Automatic indexing and reformulation of ancient dictionaries

Abdel Belaïd, Isabelle Turcan, Jean-Marie Pierrel, Yolande Belaïd, Yves Rangoni, Hassen Hadjamar

Automatic indexing and reformulation of ancient dictionaries

A. Belaïd[1], I. Turcan[23], J.M. Pierrel[3], Y. Belaïd[1], Y. Rangoni[1] and H. Hadjamar[3]

[1]Loria-University of Nancy 2 France {abelaid@loria.fr, ybelaid@loria.fr, rangoni@loria.fr}
[2] University Jean Moulin, Lyon III France {Turcan@atilf.fr, Isabelle.Turcan@univ-lyon3.fr}
[3]ATILF UMR-CNRS 7118, Nancy France {Jean-Marie.Pierrel@atilf.fr, hassen.hadjammar@atilf.fr}

Abstract

This paper is related to automatic indexing and reformulation of ancient dictionaries. The objective is to make easy the access to ancient printed documents from XVI to XIX century for a diversified public (historians, scientists, librarians, etc.). Since the facsimile mode is insufficient, the aim is to look further for the use of the indexing based on the formal structure representative of some contents in order to optimize their exploration. Starting from a first indexing experiment operated on more recent documents, the TLF ("Trésor de la Langue Française": Treasure of the French Language) in the ATILF laboratory (Nancy, France), we extended the indexing technique to automatic reformulation and reedition of ancient dictionaries. However, face to the problem extent, we limited our investigations to a very specific collections of the ATILF laboratory, the "Trévoux" dictionary (defined later).

1 Introduction

1.1 Ancient dictionary valorization

The valorization aim of ancient documents and more specially of ancient dictionaries can be summarized in the following items, already outlined in [1,2,3]:

1. Make available texts that are often presented in complex typography forms, with different levels of difficulty depending on the centuries. In XVII and the beginning of XVIII century, there are problems of page-setting (columns, marginal texts), characters difficult to read (distinction between F and long S, long or double long S (see **Figure 3, 4, 5**), "esperluètes" (&), bindings, sign classifiers, the hand with stretched index (☛) which indicates a new entry in "Trévoux" dictionary and others). In the second half of the XVIII century and XIX century, some other specific difficulties appeared such as character size, reference sign increase, domain name and author of example name abbreviations, multicolumn (often 4 as in the "Grand Dictionnaire Universel" (GDU)), etc. On all the two centuries, and beyond, there is the problem of character restitution in non-Roman character sets.
2. Allow the auto-correction of the digitization errors by drawing up the typology of the errors noted on a representative sample, and to use them for the automatic correction of new occurred errors without any modification of the original text (except the characteristics of the typography: reference letter (see **Figure 6),** ancient characters, etc.).
3. Allow the fast comparison, by alignment, of texts considered "a priori" identical, but comprising differences not easily localizable with the naked eye. It is the case of many dictionary editions and republications of the XIX century, whose texts of escorts (forewords, introductions, lecturer point of view, etc.) and nomenclatures vary.
4. Finally, to define useful and effective indexing modes (definition problem of linguistic key words, lexical, and meta-linguistics, definers, etc.).

Proceedings of the First International Workshop on Document Image Analysis for Libraries (DIAL'04)
0-7695-2088-X/04 $20.00 © 2004 **IEEE**

1.2 The Trévoux Dictionary

Presentation of the collection

The Trévoux dictionary is also called "Dictionnaire Universel François & Latin" (1704-1771). This collection, rare by its coherence, is of a great worth [2,3]: it reassembles 36 volumes in folio and one volume in quarto corresponding to six editions, with complete and uniform series as well as specimens duplicated. The whole constitutes, at the same time by its width and its singularity, a real treasure within the ancient and regional funds preserved at Trévoux (City hall of Trévoux, Saône, France). **Figure 1** shows an example of the Trévoux dictionary digitized at 300 dpi in B&W.

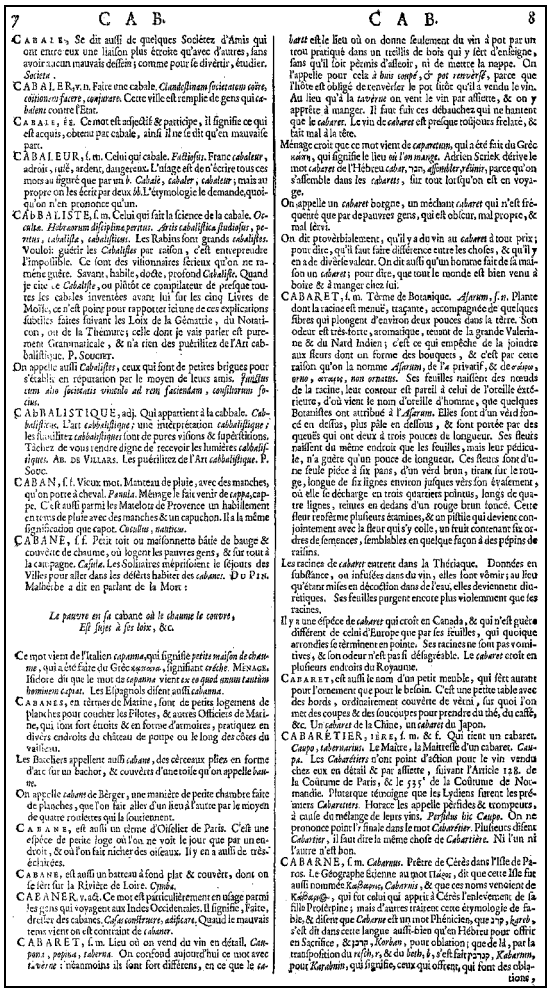


Figure 1: Example of Trévoux Dictionary.

2 The system overview

Figure 2 shows the reformulation system configuration. It is composed of four main steps related to: 1) document capture, OCR and document analysis, 2) document indexing and meta-data extraction, 3) document reformulation and user adaptation, 4) document reformatting for electronic reedition on specific electronic supports. All these phases will be detailed in the following.

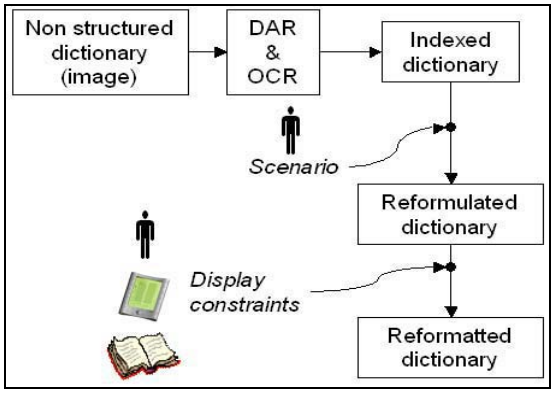


Figure 2: The complete formulation chain.

3 OCR and DAR

3.1 Digitization

Several problems can be encountered during the digitization of ancient dictionaries:

1. Background: the direct binarization of the image may introduce some errors. This is due to a degraded background, yellowed by age, which creates some dark blobs. Some of them with a big size are difficult to erase when they are connected to characters.
2. Noise: it corresponds to micro blobs introduced by ink during the printing process. There are close by size and position to some characters and can be confused with punctuation.
3. Character cutting: thin characters, which are very numerous in this kind of document, tend to be broken at the 300 dpi resolution.
4. Ligatures and kerning: some specific characters like S, f, t with long and curved shape tend to be connected or to overlap the next character in the word.

