Citation recognition for scientific publications in digital libraries

Dominique Besagni, Abdel Belaïd

Citation recognition for scientific publications in digital libraries

Dominique Besagni[1] and Abdel Belaïd[2]
[1]*URI, INIST-CNRS, 2 Allée du parc de Brabois, 54514 Vandœuvre-lès-Nancy Cedex, France*
[2]*LORIA, Campus scientifique, BP 239, 54506 Vandœuvre-lès-Nancy Cedex, France*
besagni@inist.fr, abelaid@loria.fr

Abstract

In this paper, a method based on part-of-speech tagging (PoS) is used for bibliographic reference structure. This method operates on a roughly structured ASCII file, produced by OCR.. Because of the heterogeneity of the reference structure, the method acts in a bottom-up way, without an a priori model, gathering structural elements from basic tags to sub-fields and fields. Significant tags are first grouped in homogeneous classes according to their categories and then reduced in canonical forms corresponding to record fields: ``authors'', "title", "conference name", "date", etc. Non labeled tokens are integrated in one or another field by either applying PoS correction rules or using a inter- or intra-field model generated from well-detected records. The designed prototype operates with a great satisfaction on different record layouts and character recognition qualities. Without manual intervention, 96.6% words are correctly attributed, and about 75,9% references are completely segmented from 2,575 references.

1. Introduction

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.

Concerning the measure of impact factors in bibliometrics, the bibliographic references we study are those present at the end of a scientific publication (article, conference, book …) that refer to the work of an author cited in the body of the text. The term "citation" is also used, the difference between "citation" et "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.

LORIA is associated to this experiment because of their study in methods of retro-conversion adapted to bibliographical resources. This work represent for LORIA another experience in retro-converting micro-structural documents that started with bibliographical records on index cards [1] and continued with the recognition of tables of content to feed the document server Calliope [2].

2. Citations

Citations correspond to the bibliographic references present at the end of a scientific publication. It is a structural element of a standard scientific article that can be used for analysis.

The foundation by Eugene Garfield in the 1960s of the Institute for Scientific Information (ISI) at Philadelphia (USA) was instrumental in the use of citations as a unit of measure. At first, those citations were used exclusively for information retrieval. They were deemed more objective than keywords either extracted from the text or proposed by an indexer. Nowadays in some databases of electronic publications, hypertext links are established between the references in a scientific article to the text of the corresponding publications.

In order to analyze the scientific production with objective indicators, the measurement of that production was soon reduce to publication either scientific (articles, conferences, reports …) or technological (patents). 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 as it 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-occurrences analysis of those citations. The main method: co-citations analysis developed by Small [12] 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 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.

3. Data

The experiment was carried out on 140 journals from the field of pharmacology. The digitization of the bibliographical references was realized by a subcontractor and the final result obtained by optical character recognition (OCR) is under the form of "well-formed" XML documents where each reference is well individualized amongst a set representing all the references extracted from the same article (see Figure 1). 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 standard (ISO 8879:1986) as well as in the XML standard, i.e. **Ş** for the character **Ş**.

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

- those due to the digitization: non-recognized characters, badly recognized characters (especially the uppercase letter **D** that sometime gives the uppercase letter **I** followed by a right parenthesis) or missing characters (mostly punctuation marks);
- those due to the heterogeneity of the data: the structure of a reference depends of the type of the cited document and of the origin of the citing article since the model of the citation depends of 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 strictness and that the form of the references may vary from one article to the other in the same issue of some journals;
- to that, one must add the typing errors, the omissions and sometime the presence of notes that have nothing to do with bibliographic references.

```
<INFCOM fic="1998/refm278.dat">
<NUMACQ><CLEA>35400007110423</CLEA><CLEB>0030</CLEB></NUMACQ>

 <REFBIB copie="0" >1 American Cancer Society. Cancer Facts and Figures-1997, American Cancer Society: Atlanta, 1997.</REFBIB>

<REFBIB copie="0" >2 Bonnadonna G, Valgussa P, Moliteri A, Zambetti M, Brambilla C. Adjuvant cyclophosphamide, methotrexate, and fluorouracil in node-positive breast cancer: The results of 20 years of follow-up. N Engl Med 1995; 332: 901-906.</REFBIB>

<REFBIB copie="0" >3 Booser DJ, Hortobagyi GN. Treatment of locally advanced breast cancer. Seminars in Oncology 1992; 19: 278-285.</REFBIB>

<REFBIB copie="0" >4 Rouëssé J et al. J Clin Oncol 1986; 4: 1765-1771.</REFBIB>

<REFBIB copie="0" >5 Swain SM et al. Neoadjuvant chemotherapy in the combined modality approach of locally advanced non-metastatic breast cancer. Cancer Res 1987; 47: 3889-3894.</REFBIB>

</INFCOM>
```

