Meta-Data Extraction from Bibliographic Documents for Digital Library

Abdel Belaïd, Dominique Besagni

Metadata Extraction from Bibliographic Documents for Digital Library

A. Belaïd[1] and D. Besagni[2]

[1]University Nancy 2- LORIA, F-54506 Vandoeuvre-lès-Nancy, France
abelaid@loria.fr

[2]INIST 2, Allée du Parc de Brabois, F-54514 Vandoeuvre-lès-Nancy, France
besagni@inist.fr

Abstract

This chapter addresses the problem of automatic metadata extraction within digitized documents by retro-conversion techniques. The focus is on bibliographic documents as they are by nature a source of such metadata. They are strongly structuring for a digital library (DL), their automatic recognition presents an obvious interest. However as their origin is very different (references, citations, tables of content, index cards), a generic methodology is proposed for their structure. Based on a first morphological labeling of the text, it looks for syntactic elements (syntagmas) revealing the bibliographic field nature (title, authors, date, publication source, etc.). Depending on the case, the syntax is validated either by a given grammar or by occurrence analysis in the different document elements (i.e. several references in a bibliography, or articles in a table of content). In the later, the bottom-up procedure generates a structure model from the well-recognized elements and applies it on the rest. The modeling requires taking into consideration the inter- and intra-fields relationships. The experiments performed on different types of documents confirm the interest of this approach.

1. Introduction

The digital library (DL) [1] has become more and more a common tool for everyone, a trend accentuated by the success of the Web and the easy access to every kind of information. Among the most important DL projects [2], we can mention “Project Gutenberg”, the oldest producer of free electronic books [3], the “Million Book Project” [4], and more recently Google announced its intent to digitize the book collections from several famous universities (Michigan, Harvard, Stanford, Oxford) and from the New York Public Library [5].

The DL provides an information located in one specific place to anyone, anywhere in the world, as long as the information can be retrieved. Contrary to the

Web at large, the DL offers a more organized access to selected information which is often validated, filtered and structured. With this trend, documents not registered in electronic form will risk to become invisible. It is the Google effect: “if it isn’t in Google then it doesn’t exist!”. This electronic registration is not sufficient enough to define a DL: the document itself must be in electronic form which does not mean it is machine readable.

A good DL is not only a good document retrieval system but the content must be at the same time accessible as well by the machines than by the users responding to their multiple needs. There are different aspects revealing the DL qualities relative to the:

a) content: the more structured a document is, the more useful it is. Added to this, the quality of the metadata accompanying the document is essential.
b) organization: the more standardized a format is, the more usable and durable the document is.
c) updating and patrimony valorization: the main problem is not in the feeding of new digital document but in importing digital documents from other DLs, or in adding patrimonial non-electronic documents.
d) use: the DL function does not stop with the document consultation, but the more options the better.

Since all the documents are not specifically generated to enter a DL, we need cost effective tools such as OCR, retro-conversion, hypertextualization, and metadata extraction techniques. There again the content defines the approach. Depending on the expected quality, these techniques will need more (or less) adaptation and depth.

At the DL level, independently of the origin of the document (electronic or not) it is obvious that the most important elements in terms of organization and structure, are the metadata and the hyperlinks. Although the hyperlinks, made popular by the Web, are the “icing on the cake” because they improve the navigation functionalities, the metadata are more basic, even indispensable, because they include for example the catalogue.

Although some problems remain [6], we have now a generation of OCRs capable of extracting the content with a good quality close to 100%, the research tends towards systems dedicated to structure extraction. This structure, generally of an editorial or logical nature, obviously constitutes a first step towards metadata generation.

As the documents in general are described in a DL by at least their descriptive metadata, a straightforward use of a DL can be done by correspondence between the terms outlined in bibliographic documents (like bibliographic references, citations, cards, tables of contents, etc.) and the DL metadata. Depending on the structure finesse of documents in DL and the precision of the outlined terms (which in our case are roughly recognized by OCR) the metadata recognition can be considered as a “mapping problem” between the real metadata and the recognized terms. Proper mapping of the bibliographic reference with its actual content within a DL is a challenging task of research [7].

2. The users' needs

Considering the DL as a very structured document repository that is well organized and continuously updated, we can envisage its use for some important requests similar to which they are done on the Web. But contrarily to the Web, the use of bibliographical data may offer more possibilities in the use of common DL services such as:

- **Information retrieval:** This is related to a simple DL consulting. The major need of DL is to retrieve the actual document from DL based on approximated bibliographic terms (roughly recognized by OCR or provided by the user). This corresponds to about 70% of the real DL activities where the users (such as researchers, students, etc.) are always requesting the DLs to get actual content from secondary documents (bibliographic documents). Here the impact of a misrecognition is low since the search is limited to document consulting.
- **Technological watch:** The goal of technology watch is to exploit all available information that can give indicators about the environment of any firm or organization. Among the information that are at hand, bibliographic references are an interesting source of data for such a study. The contribution of bibliographic references is of course immediate if they are in electronic and structured form, bringing out rapidly the elements directly exploitable with the techniques of bibliometric analysis. But often this bibliographic information is not available in electronic form, thus turning the analysis of information into a time-consuming task. Then comes the problem of retro-converting the information from documents on various media that is a research field in itself. Here we make the hypothesis that more the information is repeated by different authors more it is important. The search of importance can be initiated by the analysis of local DL. Here the impact of a misrecognition is average since we will pass beside significant novelties which can momentarily affect the development of the company. This minor incidence can be hurdled by frequently repeating the watch procedure.
- **Research piloting and evaluation:** It indicates the importance of evaluation principles as well as issues related to the preparation. The more spread approaches have privileged technical parameters leading to the document fabrication and diffusion. The foundation in the 1960s at Philadelphia (USA) of the Institute for Scientific Information (ISI) by Eugene Garfield was instrumental in turning the citation in a unit of measure. Used at first only as a tool for information retrieval, the citation has become an important criterion because it allows distinguishing among different publications those which received the approbation of the scientific community. By the same token, the citation is also used to appraise scientific journals especially with the impact factor calculated as the average number of citations a paper receives over a period of 2 years. Here the impact of a misrecognition may be high, since we can affect the prestige of an individual or an institution by misrepresenting their scientific output.

According to the needs, documents should be structured in such a way that they will be available from a DL in an easy manner.