5. Unknown characters: the ancient typography presents some specimens which are not known by the OCRs. For example, the letter C represented like a big epsilon (see **Figure 7**).

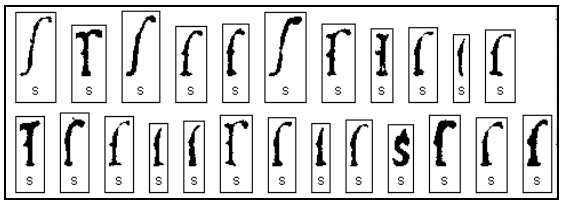


Figure 3: Long S, r and f.

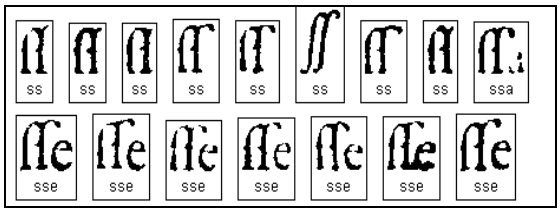


Figure 4: Examples of kerning

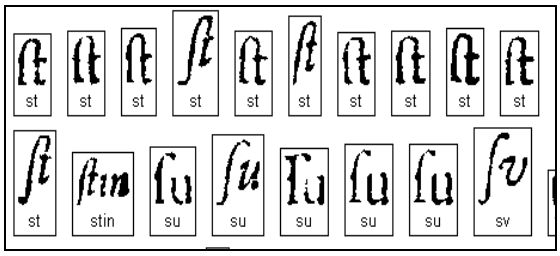


Figure 5: Examples of ligatures

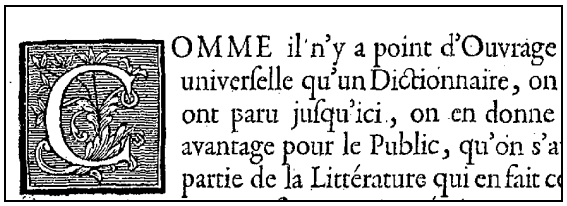


Figure 6: Example of the reference letter C in Trévoux.

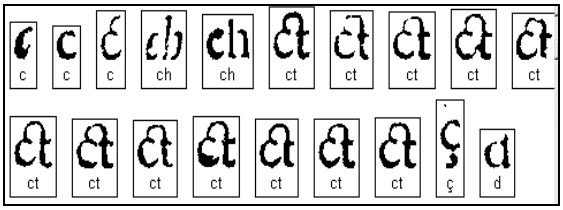


Figure 7: Example of the character C in Trévoux.

3.2 OCR Recognition

The indexing and reformulation procedures need text recognition. For this reason we have looked for an OCR solution adapted to this kind of documents. The presence of specific shapes leads us to consider them as new character classes and present them to OCRs for training. We can see in **Figures 3,4,5** and **7** the different shapes.

Furthermore, we noticed that the document typography is very complex and OCRs are often not efficient individually taken. So, we proposed an adapted approach based on OCRs combination [6]. The method consists in the development of an OCRAdapter, and an integration procedure of commercial OCRs as plug-ins. A ground truth is first established from a representative extract of the document class. Then, the first step is dedicated to the individual evaluation of each OCR. A specific alignment algorithm performs the evaluation. We have used the Myers's algorithm, based on an optimal dynamic programming matching [7,8,9].

This evaluation step allowed us to characterize for each OCR its performances and more precisely its drawbacks: bad recognized characters, recognition context considered, etc.

This knowledge on each OCR is taken into account in a second phase during the combination. The combination's philosophy is to consider only the two best OCRs and keep the better according to each kind type of error. As for evaluation, the two OCRs output files are linearized and aligned. The alignment technique is described in [9]. The principle is to choose a reference chain (also called chain of consensus) from which we will initialize the alignment. Then, this alignment is rectified continuously by comparison between the other chains and this reference chain.

The experiments realized on the Trévoux dictionary are still in progress, but similar works, operated on more recent documents, show that, for a basic combination, a score of 20/10000 errors is obtained. Furthermore, considering some specific heuristics, we can easily reach 1/10000 error. We hope that, from the first results ob-

tained on ancient dictionaries, we will obtain identical results. The heuristics employed are issued from the observation of the confusions provided by the OCRs in competition and attempt to correct them. For example, each time that OCR1 is giving "I" (always confused by "1"), we replace this answer by that one given by OCR2 more efficient on the recognition of digits.

4 Dictionary indexing

4.1 Dictionary indexing problems

The ancient dictionaries were neither written nor composed as the current dictionaries [4,5]. They don't follow the same systematic structure principles for the entry definitions. These definitions are made up of informational fields that designate various indication categories which constitute the article microstructure. This, for example, gives grammatical category indications (noun, verb, adjective, article, adverb, conjunction, etc.), pronunciation category indications, orthography difficulties, information relating to the etymology, definitions organized or not according to the semantic logic of the proper sense illustrated or derived, examples, synonyms and antonyms, particular definitions, etc.

4.2 Data indexing difficulties

The systematic absence of the speech structure in these dictionaries causes two major difficulties for the automatic processing [6]:

1. On one hand, the geographic repartition informational fields is irregular, often random, unequally differentiated by formal criteria such as the typography (for example, semiotics of the italic or capitals) or the punctuation (for example, semiotics of the point not followed by a capital letter in an enumeration) and the position in the article layout (information structure according to the subparagraphs);
2. On the other hand, the polysemous statute of certain statements presents a functional ambiguity of the fields. In fact, we can wonder to which informational fields belong the synonymic definitions or the synonyms capable to be interpreted like definitional. Similarly for the definitional etymologies or the etymology definitions, can we know if they concern the definition field or the etymology field?

This structural blur led us to create the concept of keywords and meta-linguistic sequence-keys to facilitate the information access without using a finest labeling. These tools showed their effectiveness in exploration procedures of lexicographical textual bases once implemented in a rigorous way, not by applying arbitrary labels to the text, but by extracting objective information (combinatory of typographical criteria, positional and lexical or phraseological).

In spite of the two difficulties previously stated, it is not utopian to envisage some automatic processing modalities even for a part of the information. This can help to derivate the principle of an automatic content reformulation.

In fact, a first set of investigations carried out on our corpus shows that it is possible to identify within the definitions some specific structures. These structures that correspond to specific information in the microstructure, are easily detectable thanks to the typography and positional indices of some elements.

As it is pointed out previously, this kind of formal identification is however limited to the physical aspect of the information. In fact, it is very difficult to identify the logical structures because this needs interpretation.

Nevertheless, in spite of this reserve, we consider that it will be possible to proceed to an automatic reformulation approach, based only on physical criteria. We are however conscious that in this case, a user can exploit only a part of the articles, such as the synonyms, the examples, the named sources or the grammatical terminologies.

4.3 Indexing methodology

The goal of the reformulation is thus to be able to extract, from a textual set, definite parts of the content, i.e. fields or information in "coded" position, and to reorganize them in different manner than in the original text [10,11].

This is why, considering the work realized these last years concerning the electronic publishing methodology of ancient dictionaries and the possible procedures of exploration and exploitation of their content, we consider that it is important to take into account the following elements:

1. The explicit structure of the lexicographic speech.
2. The main informational fields: typological proposition.

    A) The explicit structure of the lexicographic speech

In a first time, we discard all the informational fields functional ambiguity cases whose the logic design is

implicit or registered in an editorial process implying an interpretation accompanied with demonstrations. This excludes any prospect of automatic treatment based on formal criteria specific to the considered texts except if we admit a lexicographical bases preliminary treatment by a fine labeling based on specialists meta-lexicographic competences of these texts.

