute with its values. Figure 4.11 shows an example of XML node that is transformed using both *ignore* and *original* elements.



Element of the original document

```
<wrd x=… y=… txt="c"/>
```

a)

ignore element

```
<ignore>
  <wrd x=… y=… txt="c"/>
</ignore>
```

b)

original element

```
<box x=… y=…>
  <original>
    <element name="wrd"/>
    <attributes txt="c"/>
  </original>
</box>
```

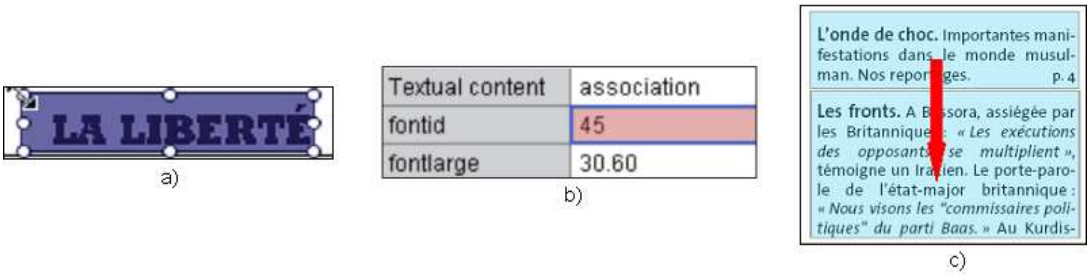
c)

Figure 4.11: An element of the original document (a) is completely ignored (b) or partially used (c).

The second condition of reversibility is resolved adding a unique identification value to each node of the original document, which represents its position in the tree before the XSL transform. The proposed solution is sufficient for guaranteeing that the original input document is exactly recreated. At opposite, manipulated documents are not always reversible after users interactions, for instance after merge and split operations.

Inquisitor's Interactive Functionalities

Inquistor proposes various manipulations, useful for validating the physical structures extracted by Xed (confer Subsection 4.2.3). Firstly, the bounding boxes of text blocks, text lines and words can be resized or moved (cf. Figure 4.12a). Further, words that have not been correctly extracted can be manually edited in a properties dialog (as shown in Figure 4.12b). Finally, Inquisitor allows to split under-segmented blocks or, at opposite, to merge those that are over-



Textual content	association
fontid	45
fontlarge	30.60

L'onde de choc. Importantes manifestations dans le monde musulman. Nos reportages. p. 4

Les fronts. A Bassora, assiégée par les Britanniques : *« Les exécutions des opposants se multiplient »*, témoigne un Irakien. Le porte-parole de l'état-major britannique : *« Nous visons les "commissaires politiques" du parti Baas.* » Au Kurdis-

Figure 4.12: Inquisitor proposes various functionalities for manipulating physical structures: resizing (a), textual content editing (b) and merging (c).

segmented (cf. Figure 4.12c). These two last operations are critical while reversing an IML file into the format of the original document. A priori, it is not possible to split or merge the *original* and *ignore* elements: these operations depend on user's requirements. Currently, when Inquisitor applies merge and split operations, it only stores this information within the original elements.

Inquisitor can be used for validating logical structures and for eliciting models of documents. In this thesis, we manually create these annotations (see Figure 4.13). Firstly, we create logical labels that are applied to the physical blocks. Then, we create a hierarchy of entities that correspond to the logical functions of text blocks. Finally, we connect together the labelled blocks and the entities using intra-aggregatus links.

To wrap up, Inquisitor allows to supervise the analysis of static documents at intra-document level, which has been presented all along this section, in order to ensure the quality of automatically produced annotations. After validation, static documents are ready for being compared to other media during events analysis, as presented in Section 4.3.

4.3 Meeting Analysis

At intra-event level, multimedia documents belonging to a meeting are aligned (alignment has been introduced in Section 2.1). The alignment uses the annotations and metadata added at intra-document level, in order to elicit the correspondences between those documents. Thus, the analysis of an event enriches the aggregati with links that represent thematic and temporal relationships (confer Chapter 3).

Multimodal alignment techniques and their errors have been studied in Mekhaldi's thesis [68]. Our work is not focused on multimodal alignment, but we aim at integrating an existing technique in our automatic system for meetings indexing. Moreover, we propose a tool for guaranteeing that alignment results can be validated by users. However, a brief overview of the integrated multimodal alignment technique and of its errors is necessary for understandability purposes.

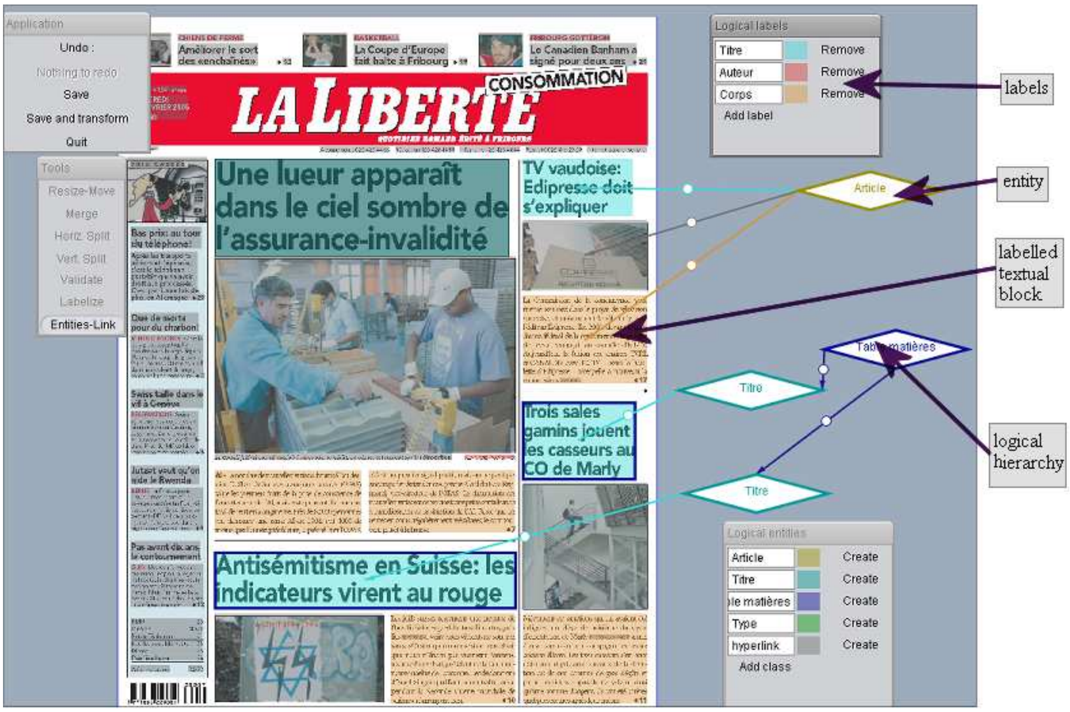


Figure 4.13: Inquisitor offers a mechanism for validating, eliciting, and creating logical structures.

Subsection 4.3.1 presents the properties of the meetings corpora taken into account in this thesis. Subsection 4.3.2 explains how a meeting is analyzed, detailing the annotations of multimedia documents and the technique of multimodal alignment. Finally, Subsection 4.3.3 presents the tools and the requirements for supervising events analysis.

4.3.1 Meeting Data

A meeting can be considered time-based information, in which all the media are synchronized. Media such as static documents, which do not naturally possess temporal dimension, are also to be related to meeting's time. A meeting groups together the documents prepared in the production phase, such as slideshow presentations, static documents, etc., and multimedia recorded during the event. Recordings consist of audio and videos of participants, and of global overviews. After recording, the meetings are annotated with descriptors, which contain locations, dates, starting times, participants' names and the set of multimedia resources. Frequently, the recorded resources have already been synchronized in an automatic manner. Finally, persons or automatic systems produce the annotations and the metadata used by alignment techniques during the meeting analysis.

The meetings included in our database belong to two different corpora. The first one is the IM2.DI corpus that contains 22 meetings in French, recorded in the Smart Meeting Room at University of Fribourg. All recorded meetings share the press reviews scenario, in which the participants discuss the front page of one or more newspapers [62]. The second corpus is provided by the AMI project and contains 171 meetings, which have been produced in the meeting rooms at IDIAP in Switzerland, Edinburgh University in Scotland, and at TNO center in Netherland. The most part of the AMI corpus is scenario-based, i.e. groups of four meetings belong to a unique session, discuss an identical topic, and are sequentially ordered in time.

From IM2.DI's corpus, we take into account meeting descriptors, newspapers, slideshows, audio files with mixed channels, and videos of participants and rooms. The AMI corpus has been annotated with more high-level information than IM2.DI corpus, but we selected the same categories of documents. Information such as whiteboard marks, pens strokes, and emotions have been ignored in the current environmental setup.

4.3.2 Meeting Analysis and Alignment

Both the IM2.DI and AMI corpora are enriched with supplemental information, presented in the following paragraphs. However, only part of this information has been retained for multimodal alignment of meetings, because various annotations lack of semantics or because they are not interesting for our purposes. For example, although gestures and gazes allow to detect on what users focus in a meeting, we do not use this information for further cross-event indexing. Once the interesting metadata and annotations presented, this subsection explicates how temporal and thematic links are created using multimodal alignment technique.

Annotation of Media in the Meeting

Meetings' documents are annotated with a multitude of additional information [89], added by persons or calculated by automatic systems. Audio files, which contain either the recording of individual speakers or a mix of all sources, are analyzed with automatic speech recognizers. Their annotations are speech segments distinguishing silences from spoken parts, textual transcription, speakers' identities and non-verbal acoustic clues such as laughing. These annotations are then analyzed in order to structure the dialog. The new derived annotations are utterances and dialog acts, similar to written sentences, which are further grouped in adjacency pairs and turns. Adjacency pairs are couples of related utterances, such as question and answer, whereas turns group successive utterances of the same speaker. Turns compose thematic episodes that, in general, correspond to meeting topics. Finally, named entities are references to persons, ideas and so on. All the annotations added to audio recordings are stamped with start and end times. In this thesis, we mainly take into account utterances, because they transmit the required thematic information although their length is generally small. Utterances of IM2.DI corpus have been manually written, while those of AMI corpus have been both automatically and manually transcribed. Assuming that users can supervise systems for automatic analysis of speech or correct their results, we only use manual annotations in the current implementation.

Videos and images are automatically analyzed in order to extract participants' gestures, facial expressions and eye gaze. Gestures are annotations describing non-verbal actions, such as pointing. Facial expression analysis attempts to detect participants' emotions. Eye gaze annotations describe where participants are looking at. Successive analysis shall define at what they are looking. Furthermore, videos are analyzed in order to extract projected slides and documents. The calculated annotations, such as slide changes detection, enrich slideshows and static documents with timestamps. Finally, in both the corpora the videos are manually annotated with speaker's name or id code, necessary for indexing this type of document.

Static documents' annotations have been widely discussed in Section 4.2. Static documents in the IM2.DI corpus are manually annotated with logical structures, whereas in the AMI corpus their physical structures have been automatically extracted using Xed (see Subsection 4.2.3).

Computing Links between Multimedia Documents

At intra-event level, static documents and spoken dialogs are compared with multimodal alignment technique [63, 68]. The alignment accesses the documents through their annotations, i.e. the utterances composing audio transcript and the tiled textual content of static documents. The technique consists in the following steps:

- *Stop-words removal.* Stop-words are recurrent terms that are useful in the oral and written communication, but not significant from a thematic point of view. For instance, article, adverbs, pronouns, and so on are classified as stop-words. Thus, indexing systems remove these terms from the textual content, in order to reduce computational time in the successive calculations. A stop-word list depends on the document's language, category and on the application. For instance, two different lists are necessary for filtering static documents and speech transcript, because the latter contains acoustic clues that do not exist into the written communication.

- *Stemming.* The stems are reductions of words that are obtained by eliminating suffixes and prefixes and by substituting pattern. Thus, a unique stem can represent similar words: for instance, the "view" string can represent "view", "viewer" and "overview". Unfortunately, a stem is not equivalent to the word's root and can characterize uncorrelated terms. In this step, the text already filtered from stop-words is transformed into a list of stems.

- *tf.idf calculation.* The *term frequency* and the *inverse document frequency* are measures that indicate the relevance of terms respectively into a document and into the whole archive. In particular, the *tf* value represents the amount of words occurrences within a specific document or part of it. Thus, a word used several times is representative for the document. At opposite, the *idf* measures the total amount of words occurrences into the whole archive of documents. Consequently, a word that frequently appears in the archive is less significant for identifying a document than a rare one. Finally, the ratio between *tf* and *idf* allow weighting the importance of the terms contained into a document relatively to the global context. In this step, firstly we calculate the *idf* value for each stem extracted from static documents and speech transcripts. Secondly, we calculate a list associating stems and *tf.idf* values for each physical text block and utterance.

- *Similarity calculation.* The similarity score represents the degree of correlation between fragments extracted from static documents and utterances. Let us consider F_1 and F_2 as being two fragments of two different documents. Their similarity score is calculated thanks to the cosine formula 4.1:

$$Cosine(F_1, F_2) = \sum_{s=1}^{N} w_{s,F_1}.w_{s,F_2} / \sqrt{\sum_{s=1}^{N} w_{s,F_1}^2 . \sum_{s=1}^{N} w_{s,F_2}^2} \tag{4.1}$$

  where w_{s,F_i} corresponds to the *tf.idf* value of the stem s, contained into the fragment F_i. The calculated similarity score is in the range from 0 to 1, and only takes into account the stems that exist in both fragments F_1 and F_2. A thematic link between the two fragments is created when the similarity score exceeds a given threshold, empirically fixed at 10%.

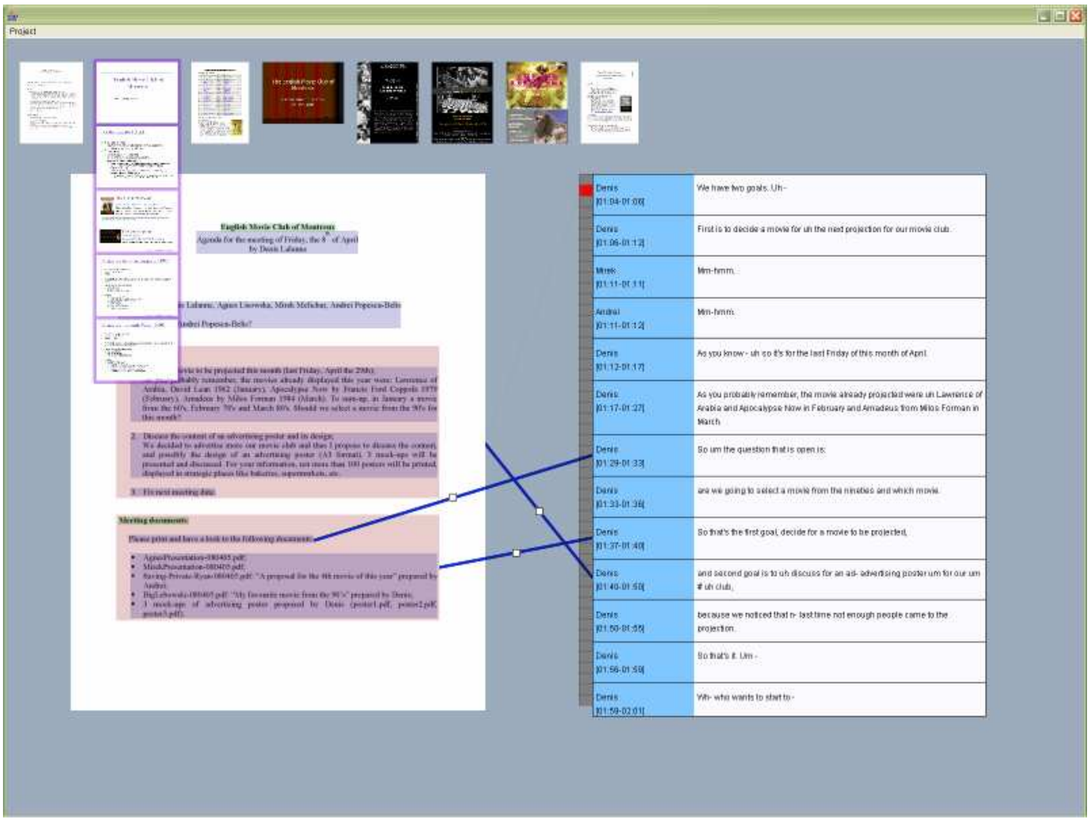
The multimodal alignment enriches the event's aggregati with thematic links representing similarities between documents. Moreover, a link between an utterances and a textual block of a static document allow to enrich the latter with the timestamps of the former. Although this technique provides in general good results [68], we suggest to validate them to avoid errors. Subsection 4.3.3 presents the alignment errors and how to correct them.

4.3.3 Wizard of Faerie: a Tool for Validating and Editing Intra-event Links

In Subsection 4.2.4, we presented the relevance of supervised analysis at intra-document level. Similarly, at intra-event level, we propose a tool to support the validation and correction of multimodal alignment results, in order to ensure their validity and to avoid the propagation of errors in the indexing of aggregati at cross-event level.

The alignment can provoke several categories of errors: documents that have not been linked, wrong similarity scores and false similarities detected. The stemming and the lack of a thesaurus are the main origins of these errors. Firstly, the stemming produces strings that represent similar words from syntactic point of view, without considering their semantics, e.g. "seed" and "seer". The alignment algorithm can match two uncorrelated words and, consequently, the similarity score is erroneously increased. Secondly, the lack of a thesaurus avoids the multimodal alignment to detect synonyms, hyponyms, hypernyms, meronyms and in general periphrasis. Thus, only words represented with the same stems are considered similar.

The tool illustrated in Figure 4.14 is called the *"Wizard of Faerie"* and allows to correct the errors presented above. In the windows' top part, the documents involved in the multimodal



English Movie Club of Montreux

Agenda for the meeting of Friday, the 8th of April
by Denis Lalanne

Meeting documents:

Please print and have a look to the following documents:

Denis [01:04-01:06]	We have two goals. Uh-
Denis [01:06-01:12]	First is to decide a movie for uh the next projection for our movie club.
Mirek [01:11-01:11]	Mm-hmm.
Andrei [01:11-01:12]	Mm-hmm.
Denis [01:12-01:17]	As you know - uh so it's for the last Friday of this month of April.
Denis [01:17-01:27]	As you probably remember, the movie already projected were uh Lawrence of Arabia and Apocalypse Now in February and Amadeus from Milos Forman in March.
Denis [01:29-01:33]	So um the question that is open is:
Denis [01:33-01:36]	are we going to select a movie from the nineties and which movie.
Denis [01:37-01:40]	So that's the first goal, decide for a movie to be projected,
Denis [01:40-01:50]	and second goal is to uh discuss for an ad- advertising poster um for our um # uh club,
Denis [01:50-01:55]	because we noticed that n- last time not enough people came to the projection.
Denis [01:56-01:59]	So that's it. Um -
Denis [01:59-02:01]	Wh- who wants to start to -

Figure 4.14: The Wizard of Faerie allows to validate the results of multimodal alignment at intra-event level.

alignment are visualized in the form of icons. When the user moves over the icon corresponding to a static document, its pages are displayed, as well in the form of icons. Documents' pages can be selected and focused in the main view on the left. A page is displayed as raster image, with physical and logical structures drawn as overlapped rectangles. At contemporary, the audio transcript of meeting is represented in a scrollable panel, with its utterances sorted by time. Finally, the focused multimedia documents are connected with links that can be removed or added by users. For instance, a user can create a new link by first clicking an utterance and then a similar article.

4.4 Conclusions

In this chapter, we have presented the analysis of static documents at intra-document level and of meetings at intra-event level. The analysis of static documents is focused on PDF files, which are automatically cleaned from inconsistencies and annotated with their physical structure. These tasks are accomplished by Xed. After analysis, the results can be validated by users, using Inquisitor tool. Moreover, this tool allows to manually annotate static documents with their logical structures.

At intra-event level, the different annotations calculated by monomodal techniques are involved in the analysis of a single meeting. In particular, multimodal alignment technique is applied over the meeting, in order to link together its audio transcript and its static documents. The results of alignment can be validated by users with a tool called the Wizard of Faerie.

After the validation, the meetings are ready for being indexed at cross-event level, as presented in Chapter 5.

Chapter 5

Meetings Indexing

The indexing of meetings is the process that produces the aggregationi, as described in Section 3.2. In this chapter, we illustrate the technologies that act at cross-event level, in order to create indexes useful for retrieving multimedia information. Our system supports two categories of indexes: stems and links. The former are used for matching textual queries submitted by users with the documents in the aggregatio. The links are a new type of indexes that enables the browsing of the aggregatio.

Moreover, we suggest to include additional features such as extraction of documents' themes, ranking mechanisms, and clustering. These features are useful to organize the aggregatio and to present the users a global view of its contents.

In Section 5.1, we discuss the issues and the challenges related to the creation of a relational search engine, i.e. a system in which the documents are indexed by links. Then, Section 5.2 presents the architecture of our indexing system, explaining the creation of the aggregatio. In particular, we detail the importation of a meeting corpus, its indexing, the automatic creation of links using multimodal alignment, the supplemental features added in our engine, the database structure, and the data returned after queries submission. Section 5.3 illustrates the opportunities offered to the users for supervising the indexing process and for editing new documents. Finally, Section 5.4 concludes this chapter and summarizes the system's performances, obtained by indexing of IM2 and AMI corpora.

5.1 Issues of a Relational Search Engine

Existing search engines retrieve information using indexes, which habitually are stems representing textual content or metadata. The efficiency of search engines can be enhanced by techniques that determine documents relevance, e.g. by counting the amount of hyperlinks pointing at them [59], of users accesses, etc. In our approach, we propose to use links between aggregati as additional indexes for browsing the data. We call a system offering this functionality a "*relational search engine*". Moreover, in this thesis, our system must be able to organize and retrieve each category of multimedia documents. The design and development of such a relational search engine open the following main challenges:

- The system is expected to detect and manage all the relationships between documents, in order to allow users to efficiently browse in the aggregatio;
- The indexing system intend to handle several multimedia sources, in different file formats;
- The system must permit aggregati to be dynamically updated, created, or deleted;
- The retrieval of information shall be very rapid for being accepted by users.

Section 5.2 proposes an architecture and an indexing technique that take into account all the aspects presented above.

5.2 Indexing System Architecture

This section describes the architecture of our indexing system for meetings aggregationi and presents in details the different modules composing it. Figure 5.1 is an overview of the system architecture, illustrating the flow of multimedia data through the different modules.

