Encoding the Specificities of Encyclopedias

Alice Brenon

Encoding the Specificities of Encyclopedias

Alice Brenon [1,2]

[1] ICAR, CNRS, UMR5191, 69342
[2] Univ Lyon, INSA Lyon, CNRS, UCBL, LIRIS, UMR5205, F-69621

Abstract This chapter illustrates the fundamental differences between dictionaries and encyclopedias by documenting the process of devising an encoding scheme and applying it to a late-19th century encyclopedia, "La Grande Encyclopédie" (hence *LGE*). The effort, made in the context of project DISCO-LGE, consisted in working from an OCRised version of the pages in XML-ALTO to produce a fully XML-TEI-compliant encoding of the individual articles. Although the TEI guidelines include a specialised module for dictionaries which was identified as a promising tool for the task, systematic traversal of the schema using graph search methods revealed some limitations when used to encode this text. These shortcomings are reviewed and illustrated on a series of examples. An alternative encoding remaining within the *core* module of TEI is then proposed and demonstrated on articles from *LGE* containing key features. Finally, different strategies followed by other projects are discussed.

Keywords digital humanities, XML-TEI, dictionaries, encyclopedias

1 Introduction

Although both terms have been used rather interchangeably over the past few centuries, a dichotomy is now commonly being made between dictionaries and encyclopedias. A simple opposition can easily justify this distinction: dictionaries define words and tell one how to use them while encyclopedia usually go into longer development to give a more comprehensive and scientific understanding of the concept being defined. This common intuition links back to the entry written in the *Encyclopédie ou Dictionnaire raisonné des sciences des arts et des métiers* (hence *EDdA*) by d'Alembert (n.d., article DICTIONNAIRE, volume 4) who opposes three kinds of dictionaries: one to define *words*, the second to define *facts* and the last one to define *things*, corresponding respectively to language, history, and science and arts dictionaries. The first type corresponds to modern dictionaries while the two others are similar to what one expects to find in an encyclopedia.

However, d'Alembert himself doesn't think of these boundaries as very strict and he hints at the extreme difficulty in merely defining words without going into semantics and philosophical considerations:

> un dictionnaire de langues, qui paroît n'être qu'un dictionnaire de mots, doit être souvent un dictionnaire de choses quand il est bien fait

("a language dictionary, which appears to be only a word dictionary, must often be a thing dictionary when it is made properly"). A similar criticism is made by Haiman (1980, p. 331)

who attacks no less than six criteria on which dictionaries and encyclopedias are generally opposed to reach the conclusion that there is no distinction between them because "dictionaries *are* encyclopedias". Regardless of the validity of his reasoning, it only proves one inclusion: that perhaps, dictionaries would be a special case of encyclopedias. This, as will be shown, does by no means imply that conversely encyclopedias are dictionaries.

XML-TEI is a set of guidelines, tools and tranining resources collectively developped by the TEI Consortium (2023) to represent text in a highly structured and machine-readable format. Its toolbox has a modular structure consisting of optional parts each covering specific needs such as the physical features of a source document, the transcription of oral corpora or particular requirements for textual domains like poetry, or, in the case at hand, dictionaries. The intrinsic complexity of dictionaries has been well identified since the inception of the project (*Previous Drafts of the Guidelines*, n.d.) and Ide & Véronis (1995) underline the amount of work which went into the third version of the guidelines (P3) to provide a toolbox both general and expressive enough to account for the variety of conventions found in dictionaries. This module has been successfully used to encode both historical (Bohbot et al., 2018; Williams, 2017) and digitally native dictionaries (Bowers & Romary, 2018). In addition, a specific guidelines tailored at encoding dictionaries named TEI-Lex0 has also been published (Bański et al., 2017).

The TEI effort is described by Ide & Véronis (1998) as "first steps" to reach a standard to encode corpora and lay a common basis for corpora comparison and reuse. They point some light inconsistencies in the design, remark that there is generally more than one way to encode a given text in XML-TEI and identify nine criteria to design a sound standard. Their claims are backed by concrete examples of encoding situations but give no idea of the prevalence of the issues reported. In fact, the sheer complexity of the guidelines can make it hard to ascertain whether a particular element structure is impossible to represent (not finding a suitable encoding is not a proof that there is none). This chapter will use results from graph theory to make a systematic study of the possibilities and shortcomings of the TEI *dictionaries* module, hence providing an additional proof that encyclopedias are not dictionaries and that the inclusion claimed by Haiman is a strict one.

2 Context of the study

To give a better understanding of this research, this section describes the aims of the project from which it stems before giving a short history of the term *encyclopedia* and underlining the known differences between dictionaries and encyclopedias which constitute the starting point of this investigation.

2.1 CollEx-Persée Project DISCO-LGE

The project (https://www.collexpersee.eu/projet/disco-lge/) set out to study *La Grande Encyclopédie, Inventaire raisonné des Sciences, des Lettres et des Arts par une Société de savants et de gens de lettres* (hence *LGE*), an encyclopedia published in France between 1885 and 1902 by an organised team of over two hundred specialists divided into eleven sections. This text comprises 31 tomes of about 1200 pages each and according to Jacquet-Pfau (2015, p. 88 et seq.) was the last major french encyclopedic endeavour directly inheriting from the prestigious ancestor that was the *EDdA* published by Diderot and d'Alembert 130 years earlier, between 1751 and 1772.

The aim of the project was to digitise and make *LGE* available to the scientific community as well as the general public. A previous version of this encyclopedia was partially available on Gallica (https://gallica.bnf.fr/services/engine/search/sru?operation=searchRetrieve&collapsing=disabled&query=dc.relation%20all%20%22cb377013071%22) but lacked in quality and its text had not been fully extracted from the pictures with an Optical Characters Recognition (OCR) system. This prevented an exhaustive study of the text with textometry tools such as TXM (Heiden, 2010). As a prelude to project GEODE (https://geode-project.github.io/), the goal of DISCO-LGE was to produce a digital version of *LGE* with a quality comparable to the one of l'*EDdA* provided by the ARTFL (http://artfl-project.uchicago.edu/) project in order to conduct a diachronic study of both encyclopedias.

2.2 *Encyclopedia*

If the word "encyclopedia" is now part of everyday vocabulary and has a slightly different meaning from dictionary, it was much more unusual and in fact controversial when Diderot and d'Alembert decided to use it in the title of their book, while having to coordinate them both in the full title of the *EDdA* which is probably the most famous work of the genre and a symbol of the Age of Enlightenment.

