CIDE 11, ARMARIUS- A Living Online Archive for Ancient Manuscripts

Reim DOUMAT[1] , Elöd EGYED-ZSIGMOND[1] , Emese CSISZÁR[2] , Jean-Marie PINON[1]

[1]Laboratoire LIRIS, INSA de Lyon, Villeurbanne, France
{reim.doumat,elod.egyed-zsigmond, jean-marie.pinon}@liris.cnrs.fr

[2]Sapientia EMTE, Tîrgu-Mres, Romania. csiszarmeso@yahoo.com

Abstract. Many museums and libraries digitize their collections of historical manuscripts, to preserve the historic documents and to make them public. The collections are available in image format and they need annotation to be accessible and exploitable. The annotations can be created manually, automatically or semi-automatically. The problem with the manual annotation is that they are expensive and tedious. Hence the reuse of users' experiences, by tracing their actions during the annotation process, helps other users to accomplish repetitive tasks in a semi-automatic manner, and assists difficult tasks. In this article we present a digital archive model and prototype of a collaborative system for the management of online ancient manuscript. The application offers an online annotation service for this type of documents, an assistant for a semi-automatic annotation, and a tracing system that saves traces of important actions in order to reuse them later in a recommender system.

Keywords. Living digital archive, manuscript annotation, assistant, tracing system.

Résumé. Plusieurs musés et bibliothèques numérisent leur collections de manuscrits historiques pour les conserver et les rendre publiques. Les collections sont disponibles en format image et ont besoin d'annotations pour être accessibles et exploitables. La création des annotations peut être manuelle, automatique ou assistée. Le problème avec l'annotation manuelle qu'il est chère et fastidieuse, donc la réutilisation de l'expérience de l'utilisateur, en se basant sur des expériences tracées, permet d'en aider d'autres à réaliser des tâches répétitives de manière semi-automatique, ou d'effectuer des tâches non triviales de manière assistée. Dans cet article nous présentons une archive numérique de manuscrits anciens en ligne. Cette application offre un service d'annotation, un système de traçage gardant les traces de certaines actions et un système d'assistance qui exploite ces traces.

Mots-clés. Archive numérique vivante, annotation des manuscrits, système d'assistance, système de traçage.

1 Introduction

Many museums and libraries digitize their collections of historic manuscripts to protect these precious documents and to make them accessible to a large public. These collections are available online in image format and they need annotations to be accessible and exploitable.

Actually, the consultation of collections on the internet is increasing progressively, because it meets the various needs of all user types, and because it offers users with services to search rapidly the information, to mark their favourite pages and to personalize their environment (Vivarium the online digital collections of Saint John's university and the College of Saint Benedict, ContentDM collection). However, these operations might be considered as non creative operations since users do not work directly on documents. The interfaces do not allow the easy communication and publication of ideas, comments, and interpretations. The importance of the annotations according to (Bottoni and al. 2004) is that they form a support to the intellectual activities, like: a highlight of interesting parties of a text, an indication to the user reflection and an enhancement of the document with new information.

Consequently, users need to annotate documents online independently from their media type (images, audios, videos, web pages, etc.). Annotations represent primordial actions that offer to users the possibility to react directly on their documents in order to enrich them. Additionally, every annotation made by the user can generate a trace in the system in order to be reused lately. This could be beneficial for all persons who do not know the domain or for those who miss the experience. According to (Egyed-Zsigmond and al. 2003), the reuse of user's experience during the annotation process permits other users to realize repetitive tasks in a semi automatic manner, or to realize difficult task in an assisted way.

In this paper we present an online archive application to manage and annotate ancient manuscripts. We incorporate within this archive some image treatment tools and web services to annotate remotely these manuscripts. Our application is enriched with an experience capitalization layer that traces the important actions, and then it integrates traces in an assistant system to help users during the annotation procedure.

The article is organized as follows: in the next section, we expose the state of the art about some of the popular annotating systems. In section 3 we present our project, called ARMARIUS and emphasise on annotating the manuscripts online by different types of users, and then we illustrate a prototype of our web application. At the end we conclude and give some perspectives.

