NLP and Digital Library Management

Lyne Da Sylva

École de bibliothéconomie et des sciences de l'information, Université de Montréal, Canada

ABSTRACT

The field of study of Natural Language Processing (NLP) has developed over the past 50 years or so, producing an array of now mature technology such as automatic morphological analysis, word sense disambiguation, parsing, anaphora resolution, natural language generation, named entity recognition, etc. The proliferation of large digital collections (evolving into Digital Libraries) and the emerging economic value of information demand efficient solutions for managing the information which is available, but which is not always easy to find. This chapter presents the requirements for handling documents in digital libraries and explains how existing NLP technology can be used to facilitate the task of document management.

Keywords: *Digital Libraries, Document Management, NLP applications for Digital Libraries, Metadata, Content Processing, Content analysis, Automatic classification, Named entity recognition for Digital Libraries, Thesauri, Document and information retrieval*

INTRODUCTION

The field of study of Natural Language Processing (NLP) has developed and ripened in the past 50 years or so, from the first machine translation and information retrieval applications to the present. These two areas of research have been far-reaching and pervasive. In the process of resolving issues of understanding natural language, for both translation and retrieval, many sub-areas of NLP have emerged: automatic morphological analysis, word sense disambiguation, parsing, anaphora resolution, natural language generation, named entity recognition, etc.

In today's research in NLP, attention has shifted from machine translation over to different versions of information retrieval (IR) applications. The increasing availability of large collections of digital documents has spurred interest in devising useful technology to handle these. Specifically, the notion of "digital libraries" (Adams, 1995; Fox et al., 1995; Arms, 2000) has emerged, with specific architecture and functionality. This is an area where many mature NLP applications can be brought into play. It is an area mostly associated with IR, which has traditionally used little NLP and yet produced efficient tools; methods needed to include more sophisticated, NLP-based approaches were, up to recently, beyond the reach of IR systems. But digital libraries are much more than simply IR.

This chapter has the following three objectives: (i) to describe the issues relating to the task of managing a digital library; (ii) to explore various NLP applications which can be applied to the task; (iii) to identify new research problems related to these issues.

BACKGROUND: DIGITAL LIBRARIES, DOCUMENT MANAGEMENT AND NLP

Digital Libraries

Digital collections existed long before the advent of the Web and the coinage of the term "digital library". NetLib (http://www.netlib.org/), created in 1985, contains a collection of freely available software, documents, and databases of interest to the numerical, scientific computing, and other communities. The Perseus project (http://www.perseus.tufts.edu/hopper/) was created in 1985 to host a collection of

resources on Ancient Greece: documents, images of artefacts, maps and the like, all linked together to allow a better understanding of Ancient Greek texts. Cornell University's e-prints archive (http://arxiv.org/), formerly the Los Alamos E-print Archive, dates from 1991. It contains prepublications in the field of physics and related disciplines. These are but a few examples among many. They were, however, isolated initiatives, suffering from minimal interfaces providing access to resources over less than efficient networks. Improvements in interface design and in network configurations, the advent of the WWW and increasing publication of materials on the Web led naturally to the creation of communities of users wishing to share and publish resources – and of technology to support it.

From a computer science perspective, digital libraries are an extension of network technologies, databases and search engines. From an information science viewpoint however, digital libraries are institutions and not machines; they are a logical extension of traditional libraries, whose mission is to acquire, organise and disseminate information. They also mean other things to other groups: a new outlet for content providers, publishers, museums and commercial vendors; a democratization tool for governments; a new service channel for educators. And from the viewpoint of NLP they represent a new opportunity, a new area of application in which to deploy existing technology, perfect it and invent more.

Definitions for what constitutes a digital library are many, and reflect the fact that work on digital libraries stems from a number of different fields, including computer science and information science of course. Relevant literature on new research is to be found in topical conference proceedings: the European Conference on Digital Libraries (ECDL), the Joint (ACM-IEEE) Conference on Digital Libraries, the International Conference of Asian Digital Libraries, the International Conference on Digital Libraries and the new Theory and Practice of Digital Libraries (formerly ECDL). It also is present in library association conferences and pre-existing conferences of information scientists, publishers, abstracting and indexing services, and online database providers (Bearman, 2008).

An early definition, still cited today, comes from Borgman (2000, 42), in which a digital library is as follows:

> *... a set of electronic resources and associated technical capabilities for creating, searching, and using information. In this sense they are an extension and enhancement of information storage and retrieval systems that manipulate digital data in any medium (text, images, sounds; static or dynamic images) and exist in distributed networks. The content of digital libraries includes data, metadata that describe various aspects of the data (e.g. representation, creator, owner, reproduction rights), and metadata that consist of links or relationships to other data or metadata, whether internal or external to the digital library.*

Digital libraries may also be viewed according to the so-called 5S model comprising Streams, Structures, Spaces, Scenarios and Societies (Gonçalves et al., 2004). As a digital library may mean different things to different people, it may be useful to draw on the model proposed by the DELOS Digital Library Reference Model (Candela et al., 2007), which separates levels of application. A Digital Library may be defined as follows:

> *An organisation, which might be virtual, that comprehensively collects, manages and preserves for the long term rich digital content, and offers to its user communities specialised functionality on the content, of measurable quality and according to codified policies. (Candela et al., 2007, p. 17)*

According to this, a Digital Library is an abstract entity, with the specific purpose of catering to a community of users. We focus on three crucial entities in the DELOS model: (i) collections of digital resources, (ii) users who access these resources and (iii) intermediaries that provide functionality for accessing them. The quality requirements and policy issues will not be addressed here. The aim of this chapter is to explore the NLP applications which may be included in a DL to provide useful functionality

to its users. The resources we will mainly be concerned with are those expressed in natural language: text documents are the prime example, but audio recordings, scanned text, descriptions of images or video and the like are also relevant. The users are understood to be human (and not machine agents). The functionalities we are primarily interested in are those which facilitate document access and retrieval by users (as opposed to long-term preservation, for example). The goal of ensuring access to documents involves the task of document management: describing, organising and storing them in such a way that their retrieval is facilitated. Which type of functionality is possible and desirable for document management (and ultimate retrieval) is described in the next section.

Different types of digital libraries exist (see for instance Bearman, 2008). A digital library may be thematic, containing resources linked to a particular theme or discipline. Or it may be genre- or format-based, such as libraries of images or video. Mission- and audience-oriented digital libraries are viewed as a service, such as digital libraries supporting distance education instruction, or digital libraries for children. Another type is institutional repositories, which contain publications and resources of various kinds emanating from a particular institution; many universities have such institutional repositories.

Given this definition, the World Wide Web is not a digital library, lacking a focussed community of users, a curator and a central service provider. Nonetheless, many research efforts applied to the Web as a whole can be fruitfully applied to a digital library context.

Some notable digital libraries include: the National Science Digital Library (NSDL) in the United States; Europeana, the ambitious library of Europe's documented cultural materials currently featuring more than 15 million works of art, books, music, and film; the Gutenberg Project, the first collection of free electronic books; the ERIC (Education Resources Information Centre) collection; the Internet Public Library; the Hathi Trust (a partnership of major research institutions and libraries in the United States); and the National Library of Australia.

Document Management

The metaphor chosen to describe collections of digital content has been the library, not only because of the fact it houses a collection of documents, but also because its aim is that of the traditional library: to allow its users to access its contents (a set of digital resources) efficiently. It follows naturally that the desired functionality from a digital library can be inspired by its traditional counterpart.

Document management as performed in a traditional library setting (as described in Lancaster, 2003, for example) involves a series of steps. First, from an initial potentially infinite source of resources (the Web, for example), a selection is made by the library's managers to retain a certain type or a certain number of resources, hereafter referred to as documents, to make up the library's collection. On the representational axis, these documents need to be represented by a formal description, including title or name, author or creator, source, location, format, etc., i.e. with descriptive metadata. The descriptions are then inserted in a local organisational system: a catalogue; they may have additional metadata attached to them, such as index terms or classification codes, a short summary or description (semantic metadata). On the physical axis, the documents (or their representation) are stored (or accessible via hyperlinks). Finally, functionality is provided to the user for searching or accessing these documents: a search engine, a browsable index or classification scheme, etc., which provide access to the descriptions and/or the documents. In addition, the library, or rather its agents, can disseminate information (such as new acquisitions) to its users. The steps are thus: document selection and acquisition, description, classification, indexing and abstracting, storage, and distribution or presentation to users.

In the digital realm, this so-called "document chain" is a closed one, as users are very often document creators themselves. In addition, with today's facilities for document annotation and tagging, the user may even provide descriptions of various kinds, thus taking an even more important role in the chain, which may not be best described as a chain at all. This is represented schematically in Figure 1,

where blocks contain the documents and users and the annotated arrows represent the processes, where NLP technology can be deployed.