The definition given by Furetière in his *Dictionnaire Universel* in 1690 is still close to its greek etymology: a "ring of all knowledges", from *κύκλος*, "circle", and *παιδεία*, "knowledge". This meaning is the one used for instance by Rabelais in *Pantagruel*, when he has Thaumaste declare that Panurge opened to him "le vray puys et abisme de Encyclopedie" ("the true well and abyss of Encyclopedia"). At the time the word still mostly refers to the abstract concept of mastering all knowledges at once. Furetière adds that it's a quality one is unlikely to possess, and even seems to condemn its pursuit as a form of hubris: "C'est une témérité à un homme de vouloir posséder l'Encyclopédie" ("it is a recklessness for a man to want to possess Encyclopedia").

Beyond this moral reproach, the concept that pleased Rabelais was somewhat dated at the end of the 17th century and attacked in the *Dictionnaire Universel François et Latin*, commonly refered to as the *Dictionnaire de Trevoux*, as utterly "burlesque" ("parodic"). The entry for "Encyclopédie" remained unchanged in the four editions issued between 1721 and 1752, mocking the use of the word and discouraging his readers to pursue it. In that intent, he quotes a poem from Pibrac encouraging people to specialise in only one discipline lest they should not reach perfection, based on an argumentation that resembles the saying "Jack of all trades, master of none". It is all the more interesting that the definition remains unaltered until 1752, one year after the publication of the first volume of the *EDdA*. The Jesuites who edited *Dictionnaire de Trevoux* frowned upon the project of the *EDdA* which they managed to get banned the same year by the Council of State on the charge of attempting to destroy the royal authority, inspiring rebellion and corrupting morality in general. There is much more at stake than words here, but the attempt to deprecate the word itself is part of their fight against the philosophers of the Enlightenment.

The attacks do not remain ignored by Diderot who starts the very definition of the word "Encyclopédie" in the *EDdA* itself by a strong rebuttal. He directly dismisses the concerns expressed in the *Dictionnaire de Trevoux* as mere self-doubt that their authors should not generalise to anyone, then leaves the main point to a latin quote by chancelor Bacon (Lojkine, 2013, p. 5), who argues that a collaborative work can achieve much more than any talented man could: what could possibly not be within reach of a single man, within a single lifetime

may be achieved by a common effort throughout generations.

History hints that Diderot's opponents took his defence of the feasability of the project quite seriously, considering the fact that they got the *EDdA*'s privileges revoked again six years after its publication was resumed (Moureau, 2001). As a consequence, the remaining ten volumes containing the text of the articles had to be published illegally until 1765, thanks to the secret protection of Malesherbes who — despite being head of royal censorship — saved the manuscripts from destruction. They were printed secretly outside of Paris and the books were (falsely) labeled as coming from "Neufchâtel" (*sic*). Following the high demand from the booksellers who feared they would lose the money they had invested in the project, a special privilege was issued for the volumes containing the plates, which were released publicly from 1762 to 1772.

In any case, in their last edition in 1771 the authors of the *Dictionnaire de Trevoux* had no choice but to acknowledge the success of the encyclopedic projects of the 18th century. In this version, the definition was entirely reworked, mildly stating that good encyclopedias are difficult to make because of the amount of knowledge necessary and work needed to keep up with scientific progress instead of calling the effort a parody. It credits Chambers' *Cyclopædia* for being a decent attempt before referring anonymously though quite explicitly to Diderot and d'Alembert's project by naming the collective "Une Société de gens de Lettres" and writing that it started in 1751. Even more importantly, two new entries were added after it: one for the adjective "encyclopédique" and another one for the noun "encyclopédiste", silently admitting how the project had changed its time and the relation to knowledge itself.

2.3 A different approach

If encyclopedias are thus historically more recent than dictionaries they also depart from the latter on their approach. The purpose of dictionaries from their origin is to collect words, to make an exhaustive inventory of the terms used in a domain or in a language in order to associate a *definition* to them, be it a phrase explaining it or a translation in another language for a foreign language dictionary. As such, they are collections of *signs* and are more concerned with the linguistic level of things. Entries in a dictionary often feature information such as the part of speech, the pronunciation or the etymology of the word they define.

In the full title of the *EDdA*, the concept of encyclopedia is more or less equated by means of the coordinating conjunction "ou" to a *Dictionnaire raisonné*, "reasoned dictionary", introducing the idea that encyclopedias are dictionaries with some additional structure and a philosophical dimension.

Back to the "Encyclopédie" article one can read that a dictionary remaining strictly at the language level, a vocabulary, can be seen as the empty frame required for an encyclopedic dictionary which will fill it with additional depth. Given how d'Alembert insists on the importance of brevity for a clear definition in the "Dictionnaire de Langues" entry, it is clear that the *encyclopédistes* did not consider encyclopedias superior to dictionaries but really as a new subgenre departing from them in terms of purpose.

3 The *dictionaries* TEI module

One of the main motivations behind project DISCO-LGE was to produce data useful to future scientific projects, which in particular requires it to be *interoperable* and *reusable*. These are

the two last key aspects of the FAIR (https://www.go-fair.org/fair-principles/) principles (*findability*, *accessibility*, *interoperability* and *reusability*) which are important guidelines for efficient, high-quality research. This section starts by describing the existing toolset provided by the XML-TEI guidelines to achieve this goal, before introducing some notations and tools from graph theory which will be used to browse the guidelines in a systematic and thorough way in section 4.

3.1 A good starting point

The *dictionaries* module has been leveraged to encode dictionaries in projects NENUFAR (https://cahier.hypotheses.org/nenufar) and BASNUM (https://anr.fr/Projet-ANR-18-CE38-0003) to encode respectively the *Petit Larousse Illustré* published by Pierre Larousse in 1905 (Bohbot et al., 2018, p. 1), roughly contemporary to *LGE*, and the *Dictionnaire Universel* by Furetière, or rather its second version edited by Henri Basnage de Beauval, an encyclopedic dictionary from the very early 18th century (Williams, 2017, p. 1). These successes suggested it to be a useful tool to encode encyclopedias but a few differences remained between both projects and DISCO-LGE: the text studied by NENUFAR does not have the encyclopedic dimension *LGE* has and BASNUM studies a much older text which had a tremendous influence on the european encyclopedic effort of the 18th century but is not as clearly separated from the dictionaric stem as *LGE* is. For these reasons, the encoding schemes used in these projects could not be reused directly, prompting for a systematic exploration of the XML-TEI schema to devise a new one.