3. Bibliographic elements as descriptive metadata

For a long time, librarians redacted bibliographical records or indexes to describe the available documents. To refer to the computer lingo, these records constitute data which serve to describe other data (e.g. book contents): they are called metadata. We can encode some essential information about the documents in a clear fashion: title, author, date of publication, keywords, etc. Table 1 summarizes the several categories of metadata used in DL as reported in [8].

Table 1. Metadata typology

Type	Purposes	Examples	Implementations
Descriptive or Intellectual	Describe and identify information objects and resources	- Unique identifier (URL, access number, ISBN) - Physical attributes (media, size…) - Bibliographical attributes (title, authors, language, keywords…)	- PURL, MARC - Dublin Core - Controlled vocabularies as thesaurus.
Structural	- Facilitate navigation and display of information objects. - Give data on the internal structure of the information objects as page, section, chapter, index and ToC, - Describe the relationships between materials (e.g. picture B is inserted in manuscript A) - Link connected files and scripts (e.g. file A is the JPEG image of archive file B)	Structuring tags	- SGML - XML - Encoded Archival Description (EAD) - MOA2, Structural Metadata Elements Electronic Binding (Ebind)
Administrative	- Facilitate the management and treatment of electronic collections. - Include technical data on creation and quality control, - Include right management, access control and required use conditions.	Technical data as scanner type and model, resolution, bit size, colorimetric space, file format, compression, owner, copyright date, use limitations, license information, conservation consigns, etc.	MOA2, Administrative Metadata Elements National Library of Australia, Preservation Metadata for Digital Collections CEDARS

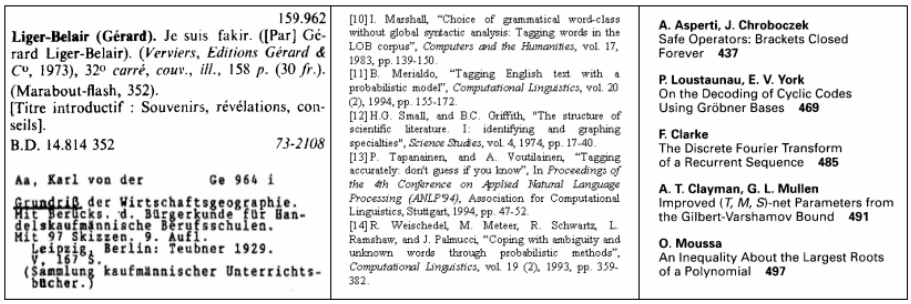
As the descriptive metadata are mandatory, their constitution is necessary for all types of documents.

- For raster image, it is mostly done manually even if some information (e.g. bibliographical record) can be imported from a specific database.
- For simple text (ASCII, Unicode), some elements (e.g. author, page, etc.) can sometimes be deduced by analyzing the text layout. For example, a centered nominal sentence in the beginning of the text can be interpreted as a title; a short text justified on the left and preceded by a number can be considered as a section title, under condition that this style is somewhat redundant in the text. On the other hand, as long as we have a text we can apply automatic indexing techniques to extract keywords.
- For structured text (e.g. all the structural elements are tagged), all the descriptive metadata are directly extractable. If the structure is well identified for the text class, then the automatic DIA techniques can be applied to mark-up this structure.

Document image analysis (DIA) can also be employed in the structural metadata extraction. For the internal structure, its contribution is similar to the one used for descriptive metadata. For the relationship between documents, DIA techniques can be used in some cases to find such links. For example, different entries of a table of content (ToC) can be related to each separate document. In case of citations in a scholarly work, a link can be generated at first between the body of the text and the bibliography section, as well as a second link between the citing document and the cited document.

4. Metadata extraction in bibliographic documents

Here, the structure granularity is relatively fine, reduced to some words and symbols. The punctuation plays an important role in its structure which is a source of problem if the recognition process is not efficient enough. Figure 1 shows three different kinds of bibliographic documents. We can notice the poverty of the text and the fact that the punctuation and the typographic style can be used as syntactic clues to separate the different fields.


159.962
Liger-Belair (Gérard). Je suis fakir. ([Par] Gérard Liger-Belair). (*Verviers, Editions Gérard & C°*, 1973), 32° *carré, couv., ill.*, 158 *p.* (30 *fr.*). (Marabout-flash, 352).
[Titre introductif : Souvenirs, révélations, conseils].
B.D. 14.814 352 *73-2108*

Aa, Karl von der Ge 964 i
Grundriß der Wirtschaftsgeographie. Mit Berücks. d. Bürgerkunde für Handelskaufmännische Berufsschulen. Mit 97 Skizzen. 9. Aufl.
Leipzig, Berlin: Teubner 1929.
V, 167 S.
(Sammlung kaufmännischer Unterrichtsbücher.)

[10] I. Marshall, "Choice of grammatical word-class without global syntactic analysis: Tagging words in the LOB corpus", *Computers and the Humanities*, vol. 17, 1983, pp. 139-150.
[11] B. Merialdo, "Tagging English text with a probabilistic model", *Computational Linguistics*, vol. 20 (2), 1994, pp. 155-172.
[12] H.G. Small, and B.C. Griffith, "The structure of scientific literature. I: identifying and graphing specialties", *Science Studies*, vol. 4, 1974, pp. 17-40.
[13] P. Tapanainen, and A. Voutilainen, "Tagging accurately: don't guess if you know", In *Proceedings of the 4th Conference on Applied Natural Language Processing (ANLP'94)*, Association for Computational Linguistics, Stuttgart, 1994, pp. 47-52.
[14] R. Weischedel, M. Meteer, R. Schwartz, L. Ramshaw, and J. Palmucci, "Coping with ambiguity and unknown words through probabilistic methods", *Computational Linguistics*, vol. 19 (2), 1993, pp. 359-382.

A. Asperti, J. Chroboczek
Safe Operators: Brackets Closed Forever **437**

P. Loustaunau, E. V. York
On the Decoding of Cyclic Codes Using Gröbner Bases **469**

F. Clarke
The Discrete Fourier Transform of a Recurrent Sequence **485**

A. T. Clayman, G. L. Mullen
Improved (*T, M, S*)-net Parameters from the Gilbert-Varshamov Bound **491**

O. Moussa
An Inequality About the Largest Roots of a Polynomial **497**

a) Catalog Cards b) References c) Table of Content

Figure 1: Examples of bibliographic elements

Coming back to the classical DIA schema, the system steps will be oriented more towards word detection than text recognition. This phenomenon will be explained in the following.