2 Related works

Digital annotations that are attached to digital collections represent two elements: metadata and content. The first is a group of attributes like (author, title, creation date, modification date,…) that could be defined by a standard (Dublin Core, Marc, MODS, TEI…) or by the environment of the annotation. The second element is the content that is created by users and is composed of textual information, images, hyperlinks, etc. Annotations vary depending on the system and the context where they are used. Many projects are interested in the annotations, in this section we refer to some of them and compare their characteristics.

2.1 Document annotation projects

Many annotation projects have developed diverse tools to annotate web pages, multimedia objects, or documents, the objectives of these projects varied between: creating repositories with web services that are adaptable to comprise different types of collections to form digital libraries like Fedora, offering image mark-up tool like the project UVic, integrating plug-ins in the web browsers to provide annotation tools such as Annotea (Kahan and Koivunen, 2001) that permits to exchange web annotations and bookmarks between users, TafAnnote (Cabanac and al., 2007), and MADCOW (Multimedia Annotation of Digital Content Over the Web) (Bottoni and al., 2004) for multimedia annotation over the web.

Some of the previous systems have a collaborative environment that permits different users to work on a group of documents and to share their knowledge, as mentioned in Table 1. The table summarizes the differences between the characteristics of these projects.

System Feature	**Fedora**	**UVic**	**MADCOW**	**Annotea**	**TafAnnote**
Document type	Digital collections	Images	Web pages, multimedia objects	Web content	Web content
Annotation type	Defined by the digital library	Keywords, comments	Many types of comments	Notes, explanations, bookmarks	Comments (discussion)
Collaboration work	Between systems	No	Yes	Yes	Yes
Recommender system	No	No	No	No	No
Type of the application	Web application	Standalone application and web based viewer	Plug-in client in standard web browser	Plug-in client, proxy	Plug-in client in Mozilla Firefox web browser

Table 1. *Comparison between annotating projects*

The main disadvantage of these systems is that they handle XML documents while it is not able with images of ancients' manuscripts. In Annotea the web pages and their contents of objects (images, texts, hyper links, etc.) are identified by URLs while scanned images of the manuscripts are identified by IDs. Annotea, TafAnnote and MADCOW permit the information exchange between user groups; this service enhances the collaboration work in order to facilitate the realization of difficult user tasks. Fedora does not enhance the information exchange between users.

Other projects interested in the annotation of the ancient manuscripts like: Bambi (Calabretto and al., 1998) which is an ancient project to annotate manuscripts on a local machine; users can work in collaboration but on the same computer. Other systems are web applications that could be used to visualize and to annotate documents remotely as IPSA (Agosti and al., 2003), Scraps (documents from the World War I) offers the access to rare books online. Annotations in Debora are extracted by image treatment tools and are classified in three levels (description, structure and contents), while IPSA works on manual image annotation of Herbal manuscripts. Table 2 summarizes the differences between these projects.

System Feature	Bambi	Debora	Scraps	IPSA
Document type	Images	Images	Images	Images
Annotation type	Manual	Automatic extraction, predetermined	Predetermined	Manual (Textual and linking annotations)
Collaboration work	Yes	No	No	Yes
Recommender system	No	No	No	No
Type of the application	Standalone application	Standalone application	Partially web application	Web application

Table 2. *Comparison between manuscripts annotating projects*

The listed systems do not contain assisting tools to facilitate the manuscripts annotation and the use of other services, or collaborative recommender tools to assist users in realizing difficult tasks.

2.2 Tracing and recommender systems projects

Tracing system registers important events and actions made by users while using the application, traces are used to build users experiences like (Hilbert and Redmiles, 2000), Trèfle (Egyed-Zsigmond and all., 2003) that generate an assistant system from experienced user actions, and (eMédiathèque) which is a collaborative platform for virtual classroom developed by eLycée, it includes a tracing infrastructure with a collaborative tools to help users in remote learning. Traces are also used to build intelligent applications such as recommender systems, which assist and give advice to users during his interaction with the application (Champin, 2003). Some recommender systems base on the user profile and his history to determine the interesting documents or web pages of each user, such as Personal Web Watcher (Mladenic, 1999), ITR recommender system (Semeraro and al., 2007).