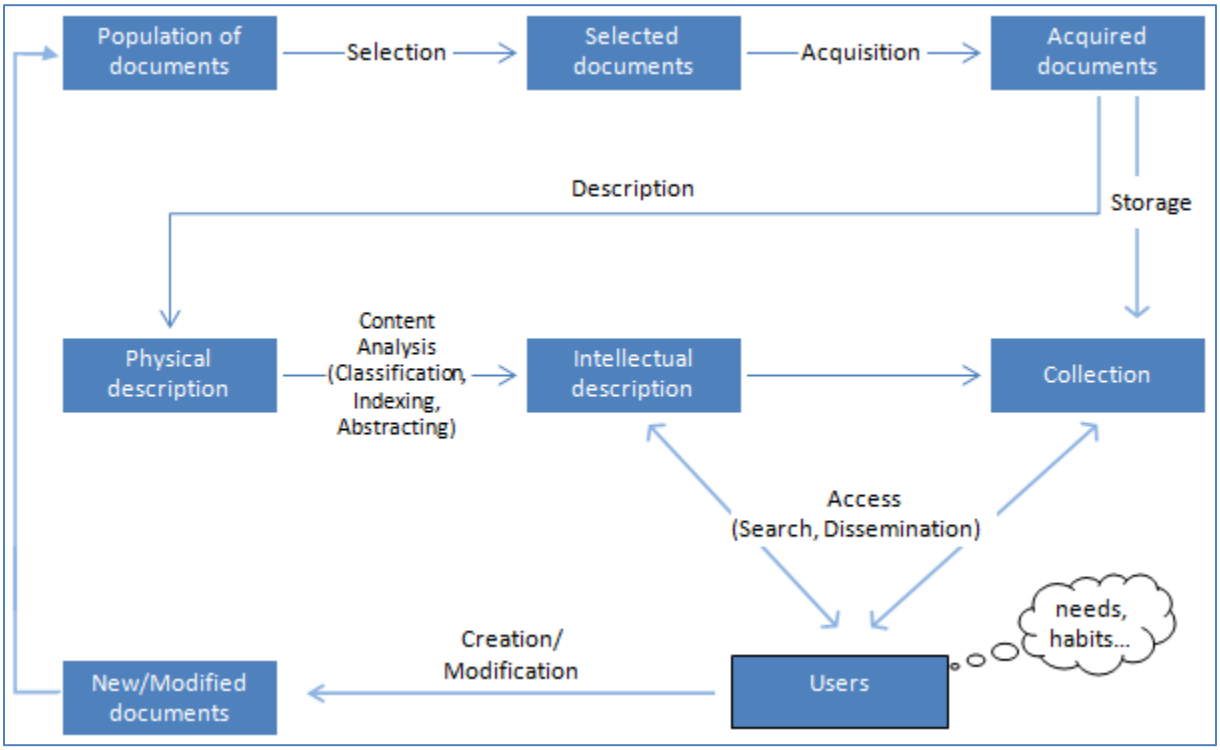

Figure 1: Document processing chain or cycle

Document management has experienced major changes with the spread of personal computers, the development of the Internet and the proliferation of digital text collections. Many operations are done on or by computers, even in a traditional setting. In the context of a digital library, all these operations are of course performed on digital content, which opens a host of possibilities for NLP. And the growing number and size of digital libraries demand management efforts that are time-efficient and consistent, which is what computational methods can offer.

The basic requirements for a digital document management system include the following: a repository of documents (able to handle multimedia content); a system providing access points to documents (i.e. indexing terms); optionally, surrogate representations of documents (e.g. as summaries, especially for non-textual items); retrieval tools (e.g. a search interface); a browsing facility; distribution tools (to disseminate information to existing or new users). The latter is optional but may offer a definite economic advantage.

A note on metadata

From a library and information science (LIS) perspective, metadata corresponds to cataloguing information; that is, the description of a resource by (mainly) its physical or "external" attributes: title, author, publication or creation date, format, length (page numbers for texts, minutes for video and audio), etc. From a computer science perspective, an early definition:

> *Metadata is data associated with objects which relieves their potential users of having to have full advance knowledge of their existence or characteristics. It supports a variety of operations. A user could be either a program or a person. (Dempsey & Heery, 1998)*

Until the middle of the 1990s, the term was used by the data management and systems design communities with a narrower interpretation, relative to a set of standards (Gilliland-Swetland, 2000). Today, its meaning extends to normalized descriptions of resources, digital or other (catalogues, indexes, archival search tools, museum documentation, etc.).

The Dublin Core metadata scheme (http://dublincore.org/documents/dces/), although intended to describe any online resource, contains much of the basic information that librarians recognize as cataloguing information. Its fifteen elements are: contributor, coverage, creator, date, description, format, identifier, language, publisher, relation, rights, source, subject, title and type. The subject and description metadata element correspond to indexing and summarization. Supplying values for these elements requires more than a perusal of superficial document properties, but rather a relatively thorough examination and description of the resource's topic, focus and content.

Metadata comes in two flavours: terse elements encased in a highly-structured container, such as author names and dates of various kinds; or unstructured, perhaps lengthy content, such as summaries. The latter may not always be easily apprehended and may require sophisticated NLP technology to analyse it or indeed to produce it. The former is sometimes straightforward and taken from a controlled vocabulary (for instance, language can be expressed using the ISO 639 standard). But in the case of persons' names (authors, creators and the like), predictable variations on names may make the process more difficult.

And metadata can be wrong. Automatic checking of the accuracy of the metadata can be a useful application in and of itself. The Google Books collection might join the ranks of digital libraries, were it not for its unreliable metadata and its lack of "added-value" functionalities (see a critical review namely in Nunberg, 2009).

NLP and Document Management

The role that NLP can play in document management was realized early on (e.g. Masterman et al., 1958; Sparck Jones, 1967), particularly for document retrieval. The interest is growing (see for example Ambroziak & Woods, 1998; Strzalkowski 1999; Voorhees 1999; Perez-Carballo & Strzalkowski 2000; Oard et al., 2001; Todirasçu & Rousselot, 2001; Ruch, 2003; Radev & Lapata, 2008; Kastner, 2009). There are important links to be made with the semantic Web, aimed at improving retrieval based on semantic grounds rather than on the presence of character strings in documents. See for instance the International Conference on Digital Libraries and the Semantic Web (http://www.icsd-conference.org/). A new development, with the advent of powerful players like Google and the like, is that there are very important stakes involved, due to the growing economic value of digital information.

Practically all NLP applications are relevant and potentially useful in a digital library setting. In particular, methods for information retrieval are an integral part of search engines, and as such are incorporated in virtually any digital library along with all supporting technologies such as word-sense disambiguation, etc.

Here are the basic characteristics of digital libraries and how they may influence the method of deployment of NLP approaches:

- Large collections of texts: sentence-based processing is of limited use if not coupled with other types. Semantic processing, lexical as well as textual, is of utmost importance; discourse analysis is also useful, especially for summarization.

- Digital material comprising text as well as image, sound and multimedia documents: the resources come in different formats, but the metadata associated with them is homogeneous – and text based.
- A variety of genres: digital libraries' text documents are not restricted to newswire nor to scientific articles, which have received special treatment in the recent history of NLP (for TREC, DUC and TAC conferences, among others); the NLP technology deployed in digital libraries must be adaptable to different settings, different types of documents, with different structural properties.
- Existence of multilingual collections: resources and technologies for various languages are needed to give access to these.
- Substantial links among documents: a digital library's collection usually houses material which is thematically linked, or linked in some other special way. This is what justifies and allows the classification of resources. In published research articles, these links enable the study of various types of relationships between sets of documents, i.e. bibliometrics, or the statistical analysis of published literature. This may use citation analysis (the study of citation patterns and citation ancestry for a given document). A common application of this LIS field is to study the impact of a particular paper (how many publications cite it?), of a group of authors (from a given research laboratory, for instance) or of a given field. For digital libraries, it can allow broadened search (searching for papers on a topic and the papers that each original paper cite, for example). Within thematic-based collections, this means that the documents may address the same topic in subtly different ways, which imply that indexing and retrieval must make finer distinctions among related topics.
- Reliance on metadata to describe the resources: a tendency to embrace normalized schemes and encodings for metadata cohabits with user-defined and user-supplied, locally-relevant metadata of various types. The creation of the metadata can be a focus for NLP, as can be operations of classification on resulting metadata.
- A mandate or mission for the collection and a community of users with fairly specific needs, habits and search behaviour (as opposed to the heterogeneous Web): this defines the requirements for a given digital library. Note that this does not exclude the degenerate example of personal collections of digital images or music, where the target community is a single user. But the mandate can focus the processing by constraining it. One important difference between the Web as a whole and a digital library is that of "community": a digital library is designed by and for a community, which determines the contents of the collection (via discriminating criteria for the inclusion of documents) and the services that are to be provided. The latter notion of "services" is indeed crucial in the definition of what a digital library is, in the information science community. A digital library cannot be a mere gathering of documents with minimal functionality.

These characteristics constrain and shape features of the NLP applications that are needed and useful.

Previous work has addressed particular features of digital libraries and has applied NLP in various implementations, especially in information retrieval. Also, data mining (Cohen, 2006) and text mining (Witten et al., 2003; Sanderson and Watry, 2007) tools are used to extraction information from text documents in order to perform tasks such as classification, metadata extraction and the like (Li et al., 2010). A collection which draws much attention is Wikipedia (Kanhabua&Nørvåg, 2010; Popescu & Grefenstette, 2010).

After 15 years or so of work on digital libraries, and in the face of a wide array of mature NLP technology, a comprehensive overview of how the two can be brought together is timely.

OVERVIEW OF NLP TOOLS IN DOCUMENT MANAGEMENT

This section sketches the spectrum of NLP applications for document management, grouped according to four aspects:

- Resource acquisition (including creation, representation and storage)
- Content processing
- Getting users in touch with documents
- Knowledge organisation tools

Resource acquisition

This aspect covers issues dealing with the acquisition of resources and the related questions of the representation of document files which are sensitive to language.

A library's collection is never final; it is continually augmented by newly acquired material. Which material is added is determined by library policy, based on a number of criteria. Leaving economic matters aside, the criteria may include the following:

- topic (e.g. ornithology for a bird-watching club documentation center; business-related literature for a financial institution's library);
- genre (biographies or novels for a public library; conference proceedings for a university or research library; personal correspondence for an archival library; movies for a cinema school's library);
- intended audience (picture books for a preschool library; junior dictionaries for a school library);
- author (for government libraries).

Documents can be added to a digital collection by downloading, creation, digitization, transformation (from one format to another), etc.

Acquiring documents

In some cases, the acquisition of new documents to be added to a digital library can be automated using NLP tools. This is especially true when the selection criteria involve topic: a profile can be defined which expresses the selection criteria for the digital library, as features of the documents; new documents' contents can compared to the profile and processed by an automatic classification algorithm. Joorabchi & Mahdi (2008) describe an implementation of such functionality for a national repository for course syllabi (see also references therein). A very similar task is also performed by so-called « information-filtering systems » (see among others Belkin & Croft, 1992, Hanani et al., 2001), which intervene between an automated retrieval system and a user, to restrict the number of documents retrieved.