5. General overview of the work

Figure 2 illustrates the general overview of the mapping procedure. This schema is composed of three main elements: 1) the object document (e.g. document containing the keywords for searching, such as bibliographic references), 2) the search engine, 3) the target document (e.g. the documents to be retrieved from the DL) described below. This schema poses the real question for which the paper will give some answers: "how to make the link between the extracted terms in secondary

documents and the corresponding documents in a DL which have been differently and soundly structured?"

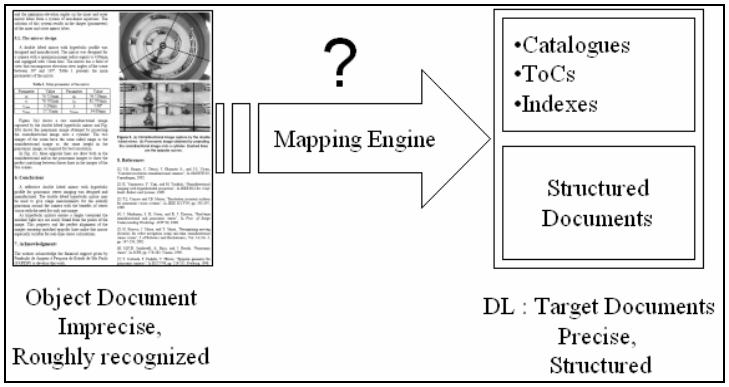


Figure 2: Overview of the mapping procedure

- **Object document:** contains some related information to the target document to retrieve the document from the DL properly. Considering the secondary documents, this information may contain author names, title, conference name, editor, etc. or other information regarding the date of publication, etc.
- **Target document:** corresponds to the DL and contains at the same time illustrative documents like catalogues, ToCs and Indexes, and structured documents.
- **Mapping Engine:** is a query interpreter that allows us to map the terms with document metadata and to find the closer document answering the query. The result accuracy is strongly tributary of the query precision and by consequence of the terms recognition accuracy.

In the following, we will try to instantiate this schema relative to the three needs.

6. Bibliographic element recognition for library management

Here we favour the meta-data like those proposed by Dublin Core (author name, date, conference title, editor, etc.) because we want to satisfy some easy bibliometric needs. The question often asked is: "I structure my document and I'd like that my colleague will retrieve it very quickly". This is also the attitude of some librarians offering a helpful tool for scientists to access very quickly to their collections by giving some keywords.

6.1 Catalogue retro-conversion

Several library programs that have been launched in the 1990's in Europe, placed emphasis on the enhancement and harmonization of machine-readable bibliographies and catalogues in Europe, thus contributing to the efficiency of libraries and improving resource sharing between them. Among the projects proposed, FACIT and MORE [9] were requested to determine the feasibility of converting older card catalogues into modern OPAC[1] using scanning, OCR, and automatic formatting into a bibliographic format, such as UNIMARC[2].

6.1.1. Object Document

The object document corresponds to the data as produced by OCR and arranged in XML. Character extraction is based on the use of commercially available OCR packages in the lower or middle price range. A combination procedure is applied on three OCRs looking for a substantial improvement of individual performances which are minimal on this kind of material. We have used the Myers's algorithm, based on an optimal dynamic programming matching [10]. The result is represented in XML where elements correspond to the text words and attributes are related to their topographical and lexical properties.

6.1.2 Target Document

The target document is related to the UNIMARC format which is a generalization of MARC (acronym for Machine Readable Catalogue or Cataloguing). Initially, UNIMARC was used for the exchange of records on magnetic tape but has since been adapted for use in a variety of exchange and processing environments. The fields, which are identified by three-character numeric tags, are arranged in functional blocks. These blocks organize the data according to its function in a traditional catalogue record. In the table below, fields 0--1-- hold the coded data while fields 2--8-- contain the bibliographic data:

Block	**Example**
0-- Identification block	010 International Standard Book Number
1-- Coded information block	101 Language of the work
2-- Descriptive information block	205 Edition statement
3-- Notes block	336 Type of computer file note
4-- Linking entry block	452 Edition in a different medium
5-- Related title block	516 Spine title
6-- Subject analysis block	676 Dewey Decimal Classification
7-- Intellectual responsibility block	700 Personal name - primary intellectual responsibility
8-- International use block	801 Originating source
9 - Reserved for local use	

[1] On line Public Access Catalogue

[2] Universal Machine-Readable Cataloguing

In addition to the 9-- block any other tag containing a 9 is available for local implementation. The fields defined by UNIMARC provide for different kinds and levels of information. This can be shown by looking at a typical record in the UNIMARC format. Figure 3 illustrates the UNIMARC conversion on one example.

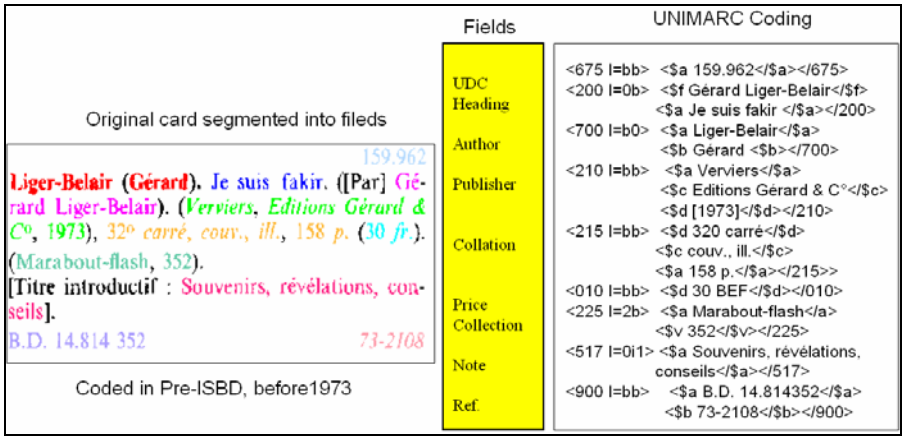


Figure 3: Example of bibliographic card (left side) and its UNIMARC representation (right side)

The target document is represented by a model which is formally described by a context-free grammar. The format of a production rule is as follows:

TERM ::=	CONSTRUCTOR SUBORDINATE_OBJECTS[QUALIFIER] | CONSTANT | TERMINAL
CONSTRUCTOR ::=	TD-SEQ | LR-SEQ | SEQ | AGGREGATE | CHOICE | IMPORT

The constructor describes the element arrangement among sequence (SEQ) (TD for top-down, LR for left-right) and AGGREGATE. The constructor "IMPORT" is used to inherit for the term some or the total description of another existent and similar term. As in XML DTDs, quantifiers such as “optional” (?), “repetitive” (+) or “optional-repetitive” (*) are used to specify the terms occurrences. Furthermore, because of the weakness of the physical structure and the multitude of choices represented in the model, some attributes, given by the library, are added to the previous description. These attributes are associated to each rule in the following format. The following example describes the term “TITLE” as a logical sequence of two objects: PROPER-TITLE and REST-OF-TITLE where the style is not italic (may be bold or standard), which is located at the beginning of the line (BEGLINE) and whose separator is a comma.

