WikiLexicographica. Linking Medieval Latin Dictionaries with Semantic MediaWiki

Bruno Bon, Krzysztof Nowak

Wiki Lexicographica. Linking Medieval Latin Dictionaries with Semantic MediaWiki

Bruno Bon[1], Krzysztof Nowak[2]
[1]Institut de recherche et d'histoire des textes (CNRS), [2]Institute of Polish Language (Polish Academy of Sciences)
[1]Paris (France), [2]Kraków (Poland)
E-mail: bruno.bon@irht.cnrs.fr, krzysztofn@ijp-pan.krakow.pl

Abstract

The paper demonstrates results of the survey of which the main aim was to scrutinize consequences of adopting wiki model in alignment of medieval Latin dictionaries. In the first section, the objectives of the project as well as its methodology are presented. As a framework Semantic MediaWiki (SMW) has been chosen. For the purpose of the research several entries from four dictionaries were selected. In the following chapters authors scrutinize presentation, search, and collaboration features provided by SMW. Authors demonstrate how intrinsic wiki concepts, such as namespaces, templates, property-value pairs etc., may be employed in macro- and microstructure display. Next, alternative modes of accessing lexicographical data are presented such as maps, timelines, charts etc. After that search capabilities are analysed, among which the most important appear to be semantic properties search and faceted browsing. Lastly, the paper focuses on different ways SMW can encourage researchers to collaborate and enrich dictionary content.

Keywords: medieval Latin; wiki-interface; multilingual dictionaries linking; dictionary alignment

Introduction

It is in 1913 that the idea of Pan-European dictionary of medieval Latin had been clearly expressed by research community, but it is not until early 1920s that the work began on preparing *Novum Glossarium Mediae Latinitatis* which would cover four centuries (IX-XII) of Latin language use (Langlois, 1924), replacing older and at that time already obsolete *Glossarium* of Charles du Fresne, sieur du Cange (Du Cange, 1883). From the very beginning it was also clear that, due to various periodization of Middle Ages, national dictionaries should be compiled as well. This is the reason why there now exist a dozen of dictionaries which vary not only in their chronological (500-1600 AD) and regional (from Spain to Poland; from Sweden to Italy) coverage, but also in their advancement (three of them completed, but the majority of projects still in progress).

Yet, it is with the advent of the e-lexicography that the founding idea of the common dictionary of European Latin should be reconsidered. A first step was made during the congress of medieval Latin lexicography held in Barcelona in 2004, where several elements of microstructure were proposed as a basis for dictionaries' alignment, among them headword, etymology and sense definitions (Heid, 2004). This proposal, however, was put forward without major consideration of such „technical" issues as software framework, encoding schema or data structure. During the next years community witnessed emergence of several e-lexicography and e-corpora projects, among which one should mention:

(1) electronic editions of Du Cange's *Glossarium*[1], *Novum Glossarium Mediae Latinitatis*[2], dictionaries

1Accomplished, available at http://ducange.enc.sorbonne.fr/.
2In progress, due to be finished in 2013, more information on http://glossaria.eu.

of medieval Latin from Polish[3] and Catalan[4] sources; (2) corpora of medieval Latin in Catalonia[5], Galicia[6], and Poland[7].

This rapid development, in turn, has aroused interest in encoding standards and lexicographical data interoperability. In the same time, several institutional enterprises have been launched in order to foster research collaboration and sustain data exchange, one of them being COST Action 1005 „Medioevo Europeo"[8]. Its goal, as the project's description says, is the development of a so-called „Virtual Centre of Medieval Studies", a common interface for querying until now dispersed databases, text collections, library catalogues etc. After being appointed as experts of the project on behalf of medieval Latin dictionaries teams, we put forward the idea of a wiki-based tool. In the present paper, we discuss a working prototype of such a wiki, an interface and research environment which could potentially serve as a means for unified edition and query of medieval Latin dictionaries and lexical databases.

Objectives and procedures

As a framework for our survey we choose MediaWiki (MW)[9] which is best known as an application running in the background of Wikipedia. Once installed, it was subsequently supplemented with a bunch of plugins of which the essential one was Semantic MediaWiki (SMW)[10], an extension that enhances MediaWiki with semantic dimension, enabling advanced data annotation and as a consequence their finer retrieval. It allows explicit declaration of exact meaning of the data contributed to the wiki page by annotating it with property name of which value are the data:

[[Property_name::Property_value]]

eg. [[Headword::Mandragora]].