2.3 State of the art conclusion

We can notice that not all of these systems are capable to organize and annotate remotely images of ancient manuscripts; stand alone applications are not useful if user groups need to work in collaboration to annotate the images. Other online projects do not offer precise annotation and collaborative space to facilitate the communication between different users (i.e. confrontations of points of views, correction to annotation done by other users). Furthermore, some projects (Annotea, MADCOW) while they have collaborative functionalities, they enable to annotate only text based web pages and not images or image fragments. Other projects concern the visualization of the rare scanned documents; they do not allow users to add annotations. All these applications do not contain recommender systems; we think that they are important to users who perform difficult tasks. Recommender systems can be developed basing on the traces of user actions. We search to annotate images of manuscripts or fragments of them, by using a web application accessible by web browsers and providing services to annotate manually and semi-automatically the manuscripts images.

In the next section, we describe a model of online archive to manage digitized documents of ancient manuscripts. This model can handle also annotations, users, and their access rights, interactions, and preferences. Our model contains a tracing layer, an assisting system and a collaborative system that permits professional and expert users to work in groups in order to complete difficult tasks.

3 Online archive for ancient manuscripts (ARMARIUS)

ARMARIUS is an online document management system, which offers a framework for a living (dynamic) archive. We aim in this framework to manage and to put online collections of ancient manuscripts, and to provide remote and collaborative annotation tools. We have several test collections: the manuscripts of a mathematician from the XIX century, other collections contain Arabic ancient manuscripts that are found in Timbuktu and some Syrian manuscripts (in Arabic and in Syriac). These documents do not have an organization; therefore it is important that users can organize the pages according to their requirements.

3.1 ARMARIUS modules

The architecture of the ARMARIUS is shown in Figure 1. The system is composed of the following modules:

- *Collections of scanned documents*: digitized images of the manuscripts structured in collections/ sub-collections depending on different factors (date, theme, etc.), the images of manuscript pages are of different forms (JPEG, PNG, TIFF…) and stored in a rational database. Each image has three versions: *thumbnail* for a low resolution, *access* for an intermediate resolution, and *real* for high resolution. The advantage of this system is to provide users with *thumbnails* when they ask for an image review, and with an intermediate resolution of the image *access* when the user does not choose the image version. Each page image may contain many *document units*. Document units represent image fragments, whole images or collections. An image fragment *document unit* can be defined by the user and has coordinates linking it to its original image. However, these coordinates change in correspondence to the image size and keep the document unit in the same place in all image versions.
- *Annotations*: many types of annotations are defined (keywords, comments, transcriptions, digital signatures, administrative or descriptive metadata) with a possibility to add other types dynamically; annotations are created by users and associated to document units. We plan to add OAI-PMH and other metadata standard (Dublin Core, TEI P5, METS…) compliant annotation import/export.
- *Application*: a Web application that is accessible through web browsers
- *Web services*: we implemented a web service based image processing tool architecture. An identified user with sufficient rights can initiate an image processing treatment on a collection. An image processing treatment starts a session and lets the user to go on with his work. On the personal space interface of the *Application*, the user can consult his image processing sessions and validate

the results of the finished ones (e.g. word-spotting). In this way the system can carry out long lasting treatments.