This chapter discusses XML elements and hence needs to name and manipulate them. They will be represented in a monospace font, in the standard XML autoclosing form within angle brackets and with a slash following the element name like `<div/>` for a `div` element (https://tei-c.org/release/doc/tei-p5-doc/en/html/ref-div.html). This notation does not mean to imply that they cannot contain raw text or other XML elements, it merely denotes such an element, without any additional assumption. In the context of a concrete document instance this can refer to the markup with all the subtree that possibly spans from it, but the same notation will be used when considering the abstract element and the rules that govern its use in relation to other elements or its attributes.

3.2 A graph problem

The XML-TEI specification contains 590 elements, which are each documented on the consortium's website in the online reference pages. With an average of almost 80 possible child elements (79.91) within any given element, manually browsing such an massive network can prove quite difficult as the number of combinations sharply increases with each step.

The problem can be advantageously transformed to benefit from the results of graph theory by representing the network of the XML elements as a directed graph which nodes are connected or not depending on the inclusion rules of the guidelines. Classical, well-known traversal techniques such as Dijkstra's algorithm (Dijkstra, 1959) which computes the shortest path between two nodes in a graph and reports when they are not connected can then be applied to compute systematically all the possible ways to nest a given element under another without any risk to forget a route because of human error.

Though a particular caution should be applied on the results provided by this algorithm because there is no guarantee that the shortest path is meaningful in general, it at least provides an efficient way to check whether a given element may or not be nested at all under another

one and gives a lower bound on the length of a meaningful path if it exists. The accuracy of this heuristic decreases as the length of the path increases in the perfect graph representing the intended, meaningful path between two nodes that a human specialist of the TEI framework could build.

The XML-TEI guidelines graph will hence be defined as follows. One node is created for each one of the 590 elements found in the specification. Then, an edge is placed between source node `A` and destination `B` if the schema states that the element represented by `B` can be contained directly by the element represented by `A`. That is, the edges in the graph represent the relation "is an admissible direct parent of" (written infix, as in "A is connected to B" if and only if "A is an admissible direct parent of B"). Please note that the word "element" is here used with the same meaning as in the TEI documentation to refer to the conceptual device characterised by a given tag name such as `p` or `div` and not to a particular instance of them that may occur in a given document. Figure 1, by using this transformation to display only the *dictionaries* module, hints at the overall complexity of the whole specification.

With this definition, moving from one node to another on the graph has an XML-TEI counterpart. Following an edge from `A` to `B` can be understood as preparing an XML structure of an `<A/>` element containing a `<B/>` element like this:

```
<A>
    <B/>
</A>
```

By iterating several times the operation of moving on that graph along one edge, that is, by considering the transitive closure of the relation "be connected by an edge" one defines *inclusion paths*, allowing to explore which elements may be nested (arbitrarily deep) under which other. The nodes visited along the way represent the intermediate XML elements required to construct a valid XML tree according to the TEI schema. Given the top-down semantics of those trees, the length of an inclusion path will be called its *depth*.

The ability for an element to contain itself corresponds directly to loops on the graph (that is an edge from a node to itself) as can be illustrated by the `<entry/>` element on figure 1: an `<entry/>` element may directly contain another one.

The generalisation of this to inclusion paths of any length greater than one is usually called a cycle and it appears natural to refine this and name them *inclusion cycles*. The `<address/>` element provides an example for this configuration: although an `<address/>` element may not directly contain another one, it may contain a `<geogName/>` which, in turn, may contain a new `<address/>` element. From a graph theory perspective, one can say that it admits an inclusion cycle of length two.

Using inclusion paths lets one find for instance that although `<pos/>` may not be directly included within `<entry/>` elements to include information about the part-of-speech of the word that an article defines, the correct way to do so is through a `<form/>` or a `<gramGrp/>` because a thorough traversal reporting all the possible paths will contain `entry-form-pos` and `entry-gramGrp-pos`. It is left to the human encoder to rate the relevance of the path found and to select an appropriate one. A total lack of path proves the impossibility of an inclusion; an abnormally high depth for the shortest path is a serious hint that the inclusion should not be possible and is not meaningful.

Another relevant example of the use of these methods can be given by querying the shortest