Software choice was driven by several project's goals and objectives which can be summarized as follows:

1. Software should already exist and be free. There were no funding envisaged in the project for writing software from scratch, since it is treated as a means of fostering discussion rather than proper goal of the project.
2. Software should be open-source. The lexicographical and corpus data in the emerging projects, in majority of cases, are (or will soon be) available under liberal licensing models, so should be the tools used in their retrieval. Since the goal of the project is to foster collaboration and data exchange, participating projects cannot be excluded or limited by the use of binary file formats or infrastructures closed to further refinement.
3. Stable development, community support. In order to ensure project longevity, the tool should be actively developed and supported by stable number of code contributors.
4. Compliant with dictionary-type data.
5. Multilingual interface.
6. Collaboration oriented.
7. Easy to use.

MW and SMW are not only free and fully open-source, but they have been also created with encyclopedia-like data in mind and provide internationalized interface. Thanks to its popularity, MW may encourage less advanced users to actively collaborate.

SMW, although steadily gaining popularity, has not been yet employed in vast lexicographic projects. According to the list of sites using SMW[11], extension has been implemented in such projects as *Liddell-Scott-Jones Ancient Greek Lexicon Edition*[12], *An interactive online etymological dictionary of Lepontic*[13], or *Neuroscience Lexicon*[14]. Any of them was known to the authors in the early 2012 when works on integrated query interface were about to start.

For the purpose of the present paper, 4-6 entries from four dictionaries were chosen and subsequently encoded by typing wiki syntax code. Whenever possible, lexicographical content was being passed to the formerly created templates[15] which automatize not

3 *eLexicon Mediae et Infimae Latinitatis Polonorum*, in progress, due to be finished in mid-2014, http://scriptores.pl.
4 In progress, more information at http://gmlc.imf.csic.es/.
5 *CODOLCAT. Corpus Documentale Latinum Cataloniae*, http://gmlc.imf.csic.es/codolcat/index.php.
6 *CODOLGA. Corpus Documentale Latinum Gallaeciae*, http://www.cirp.es/codolga/.
7 *Fontes Mediae Latinitatis Polonorum*, in progress, due to be finished in 2016, more information at http://scriptores.pl.
8 http://www.medioevoeuropeo.org/
9 http://www.mediawiki.org
10 http://semantic-mediawiki.org/
11 http://smw.referata.com/wiki/Special:BrowseData/Sites.
12 http://lsj.translatum.gr/.
13 http://www.univie.ac.at/lexlep.
14 http://neurolex.org/
15 http://www.mediawiki.org/wiki/Help:Templates.

only text formatting, but also semantic annotation of data. For instance, when a content author types {{headword| mandragora}}, a template „headword" is called with first argument set to „mandragora". Once triggered, the template:

(1) sets property „headword" to „mandragora" and displays text string „mandragora" on entry page [16] ;

(2) sets property „headword_canonical" to „mandragora" without displaying word itself on the entry page[17] .

Annotation task has been primarily conducted by authors of this paper with the help of Renaud Alexandre (IRHT CNRS). Next, several members of other lexicographical teams have been familiarized with wiki editing interface (especially wiki syntax) and asked to correct or edit entries from scratch.[18]

Macrostructure

Although the main goal of the presented database is enabling unified retrieval of dispersed dictionaries, the provenance of lexicographical data should be always easily traceable. Firstly, it permits to acknowledge the institutional and research teams' effort of which machine-readable dictionaries are the result. Secondly, it provides users with the possibility of limiting their search results. In our prototype separation of dictionary entries has been assured by resorting to the mechanism of wiki namespaces[19] , each entry being preceded by 2-letter prefix indicating the dictionary it originates from: namespace:entry_headword, eg. for Latin word *decipula* 'a snare, trap', a full page title is *PL:Decipula* which results in following entry link: .../index.php?title=PL:Decipula. This separation allows user to browse each dictionary in traditional way by looking at the entry list (.../index.php?title=PL).