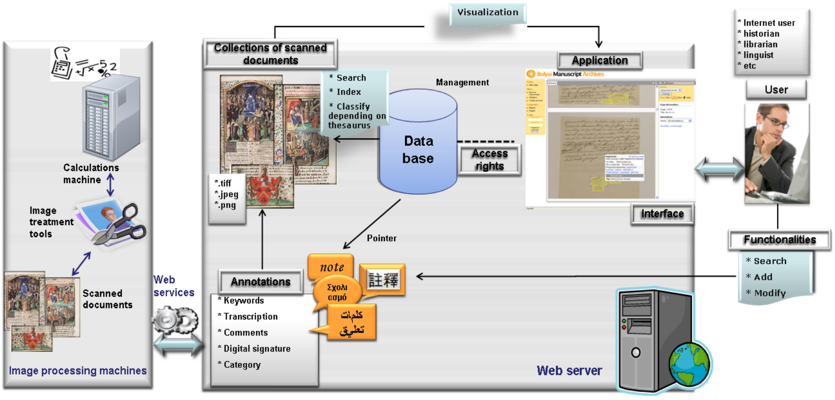


Figure 1. *General view of ARMARIUS*

- *Functionalities:* Online services (research, visualization, annotation, manual transcription, adding comments…). Users have to identify themselves to access manuscript images or digitized collections, and to search images depending on their annotations, transcriptions or other metadata. Users can also annotate manually the documents with new keywords, transcriptions or comments enriching the documents with additional information.
- *Users:* Users in ARMARIUS are classified into three categories: non-identified users (like internet users) who can only see a demo selected by the administrator about the collections, registered users belong to groups, and the administrators who manage the system and upload images.
- *Database*: contains image collection information, metadata, users, users groups, and access rights.

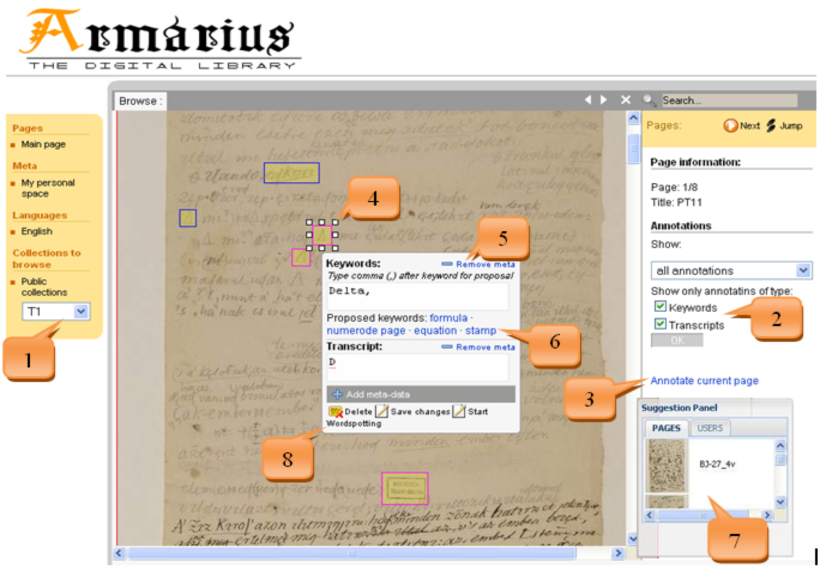
Access rights concern collections and their content, annotations and user groups. They are defined by the system administrator. Access rights define view and modification rights.

ARMARIUS registered users can create a personal space, which contains collections or pages chosen by the user. This space provides the user with all the functionalities that he needs to accomplish his work. Personal collections reflect a topic of user interests; user may organize his/her own defined collections into certain categories.

Some users' actions are registered in the system *as traces*. These traces are integrated in an assisting system in order to help other users during the search and the annotation. We created a task and object model of the application and chose from these models the tasks to be traces and the objects they modify.

3.2 ARMARIUS Functionalities

ARMARIUS application permits users to annotate remotely images and image fragments, to define objects (collections, document units, pages, keywords or other users and groups). Image annotation is done by users in a collaborative environment that permits users to work together. The collaborative system provides users with tools to see the work of other users, and to add comments on the document units of other persons or on their own work. Users can also modify the annotations of other document units if they have the permission in their groups. ARMARIUS offers also a recommender system based on users experiences.



Page information:

Page: 1/8
Title: PT11

Keywords: Delta,

Proposed keywords: formula · numerode page · equation · stamp

Transcript: D

Select a collection; 2- filter the annotation/transcription; 3- create a new annotation on the current page; 4-draw a rectangle around a fragment; 5- add various metadata; 6- use the keywords that are suggested by the recommender system; 7- suggestion panel; 8-launch a session of Word Spotting with the selected fragment

Figure 2. *Annotation Screenshot in ARMARIUS*

3.2.1 Document and user management

First of all, ARMARIUS is an image management system. It enables to upload images, creates automatically different versions, and enables their classification into collections and sub collections and views. A view is composed of images from a given collection in a given order. An image can belong to several views but to only one collection. Users usually navigate through views.