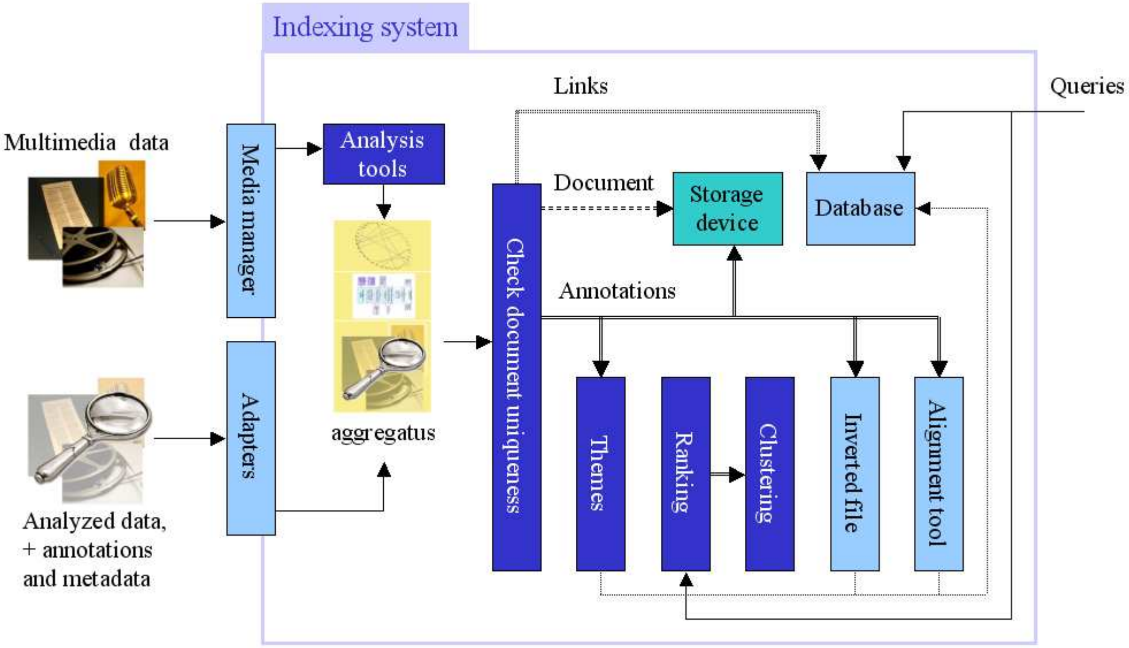


Figure 5.1: The architecture of the indexing system and the data flow.

The system takes as input either "raw" multimedia documents or data that has already been analyzed and annotated. Each document is manipulated in order to create an aggregatus (see Subsection 5.2.1). The uniqueness of aggregati is checked, because a document cannot possess multiple instances. If an aggregatus has already been imported, several options are taken into account (cf. Subsection 5.2.2). At opposite, when the aggregatus is imported and indexed for the first time, the document is stored on the physical support and its annotations are further analyzed. Firstly, the inverted file is updated and stored in the database. The inverted file associates stems to document's content in order to create indexes (confer Subsection 5.2.3). Secondly, annotations are used by the multimodal alignment module, which creates the links at the cross-event level (see Subsection 5.2.4). Furthermore, annotations are analyzed in order to extract document's keywords (cf. Subsection 5.2.5). The database is updated all along the indexing process with the information concerning the aggregatus (cf. Subsection 5.2.6). After indexing, users and applications submit queries to the database for retrieving information (cf. Subsection 5.2.7). Two auxiliary modules allow to create a ranking of documents relative to a query (see Subsection 5.2.8) and to group the aggregati into clusters (see Subsection 5.2.9).

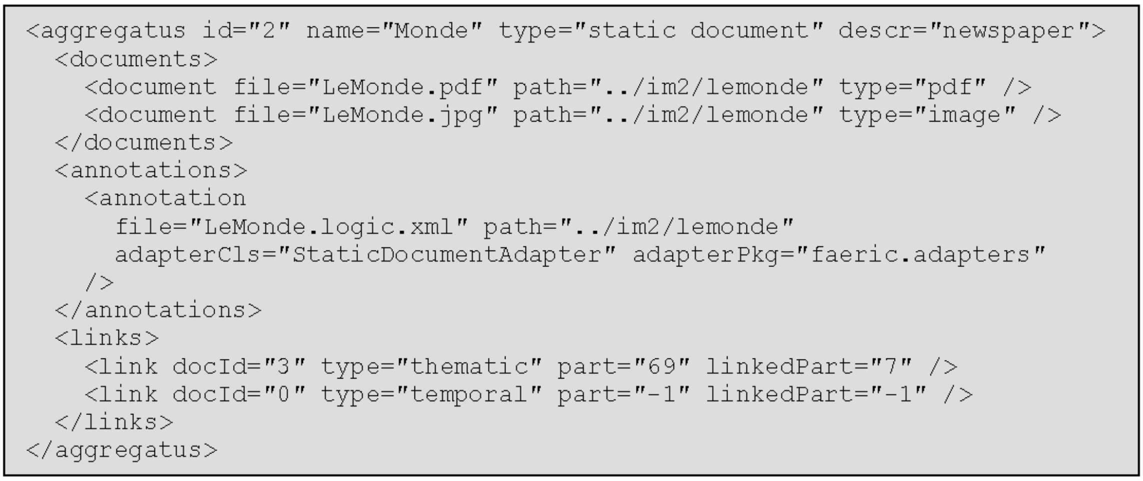
5.2.1 Importing Data Through Adapters

Our relational search engine aims to manage heterogeneous multimedia documents, represented in various formats (cf. Section 5.1). A classical solution, which has been successfully integrated in projects such as DocMining [27], consists in using adapters for transforming the original data into a unique and system-specific representation.

Our system provides two solutions for importing multimedia data and converting them into an aggregatus. On one hand, specific adapters handle documents that have already been analyzed and annotated and that do not require supplemental manipulations. The adapters convert the input information into the specific data representation of the system. On the other hand, the indexing system offers a media manager that allows to import and analyze raw multimedia documents. In fact, analysis tool such as Xed (cf. Section 4.2.3) are encapsulated in the system in the form of plug-ins. When a raw document is imported, the media manager detects its category and selects the appropriate tool for analyzing it. After analysis, the tool creates the aggregatus into the system-specific representation.

This double mechanism facilitates the integration of mature techniques into the system, in order to automatically analyze the data. For instance, a researcher who has developed a new supervised method. In a first phase, she implements an application that is not encapsulated in the indexing system. The application analyzes the data that are further validated and imported through adapters. In a second phase, when the method is reliable, the application can be integrated as plug-in in the indexing system, in order to automatically analyze the raw documents.

Figure 5.2 illustrates the aggregatus of a static document, as represented into the indexing system. For understandability purposes, the example is given in an XML format. The aggregatus contains metadata such as its type, descriptors, etc. and it is composed of documents, annotations and links. Each "document" element references a physical files. The aggregatus can contain one or more representations of the same document, e.g. a PDF file or the equivalent



```
<aggregatus id="2" name="Monde" type="static document" descr="newspaper">
  <documents>
    <document file="LeMonde.pdf" path="../im2/lemonde" type="pdf" />
    <document file="LeMonde.jpg" path="../im2/lemonde" type="image" />
  </documents>
  <annotations>
    <annotation
      file="LeMonde.logic.xml" path="../im2/lemonde"
      adapterCls="StaticDocumentAdapter" adapterPkg="faeric.adapters"
    />
  </annotations>
  <links>
    <link docId="3" type="thematic" part="69" linkedPart="7" />
    <link docId="0" type="temporal" part="-1" linkedPart="-1" />
  </links>
</aggregatus>
```

Figure 5.2: The aggregatus contains all the information associated with a static document: the files, the annotations and the links.

raster image. The "annotation" elements allow to locate the files containing the annotations and metadata, which require specific adapters for being interpreted by the system. The "link" elements contain the identifier of the linked document, the type of the link (for instance, "temporal", "thematic", and so on), and the identifiers of connected parts.

This structure is used for representing each type of aggregati. They are not required to be complete: for instance, persons recordings do not contain any "document" element.

5.2.2 Multiple Document Copies Removal

When an aggregatus has been created, the system verifies its uniqueness in the aggregatio. In fact, multiple occurrences of the same document could change its relevance: for instance, multimodal alignment could detect different similarities between multimedia documents, because the *tf-idf* values of textual information lowered (confer Subsection 4.3.2 for more details about *tf-idf*). Obviously, if the collection of documents is very large and the amount of multiple copies is low, this problem is irrelevant. However, we propose to remove multiple copies of the same aggregatus. Checking the uniqueness of documents is a non-trivial task. The differences between binary files are not a valid criterion: for instance, videos can be encoded in different formats such as *mpg* and *avi*, whereas a static document can be represented as PDF file or bitmap image. Similarly, information such as document's name is not reliable for discriminating two documents. The solution we propose consists in accessing the document's content using annotations and to check its uniqueness. The unavoidable precondition is that annotations access and structure the content in a unique and unambiguous manner. For instance, static documents converted in the XCDF format (see Subsection 4.2.3) have a unique representation, regardless of their original data representation. Similarly, the distance between the utterances of two transcripts can guarantee the uniqueness of an audio file.

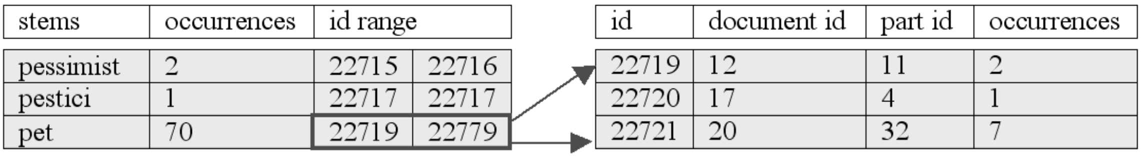
Checking documents uniqueness induces three situations:

- The aggregatus does not exist. In this case, the standard indexing process continues.
- The aggregatus already exists: it is ignored and discarded.
- Although the aggregatus already exists, it is included in a new multimedia document or event. The system creates a link connecting the existing aggregatus with the multimedia document or event. For instance, the aggregatio contains the aggregatus of a static document. The system tries to index a meeting, in which this static document is presented. Instead of creating a new aggregatus for the static document, the system links the existing one with meeting's aggregatus.

The following subsections describe only the manipulations concerning a new aggregatus. If the aggregatus already exists, no further process is required.

5.2.3 Inverted File

The inverted file is the most classical method for indexing textual documents in state-of-the-art retrieval systems [103]. The inverted file contains the list of all the terms used into an archive, coupled with the references to the documents containing them. When an aggregatus is imported, the system accesses its content using annotations and decomposes it into fragments. The terms of each fragment are then filtered from stop-words and stemmed (confer Subsection 4.3.2 about stemming). The inverted file is updated with the new document's stems. An example of our inverted file is illustrated in Table 5.1.



stems	occurrences	id range	
pessimist	2	22715	22716
pestici	1	22717	22717
pet	70	22719	22779

id	document id	part id	occurrences
22719	12	11	2
22720	17	4	1
22721	20	32	7

Table 5.1: The inverted file is composed of two tables. The left one contains the stems, whereas the right one includes the references to documents' fragments.

The first table contains the stems, the total amount of times they occur into the aggregatio, and two values defining a range. In the second table, the values of the interval are used as addresses for retrieving the documents that contain the stems. An entry of the second table is composed of a unique address, the identifiers of the document and fragments containing the stem, and the number of occurrences contained in each fragment.

When persons submit textual queries to the system, the inverted file is consulted for retrieving documents. For instance, a user submits a query. The system transforms query's keywords in stems, used as keys for retrieving addresses in the first table. Those addresses allow to identify the documents to restitute as result.

Moreover, the inverted file allows to calculate the *tf-idf* values of terms, which correspond to their occurrences in a fragment of document over the total occurrences in the whole aggregatio (see Subsection 4.3.2). The *tf-idf* values are necessary for cross-event alignment.

5.2.4 Cross-Event Alignment

Cross-event alignment is based on the multimodal alignment technique described in Subsection 4.3.2. At intra-event level, the alignment only concerns static documents, whereas at cross-event level it takes into account all the types of multimedia documents.

A new aggregatus is accessed by annotations and metadata. Its fragments are compared with those of the other aggregati, in order to create thematic links. The alignment process consists of stop-words filtering, stemming, and calculation of thematic similarity using the *tf-idf* values extracted from inverted file. Since the multimodal alignment takes into account documents' fragments, two aggregati can be linked several times.

5.2.5 Document Themes

The extraction of document's themes consists in selecting the words that are representative of its content. The heuristic integrated in our indexing system tries to detect the most relevant words in the document, taking however into account their frequency in the aggregatio.

First, the document's content is filtered from stop-words and stemmed. Then, for each stem the *tf-idf* value is calculated. Finally, the values are sorted from the highest to the lowest. The most relevant words are at the beginning of the ranking.

5.2.6 Indexed Information Representation

The representation of the aggregatio is determinant for accessing and retrieving information in a fast and efficient way. In our first attempts, the data were represented in XML files, but they were not adapted for managing the set of links between documents. Thus, we replaced the XML files with a relational database, further optimized for accelerating the access to the aggregatio. Figure 5.3 illustrates the current database structure.

Database's tables belong to four categories:

- The *documents* table contains all the aggregati and their properties, i.e. an identifier, a name, a type and an optional description field. Multimedia documents' and events' aggregati are even represented into the *compounds* table, in which their individual resources are listed.

- Each aggregatus is split into tables that describe its document, its annotations, and its links. These tables exactly correspond to the XML representation discussed in Subsection

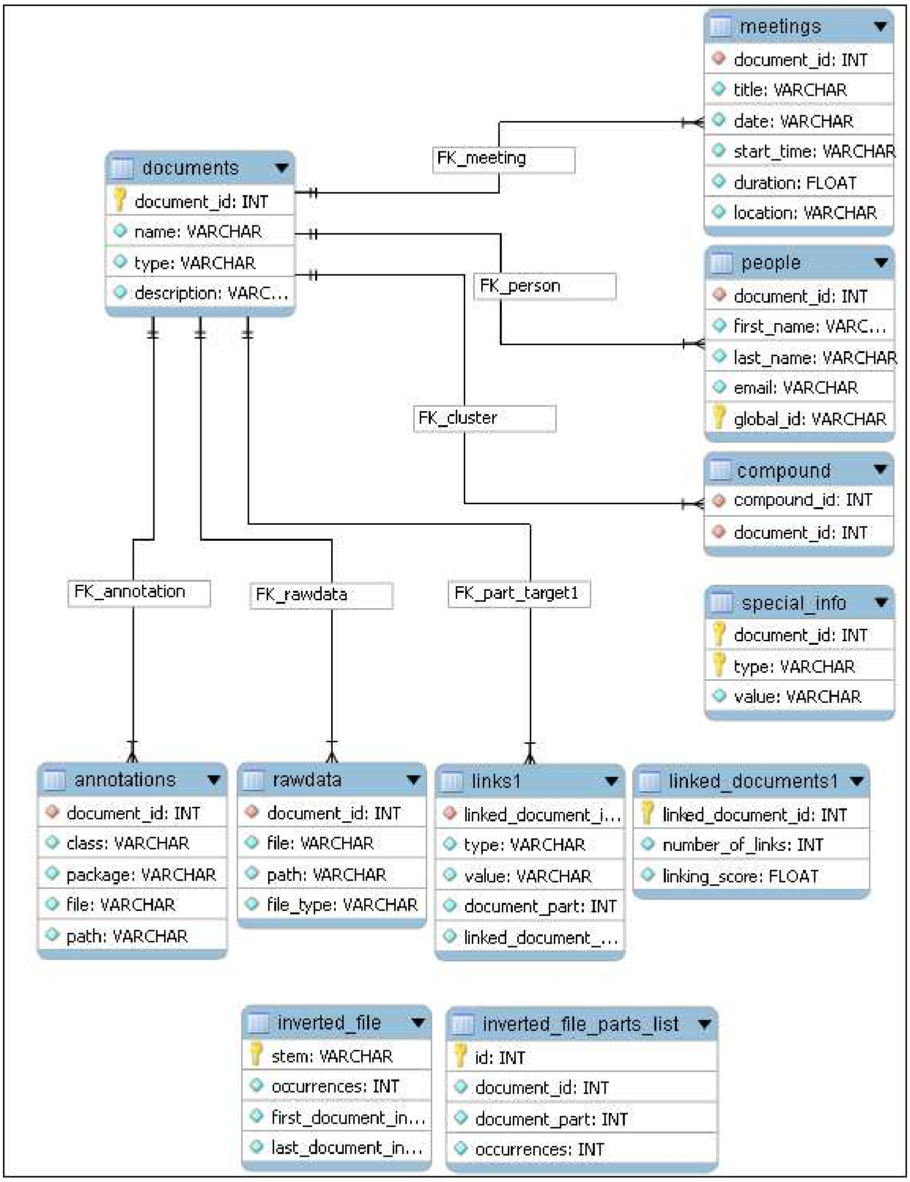


Figure 5.3: The tables allow to represent the aggregatio and to efficiently access it.

5.2.1. Each aggregatus possesses its own *links* and *linked_ documents* tables: the former includes the detailed list of its links, whereas the latter contains the amount of links with other aggregati.

- The *meetings*, *people* and *special_ info* tables are specialized for metadata and annotations. In particular, the *meetings* table contains information required by the browsers presented in Chapter 6. The *people* table allows to store the profiles of users browsing the aggregatio. An entry of this table can be associated with the aggregatus of a person recording. Finally, *special_ info* allows to enrich the aggregati with additional properties.

- Finally, the *inverted_file* and *inverted_file_parts_list* tables contain the inverted file, as described in Subsection 5.2.3.

5.2.7 Documents Retrieval

The system offers two mechanisms for retrieving documents: the classical full text search and the link-based browsing. In both cases, the users interact with the indexing system. The latter interrogates the database, which restitutes a set of aggregati as result.

Full text search consists in submitting a textual query composed of one or more keywords. The query is filtered from stop-words and the remaining keywords are stemmed. Then, the system consults the inverted file's tables, in order to retrieve indexed documents and parts. The documents matching the query are organized in both a ranking and N-ary trees, which are respectively described in Subsections 5.2.8 and 5.2.9.

Links-based browsing consists in asking the system to retrieve all the aggregati connected to an aggregatus selected by the user. These aggregati are identified using the link tables and, similarly to results of full text search, arranged in a ranking and clustered.

In both cases, the answer of the system contains the documents matching the query, as well as the ranking and the clusters.

5.2.8 Documents Ranking

The ranking targets to find and sort the most significant documents in respect of a query. Rankings are useful for presenting documents to the users in a Google-like view. Our system integrates two mechanisms for creating rankings, based on textual queries or links.

Textual queries can be submitted by users or automatically calculated, e.g. by selecting as keywords the most recurrent terms into the aggregatio. The ranking tool takes into account the documents containing at least one keyword of the query. Let us consider Q and D as being respectively a query and a document matching the query. For each document, the tool calculates a score using Formula 5.1:

$$Relevance(Q, D) = \sum_{q=1}^{Q_l} w_{q,D} \tag{5.1}$$

Where Q_l is the length of the query Q, q is a keyword of the query, and $w_{q,D}$ is the *tf-idf* value of a stemmed keyword for the document D. When the relevance to the query has been calculated for interesting documents, they are sorted by theirs scores.

Similarly, when users browse and select a document, the system retrieves its linked aggregati. These results are also organized into a ranking. In this case, the retrieved aggregati are sorted regarding at the amount of links they possess with the one selected by the user. The most linked document is the first in the ranking.

5.2.9 Clustering

The clustering of data allows to group similar information and, for instance, is useful for visualizing the aggregatio. Our indexing system provides a clustering mechanism that produces a set of N-ary trees from a ranking of documents, as described in Subsection 5.2.8.

The ranking is traversed starting from the document with the highest score. This document becomes the root of a tree. While traversing the list, each document is compared to the documents already in the tree(s). After comparison, the current document is added as child node to the most similar one. In the case of clustering, the similarity between two documents is given by the amount of links connecting them. If the best similarity is lower than a threshold, the current document becomes the root of a new tree.

5.3 Edition and Validation of Aggregati

At cross-event level, the aggregati and the links can be manipulated by users. In the first case, aggregati can be imported, removed, and edited at intra-document or intra-event level. These operations are very expensive, because the system is forced to recalculate the inverted file, the *tf-idf* values, and the correlations between multimedia documents using multimodal alignment.

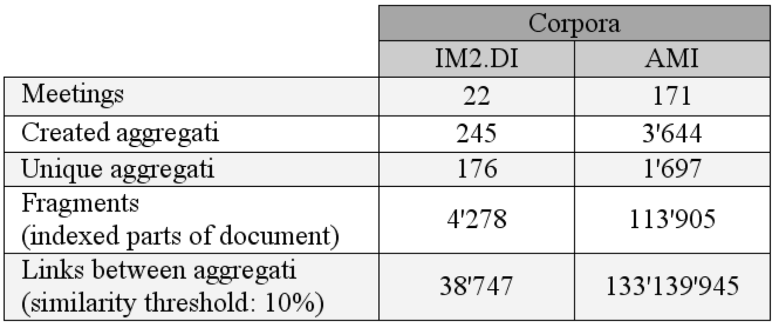
Links can be added, edited or removed without realigning the aggregati. The manipulation of links aims either at validating and correcting the results of cross-event alignment or at creating new documents. The validation and the correction of alignment results have already been discussed in Subsection 4.3.3: these tasks are identical at both intra-event level and cross-event level.

The creation of new documents mainly consists in producing a new aggregatus, composed of hyper-links towards fragments of documents in the aggregatio. For instance, a user who participates to a meeting is in charge to produce its minutes. Firstly, the user creates a new aggregatus or selects an existing one to be completed. Then, she links the items on meeting agenda with this aggregatus, in order to define the discussed topics. The topics are further developed, by connecting opinions and decisions extracted from audio transcript's utterances. Finally, the user links some images to the aggregatus, completing the minutes.

When users produce such a document, the system creates the new links and updates the inverted file with the references to the new aggregatus, but it does not realign the aggregatio. In fact, the terms frequency does not change and the new aggregatus inherits the links of the fragments hyper-connected by users.

5.4 System Performances: Creation of Aggregationi

The performances of the indexing system have been tested using IM2.DI and AMI corpora (cf. Subsection 4.3.1). The two corpora have been automatically indexed, applying the different steps presented in Section 5.2. Firstly, the meetings have been imported, creating all the required aggregati. Static documents of AMI corpus have been analyzed by Xed and successively transformed in aggregati. The other categories of multimedia documents have been imported through adapters. After the creation of aggregati, redundant documents were removed. The uniqueness of static documents was proved thanks to XCDF format, whereas that of persons recordings was established verifying first names, last names and identification codes. The uniqueness of other multimedia documents has been manually checked before importing them. Retained documents have been fragmented into their parts, using their annotations and metadata. Successively, the aggregati have been cross-connected, thanks to thematic alignment of fragments and to lexical analysis. The details about resulting aggregationi are presented in Table 5.2.



	Corpora	
	IM2.DI	AMI
Meetings	22	171
Created aggregati	245	3'644
Unique aggregati	176	1'697
Fragments (indexed parts of document)	4'278	113'905
Links between aggregati (similarity threshold: 10%)	38'747	133'139'945

Table 5.2: The table resumes the characteristics of the IM2.DI and AMI corpora and of their aggregationi.

For the creation of aggregati, a PC with a Pentium 4 CPU at 2.40 GHz and 512 MB of RAM was used. The indexing system has been implemented with Java 2 SDK [54] and uses a MySQL database [72]. The time required for indexing IM2.DI and AMI corpora corresponds respectively to 50 minutes and 26 hours. The times indicate that calculations are longer if the amount of documents augments: in fact, the time complexity of multimodal alignment algorithm is $O(n^2)$, where n corresponds to the number of aggregati. Thus, the alignment is the bottleneck that, however, can be enhanced thanks to parallel calculations and other techniques. These improvements are discussed with more details in Subsection 7.2.1.

With the AMI aggregatio, submitting a query of 8 words takes in mean less than half a second to retrieve all the documents of interest. Similarly, retrieving the documents linked to a document of interest requires in general less than 1 second.

5.5 Conclusions

In this chapter, we have presented our relational indexing system, designed for managing meeting collections. The system is composed of several modules, which allow to automatically import, analyze, and index multimedia information.

On one hand, the indexing system represents textual information into the inverted file, a state-of-the art technique. The latter is well suited for searching information by querying. On the other hand, the indexing system calculates and stores all the links between multimedia documents, in order to enable the browsing task. When the user submits a query or browses within the aggregatio, the indexing systems retrieves documents that are ranked, clustered, and returned as results.