TITLE ::=	SEQ PROPER-TITLE REST-OF-TITLE
STYLE	-ITALIC
POSITION	BEGLINE

SEP	COMMA

In the following table, the optional object PARTICULE (which weight is A) is more important than LCAP (which weight is G). This specification is logical since an optional object normally helps in reinforcing the possible presence of a term more than an object that is always present.

ZPB ::=	SEQ LCAP RP PARTICULE?
WEIGHT	PARTICULE A LCAP G

In the following, the production rule describes a choice between two terms neither of which should contain any of the strings in the lexicon Abn (expressed by the attribute Clex). There are two actions which will be activated during the rule analysis. The first one indicates to verify before the rule analysis that the search zone does not contain the string "fr.". In the event of the hypothesis being verified, the second function is executed to create a UNIMARC tag before restituting the result in the required format.

FIP ::=	CHO FIP1 IPA
CLEX	-ABN
ACTION	+VERIFYSTRINGINFIELD(FR.,FALSE) RESTITUTE(215,BB)

6.1.3. Research Engine

The research engine corresponds to a syntactic analyzer whose role is to structure the input data according to the model description. The structure recognition is preceded by Index recognition (e.g. author names or title terms existing at the end of the catalogue). Then, structural analysis operates by hypothesis verification. Weights associated to the attributes contribute to the hypothesis ranking and strategy ordering during the analysis. The UNIMARC code is generated thanks to some special remedial functions, prepared by the librarians and establishing the different translation rules. These functions, associated with some terms in the model, try to establish the correspondence between the real structure and the recognized structure. The objective of this approach is to introduce qualitative "reasoning" as a function of the recognition evaluation. This evaluation allows: 1) the reduction of errors and ambiguities dues to faulty data OCR errors, data not fitting the model specification, etc.), 2) taking into account what is important to recognize, the qualitative evaluation of the obtained solutions, 3) the isolation and separation of doubtful areas.

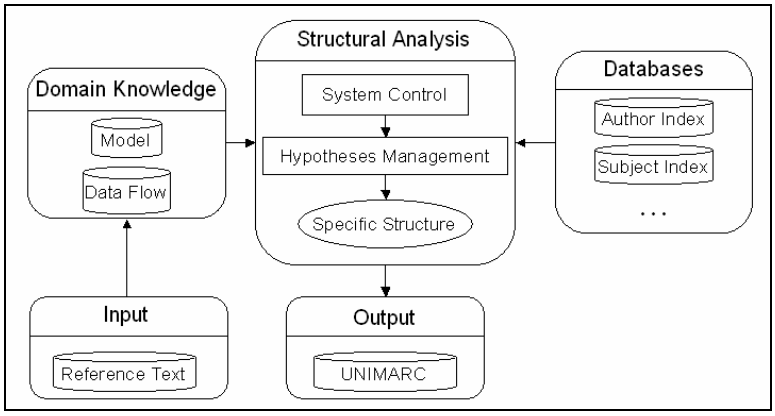


Figure 4: System overview of card recognition

6.2. ToC recognition for electronic consulting of scientific papers

Tables of Content (ToC) are the most synthetic meta-data and most introductory with the contents. They provide rapid indicators on the main document components. Their automatic identification may help to a rapid document organization and consulting in Digital Library.

6.2.1 Object Document

A textual ToC is a document composed of one header, a list of sections, article references and footnotes (see Figure 1.c). An article reference is made of an article title, a list of authors and a page number, written in one, two or three separate columns. Article references may have high or low complicated structure, but must be separated from each other. Sections corresponding to session names are given to group of papers. Although the structure is straightforward, some peculiarities can introduce ambiguities during recognition. For example: 1) the titles can contain author names, 2) the author names can be confused with common nouns, 3) the fields can follow each other and their separation is done by a random punctuation, 4) the successive articles can follow without a real distinctive sign, only the context or the semantics can help.

An automatic processing of ToC needs to recognize the content and to structure it in order to recover the article fields. For content recognition, OCRs are often used and an XML file is always generated highlighting the terms and their typographical and lexical attributes.

6.2.2 Document Target

The document target is usually very simple, corresponding to gathering of two or three fields. The difficulty comes from matching authors' names when their syntax

can change from one document to an other. For example, the first name can be given in full, the middle initial or an hyphen can be missing.

6.2.3 Mapping Engine

The mapping engine corresponds to two operations: ToC recognition and article searching in the database from the fields recognized in the ToC.

Few ToC recognizers have been proposed in the literature. On the one hand, Takasu et al. [11,12] proposed a system named CyberMagazine, based on image segmentation into blocks and syntactic analysis of their contents. The article recognition combines the use of decision tree classification and a syntactic analysis using a matrix grammar. On the other hand, Story and O'Gorman [13,14] proposed a method that combines OCR techniques and image processing.

We proposed a method based on text coding which in turn is based on Part of Speech tagging (PoS) [15]. The idea of this method, employed in language processing and text indexing, is to reassemble nouns in nominal syntagmas (syntactic element) representing the same information (see Figure 5). The nouns are given by a specific morphological tagging. This method can be applied in ToC recognition for article field identification by reassembling in the same syntagma "Title" or "Author", words having similar tags. The process of tagging consists of three stages: tokenization, morphological analysis, and syntactical grouping and disambiguation. The tokenizer isolates each textual term and separates numerical chains from alphabetic terms. The morphological analyzer contains a transducer lexicon. It produces all the legitimate tags for words that appear in the lexicon. If a word is not in the lexicon, a guesser is consulted. The guesser employs another finite-state transducer which examines the context and decides to assign the token to "Title" or to "Author" depending on prefixes, inflectional information and productive endings that it finds.