Users belong to groups; a user can belong to several groups. The rights are defined between collections and groups. As an image belongs to one collection the rights are easy to be calculated. If a user is member of different groups which have different rights on a given image, the user rights are added. Annotations can be private, restricted to group members or public visible to anyone.

Each identified user has a *Personal space* on which he can select the collections to view or to annotate, set preferences, start a search, consult the image processing sessions, see favorites, manage personal views.

3.2.2 Annotating

Users can annotate new image fragments (document units) by creating a rectangle representing the document unit then adding annotations. The annotation is done via the web browser interface. Once the document unit has been created, a dialog box appears allowing user to add keywords and/or transcriptions or other metadata as shown in Figure 2.

The list of metadata types can be extended dynamically, and for each metadata type we can specify an export translation in order to be exported according to a given metadata standard syntax.

Another way to annotate documents in an assisted manner is the use of image processing services. Some of the image processing tools are implemented as asynchronous services. For example, the word spotting in ARMARIUS helps in finding the fragments that are similar to the fragment précised by the user within a collection. It is handled as sessions: a user can select a fragment of document and launches the word spotting session that could take hours to be finished. On the main page, the user has a list of current image processing sessions. A session can be in different states: launched, finished, validated. A finished session can be visualized: its results are shown and the user is asked to validate them. She can modify, delete or accept results one by one, by page, or for the whole collection. In Figure 3 we present the results of a word spotting session in ARMARIUS.

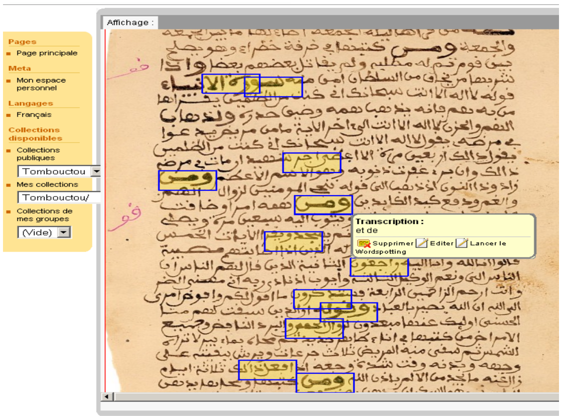


Figure 3. *The use of word spotting in ARMARIUS*

3.2.3 Recommender system for annotations

For better understanding of how a recommender system works: let us imagine the next scenario when a user (Anny) connects to ARMARIUS web application. If this is her first connection and she has no account in the database, she will be able to see just the demo collections proposed by the system administrator. If Anny is a regular user who has an ID, she will be able to search and browse the collections that are permitted to the groups she belongs to. After her login, the tracing system begins to register her actions (connect, search, browse, chose, create…), besides the objects that are affected by these actions.

The recommender system is based on a tracing layer that tracks the actions of identified users during their work session; traces are stored in a relational database together with the affected objects (collections, pages, metadata…). In order to create an experience based user assistance we have to go through different phases.

We consider that user manipulates objects through procedures thus the use of the system is traced according to Trèfle♣ model (Egyed-Zsigmond and al. 2003). For this tracing we need to formalize these procedures as well, so a user-task-model is built. Firstly we need to build an object model, which is composed of collections, pages, document units and metadata, a tree structure which holds the relations between different types of objects (Figure 4). The instantiation of this model gives us the actual structure of a collection.

The next step in constructing our recommender system was deciding which tasks to trace in order to create the observation model. These will be the tasks, which, together with the manipulated objects will create our experience and knowledge pool. Some user tasks like registration or signing in are not relevant in future recommendations, whilst other tasks, mainly those which manipulate objects in the collection's structure, will be the basis of the user assistance methods.