Figure 1. References extracted from the same article

4. Methodology

In the literature, we identified a similar work done at the NEC Research Institute as part of the CiteSeer system [9]. 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 reportedly accurate, its functioning is not explicit enough 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 [4,7].

Conversely, we consider, the retro-conversion of bibliographic references is better not done in a top-down way guided by a structure generic model. The variations are to important from one bibliographic document to the other 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.

Indeed, the study of a bibliographical document teaches us two important things:

- there is a lot of regularities in the same set of references;
- every field (or bibliographic element) includes certain specific keywords or parts of speech.

So the methodology of retro-conversion of bibliographic references is based on exploiting those two particularities.

4.1. Notion of regularity

Being written by the same author or produced by the same word processor, the bibliographical references present some regularities we are using. Those regularities affect essentially the structure of the references and are expressed as follows:

- unicity of the form for the same reference type;
- respect of the position for the same fields: when author names are given (as they are usually), it is always at the beginning of the reference;
- for the reference of a journal article, a field like the date of publication can only be in a very limited number of positions:
    - after the authors,
    - after the journal title
    - after the page number,

but always in the same position for the same set of references. Likewise, the article title (if present) is always before the journal title.

4.2. Notion of part of speech

In face of the complexity of representing the structure of a text and extracting information from that structure, several methods of document analysis were proposed on the base of tagging that information by extracting keywords. Then a linguistic model is sought from those keywords to highlight the informational content of the text fields and find linguistic units whose reference to reality is stable. For instance, the primary hypothesis of the SYDO model [6] is that the "parts of speech" (PoS) built around a noun (or a noun syntagma) are those carrying references to objects from the universe of discourse and therefore those to identify. The proposed linguistic model reflects the mechanism allowing to go from a predicate noun to a noun syntagma.

There are two families of automatic tagging method of PoS: rule-based methods [3,13] and stochastic methods [5,10,11] functioning in both supervised and unsupervised mode. The former uses typically a contextual information to attribute tags to unknown or ambiguous words. These rules are usually known as contextual frame rules. In addition to that contextual information, some systems use the morphologic information to resolve ambiguity due to unknown words. A few systems go even beyond the contextual and morphologic information by including rules taking inta account factors like punctuation or the use of uppercase letters.

The latter integrates frequency or probability in the validation process. The simplest ones use the probability that a word comes with a particular tag to identify it, but that approach has a uncertain behavior because of its local view of things that can lead to unacceptable sequence of tags. An alternative to that approach is to compute the probability that a given sequence of tags occurs. It is based on the n-grams method that considers that the best tag for a given word is determined by the probability it happens with the n preceding tags [8]. The method generally preferred for a stochastic tagger combines the two above-mentioned approaches, using the probabilities of tag sequences and the measured frequencies of words. It is known as Hidden Markov Models [11,14].

4.3. Application to bibliographic references

The notion of PoS is not as deep as in indexing applications. It consists simply in separating the text elements of the reference as the title of the article or the name of a conference by assuming those fields contain elements which are morphologically close. The rules used for that group together words tagged previously and if possible include unknown words (not recognized by the OCR, proper names, technical nouns) in order to define the extent of the fields. That technique for grouping words can also be used for other fields like the author names by putting together proper names, initials or first names and connectors.