Main namespace has been reserved for so-called „super-entries", i.e. entries of unified dictionary which serve as an index for all headwords. Super-entry page for headword *depost*, then, will provide the list of dictionaries the word is attested in (with appropriate links), as well as other information about the word in question that can be retrieved by means of the embedded queries. This information is presented in form of timelines and maps of the attested word occurrences which have been extracted from respective dictionary entries:

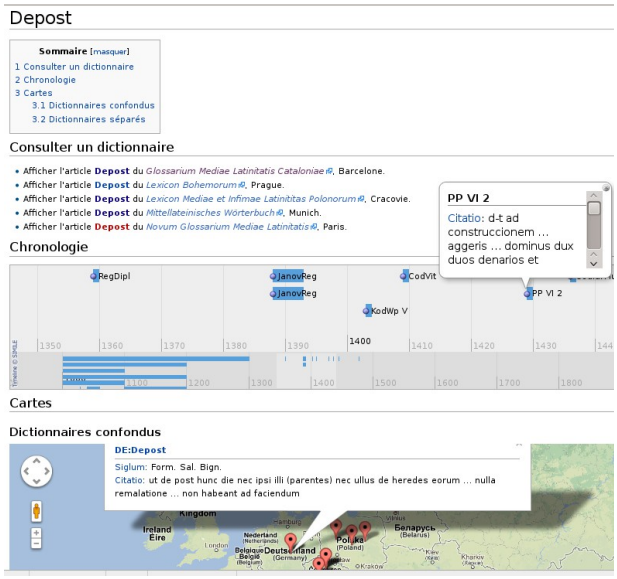

Figure 1: Super-entry page

Spatio-temporal information retrieval which is in fact a heart of the WikiLexicographica was possible, because each source quotation is stored as so-called „semantic internal object" (SIO)[20] , a complex data structure which permits encapsulation of multiple property-value pairs. SIOs include in particular:

(1) reference to the entry they belong to;

(2) source reference abbreviation;

(3) bibliographical data (page, verse etc.);

(4) proper citation;

(5) date of text composition;

(6) geographical provenance of the text.

```
{{#set_internal:CitInt
|Cit_siglum={{{1}}}
|Cit_ref={{{2}}}|
|Cit_datum={{{3|}}}
|Cit_citatio={{{4|}}}
|Cit_start={{#if:{{{5|}}}|{{{5|}}}|{{#show:{{NAMESPACE}}:{{{1|}}}|?Fons start}} }}
|Cit_end={{#if:{{{6|}}}|{{{6|}}}|{{#show:{{NAMESPACE}}:{{{1|}}}|?Fons end}} }}
|Cit_genus={{#show:{{NAMESPACE}}:{{{1|}}}|?Fons genus}}
|IsPartOfLemme={{NAMESPACE}}:{{PAGENAME}}
|Cit_geo={{#show:{{NAMESPACE}}:{{{1|}}}|?Fons geo}}
|Cit_coordinates={{#geocode:{{#show:{{NAMESPACE}}:{{{1|}}}|?Fons geo}} }}
}}
```

Figure 2: SIO structure

16The code in template is [[Headword::{{{1}}}|{{{1}}}]], where number stands for argument order number.

17The code being {{#set:Headword_canonical={{#regex: {{{1}}}|/\w+\s*)+/}}}}. Canonical form of entry headword is computed by applying to a full headword a simple regular expression which gets rid of symbols, numbers etc.

18Their names can be found in Acknowledgment section of the present paper.

19 http://www.mediawiki.org/wiki/Namespaces.

20http://www.mediawiki.org/wiki/Extension:Semantic_Internal_Objects.

Since in medieval Latin dictionaries neither chronological (5), nor geographical (6) data are explicitly declared for each quotation, values of these properties are usually computed from information provided in source description pages which form an essential part of integrated dictionary macrostructure.

Source pages belong to the same namespaces as the entries themselves. They are distinguished from them by category attribution: whereas entries belong to the category *Voces* (*lat.* 'words'), source description pages are marked as *Fontes* (*lat.* 'sources'). Source description page consists of manually typed or database extracted metadata which is subsequently passed to a bunch of embedded queries. So, for instance, wiki syntax:

{{fons|EU|France, Lille||1150|1200|Alan. Ins. elucid.| Elucidatio in Cantica canticorum. – PL 210 col. 51-110| commentarius}}

results in a page (Figure 3) where one can see a source provenance map, bibliographical record and so on. The most interesting, however, is the section *Citationes* (*lat.* 'quotations') where user can find a list of headwords in which the source in question is referenced with its appropriate quotations. As long as we have not at our disposal complete, research-driven corpus of medieval Latin, dictionaries can be considered provisional corpora including a selection of medieval Latin literature in seemingly good editions.