In addition, it is sometimes necessary to transform non textual documents into textual documents, by NLP means: optical character recognition (see Mello & Lins, 1999, for a comparative study of OCR tools and their performance level), handwriting recognition (Plamondon & Srihari, 2000) for certain historical archives, transcription of audio materials, machine translation, extraction of text from HTML or PDF formats, etc. An additional step of checking for spelling, grammar and style of documents can be performed, when they are acquired by these types of transformations, and thus quite error-prone,.

Determination of proper processing tools

Tools which will be used to process the documents, for example term extractors, part-of-speech taggers, summarizers, etc, are language-sensitive: German texts for instance require different tools than Chinese texts. It is a reasonable assumption in today's understanding of digital libraries that they are intended to be multilingual. To optimize the overall functioning of the library management system, it is desirable to include in the system functionalities for the automatic identification of language and encoding. Such systems have been developed in the past 15 years, based on character *n*-gram profiles. Řehůřek & Kolkus (2009) provide an up-to-date presentation in the context of the Web.

Document description

To represent and store documents in a digital library, it is necessary to produce some sort of record by which they are accessed. This corresponds to a traditional library's bibliographic entry, or a metadata record (i.e. descriptive metadata). This record is typically produced explicitly, either hand coded or automatically produced by extracting metadata from the resource. No semantics is involved and usually very little NLP technology. However, the normalisation of author names and titles is a reasonable objective, and would require NLP tools similar to those for the normalisation of named entities (see for instance Andréani & Lebarbé, 2010). See also Kanhabua & Nørvåg (2008) on automatic means of determining a timestamp for documents which lack one. Also, one can imagine including here the results of automatic identification of document language and encoding, or of date formats.

The descriptive or "physical" metadata described above is often not sufficient, or not ideal, for retrieval by a library's users. Additional metadata can be produced automatically by content processing.

Content processing

Content processing is a major part of the document management endeavour. It consists in producing enhanced metadata descriptions, in order to facilitate document retrieval by users, in addition to the retrieval capabilities provided by full-text searching. Resulting metadata is to be included in the digital library's knowledge organisation system. Content processing implies performing an analysis of the linguistic and/or conceptual contents of the text documents, and produces appropriate representations for these documents (such as indexing terms, summaries, classification codes, etc.). Content processing thus covers the traditional tasks of classifying, indexing and summarizing documents. Classifying implies grouping together documents on similar topics, and usually makes use of a classification scheme (such as the Dewey Decimal Classification or the Universal Decimal Classification, etc.); its analog in the digital world would be the hierarchical presentations of directories. Indexing (which may be interpreted differently by different communities) involves here the description of documents with a short list of terms or keywords representing the main topics discussed in the document. Summarization yields a shortened form of documents in a (usually) narrative style.

These content processing tasks are tackled by three basic NLP technologies. First comes the triplet of automatic classification, categorization and clustering (Yang, 1999; Boutella et al., 2004; Tsoumakas & Katakis, 2007). An overview of techniques for automatic text classification is presented in Sebastiani (2002). Toms & McCay-Peet (2009) discuss the relevance for this topic for digital libraries, in that it may enable serendipitous discovery such as is possible when browsing a physical library's shelf, and may enhance focused search. This is the reasoning, of course, behind approaches which cluster a search engine's results. Classification using an established bibliographic-type classification (such as Dewey) has been gaining attention: Vizine-Goetz (1996), Thompson et al. (1997), Jenkins et al. (1998), Prabowo et al. (2002), Hodge et al. (2003), Golub (2006). Wang (2009) notes the challenges posed to state-of-the-art text categorization technologies by library classification systems, such as the Dewey classification, with its deep hierarchy, data sparseness, and skewed distribution; they offer reasons why

classification is desirable in the context of growing digital collections and describe previous approaches before offering their machine-learning solution to the problem. On clustering, see for instance Aas & Eikvil (1999), Steinbach et al., (2000) and Grira et al. (2006) for general presentations. Yoo (2006) performs a comprehensive comparison study of various document clustering approaches on MEDLINE and also applies a domain ontology such as MeSH to document clustering; this is done in order to investigate if the ontology improves clustering quality for MEDLINE articles. Chengzi & Dan (2008) introduce a new approach for building a topical digital library, using concept extraction and document clustering; thus clustering here is used for collection creation. Note that automatic classification can also be applied to search results, as variants in the presentation (see for instance Palmer et al., 2001).

The second content processing task is indexing; it is implemented by search and retrieval methods. All forms of information retrieval (Van Rijsbergen, 1979; Perez-Carballo, J. & T. Strzalkowski, 2000; Sparck-Jones, 2007, to name only a few), and the related topics of automatic annotation (essentially a synonym for automatic indexing) and metadata extraction (e.g. Edvardsen et al., 2009; Ciravegna et al., 2004; Kelly, 2004;; Péter, 2004) are highly relevant to digital libraries; see also Rasmussen (2004) on information retrieval challenges for digital libraries. Examples of applications for indexing include the following. Krapivin et al. (2010) add NLP techniques to machine learning (Support Vector Machines (SVM), Local SVM, Random Forests) to improve the extraction of keyphrases from scientific documents; the digital library to which they apply their algorithm consists of ACM papers from the Computer Science domain. Tahmasebi et al. (2010) study word sense discrimination on a historical document collection to improve understanding and accessibility of this particular digital archive. For search and retrieval: Batjargal et al. (2010) use a translation dictionary to enable retrieval of ancient historical documents written in traditional Mongolian using a query in modern Mongolian; Gou et al. (2010) use a combination of tf-idf measures and social networks (of the user community) to improve ranking algorithms for retrieval. Finally, recommender-type systems (Hwang et al., 2003; Krottmaier, 2002; Faensen et al., 2001; Huang et al., 2002; Smeaton & Callan, 2005; Avancini et al., 2007) and user-preference based ranking (Manolopoulos & Sidiropoulos, 2005; Mutschke, 2003) can also bring much-appreciated functionality to the search facilities of digital libraries. For recommender systems, the underlying technology can be document classification, i.e. determining whether a new document belongs to the (theoretical, virtual) class of "documents interesting to this user". Note however that Bearman (2008:242) remarks that no recent studies have examined user satisfaction with different methods of ranking.

The third content processing task, automatic summarization, tries to replicate and improve on human summarization of documents. Sparck Jones (2007) presents an overview of present-day summarization technology; Kan & Klavans (2002) use librarians' techniques to produce summaries for information retrieval; Ou et al. (2009) describe summarization in the context of a digital library. Jaidka et al. (2010) perform multi-document summarization of research papers based on techniques drawn from human summarization behaviour and guided by discourse analysis. Wan et al. (2009) use properties of scholarly articles, namely citations. They construct a summary for a cited text which is focused on the context: sentences from the cited document are extracted based on elements from the citation context.

Within a digital library framework, content processing (specifically: indexing, summarizing and classifying documents) enables the system to add information to the basic bibliographic entries containing metadata such as a document's title, author, date of creation, URL, format, etc. The result of content processing adds indexing keywords or classification codes, i.e. additional access points which should enable easier retrieval, or summaries which make it easier to ascertain the document's relevance to the user's needs.

Getting users in touch with documents

This aspect deals with the *raison d'être* of libraries: access to documents by users, either by their own initiative (retrieval) or by the information system's ability to broadcast news out to a community of users.

Document/information retrieval

In a traditional library setting, actual document retrieval is often preceded by a "reference interview", where a librarian tries to ascertain the exact information needs of the user and thus to develop a successful search strategy which will include online search as well as searches in other sources. In a digital library world, this initial phase is non-existent. Users refine their search strategy themselves, gradually, as a reaction to the responses of the system and to what they discover about the contents of the collection. In addition, certain features of the digital library system have been designed to simulate the broadening or sophistication of the search that a librarian would perform. And thus document retrieval in a digital setting is reducible to so-called "information retrieval". This is probably the best-researched field in document management. The presentation here will only aim to underline the array of NLP technology used (this is also addressed by Mustafa el Hadi, 2004).

Search engines minimally tackle basic issues of matching terms or concepts between queries and documents. More specifically, the match is performed between a query and a previously-compiled index of terms and expressions extracted from the document collection (see Indexing, above). The query may also undergo the same processing as documents did during the indexing phase, i.e. stemming or lemmatization, disambiguation, etc. Search engines of all types perform this daily.

Query expansion refers to the process of adding terms to a query, to broaden a search for example, or on the contrary to further specify one of the query terms; this, incidentally, would be done naturally by a librarian devising a search strategy. This may be achieved by using a thesaurus to capture synonyms to add to the query (thus adding words like "building" to a query containing "construction", to capture related items). Or more general terms can be added, such as adding "material" to a query containing "concrete" and "plaster", to capture other types of building materials. Finally, more specific terms can be added: "lark" or "finch" could be added to a query containing "bird", in case some documents mention only specific breeds. This can be performed if the system contains an appropriate thesaurus. Additional semantic processing may be required, to determine which strategy should be taken for a given query. See Song et al. (2006) for an application of query expansion to digital libraries.

A multilingual document base can present a challenge to document retrieval: the user's query may not contain words used to index the documents, because they are in a different language. Cross-linguistic information retrieval (CLIR) relies on translation dictionaries, or other translation technology, to bridge the gap between users' queries and documents. See for instance Nie (2010) for a presentation of the field, and Oard (1997) for an early recognition of its relevance to digital libraries.

A librarian would verify that the information supplied to a user does indeed answer his or her needs, by asking whether the supplied documents are deemed relevant by the user. In a digital information retrieval setting, so-called relevance feedback is an automated version of this exchange. It can be used to improve search results by using additional knowledge sources. The most basic type of relevance feedback relies on a user's judgment of relevance of selected documents. These judgments are used to issue a new query which includes terms extracted from the relevant documents. Simple co-occurrence statistics on words or terms can be used, but there is also an opportunity for more elaborate semantic processing to be performed in ascertaining the relevance of documents for a given user in a specific community.