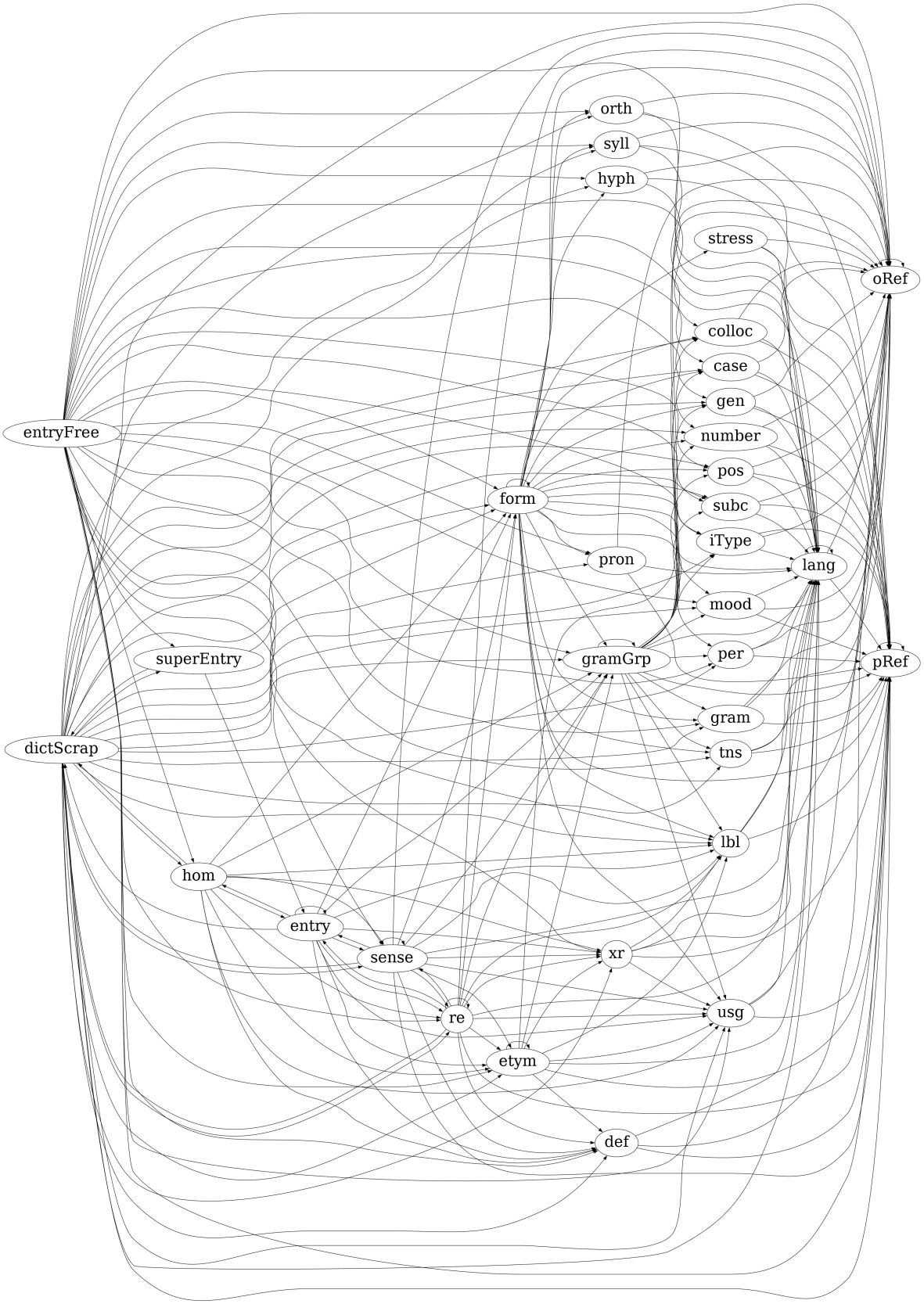


Figure 1: The subgraph of the *dictionaries* module

inclusion path of a <pos/> under the <body/> of the document: it yields an inclusion directly through <entryFree/> (with an inclusion path of length 2), which unlike <entry/> accepts it as a direct child node. Possibly not what is wanted depending on the regularity of the articles being encoded and the occurrence of other grammatical information such as <case/> or <gen/> to justify the use of the <gramGrp/>, but searching exhaustively for paths up to length 3 returns as expected the path through <entry/>, among others. The big picture starts to appear: <pos/> does not need to be nested very deep, it can appear quite near the "surface" of article entries.

3.3 Content of the module

The central element of the *dictionaries* module is the <entry/> element meant to encode one single entry in a dictionary, that is to say a head word associated to its definition. It is the natural way in from the <body/> element to the *dictionaries* module: indeed, although <body/> may also contain <entryFree/> or <superEntry/> elements, the former is a relaxed version of <entry/> while the latter is a device to group several related entries together. Both can contain an <entry/> directly while no obvious inclusion exists the other way around: most (> 96.2%) of the inclusion paths of "reasonable" depth (which will be arbitrarily defined as strictly inferior to 5, that is twice the average shortest depth between any two nodes) either include <figure/> or <castList/>, two very specific elements which should not need to appear in an article in general, showing that the purpose of <entry/> is not to contain an <entryFree/> or <superEntry/>. Hence, not only the semantics conveyed by the documentation but also the structure of the elements graph evidence <entry/> as the natural top-most element for an article. This example demonstrates again how a graph-centred approach can provide insights about the XML-TEI schema.

Once a block for an article is created, it may contain elements useful to represent various of its features. Its written and spoken forms are usually encoded by <form/> elements. Grammatical information like the <case/>, <gen/> or <number/> and <pers/> can be contained within a <gramGrp/>, along with information about the categories it belongs to like <iType/> for its inflection class in languages with a declension system or <pos/> for its part-of-speech. The <etym/> element is made to hold the etymology of an entry. In the case when there are alternative spellings in varieties of the language or if the spelling has changed over time, <usg/> can be used.

All these examples are by no means an exhaustive list; the complete set provides the encoder with a toolbox to describe all the information related to the form the entry is found at and seems general enough to accomodate the structure of any book indexing entries by words.

A common feature shared by dictionaries and encyclopedias is the ability to connect entries together by using a word or short phrase as the link, referring the reader to the related concept. This is known as cross-references and can appear either when the definition of a term is adjacent to another one or to catch alternative spellings where some readers might expect to find the word and redirect them to the form chosen as the reference. In XML-TEI, this is done with the <xr/> element. It usually contains the whole phrase performing the redirection, with an imperative locution like "please see [...]".

The "active" part of the cross-reference, that is the very word within the <xr/> that is considered to be the link or, to make a modern-day HTML metaphor, the region that would be clickable, is represented by a <ref/> element. Though it is not specific to the *dictionaries*

module, it is included in this description of the toolbox because it is particularly useful in the context of dictionaries. This element may have a target attribute which points to the other resource to be accessed by the interested reader.

The remaining part of entries is also usually the largest and represents the content associated to the headword by the entry. In a dictionary, that is its meaning.

The <sense/> element is a valid child for <entry/> and groups together a definition of the term with <def/>, usage examples with <usg/> (another use of this versatile element) and other high-level information such as translations in other languages. Both <def/> and <usg/> elements may appear directly under the <entry/>.

Before concluding this description of the *dictionaries* module from the perspective of someone trying to concretely encode a particular dictionary or encyclopedia, the graph approach is again leveraged to evidence some of its aspects in terms of inclusion structure.

First, it is remarkable that all elements in the *dictionaries* module have a cyclic inclusion path, that is to say, there is an inclusion path from each element of this module to itself. Although having such a cycle is a widespread property in the remainder of XML-TEI elements shared by 73.8% of them (411 out of the 557 elements in the other modules), all 33 elements of the *dictionaries* module having one is far above this average. In addition, the cycles appear to be rather short, with an average length of 2.00 versus 2.50 in the rest of the population. This observation is all the more surprising considering the fact that the *dictionaries* module contains short "leaf" elements like <pos/> which should not obviously need to admit cycles since one rather expects them to contain only one word, like <pos>adj</pos> in the example given in the official documentation. Among those (shortest) cycles, 20 include the <cit/> element made to group quotations with a bibliographic reference to their source which should clearly be unnecessary to encode an article in the general case.