In Chapter 6, we present two browsers that allow users to navigate in the aggregatio at intra-event level and at cross-event level. Both browsers access the aggregatio through the indexing system.

Chapter 6

Meetings Browsing

The efficiency of retrieving information depends not only on the quality of data mining and indexing, but also on opportunities proposed to the users for searching and browsing. The user is the final consumer of information and the only who can decide what is relevant or not. Furthermore, each person has a different perception of information, influenced by its cultural context, its personal experiences and sensibility, and so on. Unfortunately, although search engines integrate more and more powerful techniques for indexing data, their user interfaces last primitive. In particular, we criticize the *presentation* of retrieved information and the limited functionalities offered to users for *interacting* with the indexing system.

On one hand, Google-like interfaces display few results at the same time, although users are able to rapidly interpret visualizations representing hundreds of documents. Furthermore, current visualizations do not restitute visual feedbacks about archive's structure. In this thesis, we propose to integrate infovis techniques (as defined in Section 2.4) in graphical user interfaces, in order to represent as many documents as possible and to elicit the relationships between them.

On the other hand, most existing systems associate the retrieval task to full text search, which consists in submitting a textual query and in refining it iteratively, until the desired information is found. Unfortunately, the browsing task is not taken into account. Our relational indexing system is well-suited to integrate both searching and browsing functionalities.

In this chapter, we present our browsers that take into account these criteria of *presentation* and *interaction*. Section 6.1 illustrates users' requirements, highlighting the functionalities they need to consult and manipulate the aggregationi. Section 6.2 describes *JFriDoc*, a meeting browser that acts at intra-event level. Section 6.3 presents *FaericWorld Browser* that allows to navigate in the aggregationi at cross-event level. Section 6.4 summarizes the whole architecture of FaericWorld, illustrating how to extend the existing system. Finally, Section 6.5 is dedicated to three user evaluations of JFriDoc meeting browser.

6.1 User Requirements

Section 3.3 presented a list of user tasks related to meetings recordings, e.g. information retrieval, replaying, and so on. In this section, we propose some visual and interactive features that a graphical user interface should integrate, in order to enable or, at least, to facilitate those tasks.

- *Replaying*. When users replay a meeting, the browser must reproduce synchronized multimedia data. Those data are subjected to meeting's time, which can be controlled by users using commands such as "play", "stop", "pause", "forward", and "backward". Furthermore, meeting's time can be controlled by interacting with focused multimedia documents. For instance, when a user clicks an utterance of the audio transcript, all the media should be synchronized at its time.

- *Information retrieval*. The retrieval is the process that matches a query with the content of the aggregatio. Queries are submitted by users, when they search textual content or browse by links. Meeting browsers must support both functionalities with interactive visualizations. After retrieval, the documents found by the system should be displayed in interactive visualizations, which elicit the relevance of the results in respect of the query. Information such as documents' properties, relationships, and relevance can be expressed by colors, shapes, locations, etc. These techniques help users to find the documents of interest. Moreover, filtering methods [4] allow to reduce the amount of visualized results, whereas *Link & Brush* mechanism [99] permits to interconnect different views displaying the same documents.

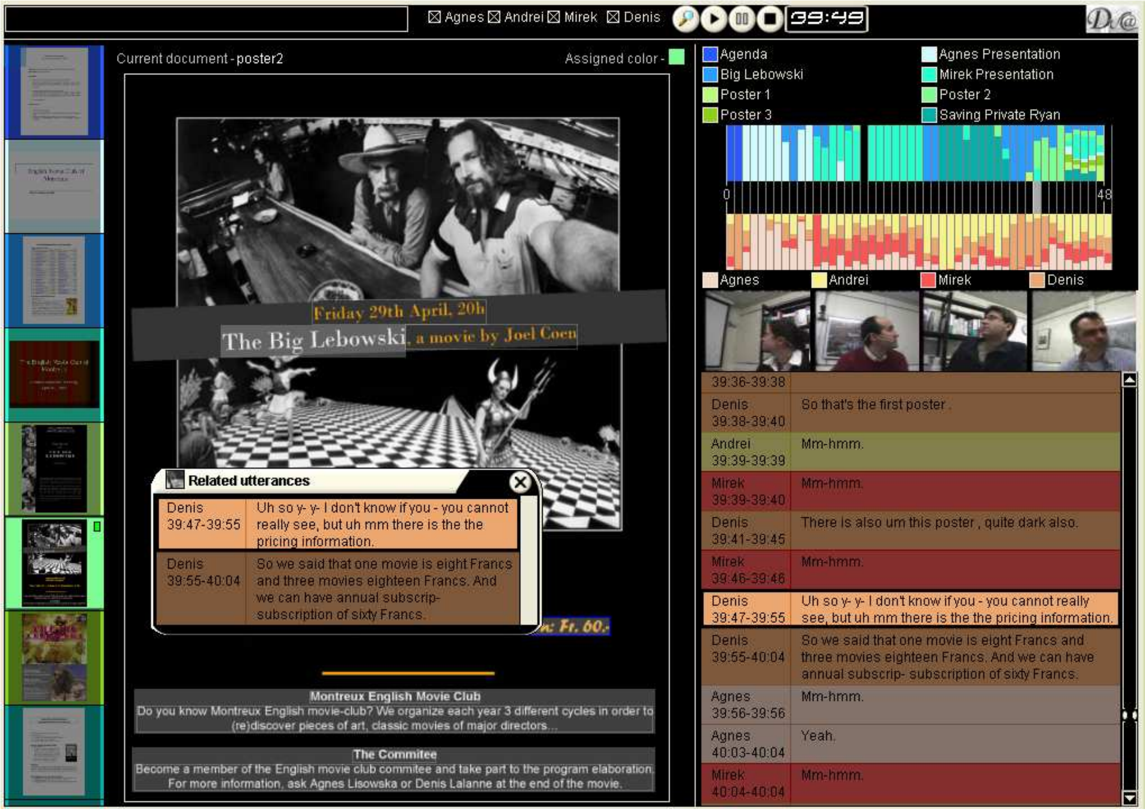
- *Minutes creation*, *edition*, and *meeting preparation*. These tasks imply that aggregati and links can be dynamically created, edited or deleted. These opportunities should be provided at different level of granularity, in order to allow to edit parts of documents, as well as to modify or create links between aggregati.

- *Survey*. Survey consists in studying the development of information in the aggregatio. For instance, the evolution of themes over a long period of time can be elicited through temporal visualizations.

The main challenge of our meeting browsers consists in offering all these functionalities in the same user interface. For instance, a user can submit a query, browse into the retrieved results, preview a meeting, and finally play it, without changing of interface or modality. In Sections 6.2 and 6.3, we present the browsers that implement the user tasks described above at intra-event level and at cross-event level.

6.2 JFriDoc, the Meeting Browser

JFriDoc is a document-centric browser that allows to navigate in a meeting aggregatus using cross-aggregati links. It visualizes the presented static documents, the videos of participants, the utterances of dialogue's transcript and it allows to interact with all these synchronized resources. For instance, a user can replay the whole meeting: while she lists at the audio recordings, the browser automatically highlights the corresponding utterances in the transcript and displays the static documents in the verbal focus. The main novelties of JFriDoc, in comparison with the meeting browsers presented in Section 2.2, are 1) the use of static documents and, in particular, of their physical and logical structures as artifacts for browsing the meeting and 2) the integration of visualization techniques for eliciting correlations between multimedia documents. Figure 6.1 is an overview of the meeting browser.



Denis 39:47-39:55	Uh so y- y- I don't know if you - you cannot really see, but uh mm there is the the pricing information.
Denis 39:55-40:04	So we said that one movie is eight Francs and three movies eighteen Francs. And we can have annual subscrip- subscription of sixty Francs.

Montreux English Movie Club
Do you know Montreux English movie-club? We organize each year 3 different cycles in order to (re)discover pieces of art, classic movies of major directors...

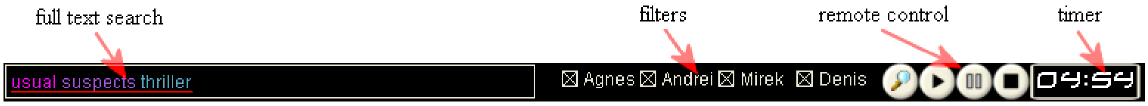
The Commitee
Become a member of the English movie club commitee and take part to the program elaboration. For more information, ask Agnes Lisowska or Denis Lalanne at the end of the movie.

39:36-39:38	
Denis 39:38-39:40	So that's the first poster .
Andrei 39:39-39:39	Mm-hmm.
Mirek 39:39-39:40	Mm-hmm.
Denis 39:41-39:45	There is also um this poster , quite dark also.
Mirek 39:46-39:46	Mm-hmm.
Denis 39:47-39:55	Uh so y- y- I don't know if you - you cannot really see, but uh mm there is the the pricing information.
Denis 39:55-40:04	So we said that one movie is eight Francs and three movies eighteen Francs. And we can have annual subscrip- subscription of sixty Francs.
Agnes 39:56-39:56	Mm-hmm.
Agnes 40:03-40:04	Yeah.
Mirek 40:04-40:04	Mm-hmm.

Figure 6.1: JFriDoc meeting browser allows to navigate in a meeting.

JFriDoc is based on JFerret framework [101, 55] and is composed of different synchronized modules that communicate together and integrate the individual media and modalities. These modules are presented in the following paragraphs.

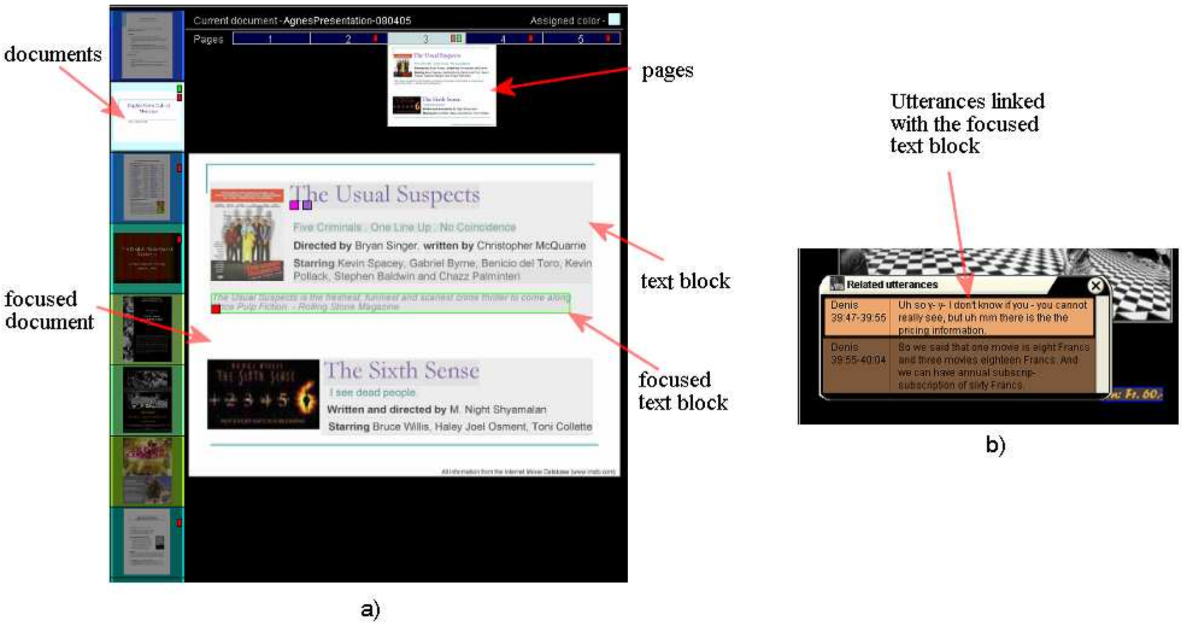
The *console module* allows to control the replaying of the meeting and integrates search functionalities. It is composed of a remote control, a timer, a component for full text searching, and of filters (confer Figure 6.2). The remote control allows to play, pause and stop the meeting recording. The timer indicates the current meeting's time. The component for full text searching permits to submit queries to the meeting browser, which retrieves the static documents and the utterances fulfilling the request. Each term of the query is associated with a specific color, in order to help users to immediately find the interesting information. Finally, filters allow to tune the query, in order to search only in the static documents and utterances belonging to one or more speakers.



usual suspects thriller

☒ Agnes ☒ Andrei ☒ Mirek ☒ Denis

Figure 6.2: The console proposes search, filtering, and replaying functionalities.



Current document - AgnesPresentation-080405

Assigned color -

The Usual Suspects

Five Criminals . One Line Up . No Coincidence

Directed by Bryan Singer, written by Christopher McQuarrie

Starring Kevin Spacey, Gabriel Byrne, Benicio del Toro, Kevin Pollack, Stephen Baldwin and Chazz Palminteri

The Usual Suspects is the freshest, funniest and scariest crime thriller to come along since Pulp Fiction. - Rolling Stone Magazine

The Sixth Sense

I see dead people

Written and directed by M. Night Shyamalan

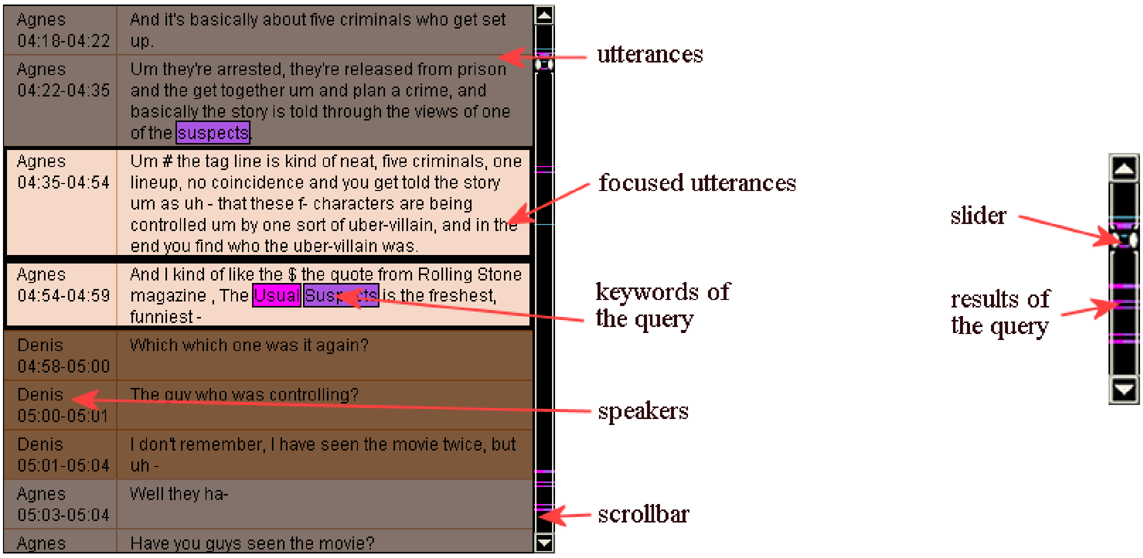
Starring Bruce Willis, Haley Joel Osment, Toni Collette

All information from the Internet Movie Database (www.imdb.com)

Denis 39:47-39:55	Uh so y- y- I don't know if you - you cannot really see, but uh mm there is the the pricing information.
Denis 39:55-40:04	So we said that one movie is eight Francs and three movies eighteen Francs. And we can have annual subscrip- subscription of sixty Francs.

Figure 6.3: The module allows to browse the meeting using static documents (a) and their links with audio transcript (b).

The *static documents module* is JFriDoc's main component (see Figure 6.3a), which allows to browse the meeting using static documents. They are represented as icons that can be clicked in order to open a document and to consult its pages. Each page can be selected and zoomed in the main area, in which it is displayed with its physical and logical structures. Static documents' structures are visualized as overlapped rectangles and can be clicked for accessing to the linked speech's utterances (cf. Figure 6.3b). When users replay the meeting, the documents or pages in verbal focus are displayed in the main surface. Similarly, the text discussed at a given moment is highlighted. Finally, the textual content corresponding to a query is emphasized with the colors associated to the keywords.



Agnes 04:18-04:22	And it's basically about five criminals who get set up.
Agnes 04:22-04:35	Um they're arrested, they're released from prison and the get together um and plan a crime, and basically the story is told through the views of one of the suspects.
Agnes 04:35-04:54	Um # the tag line is kind of neat, five criminals, one lineup, no coincidence and you get told the story um as uh - that these f- characters are being controlled um by one sort of uber-villain, and in the end you find who the uber-villain was.
Agnes 04:54-04:59	And I kind of like the $ the quote from Rolling Stone magazine , The Usual Suspects is the freshest, funniest -
Denis 04:58-05:00	Which which one was it again?
Denis 05:00-05:01	The guy who was controlling?
Denis 05:01-05:04	I don't remember, I have seen the movie twice, but uh -
Agnes 05:03-05:04	Well they ha-
Agnes	Have you guys seen the movie?

Figure 6.4: The transcript is displayed in a scrollable panel. Visualization methods help users to identify the focused utterances and the results of submitted queries.

The *transcript module* (see Figure 6.4) allows to navigate in the utterances of the spoken dialogue, chronologically sorted from the beginning to the end of the meeting. Utterances' background is filled with the color associated to the speaker that pronounced them. The transcript is synchronized with meeting's time: the focused utterances are highlighted, whereas the others are slightly obscured. Finally, the results of full text search are high spotted in both the utterances and the scrollbar, with the colors assigned to the keywords.

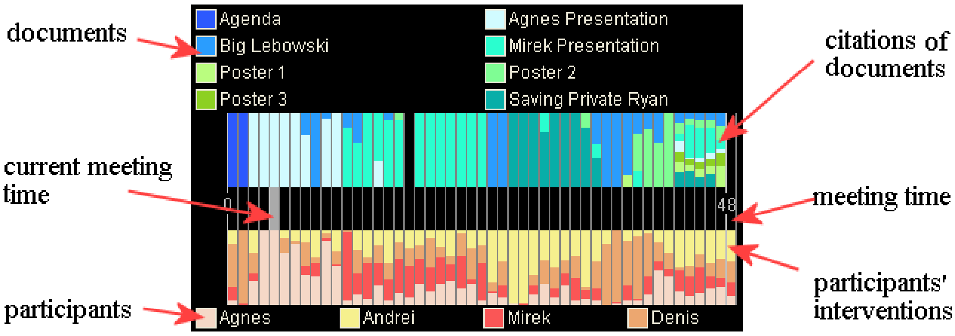


Figure 6.5: The meeting overview elicits the importance of documents and speakers during the meeting.

The *overview module* (cf. Figure 6.5) visualizes the distribution of speakers' turns and of focuses on static documents, relative to the whole meeting's time. The top histogram indicates when and how long the static documents were focused during the meeting. Similarly, the bottom histogram highlights the dominant speakers during the corresponding time interval. The combination of the two visualizations elicits correlations between the speakers and the documents: for instance, the histograms indicate who presented a static document or which one was the most discussed by participants.

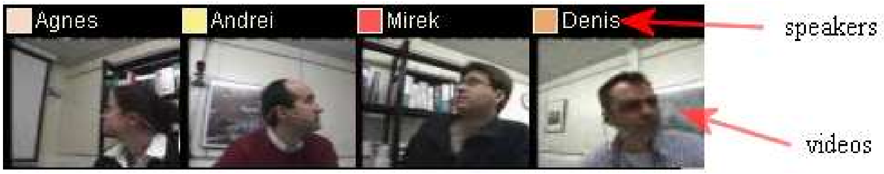


Figure 6.6: The videos of participant are labelled with their identities.

The *videos module* only reproduces videos, which are synchronized with the meeting's time (see Figure 6.6). Each video is labelled with speaker's name or identity code.

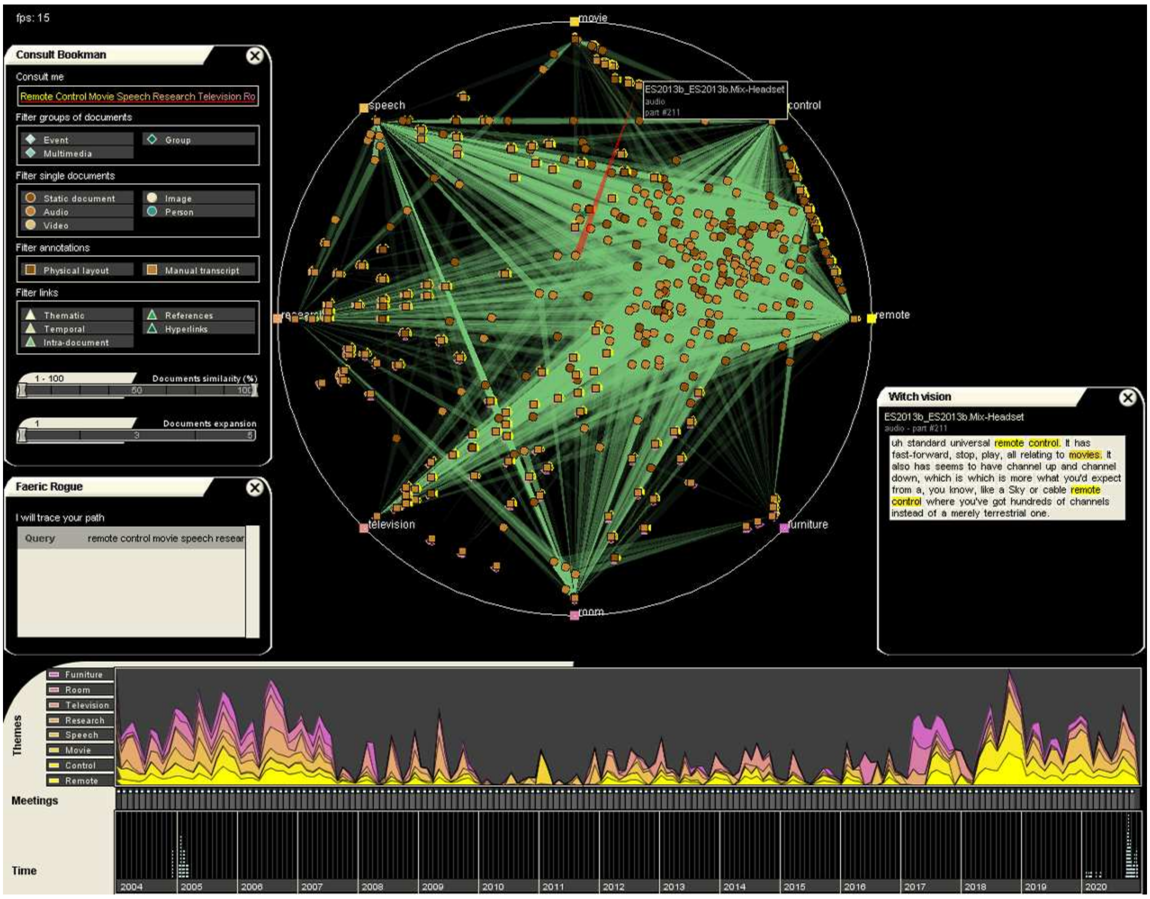
All these modules are synchronized and enable the user to switch between the different modalities thanks to temporal and thematic links. For instance, Denis is a member of a movie club, who did not participate to the last meeting. Thus, he consults the meeting recordings in order to discover the film selected for the next projection. Firstly, he opens the meeting's agenda and he finds the item concerning the selection of the next film to project. Clicking this paragraph, it opens a list of participants' utterances and he selects the first one, in order to replay the interesting part of the meeting. He discovers that the film is *Intacto*. Unfortunately, he does not know this film and thus he submits a query with film's title. Thanks to full text search, Denis retrieves both the static document describing the film and the audio recordings containing participants' opinions about it.