Broadcasting documents to users

It is customary for an information service such as a library to issue bulletins to its users, informing them of new material or special events, when appropriate. This can be done through mailing lists, billboards, etc. The equivalent in the digital world is straightforward. What is novel here, however, is that bulletins can be tailored to individual user profiles. Specifically, new documents can be analysed (indexed, classified or summarised) and compared to a user profile consisting of user-supplied or system-supplied keywords; in the event of a match, users can be notified of these new documents through appropriate

messaging technology (e-mail, RSS feed, etc.). Such a system is described in Morales del Castillo et al. (2009) while Gu et al. (2008) present a similar functionality to support learning.

Answering users' questions

A major part of every librarian's day involves answering questions for users. Some modern versions of such a reference service employ chat rooms and the like ("Ask-a-librarian" services), with a human librarian accessible over the internet. An even more modern take on the idea is to use a question answering system, such as in Mittal et al. (2005) or Bloehdorn et al. (2007).

The task of relating users to documents is obviously at the core of a library's mission and of digital libraries' functionalities. NLP tools can assist in various ways, as has been illustrated so far. We now turn to an aspect which transcends document management tasks.

Knowledge organisation tools

We refer here to linguistic resources used in the text management and processing tasks described above. The one that is most specific to document management is the thesaurus (other knowledge organisation tools relevant for digital libraries are presented in Soergel, 2009).

Properties of thesauri

Note that the term "thesaurus" means slightly different things to information professionals (librarians) and computer scientists, or to language educators for that matter. Loosely speaking, a thesaurus is some kind of synonym dictionary; in reality it is much more. It encodes not only synonymous terms but also hierarchical relationships (i.e. which terms are broader and narrower than a given term) and other types of semantic relationships, depending on the resource. Specifically, the "thesaurus" most used in NLP applications, WordNet, is not a thesaurus by LIS standards.

The LIS version of the thesaurus (defined by international standards ISO 2788 and ISO 5064) adopts a stricter definition of thesaural relationships. These are restricted to only three types: (i) hierarchy (broader/narrower terms or generic/specific terms, otherwise known as hypernym/hyponym terms); (ii) synonymy (semantic equivalents which may include spelling variants, shortened forms, etc.); and (iii) the so-called associative relationship, relating terms that are neither synonyms nor in a hypernym/hyponym relation, yet are related semantically. Thesaural relations exclude (almost all) partitive (part-whole) relationships and others which are routinely introduced in ontologies.

An innovation of the WordNet thesaurus is the declaration of *synsets*, i.e. sets of words based on their meanings which are deemed synonymous and have an equal status in the system. In a traditional thesaurus, when synonyms are identified, one term is promoted to the rank of descriptor (or "preferred term", an indexing candidate), and its synonyms are relegated to non-descriptors or non-preferred terms, not used in indexing. Only descriptors may entertain hierarchical or associative relationships with other terms. Potentially ambiguous descriptors (such as river *banks* and financial *banks*) are disambiguated not through their meaning, but through explicit descriptor modification. All descriptors in a thesaurus are formally or graphically different: thus a thesaurus would differentiate explicitly "river bank" and "financial bank", or the modified terms "bank (boundary)" and "bank (financial institution)". In addition, synonyms need not be true semantic synonyms, but merely recognized as sufficiently synonymous in a given (indexing and retrieval) context. For instance, "bow" and "arrow" can be declared synonyms in a thesaurus used to describe a collection where so few documents mention either that they are best handled together.

The associative relationship is more vague (and is seen as problematic by automatic semantic processing approaches) but is deemed easier to understand and use by humans in the context of indexing and subsequent searching.

Uses of thesauri in digital libraries

The content management tasks (automatic indexing, classification and summarization) can greatly benefit from knowledge sources such as thesauri, which encode semantic relationships among words and terms. The two most basic of these are the synonymy relation and the hypernym/hyponym relation. The two can be used to improve on content processing, such as indexing with more general or more specific terms, and bringing together synonymous expressions to enhance indexing or to allow generalizations in summarizing.

Automatic construction of thesauri

Attempts have been made to create thesauri by automatic means, to overcome the problem of the scarcity of appropriate resources. General language thesauri (such as WordNet and the like) offer a wide coverage, but have serious limitations in specialized domains. Specialized thesauri have the opposite flaw (often too narrow in scope), and are in addition fairly rare, often not available for a given specialized domain. To circumvent these problems, the automatic construction of a thesaurus is an endeavour that has been attempted by several researchers (see for instance Auger & Barrière, 2008 and others). The linguistic challenge lies in the automatic identification of semantic relations of synonymy, hypernymy/hyponymy, and other "essential" semantic relationships which may be difficult to characterize exhaustively. All of these present serious challenges. This research area is close to that of ontology learning and population from text.

Meusel et al. (2010) present a method for extending an existing thesaurus using a mixture of machine learning and NLP; they test their method on MeSH and WordNet. Eckert et al. (2010) use human expert feedback on the relatedness and relative generality of terms to construct dynamically changing concept hierarchies; although not using NLP methods, their work is relevant among other things to suggest novel ways of automating parts of it. However, Arms&Arms (2004) suggest that in heterogeneous collections, controlled vocabularies and shared ontologies are unachievable; accordingly, they recommend brute force, full-text indexing (Bearman, 2008:240).

Summary

Tools for the acquisition, description and dissemination of resources are basic requirements of a digital library system and may rely on knowledge organisation resources. Many language -related issues must be addressed for the management of a digital library, and it can indeed benefit from natural language processing tools.

A CLOSER LOOK AT SOME CHALLENGES FOR DIGITAL LIBRARY MANAGEMENT

The previous wide-ranging exposé has identified numerous possibilities for NLP applications in the context of digital library management. The rest of this chapter focuses on certain specific challenges met by digital libraries.

Named entity recognition and resolution

It is useful and often necessary to be able to determine when two similar variants of a named entity in fact designate the same one: John Smith, J. Smith, Pres. Smith, John Smith Jr., etc. Organisation names can also vary: Acme Deliveries vs. Acme Deliveries Inc; IBM vs. International Business Machines; The John Hopkins University vs. John Hopkins; etc. This problem is compounded when names come from a foreign country, possibly through transliteration from a foreign language. This has long been recognized in library cataloguing and is the focus of sections in the Anglo-American Cataloguing Rules handbook (Joint Steering Committee for Revision of AACR, 2002). In the domain of scholarly publications, names of institutions, universities, research laboratories, etc. can manifest different variants. This presents a

problem when one wants to identify named entities emanating from different sources: different publications, different libraries, in bibliographies from different documents, sometimes dictated by bibliographic styles. It is a problem for a number of endeavours and is indeed a topic of many research papers related to digital libraries.

When creating metadata on authors or creators, it is desirable to ascertain a person's name in a non ambiguous manner. Feitelson (2004), Hong et al. (2004), Wu et al. (2004) and Bainbridge et al. (2011) study the problem of name variants in digital libraries.

In citation analysis, it is also crucial to distinguish people with similar names while allowing variants for the name of each person. Fereira et al. (2010) discuss the problems encountered in such a task and propose a disambiguation method for a given name based on a two-step method: clustering citation records based on the similarity of co-author names, followed by unsupervised disambiguation. Treeratpituk & Giles (2009) use random forests (a machine learning classification algorithm) to perform the disambiguation task in academic publications. Pereira et al. (2009) use information available on the Web (curricula vitae and Web pages containing publications of the ambiguous authors) and a hierarchical clustering method that groups citations in the same document together, to disambiguate similar names and detect variants. Sugiyama&Kan (2010) tackle the task of recommending new articles to researchers based on their past works, by comparing the cited works in each. In addition, citation analysis is used as a retrieval method by Péter (2004); He et al. (2004) use citation-based retrieval rather than subject retrieval to search scholarly publications.

In another context, Haruechaiyasak & Damrongrat (2010) apply textual analysis to identify persons appearing in photographs in news articles; for this purpose, named entity recognition and disambiguation is necessary.

Not all named entities designate persons, organisations or even geographical entities (see, on the latter, Freire et al., 2011); in biomedical and chemical literature, proteins, diseases, chemical, genes, etc. are entities which must be identified in the text. They exhibit various peculiarities which make them difficult to spot consistently (Tönnies et al., 2010). Kanhabua & Nørvåg (2010) propose methods to identify variants of named entities describing time or events and present an evaluation based on TREC collections.

One can envision coupling citation analysis and content analysis (described in the previous section) to perform multidimensional classification of sets of documents, in which case again the question of recognizing variants of a named entity is essential. The ability to do so may enable a system to bridge across different digital libraries.

Tools to assist OCR

Some challenges arise due to the digitization process of certain types of documents: namely, historical documents and so-called retrospective collections of modern digital media. Access to these is hampered by the poor quality of the OCR text. Tahmasebi et al. (2010) investigate the effects of OCR errors on word sense discrimination results on historical documents; evaluations are performed on The Times newspaper archive, with documents dating from 1785 to 1985. Allen et al. (2010) tackle the task of identifying sections and regular features of historical newspapers in order to improve the automatic classification of articles; the ultimate goal is to provide improved search services for these documents.

Search and retrieval

Improved search strategies are needed. Methods which favour precision (eliminating irrelevant items) are especially sought, as we see the development of topical digital libraries – where distinctions between documents can be finer-grained than on the Web as a whole (Bethard, 2009). On the other hand, to enhance recall, the integration of lexical resources such as thesauri and ontologies should be useful.

The context of a digital library – namely, the knowledge that one may have of its users – should enable improved evaluation of the effectiveness of search technology. The evaluation can be done not in a general way, but in a manner specific to the community being serviced by the digital library. It has been suggested that useful metadata is not necessarily linked to content, but that contextual metadata, describing groups that share work processes and workflow process models, are more useful than content descriptors in some instances (Klas, Fuhr, & Schaefer, 2004). This however is not derived from the source document and thus NLP techniques will have limited impact on this topic.

Question-answering systems can act the part of a librarian, and provide answers to questions rather than documents containing the answers (Bloehdorn et al., 2007; Vakkari & Taneli, 2009).