Secondly, although examples of connections from this module to the rest of the XML-TEI have been evidenced in this section, especially to the *core* module (to which belongs for example the <ref/> element), the *dictionaries* module appears somewhat isolated from important structural elements like <head/> or . Indeed, computing all the paths of length shorter or equal to 5 from either <entry/> or <sense/> elements to the latter by a systematic traversal of the graph yields exclusively paths (respectively 8 943 and 38 649 of them excluding loops) containing either a <floatingText/> or an <app/> element. The first one, as its name aptly suggests, is used to encode text that does not quite fit the regular flow of the document, as for example in the context of an embedded narrative. Both examples displayed in the online documentation feature a <body/> as direct child of <floatingText/>, neatly separating its content as independent. The purpose of the second one, although its name — short for apparatus — is less clear, is to wrap together several versions of the same excerpts, for instance when there are several possible readings of an unclear group of words in a manuscript, or when the encoder is trying to compile a single version of a piece of work from several sources which disagree over some passage. In both cases, it appears obvious that it is not something that is expected to occur naturally in the course of an article in general.

Thus, despite a rather dense internal connectivity, the *dictionaries* module fails to provide encoders with a device to represent recursively nesting structures like .

4 A new standard ?

Studying the content of *LGE* and considering several articles in particular, one can identify structures which are specific to encyclopedias and not compatible with the *dictionaries* module presented in the previous section. It follows that this module is not able to encode arbitrary encyclopedic content and hence a new fully TEI-compliant encoding scheme is proposed. The rest of the section is concerned with the needs of automated encoding processes and compares the proposal with other strategies to overcome the issues previously identified with the dedicated module for dictionaries.

4.1 Idiosynchrasies of encyclopedias

Browsing through the pages of an encyclopedia reveals a certain number of noticeable differences. A comprehensive list would be difficult to draw because of the great variety in terms of editorial choices but the most obvious can be discussed.

The first immediately visible feature that sets encyclopedias apart from dictionaries and can be found in the *EDdA* as well as in *LGE* is the presence of subject indicators at the beginning of articles right after the headword which organise them into a domain classification system. Those generally cover a broad range of subjects from scientific disciplines to litterature, and extending to political subjects and law.

These indicators have no element in the *dictionaries* module explicitely designed to encode them. As section 3 illustrates, the elements set is geared towards the words themselves instead of the concept they represent. The tool closest to what is needed can be found in the `<usg/>` element used with a specific `type` attribute set to `dom` for "domain". Indeed several examples from the documentation encode subject indicators very similar to the ones found in encyclopedias within this element, but the match is not perfect either: all appear within one of multiple senses, as if to clarify each context in which the word can be used, as expected from the element's name, "usage". In encyclopedias, if the domain indicator does in certain cases help to distinguish between several entries sharing the same headword, the concept itself has evolved beyond this mere distinction. Looking back at the *EDdA*, the adjective *raisonné* in the rest of the title directly introduces a notion of structure that links back to the "Systême figuré des connoissances humaines" (Blanchard & Olsen, 2002, p. 1) which schematic structure is shown in Figure 2. The authors have devised a branching system to classify all knowledge, and the occurrence at the beginning of articles, more than a tool to clear up possible ambiguities also points the reader to the correct place in this mind map.

The situation regarding subject indicators is hardly better outside of the module. The `<domain/>` element despite its name belongs exclusively in the header of a document and focuses on the social context of the text, not on the knowledge area it covers. The `<interp/>` despite its name is not so much about labeling something as an interpretation to give to a context (which subject indicators could be if you consider that, placed at the beginning, they are used to direct the mind frame of the readers towards a particular subject). However, the documentation clearly demonstrates it as a tool for annotators of a document, which text content is not part of the original document but some additional result of an analysis performed in the context of the encoding, used only throughout references in XML attributes.

This point, although not the most concerning, still remains the hardest to address but all things considered the `<usg/>` element stands out as the most relevant.