5. Description of the approach

The proposed approach for retro-converting bibliographical references comprised three principal steps:

- morphological tagging of the different elements of the text;
- structural analysis in order to extract the different fields. That extraction can be done in two separate and sometime complementary ways:
    - by studying the regularity and redundancy,
    - by grouping in PoS.
- syntactical analysis in order to correct some references not completely validated before. That analysis is based on the references completely validated to generate models used for the correction of the references badly digitized, badly written or including words not recognized at the first step.

6. Morphological tagging

Table 1. Main primary tags

Tag	Meaning
AN	Alphanumeric string
CC	Connector (et, &, ...)
CWC	Common word, capital initial
CWL	Common word, lowercase
CWU	Common word, uppercase
EA	Expression “et al.”
ED	Editor (“Ed”., “Eds”.)
IN	Expression “In:”
IT	Initial
JM	Journal abbreviation
NM*n*	Number (*n* digits)
PG	Expression “p.” or “pp.”
PN	Proper name (author)
PR	Preposition
PU*s*	Punctuation mark *s*
UN	Unknown

The data set is composed of 64 articles chosen at random and contains 2,575 references. Each element, word or other, of every reference receives a tag accordingly to a pre-established list, defined a priori et susceptible to change in function of the results we obtain (see Table 1). Moreover, the tag for a number is followed by the number of digits (so “**2003**” is tagged “**NM4**”) and the tag of a punctuation mark is followed by the punctuation mark itself (so “**-**“ is tagged “**PU-**“).

Table 2. Lists used for tagging

List	Size
Proper names	626.641
French words	596.967
English words	128.679
Journals	7.279
Countries	362
Prepositions	180

We developed a simple ad hoc tagger which recognizes words in function of their form (numbers, initials ...) or the fact they belong to a list (author names, journal abbreviations, English or French common nouns ...) and give them the corresponding tag. The same word can receive several tags. Conversely, an element that can't be classified in any category receives the tag “UN”. The different lists (see Table 2) 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.

<INFCOM fic="1998/refm278.dat">
<NUMACQ><CLEA>35400007110423</CLEA><CLEB>0030</CLEB></NUMACQ>

<REFBIB copie="0" >1 American Cancer Society. Cancer Facts and Figures - 1997, American Cancer Society: Atlanta, 1997.</REFBIB>
<TAG etape="1">1/NM1 American/CWC Cancer/CWC/JM Society/CWC ./PU. Cancer/CWC/JM Facts/CWC and/CC Figures/CWC -/PU- 1997/NM4 ,/PU, American/CWC Cancer/CWC/JM Society/CWC :/PU: Atlanta/UN ,/PU, 1997/NM4 ./PU.</TAG>

<REFBIB copie="0" >2 Bonnadonna G, Valgussa P, Moliteri A, Zambetti M, Brambilla C. Adjuvant cyclophosphamide, methotrexate, and fluorouracil in node - positive breast cancer: The results of 20 years of follow - up. N Engl Med 1995; 332: 901 - 906.</REFBIB>
<TAG etape="1">2/NM1 Bonnadonna/UN G/IT ,/PU, Valgussa/UN P/IT ,/PU, Moliteri/UN A/IT ,/PU, Zambetti/PN M/IT ,/PU, Brambilla/PN C/IT ./PU. Adjuvant/CWC cyclophosphamide/CWL ,/PU, methotrexate/CWL ,/PU, and/CC fluorouracil/CWL in/CWL/PR node/CWL -/PU- positive/CWL breast/CWL cancer/CWL :/PU: The/CWC results/CWL of/CWL/PR 20/NM2 years/CWL of/CWL/PR follow/CWL -/PU- up/CWL/PR ./PU. N/IT Engl/PN/JM Med/PN/JM 1995/NM4 ;/PU; 332/NM3 :/PU: 901/NM3 -/PU- 906/NM3 ./PU.</TAG>