A related area of research consists in reconciling controlled vocabulary and natural language tagging (see for instance Seki et al., 2010): the advantages of the controlled vocabulary may be counterbalanced by those of tagging. Controlled vocabularies offer disambiguation of homonyms/homographs, grouping of synonym terms, which result in higher inter-indexer consistency and higher recall, whereas tagging manifests closeness to the vocabulary of the users, quick adaptation to neologisms, both resulting in higher precision and in some cases higher recall. Applying to natural language tags the same type of processing as that used in automatic thesaurus construction (thus bringing it automatically closer to a controlled vocabulary) could help harness the power of each type.

Retrieval of non-textual documents

One interesting aspect of digital libraries is that they bring together three formerly quite distinct disciplines, i.e. libraries, archives and museums. Digital resources in digital libraries are not limited to textual documents, nor to digital objects, but can include images, video, sound, and digital renderings of three-dimensional physical objects. The extraction of information from the text surrounding images can support automatic indexing of these images (see for instance Haruechaiyasak & Damrongrat, 2010), and the same can be applied to video, audio or multimedia resources (Da Sylva & Turner, 2005).

Genre-based processing

Genre-based processing (i.e. that which takes into account the genre or type of a document and can adjust accordingly) is an important issue that can be tackled by NLP means. For example, in automatic summarization, Saggion & Lapalme (2000) take advantage of the predictable structure of scientific articles to focus on certain sections from which to extract sentences which will appear in the final extract. Chieze et al. (2010) take a similar approach to handle specific types of legal documents (court judgement renderings, and intellectual property and tax law texts). The latter are examples of single-genre processing. To allow for processing of more than one genre would improve on existing, "off-the-shelf" technology which is geared towards a single genre.

FUTURE RESEARCH DIRECTIONS

One aspect which is less well-researched is the access to sub-document structures: how can the system help the user in targeting more precisely the information within a document? This relates to certain applications of XML retrieval (see for instance Smadhi, 2003). Various technologies can provide reading aids for digital documents, enabling a quicker perusal of document contents to ascertain relevance or to enable faster information gathering. Traditionally, this type of information search was enabled by back-of-the-book indexes. Full-text searching may have rendered some aspects of book indexing obsolete, but it can still be a useful tool as a browsable snapshot of the document's contents and as an indicator of the relationships among topics in a document. Work on automatic back-of-the-book indexing has been extremely scarce in the past few decades, although it was experimented with early on (Artandi, 1963; Earl, 1970; Salton, 1988). See however Da Sylva (2004), Da Sylva & Doll (2005) and Nazarenko & El Mekki (2005) for more recent implementations. Owen et al. (2010) explore ways to improve cursory

navigation in a document collection; their proposed methods include what they call "semantic" rendering, in which the document display is altered depending on scroll speed. This type of aid to navigation and information evaluation could be explored further to include some types of document summarization or indexing. Melucci (2004) describes the design and the implementation of a tool that generates networks of links within and across hyper-textbooks through a completely automatic and unsupervised procedure; this supports access to information encapsulated in textbooks.

An ongoing concern is that of providing more than the traditional library was capable of: using computer technology in general and NLP in particular to provide functionality which was impossible in the traditional setting. This includes things such as multi-document or query-based summarizing, which can be produced at will based on varying parameters (as opposed to human-produced summaries, created once and used for every type of query or need).

Other concerns for digital libraries that are not *a priori* the concerns of NLP, but which are ubiquitous and can impact application of technologies: the legal aspects linked to intellectual property for documents included in digital libraries; information-seeking behaviour; processing in distributed architectures (sometimes involving different systems); long-term preservation of digital materials. The latter two aspects may be lessened by adopting recognized, open standards.

Long-term preservation in particular is a serious question, given the non perennial nature of computer media (including CDs, hard drives, etc.). A traditional library's print collections will last for hundreds of years, but our digital files may not. As mentioned in the DELOS Reference model, preservation may also be viewed as interoperability over time (Candela et al., 2007, p. 57): ensuring that the digital files of today can still be read and understood correctly in the future. How NLP can contribute to the solution to this problem remains to be seen.

CONCLUSION

The digital library setting represents an interesting opportunity for computational linguistics: it can use many new applications with great potential (notably, a great financial or economic potential, given the new economic value of information). Current focus on very large digital libraries may test the robustness of seemingly mature NLP technology.

In the past, syntax has played a large role in NLP development, notably in symbolic approaches to machine translation, where systems were developed with translation rules from one language's syntactic constructions to another. So far, syntax has played a very small part in NLP for document management (see however Spagnola & Lagoze, 2011). Research must now focus on computational semantics: lexical, phrasal and sentential semantics, and in even higher level units. Indeed, text linguistics or discourse analysis will drive new research, especially for summarization and certain approaches to classification. In the long term, the ultimate challenge will be to model more than merely the linguistic dimensions of digital library management, adding also cognitive, communicational, pragmatic, social or semiotic dimensions, etc. These can appeal to cognitive science and artificial intelligence in general; but even in the linguistic dimensions, challenges abound.

REFERENCES

Aas, K., & Eikvil, L. (1999). *Text Categorisation: A Survey*. Technical report, Norwegian Computing Center. Retrieved October 7, 2010, from http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.41.2236.

Adam, N. R. (Ed.) (1995). *Digital libraries : research and technology advances : ADL'95 Forum*, McLean, Virginia, USA, May 15-17, Forum on Research and Technology Advances in Digital Libraries. Berlin: Springer, 1996.

Allen, R. B., & Hall, C. (2010). Automated Processing of Digitized Historical Newspapers beyond the Article Level: Finding Sections and Regular Features. In *Proceedings of ICADL2010*, pp. 91-101.

Ambroziak, J., & Woods, W. A. (1998). Natural Language Technology in Precision Content Retrieval. In *Proceedings of the International Conference on Natural Language Processing and Industrial Applications (NLP+IA 98)*, August 18-21, 1998, Moncton, New Brunswick, Canada. Retrieved October 7, 2010 from http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.21.9236.

Andréani, V. & Lebarbé, T. (2010). Named entity normalization for termino-ontological resource design: mixing approaches for optimality. In *Proceedings of 10th International Conference Journées d'Analyse statistique des Données Textuelles*, 9-11 June 2010 (pp. 163-172).

Arms, W. Y. (2000). *Digital libraries*. Cambridge, Ma.: MIT Press.

Arms, W. Y., & Arms, C. R. (2004). Mixed content and mixed metadata: Information discovery in a messy world. In D. Hillmann & E. Westbrooks (Eds.), *Metadata in practice* (pp. 223-237). Chicago: American Library Association.

Artandi, S. (1963). *Book indexing by computer*, New Brunswick, N.J.: S.S. Artandi.

Auger, A., & Barrière, C. (2008). Pattern based approaches to semantic relation extraction : a state-of-the-art. *Terminology*, John Benjamins , 14(1), 1-19.

Avancini, H., Candela, L., & Straccia, U. (2007). Recommenders in a personalized, collaborative digital library environment. *Journal of Intelligent Information Systems*, 28(3), 253 – 283.

Bainbridge, D., Twidale, M.V., &Nichols, D.M. (2011). That's 'é', not 'þ' '?' or '□': A user-driven context-aware approach to erroneous metadata in digital libraries. In *Proceedings of JCDL 2011*, Ottawa, Canada, June13-17, 2011.

Batjargal, B., Khaltarkhuu, G., Kimura, F.; & Maeda, A. (2010). Ancient-to-modern Information Retrieval for Digital Collections of Traditional Mongolian Script. In *Proceedings of ICADL2010*, pp. 25-28.

Bearman,.D. (2008). Digital Libraries. *Annual Review of Information Science and Technology*, 41(1): 223-272.

Belkin, N., & Croft, B. (1992). Information Filtering and Information Retrieval: Two Sides of the Same Coin?, *Communications of the ACM*, 35(12), 29-38.

Bethard, S., Ghosh, S., Martin, J. H., & Sumner, T. (2009). Topic model methods for automatically identifying out-of-scope resources. In *Proceedings of JCDL2009: 9th ACM/IEEE-CS Joint Conference on Digital Libraries* (pp. 19-28), Austin, TX,USA, June 15-19, 2009.

Bloehdorn, S., Cimiano, P., Duke, A., Haase, P., Heizmann, J., Thurlow, I., & Völker, J. (2007). Ontology-Based Question Answering for Digital Libraries. In L. Kovács, N. Fuhr, & C. Meghini (Eds.), *Research and Advanced Technology for Digital Libraries*. Lecture Notes in Computer Science, Volume 4675 (pp. 14-25).

Borgman, C. L. (2000). *From Gutenberg to the Global Information Infrastructure: Access to Information in the Networked World*. Cambridge, MA: The MIT Press.

Boutella, M. R., Luob, J., Shena, X., & Brown, C. M. (2004). Learning multi-label scene classification. *Pattern Recognition*, 37, 1757 – 1771.

Chieze, E., Farzindar, A., & Lapalme, G. (2010). An Automatic System for Summarization and Information Extraction of Legal Information. In E. Francesconi, S. Montemagni, W. Peters, & D. Tiscornia (Eds.), *Semantic Processing of Legal Texts: Where the Language of Law Meets the Law of Language*, vol. 6036, series. Lecture Notes in Computer Science (pp. 216-234), Berlin: Springer, June 2010.

Ciravegna, F., Chapman, S., Dingli, A., & Wilks, Y. (2004). Learning to harvest information for the Semantic Web. In *Proceedings of the 1st European Semantic Web Symposium*, pp. 312-326.

Cohen, D. J. (2006). From Babel to knowledge: data mining large digital collections. *D-Lib Magazine*, 12(3).

Da Sylva, L. (2004). A Document Browsing Tool Based on Book Indexes. In *Proceedings of Computational Linguistics in the North East (CliNE'04)* (pp. 45-52), Université Concordia, Montréal, 30 Aug 2004.