From a strict formal point of view, three series of combined criteria must be taken into account to determine the first principles of fields location. This is a necessary precondition to any automatic extraction of their contents:

1. Typographical criteria: character size (large / small right Roman, large / small capitals, etc.), character status (upper case / lower case), font style (italic / straight Roman, etc).
2. Positional criteria of the subparagraphs within the same article and of the information present within the same subparagraph:
   - In the article area: identification and delimitation of the subparagraphs. Each subparagraph must be numbered and profit from a mark of subparagraph beginning / end, for each subparagraph withdrawal / absence of withdrawal.
   - In the area of a subparagraph: all the punctuation marks must be codified (specially the points not followed by a capital letter, commas followed by italic and certain semicolons). The typography (in particular upper case / lower case, straight / Italic), when having a diacritic function, must be codified because it allows to determine the beginning and the end of an informational field or particular sequence identifiable within a field.
3. Lexical or phraseological criteria: we find here the logic of keywords and metalinguistic sequence-keys which allow to isolate the grammatical category marks (often shortened in coded position), the field marks (sometimes with alternatives), the synonyms with definitional function (especially locatable from a positional point of view), the formulas specific to the definitional or etymological speech, etc.

B) The main informational fields: typological proposition

The main fields encountered in the ancient dictionaries are as follows:

1. Entries or "vedette".
2. Grammatical category marks.
3. Indications of pronunciation and of graphic variant.
4. Usage, domain, technical, language level marks, etc.
5. Synonymous with definitional function.
6. Elaborated definitions.
7. Translations of foreign languages.
8. Etymologies.
9. Marked bibliographic references (author names cited and concerned entitled texts)
10. Texts cited with and without reference.
11. Intra-textual returns (to other articles within the same piece of work) or inter-textual (to a set of text sources).

We propose a modeling in the form of a table as following:

Criteria	Typographical	Positional	Phascoligical
1.Entry	Capitals	Beginning sub-paragraph	
2.Pronunciation	Sometimes in Italic or between parentheses	Just after the entry or the field 3	Key-words «on / se prononce»
3. Grammar Cat.	Italic	Just after the entry or the field 2	Typology of the abbreviation of metalinguistic Keywords
4.Domain Cat.	Straight Roman	After fields 2 or 3	Metalinguistic keywords: «terme(s) de», «mot de», «en» + domain name
5.Definition or definitional synonyms	Straight Roman	After fields 3 or 4	a) Metalinguistic keywords, «sorte de», «espèce de», «signifie, veut dire, comme quand on dit, etc.» b) Def. by synonymous (hyperonymous), or derivated.
6.Examples / ref. & quotations	Sometimes in Italic or between brackets	After field 5	The author names and work titles can be listed
7.Etymology	The etymology is sometimes in italic or in straight roman in small capitals	Random, sometimes in position 1, sometimes after fields 2-6	Metalinguist. Keywords or keyword sequences: «vient de», «formé de», «dérivé de», «d'origine»
8.References	Sometimes in Italic	Position at the end of an article or a subparagraph	Metalinguistic keywords. «voyez», «voy.»

A number is given to each automatizable field category. We associate to this number the principal criteria, according to the corpus on which the work is applied (French language dictionaries of XVII, XVIII and XIX centuries), knowing that, it is obviously imperative to

distinguish the fields numbers of those corresponding to the order in which appear the fields within an article. The proposed table is, of course, only an indication.

Illustration of the typologies:

We are going to detail the previous presented typologies with some extracts chosen for representative dictionaries:

Extract from « Dictionnaire françois de Richelet, 1680 »
« Dictionnaire, s. m. Livre qui contient les mots d'une langue, d'un art, ou d'une science par ordre alphabétique. [Un bon dictionnaire est tres-dificile à faire. Un dictionnaire de droit, un dictionnaire de médecine, un dictionnaire de rimes.] »
The corresponding modeling table is as follows:

1. Entry	*Dictionnaire*
2. Grammatical category	s.m.
3. Pronunciation	0
4. Domain Category	0
5. Definition	*« livre qui ...»*
6. Example	*« [Un bon dictionnaire est tres-dificile à faire. Un dictionnaire de droit, un dictionnaire de médecine, un dictionnaire de rimes.] »*
7. Etymology	0
8. Reference to author and work	0

Extract from « Dictionnaire de l'Académie française, 1694 »
« DICTIONNAIRE. s. m. Vocabulaire. Recueil de tous les mots d'une Langue, mis par ordre. Dictionnaire François. dictionnaire Latin. dictionnaire François-Latin. dictionnaire par ordre alphabétique. dictionnaire par l'ordre des racines, par racines. bon dictionnaire. ample dictionnaire..».
The corresponding modeling table is as follows:

1. Entry	*DICTIONNAIRE*
2. Grammatical category	s.m.
3. Pronunciation	0
4. Domain Category	0
5. Definition	*« Vocabulaire. Recueil de tous les mots d'une Langue »*
6. Example	*« Dictionnaire François. Dictionnaire Latin. dictionnaire François-Latin... »*
7. Etymology	0
8. Reference to author and work	0

Extract from « Dictionnaire étymologique de Ménage, 1694 »
« AMARER. Terme de Marine, qui signifie attacher, ou lier. Voyez le Sr Guillet dans son Dictionnaire de la Marine : Et Mrs de l'Academie dans leur Dictionnaire de la Langue Françoise. »
The corresponding modeling table is as follows:

1. Entry	*AMARER*
2. Grammatical category	0
3. Pronunciation	0
4. Domain Category	*« Terme de Marine »*
5. Definition by synonymous	*« qui signifie attacher, ou lier »*
6. Example	*0*
7. Etymology	*0*
8. Reference to author and work	*« Voyez le Sr Guillet dans son Dictionnaire de la Marine : Et Mrs de l'Academie dans leur Dictionnaire de la Langue Françoise. »*

Extract from « Dictionnaire de Ménage, 1694 » and « Dictionnaire de Trévoux, 1743 » the same word : « Ménage »

« LARGUE. Terme de Marine. Vent largue. De l'Italien largo. On a dit largue, pour large ; comme cargue, pour charge. »
The corresponding modeling table is as follows:

1. Entry	*LARGUE*
2. Grammatical category	*[adj]*
3. Pronunciation	0
4. Domain Category	*« Terme de Marine »*
5. Definition	*0*
6. Example	*« Vent largue »*
7. Etymology	*« De l'Italien largo »*
8. Reference to author and work	0

« LARGUE, s. m. Quelquefois on donne un article féminin à ce mot, & on dit la largue. Terme de Marine. Haute mer. Altum mare, altum. Il n'a guère d'usage qu'en ces phrases ; Prendre le largue, tenir le largue, faire largue; pour dire, Prendre la haute mer, tenir la haute mer, aller en haute mer. On dit aussi adverbialement qu'ils se sont mis à la largue, qu'ils se sont mis en haute mer, de peur d'être jettés sur les côtes. Tous les autres vaisseaux qui étoient dans le port s'étant mis à la largue, saluerent ces nouveaux venus de toute leur artillerie. Du Loir, Voyage du Lev. page 107. Ce mot est aussi adjectif ; ainsi on appelle vent largue, ou vent de quartier, Obliquus ventus, l'aire de vent qui est comprise entre le vent

arrière, & le vent de bouline. C'est le plus favorable des vents pour le sillage, car il donne dans toutes les voiles : au lieu que le vent en poupe ne porte que dans les voiles d'arrière, qui dérobent le vent aux voiles des mâts d'avant. Un vaisseau qui fait trois lieues par heure de vent largue, n'en fait que deux de vent en poupe. Au lieu de nous tenir au plus près, nous courions vent largue de deux aires de vent, afin d'être mieux en ligne. M. le C. de Toulouse. Nous avons un peu de vent, mais il est largue. Toutes nos voiles portent, & nous ne roulons plus. L'Ab. de Choisi. Ce mot est la même chose que large, il n'y a que la prononciation de la dernière syllabe qui soit différente ; mais il ne faut s'en servir qu'en termes de Marine. »