A special release of JFriDoc has been integrated in FaericWorld Browser, which acts at cross-event level and is presented in Section 6.3. Moreover, JFriDoc has been used in the evaluations described in Section 6.5.

6.3 FaericWorld Browser

This section presents FaericWorld Browser and, in particular, its interactive visualizations.

FaericWorld Browser allows to navigate into the aggregationi at cross-event level and takes full benefit of the indexing system presented in Chapter 5. Searching, browsing and replaying functionalities are integrated in the same graphical user interface. Figure 6.7 is an overview of FaericWorld Browser and its graphical components, which are described in the following subsections.



fps: 15

Consult Bookman

Consult me: Remote Control Movie Speech Research Television Ro

Filter groups of documents: Event, Multimedia, Group

Filter single documents: Static document, Audio, Video, Image, Person

Filter annotations: Physical layout, Manual transcript

Filter links: Thematic, Temporal, Intra-document, References, Hyperlinks

Documents similarity (%): 1 - 100

Documents expansion: 1

Faeric Rogue

I will trace your path

Query: remote control movie speech resear

Witch vision

ES2013b_ES2013b.Mix-Headset

audio - part #211

uh standard universal remote control. It has fast-forward, stop, play, all relating to movies. It also has seems to have channel up and channel down, which is which is more what you'd expect from a, you know, like a Sky or cable remote control where you've got hundreds of channels instead of a merely terrestrial one.

Figure 6.7: FaericWorld browser with its interactive visualizations.

6.3.1 Searching into the Aggregatio

Full text search is the retrieval mechanism integrated in the majority of existing browsers and search engines. In the same way, FaericWorld Browser allows users to submit textual queries to the indexing system presented in Chapter 5. The system consults the inverted file, in order to retrieve and to return the aggregati matching the query. The results, corresponding to the entry point in the aggregatio, are displayed on a radial visualization [46], also known as RadViz, and into a ThemeRiver [43]. The browser also proposes a default query for automatically accessing the aggregatio and for visualizing its global thematic structure. This default query is composed of the most recurrent terms in the inverted file that belong to disjoint sets of aggregati.

Figure 6.8 shows how the multimedia documents matching the query "Raffarin Chirac Bush Irak" are displayed into the RadViz. The keywords of the query are located on the frontiers of the RadViz and they act as attractors for the aggregati. The documents are disposed into the RadViz regarding at the amount and at the relevance of the keywords they contain. To be exact, a keyword attracts an aggregatus when its *tf-idf* value is high (see Subsection 4.3.2

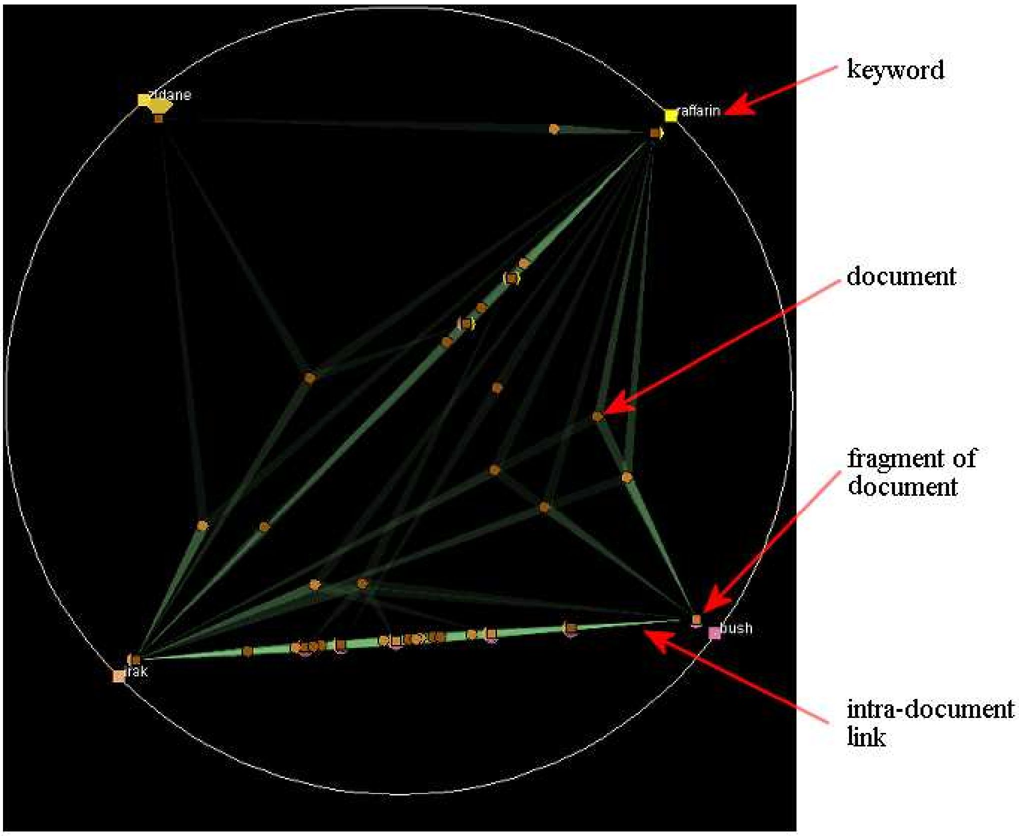
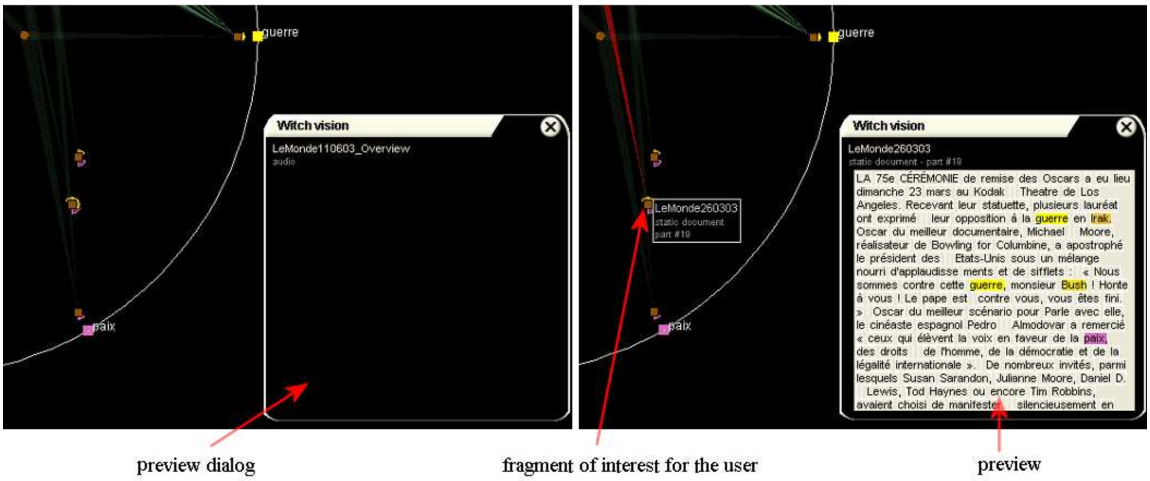


Figure 6.8: The RadViz visualizes the result of a query submitted to the indexing systems. The circles symbolize the aggregati and the squares represent their fragments.

about *tf-idf*). Thus, the aggregati located in the center of the RadViz ideally contain all the keywords of the query and these keywords have an equivalent relevance. At opposite, when a keyword is more influencing than the others, the aggregati are visualized in its proximity. For instance, Figure 6.8 shows that the "Zidane" name rarely occurs in the aggregatio. At opposite, "Raffarin", "Bush", and "Irak" keywords are frequently used, because several aggregati are attracted by their anchors. Furthermore, "Bush" and "Irak" are correlated in most of the documents retrieved by the system.

The fragments belonging to the aggregati are also visualized in the RadViz, using the *tf-idf* values of the keywords they contain for calculating their location. Each fragment is linked with its aggregatus by intra-aggregatus links, as shown in Figure 6.8.

The users can move the anchors of RadViz, in order to dispose the aggregati in a different way and to discover thematic groups. A mouse over an aggregatus of interest, or over one of its fragments, activates a preview of its content (see Figure 6.9), enriched with additional information such as document's name, amount of links, topics, etc. When the previews contain textual information matching queries, it is highlighted with the colors associated to the keywords.



Witch vision

LeMonde260303
static document - part #19

LA 75e CÉRÉMONIE de remise des Oscars a eu lieu dimanche 23 mars au Kodak Theatre de Los Angeles. Recevant leur statuette, plusieurs lauréat ont exprimé leur opposition à la guerre en Irak. Oscar du meilleur documentaire, Michael Moore, réalisateur de Bowling for Columbine, a apostrophé le président des Etats-Unis sous un mélange nourri d'applaudisse ments et de sifflets : « Nous sommes contre cette guerre, monsieur Bush ! Honte à vous ! Le pape est contre vous, vous êtes fini. » Oscar du meilleur scénario pour Parle avec elle, le cinéaste espagnol Pedro Almodovar a remercié « ceux qui élèvent la voix en faveur de la paix, des droits de l'homme, de la démocratie et de la légalité internationale ». De nombreux invités, parmi lesquels Susan Sarandon, Julianne Moore, Daniel D. Lewis, Tod Haynes ou encore Tim Robbins, avaient choisi de manifester silencieusement en

Figure 6.9: The user focuses a fragment of interest, which is visualized in the preview dialog. The terms belonging to the fragments that match the query are highlighted.

A drawback of the RadViz is that it can generate ambiguous visualizations of the results. In fact, an aggregatus can be displayed in the center of the radial visualization, although it contains only a part of the searched terms. This problem occurs when these keywords have the same *tf-idf* values and their attractors are placed at the opposite poles on the RadViz. In order to overcome this ambiguity, we propose to use a solution inspired to ShapeViz [93], for explicitly indicating as pies the *tf-idf* values of keywords. Figure 6.10 illustrates the two advantages of the visualization, indicating which and how many keywords are contained in the aggregatus or in a part of it. For instance, in Figure 6.10a, the pies suggest that the keywords are frequently used in the aggregatio, while in Figure 6.10b the pie is accentuated because a unique aggregatus contains the searched keyword.

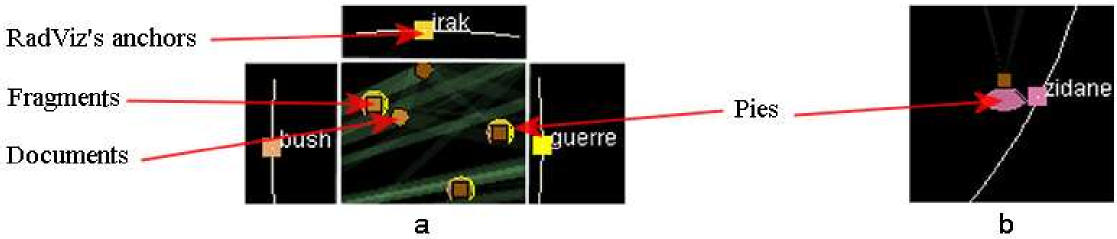


Figure 6.10: The pies representing *tf-idf* values surround the fragments of aggregati. In (a), the terms have the same relevance and the pies are small. At opposite, in (b) the pie is accentuated because the keyword is contained into a unique aggregatus.

The second main visualization of FaericWorld Browser is based on ThemeRiver [43]. This temporal visualization targets at displaying the evolution of themes in the aggregatio throughout time. Figure 6.11a shows that our ThemeRiver implementation is composed of three parts: the calendar, the list of meetings, and the themes plot. The calendar shows when the meetings have been recorded. The list of meetings contains the events, which are chronologically sorted from the oldest to the most recent one. Finally, the plot represents the evolution of the themes defined by the query. Each theme is represented such as a current. The strength of each river's current indicates the importance of a theme at a precise moment: it depends on the *tf-idf* values of the textual content contained in meetings' descriptors and in their documents. Thus, the meeting having the largest river is potentially the most interesting, relatively to the submitted query. For instance, in Figure 6.11a, the sixth meeting is the best candidate for fulfilling user's expectations.

The users can select a meeting in the calendar or in the list: its aggregati are thus highlighted in the RadViz, thanks to *Link & Brush* mechanism [99]. Furthermore, we propose a *focus + context* approach that allows to zoom into a meeting and to stretch the rest of the visualization (cf. Figure 6.11b). This functionality allows to visualize the themes evolution relative to the meeting's time, e.g. minute after minute.

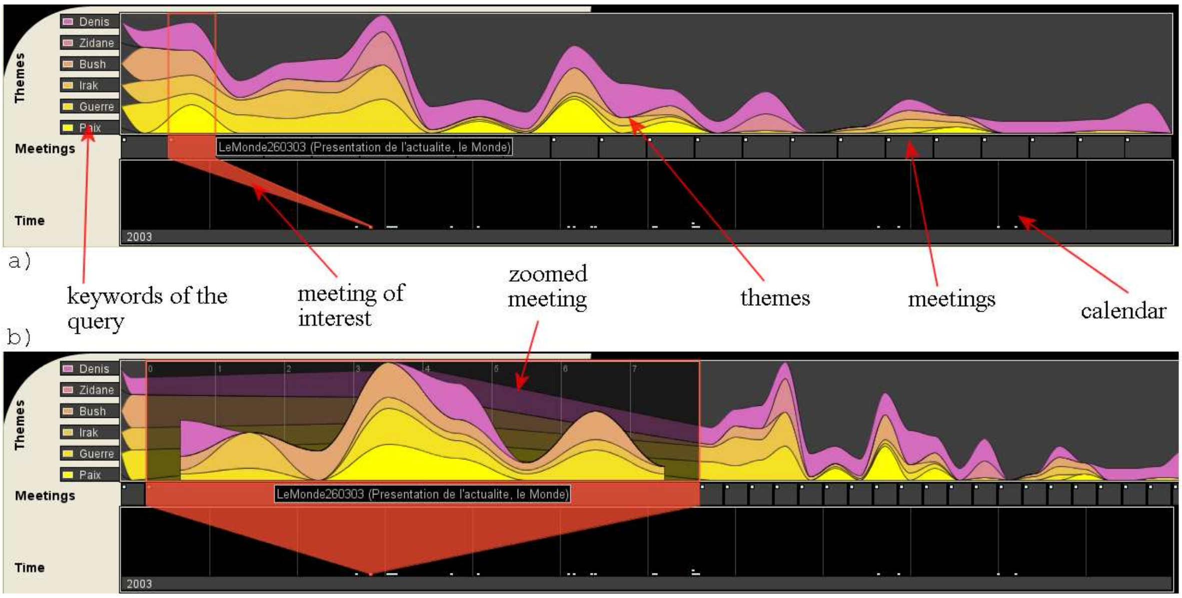


Figure 6.11: The visualization permits to study thematic evolution of meetings (a). Similarly, the zoom functionality allows to survey how topics evolve during a single meeting (b).

6.3.2 Browsing the Aggregatio Using Cross-aggregati Links

One of the major contributions of this thesis is the use of links for navigating in the aggregationi. In fact, users can find an interesting document after the submission of a textual query, but frequently full text search is not enough to satisfy their expectations. Similarly, sometimes the users cannot retrieve media such as videos by searching, because these documents are poorly indexed. Thus, links between aggregati are the solution we propose for discovering new information.

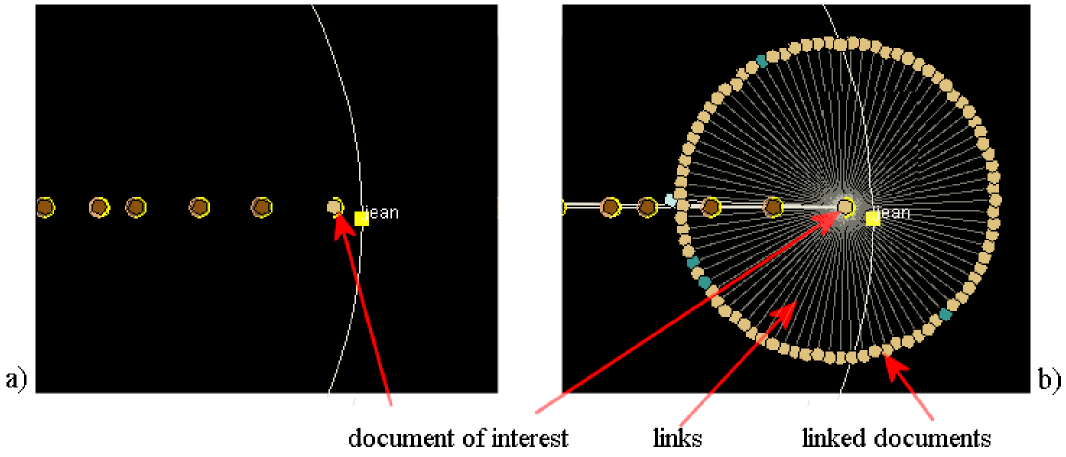


Figure 6.12: The user selects an aggregatus (a) and she discovers all related documents (b).

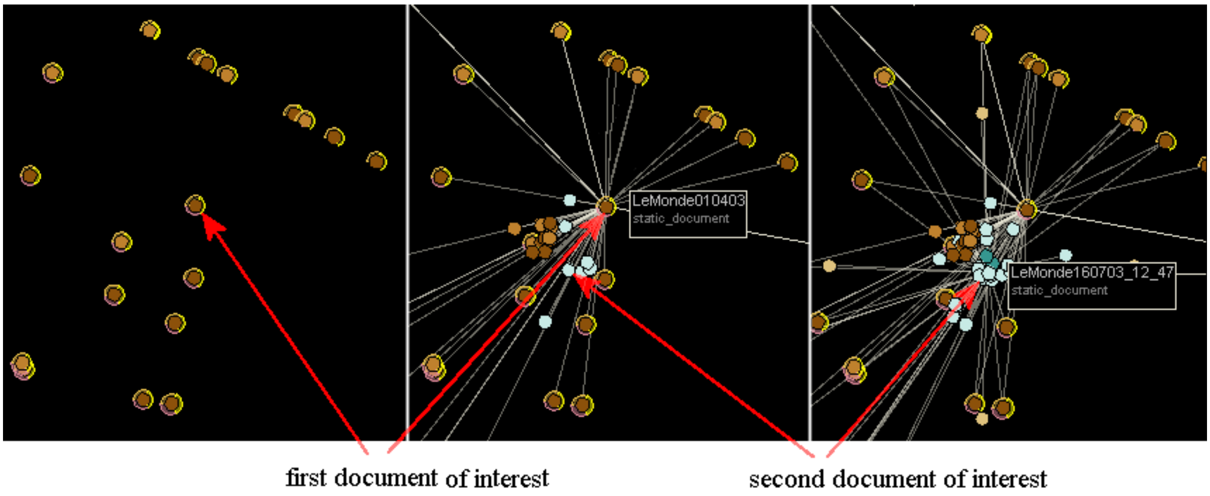


Figure 6.13: The user sequentially selects two aggregati for discovering their similar documents. The latter are linked to several aggregati already displayed in the RadViz.

When the users believe that an aggregatus returned by full text search is interesting, but does not completely fulfill their needs, they can select it for discovering similar documents. The linked aggregati are directly displayed in the RadViz: their locations depend on the relationships they have with already visualized results. Two different cases are taken into account: the new aggregati are linked either only with the document selected by the user, or with more aggregati already displayed in the RadViz. In the first case, the new documents are disposed around the aggregatus selected by the user. The distance between this aggregatus and its linked documents is inversely proportional to their similarity. For instance, Figure 6.12 shows the selection of an aggregatus (a) and the similar documents that appear in its proximity, because they do not possess any link with other visualized aggregati (b).

The second case is illustrated in Figure 6.13. If a new document is linked with two or more aggregati already displayed in the RadViz, its location is defined as the center of gravity of those aggregati.

Browsing the aggregatio increases the amount of information displayed in the RadViz. After a while, the visualization can be overloaded with aggregati that do not interest the user. We propose two solutions for overcoming this problem: documents removal and filtering. *Document removal* allows users to eliminate an uninteresting aggregatus from the RadViz. After this operation, FaericWorld Browser also removes its most similar aggregati. The users can adjust the similarity threshold using a slider.

Filters allow to select which categories of aggregati and links are visible (cf. Figure 6.14). For instance, users can hide static documents and temporal links. A slider allows to control similarity thresholds, permitting to make visible or not the links between aggregati. When an aggregatus discovered by browsing does not possess visible links anymore, it is automatically hidden.

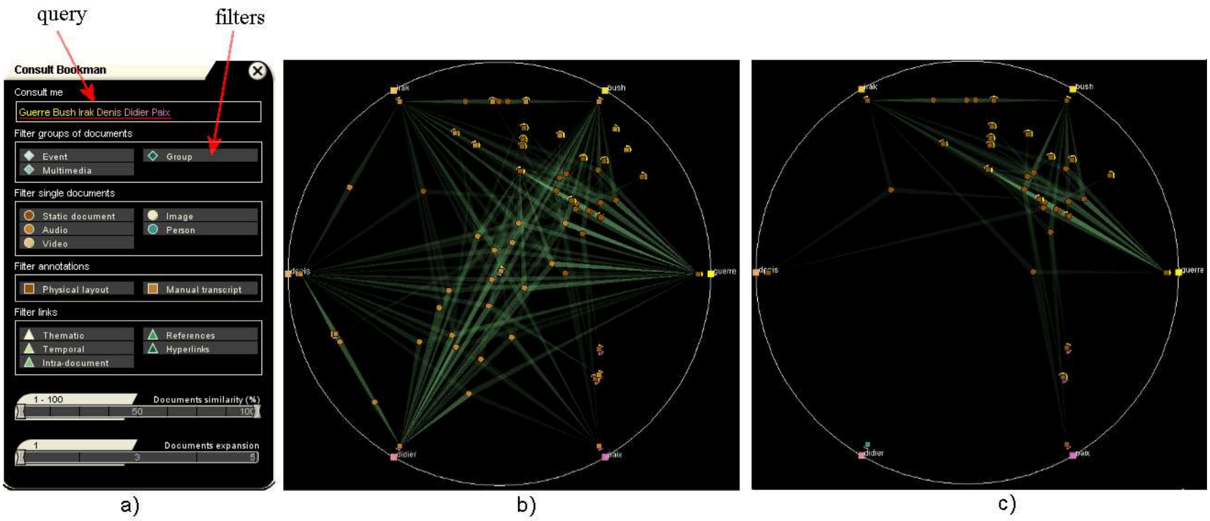
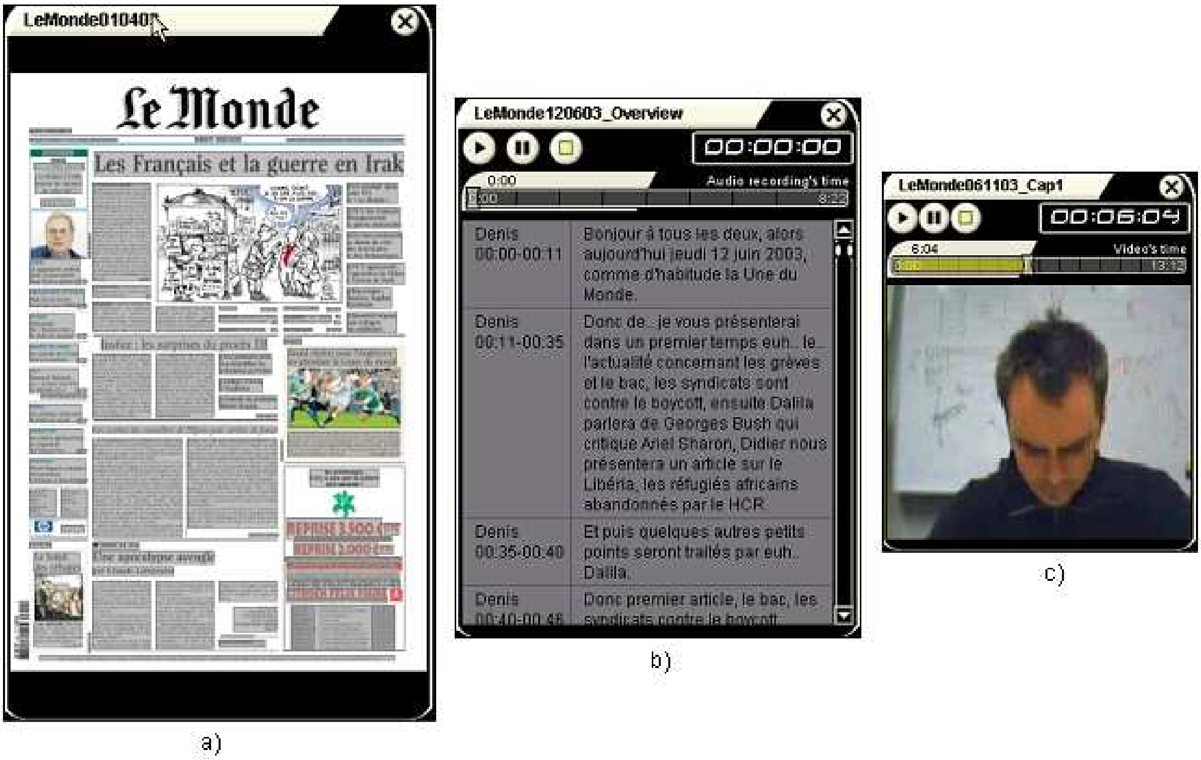

Consult me: Guerre Bush Irak Denis Didier Paix

Figure 6.14: The dialog provides filtering functionalities (a), which are useful for tuning the visualization (b). For instance, in (c) all audio aggregati have been hidden.