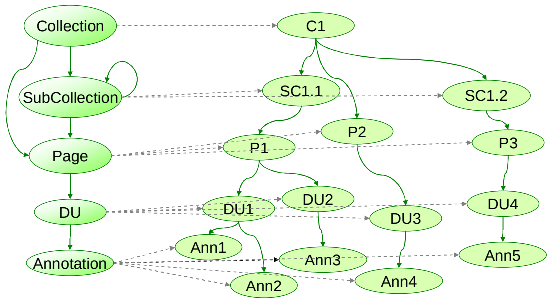


Figure 4. *Simplified object model and an instance fragment*

In order to be able to compare these user traces among themselves, we need to formalize the distances between objects, as well as the distances between different types of tasks. This process of comparison always involves two objects of similar or different types. In case of comparing two different types of objects or two metadata we rely only on the structure specified by the collection. The similarity of the two objects will be a number, equal with the distance between the two nodes in the collection's structure tree, representing the two objects. In case of similar object types, such as document units, pages and collections, besides these physical distances we also need to take in consideration the content similarity between them. That is, for example, in case of two document units, we not only take in consideration whether they are on the same page or in the same collection, but we analyse the metadata associated with them, and their similarity. Based on these we can put together a similarity measure for calculating content distances between two different document units:

$$DC_{du}(du1,du2)=\min\left(\sum_{\substack{mdi\in A1\\ mdj\in A2}} D_{md}(mdi,mdj)\cdot D_{mt}(mt_{mdi},mt_{mdj})\right)$$

where *du1, du2* are the two document units to compare, *A1* and *A2* are the metadata associated with them, D_{md} is the physical distance between two metadata, and D_{mt} tells us whether the two metadata are of the same type or not.
Analogically we can calculate distances between pages, based on the document units positioned on them.

$$DC_p(p1, p2) = \min\left(\sum_{\substack{dui \in B1 \\ duj \in B2}} DC_{du}(du_i, du_j)\right)$$

where *p1* and *p2* are the pages to be compared, *B1* and *B2* are the document units associated with the pages.

The user tracing itself is a process of determining the action of the user and inserting this together with the user identifier, the objects manipulated and other parameters such as sate, and session identifier into the system's database. Each type of task has different number and type of parameters.

Upon identification, the recommender system will exploit the stored traces, to provide step by step assistance to user actions. At different points of the navigation the system provides different types of recommendations, one of them relies only on the distances between objects and it is used to suggest similar pages and document units while browsing, by recommending the objects closest to the currently browsed page or document unit, to suggest similar metadata in case of document annotation, or to add the current collection into the user's personal space. For example if the user creates an annotation on a page, and adds the metadata "paragraph" to the document unit, and there is already a document unit annotated with the metadata "paragraph", "number" and "section" the system will recommend the words "number" and "section" to the user.

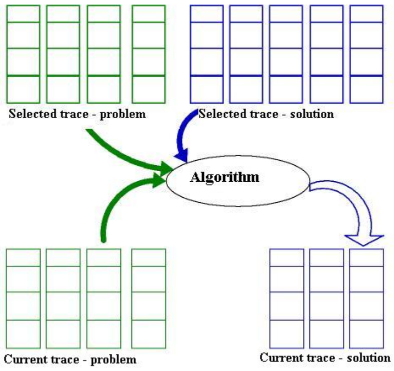


Figure 5. *Finding and reusing similar cases*

A more complex algorithm is used for other user assisting functionalities, such as recommending search results, or recommending similar pages based on user actions. During this procedure we cut the traces in reusable and adaptable episodes according to the case based reasoning paradigm. Each case is composed of two parts: problem and solution; the current trace of the user is our current problem.

The system searches for similar problems among these cases, and suggests the solution part of the case, or those objects that were manipulated by the tasks of this solution.

This similarity is calculated based on the distances of the tasks' parameters. For example: Anny performs a search with the keywords "stamp" and "round", she selects a page from the result set, views the suggested pages, those pages that the system finds similar to the current page. She navigates to one of these pages, and she selects a document unit on this page. From the list of similar document units she again navigates to a page holding one of the recommended document units. All of these tasks and objects are traced by the system, so when another user signs in and performs a search with the word "stamp", the recommender algorithm will extract all those pages that Anny had viewed earlier and show him.

4 Conclusion and future works

In this article, we have presented a "living" digital archive model of ancient manuscripts: ARMARIUS, with a web application prototype. Our proposed model could be also used in other domains (scientific, medical…). In this paper, we treated the following problems:

- How to represent the digitized document in a living archive? This concerns annotations creation, documents structuring, the storage in a database, and a model to access the documents. And the need for a user collaborative work space to create a discussion environment concerning the collections.
- User assistant integration, this assistant proposes different help types during the annotation, the document search, and the creation of a personal space.
- The assisting system, the collaborative system, and the discussion space are important to annotate the manuscripts. Especially that this type of document requests lot of explanations and image treatment tools are not very efficient.