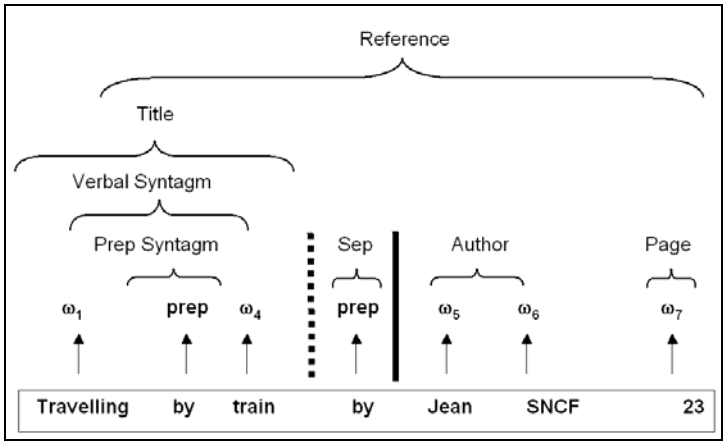


Figure 5: Syntagm extraction principle, where : ω_i : words, prep : proposition

As an application of ToC recognition system, to the best of our knowledge Calliope [15] is the most straightforward application used for article consulting. Calliope allows the management of distributed scientific documentary resources (related to computer science and applied mathematics) through Internet, composed of papers or electronic documents circulating among a community of researchers.

Calliope operating principles include: 1) consultation via navigation of a list of scientific journals (about 650 titles) as well as their corresponding tables of contents, 2) consultation of new tables of contents, further to a weekly updating of the server, 3) a personalized subscription by electronic mail, to electronic tables of contents of selected journals; the user will thereby receive the contents of the latest published issue. At any time the user is able to consult his list of subscriptions, cancel some or other, or subscribe to new journals, 4) the search by word (title of article or journal, or author's name).

As illustrated in Figure 6, Calliope is based on Rank Xerox XDOD (Xerox Document On Demand) software for scanning, and DocuWEB, the XDOD WEB server, for display and printing of articles. The images of scanned documents are stored and managed on several servers of documents, not necessarily near the scanning sites. The electronic tables of contents supplied by OCR-DIA are reformatted in HTML before being integrated into the Calliope WEB server.

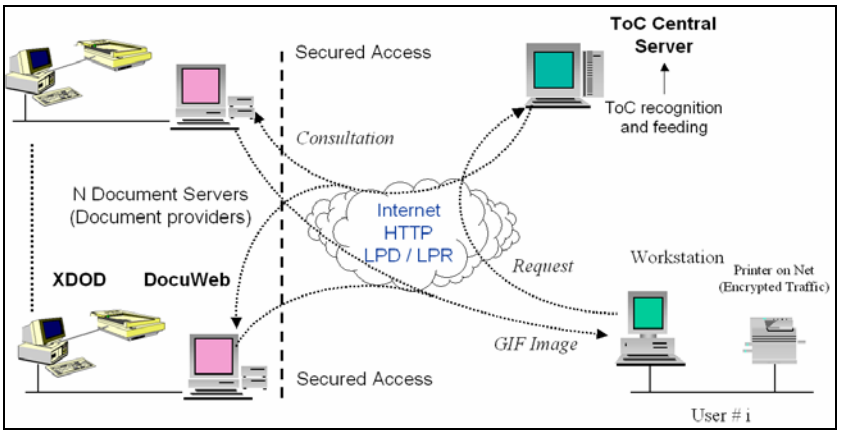


Figure 6: The scientific paper server: Calliope

7. Bibliographic reference structure in technological watch

The technological watch based on bibliographic references consists in analyzing the citations at the end of scientific papers or on the Web and in deducing interesting information in terms of new topics, famous authors, good conferences. The investigation is more oriented towards the search of stability of selected elements alone and always within the same context, than to the exhaustive recognition of all the reference. The syntactic methods, as shown before, are not adapted. F. Parmentier [16] proposed an AI based approach for bibliographic term extractions. The approach is based on a conceptual model.

7.1. Object document

Bibliographic references correspond to the citations at the end of scientific papers. They are extracted from existing databases or searched on the Web. The basic format can be the one used in BibTeX but the real structure can change according to the writer and the editor.

The model is determined from a reference database in BibTeX format (see Figure 7) and from the printed file of this database, using LaTeX.

```
@ARTICLE{joseph92a,
       AUTHOR = {S. H. Joseph and T. P. Pridmore},
       TITLE = {Knowledge-Directed Interpretation of Mechanical
          Engineering Drawings},
       JOURNAL = {IEEE Transactions on PAMI},
       YEAR = {1992},
       NUMBER = {9},
       VOLUME = {14},
        PAGES = {211--222},
       MONTH = {September},
       KEYWORDS = {segmentation, forms},
       ABSTRACT = {The approach is based on item extraction}
}
```

Figure 7: reference in BibTeX format

In order to have at our disposal all the possible optional fields, we generate automatically an artificial base using a BibTeX reference structure where field contents are replaced by field names. This reference has been used for the model construction. Each reference is then formatted in Postscript by BibTeX/LaTeX and converted into XML. Then, all that is needed is to search for the field contents in order to locate them, and deduce the separators. Figure 8 represents the result of the application of LaTeX in plain printing style on the artificial reference of Figure 7. It can be noticed that some fields are missed such as "editor" and "number".

7.2 Target Document

We used a concept network in order to represent the generic structure. The concept network is composed of nodes and links. Nodes are divided into two categories: generic, corresponding to the structure components, and specific, dealing with examples contained in the database. The links represent the conceptual proximity of the nodes (e.g. author and co-authors in the AUTHOR field, VOLUME and NUMBER, words in the TITLE, etc.). Figure 9 gives a global view of an example of such concept network.

7.3. Retrieval Engine

As the model is more conceptual than syntactic, translating a less rigorous manner of structure produced by human, the analysis approach is based on the extraction and the validation of terms. Fields are validated by studying the coherence of each term with its neighbors in the same field and with terms in the other fields. The system architecture is drawn from the work of Hofstadter and Mitchell [17,18] on emergent systems. In this system, knowledge is represented by a slip net, and evolves dynamically during the treatment in order to adapt the architecture to the given problem. The evolution is managed by a mechanism of propagation of activation pointing out the more pertinent concepts. These concepts (called "emerged") lead to the execution of specific agents.