<REFBIB copie="0" >3 Booser DJ, Hortobagyi GN. Treatment of locally advanced breast cancer. Seminars in Oncology 1992; 19: 278 - 285.</REFBIB>
<TAG etape="1">3/NM1 Booser/PN DJ/PN/IT ,/PU, Hortobagyi/PN GN/PN/IT ./PU. Treatment/CWC of/CWL/PR locally/CWL advanced/CWL breast/CWL cancer/CWL ./PU. Seminars/CWC in/CWL/PR Oncology/CWC/JM 1992/NM4 ;/PU; 19/NM2 :/PU: 278/NM3 -/PU- 285/NM3 ./PU.</TAG>

Figure 2. Example of primary tagging

The file obtained after tagging is distinguished from the original file by the presence of an element "TAG" containing the tagged words (see Figure 2).

7. Structural analysis

That analysis is done by studying the regularity and the redundancy or by grouping together words from the text in function of their tag. 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. At the end of that process, each term or group of terms gets a new tag giving a more explicit identification of the field to which it belongs (see Table 3).

Table 3. Main secondary tags

Tag	Meaning
AU	Authors
DA	Publication date
JN	Journal title
PG	Page number
TIP	Article title
VOL	Volume number

7.1. Grouping rules

Amongst the different types of rule, we have:

- reducing rules: that lead to the grouping of consecutive identical tags like 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 that may be present in such a context (see Table 4). The basic syntax of these rules is simple: a sequence of tag separated by the plus sign "+" receives the tag appearing on the right of the symbol "**=>**". We also use the syntax of the regular expressions from the Perl programming language: the sign "**!**" indicates a negation, the square brackets give a range of possible characters and the vertical bar "|" an alternative between several tags or several rules.
- forming rules: initiate the beginning of a field by associating tags that are complementary in their description as for example the association of the name and the initials in the formation of an author's name (**AU**) (see Table 5).

Table 4. Example of reduction rules

Reduction rules for IT
IT!PN + IT!PN => IT |
IT!PN + PU[.,] + PU- + IT!PN + PU[.,] => IT |
IT + PU- + IT => IT |
IT!PN + PU[.,] + IT!PN + PU[.,] => IT
Reduction rules for PN
PN!IT + PN!IT => PN |
PN!IT + PU- + PN!IT => PN |
UN + PU- + PN!IT => PN |
PN!IT + PU- + UN => PN

- extending rules: concatenating sub-fields recognized independently as for example an author and the expression "et al." which confirms the field "**AU**" (see Table 6) or the expansion of the article title from an initial core composed of three common nouns by adding nouns, connectors and prepositions to let the field grow as much as possible. Note the secondary tag "**TMP**" that can be defined as the result of a rule and is always erased at the end of the treatment. Its purpose is to hold temporary groups of terms which can exist only during a part of the process so that another set of rules can use its components as input.

Table 5. Example of forming rules

Forming rules for AU
IT + PN => AU |
PN + IT => AU
Idem with punctuation marks
PN + PU[.,] + IT + PU[.,] => AU |
PN + PU[.,] + IT => AU |
IT + PU[.,] + PN + PU[.,] => AU |
IT + PU[.,] + PN => AU

- agglutination rules: allow the unknown terms (**UN**) to be absorbed if the conditions are right, as for example between two author names (see Table 7).

Table 6. Example of extending rules

With connectors
AU + CC + TMP => AU
(AU|TMP) + CC + AU => AU
AU + PU[,.] + CC + TMP => AU
(AU|TMP) + PU[,.] + CC + AU => AU
With the expression "et al."
(AU|TMP) + PU[.,] + EA => AU
(AU|TMP) + EA => AU

Table 7. Example of agglutination rules

Agglutination rules for UN in the field "Authors"
(UN|CWC|CWU|IT|PN) + AU => AU |
(UN|CWC|CWU|IT|PN) + PU[,.] + AU => AU |
(AU|UN|CW[CU]|IT|PN) + PU[,.] + PU[,.] + AU => AU

- mixed rules: combining a set of grouping rules to detect potential candidates and regularity to select the best amongst them. It is notably the case with page numbers the structure of which is very variable (see Table 8). In fact, that structure may be preceded by a page indicator like "p." and the hyphen may be missing.