In our future works, we aim at integrating technologies of type "push" and RSS to track the evolution of certain documents, themes, collections, etc. we aim also to offer and to assist the discussion space, hence users can confront straightforwardly their points of view about a document. We are also interested in developing a module that allows users to exchange messages between each others, to discuss the collections and their content. Finally, we are concerned to enrich the system with image treatment tools that are especially adapted to this type of manuscripts (other than word spotting).

5 References

« CONTENTdm Digital Collection Management Software by OCLC. » http://www.contentdm.com/ (Accessed at 5 may 2008).

«Dublin Core Metadata Element Set. » http://dublincore.org/documents/dces/ (Accessed at 5 may 2008).

«Fedora Commons- Home. » http://www.fedora-commons.org/ (Accessed at 13 may 2008).

«Le protocole OAI et ses usages en bibliothèque ». http://www.culture.gouv.fr/culture/dll/OAI-PMH.htm (Accessed 14 may 2008).

«MARC STANDARDS. » http://www.loc.gov/marc/index.html (Accessed at 13 may 2008).

«Metadata Object Description Schema: MODS (Library of Congress) ». http://www.loc.gov/standards/mods/ (Accessed at 13 may 2008).

«The UVic Image Markup Tool Project. » http://www.tapor.uvic.ca/%7Emholmes/image_markup/index.php (Accessed at 5 may 2008).

«Vivarium. » http://cdm.csbsju.edu/ (Accessed at 13 may 2008).

Agosti M., Benfante L., Orio N. (2003). "IPSA: A digital archive of herbals to support research." In *Digital Libraries: Technology and Management of Indigenous Knowledge for Global Access*, *Lecture Nottes in Computer Science*. Springer Berlin/Heidelberg , vol. 2911/2003, 253-264

Bottoni P. And al. (2004). MADCOW: a multimedia digital annotation system." In *Proceedings of the working conference on Advanced visual interfaces*. Gallipomi, Italy: ACM.55-62

Cabanac G., Chevalier M., Chrisment C., Julien C. (2007). « An Original Usage-Based Metrics for Building a Unified View of Corporate Documents. » In *Database and Expert Systems Applications*, vol. 4653/2007, *Lecture Notes in Computer Science*. Springer Berlin/Heidelberg. 202-212

Calabretto S., Jean-Marie P., Bozzi A. (1998). "BAMBI: système de gestion de manuscrits anciens pour historiens." *Les bibliothèques numériques*, vol 2. 31-50.

Champin P.-A. (2003). « Ardeco: an assistant for experience reuse in Computer Aided Design. » *In From structured cases to unstructured problem solving episodes*. Trondheim (NO). 287-294.

Egyed-Zsigmond E., Mille A., Prié Y. (2003). « Club ♣(Trèfle) : A Use Trace Model. » In *Case-Based Reasoning Research and Development*, vol. 2689/2003, *Lecture Notes in Computer Science*. Springer Berlin/Heidelberg. 1056

Hilbert, D. M. and Redmiles, D. F. 2000. Extracting usability information from user interface events. *ACM Comput. Surv.* 32, 4 (Dec. 2000), 384-421.

Kahan J., Koivunen M. (2001). « Annotea : an open RDF infrastructure for shared Web annotations." In *International World Wide Web Conference*. Hong Kong: ACM. 623-632

Le Bourgeois F., Emptoz H. (2007). « DEBORA : Digital AccEss to Books of the RenAissance. » *International Journal on Document Analysis and Recognition*. vol. 9 193-221.

Mladenic D. (1999). « Text-Learning and related intelligent agents: a survey.» *IEEE Intelligent Systems*. 44-54.

Semeraro G., Basile P., Degemmis M., Lops P. (2007). «Content-based recommendation services for personalized digital libraries. » *In Digital Libraries: Research and Development*. Vol. 4877/2007. 77-86.