The corresponding modeling table is as follows:

1. Entry	*LARGUE*
2. Grammatical category	s.m. *« Quelquefois on donne un article féminin à ce mot, & on dit la largue »* *« On dit aussi adverbialement »* *« Ce mot est aussi adjectif »*
3. Pronunciation	*«il n'y a que la prononciation de la dernière syllabe qui soit différente »*
4. Domain Category	Terme de Marine
5. Definition by synonymous	*«Haute mer»* *« pour dire »* *«on appelle»*
6. Example	*« Prendre le largue, tenir le largue, faire largue; »* *« ils se sont mis à la largue »* *« vent largue, ou vent de quartier »*
7. Etymology	0
8. Reference to author and work	*« Du Loir, Voyage du Lev. page 107 »* *« M. le C. de Toulouse »* *« L'Ab. de Choisi »*

We will thus remember that:

1. It is necessary to label the fields that will enter in the reformulation functioning logic, therefore identifiable in an absolutely objective manner without lending to discussion because of semantic or functional ambiguities subjected to a subjective interpretation.
2. It is necessary to define, for each specific corpus, the criteria set, combinatory or not, that allows the implementation of an automatic labeling based only on formal reference marks, therefore objectively defined.

However, according to the apprehended corpus, the formal criteria will be able to vary. This implies to determine, in addition to the minimal formal criteria, the facts of random distribution and different variant typologies.

Fields isolated according these minimal formal criteria

The fields that can be automatically recognized are:

1. Entries or "vedettes" => position + typography (subparagraph beginning, shifting, big capitals).
2. Grammatical category marks => position + typography + meta-linguistic keywords (just after the entry, sometime in italic, abbreviation system).
3. Indication of pronunciation and of graphic variants => position + typography + meta-linguistic keywords (before or after the mark of grammatical category, pronunciation indication often between parentheses, keyword: the verb « prononcer »).
4. Usage marks, domain marks, technical marks, language level marks … => position + typography + metalinguistic keywords (« terme de », « en termes de », « en » + domain name, « mot de », …) provided taking into account the trade names (« les tonneliers disent », « se servent de »…).
5. Synonyms with definitional functions => position + typography ("par exemple dans l'Académie", capital).
6. Intra-textual references (to other articles within the same work) or inter-textual (to a set of text sources) => position + typography + metalinguistic words ("voyez").

Difficult fields

In the case of the difficult fields that are very delicate to mark out in an automatic way, we will retain, as an example, the principal complexities specific to the various texts for some categories of fields only. Are concerned the following fields:

7. The elaborated definitions.
8. The translations in foreign languages.
9. The etymologies.
10. The marked bibliographic references (author names cited and concerned text entitled).
11. The cited texts with or without reference.

The complexity report of the identifying information of the fields is as follows:

12. The elaborated definitions: the metalinguistic keyword of reference « signifie » sometimes knows complex alternatives (« on dit de … que… parce que… ") or is not always expressed, or is taken again during the same article in the case of multiple references words treatment.
13. The translations in foreign languages are often in italic, but as the italic is polysemous in many an-

cient dictionaries, it is necessary to associate another criterion. Indeed, the labeling of the translation languages is very random. Therefore, except considering works announcing a translation of principle (like the "Dictionnaire Universel François & Latin de Trévoux" = DUFLT), only the linguistic ability of the dictionaries reader-consultant authorizes an identification of the translation languages. Knowing that there is often absolute ambiguity of the forms belonging to the whole of the Romance languages, and that in any event even in a dictionary like the DUFLT, it is difficult automatically to select the shapes in italics corresponding to Latin translations.

14. The etymologies are not easily automatically isolable not only because of the variety and complexity of the metalinguistic key-sequences « a pour origine, est formé de, est composé de, est fait de, est emprunté de, vient de, etc… », but also especially because of polysemous operations of the formulations... to what we can add the randomly presence of an etymological speech, the relative blur statute semantic of certain etymologist speeches, without counting the alternatives of positions and the delimitation difficulties of the fields of a speech relating at the same time to the etymology, the history of the directions, and the socio-cultural evolutions of notions…
15. Marked bibliographical references (names of cited authors and headings of the concerned texts). Without a precise knowledge of the authors, we can trust neither the typography semiotics (proper names authors are always in capital letters or in italic), neither with the position at the end of the quoted text, provided that this last either identifiable like such by italics, nor with the association of a proper name and a more or less identifiable reference because of the abbreviations, etc.
16. Texts cited with or without reference: identification difficulty of the precise limits of a citation, even when the authors and printers of dictionaries choose the italic for the quoted texts, without neglecting the complex facts of citations inserted in quotations, quotations not marked and sometimes easily locatable by a reader accustomed to the reading of certain texts.

5 Automatic Indexing

5.1 Entry Recognition

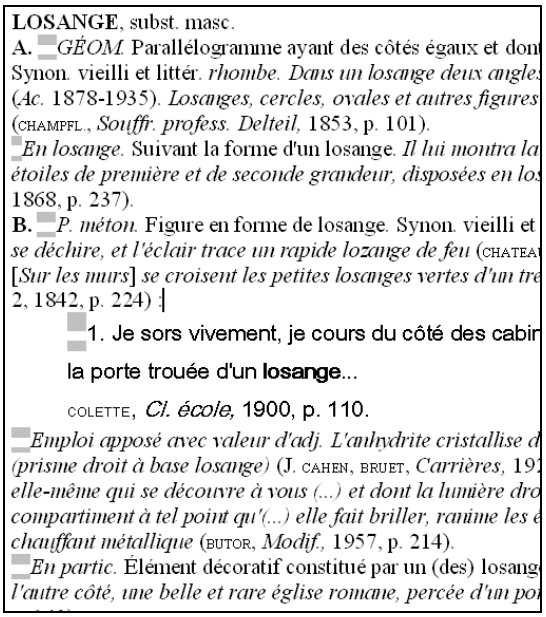
Table 1 shows in the first column the different kind of entries, in the second column some examples occurred, and in the last column some rules used for the entry extraction. As the entries are very specific in terms of typography and composition, their extraction doesn't give major errors.

Entry	Example	Characteristics
Word	ABANDON, ſ.m. Mépris; délaiſſement *Derelictio, deſtitutio. Neglectus rei alicujus.* uſage. On ne le trouve guére que dans Mol	Upper case bancient, spaced, followed by a coma, left justified
Subword with ≠ meaning	ABANDON, ſe dit d'ordinaire ſon à l'*abandon*, au pillage. I	Small capitals similar typo. but included in bloc def.
Multiple words: several orthog. for same word	ABASSI, ou ABASSIS, ſubſt. m. qui eſt ronde & qui a cours en Pèrſe viron dix-huit ſols ſix deniers. Il fa	Similar single word, separated by « ou » and a coma
Particular case	AbBREUVOIR, ſ.m.Lieu où *rium. Mener les chevaux à l'abbr* d'un glacis le plus ſouvent pavé	2d character in lower case (sometimes in small capital) if main entry
Sub-definition	ABBREUVOIR, en tèrme de Maçonner. que les Maçons laiſſent entre les joints c entrer du mortier. En ce ſens l'on ſe ſert	In a secondary entry, small capitals

Table 1: Entry characterization.