6.3.3 Viewing and Playing: Browsing in Time

When the user finds a document containing the searched information, she can view or play it. FaericWorld Browser offers specific views for visualizing each type of media, as illustrated in

Figures 6.15 and 6.16.



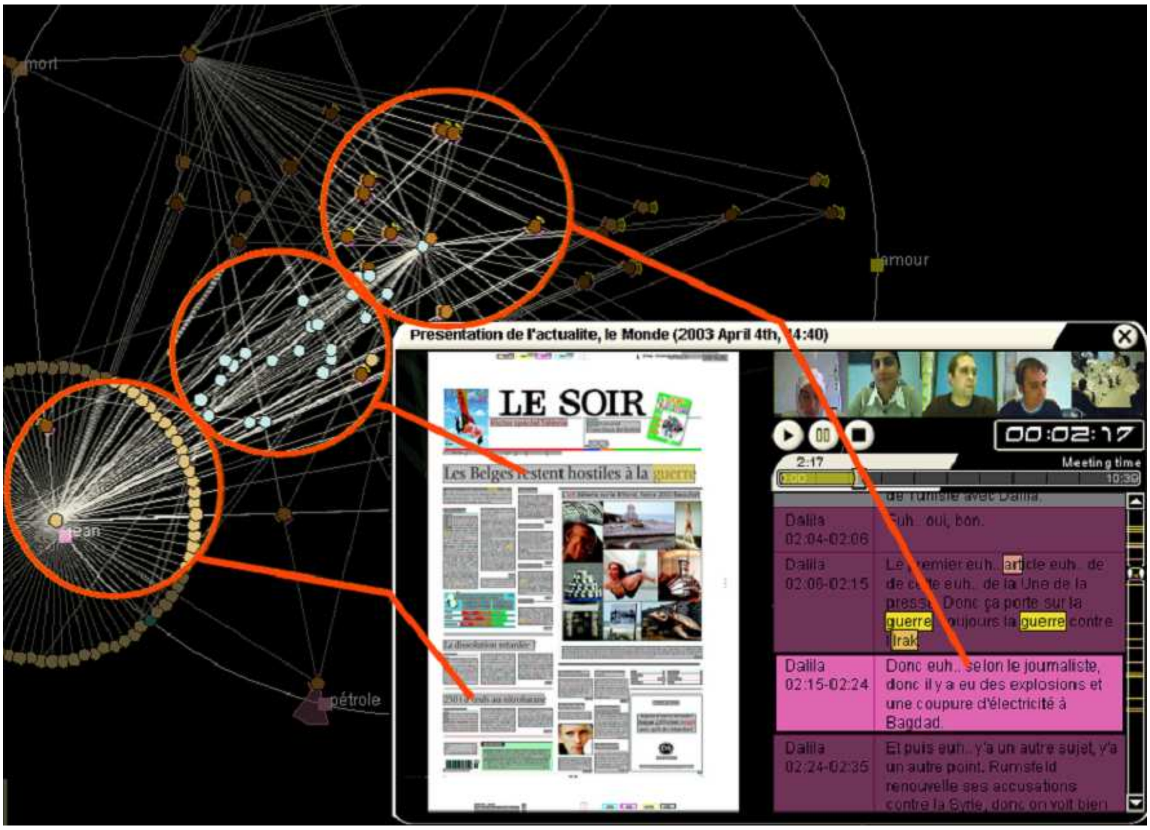
Denis 00:00-00:11	Bonjour à tous les deux, alors aujourd'hui jeudi 12 juin 2003, comme d'habitude la Une du Monde.
Denis 00:11-00:35	Donc de.. je vous présenterai dans un premier temps euh.. le.. l'actualité concernant les grèves et le bac, les syndicats sont contre le boycott, ensuite Dalila parlera de Georges Bush qui critique Ariel Sharon, Didier nous présentera un article sur le Libéria, les réfugiés africains abandonnés par le HCR.
Denis 00:35-00:40	Et puis quelques autres petits points seront traités par euh.. Dalila.
Denis 00:40-00:48	Donc premier article, le bac, les syndicats contre le boycott

Figure 6.15: The views allow to consult static documents (a), audio transcript (b), and videos (c).

Document retrieval and consulting are only a part of the browsing task. In fact, users examining documents are stimulated by new topics that protract the browsing experience. Consequently, when users consult the content of a document, FaericWorld Browser automatically proposes new similar aggregati thanks to the linking mechanism.

This functionality is especially interesting for meetings, because their themes and focused documents continuously change. Thus, the set of visualized aggregati is as well dynamic and evolves throughout the meeting's time. Figure 6.16 exemplifies this principle. For instance, a user retrieves and replays a meeting concerning the world politics. When she clicks the title of an article belonging to a PDF newspaper, the whole meeting is synchronized to the moment at which the article was discussed (cf. Section 6.2). At contemporary, a set of aggregati connected by thematic links is visualized into the RadViz. When the user decides to consult a new topic, she clicks another title and the thematic space changes. The aggregati previously focused into the RadViz are substituted with documents corresponding to the new theme.

Section 6.4 presents the entire architecture of FaericWorld and shows how its browsers are integrated with the systems presented in chapters 5.


mort
amour
ean
pétrole
Presentation de l'actualite, le Monde (2003 April 4th, 14:40)
LE SOIR
Les Belges restent hostiles à la guerre
00:02:17
2:17
Meeting time
0:00
10:30

de Tunisie avec Dalila.

Dalila 02:04-02:06	Euh.. oui, bon.
Dalila 02:06-02:15	Le premier euh.. article euh.. de de cette euh.. de la Une de la presse. Donc ça porte sur la guerre toujours la guerre contre l'Irak
Dalila 02:15-02:24	Donc euh.. selon le journaliste, donc il y a eu des explosions et une coupure d'électricité à Bagdad.
Dalila 02:24-02:35	Et puis euh.. y'a un autre sujet, y'a un autre point. Rumsfeld renouvelle ses accusations contre la Syrie, donc on voit bien

Figure 6.16: The topics evolve throughout the meeting, updating the links and the aggregati displayed in the radial visualization.

6.4 FaericWorld's Dataflow and Architecture

This section presents the dataflow and the full architecture of FaericWorld, i.e. the system that implements and integrates the analysis, indexing and browsing technologies presented in this thesis.

6.4.1 Dataflow

Figure 6.17 is an overview of the dataflow, summarizing the flow of information in FaericWorld. Firstly, the IM2.DI and AMI corpora are imported in the analysis module. Xed analyzes the static documents in PDF format and restructures them in XCDF (cf. Subsection 4.2.3). The physical structures extracted by Xed are validated by users using the Inquisitor tool (confer Subsection 4.2.4), which furthermore allows to add logical structures. Other media than static documents are imported through adapters and transformed into aggregati (cf. 5.2.1). After import, each meeting is automatically analyzed by multimodal alignment tool (see Subsection 4.3.2). The resulting links are validated using the Wizard of Faerie tool (cf. Subsection 4.3.3).

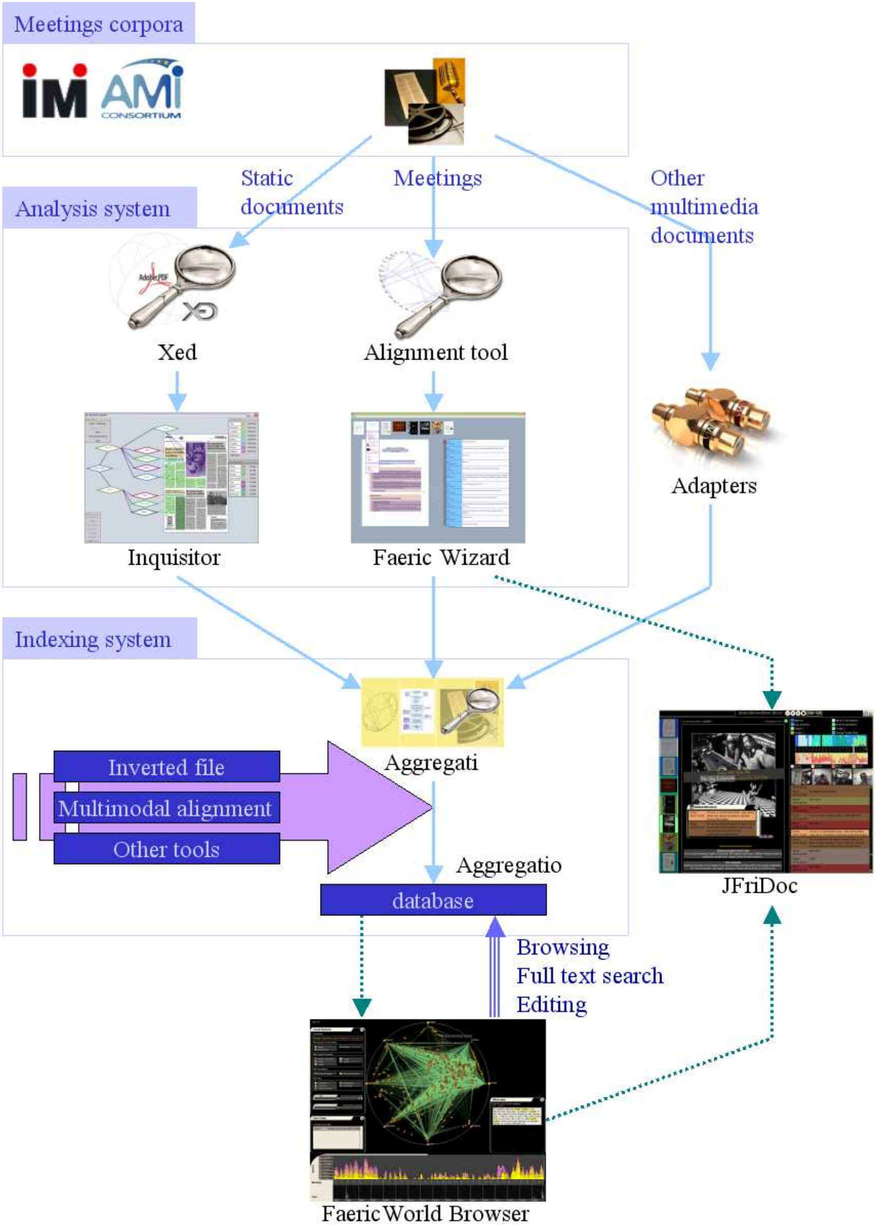


Figure 6.17: FaericWorld is our system that integrates the technologies for managing meetings collections, from analysis to browsing.

At this stage, it is already possible to browse the individual meetings with JFriDoc (as described in Section 6.2).

At indexing stage, the inverted file is updated with the textual information of the new aggregati, extracted from documents' content, annotations and metadata (cf. Subsection 5.2.3). Each aggregatus is thus completed with the cross-aggregati links calculated by the multimodal alignment tool presented in Subsection 5.2.4. Finally, the aggregatio and the indexes are stored into the database (see Subsection 5.2.6).

FaericWorld creates views of the aggregatio (cf. Section 6.3). These views are the results of manual and automatic queries (see Subsection 5.2.7), combined with techniques of ranking and clustering on retrieved documents (as presented in Subsection 5.2.5 and in Subsection 5.2.9). FaericWorld continuously updates its visualizations with the aggregati retrieved while searching and browsing. Finally, when the user consults a document, FaericWorld opens a specific view for each category of aggregati.

6.4.2 FaericWorld's Architecture

Figure 6.18 illustrates FaericWorld's architecture, with its configuration files on the left and the different components of the system on the right. Existing modules can be easily extended, in order to take into account new categories of documents, to encapsulate analysis tools, and to configure the visualizations of the browsers.

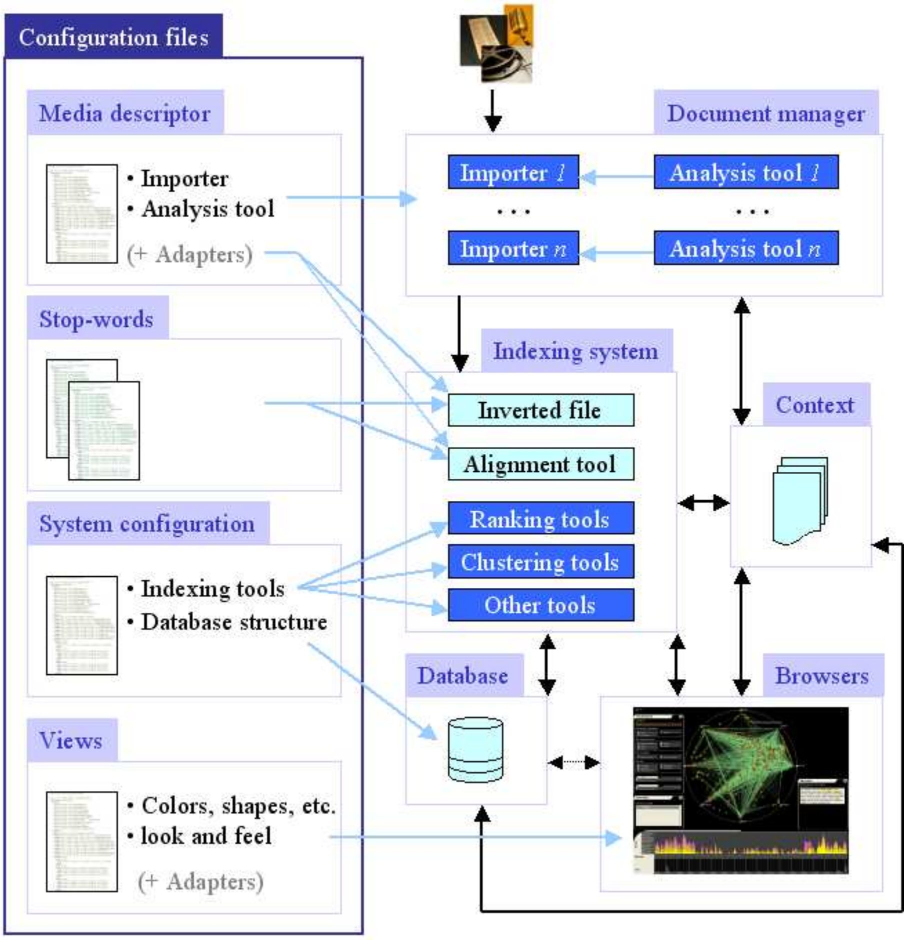
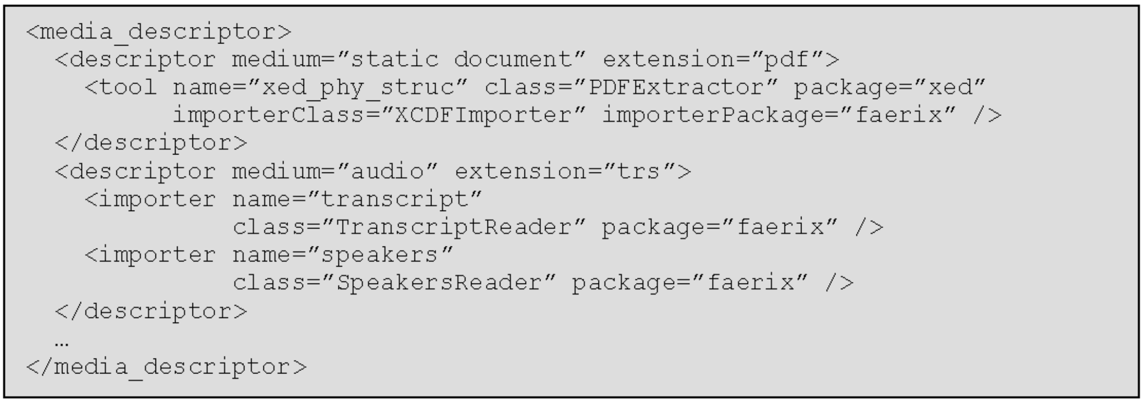


Figure 6.18: FaericWorld's architecture can be easily configured and extended.

Four configuration files help users to plug their modules in the system: the media descriptor, the system configuration file, the stop-words lists, and the views configuration file. Configuration properties can be accessed from everywhere, through a global context object.

The Media Descriptor

The *media descriptor* is an XML file that associates the document's formats with importers and analysis tools (cf. Figure 6.19).



```
<media_descriptor>
  <descriptor medium=”static document” extension=”pdf”>
    <tool name=”xed_phy_struc” class=”PDFExtractor” package=”xed”
          importerClass=”XCDFImporter” importerPackage=”faerix” />
  </descriptor>
  <descriptor medium=”audio” extension=”trs”>
    <importer name=”transcript”
              class=”TranscriptReader” package=”faerix” />
    <importer name=”speakers”
              class=”SpeakersReader” package=”faerix” />
  </descriptor>
  …
</media_descriptor>
```

Figure 6.19: An example extracted from the media descriptor.

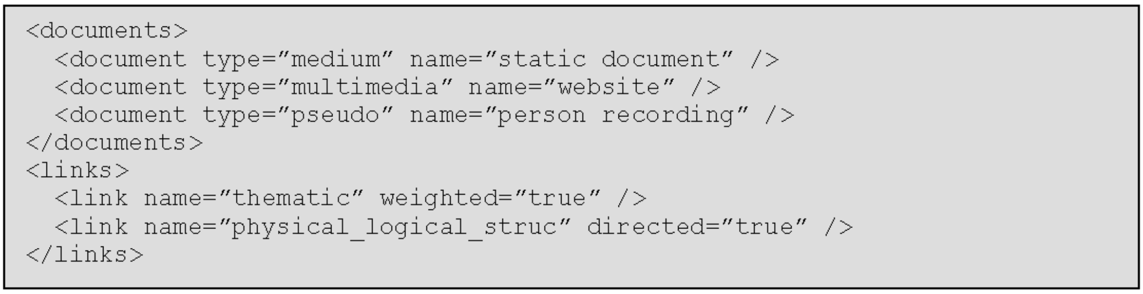
The importers convert an analyzed document into an aggregatus, containing the original file, its annotations, and its links, as explained in Subsection 5.2.1. For each annotation, the importer must define the corresponding adapter, used for accessing document's content and information. These adapters are further required by indexing system and by specific views of the browsers.

The importers are useful when users want to index new types of media. Similarly, when a researcher is developing a new analysis tool, she can first test its efficiency by importing its annotations. Developing a new importer consists of updating the media descriptor's XML and of extending the *Importer* abstract Java class. The new class must update the database with document's information. Further, the indexing system consults the database in order to retrieve the imported documents. The documents and their annotations are accessed using the declared adapters. The latter can be developed by users, which must implement the *Adapter* Java interface.

Users can also integrate reliable analysis techniques in the system. The main class of the analysis tool must implement the *AnalysisTool* interface. In this case, users will update the media descriptor, as well as select an importer for reading analysis tool's results. Alternatively, the users can develop an analysis tool that directly produces the aggregati, but this solution is discouraged because of modules reusability and maintenance reasons.

System Configuration File

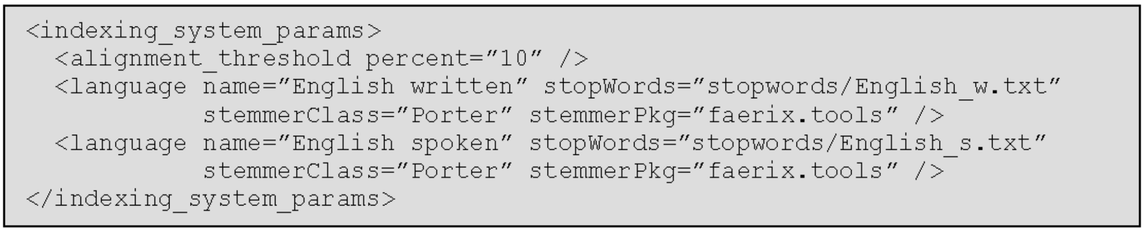
The *system configuration* is insured by an XML file, which allows to configure the indexing system and the database. First, the file contains the list of all the media and links managed by the system (see Figure 6.20). When users create importers for new media, they must also update the configuration file.



```
<documents>
   <document type=”medium” name=”static document” />
   <document type=”multimedia” name=”website” />
   <document type=”pseudo” name=”person recording” />
</documents>
<links>
   <link name=”thematic” weighted=”true” />
   <link name=”physical_logical_struc” directed=”true” />
</links>
```

Figure 6.20: Managed documents and links are enumerated in the configuration file.

Second, the configuration file contains the parameters for tuning the inverted file, the alignment tools, and textual querying mechanism (confer Figure 6.21). More precisely, it describes the files containing stop-words lists, the thresholds for alignment, and the classes used by stemming algorithms (cf. Subsection 4.3.2).

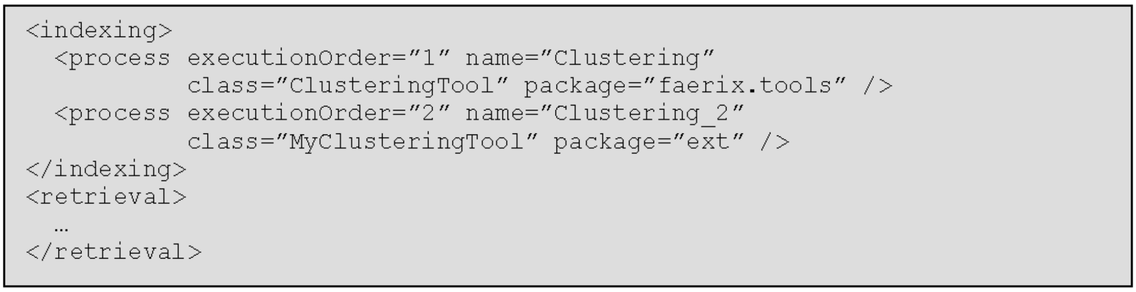


```
<indexing_system_params>
   <alignment_threshold percent=”10” />
   <language name=”English written” stopWords=”stopwords/English_w.txt”
             stemmerClass=”Porter” stemmerPkg=”faerix.tools” />
   <language name=”English spoken” stopWords=”stopwords/English_s.txt”
             stemmerClass=”Porter” stemmerPkg=”faerix.tools” />
</indexing_system_params>
```

Figure 6.21: Inverted file’s and alignment tool’s properties can be set in the configuration file.