Da Sylva, L., & Doll, F. (2005). A Document Browsing Tool: Using Lexical Classes to Convey Information. G. Lapalme, B. Kégl (Eds), *Advances in Artificial Intelligence: 18th Conference of the Canadian Society for*

Computational Studies of Intelligence, Canadian AI 2005 (Proceedings), New York: Springer-Verlag, 2005, pp. 307-318.

Da Sylva, L., & Turner, J. M. (2005). Using ancillary text to index web-based multimedia objects. *Literary and Linguistic Computing*, 21(2):219-228. Oxford : Oxford University Press.

Dempsey, L., & Heery, R. (1998). Metadata: A current view of practice and issues. *Journal of Documentation*, 54(2):145-172.

Earl, L.L. (1970). Experiments in automatic extraction and indexing. *Information Storage and Retrieval*, 6, 313-334.

Eckert, K., Niepert, M., Niemann, C., Buckner, C., Allen, C., & Stuckenschmidt, H. (2010). Crowdsourcing the Assembly of Concept Hierarchies. In *Proceedings of JCDL2010, 10th ACM/IEEE-CS Joint Conference on Digital Libraries* (pp. 139-148), Surfer's Paradise, Australia, June 21-25, 2010.

Edvardsen, L. F. H., Sølvberg, I. T., Aalberg, T., & Trætteberg, H. (2009). Automatically generating high quality metadata by analyzing the document code of common file types. In *Proceedings of JCDL2009: 9th ACM/IEEE-CS Joint Conference on Digital Libraries* (pp. 29-38), Austin, TX,USA, June 15-19, 2009

Faensen, D., Faultstich, L., Schweppe, H., Schweppe, H., Hinze, A., & Steidinger, A. (2001). Hermes: a notification service for digital libraries. In *Proceedings of the 1st ACM/IEEE-CS Joint Conference on Digital Libraries* (JCDL '01).

Feitelson, D. G. (2004). On identifying name equivalences in digital libraries. *Information Research*, 9(4).

Ferreira, A., Veloso, A., Goncalves, M., & Laender, A. (2010). Effective Self-Training Author Name Disambiguation in Scholarly Digital Libraries. In *Proceedings of JCDL2010, 10th ACM/IEEE-CS Joint Conference on Digital Libraries* (pp. 39-48), Surfer's Paradise, Australia, June 21-25, 2010.

Fox, E.A., Akscyn, R.M., Furuta, R., & Leggett, J.J. (1995). Digital Libraries. *Communications of the ACM*, 38(4), 23-28.

Freire, N., Borbinha, J., Calado, P., & Martins, B. (2011). Metadata Geoparsing System for Place Name Recognition and Resolution in Metadata Records. In *Proceedings of JCDL 2011*, Ottawa, Canada, June13-17, 2011.

Gilliland-Swetland, A.M. (2000). Setting the stage. In M. Baca (Ed.), *Introduction to Metadata: Pathways to Digital Information*. Los Angeles: Getty Information Institute.

Golub, K. (2006). Automated subject classification of textual web documents. *Journal of Documentation*, 62 (3), 350-371.

Gonçalves, M. A., Fox, E. A., Watson, L. T., & Kipp, N. A. (2004). Streams, Structures, Spaces, Scenarios, Societies (5S): A Formal Model for Digital Libraries. *ACM Transactions on Information Systems*, 22 (2), 270-312.

Gou, L., Chen, H.-H., Kim, J.-H., Zhang, X.L., & Giles, C. L. (2010). Social network document ranking. In *Proceedings of JCDL2010, 10th ACM/IEEE-CS Joint Conference on Digital Libraries* (pp. 313-322), Surfer's Paradise, Australia, June 21-25, 2010.

Grira, N., Crucianu, M., & Boujemaa, N. (2006). Unsupervised and Semi-supervised Clustering: a Brief Survey. In S. Boughorbel, (Ed.), *A Review of Machine Learning Techniques for Processing Multimedia Content*, Report of the MUSCLE European Network of Excellence (6th Framework Programme). Retrieved October 7, 2010 from http://www-rocq.inria.fr/~crucianu/src/BriefSurveyClustering.pdf.

Gu, Q., de la Chica, S., Ahmad, F., Khan, H., Sumner, T., Martin, J. H., & Butcher, K. (2008) Personalizing the Selection of Digital Library Resources to Support Intentional Learning. In B. Christensen-Dalsgaard, D. Castelli, B. Ammitzbøll Jurik, & J. Lippincott (Eds.), *Research and Advanced Technology for Digital Libraries*. Lecture Notes in Computer Science, Volume 5173 (pp. 244-255).

Hanani, U., Shapira, B., & Shoval, P. (2001). Information filtering: Overview of issues, research and systems. *User Modeling and User-Adapted Interaction* 11, 203–259.

Haruechaiyasak, C., & Damrongrat, C. (2010). Identifying Persons in News Article Images Based on Textual Analysis. In *Proceedings of ICADL2010*, pp. 216-225.

He, Y.; Hui, S. C.; & Fong, A. C. M. (2003). Citation-based retrieval for scholarly publications. *IEEE Intelligent Systems*, 18(2), 58-65.

Hodge, G. M., Zeng, M. L., & Soergel, D. (2003). Building a meaningful Web: from traditional knowledge organization systems to new semantic tools. In *Proceedings of the 3rd ACM/IEEE-CS joint conference on Digital libraries*, Houston, Texas (pp. 417-417).

Hong, Y., On, B.-W., & Lee, D. (2004). System support for name authority control problem in digital libraries: OpenDBLP approach. In *Proceedings of the 8th European Conference on Digital Libraries*, pp. 134-144.

Huang, Z., Chung, W., Ong, T.H., & Chen, H. (2002). A graph-based recommender system for digital library. In *Proceedings of the 2nd ACM/IEEE-CS Joint Conference on Digital Libraries* (JCDL '02).

Hwang, S. Y., Hsiung, W. C., & Yang, W. S. (2003). A prototype WWW literature recommendation system for digital libraries. Online Information Review, 27:169-182.

International Standards Organization (ISO). (1986). *ISO 2788 Documentation – Guidelines for the establishment and development of monolingual thesauri*. Geneva: ISO.

International Standards Organization (ISO). (1985). *ISO 5964 Documentation -- Guidelines for the establishment and development of multilingual thesauri*. Geneva: ISO.

Jaidka, K., Khoo, C., & Na, J.-C. (2010). Imitating Human Literature Review Writing: An Approach to Multi-Document Summarization. In *Proceedings of ICADL2010*, pp 116-119.

Jenkins, C., Jackson, M., Burden, P., & Wallis, J. (1998). Automatic classification of Web resources using Java and Dewey decimal classification. *Computer Networks and ISDN Systems archive*. 30(1-7), 646-648.

Joint Steering Committee for Revision of AACR. (2002). *Anglo-American Cataloguing Rules*, 2nd Ed, 2002 Revision. Ottawa: Canadian Library Association.

Joorabchi, A., & Mahdi, A. E. (2009). Leveraging the Legacy of Conventional Libraries for Organizing Digital Libraries. In Agosti, Maristella et al. (Eds) *Proceedings of the 13th European Conference, ECDL 2009*, Corfu (Greece), 27 Sept.-2 Oct 2009 (pp. 3-14).

Joorabchi, A., & Mahdi, A. E. (2008). Development of a National Syllabus Repository for Higher Education in Ireland. In B. Christensen-Dalsgaard, D. Castelli, B. Ammitzbøll Jurik, & J. Lippincott (Eds.), *Research and Advanced Technology for Digital Libraries*, Lecture Notes in Computer Science, 2008, Volume 5173 (pp. 197-208).

Kan, M.-Y., & Klavans, J. L. (2002). Using Librarian Techniques in Automatic Text Summarization for Information Retrieval. In *Proceedings of JCDL'02*, July 13-17, 2002, Portland, Oregon, USA.

Kanhabua, N., & Nørvåg, K. (2008) Improving Temporal Language Models for Determining Time of Non-timestamped Documents. In B. Christensen-Dalsgaard, D. Castelli, B. Ammitzbøll Jurik, & J. Lippincott (Eds.), *Research and Advanced Technology for Digital Libraries*. Lecture Notes in Computer Science, 2008, Volume 5173 (pp. 358-370).

Kanhabua, N., & Nørvåg, K. (2010). Exploiting Time-based Synonyms in Searching Document Archives. In *Proceedings of JCDL2010, 10th ACM/IEEE-CS Joint Conference on Digital Librarie* (pp. 79-88)*s*, Surfer's Paradise, Australia, June 21-25, 2010.

Kastner, I. (2009). Developments in Information Retrieval: Part 1. *Library + Information Update*, Dec 2009, 17-19.

Kelly, B. (2004). Interoperable digital library programmes? We must have Q&A! In *Proceedings of the 8th European Conference on Digital Libraries*, pp. 80-85.

Krottmaier, H. (2002). Automatic references: Active support for scientists in digital libraries. In *Proceedings of the 5th International Conference on Asian Digital Libraries*, 254-255.

Krapivin, M., Autayeu, A., Marchese, M., Blanzieri, E., & Segata, N. (2010). Keyphrases Extraction from Scientific Documents: Improving Machine Learning Approaches with Natural Language Processing. In *Proceedings of ICADL2010*, pp. 102-111.

Lancaster, F. W. (2003). *Indexing and Abstracting in Theory and Practice, 3rd edition*. Champaign, IL: U of Illinois Press.

Li, N., Zhu, L., Mitra, P., & Giles, C. L. (2010). oreChem ChemxSeer: A Semantic Digital Library. In *Proceedings of JCDL2010, 10th ACM/IEEE-CS Joint Conference on Digital Libraries* (pp. 245-254), Surfer's Paradise, Australia, June 21-25, 2010.

Manolopoulos, Y., & Sidiropoulos, A. (2005).A new perspective to automatically rank scientific conferences using digital libraries. *Information Processing & Management*, 41, 289-312.

Mas, C.-P., Fuhr, N., & Schaefer, A. (2004). Evaluating strategic support for information access in the DAFFODIL system. In *Proceedings of the 8th European Conference on Digital Libraries*, 476-487.