5.2 Information field identification

Concerning the identification and the tagging of the informational fields, an optimal solution could consist in implementing a procedure similar to that performed for the computerization of the TLFi (Treasure of the French Language computerized: www.atilf.fr/tlfi) [18] which is composed into various semi-automatic steps [14]. The first step is dedicated to the typography style labeling. **Figure 9** shows the result of this labeling procedure on the example of the **Figure 8.** The text is cut according to the punctuation. Each text chain is surrounded by a specific XML tag corresponding to the font style, for example <I> … </I> for italic. The second step is dedicated to the semantic labeling **Figure 10.** By using some dictionaries and by the help of the operator, some semantic tags are assigned to the same text chains previously labeled. Among the semantic labels used, we can find: article (art), entry (ved), word (mot), domain (dom), etc. The last step is related to the structural labeling **Figure 11.** The structure identifies the definition levels (main definition, sub-definitions, etc.).



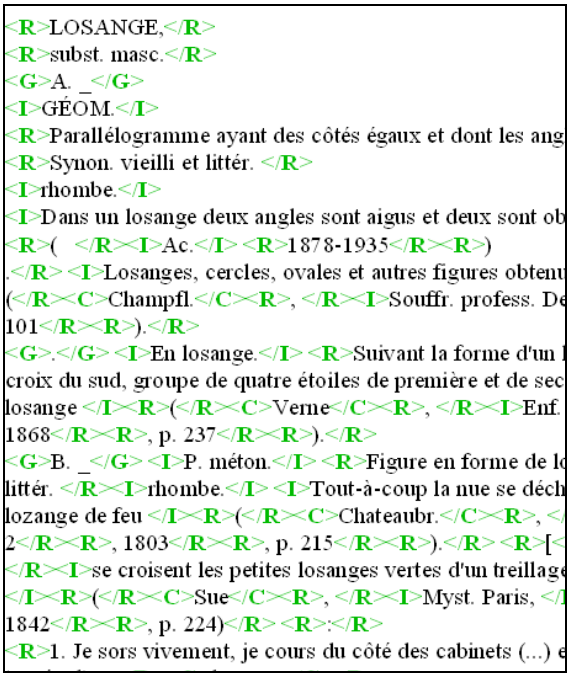
LOSANGE, subst. masc.
A. *GÉOM.* Parallélogramme ayant des côtés égaux et dont
Synon. vieilli et littér. *rhombe. Dans un losange deux angles*
(*Ac.* 1878-1935). *Losanges, cercles, ovales et autres figures*
(CHAMPFL., *Souffr. profess. Delteil,* 1853, p. 101).
En losange. Suivant la forme d'un losange. *Il lui montra la*
étoiles de première et de seconde grandeur, disposées en los
1868, p. 237).
B. *P. méton.* Figure en forme de losange. Synon. vieilli et
se déchire, et l'éclair trace un rapide lozange de feu (CHATEA
[*Sur les murs*] *se croisent les petites losanges vertes d'un tre*
2, 1842, p. 224) :

1. Je sors vivement, je cours du côté des cabin
la porte trouée d'un **losange**...

COLETTE, *Cl. école,* 1900, p. 110.

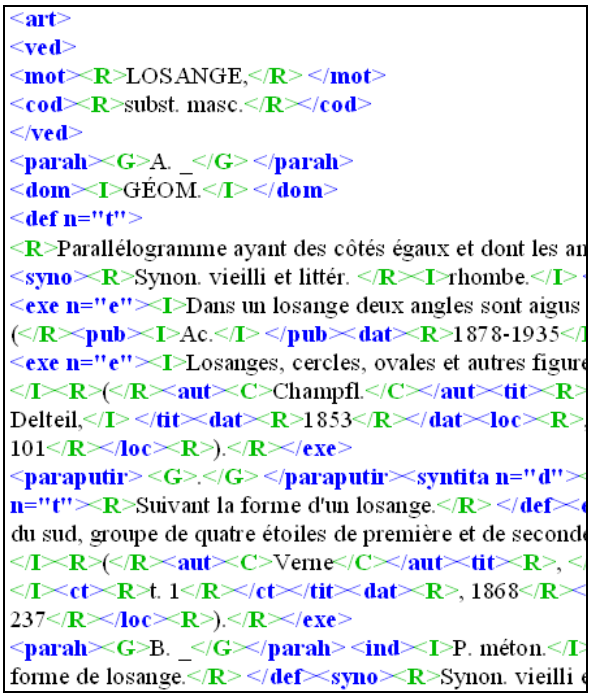
Emploi apposé avec valeur d'adj. L'anhydrite cristallise d
(prisme droit à base losange) (J. CAHEN, BRUET, *Carrières,* 19
elle-même qui se découvre à vous (...) et dont la lumière dro
compartiment à tel point qu'(...) elle fait briller, ranime les é
chauffant métallique (BUTOR, *Modif.,* 1957, p. 214).
En partic. Élément décoratif constitué par un (des) losang
l'autre côté, une belle et rare église romane, percée d'un po

Figure 8: Image of a dictionary example.



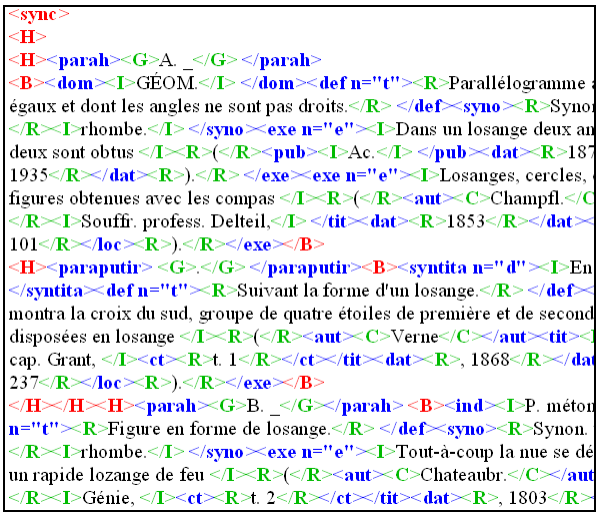
<R>LOSANGE,</R>
<R>subst. masc.</R>
<G>A. _</G>
<I>GÉOM.</I>
<R>Parallélogramme ayant des côtés égaux et dont les ang
<R>Synon. vieilli et littér. </R>
<I>rhombe.</I>
<I>Dans un losange deux angles sont aigus et deux sont ob
<R>(</R><I>Ac.</I> <R>1878-1935</R><R>)
.</R> <I>Losanges, cercles, ovales et autres figures obtenu
(</R><C>Champfl.</C><R>, </R><I>Souffr. profess. De
101</R><R>).</R>
<G>.</G> <I>En losange.</I> <R>Suivant la forme d'un
croix du sud, groupe de quatre étoiles de première et de sec
losange </I><R>(</R><C>Verne</C><R>, </R><I>Enf.
1868</R><R>, p. 237</R><R>).</R>
<G>B. _</G> <I>P. méton.</I> <R>Figure en forme de lo
littér. </R><I>rhombe.</I> <I>Tout-à-coup la nue se déch
lozange de feu </I><R>(</R><C>Chateaubr.</C><R>, <
2</R><R>, 1803</R><R>, p. 215</R><R>).</R> <R>[<
</R><I>se croisent les petites losanges vertes d'un treillage
</I><R>(</R><C>Sue</C><R>, </R><I>Myst. Paris, <
1842</R><R>, p. 224)</R> <R>:</R>
<R>1. Je sors vivement, je cours du côté des cabinets (...) e

Figure 9: Typography style labeling.