Table 8. Page number formats

numeric – numeric
alphanumeric – alphanumeric
alphanumeric – numérique
numeric
alphanumeric

7.2. Application

We present here the application of these rules in order to find the important fields of a bibliographic reference.

Number	Type	Punctuation	Length	Increment	Next tag
1	NM	0	1	NA	CWC
2	NM	0	1	+	UN
3	NM	0	1	+	PN
4	NM	0	1	+	UN
5	NM	0	1	+	CWC/PN
6	NM	0	1	+	PN
7	NM	0	1	+	PN
8	NM	0	1	+	PN

Figure 3. Parameters used in search of a key

7.2.1. Search for keys. A bibliographic reference may be preceded by a key as a number or the abbreviation of the first author. The presence of that key implies that the date can only be in a simple numeric form, i.e. **1998**, but never followed by a letter, i.e. **1998b**. The key, if any, is necessarily at the beginning of the reference. It may be surrounded by a double punctuation mark (square or curly brackets or parentheses) or followed by a single punctuation mark (dot, hyphen or right parenthesis). Moreover if the first element, punctuation marks not included, is numeric then it is supposed to be incremented at each reference. All these parameters are analyzed for the whole set of references from a same article and a simple statistic calculation determines if there is a key and what is its form (see Figure 3).

XML tagging of a reference

<CLE> 1/NM1 ./PU. </CLE> Weinstein/PN MC/PN/IT ,/PU, Stason/PN WB/IT :/PU: Foundations/CWC of/CWL/PR cost/CWL -/PU- effectiveness/CWL analysis/CWL for/CWL/PR health/CWL and/CC medical/CWL practices/CWL ./PU. N/IT Engl/PN/JM J/JM/IT Med/PN/JM 1977/NM4 ;/PU; 296/NM3 :/PU: 716/NM3 -/PU-

Highlighting of the key in references

1. Weinstein MC, Stason WB: Foundations of cost - effectiveness analysis for health and medical practices. N Engl J Med 1977;296: 716 - 721.
2. Russell LB. Gold MR, Siegel JE, Daniels N, Weinstein MC: The role of cost - effectiveness analysis in health and medicine. JAMA 1996;276:1172 - 1177,
3. Alexander B, Nasrallah HA, Perry PJ, Liskow BI, Dunner FJ: The impact of psychopharmacology education on prescribing practices. Hosp Community Psychiatry 1983;34:1150 - 1153.

Figure 4. Search for a key

Then in every reference the first element of which matches with the obtained form, we add a tag "CLE" (French for key) around that element. Figure 4 shows an example of such a tagging of the field "key" and the result of the extraction of that field on the three first references of an article. The field is highlighted in color (pink in that case) when displayed by a W3 browser.

XML tagging of a reference

<AU> BEAN/CWU/PN ,/PU, B.P./IT </AU> (/PU(1985/NM4)/PU) ./PU. Two/CWC/PN kinds/CWL of/CWL/PR calcium/CWL channels/CWL in/CWL/PR canine/CWL atrial/CWL cells/CWL ./PU. Differences/CWC in/CWL/PR kinetics/CWL ,/PU, selectivity/CWL and/CC pharmacology/CWL ./PU. <AU> J/JM/IT ./PU. Gen/PN/JM ./PU. </AU> Physiol/JM ./PU. ,/PU, 86/NM2 ,/PU, 1/NM1 30/NM2 ./PU.

Highlighting of the authors in references

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.

BEAN, B.P., STUREK, M., PUGA, A. & HERMSMEYER, K. (1986). Calcium channels in muscle cells isolated from rat mesenteric arteries: modulation by dihydropyridine drugs. Circ. Res., 59, 229 - 235.

BEUCKELMANN, D.J., NABAUER, M. & ERDMANN, E. (1991). Characteristics of calcium - current in isolated human ventricular myocytes from patients with terminal heart failure. J. Mol. Cell. Cardiol., 23, 929 937.