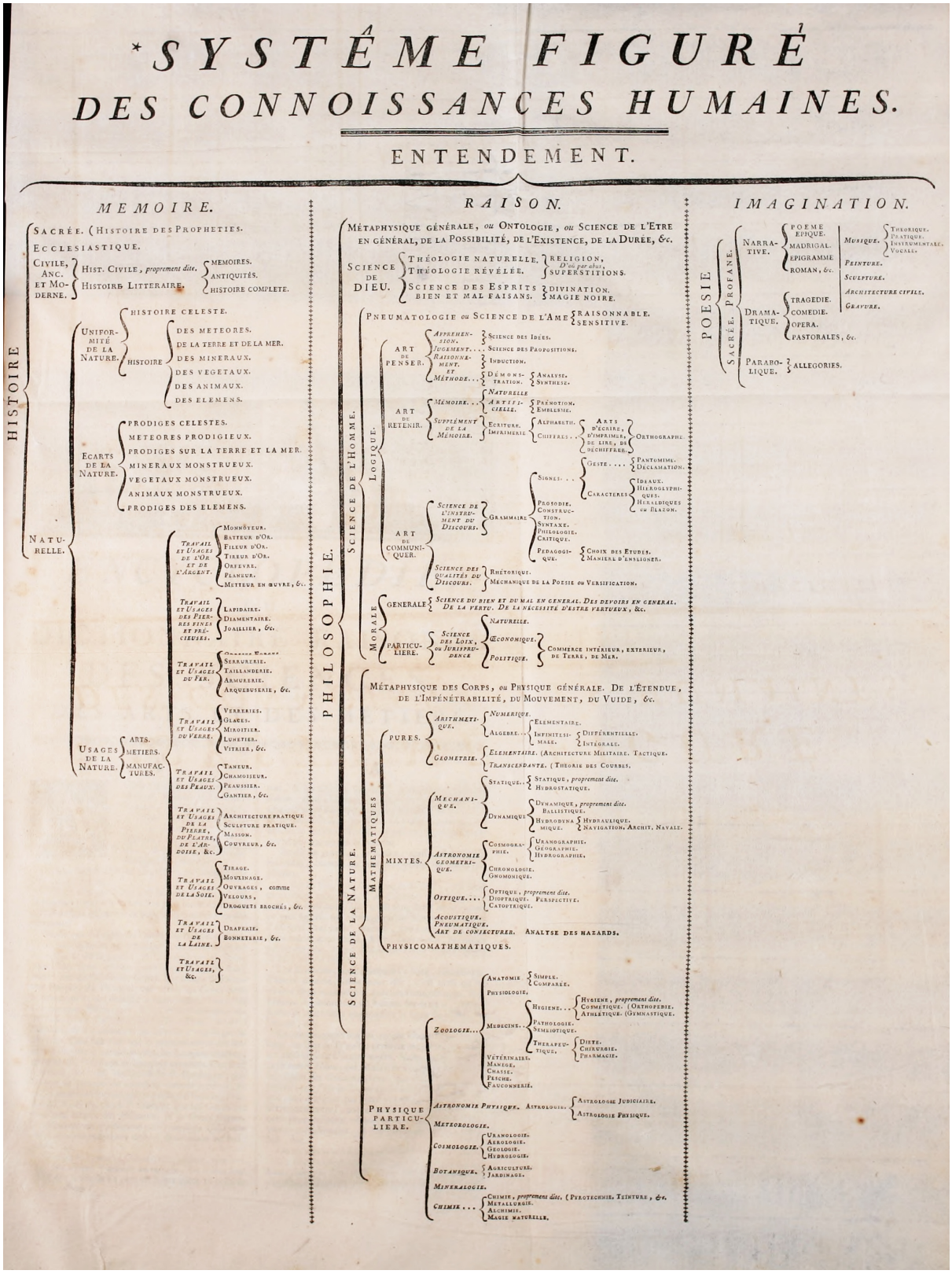


Figure 2: "Système figuré des connoissances humaines", the taxonomy at the heart of the Encyclopédie (Wikimedia Commons)

Notwithstanding the correct way to represent domains of knowledge, their extent itself raises concerns regarding the *dictionaries* module. Indeed, among the vast collection of domains covered in encyclopedias in general and in *LGE* in particular are historical articles and biographies. If the notion of meaning can appear at least ill-fitting for a text describing a series of historical events, one may still argue that it groups them into a concept and associates it to the name of the event. But when it comes to relating the life of a person, describing their relation to events and other persons comes out even further from the notion of meaning. Entries such as the one about SANJO Sanetomi (see Figure 3) do not constitute a *definition*.

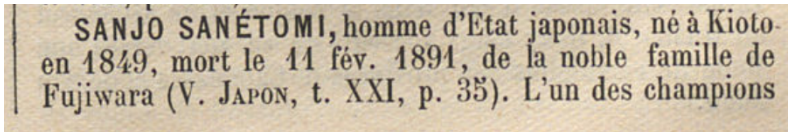
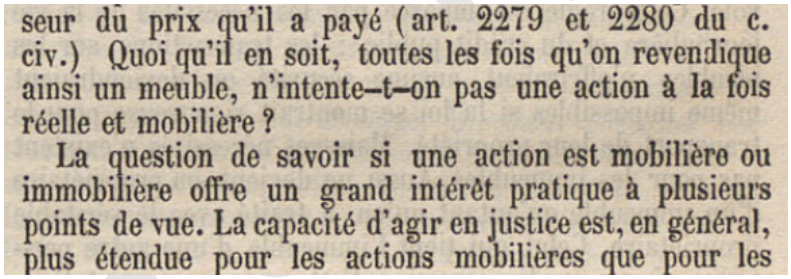

SANJO SANÉTOMI, homme d'Etat japonais, né à Kioto en 1849, mort le 11 fév. 1891, de la noble famille de Fujiwara (V. JAPON, t. XXI, p. 35). L'un des champions

Figure 3: Beginning of the article relating the life of SANJO Sanetomi, in La Grande Encyclopédie, tome 29 (BnF - Gallica)

Moreover, encyclopedias, because of all that they have inherited from the philosophical Enlightenment, are not only spaces designed to assert, they also intrinsically include an interrogative component. Some articles lay down the basis required to understand the complexity of an issue and invite the reader to consider it without providing a definitive answer, going as far as to explicitly use question marks as in the article "Action" displayed in Figure 4.


seur du prix qu'il a payé (art. 2279 et 2280 du c. civ.) Quoi qu'il en soit, toutes les fois qu'on revendique ainsi un meuble, n'intente-t-on pas une action à la fois réelle et mobilière ?

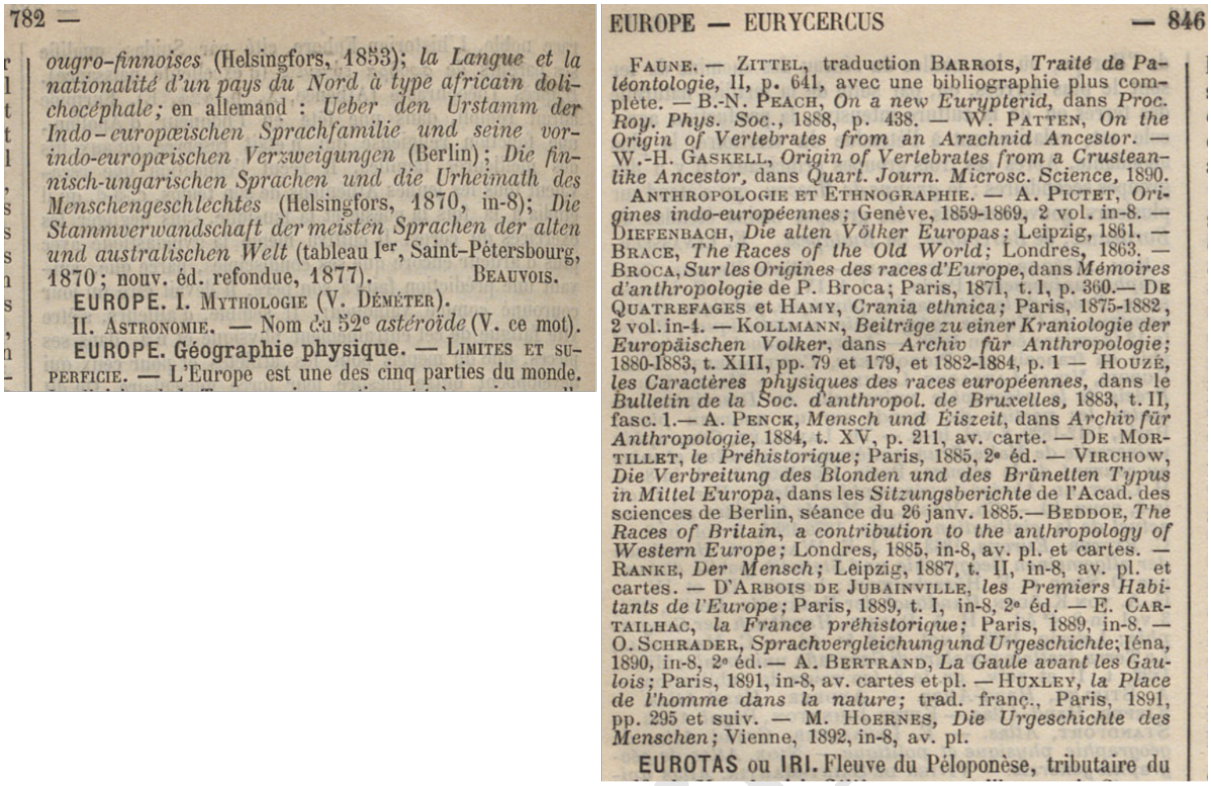
La question de savoir si une action est mobilière ou immobilière offre un grand intérêt pratique à plusieurs points de vue. La capacité d'agir en justice est, en général, plus étendue pour les actions mobilières que pour les