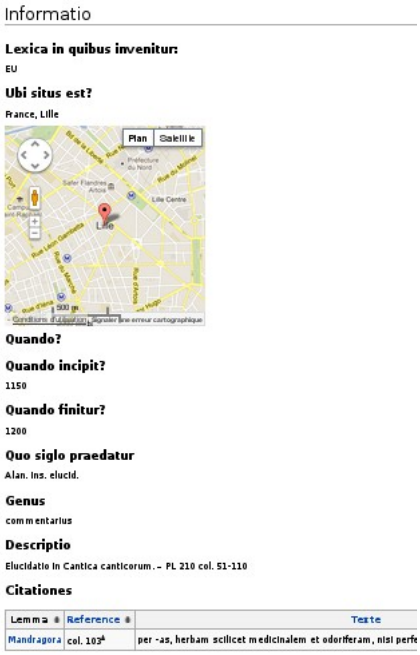

Figure 3: Source description page

Sources and their quotations can be subsequently browsed traditionally in the form of alphabetically ordered lists. However, user can also:

(1) sort by frequency in dictionaries;

(2) browse them on a timeline;

(3) access them on a google map.

In our survey, the last form of lexicographical data analysis has been enriched (Figure 4) thanks to the map layers provided by the project *Digital Atlas of Roman and Medieval Civilization*[21] . One is now able to view source citations in the context of administrative boundaries of the medieval world and in the light of regional variation of medieval intellectual culture.[22]

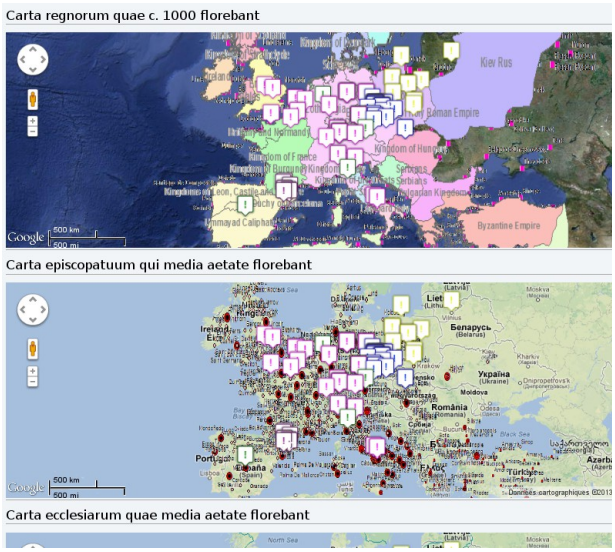

Figure 4: Map layers

Microstructure

Medieval Latin dictionaries have as their primary public the research community, a fact which too often means that their entry structure is far from being user friendly. In our wiki we attempt to address this problem by providing two parallel access points to dictionary microstructure. The first perspective that the user is faced with when visiting entry page is a basic one (Figure 5). It comprises essential lexicographical information, such as graphic forms, inflection type, gender, abbreviated sense definitions. The basic view

21 http://darmc.harvard.edu/icb/icb.do.

22 This was possible due to the use of such layers as „medieval kingdoms", „universities C12-C15" etc.

tab, though, is also a place where user is given an overall picture of word occurrences. Entry source citations are here conveniently epitomized in text type chart, timeline, and map. Now, only a quick glance should suffice to estimate in what medieval genres, when and where the word in question would be cited most frequently.

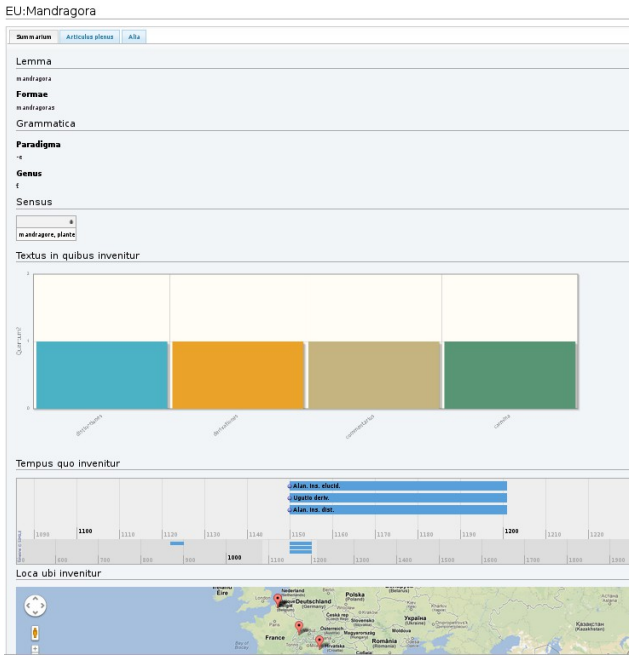

Figure 5: Dictionary entry (basic view)