<art>
<ved>
<mot><R>LOSANGE,</R> </mot>
<cod><R>subst. masc.</R></cod>
</ved>
<parah><G>A. _</G> </parah>
<dom><I>GÉOM.</I> </dom>
<def n="t">
<R>Parallélogramme ayant des côtés égaux et dont les an
<syno><R>Synon. vieilli et littér. </R><I>rhombe.</I>
<exe n="e"><I>Dans un losange deux angles sont aigus
(</R><pub><I>Ac.</I> </pub><dat><R>1878-1935</
<exe n="e"><I>Losanges, cercles, ovales et autres figure
</I><R>(</R><aut><C>Champfl.</C></aut><tit><R>
Delteil,</I> </tit><dat><R>1853</R></dat><loc><R>,
101</R></loc><R>).</R></exe>
<paraputir> <G>.</G> </paraputir><syntita n="d">
n="t"><R>Suivant la forme d'un losange.</R> </def><
du sud, groupe de quatre étoiles de première et de seconde
</I><R>(</R><aut><C>Verne</C></aut><tit><R>, <
</I><ct><R>t. 1</R></ct></tit><dat><R>, 1868</R><
237</R></loc><R>).</R></exe>
<parah><G>B. _</G></parah> <ind><I>P. méton.</I>
forme de losange.</R> </def><syno><R>Synon. vieilli e

Figure 10: Semantic labeling.



<sync>
<H>
<H><parah><G>A. _</G> </parah>
<B><dom><I>GÉOM.</I> </dom><def n="t"><R>Parallélogramme
égaux et dont les angles ne sont pas droits.</R> </def><syno><R>Syno
</R><I>rhombe.</I> </syno><exe n="e"><I>Dans un losange deux an
deux sont obtus </I><R>(</R><pub><I>Ac.</I> </pub><dat><R>187
1935</R></dat><R>).</R> </exe><exe n="e"><I>Losanges, cercles,
figures obtenues avec les compas </I><R>(</R><aut><C>Champfl.</C
</R><I>Souffr. profess. Delteil,</I> </tit><dat><R>1853</R></dat><
101</R></loc><R>).</R></exe></B>
<H><paraputir> <G>.</G> </paraputir><B><syntita n="d"><I>En
</syntita><def n="t"><R>Suivant la forme d'un losange.</R> </def><
montra la croix du sud, groupe de quatre étoiles de première et de second
disposées en losange </I><R>(</R><aut><C>Verne</C></aut><tit><
cap. Grant, </I><ct><R>t. 1</R></ct></tit><dat><R>, 1868</R></dat
237</R></loc><R>).</R></exe></B>
</H></H><H><parah><G>B. _</G></parah> <B><ind><I>P. méton
n="t"><R>Figure en forme de losange.</R> </def><syno><R>Synon.
</R><I>rhombe.</I> </syno><exe n="e"><I>Tout-à-coup la nue se dé
un rapide lozange de feu </I><R>(</R><aut><C>Chateaubr.</C></aut
</R><I>Génie, </I><ct><R>t. 2</R></ct></tit><dat><R>, 1803</R>

Figure 11: Structural Labeling.

6 The reedition chain

Once the indexation phase is finished, the following phase concerns the reformulation and the reformatting taking into account the user wishes, the electronic support limits (physical and software characteristic) and the usage rules.

6.1 Reformulation

Here we are interested in the document logical structure, which we wish to modify in order to satisfy a personal use (expressed by a scenario). This modification concerns at the same time the document organization (layout) and its content. For example, the user can ask for a synthetic view of the document, to see only some definitions, by changing the original order, etc. In the case of the dictionaries, the user can select only the synonymous, the grammatical forms, etc.

The scenario is a collection of queries. Considering the document as a structure hierarchy of elements (XML, see **Figure 13**), the queries express two types of operations: tree element selection and tree element reorganization.

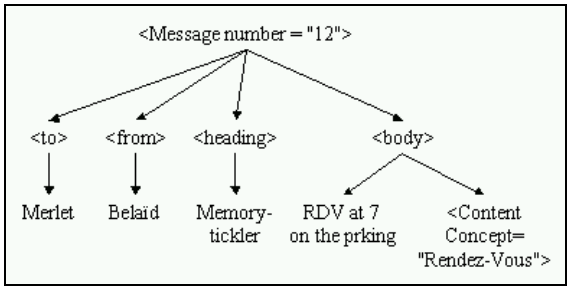


Figure 13: Logical description XML of the indexed document.

There are two kind of selection: direct and indirect. The direct selection permits to indicate an element starting from its absolute position in the logical structure (for example, the sub-definition of the 2nd dictionary entry). The indirect selection (or conditional) gives access to all the elements satisfying certain conditions. Those relate either to the content or its meaning (for example, to reach all the definitions speaking about the concept of "navy"), or function of the elements relative position to be selected in the document logic structure (to reach all the definitions of an entry). We defined an element designation language, which is detailed in **Table 2** and illustrated by examples in **Table 3.**

select_direct(list of target nodes)
select_remove_direct (list of target nodes) direct selection/unselection of nodes and their sub-structures
select_prune() Pruning of empty branchs
select_keep_parent_from_tag(node, list of tags to search) *select_remove_parent_from_tag(node, list of tags to search)* selection/unselection of a node and its substructure according to the presence of specific descendants;
select_in_parent_keep_tag(node, list of tags to search) and *select_in_parent_remove_tag(parent, node, list of tags to search)* selection/unselection of the part of the node substructure answering to some conditions

Table 2: Selection operators.

select_direct([Entry/Entry definition entry/synonymous]) to keep only the definition and the synonymous of a definition
select_remove_direct([entry/definition [2], entry/definition [4]) to keep only the definitions 2 and 4 of an entry
select_keep_parent_from_tag(entry/definition, [synonymous]) to keep only the entries of a n entry containing a synonymous
select_remove_parent_from_tag(entry/definition, [content [@concept = "navy"]]) to not keep the entry definitions speaking about the concept "navy"
select_in_parent_keep_tag(definition, [synonymous, illustration]) to keep, in each definition, only synonymous and illustrations
select_in_parent_remove_tag(definition, [illustration]) to erase, in each definition, all the illustrations

Table 3: Example of selection operators.