Figure 4: Excerpt from article "Action", in La Grande Encyclopédie, tome 1 (BnF - Gallica)

In this extract, the author devises a hypothetical situation to illustrate how difficult it is to draw the line between two supposedly mutually exclusive subcategories of legal actions. The whole point of the passage is to convey the idea that the term eludes definition, wrapping it in a `<sense/>`, or worse, a `<def/>` element would be an utter misnomer.

As a result, the use of `<sense/>` and `<def/>` is not appropriate for encyclopedic content in general.

The final difficulty can be considered a partial consequence of the previous one on the structure of articles. The difficulty to define complex concepts is the very reason why authors approach their subjects from various angles, circumnavigating it as a best approximation. This strategy favours long, structured developments with sections and subsections covering the multiple aspects of the topic: from a historical, political, scientific point of view… The longest articles, such as article "Europe" shown in Figure 5, can thus span several dozens of pages. They can contain substructures with titles on at least three levels (for instance, a `a)` under a `1)` under a `I.`), each of which are in turn generally developed over several paragraphs.



782 —

ougro-finnoises (Helsingfors, 1853); *la Langue et la nationalité d'un pays du Nord à type africain dolichocéphale;* en allemand : *Ueber den Urstamm der Indo-europæischen Sprachfamilie und seine vorindo-europæischen Verzweigungen* (Berlin); *Die finnisch-ungarischen Sprachen und die Urheimath des Menschengeschlechtes* (Helsingfors, 1870, in-8); *Die Stammverwandschaft der meisten Sprachen der alten und australischen Welt* (tableau Ier, Saint-Pétersbourg, 1870; nouv. éd. refondue, 1877). BEAUVOIS.

EUROPE. I. MYTHOLOGIE (V. DÉMÉTER).

II. ASTRONOMIE. — Nom du 52e *astéroïde* (V. ce mot).

EUROPE. Géographie physique. — LIMITES ET SUPERFICIE. — L'Europe est une des cinq parties du monde.

EUROPE — EURYCERCUS — 846

FAUNE. — ZITTEL, traduction BARROIS, *Traité de Paléontologie*, II, p. 641, avec une bibliographie plus complète. — B.-N. PEACH, *On a new Eurypterid*, dans *Proc. Roy. Phys. Soc.*, 1888, p. 438. — W. PATTEN, *On the Origin of Vertebrates from an Arachnid Ancestor.* — W.-H. GASKELL, *Origin of Vertebrates from a Crustean-like Ancestor*, dans *Quart. Journ. Microsc. Science*, 1890.

ANTHROPOLOGIE ET ETHNOGRAPHIE. — A. PICTET, *Origines indo-européennes;* Genève, 1859-1869, 2 vol. in-8. — DIEFENBACH, *Die alten Völker Europas;* Leipzig, 1861. — BRACE, *The Races of the Old World;* Londres, 1863. — BROCA, *Sur les Origines des races d'Europe*, dans *Mémoires d'anthropologie* de P. Broca; Paris, 1871, t. I, p. 360.— DE QUATREFAGES et HAMY, *Crania ethnica;* Paris, 1875-1882, 2 vol. in-4. — KOLLMANN, *Beiträge zu einer Kraniologie der Europäischen Volker*, dans *Archiv für Anthropologie;* 1880-1883, t. XIII, pp. 79 et 179, et 1882-1884, p. 1 — HOUZÉ, *les Caractères physiques des races européennes*, dans le *Bulletin de la Soc. d'anthropol. de Bruxelles*, 1883, t. II, fasc. 1.— A. PENCK, *Mensch und Eiszeit*, dans *Archiv für Anthropologie*, 1884, t. XV, p. 211, av. carte. — DE MORTILLET, *le Préhistorique;* Paris, 1885, 2e éd. — VIRCHOW, *Die Verbreitung des Blonden und des Brünetten Typus in Mittel Europa*, dans les *Sitzungsberichte* de l'Acad. des sciences de Berlin, séance du 26 janv. 1885.— BEDDOE, *The Races of Britain, a contribution to the anthropology of Western Europe;* Londres, 1885, in-8, av. pl. et cartes. — RANKE, *Der Mensch;* Leipzig, 1887, t. II, in-8, av. pl. et cartes. — D'ARBOIS DE JUBAINVILLE, *les Premiers Habitants de l'Europe;* Paris, 1889, t. I, in-8, 2e éd. — E. CARTAILHAC, *la France préhistorique;* Paris, 1889, in-8. — O. SCHRADER, *Sprachvergleichung und Urgeschichte;* Iéna, 1890, in-8, 2e éd. — A. BERTRAND, *La Gaule avant les Gaulois;* Paris, 1891, in-8, av. cartes et pl. — HUXLEY, *la Place de l'homme dans la nature;* trad. franç., Paris, 1891, pp. 295 et suiv. — M. HOERNES, *Die Urgeschichte des Menschen;* Vienne, 1892, in-8, av. pl.

EUROTAS ou IRI. Fleuve du Péloponèse, tributaire du

Figure 5: La Grande Encyclopédie, tome 16, article "Europe", spanning from p.782 to p.846, that is 64 pages, and ending after a bibliography longer than one column of text (BnF - Gallica)

The nested structure that have just been evidenced demands of course a nesting structure to accomodate it. More precisely, it guides the search of XML elements by adding several constraints: what is required is a pair of elements. The first one representing a (sub)section must be able to include both itself and the second one, which does not have any special constraint except the one to have a semantics compatible with the purpose of being used to represent section titles. In addition, the first element must be able to contain several <p/> elements, <p/> being the reference element to encode paragraphs according to the XML-TEI documentation.

The *dictionaries* module has been shown to be equiped with a questionable but possible element for subject domains. However, it does not include any element for section titles. In the rest of the TEI specification, the elements <head/> and <title/> — the latter with the possibility to set its type attribute to sub — stand out as the best candidates for the semantics condition on the second element.