In the next tab, one can consult full entry with all the idiosyncrasies each dictionary editorial system is affected with. The relative variety of typographic conventions, as well as different levels of data explicitness render preserving original entry display for each dictionary a very difficult, if not impossible task. This is one of the reasons why, in our opinion, wiki interface should not be considered a means of text-oriented digitization of lexicographical work. The other major problem stems from the relatively flat textual data representation in Semantic Mediawiki: it is not easy (if possible) to properly reflect nested tree structure of the dictionary entry by means of wiki syntax only. It can seem a serious limitation, considering that medieval Latin lexicographers tend to make a heavy use of sense nesting in order to account for semantic change. Thus, more appropriate approach seems to be data-oriented recompilation of source data into desired output format, even if some of the original data (in particular, formatting) are lost. Burden of preserving original work in its typographic, sequential etc. order may, then, be arguably shifted towards each separate project rather than integrated query interface.

Next tab of each entry gathers links to other linguistics resources (Figure 6). First of all, user can easily verify whether the same headword exists in other dictionaries included in the wiki. Secondly, one is proposed to search the headword in other Latin dictionaries. Lastly, links to textual corpora and text collections are provided. This is also where lexicographical data may be enriched with world knowledge. The example of *mandragora* 'mandragora' shows possible fields of lexicon-encyclopaedia interface enrichment: links point to plant taxonomy pages, to Wikipedia entry on *mandragora*, and to the images accessible in Wikimedia Commons.

Figure 6: Dictionary entry ('Other resources' tab)

Search and Browse Capabilities

Search and browse capabilities of the presented infrastructure are partly known from Wikipedia and its derivatives. It comes of no surprise that entries may be retrieved by means of simple full text search. As in Wikipedia, when typing word beginning in ajax-based search form, user is given suggestions. Naturally, it is

also possible in advanced mode to limit search results to specific namespaces, i.e. dictionaries.[23]

Framework which is object of the present study seems to reach its full potential, however, thanks to semantic layer provided by SMW. Semantic properties embedded in each entry can be browsed, for instance, thanks to the factbox displayed on the bottom of the entry page (Figure 7).

Figure 7: Entry page factbox

After clicking „magnifying glass" icon near the value of each property, user is taken to the page where all the entries with the same value set for selected property will be listed. So, for example, in the case of *mandragora*, if one clicks the zoom icon near the value *f.* (*femininum*, *lat.* 'feminine') of property „gender", one is redirected to the page where all the feminine substantives from all the dictionaries included in the wiki are listed. Similar results can be obtained from the „Special Page" on which user can process a simple semantic search, by directly specifying in two-fields form the property and its value s/he is looking for (Figure 8).

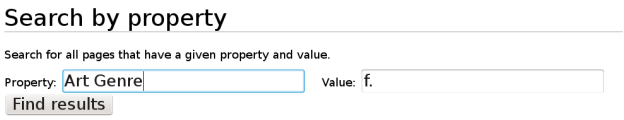

Figure 8: Direct search for semantic properties' values

More advanced semantic queries can be formulated from within two other search interfaces available in SMW-based wiki which are accessible from „Special Pages": Special:Ask and Special:BrowseData. The first (Figure 9) requires from the users a basic knowledge of SMW syntax, but it also provides them with numerous output formats they can choose from, e.g. different types of charts, timelines, maps, tables, slides etc.[24]

Figure 9: Semantic search (Special:Ask page)

„Special:BrowseData" (Figure 10), on the other hand, includes search patterns envisaged by each wiki creators and depends only on their creativity, user requirements, and last but not least, time or funding limitations. It enables faceted browsing of semantic properties of wiki data. In the case of our framework, the data in question are source and entry pages. The latter can be browsed according to part of speech they represent, inflectional type, gender, domain of use etc., while the first according to all the metadata which were previously mentioned.

Figure 10: Semantic faceted browsing (Special:BrowseData page)

23Full-text search capabilities may be enhanced by using plugins list at http://www.mediawiki.org/wiki/Fulltext_search_engines. They have not been subject to the tests in the present survey.

24Display of search results is provided by Semantic Result Formats plugin (http://semantic-mediawiki.org/wiki/Semantic_Result_Formats).