Figure 5. Search for authors

7.2.2. Search for authors. As said previously, the search for authors is done by grouping together words in function of a series of rules. For that, we developed a program able to read these rules after a simple syntax and to apply them to the references. As indicated before, it is possible to characterize tags with a symbol of negation. So the tag "IT!PN" means an initial (or a series of initials) that cannot be a proper name. It is although possible to indicate a context as the beginning of the reference or being preceded and/or followed by one or more elements. For instance, the rule "**<CLE> + UN + IT => AU**" indicates "unknown word plus initial give

author" if there is an element "**CLE**" before that unknown word. While we are looking for authors, we also try to find the editors which receive the tag "**ED**". At that stage, it is possible to have errors that will be corrected later on (see Figure 5)

7.2.3. Search for dates. A bibliographic reference usually includes a date that may either be in a simple numeric form, i.e. **1998**, or followed by a letter, i.e. **1998b**. As for the key, we look for the date by studying the regularity except we don't have an a priori position. The only indication is that it is a four-digit number within a range from 1850 to the current year. So we analyze the environment of all the potential date to determine their position and the type of the element around them.

7.2.4. Search for article titles. Titles are a particular challenge for they may be made up of about everything, including author names, dates and even journal or monograph titles. The goal here is above all to locate at least part of the title to determine its position. Locating that field allow us to correct some errors on the field "Authors" because that field can only be in front of the field "Title". The starting rule to group words is to find three lowercase common nouns in a row; then we add other common nouns, connectors, prepositions to expand the field as much as we can.

7.2.5. Search for page numbers. Despite what is usually though about it, the format of page numbers is not limited to a number followed by an hyphen and another number even if it is the most often encountered form. We may have alphanumeric strings, sequences of digits and letters, instead of numbers or a single number or alphanumeric string (see Table 8). Moreover, the hyphen may have been misrecognized or not recognized at all. For these reasons, we use both the grouping method to collect elements that may be page numbers before making a choice based on regularity.

7.2.6. Search for volume number. Volume numbers are usually present as a simple number directly in front of the page numbers. However, it may be followed by an issue number between parentheses or by a mention like "special edition". It is possible with a few simple rules using the context to retrieve these volume numbers.

7.2.7. Search for journal titles. At that stage, a great part of the bibliographic reference has already been tagged so we can use the context. Moreover, a journal title is generally composed of standardized abbreviations (tagged "JM") or common noun with a capital initial. To limit errors, we do not process references concerning monographs that we can spot by the presence of tags like "ED" or "IN". The example in Figure 6 summarizes the result of the extraction of these different fields. Each field is identified by a particular color on the W3 server we use to assess that extraction.

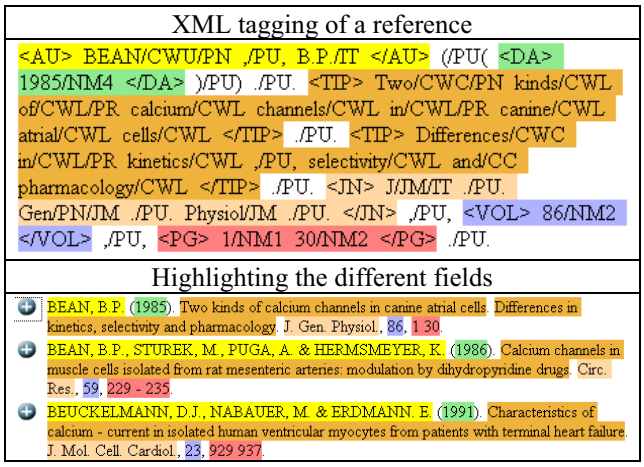

Figure 6. Result of structural analysis

8. Syntactic analysis

The structural analysis shows some limits and for several reasons: some terms are unknown and cannot be integrated in a field, the structure is too complex, there is a confusion between numbers (page, volume and date), etc. The idea of the syntactic analysis is to exploit what has been well recognized to create a model of how the different fields of the reference are supposed to be joined together, in other words the syntax of the reference. So the process consists in extracting such a model and using it for the correction.