Third, the configuration file allows to define chains of analysis processes, as illustrated in Figure 6.22. After the alignment, the users can integrate new plug-ins for analysis of data, such as clustering algorithms, ranking heuristics, and so on. Similarly, the users can implement new methods of classification for the documents retrieved while searching or browsing.

Finally, the configuration file contains the information for accessing the database, such as user’s name and password, server’s address, and used ports. Furthermore, it describes the database’s tables and their structure, which can be modified and extended for including new categories of documents and links. Additional tables can also be created in order to store the analysis results of tools created by researchers.



```
<indexing>
  <process executionOrder=”1” name=”Clustering”
           class=”ClusteringTool” package=”faerix.tools” />
  <process executionOrder=”2” name=”Clustering_2”
           class=”MyClusteringTool” package=”ext” />
</indexing>
<retrieval>
  …
</retrieval>
```

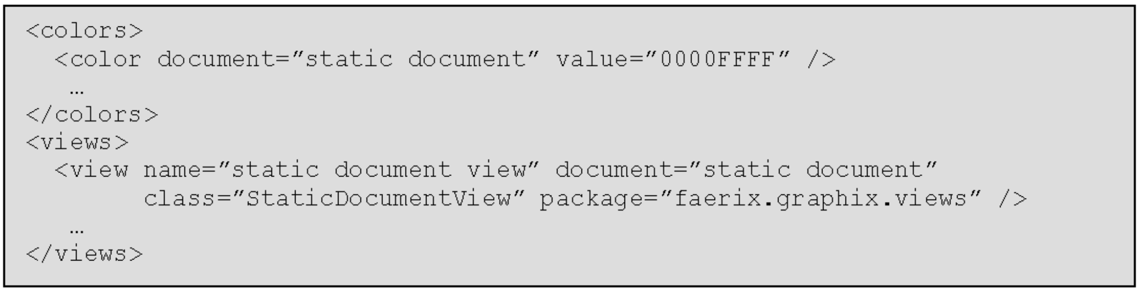
Figure 6.22: Users can define additional analysis processes after indexing and retrieving.

Stop-Words Lists

Stop-words lists are used for creating the inverted file, aligning the aggregati, and for filtering textual queries (cf. respectively Subsections 5.2.3, 5.2.4, and 5.2.7). These lists allows to filter from text words that are not considered thematically relevant, such as articles, adverbs, etc. Each stop-words list is related to an idiom or a unique modality of communication (e.g. written, oral, etc.) Our system supports multiple lists for managing multimedia corpora in different languages and, currently, it contains three sets of stop-words, i.e. for written English, spoken English, and French. Users can add new lists, by creating ASCII files containing a stop-word per line. The new lists must be referenced in the system’s configuration, as presented above, for being loaded by the indexing system.

Views Configuration File

The *views configuration file* allows users to associate the documents and links with colors and shapes, which are used by browsers. Furthermore, users can create new views for displaying and previewing the documents. In this case, they must implement the *ViewAdapter* Java interface and reference the new classes in the configuration file. Figure 6.23 illustrates both those functionalities.



```
<colors>
  <color document=”static document” value=”0000FFFF” />
   …
</colors>
<views>
  <view name=”static document view” document=”static document”
        class=”StaticDocumentView” package=”faerix.graphix.views” />
   …
</views>
```

Figure 6.23: This configuration file allows to integrate new views and define documents’ appearance.

The view configuration file also contains additional properties for describing the browsers' look and feel. For instance, users can change fonts, characters' size, borders, cursors, and so on.

In this subsection, we have presented the configuration files that help users to integrate new technologies. In fact, FaericWorld's architecture is flexible and can be easily extended with new plug-ins, integrating other categories of documents, analysis techniques, indexing and retrieving methods, and interactive visualizations. The new components can be developed by researchers locally, without requiring that they are familiar with the entire architecture. Once a new plug-in developed, the users will simply update the configuration files, without modifying system's components.

In Section 6.5, we present three evaluations centered on JFriDoc, aiming to test the efficiency of alignments for browsing, of static documents for retrieving information in a meeting, and of the modules composing the browser.

6.5 Evaluations of Document-centric Meeting Browsers

In this section, we present two experiments ran with JFriDoc and two methodologies for evaluating meeting browsers. In Subsection 6.5.1, we describe an evaluation of JFriDoc that aims at measuring the efficiency of multimodal alignments for browsing. The BET is a methodology for evaluating meeting browsers that is discussed in Subsection 6.5.2. Finally, in Subsection 6.5.3 we discuss the CIET, an evaluation method inspired by the BET that aims at measuring the usefulness and usability of different components belonging to a graphical user interface. This method has been applied for evaluating two components of JFriDoc.

6.5.1 Measuring Document Alignments Usefulness for Browsing

A user evaluation of JFriDoc has been performed on 8 users by Fati Arabchahi and Denis Lalanne [7]. Its goal was to test the usefulness of alignments between documents, in order to browse meetings at intra-event level. More precisely, the evaluation was focused on the thematic and temporal links existing between audio transcript and static documents presented during the meeting.

Actors

The evaluation was accomplished on two conditions:

- Users browse a meeting using JFriDoc, which allows users to take benefit of links between documents (*C1*);

- Users browse a meeting using a special release of JFriDoc, where the links between documents are not enabled (*C2*).

All users have tested both conditions, in order to control users' ability skills. The two browsers *C1* and *C2* are identical from a visual point of view, but the former supports document alignments, at the opposite of the latter.

Three meetings were involved in the evaluation: in the learning phase, *MT* has been used for training users, whereas *M1* and *M2* have been used for running the experiment. These meetings belong to IM2.DI archive (cf. Subsection 4.3.1), which is based on press reviews scenario. In *M1* the articles were presented regarding to newspapers' reading order, whereas in *M2* their were randomly discussed.

For each meeting, a list of four questions has been prepared. Two questions are mono-modal, i.e. their answer is either in static documents or in dialogues transcript. The other questions are multimodal: users can only retrieve their answer by looking at both the static documents and the transcript. For instance, a question such as "Which items of the agenda have been discussed during the meeting?" requires that users firstly retrieve the items in the agenda and, secondly, compare them with audio transcript. The questions of each meeting have been furthermore organized in two different groups: in the first group (*mM*), users must answer to monomodal questions and then to multimodal ones, whereas in the second set (*Mm*) their order is reversed. Two groups have been defined for avoiding a decrease of attention and thus to control a variable of the test.

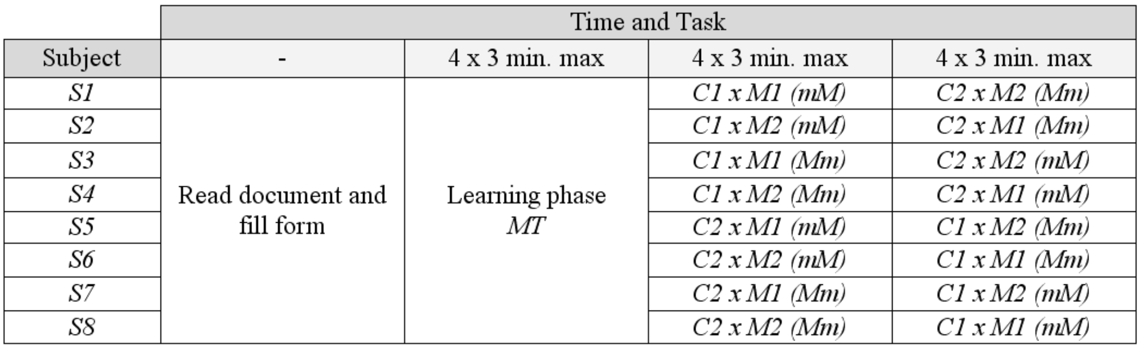
Protocol

The protocol consisted in testing 8 subjects during 20-45 minutes. A maximum of 3 minutes has been given for answering each question.

The protocol includes the following steps:

1. Each subject read a document explaining that "we evaluate browsers and not persons".
2. The participants fill a form with personal information.
3. The users read a document describing JFriDoc's visualizations and functionalities.
4. During the learning phase, the users browse the *MT* meeting with *B1*, the browser supporting alignments.
5. The users test the first condition with *C1* or *C2* (cf. Table 6.1).
6. The subjects test the second condition with *C2* or *C1* (see Table 6.1).
7. Finally, users are interviewed about the experiment, in order to evaluate their satisfaction.

The complete protocol is summarized in Table 6.1, describing the conditions, the meetings, and the questions planned for each users.



	Time and Task			
Subject	-	4 x 3 min. max	4 x 3 min. max	4 x 3 min. max
S1	Read document and fill form	Learning phase *MT*	*C1 x M1 (mM)*	*C2 x M2 (Mm)*
S2			*C1 x M2 (mM)*	*C2 x M1 (Mm)*
S3			*C1 x M1 (Mm)*	*C2 x M2 (mM)*
S4			*C1 x M2 (Mm)*	*C2 x M1 (mM)*
S5			*C2 x M1 (mM)*	*C1 x M2 (Mm)*
S6			*C2 x M2 (mM)*	*C1 x M1 (Mm)*
S7			*C2 x M1 (Mm)*	*C1 x M2 (mM)*
S8			*C2 x M2 (Mm)*	*C1 x M1 (mM)*

Table 6.1: The protocol is detailed for each subject.

Results

The evaluation measured both qualitative and quantitative aspects: success and failures in answering questions, the time required for each answer, the number of clicks, and so on. With documents alignments (*C1*), the users have solved the 76% of the questions, against the 66% of correct answers with *C2*.

When focusing on multimodal questions, the difference between *C1* and *C2* are particularly significant. Alignments have allowed subjects to solve 70% of questions. At opposite, only 50% of answers were correct without alignments.

This evaluation indicates that alignments between documents improve user performances while browsing meetings. Furthermore, it suggests that static documents can be used as interactive visualization for accessing multimedia information.

6.5.2 The Browser Evaluation Test

The user evaluation of meeting browsers is still an open domain. Up to now, the researchers exclusively measured users' satisfaction instead of their performances and did never define a strict evaluation protocol [94]. Recently, the Browser Evaluation Test (BET [102, 100]) has been designed by IM2.HMI IP[1] and AMI project[2], in order to fill this lack. The BET aims at objectively comparing the performances of different meeting browsers and at evaluating the individual mono-modal components, e.g. scrollable audio's transcripts, interactive visualizations of documents, etc.

The BET method consists in the following steps (as illustrated in Figure 6.24):

- *Corpus creation.* The meetings are recorded in the meeting rooms. Their multimedia documents are further enriched with annotations, either manually added by researchers or

[1] Human Machine Interaction Individual Project, Swiss National Center of Competence in Research on Interactive Multimodal Information Management (IM2), http://www.im2.ch

[2] Augmented Multi-party Interaction, http://www.amiproject.org

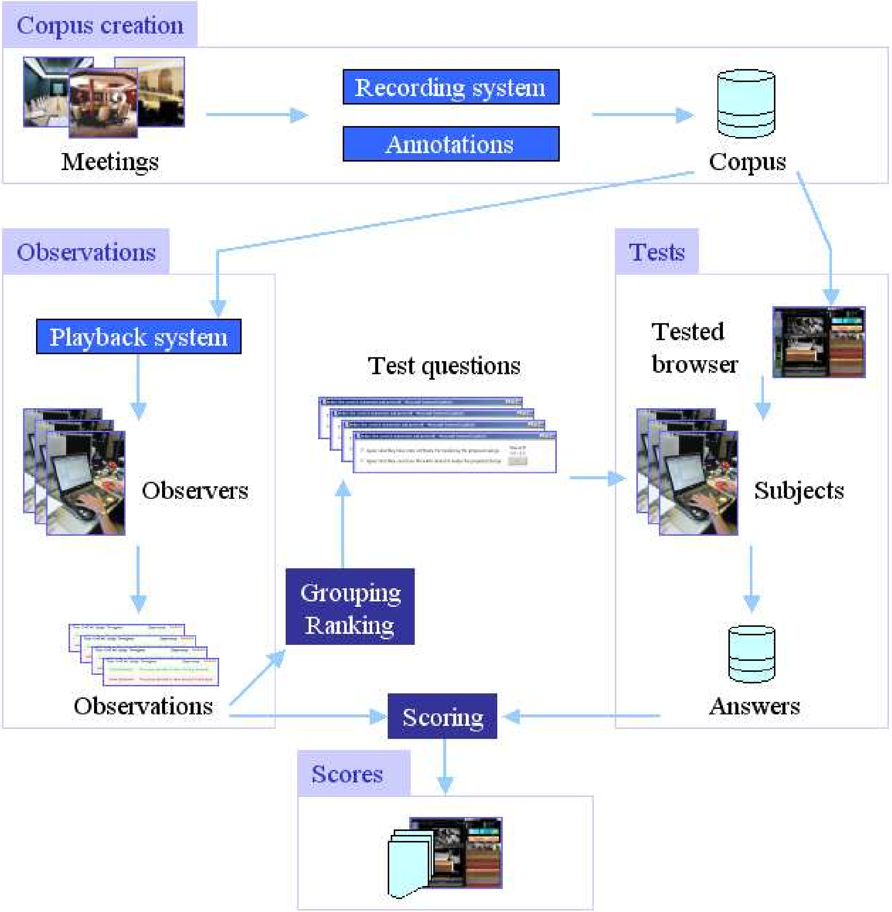


Figure 6.24: The BET method.

calculated by analysis systems. Currently, the BET corpus contains two meetings produced in the context of IM2 and AMI projects: the *IB4010* and *IS1008c*. The characteristics of these meeting will be described in Subsection 6.5.3.

- *Observations production.* External observers replay the meetings and define observations that are of interest for meeting participants. Observations are then turned into questions. The questions are true or false assessments and are stored in a database. Engaged observers did not participate to the meetings and are remunerated.

- *Tests production.* Part of recorded observations is selected for producing the tests. The selection has been accomplished by regrouping similar observations, selecting the most popular, and by defining a representant per group.

- *Browsers tests.* The subjects use the different browsers, answering to a maximum of test questions as faster and well as possible. The subjects are persons that did not participate neither to browsers development nor to observations production.

- *Scoring.* The results of test are compared with the correct answers of original observations and the browser's performance is scored. The scores allow to compare tested browsers.

The DIVA group has been involved in the definition of the method presented above and has participated to the production of tests from recorded observations. JFriDoc has been developed for participating in the BET test campaign, in order to test the efficiency of static documents for browsing a meeting.

Currently, the BET is in its final stage and some browsers of IM2 partners have already been evaluated. JFriDoc will be tested in the next future. However, while designing the BET, we have decided to evaluate of JFriDoc with another method, presented in the Subsection 6.5.3.

6.5.3 Component Integration Evaluation Technique

The Component Integration Evaluation Technique (CIET) is a method inspired by the BET. It has been designed to test the usefulness of the individual components belonging to a graphical user interface (GUI). The main idea is to compare a base-GUI with simplified releases of it. Each release contains all the components except one, which can be resized, substituted or eliminated. During the test, each release is compared to the base-GUI, which allows to normalize the scores of different tests. In this manner, it is possible to study the usefulness of a component relatively to the others. Furthermore, the experiment can be asynchronously ran in time, e.g. by testing a component per month with different groups of users.

In the next paragraphs, we explain the CIET method, by applying it to the evaluation of two JFriDoc's components, as illustrated in Figure 6.25.

Actors

JFriDoc plays the role of base-GUI (*B1*). The browser has been also simplified, obtaining two releases of it, respectively without audio transcript (*B2*) and without static documents (*B3*) modules. Although *B2* is static document centric, it allows to access speech thanks to the links created by multimodal alignment. For instance, the user can click a paragraph and the linked utterance is played.

The two meetings belonging to BET corpus have been selected for this evaluation. The first one is the *IB4010* (*M1*) that simulates the meeting of a movie club, in which the members select a film to project and the design of its advertising poster. The *IS1008c* is the second meeting (*M2*) and replicates the debate between the leaders of a company producing remote controls. Both meetings are in English, but they have various structural differences: *M1*'s duration is longer and the meeting contains more static documents than *M2*. At opposite, *M2*'s dialogues are less fragmented than those of *M1*. These two meetings have been selected for balancing the tests with two meeting browsers and for controlling users' bias.

Defined questions simulate a task, in which persons that did not participate to a meeting

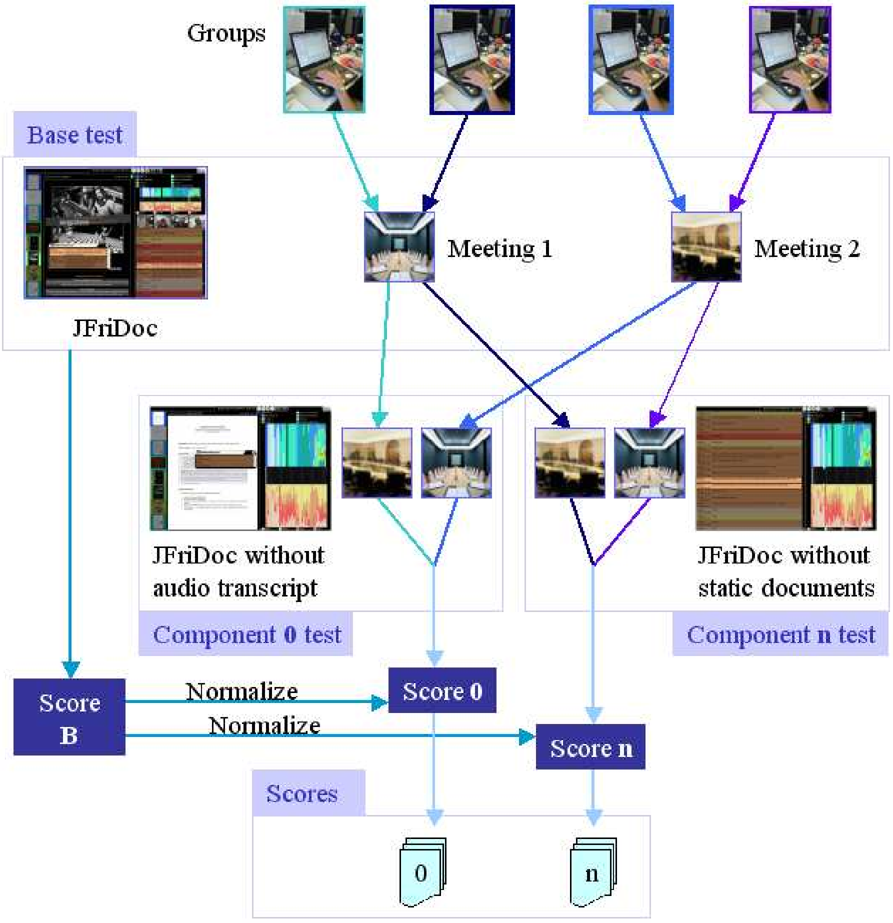
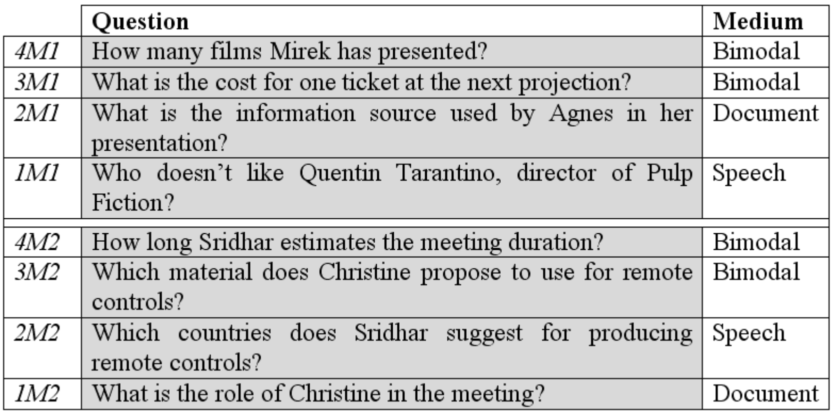


Figure 6.25: The CIET method used for evaluating JFriDoc components.

search interesting information. Since in this evaluation a component integrates a unique modality, we also defined monomodal and bimodal questions about the two meetings. The answers to the first category of questions are exclusively contained in one media, i.e. in static documents or in audio recordings. At opposite, the answers to bimodal questions exist in both the media. Fours questions per meeting have been selected (cf. Table 6.2).

The questions are expressed in a neutral way: they do not suggest the modality in which users can find the answer and, thus, the interactive component to use. For instance, explicit questions such as "Who told..." or "Where has been written..." have been avoided. Moreover, the questions are independent from participants' personal knowledge and the answer can be extrapolated from the meeting content only. For example, a question such as "Who is the actor playing the leading role in the film..." is not relevant for this evaluation.

The participants to the evaluation were students in economic sciences and biology, which are familiar with computers without being specialist, and no one was native English speaker.



	Question	Medium
4M1	How many films Mirek has presented?	Bimodal
3M1	What is the cost for one ticket at the next projection?	Bimodal
2M1	What is the information source used by Agnes in her presentation?	Document
1M1	Who doesn't like Quentin Tarantino, director of Pulp Fiction?	Speech
4M2	How long Sridhar estimates the meeting duration?	Bimodal
3M2	Which material does Christine propose to use for remote controls?	Bimodal
2M2	Which countries does Sridhar suggest for producing remote controls?	Speech
1M2	What is the role of Christine in the meeting?	Document

Table 6.2: Four questions have been selected for each meeting.

Protocol

The protocol of this evaluation consists in testing the browsers with 8 users, during 45 minutes per person. The participants are separated into four groups, defined by the first tested meeting, i.e. *M1* or *M2*, and by the used browsers, *B1B2* or *B1B3*.

The protocol is composed of the following steps:

1. The participants read a document explaining that "we evaluate multi-modal browsers and not persons' capacities".

2. They fill a form with their generalities, skills and so on.

3. The users watch a video-tutorial describing JFriDoc and its features. Then, they receive a document that summarizes the tutorial and that can be consulted all along the test. We do not provide any oral explanation.

4. All the users browse into the first meeting (*M1* or *M2*) with JFriDoc (*B1*) and answer to the corresponding questions presented in Table 6.2. *B1* plays a double role: it allows users to gain in familiarity with JFriDoc and enables to normalize the scores of the following step.

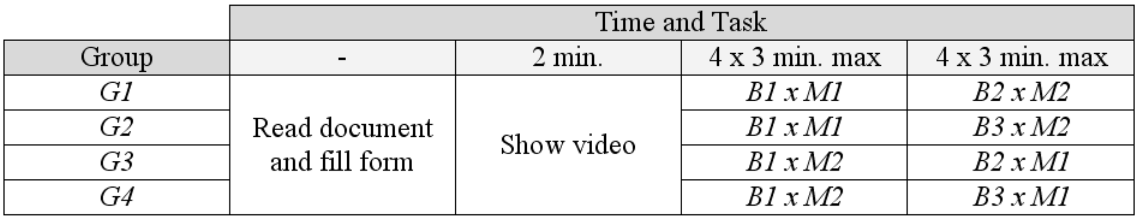
5. The participants write a summary of the meeting throughout 3 minutes.