Collaboration

From its beginning wiki-based interface that is the subject of present study has been conceived as a means of promoting collaboration between researchers of different expertise in medieval studies. Lexicographical data enriched with encyclopaedic information extracted from knowledge databases may be a good starting point for prospective framework of medieval culture research. Users' contributions, as we hope, should be encouraged by the very fact of reusing Wikipedia-like interface, with its well-known collaboration feature, namely „Discussion page". Despite the fact that MediaWiki was created for projects in which anonymous editing is welcomed, one can, though, (1) impose access and edition limitations in order to get rid of the acts of vandalism and (2) provide admin users with right to accept or deny any changes. Users who are familiar with wiki syntax can be assigned edit rights and contribute to entries or source pages without any difficulty. However, even non-technical oriented users may contribute to the wiki, if they are given chance to use simple edit forms. In the framework demonstrated in the present paper, this is the case of the source pages which can be modified in traditional way, by typing wiki syntax code, or by filling in forms provided by wiki developers (Figure 11).[25]

One can expect that the wiki will be fed at first with data from ongoing lexicographic projects, and next enriched by the users themselves. Batch import of lexicographical information from existing dictionaries may be carried out, e.g., by means of RDF import plugin. In spite of the fact that the dictionaries under analysis may seem to differ essentially, at least as far as their micro- and macrostructure are concerned, their electronic versions are considered to be TEI compliant, and follow rules indicated in the chapter of TEI Guidelines devoted to encoding of machine-readable dictionaries (TEI Consortium, 2013).[26] Since WikiLexicographica has to serve as a common interface of data retrieval, shared information schema should be conceived as well. Extent of data extraction could be then decided according to time or financial limitations, however burden of mapping between particular schemas and the common one need to be shifted to each lexicographic team.

The second main contributor of WikiLexicographica is expected to be research community, namely philologists, linguists, historians, palaeographers, briefly, all those who work with medieval Latin texts. Apart from simple form or meaning corrections and additions, users may be encouraged, e.g., to propose adding new words found in their sources or deleting existing ones if manuscripts deny lexicographer's reading; to supply entries with world knowledge facet which in turn can greatly support text comprehension; to create links between words by making their relations explicit, and so on.

Conclusions

MediaWiki, the software underlying Wikipedia, enhanced with semantic data annotation capabilities offered by Semantic MediaWiki extension appears to be a tool mature enough to serve as an interface for lexicographical data retrieval. It provides presentation and collaboration features an average Wikipedia user can already be familiar with. Interface popularity itself is likely to encourage contributions even from those less technical-oriented researchers. It is, however, retrieval of semantic properties that should attract a major interest of researchers since charts, timelines and maps, as well as embedded queries, offer a fresh and inventive look at the lexicographical data.

Acknowledgment

This work was supported by following grants: ANR Omnia "Outils et Méthodes Numériques pour l'Interrogation et l'Analyse des textes médiolatins"; NCN 3736/B/H03/2011/40 „eLexicon Mediae et Infimae Latinitatis Polonorum".

Collaboration on this paper was made possible by support of COST Action 1005 „Medioevo Europeo" (www.medioevoeuropeo.org).

Content of the WikiLexicographica was partially typed by members of respective dictionaries teams: Renaud Alexandre (Novum Glossarium), Susanna Allés Torrent (Glossarium Mediae Latinitatis Cataloniae), Pavel Nývlt

25This is possible thanks to the plugin called Semantic Forms (http://www.mediawiki.org/wiki/Extension:Semantic_Forms).

26So far no attempt has been made in order to standardize medieval Latin dictionaries schema according to, for instance, Lexical Markup Framework.

(Lexicon Bohemorum), Marta Segarrés Gisbert (GLMC).

References

Du Cange, C.D.F. (1883). *Glossarium mediae et infimae latinitatis (Editio nova aucta pluribus verbis aliorum scriptorum a Leopold Favre) conditum a C. Du Fresne, domino Du Cange, auctum a monachis ordinis sancti Benedicti; cum supplementis integris D. P. Carpenterii, Adelungii, aliorum suisque digessit G. A. L. Henschel*, Niort: L. Favre.

Heid, C. (2004). Table ronde "Lexicographie et informatique" (Barcelone, 2 juin 2004). *ALMA, Bulletin du Cange*, 62, pp.327–332.

Langlois, C.-V. (1924). Historique sommaire de l'entreprise de 1920 à janvier 1924. *ALMA, Bulletin du Cange*, 1, pp.1–15.

TEI Consortium (2013). TEI P5: Guidelines for Electronic Text Encoding and Interchange. Version: 2.5.0. Last modified: 26th July 2013. Accessed at: http://www.tei-c.org/Guidelines/P5/.