The purpose of the reorganization operators is to authorize the original document reading order modification. For example, the user can wish to read synonymous before definition, or to replace all the illustrations at the end of a definition. We can see some example of such operators in **Table 4.**

arrange_change_order(common parent, tag element1, tag element2) to change the position of two elements (regarding the corresponding tree root, or regarding a specific node)
arrange_move_after(tag element to move, tag element used as a reference, list of eventual tags to add) to move an element and to situate it after an other
arrange_up_one_level(element to climb up, list of tags to add) *arrange_change_level(element to climb up, new parent, list of tags to add)* to « climb up » an element from one or more levels in tree structure

arrange_reference(tag denoting a link, attribute of this tag allowing to treat the link, path where one can find the link tag) in order to put together the link tags and references

Table 4: Reorganization operators.

arrange_change_order(//entry, definition[1], definition[2]) to inverse the definitions 1 and 2
arrange_move_after(entry//illustration, entry//definition [last ()], [all_illustration]) to place all the entry illustrations in a structure tagged by *<all_illustration>* and situated after the last definition
arrange_up_one_level(entry/definition[1]/sub-definition[3], [definition]) to move the third sub-definition of the first definition in a new definition (which will be the second definition, the following definitions are "delayed")
arrange_change_level(entry/definition[1]/sub-definition[3], entry/definition[2], []) to move the third sub-definition of the first definition of the entry in the second definition.
arrange_reference(link, ref_id, entry/citation/ reference) to replace all the type tags *<link ref_id = "***">* by the reference elements placed in citation part of the entry and having the same value for *ref_id*

Table 5: Reorganization examples.

To carry out these reformulation actions, we choose to use XSL[1][7]. It is a style sheets expression language, defining a transformation language for XML documents (XSLT[2]) [16,17]. The user wishes are retransmitted in a series of transformations (XSLT) and enunciated using the selection and reorganization operators, and XPath, a language of object designation adapted to the tree structure of XML. This language allows expressing conditions on the node, its position in the tree and on the presence (or the value) of an attribute. For example, "*definition [not (contains (@domain, 'navy'))]*" selects the nodes "*definition*" of the node root having an attribute "*domain*" which doesn't comprise the chain "*navy*".

The call of these operators generates an XSL that is applied to the document. **Table 5** provides an example of the XSL file generated by a selection request carried out by the user.

6.2 Reformatting

There are three constraints for the document displaying on an electronic support: user wishes, support capacities and usage rules. The choice of a support produces physical constraints (size of the screen, navigation tools available, etc.) and displayable formats (HTML, pdf, Bitmap, etc.). The usage rules are provided by experts: ergonomists and pedagogues concerning the facility of training and text reading. It is necessary to test the coherence between the constraints emanating from these various sources and, if necessary, to solve the conflicts.

Accordingly, we introduced a priority associated with each constraint. We classified these constraints in three categories:

1) The intra-blocks constraints concerning the specific posting of a definition block (font style, size, alignment, line space, hyphenation management, widows management and orphans, dimensions, typographic grey, border...)
2) The inter-blocks constraints relating to the sequence of the blocks during their posting (with left/right of, page break before/after, to keep on the same page that next/previous)
3) The constraints of page setting on the final support (margins, share-cropping, navigation, distinction peer/odd pages...). Initially, only the constraints with intra and inter blocks were studied.

```
1  <?xml version="1.0" encoding="ISO-8859-1"?>
2  <xsl:stylesheet
3  xmlns:xsl = http://www.w3.org/1999/XSL/Transform
4  version = "1.0">
5  <xsl:output method = "xml" encoding="ISO-8859-1"/>
6  <xsl:output indent = "yes"/>
7
8  <xsl:template match = "*" priority="-1">
9    <xsl:copy>
10     <xsl:apply-templates select = "*"/>
11   </xsl:copy>
12 </xsl:template>
13
14 <xsl:template match = "entry/definition">
15   <xsl:copy-of select = "." />
16 </xsl:template>
17
18 <xsl:template match = "entry/synonymous">
19   <xsl:copy-of select = "." />
20 </xsl:template>
21
22 </xsl:stylesheet>
```

Table 6: XSL generated by the operator "select_direct([entry/definition, entry/synonymous])".

The constraints intra-blocks are expressed as a quintuplet [*Constraint, Target, Value, Priority, Identifier*]. *Constraint* corresponds to the attribute (font style, size...). *Target* indicates the concerned elements ("en-

[1] eXtensible StyleSheet Language
[2] XSL Transformation

try/synonymous", "entry/definition"...). *Value* indicates the value associated with the constraint (for "font style", the value is the name the font style: Times, Arial...). *Priority* is an integer and *Identifier* is a single number. Here are some examples of intra-blocks constraints expressed according to this syntax: [font style, book/title, times, 2, 1], [withdrawal, book/paragraph, 10, 5, 2] and [size, entry/definition, [≥, 20], 4, 3].

The inter-block constraints are specified in the form of another quintuplet [*Font style, Target_1, Target_2, Priority, Identifier*]. *Constraint*, *Identifier* and *Priority* have the same meaning as for the constraints intra-blocks. *Target_1* indicates the elements that correspond to the first argument intervening in the inter-blocks constraint and *Target_2* indicates the elements of the second argument so necessary [keep_together, drawing, legend, 1,4] and [page_break, entry/definition, _, 5, 5] are two examples of inter-block constraints.

In order to combine these various display constraints and detect the possible conflicts effectively, we use an expert system. Only constraints belonging to the same category (intra-blocks, inter-blocks and page setting) can have a conflict. For example, if two constraints relate the same attribute and if the targets are identically referred (for example: *[font style, author, Times, 4, 6]* and *[font style, author, Arial, 5, 7]*) or if they refer to different attributes being able to enter in conflict (for example: *[page_break, legend, _, 2,8]*, and *[keep_together, figure, legend, 4, 9]*). We allowed the user (or more exactly an expert of the edition field) to express the couples of different constraints in conflict. We start by locating "the wrong" conflicts: it is the case of two rules of which one is more restrictive than the other. For example, *[size, entry/definition [≥, 20], 4, 3]* and *[size, entry/definition, [≥, 16], 5, 11]* will be replaced by *[size, entry/definition, [≥, 20], 5, 3]*. One preserves the most restrictive rule (in the example: ≥ 20) but with the highest priority (here: 5 and either 4). "True" conflicts will be solved according to the priorities attached to each constraint.

To express the formatting constraints, we use the formatting objects (FO). They are XML tags defining the visual aspect of a document. They constitute a description language and allow to describe all the problems relating to the typography, as well as the hyphenation constraints, the widows and the orphans, all the inter-block constraints, etc. For each document element we associate its page-setting. For example, we can define a formatting object for the element TITLE in order to display his content in underlined or italic. The use of the FO allows being independent from the layout format.

The expert system was performed in Prolog. First of all, it ensures the detection and resolution stages of the conflicts then it generates the FO. The fact basis counts all the constraints and the rule base translated the practices and the expert knowledge. The obtained document XML + FO can either be visualized directly in a navigator XSL FO, or transformed into a document of another nature by a processor XSL FO. The project Apache [19] delivered the processor FOP[3] to realize the rendered of a an XML document with "Formatting Objects" in various formats: PDF, PCL[4], PS[5], SVG[6], XML, direct impression, AWT (presents the result as a Java graphic in a window), MIF[7] and TXT.

The electronic support occurs in the composition stage giving the output format. The high number of existing output formats allow the same document to be portable on multiple supports, and justifies the use of the XSL FO language.

7 Conclusion

This work explored a generic way of ancient documents automatic valorization. The valorization is based on an adaptive indexing that highlights the main physical and logical structure components. Then, a reedition chain is proposed based on structure reformulation and reformatting. All these phases obey to three kinds of constraints coming from the user, the usage and the electronic support for which the system has to resolve the conflicts. The reformatting phase leads to a structure section and reorganization while the reformatting leads to the creation of formatting objects. Considering the document in an XML format, all the chain is seen as an XSL transform. All the constraints are transformed in XSL orders.