Masterman, M., Needham, R.M., & Sparck Jones, K. (1958). The analogy between mechanical translation and library retrieval. In *Proceedings of the International Conference on Scientific Information* (pp. 917-935). Washington, D.C.: National Academy of Sciences – National Research Council, Vol. 2.

de Mello, C. A.B., & Rafael D. L. (1999). A Comparative Study on OCR Tools. In *Vision Interface '99*, 19-21 May, Trois-Rivières, Canada (pp. : 224-232). Retrieved October 7, 2010, from http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.12.2361.

Melucci, M. (2004). Making digital libraries effective: Automatic generation of links for similarity search across hyper-textbooks. *Journal of the American Society for Information Science and Technology*, 55, 414-430.

Meusel, R., Niepert, M., Eckert, K., & Stuckenschmidt, H. (2010). Thesaurus Extension using Web Search Engines. In *Proceedings of ICADL2010*, pp. 198-207.

Mittal, A., Gupta, S., Kumar, P. & Kashyap, S. (2005). A Fully Automatic Question-Answering System for Intelligent Search in E-Learning Documents. *International Journal on E-Learning*, 4(1), 149-166.

Morales del Castillo, J. M., Pedraza-Jimenez, A., Ruiz, A. A., Peis, E., & Herrera-Viedma, E. (2009). A Semantic Model of Selective Dissemination of Information for Digital Libraries. *Information Technology and Libraries*, 28(1), 21-30.

Mustafa el Hadi, W. (2004). Human Language Technology and Its Role in Information Access and Management. *Cataloging & Classification Quarterly* 37(1/2), 131-151.

Mutschke, P. (2003). Mining networks and central entities in digital libraries: A graph theoretic approach applied to co-author networks. In F. Pfenning, M. R. Berthold, H.-J. Lenz, E. Bradley, R. Kruse, & C. Borgelt (Eds.), *Advances in intelligent data analysis V* (*Lecture Notes in Computer Science*, 2810; pp. 155-166). Berlin: Springer.

Nazarenko, A., & Ait El Mekki, T. (2005). Building back-of-the-book indexes. *Terminology* 11(1), 199-224.

Nie, J.-Y. (2010). *Cross-Language Information Retrieval*. San Francisco: Morgan & Claypool Publishers.

Nunberg, G. (2009). Google's Book Search: A Disaster for Scholars. The Chronicle of Higher Education, Aug 31, 2009. http://chronicle.com/article/Googles-Book-Search-A/48245/.

Oard, D. W. et al. (2001). Multilingual Information Retrieval. In E. Hovy, N. Ide, R Frederking, J. Marian, & A. Zampolli (eds.), *Multilingual Information Management: Current Levels and Future Abilities* (pp. 223-256).

Oard, D.W. (1997). Serving users in many languages: Cross-language information retrieval. *D-lib Magazine*.

Ou, S., Khoo, C. S.G., & Goh, D. H.-L. (2009). *Automatic Text Summarization in Digital Libraries*. In Y.-L. Theng, S. Foo, D. Goh, & J.-C. Na (Eds). *Handbook of research on digital libraries: design, development, and impact* (pp. 1599-172). Hershey, PA: Information Science Reference, c2009.

Owen, T., Buchanan, G., Eslambolchilar, P., & Loizides, F. (2010). Supporting Early Document Navigation with Semantic Zooming. In *Proceedings of ICADL2010*, pp. 168-178.

Palmer, C.R., Pesenti, J., Valdes-Perez, R.E., Christel, M.G., Hauptmann, A.G., Ng, D., & Wactlar, H.D. (2001). Demonstration of Hierarchical Document Clustering of Digital Library Retrieval Results. In *Proceedings of JCDL'01* (p. 415), June 24-28, 2001, Roanoke, Virginia, USA.

Pereira, D. A., Ribeiro-Neto, B., Ziviani, N., Laender, A. H.F., Gonçalves, M. A., & Ferreira, A. A. (2010). Using web information for author name disambiguation. In *Proceedings of JCDL2009: 9th ACM/IEEE-CS Joint Conference on Digital Libraries* (pp. 49-58), Austin, TX,USA, June 15-19, 2009.

Perez-Carballo, J., & Strzalkowski, T. (2000). Natural language information retrieval: progress report, *Information Processing & Management* 36(1), 155-78.

Péter, J. (2004). Link-enabled cited references. *Online Information Review*, 28, 306-311.

Plamondon, R, & Srihari, S. N. (2000). On-Line and Off-Line Handwriting Recognition: A Comprehensive Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(1), 63-84.

Popescu, A., & Grefenstette, G. (2010). Spatiotemporal Mapping of Wikipedia Concepts. In *Proceedings of JCDL2010, 10th ACM/IEEE-CS Joint Conference on Digital Libraries* (pp. 129-138), Surfer's Paradise, Australia, June 21-25, 2010.

Pouliquen, B., Steinberger, R., & Ignat, C. (2003). Automatic Annotation of Multilingual Text Collections with a Conceptual Thesaurus. In *Ontologies and Information Extraction*. Workshop at EUROLAN'2003: The Semantic Web and Language Technology – Its Potential and Practicalities. Bucharest, 28 July – 8 August 2003.

Prabowo, R., Jackson, M., Burden, P., & Knoell, H.-D. (2002). Ontology-based automatic classification for Web pages: design, implementation and evaluation. In *Proceedings of the Third International Conference on Web Information Systems Engineering*, WISE 2002. (pp. 182 – 191).

Rasmussen, E. (2004) Information Retrieval Challenges for Digital Libraries. In *Proceedings of Digital Libraries: International Collaboration and Cross-Fertilization, 7th International Conference on Asian Digital Libraries*, ICADL 2004, Shanghai, China, December 13-17, 2004. Lecture Notes in Computer Science 3334. Berlin: Springer 2004.

Řehůřek, R., & Kolkus, M. (2009). Language Identification on the Web: Extending the Dictionary Method. In A. Gelbukh (Ed.), *Computational Linguistics and Intelligent Text Processing*, Lecture Notes in Computer Science, 5449 (pp. 357-368).

Ruch, P. (2003). *Applying Natural Language Processing to Information Retrieval in Clinical Records and Biomedical Texts*. Ph.D. thesis. Genève : Imprimerie des Hôpitaux Universitaires de Genève.

Saggion, H., & Lapalme, G. (2000). Concept Identification and Presentation in the Context of Technical Text Summarization. *Workshop on Automatic Abstracting, NAACL-ANLP2000*, Association for Computational Linguistics, Seattle, USA, 30 Apr 2000.

Salton, G. (1988). Syntactic approaches to automatic book indexing. In *Proceedings of the 26th annual meeting on Association for Computational Linguistics* (pp. 204-210), Buffalo, New York.

Sanderson, R., & Watry, P. (2007). Integrating data and text mining processes for digital library applications. In *Proceedings of the 7th ACM/IEEE-CS joint conference on Digital libraries* (pp. 73-79).

Sebastiani, F. (2002) Machine Learning in Automated Text Categorization. *ACM Computing Surveys*, 34(1), 1–47.

Seki, K., Qin, H., & Uehara, K. (2010). Impact and Prospect of Social Bookmarks for Bibliographic Information Retrieval. In *Proceedings of JCDL2010, 10th ACM/IEEE-CS Joint Conference on Digital Libraries* (pp. 357-360), Surfer's Paradise, Australia, June 21-25, 2010.

Smadhi, S. (2003). System of information retrieval in XML documents. In Shirley A. Becker (Ed.) *Effective databases for text & document management* (pp. 1-11), Hershey, PA: IGI Publishing.

Smeaton, A.F., & Callan, J. (2005). Personalisation and recommender systems in digital libraries. *International Journal of Digital Libraries*, 5(4), 299-308.

Soergel, D. (2009). Digital Libraries and Knowledge Organization. In S.R. Kruk, & B. McDaniel (Eds). *Semantic Digital Libraries* (pp. 3-39). Berlin: Springer.

Song, M., Song I.Y., Allen, R.B., & Obradovic, Z. (2006). Keyphrase extraction-based query expansion in digital libraries. In *JCDL '06: Proceedings of the 6th ACM/IEEE-CS joint conference on Digital libraries* (pp. 202-209), Chapel Hill, NC, USA, June 11-15, 2006.

Sparck Jones, K. (1967). Current work on automatic classification for information retrieval. *T.A. Informations*, 2, 92-96.

Sparck Jones, K. (2007). Automatic summarising: the state of the art. *Information Processing and Management*, 43(6), 1449-1481.

Spagnola, S., & Lagoze, C. (2011). Word Order Matters: Measuring Topic Coherence with Lexical Argument Structure. In *Proceedings of JCDL 2011*, Ottawa, Canada, June13-17, 2011.

Sugiyama, K., & Kan, M.-Y. (2010). Scholarly Paper Recommendation via User's Recent Research Interests. In *Proceedings of JCDL2010, 10th ACM/IEEE-CS Joint Conference on Digital Libraries* (pp. 29-38), Surfer's Paradise, Australia, June 21-25, 2010.

Steinbach, M., Karypis, G., & Kumar, V. (2000). A Comparison of Document Clustering Techniques. In *KDD Workshop on Text Mining*. Retrieved October 7, 2010 from http://glaros.dtc.umn.edu/gkhome/node/157.

Strzalkowski, T. (ed). (1999). *Natural Language Information Retrieval*. Dordrecht : Kluwer Academic Publishers.

Tahmasebi, N., Niklas, K., Theuerkauf, T., & Risse, T. (2010). Using Word Sense Discrimination on Historic Document Collection. In *Proceedings of JCDL2010, 10th ACM/IEEE-CS Joint Conference on Digital Libraries* (pp. 89-98), Surfer's Paradise, Australia, June 21-25, 2010.

Thompson, R., Shafer, K., & Vizine-Goetz, D. (1997). Evaluating Dewey concepts as a knowledge base for automatic subject assignment. In *Proceedings of the second ACM international conference on Digital libraries*, Philadelphia, Pennsylvania, United States, (pp. 37-46).