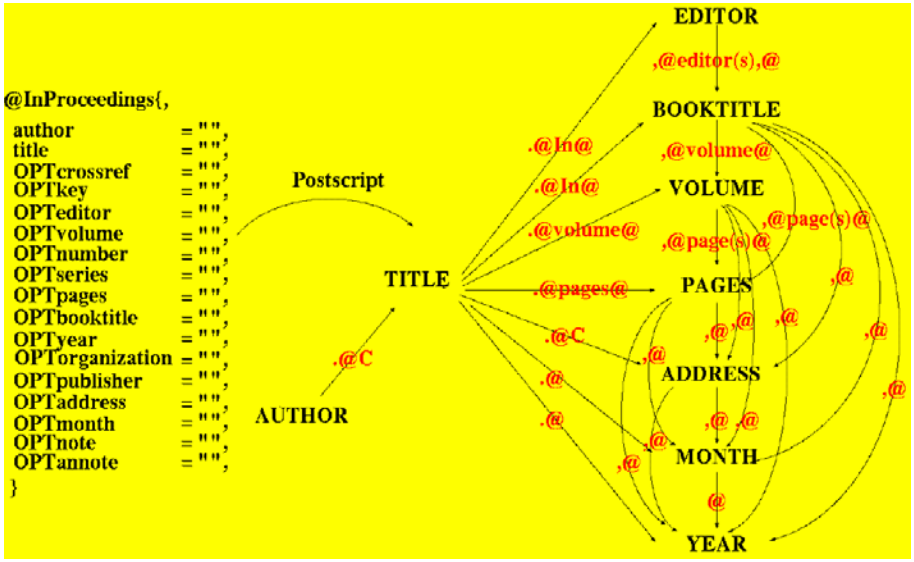


Figure 8: The physical model of BibTeX references

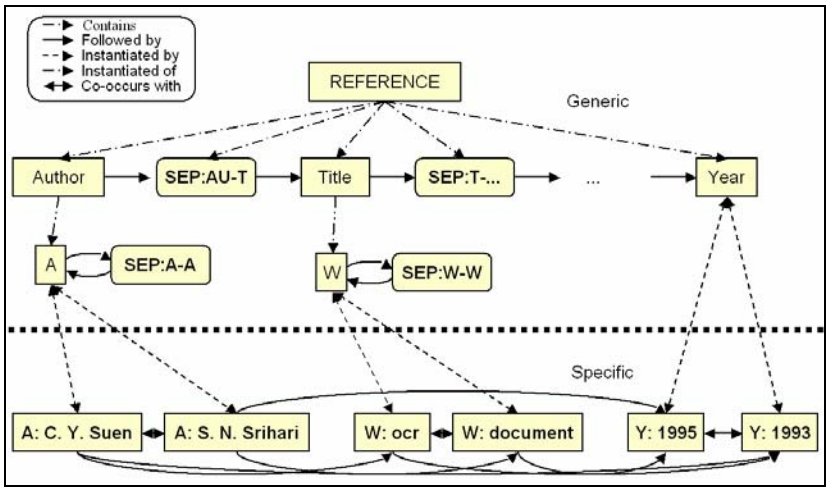


Figure 9: The concept model

A. Belaid and A. David showed in [18] how Information Retrieval techniques can be used at the same time to enhance the recognition process and to speed up the document retrieval. Figure 10 illustrates this principle. From the document structures, an Object Oriented Database (OOD) is generated. OOD allows the implementation of the concept of class from which object instances can be created. The OOD query method developed is what we call *classification* with *constraints*.

The method used is by attaching a set of instructions to each marker. The end doc marker invokes the document instance method for the generation of inverse list for each attribute which that can be used for request formulation and for calculus of indicators.
Given an attribute, the inverse list, for that attribute, gives the objects in which the given value is assigned to the attribute. He has adopted the approach that consists in transforming the logical and physical structures into their XML equivalents.

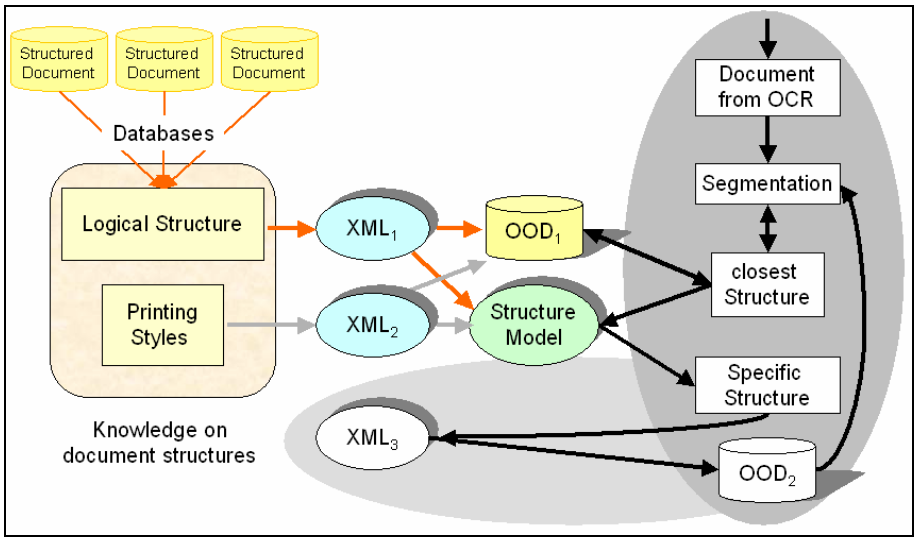


Figure 10: IRS techniques for DIA and document retrieval. XML1,2 are the normalized structure obtained from the databases and XML3 corresponds to the specific structure. OOD2 is the intermediate database for recognition.

8. Citation analysis in research piloting and evaluation

8.1 Impact factors

The measure of impact factors in bibliometrics is also operated on the bibliographic references present at the end of a scientific publication (article, conference, book, etc.) that refer to the work of an author cited in the body of the text. The difference between “citation” and “reference” being just a difference of perspective: for the citing author, it is a “reference” to the cited author; for the cited author, it is a “citation” by the citing author.