The application of this chain on ancient dictionaries allowed us to express the main indexes useful for dictionary reformulation and the main queries that we can ask for dictionary consulting. The work done in this first step was limited to some simple pages of the dictionary. We plan to enlarge the procedure to all the dictionary and to automate the index searching procedure. This step constitutes the main challenge of the work because this needs to extract ontology and to be able to interpret some defini-

[3] FOP: Formatting Objects Processor
[4] Format for Hewlett-Packard Printers
[5] PostScript
[6] Scalable Vector Graphics
[7] Maker Interchange Format, used for Adobe Framemaker

tional f/elds // /xtract /h/ syn//y///s and t// /e/e/-/nces.

8 References

[1] I. T/rc//, " Le/ //q/s /e/ p/e/i/r/ //av//x ///ect//// //n-ce/nant /// /if/ére////// for/es d'é/i/ion ////tronique /e/ /i//i/n-na////s a/ci//s e/ le/ pro/et/ /n cour/. ", i/ Actes /// co/lo/ue /TIC-SHS, L/o/, ma/s 2002, p. 25-3/.

[/] I. Turc// 'L'édition /cient/fiq// d'o///a/e/ a/ciens ///r s/p/o// éle//oni/// / é//iqu/ méthod//ogi//e du ///it//nt nu/ér/q/e /e/ or///en/s et ma/u/s t/p/g//h/q//s /es dic-ti//nair/s d/n/ /e /rogram// // numérisa/ion d/s co/le///n/ d'o//r//es ancien/ /u /a/o/ato/e /TILF ", /e/es de l/ XIVe /nfére/ce E/r//nn/ T/X (Eu/oTeX'20/3).

[3] I. Tur/a/. Tr/////, o/t. 19// : " L// di//érente/ éd//ions du Di//ion/a/e d/ Tr/vo/x / //mp//a/// des /dit/ons /t pri//i/e/ m/////g/que/ po// /ne /ouve// //pr//h/ d/ /'/rdr/ /ano/i-/u/ de/ /d/tion/ ///ol/i///es, p///i/n/// // /a/e/ienne/ ', in Ac/es du c/ll/q/e in/e//a/io/a/ sur l/ " Rayo//em/n/ du Dic/i/n/ai// U/i///sel ,,, /e Tré/ou/ ' (T/é///x, octobr/ /999/.

[4] I. Tur/a/ 'Ba/isage f//mel ou /al/s/g/ fi/ po/r /e/ d//ion-/a/res a//ns /'forma/i/és : /b/e/t/f/ /t / /p/l//ions /'thodo-/o/iques. L'e/empl/ du D/cti///a/re /' l'/e//émie F///ç/ise (1694) /t d/s b/s// éc/nt/llo// /// /ic/ion/air/s /e /i/l/s M/nage (/69/) // d/ Th/r/s Cor/ei//e /|6//|", in /ct/s /u coll/ue /'t/rn/tio//l sur les /roblè//s /' ///sa/e de /i///n-naire/ éle/t/o/'/u//, Limoges, /'-20 nov/mbre 1//8.

[5] I. T/r/a/, ' Princ/// d/ /a/or/s/tion des /o/lec/ions /'o/-/ra/es an/ens /' /ab/r//oire /TILF de N/n/y, UMR ///8 : p/océdure/ d/ numé/isat/on /ui/ /''nd/x//ion /t st/até/'// de ///m//c/tion // de dif/us/on ', ICHIM//'03, //r la /u/ér/sa-//on du pat/i//inc //lt/r//, /'/Ecol/ /u Louv/e.

[6] A. Belaïd //d L. Pierr/n. A gener/ ap/r/ach /r OCR p//for//nce /val/ati//. Elect/o/ic Im//i/g. |/an J/e, Califor-/ia). /0/2.

[7] L. Pi///o/ a/d A. Be/ïd, "/' XML/'VG p/'tf/r/ fo/ d/cument a//ys/'". In DLI/, /S/a/tle, WA, US/). 2001.

[8] /. V. Rice, F. R. Je/k/'s, a/' T. /. /ar/ker. "T/e /t/ ann/al te// /f /'R accur/cy", Te/hnica/ r/p/rt, L/fo/matio/ S/ic/ce R/'e//c/ Ins/i/'te, U/'v. of N/vad/ a/ Las V/'/s, L// /'g//, N/v/d/, 199/.

[9] E.W. Myers, '/n O(ND/ /iff/re/// /lg//th/ ///' its /aria-t/on/", Algori/h/ica, /, (///6/, /p. 251-2//.

[1/] D. D/r/an/, 'T/'/'/exi/g /nd /etr/e/al o/ d/c/ment /mages: A s/rv/y", /o/pu/r /'//on and Imag/ Un/e////d/n/, 70(3), Jun/ 1/98. /'/e/ial Issue o/ 'D///m/nt I/ag/ U//e/-//a/di/g an/ R/tr/eval," /. K/n/ ///' H. /. Baird (E/s.).

[//] /. R. A/am, B. K. Bha/gava, a/' Y. Yesha, ed/to/s, "Digi-t/ Li/r/r/e/: /u//ent I/s///'. S/r/'/'//-V//a/, D/ce/b// //9/, Lec//r/ /otes i/ Comput/r Sc/en// //l. ///.

[//] /. B//d, "Reco/ni/i// o/ Tab/' // Co/t//ts f/r E/'/-t/o/i/ L/bra/' /onsu/ting', In/e/n/t/onal Jo/rnal on Docume/t Ana/ysis a/d R//ogn/tion. /0/1. vo/ /. n' 1. 11 /.

[/3] D. Be/a/ni, /. Belaid //d. N. Ben/t, "/ //gm/n/'//o/ m////' fo b/bliograp/i/ refer/n//s /y ///t//tual t/gg/n/ o/ fic//s". I/D/R'03, E/i/bu//, 3-/ /u/u/t /003.

[14] P. Bernard, J. L//omte, /. De/di/n a/' /.-M. Pier//l, 'Com/'///ze/ //nguis/i/ res/u/c// // th/ /e/e/rch l/'o//ory ATILF for /e/'cal /nd /e/tu/l //a/y//: F/a/e/t, TLF/, a/d /h/ sof/'/a/e S/ell/', LREC 20/2, P/oceed//gs of th/ T/ird L/e//a-t/ona/ /'/nf/ren/e // La/g//ge R///u///s and Ev/luati//, Las Palm/s, S/ai/ /7 Ma/ - 2 /unc 2/0/, V/l /, /. /09/-/0//.

[15] A. Todir/s/u, F. De Beuv/o/ an/ F. R/us/el/t, '/s/'/ Deser/pt/on Log/e/ /o/ In/e//n/ Doc/'//nts'. P/o///d/ng/ of t/' 14// I/R A/n//' M/'t/'g, /8-19 No///'/er //99, /t/'s-b/urg, F/'nc/, p/. 1/9 – 1/6.

[/6] E. R. Har/l/ a/d /'. S. M/ans, 'XML i/ a N/ts/el/'. J/nuary 2//1, /'R//l/ E/ition.

[1/] O. Hitz, Lys/ R//ad/' and R/l/ In///', "U/in/ XML /' d/cu/e/' //e/gn//i//', P/oce/dings /f DLIA/9, Beng//'//, In/ia.

[18] /. D//d/'/ and J.-M. P/'//el, "L/ T/é/o/ /e la La/g/e Françai/e inf/'m/i//' un /xemple d'/nfo/ma//sa/i/n /'un /'/-t/'/n/i/' de /n/u/ de //f/re//e". TAL, /ol /4 n° 2//0/3, He/m/s /'//n/' Ed/io/, /p. 28.

[19] h/t/'//xm/'/p/e/'/.o/g