We propose two types of model: inter- and intra-field models to progressively correct the different fields. The former propose a correction of the limit between fields and the latter is used to correct the internal structure of the field.

8.1. Inter-field modeling

We used a couple modeling revealing the associations of consecutive fields. The search is done in three steps:

From each recognized field, we determine what field follows and we measure the frequency of couples of fields as well as their relative position;

As we want to delimit correctly the different fields, we look for the separators if any and we measure their frequency;

Then we build the model by sorting all the couples accordingly to their average position and stringing them like dominoes as shown in Figure 7. When confronted with several possibilities, only the couple with a significantly greater frequency is kept. Likewise, only

the separators with a significantly high frequency are considered valid.

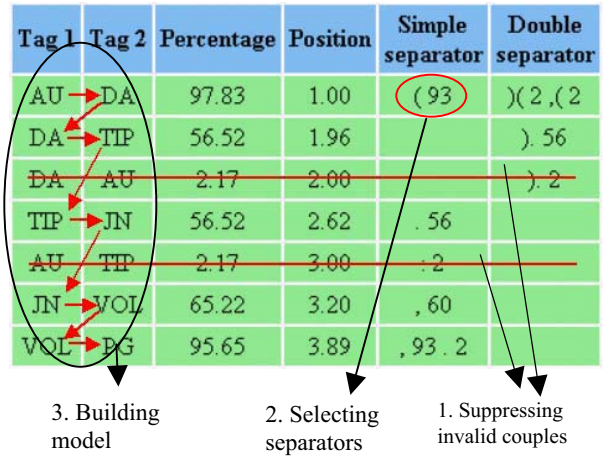

Figure 7. Extraction of an inter-field model

The complete model is given in Figure 8.

Figure 8. Inter-field model

Using that model, incomplete fields are corrected if the surrounding fields are clearly identified and delimited. In such a case, we can extend the field on the right and/or the left until the gap is closed. Put to the extreme, we can deduce the presence of an utterly unrecognised field by the presence of the correct fields and separators around it.

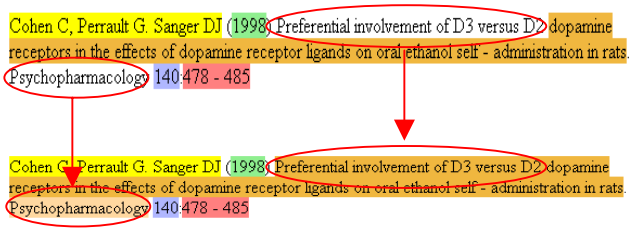

Figure 9. Example of inter-field correction

In the example in Figure 9, the title could be extended up to the right parenthesis that is the separator between date and title in the model. Moreover, always with that model, we could deduce that "Pharmacology" is a journal title from its position in the reference between the title and the volume number as well as from the presence of the expected separators.

8.2. Intra-field modeling

Once we have the right name and limits for a field, we search the kinds of element present in that field and their structure in terms of sequence and separators. In the case of the field "Authors", initials may be in front or behind an author's family name and each initial or name may be followed by a punctuation mark. That pattern may be different for the first author and the last author may be preceded by a connector. For these reasons, we consider three different cases as shown in Figure 10

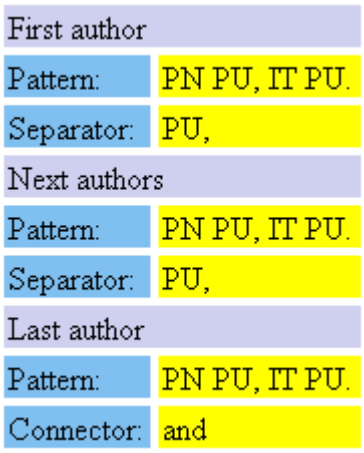

Figure 10. Example of intra-field model

With that model, we check each bibliographic reference. In the example shown in Figure 11 and corresponding to the model in Figure 10, the connector "and" indicates the position of the last author and the string "Wilson, J.M." corresponds to the pattern of an author's name. This means that the word "Gene" cannot belong to that field and it is therefore excluded. After that, a new iteration is necessary to identify the field "Title" from the inter-field model and test it with its intra-field model.