The Institute for Scientific and Technical Information (INIST) of the French National Center for Scientific Research (CNRS) has begun an experiment to digitize these bibliographic references especially because of the interest of citations in

bibliometrics and/or scientometrics. In order to analyze the scientific production with objective indicators, the measurement of that production was soon reduced to either scientific (articles, conferences, reports …) or technological (patents) publications. The first obvious indicator was the number of publications, but soon the citations were preferred because they have the advantage of representing an endorsement by the scientific community at large. That indicator gives a measure of the impact of a study, a laboratory or even a country in a particular scientific domain. Likewise, by analyzing citations from journals to journals, their impact can be assessed. It is the impact factor that is defined and supplied by the Journal of Citation Reports (JCR) from ISI.

Another use for citations is the analysis of relationships within a scientific community by a co-occurrence analysis of those citations. The main method: co-citations analysis developed by Small [20] measures the likeliness of cited documents by the number of documents citing them together. A clustering of these co-citations allows to identify islands within the mainstream scientific literature that define the research fronts. That co-citation analysis can also be used with authors instead of documents.

At the present time, all these citations are supplied by a single database which is the Science Citation Index (SCI) from ISI. Therefore, biases exist, especially:

- Only journal articles are treated (neglecting conference proceedings, reports, doctoral dissertations …);
- It strongly favors publications in English in general and US one in particular;
- Some domains as physical sciences are better covered than others as engineering, and some domains are altogether neglected as humanities.

The interest of our work is to propose a method that allows INIST to fill part of these shortcomings. We are especially looking for:

- Adding links between citing and cited documents to the bibliographic records of INIST ’s own databases PASCAL and FRANCIS for the purpose of information retrieval;
- Supplying citations for scientific journals and domains not considered by ISI.

8.2 Methodology

8.2.1. Object Document

Contrary to what is described in section 7, the bibliographic references used in this experiment come from actual articles from scientific (i.e. pharmaceutical) journals. After recognizing by OCR, all the references from the same article are in a XML file with each reference individualized. However, the different parts of those references (authors, title, journal, date ...) are not identified. The character set used in the data files is ISO Latin-1 (standard IS0 8859-1). The other alphabetic characters not

belonging to that character set are represented as character entities as defined in the SGML/XML standards, i.e. "**Ş**" for the character "**Ş**".

The problems we encountered while trying to segment the bibliographical references are of several orders:

- Problems due to the digitization: non-recognized characters, badly recognized characters (e.g. the uppercase letter **D** that sometime gives the uppercase letter **I** followed by a right parenthesis) or missing characters (mostly punctuation marks);
- Problems due to the heterogeneity of the data: the structure of a reference depends on the type of the cited document and on the origin of the citing article since the model of the citation depends on the journal where the article is published. Although on this last point, it must be said that all the journals do not enforce their own rules with the same rigor and that the form of the references may vary from one article to the other in the same issue of some journals;
- Other problems: one must add the typing errors, the omissions and sometime the presence of notes that have nothing to do with bibliographic references.

8.2.2. Target Document

For the time being, we try to identify the different fields of references from journal articles because they are more frequent (at least in the chosen domain) and are the usual subject of citation analysis. Although we use a bottom-up data-driven method, we have some generic model of the structure of a reference, e.g. when given, author names are at the beginning of the reference or the date of publication can only be in one of three positions: after the authors or the journal title or the page numbers.
Also helpful the fact that an author will respect the form of a reference type within a given article. Therefore, not only can we check the validity of the segmentation, we can also create a model for each set of references to correct and validate the final result.

8.2.3. Mapping Engine

In the literature, we identified a work done at the NEC Research Institute as part of the CiteSeer system [21]. The Autonomous Citation Indexing (ACI) uses a top-down methodology applying heuristics to parse citations. This approach employs some invariants considering that the fields of a citation have relatively uniform syntax, position and composition. It uses trends in syntactic relationships between fields to predict where a desired field exists if at all.

Even though this method is reported to be accurate, its functioning is not described enough explicitly to measure its efficiency on OCR output.

Similar works has been done in the field of mathematics to link together retro-converted articles in specialized databases looking for known patterns of author names, invariants and journal titles from a predefined list [22,23].

Conversely, we prefer not to do the retro-conversion of bibliographic references in a top-down way guided by a structure generic model. From one document to the

other, the structure of the citations shows a great variability, too important to reuse the same model. So we propose a bottom-up data-driven methodology. It is based on locally studying the common structure of all the references written in the same bibliographic document and adapting accordingly the heuristic rules. The methodology of retro-conversion of bibliographic references is based on exploiting two essential particularities: regularity of the structure and redundancy in the same document.

The proposed approach is also based on part-of-speech tagging as used for ToCs. However since the citations have a more complete structure than the ToC articles, the structure analysis is more investigated in this case. The different lists we use came from electronic resources from INIST like the Pascal database for author names, journal titles and country names and from electronic dictionaries for English and French words as well as prepositions.

That analysis is done by studying the regularity and the redundancy of specific elements of the text or by grouping some elements together. In the first case, we define qualitatively or quantitatively the chosen element, i.e. the date of publication, for each reference in function of its type but also in function of its position and of the characteristics of the elements, words or punctuation marks, surrounding it. The regularity and the frequency of these characteristics on the whole set of references from the same article allow us to locate and tag that element in most of the references. In the second case, the search is done by gathering words from the text in function of some rules:

- reduction rules: these rules are used to group consecutive identical tags. For example, the grouping of two initials (**IT**) from an author's name or two proper names (**PN**) while taking into account the type of punctuation marks (**PU**) that may be present in such a context.
- forming rules: these rules are used to initiate the beginning of a field by associating tags that are complementary in their description. For example, the association of the name and the initials in the formation of an author's name (**AU**).
- extending rules: these rules are used to concatenate sub-fields recognized independently. For example, an author and the expression "*et al.*" (**EA**) which confirms the field "**AU**" or the expansion of the article title from an initial core composed of three common nouns by adding nouns, connectors (**CC**) and prepositions to let the field grow as much as possible.
- agglutination rules: these rules are used to allow the unknown terms (**UN**) to be absorbed if the conditions are right. For example, between two author names, an unknown term should be absorbed.
- mixed rules: these rules are used to combine a set of grouping rules to detect potential candidates before using regularity to select the best amongst them. It is notably the case with page numbers (**PG**) that can be made of numbers (**NU**) and/or alphanumeric strings (**AN**).