6. The users browse the other meeting (*M2* or *M1*) with a simplified browser, either *B2* or *B3*, and they answer to the second list of questions.

7. They summarize the second meeting.

8. Finally, the users fill a satisfaction form about the tested browsers.

The protocol is illustrated in Table 6.3, in which the tasks of each single group are described in details.



	Time and Task			
Group	-	2 min.	4 x 3 min. max	4 x 3 min. max
G1	Read document and fill form	Show video	*B1 x M1*	*B2 x M2*
G2			*B1 x M1*	*B3 x M2*
G3			*B1 x M2*	*B2 x M1*
G4			*B1 x M2*	*B3 x M1*

Table 6.3: All the groups follow the same protocol, but with different browsers and meetings.

Results and Feedbacks

The target of this CIET evaluation was to compare the relevance of a JFriDoc's component, relative to the others. More precisely, we measure the loss of efficiency when removing the static document component or the audio transcript component.

Table 6.4 summarizes users' performances with the different browsers and contains the answers given by the four groups of users (*G1-G4*). Their performances are presented according to the tested browsers and meetings. The headers from 4 until 1 indicate the questions presented into Table 6.2. A correct answer is symbolized with the mark *"x"*, whereas an empty cell signifies that the users gave a wrong answer or that the time elapsed. The success rates are calculated in the rows and columns represented by *"%"* symbol and indicate the correct answers, respectively for each user and for each question.

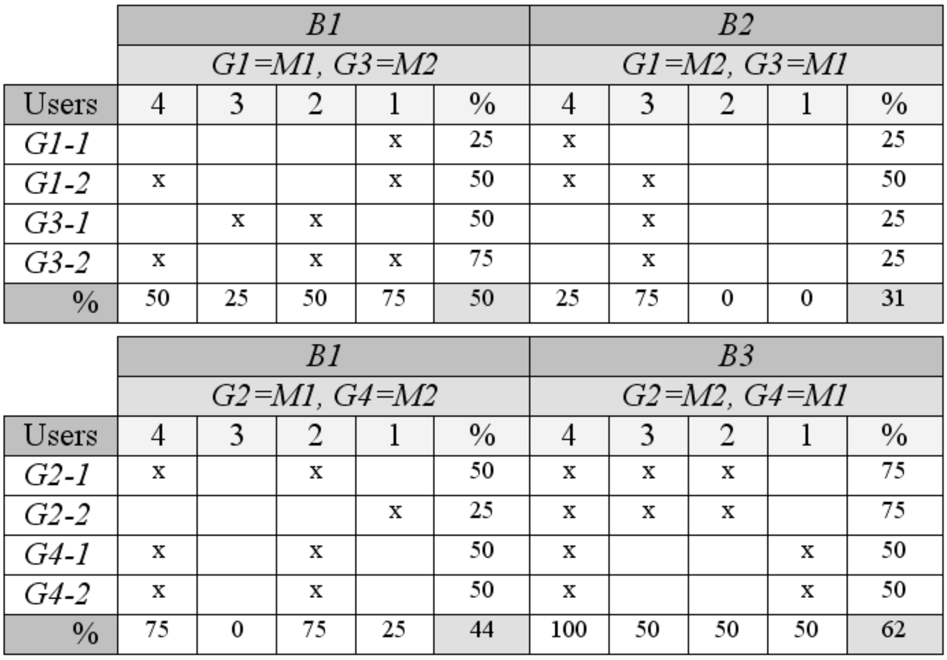
Table 6.4 summarizes the correct answers given by users. After normalization with *B1*, the performance factors of *B2* and *B3* are respectively 0.62 and 1.40. In other words, it means that removing the audio transcript component from JFriDoc reduces twice users efficiency than removing the static document component, for the given task.

The null scores of *B2* obtained with *2M1* and *1M2* mono-modal questions allowed to detect a lack of the static document component. In fact, the answers were written with small characters and, although the users found them, they were not able to read the text. Thus, zoom functionalities can be added to the component in order to improve its usability.

Finally, since in this evaluation a component integrates a modality, the non-null scores of both JFriDoc releases also suggest that static documents and audio transcript are useful for retrieving information.

The final step in the protocol was to fill a user satisfaction form. 75% of users never tried a multimodal interface before. All the users declared that the test of *B1* was "traumatic". However, 75% of them preferred this browser to *B2* or *B3* because it contains both modalities.

To wrap up, in this evaluation we assessed the CIET method, which targets at testing the



	B1					*B2*				
	G1=M1, G3=M2					*G1=M2, G3=M1*				
Users	4	3	2	1	%	4	3	2	1	%
G1-1				x	25	x				25
G1-2	x			x	50	x	x			50
G3-1		x	x		50		x			25
G3-2	x		x	x	75		x			25
%	50	25	50	75	50	25	75	0	0	31

	B1					*B3*				
	G2=M1, G4=M2					*G2=M2, G4=M1*				
Users	4	3	2	1	%	4	3	2	1	%
G2-1	x		x		50	x	x	x		75
G2-2				x	25	x	x	x		75
G4-1	x		x		50	x			x	50
G4-2	x		x		50	x			x	50
%	75	0	75	25	44	100	50	50	50	62

Table 6.4: The results obtained by users with the two combinations of browsers *B1B2* (a) and *B1B3* (b).

efficiency of GUI' individual components, relatively to a specific task. The test can be ran asynchronously and with different subjects. Furthermore, testing the base-browser with itself, ie. *B1* with *B1*, has several advantages:

- It allows to define a reference score, for comparing all individual components.
- Minor difference in the GUI can be tested: for instance, the size and the position of interactive visualizations.
- Variables of the test can be evaluated. For instance, a researcher can compare two meetings for estimating their similarity. Likewise, he can test if questions difficulty is equivalent for two experiments or not.
- It allows to test the quality of the training phase with a video tutorial. In fact, if users' performances significantly increase in the second experiment, the training phase was not adequate.

6.6 Conclusions

In this chapter, we have presented the browsers for navigating in the aggregatio, FaericWorld's dataflow and architecture, and three user evaluations based on JFriDoc.

JFriDoc is a document-centric browser that allows to replay and browse a meeting. It uses infovis techniques for representing multimedia documents, with their annotations and meta-data. Each module is interactive and can be used for browsing the synchronized multimedia information.

FaericWorld is another document-centric browser, which allows to navigate in a meeting collections. It mainly supports searching, browsing, replaying of multimedia information, and editing of documents. The main contributions of FaericWorld's are 1) the use of links for browsing and 2) its interactive visualizations for eliciting the structure of the collection.

The two browsers are integrated in the system presented all along this thesis. In this chapter, we have also presented its dataflow and architecture. In particular, we have focused on the extensibility of the architecture, by explaining how to integrate new categories of document, analysis tools, and interactive visualizations.

Finally, three evaluations based on JFriDoc have been described. The first experience proves that multimodal alignments, and thus links, improve the retrieval of information at intra-event level.

The second evaluation is the BET, which targets at providing a standard evaluation framework for comparing the efficiency of different browsers for retrieving meetings' information. The BET evaluation has not yet been applied on JFriDoc, but it is planned for a near future.

The third experience is inspired by the BET and aims at evaluating the individual components of a graphical user interface. In the case of JFriDoc, we tested the efficiency of users for retrieving meeting's information, when removing the static document component or the audio transcript component. The results have indicated that users are twice less performing without audio transcript component. Furthermore, they suggest that both audio and static document modalities are useful for browsing a meeting.

Those experiences have been ran at intra-document level. In the future, we would like to evaluate FaericWorld, by extending our user evaluations to cross-event level. In particular, since FaericWorld benefits of a modular and flexible architecture, it would be possible to test each component using the enhanced CIET method.

Chapter 7

Conclusion

In this thesis, we propose a method for structuring and browsing multimedia collections of meetings. The approach we have assessed is bottom-up: firstly, we have analyzed a single medium and, then, we have elicited its relationships with the other multimedia documents belonging to the same event. Finally, we have extracted the correlations existing in an entire collection of meetings. This method allows to create links that are optimal for successfully integrating the browsing task, which is ignored by most existing search engines.

A collection that has been analyzed in such a manner fulfills the model of aggregatio we propose in Chapter 3. An aggregatio contains the aggregati, which are composed of a document, its annotations and metadata, and its links towards other aggregati. This representation is adapted for single media (for instance, static documents, video, and so on), multimedia documents, and meta-documents (such as meetings and persons recordings).

At intra-document level (cf. Chapter 4), we mainly focused on static document analysis: this kind of medium can be analyzed with reliable methods, able to extract thematic and structural information. Static documents analysis involves three different phases. Firstly, we automatically extract document's physical structure, e.g. tokens, lines, and homogeneous text blocks. The users can validate the obtained results, in order to avoid errors propagation in the successive analyses. Considering the high recognition rates when extracting the physical structure, this phase is however not mandatory. Finally, users can manually add logical structures.

At intra-event level (presented in Chapter 4), a meeting is analyzed using multimodal alignment. This technique consists in comparing meeting's documents and in linking them when similarities are detected. During the comparison, the documents are accessed using the annotations and metadata produced by monomodal techniques of analysis. Multimodal alignment enables the transfer of temporal and thematic information between multimedia documents. After alignment, we propose that users manually validate its results, in order to prevent indexing errors.

The indexing of a meetings collection is based on two mechanisms (see Chapter 5). Textual information is indexed using an inverted file, i.e. a dictionary of the words contained in documents' content, annotations, and metadata. Each word is associated to references towards the documents containing it. The second mechanism is based on multimodal alignments and produces the links between documents. The two mechanisms enable respectively the searching and the browsing tasks.

The indexed collection can be explored using the browsers presented in Chapter 6. JFriDoc is a browser for navigating in a single event, e.g. a meeting. Its multimedia documents are represented in interactive visualizations, which are synchronized thanks to links and facilitate information retrieval. FaericWorld Browser proposes interactive visualizations that allow to explore an entire meetings collection. In, particular it permits to search textual information, to browse by links, and to replay retrieved multimedia documents.

In Section 7.1, we highlight the main contribution of this thesis. Finally, in Section 7.2, we discuss the limits of our techniques, the possible solutions, and the perspectives for improving the methods and the systems.

7.1 Contributions

The main goals of this thesis were triple. The first one was to use the high-level information extracted from static documents to complete the abstractions of media hard to mine, such as video, audio, and images. The second goal was to elicit the relationships existing between the multimedia documents in meetings collections. These relationships are useful for structuring, indexing, and browsing a collection. The last goal was to use techniques of information visualization for representing and browsing the collections of meetings. These goals have been achieved, by proposing a system that encapsulates and coordinates different technologies.

Our system has various peculiarities. Firstly, until now the researches dealing with mining, indexing, and retrieval of meetings were focused on one unique event at a time. In our work, we take into account the general context and we consider that the meeting is not an isolated event. This is a more realistic approach that opens new opportunities for end-users of meetings recordings.

Secondly, the multimodal alignment technique, applied to the entire collection of meetings, enables the effective transfer of high-level information between multimedia documents. Our work suggests that this technique can be applied to each category of multimedia archive.

Thirdly, we propose an extensible and incremental architecture for structuring an archive of multimedia documents. The underlying model is trivial, but sufficient for developing a system able to manage the individual document, as well as the complete multimedia collection. Moreover, the architecture integrates the user, which can correct analysis results, validate and edit indexes, and browse the meetings collection.

Fourthly, the new browsing paradigm takes full advantage of similarities between documents. The linking mechanism allows to browse between all categories of multimedia documents in the same visualization. Moreover, it permits to retrieve as well those media that lack of indexes.

Finally, various systems have been developed in order to integrate and validate our architecture for multimedia documents collections:

- ***Xed*** is an independent tool that extracts from PDF files their cleaned content and their physical structure. Currently, it efficiently accomplishes these tasks and it has already been used in projects such as SMAC [96].

- ***Inquisitor*** is a graphical user interface for validating and editing the physical structure and the logical structures extracted from static documents. Although the idea has already been implemented in other tools such as xmillum [45], Inquisitor explores new paradigms in its interactive visualizations.

- ***JFriDoc*** is a document-centric browser for replaying and navigating in a single meeting. Its main contribution is the use of visualization techniques for facilitating the retrieval of information. The views of the browser have been developed for managing overlaps, e.g. person speaking at contemporary, two or more static documents focused at a given time, multiple thematic links, etc. Moreover, JFriDoc is one of the first browsers proposing static documents as main artifact for navigating in a meeting.

- ***FaericWorld*** is composed of the *analysis and indexing system* and of the *browser*. The former integrates the different technologies proposed in this thesis, from the analysis to the indexing of meeting collections. The latter allows to navigate in the meetings aggregationi and implements the new browsing paradigm based on links. Furthermore, the views of the browser extend state-of-the art visualizations, such as RadViz and ThemeRiver.

7.2 Perspectives

This section suggests enhancements of the work presented in this thesis, presents unsolved issues, and proposes new perspectives for future works. Three subsections are dedicated to the improvement of indexing, the enhancement of visualizations, and to the evaluation of FaericWorld. Subsection 7.2.4 discusses how the methods presented in this thesis could be applied to the general context of the Internet.

7.2.1 Data Indexing

Data indexing is the bottleneck of FaericWorld system (cf. Section 5.4). In particular, the multimodal alignment of meetings is very expensive in computation time, because each document in the archive is compared with each other. More precisely, the task is accomplished in $O(n^2)$, where n is the amount of documents in the collection. Obviously, when this amount grows, the computational time become very significant (compare the import times of IM2.DI and AMI collections in Section 5.4). Moreover, in our current solution, the relationships between documents must be recalculated each time a new document is added to the aggregatio. The implemented system has been improved for importing a group of documents: firstly, it imports all the new documents, it updates the inverted file, and only at last it applies the alignment. Obviously, this improvement is minimal and do not accelerate the multimodal alignment technique itself. However, we propose different possible solutions to overcome this problem:

- *Selective alignment.* Instead of applying multimodal alignment over the whole aggregatio, it is possible to track the updates of *tf-idf* values in the inverted file, in order to detect the documents to be realigned. Moreover, if the modifications of *tf-idf* values are not significant, it is not necessary to reapply the alignment technique.
- *Alignment of representative documents.* When the archive becomes too big for being completely realigned, it is possible to align only characteristic documents, e.g. the most linked of the collection or the representatives of clusters.
- *Postponed alignment.* Another alternative consists in applying the multimodal alignment only after a certain amount of new imported documents.
- *Parallel calculations.* The multimodal alignment can be distributed on different machines. The algorithm is trivial and does not require any modification. The only precondition is that the inverted file must be updated before starting the alignment.

Besides performances improvement, the alignment technique can be extended to take into account the different categories of multimedia documents. For instance, textual information enriching videos can be considered more influent, because of its limited amount.

7.2.2 Visualization and Interaction Issues

A preliminary test has indicated that the RadViz visualization of FaericWorld browser can be difficult to interpret for the users. In fact, when the amount of displayed documents is huge, the visualization contains several overlapping of media and links, despite the dynamic visual filtering mechanism and the improvement of the standard RadViz. Moreover, the indexing system creates a ranking and clusters of documents that currently are not exploited. Consequently, we propose to substitute, or at least to support, the RadViz with other visualizations.

Currently, FaericWorld browser offers a set of interactive tools for consulting multimedia documents, zooming the results, filtering the visualized data, previewing information, and so on. Certain functionalities deserve to be extended or included. For instance, it would be possible to select two or more documents and to hide those being the most linked with the selection. Management of users' personal information could also be supported. For instance, users could reuse and exchange their searching and browsing histories, collaborate for retrieving information, or create their personal bookmarks and clusters for accessing and organizing the archive of meetings.

7.2.3 Evaluation

Various aspects of our document-centric approach for managing multimedia archives have been tested or validated in this thesis. First of all, our approach has been applied and assessed through multimedia meetings archives. Secondly, we demonstrated that the automatic analysis of PDF documents is reliable and provides valuable results, in order to structure multimedia archives through multimodal alignments. Thirdly, the evaluation presented in Subsection 6.5.1 assessed that multimodal alignment improves user performances while browsing a meeting. In further researches, other aspects of this thesis merit to be observed.

For instance, the evaluation presented in Subsection 6.5.1 could be extended for testing multimodal alignments' efficiency for browsing the entire collection of meetings with FaericWorld.

Similarly, the visualizations and the components integrated in FaericWorld could be tested using the CIET method presented in Subsection 6.5.3. Qualitative and quantitative evaluations could be set up in future, when alternative visualizations to the RadViz will be integrated.

7.2.4 Final assessment

From a practical point of view, the technology presented all along this work cannot be used for structuring and browsing very large archives. The indexing mechanism is in its early stage and it is too expensive in term of computational time and storage resources for being applicable. Several improvements and approximations are necessary for indexing such amount of documents: probably, the perspectives presented in the Subsection 7.2.1 are a good starting point.

The browsing paradigm based on links is an interesting alternative to the classical search task offered by existing search engine. Obviously, visualizing all the documents of a very large archive at a time is neither possible nor useful, but the mechanism could however be used for browsing in restrained similarity spaces or in a specific domain.

Although the methodology presented in this thesis is not mature for being applied to the Internet, it can already be used for archives containing personal documents [61], daily news, and other correlated multimedia information.

Finally, we believe that, in the future, information visualization will play a major role, by completing mining and indexing technologies, and by redefining multimedia browsing strategies.

Bibliography

[1] Able2Extract. http://www.investitech.com.

[2] Gregory D. Abowd, Christopher G. Atkeson, Jason Brotherton, Tommy Enqvist, Paul Gulley, and Johan LeMon. Investigating the capture, integration and access problem of ubiquitous computing in an educational setting. In *CHI '98: Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 440–447, New York, NY, USA, 1998. ACM Press/Addison-Wesley Publishing Co.

[3] Sébastien Adam, Maurizio Rigamonti, Eric Clavier, Eric Trupin, Jean-Marc Ogier, Karl Tombre, and Joël Gardes. Docmining: A document analysis system builder. In *Document Analysis Systems VI, 6th International Workshop, DAS 2004*, number 3163 in LNCS, pages 472–483, Florence (Italy), September 2004. Springer-Verlag.

[4] Christopher Ahlberg and Ben Shneiderman. Visual information seeking using the filmfinder. In *CHI '94: Conference companion on Human factors in computing systems*, pages 433–434, New York, NY, USA, 1994. ACM Press.

[5] Marita Ailomaa, Miroslav Melichar, Martin Rajman, Agnes Lisowska, and Susan Armstrong. Archivus: a multimodal system for multimedia meeting browsing and retrieval. In *Proceedings of the COLING/ACL on Interactive presentation sessions*, pages 49–52, Morristown, NJ, USA, 2006. Association for Computational Linguistics.

[6] Anjo Anjewierden. Aidas: Incremental logical structure discovery in pdf document. In *Sixth International Conference on Document Analysis and Recognition (ICDAR'01)*, pages 374–377, Seattle (USA), 2001.

[7] Fati Arabchahi and Denis Lalanne. Assessing the usefulness of document alignment for meeting browsing. Technical report, Department of Informatics, University of Fribourg, 2004.

[8] Steven R. Bagley, David F. Brailsford, and Matthew R. B. Hardy. Creating reusable well-structured pdf as a sequence of component object graphic (cog) elements. In *ACM Symposium on Document Engineering (DocEng'03)*, pages 58–67, Grenoble (France), 2003.

[9] Frédéric Bapst. *Reconnaissance de documents assistée : architecture logicielle et intégration de savoir-faire*. PhD thesis, University of Fribourg, Switzerland, 1998. thesis Nr. 1228.

[10] Ardhendu Behera. *A Visual Signature-based Identification Method of Low-resolution Document Images and its Exploitation to Automate Indexing of Multimodal Recordings*. PhD thesis, University of Fribourg, Switzerland, 2006. thesis Nr. 1529.

[11] Jean-Luc Bloechle, Maurizio Rigamonti, Karim Hadjar, Denis Lalanne, and Rolf Ingold. Xcdf: A canonical and structured document format. In *7th International Workshop, DAS'06*, number 3872 in LNCS, pages 141–152, Nelson (New Zealand), February 2006. Springer-Verlag.

[12] Jean-Luc Bloechle, Maurizio Rigamonti, Denis Lalanne, and Rolf Ingold. Xcdf : un format canonique pour la représentation de documents. In *Colloque International Francophone sur l'Ecrit et le Document (CIFED'06)*, pages 19–23, Fribourg (Switzerland), September 2006.

[13] Kurt D. Bollacker, Steve Lawrence, and Clyve Lee Giles. Citeseer: an autonomous web agent for automatic retrieval and identification of interesting publications. In *2nd International Conference on Autonomous Agents*, pages 116–123. ACM Press, 1998.

[14] Jason A. Brotherton and Gregory D. Abowd. Lessons learned from eclass: Assessing automated capture and access in the classroom. *ACM Trans. Comput.-Hum. Interact.*, 11(2):121–155, 2004.

[15] BMRC Lecture Browser. http://bmrc.berkeley.edu/frame/projects/lb/index.html.

[16] Horst Bunke and A. Lawrence Spitz, editors. *Document Analysis Systems VII*, volume 3872 of *Lecture Notes in Computer Science*. Springer Berlin / Heidelberg, February 2006.

[17] Horst Bunke and Patrick S. Wang. *Handbook of Character Recognition and Document Image Analysis*. World Scientific Publishing Company, janvier 1997.

[18] James P. Callan. Passage-level evidence in document retrieval. In *17th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 302–310. Springer-Verlag, 1994.

[19] Marco Campanella, Riccardo Leonardi, and Pierangelo Migliorati. An intuitive graphic environment for navigation and classification of multimedia documents. In *Multimedia and Expo, 2005. ICME 2005. IEEE International Conference on*, pages 743–746, 6-8 July 2005.

[20] Matthew Carey, Daniel C. Heesch, and Stefan M. Rüger. Info navigator: A visualization tool for document searching and browsing. In *Conference on Distributed Multimedia Systems DMS'03*, pages 23–38, 2003.

[21] Hui Chao and Jian Fan. Layout and content extraction for pdf documents. In *IAPR International Workshop on Document Analysis Systems (DAS'04)*, pages 213–224, Florence (Italy), 2004.

[22] Hui Chao and Lin Xiaofan. Capturing the layout of electronic documents for reuse in variable data. In *Eighth International Conference on Document Analysis and Recognition (ICDAR'05)*, pages 940–944, Seoul (Korea), 2005.

[23] Bidyut B. Chaudhuri. *Digital Document Processing*. Advances in Pattern Recognition. Springer, 2007.

[24] Patrick Chiu, John Boreczky, Andreas Girgensohn, and Don Kimber. Liteminutes: an internet-based system for multimedia meeting minutes. In *WWW '01: Proceedings of the 10th international conference on World Wide Web*, pages 140–149, New York, NY, USA, 2001. ACM Press.

[25] Michael G. Christel, Alexander G. Hauptmann, Howard D. Wactlar, and Tobun D. Ng. Collages as dynamic summaries for news video. In *MULTIMEDIA '02: Proceedings of the tenth ACM international conference on Multimedia*, pages 561–569, New York, NY, USA, 2002. ACM Press.

[26] Eric Clavier, Sébastien Adam, Pierre Héroux, Maurizio Rigamonti, and Jean-Marc Ogier. Docmining : Une plate-forme de conception de systèmes d'analyse de documents. In *Huitième Colloque International Francophone sur l'Ecrit et le Document (CIFED'04)*, pages 97–102, La Rochelle (France), June 2004.

[27] Eric Clavier, Gérald Masini, Mathieu Delalandre, Maurizio Rigamonti, Karl Tombre, and J. Gardes. Docmining: A cooperative platform for heterogeneous document interpretation according to user-defined scenarios. In *5th IAPR International Workshop on Graphics Recognition (GREC'03)*, pages 21–32, Barcelona (Spain), July 2003.