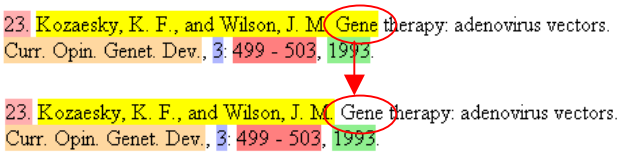

Figure 11. Example of intra-field correction

9. Experiment

The experiment was done on 140 journals in the field of pharmacology. The digitization of the bibliographical references was carried out by a subcontractor. The data set is made of 64 articles chosen at random from the original data set and contains 2,575 references.

We tagged them and we carried out the structural analysis and the syntactic analysis up to the intra-field modeling stage. At each step, the results can be visualized using a HTTP server and CGI scripts that highlight each recognized field with a specific color as shown in Figure 6.

Table 9. Results after inter-field correction

Fields	Complete	Partial	Not found	Wrong
Authors	90.2%	6.6%	0.3%	2.9%
Title	82.4%	15.4%	1.7%	0.4%
Journal	92.4%	2.9%	3.2%	1.5%
Date	97.7%	0.0%	2.3%	0.0%
Volume	93.6%	0.4%	5.8%	0.2%
Pagination	94.7%	0.6%	4.3%	0.4%
Whole Reference	75.9%	18.8%	0.0%	5.3%

In parallel, that data set was tagged by hand so we have a standard against which we can compare the results of our segmentation method. After the inter-field correction stage, 96.6% of words have been placed correctly in the right field while 0.5% have been wrongly attributed. Table 9 shows the results field by field and for the whole reference expressed as the percentage of reference where a specific field is complete, incomplete, not found or erroneous. For the whole reference, that means all fields are complete, at least one is incomplete, none are found or at least one is erroneous.

10. Conclusion

The method presented here works well on bibliographical references from most articles. For other sets of references with too few citations or too much heterogeneity, new algorithms will have to be devised to get round the problem like treating one document type at a time. Also, with the exception of the confusion between commas and periods, we do not compensate for the typical OCR errors.

For the entire process, we are following new leads to improve each stage from tagging to correcting.

For the time being, the process cannot learn from previous use, neither can it use external sources of information to help solve a problem (INIST has more then 10 million bibliographical records on-line and counting). This might also increase the efficiency of the system.

11. References

[1] A. Belaïd, "Retrospective document conversion: application to the library domain", *International Journal on Document Analysis and Recognition*, vol. 1, 1998, pp. 125-146.

[2] A. Belaïd, "Recognition of table of contents for electronic library consulting", *International Journal on Document Analysis and Recognition*, vol. 4, 2001, pp. 35-45.

[3] E. Brill, "A simple Rule-Bases Part of Speech Tagger", In *Proceedings of the third Annual Conference on Applied Natural Language Processing*, ACL, 1992.

[4] 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>

[5] S.J. DeRose, "Grammatical category disambiguation by statistical optimisation", *Computational Linguistics*, vol. 14 (1), 1988, pp. 31-39.

[6] G. Henneron, G. Lallich-Boidin, and R. Palermiti, "Analyse du français : achèvement et implantation de l'analyseur morpho-syntaxique", *Les cahiers du CRISS*, Nb. 16, 1990.

[7] 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>

[8] J. Kupiec, "Robust Part-of-speech tagging using a hidden Markov model", *Computer Speech and Language*, vol. 6 (3), 1992, pp. 225-242.

[9] S. Lawrence, C. L. Giles, and K. Bollacker, "Digital Libraries and autonomous Citation indexing", *IEEE Computer*, vol. 32 (6), 1999, pp. 67-71.

[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.