Reduction rules	IT + IT => IT IT + PU- + IT => IT
Forming rules	IT + PN => AU IT + PU, + PN => AU
Extending rules	AU + CC + AU => AU AU + EA => AU
Agglutination rules	UN + AU => AU UN + PU, + AU => AU
Mixed rules	NU + PU- + NU => PG AN + PU- + AN => PG

Figure 11: Typology of grouping rules with some simple examples.

At the end of that process, each term or group of terms gets a new tag which gives a more explicit identification of the field to which it belongs. But that kind of analysis shows some limits, so the next step consists in exploiting what has been well recognized to create a model of how the different fields of the reference are supposed to be joined together (inter-field model) and how some of them (e.g. authors, journal title) are structured (intra-field model). Then, both models are used in turn to complete and to correct the faulty references. As shown in Figure 13, the model extracted in Figure 12 is used to extend the title field up to the right parenthesis and to deduce the journal title from its position.

Reference with highlighted fields
BEAN, B.P. (1985). Two kinds of calcium channels in canine atrial cells. Differences in kinetics, selectivity and pharmacology. J. Gen. Physiol., 86, 1 30.
Corresponding inter-field model
AU (DA) . TIP . JN , VOL , PG

Figure 12: Inter-field modelling

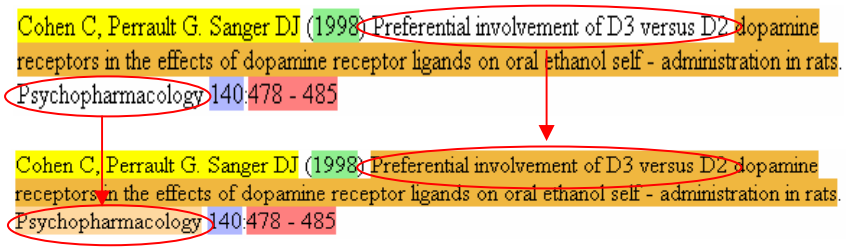


Figure 13: Example of inter-field correction

Figure13: Example of inter-field correction

9. Conclusion

As the digital libraries grow more complex and more numerous, there is a need for methods of metadata extraction to improve information retrieval and navigability within the database. Such methods can also add new usage to existing DLs as extracting information from bibliographic references for bibliometric analyses. The methods described here can not only be improved, but must be considered as part of an iterative process. The more information you extract, the more possibilities you have to refine the whole process: e.g., collecting the different models for a bibliographic references help find the correct model for a set of references which is too small or contains too many errors for the statistical analysis to be efficient.

References

1. http://www.dlib.org/metrics/public/papers/dig-lib-scope.html
2. http://en.wikipedia.org/wiki/List_of_digital_library_projects
3. http://www.gutenberg.org
4. http://www.library.cmu.edu/Libraries/MBP_FAQ.html
5. http://www.google.com/googleblog/2004/12/all-booked-up.html
6. H. S. Baird. “Difficult and Urgent Open Problems in Document Image Analysis for Libraries”, First International Workshop on Digital Image Analysis for Libraries (DIAL 2004), pp. 25-32, 2004.
7. A. Belaïd, “Retrospective document conversion: application to the library domain”, International Journal on Document Analysis and Recognition, vol. 1, 1998, pp. 125-146.
8. http://www.library.cornell.edu/preservation/tutorial/metadata/table5-1.html
9. http://www.cordis.lu/libraries/en/projects.html
10. W. Miller and E.W. Myers, A File Comparison Program, Software, Practice and Experience, Vol. 15, No. 11, pp. 1025-40.
11. A. Takasu, S. Satoh, E. Katsura, "A Document Understanding Method for Database Construction of an Electronic Library", International Conference on Pattern Recognition, pp. 463-466, 1994.
12. A. Takasu, S. Satoh, E. Katsura, "A Rule Learning Method for Academic Document Image Processing", International Conference on Document Analysis and Recognition, ICDAR'95, Vol. I, pp. 239-242.
13. L. O'Gorman. "Image and Document Processing Techniques for the Right Pages Electronic Library System", International Conference on Pattern Recognition, Vol. 2, pp. 260-263, 1992.
14. G. A. Story, L. O'Gorman, D. Fox, L. Levy Schaper, H. V. Jagadish, "The Right Pages Image-Based Electronic Library for Alerting and Browsing", Computer, 25, No. 9:17–26, September 1992
15. A. Belaïd, “Recognition of table of contents for electronic library consulting”, International Journal on Document Analysis and Recognition, vol. 4, 2001, pp. 35-45.

16. F. Parmentier and A. Belaïd, Logical Structure Recognition of Scientific Bibliographic References, International Conference on Document Analysis and Recognition (ICDAR'97), Vienna, Austria, August 1997, pp. 1072-1076.
17. D. R. Hofstadter and M. Mitchell. The Copycat Project: A Model of Mental Fluidity and Analogy-Making. In J. Barnden and K. Holyoak, editors, Advances in Connectionnist and Neural Computation Theory, volume 2, pages 31–112. Norwood, New Jersey: Ablex Publishing Corporation, 1994.
18. M. Mitchell. Analogy-Making as Perception : A Computer Model. MIT Press, 1993.
19. A. Belaïd and A. David, "The Use of Information Retrieval Tools in Automatic Document Modelling and Recognition", Workshop on Document Layout Interpretation and its Applications, Bangalore, India, DLIA'99, pp. 522-526, 1999.
20. H.G. Small, and B.C. Griffith, "The structure of scientific literature. I: identifying and graphing specialties", Science Studies, vol. 4, 1974, pp. 17-40.
21. S. Lawrence, C. L. Giles, and K. Bollacker, "Digital Libraries and autonomous Citation indexing", IEEE Computer, vol. 32 (6), 1999, pp. 67-71.
22. K. Dennis, G.O. Michler, G. Schneider, and M. Suzuki, "Automatic reference linking in digital libraries", Workshop on Document Image Analysis and Retrieval (DIAR'03), Madison, Wisconsin, June 21, 2003. <http://www.exp-math.uniessen.de/algebra/retrodig/digili b2.pdf>
23. K. Kratzer, "Automatic reference linking by means of MR lookup", Workshop on Linking and searching in distributed digital libraries, Ann Arbor, Michigan, March 18-20, 2002. <http://www.exp-math.uni-essen.de/algebra/veranstaltungen/kratzer.pdf>