[28] Eric Clavier, Gérald Masini, Mathieu Delalandre, Maurizio Rigamonti, Karl Tombre, and Joël Gardes. Docmining: A cooperative platform for heterogeneous document interpretation according to user-defined scenarios. In *Graphics Recognition, Recent Advances and Perspectives, 5th International Workshop, GREC'03*, LNCS, pages 13–24, Barcelona (Spain), 2004. Springer-Verlag.

[29] Pdf-File Converter. http://www.pdf-file.com.

[30] Ross Cutler, Yong Rui, Anoop Gupta, JJ Cadiz, Ivan Tashev, Li wei He, Alex Colburn, Zhengyou Zhang, Zicheng Liu, and Steve Silverberg. Distributed meetings: a meeting capture and broadcasting system. In *MULTIMEDIA '02: Proceedings of the tenth ACM*

international conference on Multimedia, pages 503–512, New York, NY, USA, 2002. ACM Press.

[31] Berna Erol and Ying Li. An overview of technologies for e-meeting and e-lecture. In *Multimedia and Expo, 2005. ICME 2005. IEEE International Conference on*, page 6pp., 6-8 July 2005.

[32] Florian Evéquoz, Maurizio Rigamonti, Denis Lalanne, and Rolf Ingold. Document inquisitor : un système de validation des structures et d'élicitation de modèles de documents. In *Conférence Internationale sur le Document Electronique (CIDE'06)*, pages 79–95, Fribourg (Switzerland), September 2006.

[33] Thomas M. J. Fruchterman and Edward M. Reingold. Graph drawing by force-directed placement. *Softw. Pract. Exper.*, 21(11):1129–1164, 1991.

[34] George W. Furnas. Generalized fisheye views. *SIGCHI Bull.*, 17(4):16–23, 1986.

[35] Robert P. Futrelle, Mingyan Shao, Chris Cieslik, and Andrea Elaina Grimes. Extraction, layout analysis and classification of diagrams in pdf documents. In *Seventh International Conference on Document Analysis and Recognition (ICDAR'03)*, pages 1007–1012, Edinburgh (Scotland), 2003.

[36] Google. http://www.google.com.

[37] Rudinei Goularte, José Antonio Camacho-Guerrero, Valter R. Inacio Jr., Renan G. Cattelan, and Maria da Graça Campos Pimentel. M4note: a multimodal tool for multimedia annotations. In *WebMedia and LA-Web, 2004. Proceedings*, pages 142–149, 2004.

[38] Karim Hadjar. *Une étude de l'évolutivité des modèles pour la reconnaissance de documents arabes dans un contexte interactif.* PhD thesis, University of Fribourg, Switzerland, 2006. thesis Nr. 1513.

[39] Karim Hadjar, Oliver Hitz, Lyse Robadey, and Rolf Ingold. Configuration recognition model for complex reverse engineering methods: 2(crem). In *5th International Workshop on Document Analysis Systems (DAS'02)*, pages 469–479, New Jersey (USA), 2002.

[40] Karim Hadjar and Rolf Ingold. Arabic newspaper page segmentation. In *Document Analysis and Recognition, 2003. Proceedings. Seventh International Conference on*, pages 895–899, 3-6 Aug. 2003.

[41] Karim Hadjar, Maurizio Rigamonti, Denis Lalanne, and Rolf Ingold. Xed: A new tool for extracting hidden structures from electronic documents. In *DIAL'04: Proceedings of the First International Workshop on Document Image Analysis for Libraries (DIAL'04)*, page 212, Washington, DC, USA, 2004. IEEE Computer Society.

[42] Matthew R. B. Hardy, David Brailsford, and Peter L. Thomas. Creating structured pdf files using xml templates. In *ACM Symposium on Document Engineering (DocEng'04)*, pages 99–108, Milwaukee (USA), 2004.

[43] Susan Havre, Elizabeth Hetzler, Paul Whitney, and Lucy Nowell. Themeriver: visualizing thematic changes in large document collections. *Visualization and Computer Graphics, IEEE Transactions on*, 8(1):9–20, Jan.-March 2002.

[44] Marti A. Hearst. Tilebars: visualization of term distribution information in full text information access. In *CHI '95: Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 59–66, New York, NY, USA, 1995. ACM Press/Addison-Wesley Publishing Co.

[45] Oliver Hitz. *A Framework for Interactive Document Recognition.* PhD thesis, University of Fribourg, Switzerland, 2005. thesis Nr. 1488.

[46] Patrick Hoffman, Georges Grinstein, Kenneth Marx, Ivo Grosse, and Eugene Stanley. Dna visual and analytic data mining. In *Proceedings of Visualization '97*, pages 437–441,572, 19-24 Oct. 1997.

[47] Wolfgang Hürst. Indexing, searching, and skimming of multimedia documents containing recorded lectures and live presentations. In *MULTIMEDIA '03: Proceedings of the eleventh ACM international conference on Multimedia*, pages 450–451, New York, NY, USA, 2003. ACM Press.

[48] Wolfgang Hürst and Georg Götz. Interface issues for interactive navigation and browsing of recorded lectures and presentations. In *Proceedings of ED-MEDIA 2004*, Lugano, Switzerland, June 2004.

[49] Wolfgang Hürst, Gabriela Maass, Rainer Müller, and Thomas Ottmann. The "authoring on the fly" system for automatic presentation recording. In *CHI '01: CHI '01 extended abstracts on Human factors in computing systems*, pages 5–6, New York, NY, USA, 2001. ACM Press.

[50] Adobe Systems Incorporated. Acrobat reader. http://www.adobe.com/fr/products/acrobat.

[51] Adobe Systems Incorporated. Pdf reference. http://partners.adobe.com/asn/tech/pdf/specifications.jsp.

[52] Indico. Integrated digital conference. http://indico.sissa.it.

[53] Paul Janecek and Pearl Pu. An evaluation of semantic fisheye views for opportunistic search in an annotated image collection. *Journal of Digital Libraries,Special Issue on Information Visualization Interfaces for Retrieval and Analysis*, 1(5):42–56, 2005.

[54] Java. http://java.sun.com.

[55] JFerret. http://www.idiap.ch/mmm/tools/jferret.

[56] JPEDAL. http://www.jpedal.org.

[57] Susanne Jul and George W. Furnas. Navigation in electronic worlds: a chi 97 workshop. *SIGCHI Bull.*, 29(4):44–49, 1997.

[58] Peter Klein, Frank Muller, Harald Reiterer, and Maximilian Eibl. Visual information retrieval with the supertable + scatterplot. In *Proceedings of the Sixth International Conference on Information Visualisation*, volume 00, pages 70–75, Los Alamitos, CA, USA, 2002. IEEE Computer Society.

[59] Jon M. Kleinberg. Authoritative sources in a hyperlinked environment. *J. ACM*, 46(5):604–632, 1999.

[60] Jan Kuper, Horacio Saggion, Hamis Cunningham, Thierry Declerck, Franciska de Jong, Dennis Reidsma, Yorick Wilks, and Peter Wittenburgh. Intelligent multimedia indexing and retrieval through multi-source information extraction and merging. In *International Join Conference on Artificial Intelligence, IJCAI*, pages 409–414, 2003.

[61] Denis Lalanne, Florian Evéquoz, Maurizio Rigamonti, Bruno Dumas, and Rolf Ingold. An ego-centric and tangible approach to meeting indexing and browsing. In *4th Joint Workshop on Multimodal Interaction and Related Machine Learning Algorithms (MLMI'07)*, page to appear, Brno (Czech Republic), 2007.

[62] Denis Lalanne, Rolf Ingold, Didier Von Rotz, Ardhendu Behera, Dalila Mekhaldi, and Andrei Popescu-Belis. Using static documents as structured and thematic interfaces to multimedia meeting archives. In *Machine Learning for Multimodal Interaction MLMI'04*, number 3361 in LNCS, pages 87–100. Springer-Verlag, 2004.

[63] Denis Lalanne, Stéphane Sire, Rolf Ingold, Ardhendu Behera, Dalila Mekhaldi, and Didier Von Rotz. A research agenda for assessing the utility of document annotations in multimedia databases of meeting recordings. In *3rd International Workshop on Multimedia Data and Document Engineering, in conjunction with VLDB-2003*, pages 47–55, 2003.

[64] Dar-Shyang Lee, Berna Erol, Jamey Graham, Jonathan J. Hull, and Norihiko Murata. Portable meeting recorder. In *MULTIMEDIA '02: Proceedings of the tenth ACM international conference on Multimedia*, pages 493–502, New York, NY, USA, 2002. ACM Press.

[65] LinkedIn. https://www.linkedin.com.

[66] Agnes Lisowska, Martin Rajman, and Trung H. Bui. Archivus: A system for accessing the content of recorded multimodal meetings. In *Machine Learning for Multimodal Interaction MLMI'04*, pages 291–304. Springer Berlin / Heidelberg, 2005.

[67] Simone Marinai and Andreas Dengel, editors. *Document Analysis Systems VI*, volume 3163 of *Lecture Notes in Computer Science*. Springer Berlin / Heidelberg, September 2004.

[68] Dalila Mekhaldi. *A Study on Multimodal Document Alignment: Bridging the Gap Between Textual Documents and Spoken Language*. PhD thesis, University of Fribourg, Switzerland, 2006. thesis Nr. 1521.

[69] Rainer Müller and Thomas Ottmann. The "authoring on the fly" system for automated recording and replay of (tele)presentations. *Multimedia Systems*, Volume 8(Number 3):158–176, October 2000.

[70] Nicolas Moënne-Loccoz, Bruno Janvier, Stéphane Marchand-Maillet, and Eric Bruno. Managing video collections at large. In *CVDB '04: Proceedings of the 1st international workshop on Computer vision meets databases*, pages 59–66, New York, NY, USA, 2004. ACM Press.

[71] Sugata Mukhopadhyay and Brian Smith. Passive capture and structuring of lectures. In *Seventh ACM international conference on Multimedia*, pages 477–487, USA, 1999.

[72] MySQL. http://www.mysql.com.

[73] Steven Noel, Chee-Hung H. Chu, and Vijay Raghavan. Visualization of document co-citation counts. In *In proceedings of Sixth International Conference on Information Visualisation*, pages 691– 696, 2002.

[74] Mohammad D. Paknad and Robert M. Ayers. Method and apparatus for identifying words described in a portable electronic document. U.S. Patent 5,832,530, 1998.

[75] W. Bradford Paley. Textarc: Showing word frequency and distribution in text. In *Proceedings of the IEEE Symposium on Information Visualization(Infovis '02) Poster Compendium*, Los Alamitos, CA, USA, 2002. IEEE Press.

[76] PDF2Office. http://www.recosoft.com.

[77] PDFTron. http://www.pdftron.com.

[78] Andrei Popescu-Belis and Paula Estrella. Generating usable formats for metadata and annotations in a large meeting corpus. In *ACL 2007 (45th International Conference of the Association for Computational Linguistics) Companion Volume*, page to appear, Prague, Czech Republic, June 2007.

[79] Fuad Rahman and Hassan Alam. Conversion of pdf documents into html: a case study of document image analysis. In *Conference Record of the Thirty-Seventh Asilomar Conference on Signals, Systems and Computers 2003*, pages 87–91, USA, 2003.

[80] Maurizio Rigamonti, Jean-Luc Bloechle, Karim Hadjar, Denis Lalanne, and Rolf Ingold. Towards a canonical and structured representation of pdf documents through reverse engineering. In *Document Analysis and Recognition, 2005. Proceedings. Eighth International Conference on*, volume 2, pages 1050–1054, 29 Aug.-1 Sept. 2005.

[81] Maurizio Rigamonti, Karim Hadjar, Denis Lalanne, and Rolf Ingold. Xed: un outil pour l'extraction et l'analyse de documents pdf. In *Huitième Colloque International Francophone sur l'Ecrit et le Document (CIFED'04)*, pages 85–90, La Rochelle (France), June 2004.

[82] Maurizio Rigamonti, Oliver Hitz, and Rolf Ingold. A framework for cooperative and interactive analisys of technical documents. In *5th IAPR International Workshop on Graphics Recognition (GREC'03)*, pages 407–414, Barcelona (Spain), July 2003.

[83] Maurizio Rigamonti, Denis Lalanne, Florian Evéquoz, and Rolf Ingold. Browsing multimedia archives through intra- and multimodal cross-documents links. In *Machine Learning for Multimodal Interaction MLMI'05*, number 3869 in LNCS, pages 114–125, Edinburgh (UK), July 2006. Springer-Verlag.

[84] Maurizio Rigamonti, Denis Lalanne, and Rolf Ingold. Faericworld: Browsing multimedia events through static documents and links. In *In proc. of INTERACT 2007*, LNCS, pages 102–115, Rio De Janeiro, Brasil, September 2007. Springer-Verlag.

[85] Lyse Robadey. *2(CREM) : Une méthode de reconnaissance structurelle de documents complexes basée sur des patterns bidimensionnels.* PhD thesis, University of Fribourg, Switzerland, 2001.

[86] Ivica Rogina and Thomas Schaaf. Lecture and presentation tracking in an intelligent meeting room. In *Multimodal Interfaces, 2002. Proceedings. Fourth IEEE International Conference on*, pages 47–52, 14-16 Oct. 2002.

[87] Lawrence A. Rowe, Diane Harley, Peter Pletcher, and Shannon Lawrence. Bibs: A lecture webcasting system. Technical report, Berkeley Multimedia Research Center, 2001.

[88] Ben Shneiderman. The eyes have it: A task by data type taxonomy for information visualizations. In *VL '96: Proceedings of the 1996 IEEE Symposium on Visual Languages*, pages 336–343, Washington, DC, USA, 1996. IEEE Computer Society.

[89] Stéphane Sire and Denis Lalanne. Smart meeting minutes application specification. Technical report, University of Fribourg, August 2002. version 2.4.

[90] SMIL. Synchronized multimedia integration language. http://www.w3.org/AudioVideo.

[91] John R. Smith, Milind Naphade, and Apostol (Paul) Natsev. Multimedia semantic indexing using model vectors. In *Multimedia and Expo, 2003. ICME '03. Proceedings. 2003 International Conference on*, volume 2, pages 445–448, 6-9 July 2003.

[92] Micheal J. Swain. Searching for multimedia on the world wide web. In *Multimedia Computing and Systems, 1999. IEEE International Conference on*, volume 1, pages 32–37, 7-11 June 1999.

[93] Holger Theisel and Matthias Kreuseler. An enhanced spring model for information visualization. In *Eurographics 98*, volume 17, pages 335–344. Blackwell Publishing, 1998.

[94] Simon Tucker and Steve Whittaker. Accessing multimodal meeting data: Systems, problems and possibilities. In *Machine Learning for Multimodal Interaction MLMIŠ04*, number 3361 in LNCS, pages 1–11. Springer-Verlag, 2004.

[95] Simon Tucker and Steve Whittaker. Reviewing multimedia meeting records: Current approaches. In *Multimodal multiparty meeting processing workshop, ICMI 2005, International Conference on Multimodal Interfaces*, 2005.

[96] Didier von Rotz, David Bourillot, Omar Abou Khaled, Rudolf Scheurer, Denis Lalanne, Rolf Ingold, Jean-Yves LeMeur, and Thomas Baron. Smac - smart multimedia archive for conferences. *Flash informatique, EPFL*, (1):3–10, 2006.

[97] Alex Waibel, Michael Bett, Michael Finke, and Rainer Stiefelhagen. Meeting browser: Tracking and summarizing meetings. In D. E. M. Penrose, editor, *Proceedings of the Broadcast News Transcription and Understanding Workshop*, pages 281–286, Lansdowne, Virginia, February 1998. Morgan Kaufmann.

[98] Alex Waibel, Tanja Schultz, Michael Bett, Matthias Denecke, Robert Malkin, Ivica Rogina, Rainer Stiefelhagen, and Jie Yang. Smart: the smart meeting room task at isl. In *Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP '03). 2003 IEEE International Conference on*, volume 4, pages 752–755, 6-10 April 2003.

[99] Edward J. Wegman and Qiang Luo. High dimensional clustering using parallel coordinates and the grand tour. In *Computing Science and Statistics, Proceedings of the Twenty-Eighth Symposium on the Interface*, pages pp. 361–368, 1997.

[100] Pierre Wellner and Mike Flynn. Report on the evaluation of meeting browsers. Technical report, IDIAP, 2006.

[101] Pierre Wellner, Mike Flynn, and Maël Guillemot. Browsing recorded meetings with ferret. In *Machine Learning for Multimodal Interaction MLMI'04*, number 3361 in LNCS, pages 12–21. Springer Berlin / Heidelberg, 2004.

[102] Pierre Wellner, Mike Flynn, Simon Tucker, and Steve Whittaker. A meeting browser evaluation test. In *CHI '05: CHI '05 extended abstracts on Human factors in computing systems*, pages 2021–2024, New York, NY, USA, 2005. ACM Press.

[103] Wai Yee Peter Wong and Dik Lun Lee. Implementations of partial document ranking using inverted files. *Inf. Process. Manage.*, 29(5):647–669, 1993.

[104] WordNet. http://wordnet.princeton.edu/.

[105] XML. Extensible markup language (xml) 1.0 (4th ed.). http://www.w3.org/TR/xml.

[106] XPDF. http://www.foolabs.com/xpdf/home.html.

[107] XSLT. Xsl transformations (xslt) version 1.0. http://www.w3.org/TR/xslt.

[108] Ka-Ping Yee, Kirsten Swearingen, Kevin Li, and Marti Hearst. Faceted metadata for image search and browsing. In *CHI '03: Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 401–408, New York, NY, USA, 2003. ACM Press.

[109] YouTube. http://www.youtube.com.

Curriculum Vitae

Maurizio Rigamonti

Personal Information

Languages: Italian (mother tongue), French, German, and English

Birth date: November 4, 1977

Nationality: Swiss

Education

2002-2008: PhD Student and teaching assistant, at Department of Informatics, University of Fribourg, Switzerland

1997-2002: Graduate studies in computer science, at Department of Informatics, University of Fribourg, Switzerland

1992-1997: Literary maturity (B type, with Latin), at Lyceum of Lugano 1, Switzerland

Professional Experiences

Projects: DocMining, Xed, JFriDoc, and FaericWorld

Publications: 14 papers in HCI and document analysis, 5 posters

Talks: 6 international talks, 7 public presentations

Teaching: 2 courses as teaching assistant, 2 WINS stages created, 2 courses as undergraduate teaching assistant

Webmaster: DIVA group (2003-2005), Website of Diploma Awarding Ceremony, Faculty of Sciences (2005), SDN conference (2006)

Additional Experiences

Associations: Fribot (3 years of committee, involved in logistics, writings, and sponsoring)

Army: Infantry sergeant (qualified to graduate as lieutenant)

Sport: Soccer (FC Matran)

Research Interests

- Information visualization
- Document analysis and engineering
- Multimedia data analysis
- Indexing and information retrieval

Publications

Maurizio Rigamonti, Denis Lalanne, and Rolf Ingold. Faericworld: Browsing multimedia events through static documents and links. In *In proc. of INTERACT 2007*, LNCS, pages 102–115, Rio De Janeiro, Brasil, September 2007. Springer-Verlag.

Denis Lalanne, Florian Evéquoz, Maurizio Rigamonti, Bruno Dumas, and Rolf Ingold. An ego-centric and tangible approach to meeting indexing and browsing. In *4th Joint Workshop on Multimodal Interaction and Related Machine Learning Algorithms (MLMI'07)*, page to appear, Brno (Czech Republic), 2007.

Florian Evéquoz, Maurizio Rigamonti, Denis Lalanne, and Rolf Ingold. Document inquisitor : un système de validation des structures et d'élicitation de modèles de documents. In *Conférence Internationale sur le Document Electronique (CIDE'06)*, pages 79–95, Fribourg (Switzerland), September 2006.

Jean-Luc Bloechle, Maurizio Rigamonti, Denis Lalanne, and Rolf Ingold. Xcdf : un format canonique pour la représentation de documents. In *Colloque International Francophone sur l'Ecrit et le Document (CIFED'06)*, pages 19–23, Fribourg (Switzerland), September 2006.

Maurizio Rigamonti, Denis Lalanne, Florian Evéquoz, and Rolf Ingold. Browsing multimedia archives through intra- and multimodal cross-documents links. In *Machine Learning for Multimodal Interaction MLMI'05*, number 3869 in LNCS, pages 114–125, Edinburgh (UK), July 2006. Springer-Verlag.

Jean-Luc Bloechle, Maurizio Rigamonti, Karim Hadjar, Denis Lalanne, and Rolf Ingold. Xcdf: A canonical and structured document format. In *7th International Workshop, DAS'06*, number 3872 in LNCS, pages 141–152, Nelson (New Zealand), February 2006. Springer-Verlag.

Maurizio Rigamonti, Jean-Luc Bloechle, Karim Hadjar, Denis Lalanne, and Rolf Ingold. Towards a canonical and structured representation of pdf documents through reverse engineering.

In *Document Analysis and Recognition, 2005. Proceedings. Eighth International Conference on*, volume 2, pages 1050–1054, 29 Aug.-1 Sept. 2005.

Eric Clavier, Gérald Masini, Mathieu Delalandre, Maurizio Rigamonti, Karl Tombre, and J. Gardes. Docmining: A cooperative platform for heterogeneous document interpretation according to user-defined scenarios. In *5th IAPR International Workshop on Graphics Recognition (GREC'03)*, Revised selected papers, number 3080 in LNCS, pages 13–24, Barcelona (Spain), 2004.

Sébastien Adam, Maurizio Rigamonti, Eric Clavier, Eric Trupin, Jean-Marc Ogier, Karl Tombre, and Joël Gardes. Docmining: A document analysis system builder. In *Document Analysis Systems VI, 6th International Workshop, DAS 2004*, number 3163 in LNCS, pages 472–483, Florence (Italy), September 2004. Springer-Verlag.

Maurizio Rigamonti, Karim Hadjar, Denis Lalanne, and Rolf Ingold. Xed: un outil pour l'extraction et l'analyse de documents pdf. In *Huitième Colloque International Francophone sur l'Ecrit et le Document (CIFED'04)*, pages 85–90, La Rochelle (France), June 2004.

Eric Clavier, Sébastien Adam, Pierre Héroux, Maurizio Rigamonti, and Jean-Marc Ogier. Docmining : Une plate-forme de conception de systèmes d'analyse de documents. In *Huitième Colloque International Francophone sur l'Ecrit et le Document (CIFED'04)*, pages 97–102, La Rochelle (France), June 2004.

Karim Hadjar, Maurizio Rigamonti, Denis Lalanne, and Rolf Ingold. Xed: A new tool for extracting hidden structures from electronic documents. In *DIAL'04: Proceedings of the First International Workshop on Document Image Analysis for Libraries (DIAL'04)*, page 212, Washington, DC, USA, 2004. IEEE Computer Society.

Maurizio Rigamonti, Oliver Hitz, and Rolf Ingold. A framework for cooperative and interactive analisys of technical documents. In *5th IAPR International Workshop on Graphics Recognition (GREC'03)*, pages 407–414, Barcelona (Spain), July 2003.

Eric Clavier, Gérald Masini, Mathieu Delalandre, Maurizio Rigamonti, Karl Tombre, and J. Gardes. Docmining: A cooperative platform for heterogeneous document interpretation according to user-defined scenarios. In *5th IAPR International Workshop on Graphics Recognition (GREC'03)*, pages 21–32, Barcelona (Spain), July 2003.