Todirasçu, A., & Rousselot, F. (2001). Ontologies for Information Retrieval. In *Proceedings of TALN 2001* (pp. 305-314), Tours, 2-5 July 2001.

Toms, E., & McCay-Peet,L. (2009). Chance Encounters in the Digital Library. In M. Agosti et al. (Eds.). *Research and Advanced Technology for Digital Libraries, 13th European Conference, ECDL 2009* (pp. 192-202), Corfu (Greece), 27 Sept.-2 Oct 2009.

Tönnies, S., Köhncke, B., Koepler, O., & Balke, W.-T. (2010). Exposing the Hidden Web for Chemical Digital Libraries. In *Proceedings of JCDL2010, 10th ACM/IEEE-CS Joint Conference on Digital Libraries* (pp. 235-244), Surfer's Paradise, Australia, June 21-25, 2010.

Treeratpituk, P., & Giles, C. L. (2010). Disambiguating authors in academic publications using random forests. In *Proceedings of JCDL2009: 9th ACM/IEEE-CS Joint Conference on Digital Libraries* (pp. 39-48), Austin, TX,USA, June 15-19, 2009.

Tsoumakas, G., & Katakis, I. (2007). Multi-label classification: An overview. *International Journal of Data Warehousing and Mining*, 3(3), 1-13

Vakkari, P., & Taneli, M. (2009). Comparing Google to Ask-a-Librarian Service for Answering Factual and Topical Questions. In M. Agosti et al. (Eds.). *Research and Advanced Technology for Digital Libraries, 13th European Conference, ECDL 2009* , Corfu (Greece), 27 Sept.-2 Oct 2009 (pp. 352-363).

Van Rijsbergen, C. J. (1979). *Information Retrieval*, Newton, MA: Butterworth-Heinemann.

Vizine-Goetz, D. (1996). *Using library classification schemes for Internet resources. OCLC Internet Cataloging Project Colloquium*. Retrieved October 8, 2010, from http://webdoc.sub.gwdg.de/ebook/aw/oclc/man/colloq/v-g.htm

Voorhees, E.M. (1999). Natural language processing and information retrieval. In M.T. Pazienza (Ed). *Information extraction. Towards scalable, adaptable systems* (pp. 32-48). Berlin: Springer-Verlag.

Witten, I.H., Don, K. J., Dewsnip, M., & Tablan, V. (2003). Textmining in a digital library. *International Journal on Digital Libraries*, 5:1-4.

Wang, J. (2009). An Extensive Study on Automated Dewey Decimal Classification. *Journal of the American Society for Information Science and Technology*, 60(11), 2269–2286.

Wu, P. H.-J., Na, J.-C., & Khoo, C. S. G. (2004). NLP versus IR approaches to fuzzy name searching in digital libraries. In *Proceedings of the 8th European Conference on Digital Libraries*, pp. 145-156.

Yang, Y. (1999). An Evaluation of Statistical Approaches to Text Categorization, *Information Retrieval*, 1(1-2), 69-90.

Yoo, I. (2006). A comprehensive comparison study of document clustering for a biomedical digital library MEDLINE. In *Proceedings of the 6th ACM/IEEE-CS Joint Conference on Digital Libraries* (pp. 220-229), Chapel Hill, NC, USA.

ADDITIONAL READING SECTION

Agosti, M., Borbinha, J., Kapidakis, S., Papatheodorou, C., & Tsakonas, G. (2009). In M. Agosti, J. Borbinha, & S. Kapidakis (Eds.), *Research and Advanced Technology for Digital Libraries. Proceedings of the 13th European Conference*, ECDL 2009, Corfu, Greece, September 27 - October 2, 2009. Lecture Notes in Computer Science Volume 5714. Berlin; Heidelberg : Springer-Verlag.

Andrews, J., & Law, D.G. (Eds) (2004). Digital libraries: policy, planning, and practice. Aldershot, Hants: Ashgate.

Archer, D. W., Delcambre, L. M. L., Corubolo, F., Cassel, L., Price, S., Murthy, U., Maier, D., Fox, E.A., Murthy, S., & McCall, J. (2008). Superimposed Information Architecture for Digital Libraries. *Research and Advanced Technology for Digital Libraries* (Proceedings of the 12th European Conference, ECDL 2008, Aarhus, Denmark, September 14-19), Lecture Notes in Computer Science, 2008, Volume 5173 (pp. 88-99). Berlin; Heidelberg : Springer-Verlag.

Bontcheva, K., Maynard, D., Cunningham, H., & Saggion, H. (2002). Using Human Language Technology for automatic annotation and indexing of digital library content. In *Proceedings of ECDL 2002 : European conference on research and advanced technology for digital libraries*, Rome , 2002, vol. 2458 (pp. 613-625). Berlin; Heidelberg : Springer-Verlag.

Buchanan, G., Masoodian, M., & Cunningham, S.J. (Eds.) (2008). In *Digital libraries, universal and ubiquitous access to information. Proceedings of the 11th International Conference on Asian Digital Libraries*, ICADL 2008, Bali, Indonesia, December 2-5, 2008. Berlin : Springer-Verlag.

Chengzhi, Z., & Dan,W. (2008). Concept Extraction and Clustering for Topic Digital Library Construction. In *Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology* (pp. 299-302).

Christensen-Dalsgaard, B.. Castelli, D., Jurik, B.A., & Lippincott, J. (Eds). (2008). *Research and Advanced Technology for Digital Libraries. Proceedings of the 12th European Conference*, ECDL 2008, Aarhus, Denmark, September 14-19, 2008. Lecture Notes in Computer Science Volume 5173. Berlin; Heidelberg : Springer-Verlag.

Ferro, N. (2009). Annotation Search: The FAST Way. In M. Agosti, J. Borbinha, & S. Kapidakis (Eds.), *Research and Advanced Technology for Digital Libraries. Proceedings of the 13th European Conference* (pp. 15-26), ECDL 2009, Corfu, Greece, September 27 - October 2, 2009. Lecture Notes in Computer Science Volume 5714. Berlin; Heidelberg : Springer-Verlag.

Golub, K. (2006). Using Controlled Vocabularies in Automated Subject Classification of Textual Web Pages, in the Context of Browsing. *TCDL Bulletin*, 2(2). Retrieved October 8, 2010 from http://www.ieee-tcdl.org/Bulletin/v2n2/golub/golub.html.

Kovács, L., Fuhr, N., & Meghini, C. (eds). (2007). *Research and advanced technology for digital libraries. Proceedings of the 11th European Conference on Digital Libraries*. Budapest, Hungary, September 16-21, 2007. Lecture Notes in Computer Science, Volume 4675. Berlin; Heidelberg : Springer-Verlag.

Lagoze, C., Payette, S., Shin, E., & Wilper, C. (2006). Fedora: an architecture for complex objects and their relationships. *International Journal on Digital Libraries*, 6(2), 124-138.

Mitchell, S. (2006). Machine assistance in collection building: new tools, research, issues, and reflections. *Information Technology and Libraries*, 25(4): 190-216.

Rydberg-Cox, J. A. (2006). *Digital libraries and the challenges of digital humanities*. Oxford: Chandos Publishing.

Shiri, A., & Molberg, K. (2005). Interfaces to Knowledge Organization Systems in Canadian Digital Library Collections. *Online Information Review*. 29(6), 604-620.

Soergel, D. (2002). A Framework for Digital Library Research. *DLIB Magazine*, 8(12). Retrieved October 8, 2010 from http://www.dlib.org/dlib/december02/soergel/12soergel.html.

Tedd, L.A., & Large, A. (2005). *Digital Libraries: Principles and Practice in a Global Environment*. Germany: K.G. Saur.

Tuominen, K., Talja, S., & Savolainen, R. (2003). Multiperspective digital libraries: The implications of constructionism for the development of digital libraries. *Journal of the American Society for Information Science and Technology*, 54, 561-569.

Witten, I. H., Bainbridge, D., & Boddie, S. J. (2001) Greenstone: open-source digital library software with end-user collection building, *Online Information Review*, 25(5):288-298.

KEY TERMS AND DEFINITIONS

Abstracting (or Summarization): the operation by which the subject contents of a document are expressed by a short, narrative-style text.

Classification: a system of organising library materials (books, periodicals, audiovisual materials, etc.) according to their subject. Also the process of attributing a class (or a call number) to a given information resource.

Citation analysis: the study of the frequency and patterns of citations to other works in articles and books.

Classification scheme: a descriptive system used for grouping together works on similar subjects. *See also* Classification.

Collection (or Document collection): set of documents selected and housed by a given information service for a specific user community.

Content processing: the set of operations performed on documents to describe their subject contents. This includes classification, indexing and abstracting (or summarization). The result is semantic metadata.

Controlled vocabulary: a carefully selected set of terms from a natural language, used to describe or index a document collection. It applies formal restrictions (singular number only, for example) as well as semantic restrictions (homonym disambiguation, grouping of synonyms, etc.). The vocabulary may include compound terms not usually found in language, such as "World War II – history".

Document description: a step of document management consisting of supplying descriptive metadata for a given resource.

Document management: a series of operations relevant to the use of a document collection: creation, selection, acquisition, description, content processing, organisation, storage and retrieval of documents.

Indexing: analysing the content of a document and assigning to it a small set of terms to represent its main topics; the terms are usually taken from a controlled vocabulary.

Information retrieval: the area of study concerned with searching for documents, for information within documents, and for metadata about documents.

Metadata: information about the subject contents of a resource (semantic metadata), such as keywords, or about its physical or external characteristics (descriptive metadata), such as author, date of publication or format.

Named entity: a natural language expression referring to a single entity in the world, such as persons' names, organisations, geographical locations, timestamps or other.

Thesaurus: a type of controlled language which makes explicit certain types of semantic relations among terms, namely hierarchy (hypernym-hyponym relation), equivalence (synonymy) and associative relations (covering various other semantic relations).