Filtering the content of the module to keep only the elements which can at the same time contain themselves, be included under <entry/> and include a <p/> and either the <head/> or <title/> elements yields absolutely no candidates. It is remarkable that even replacing the <entry/> element for the root of each article with an <entryFree/>, an element supposed to relax the constraints to accomodate more unusual structures in dictionaries does not bring any improvement.

The lack of results from these simple queries forces one to adopt a less restrictive approach to find an encoding. The occurrence of an intermediate element could for instance be needed between the element wrapping the whole article and the recursing one used to encode each section. This "section" element could also need a companion element to be able to include

itself, or, to formalise it in terms of graph theory, the condition that this element admits a loop could be relaxed to consider instead cycles of a given (small, this still needs to represent a fairly direct inclusion) length to be enough. Simultaneously the maximum depth of the inclusion paths between <entry/>, the pair of elements and the <p/> element will be increased to yield more results.

By setting this depth to 2, that is, by accepting one intermediate element to occur in the middle of each one of the inclusion paths that define the structure required to encode encyclopedic discourse, 21 elements can be found, none of which stands out as an obvious good solution: all paths to include the <p/> element from any *dictionaries* element either contains a <figure/> (already discussed in section 3 when practising the graph approach to search for inclusions between <entry/> and <entryFree/> and dismissed as not useful in general), a <stage/> (reserved to stage direction in dramatic works) or a <state/> (used to describe a temporary quality in a person or place), again not even close to what is wanted. The paths to either <head/> or <title/> are similarly disappointing. Again, changing <entry/> for <entryFree/> returns the exact same candidates. If that is not a definite proof that none of these elements could meet the investigated criteria, it is a fact than no element in this module stands out as the obvious good solution and a serious hint to keep looking somewhere else.

Therefore, the search is extended again to include elements outside the *dictionaries* module which could be used to encode the sections and subsections, under the same constraint as before to try and find a composite solution that would remain under the <entry/> element even if resorting to subcomponents outside of the dedicated module. Only three elements are returned: <figure/>, <metamark/> and <note/>.

The first one as has been repeatedly underlined is meant for graphic information and is not suitable for text content in general.

The purpose of <metamark/> is to transcribe the edition marks than may appear on a particular primary source in order to alter the normal flow of the text and suggest an alternative reading (deletion, insertion, reordering, this is about a human editing the text from a given physical copy of it), but it is unfortunately of no use to encode a section of an article.

The first element that might at least seem acceptable is the last one, <note/>. It is meant to contain text, is about explaining something and seems general enough (not specific to a given genre, or to the occurrence of a particular object on the page). Unfortunately, its semantics still seems a bit off compared to what is required. The documentation describes it as an "additional comment" which appears "out of the main textual stream" whereas the long developments in articles are the very matter of the text of encyclopedias, not mere remarks in the margins or at the foot of pages.

4.2 Encoding within the *core* module

The remarks made in section 3 explain why the *dictionary* module is unable to represent encyclopedias, where the notion of "meaning" is less central than in dictionaries and where discourse with nested structures of arbitrary depth can occur. Even composite encodings using elements outside of the *dictionaries* module under an <entry/> element do not meet the requirements of the project. Since the *core* module obviously accomodates these structures by means of the , <head/> and <p/> elements which have the additional advantage of carrying less semantical payload than <sense/> or <def/>, these elements will be used to devise an encoding scheme which can be recommended for other projects aiming at representing

encyclopedias.

To remain consistent with the way the *dictionaries* module was studied only what happens at level of each individual article will be considered, that is right under the `<body/>` element representing a whole volume. Everything related to its metadata happens as expected in the file's `<teiHeader/>` which is well-enough equiped to handle them. In order to present the scheme throughout the following section a reference article, "Cathète" from tome 9 — reproduced in Figure 6 — will be encoded step by step.

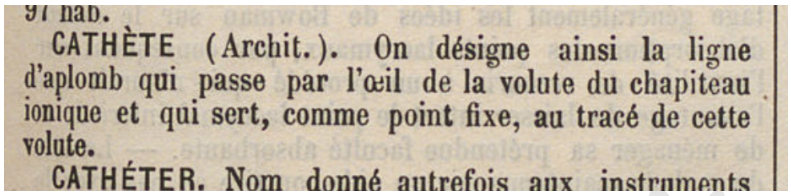

97 hab.
CATHÈTE (Archit.). On désigne ainsi la ligne d'aplomb qui passe par l'œil de la volute du chapiteau ionique et qui sert, comme point fixe, au tracé de cette volute.
CATHÉTER. Nom donné autrefois aux instruments

Figure 6: La Grande Encyclopédie, tome 9, article "Cathète" (BnF - Gallica)

Remaining within the *core* module for the structure, almost all useful elements are available and practically no additional documentation is needed beyond the official TEI guidelines. Each article is represented by a `<div/>`. Setting an `xml:id` attribute on it with a unique value will ease identify, browse and retrieve the articles from the encoded corpus. An auto-increasing serial would of course provide an appropriate value for such a unique attribute but has some drawbacks: as long as the articles segmentation isn't fixed (which could happen if choices regarding entries and sub-entries were to change along a project or if, as is the case of DISCO-LGE, the automatic segmentation went through successive improvement steps), the identifiers of articles would massively change from one version to the other, even articles segmented correctly. Given the iterative nature of many studies in digital humanities, this would make it harder to use results found early in a project. For this reason, the values used for `xml:id` in project DISCO-LGE depend only on the local quality of the segmentation and remain globally stable. They are computed as the head word of the entries normalised to lowercase, stripping spaces and replacing all non-alphanumerical characters by a dash (`'-'`) to avoid issues with the XML encoding, and suffixed by a serial to distinguish between the few entries sharing the same head. Thus, if an oversegmentation or a subsegmentation are fixed (meaning respectively that two "articles" get fusioned or that one "article" actually contained several which get split as such) only articles with the same headword are impacted. Figure 7 illustrates this choice for the container element on the article "Cathète" displayed on figure 6.

```
<div xml:id="cathète-0"></div>
```

Figure 7: The container `div` element for article "Cathète"

Inside this element should be a `<head/>` enclosing the headword of the article. The usual `<hi/>` elements are available within `<head/>` if the headword is highlighted by any special typographic means such as bold, small capitals, etc. The one disappointment of the encoding scheme being defined in this chapter is the lack of support for a proper way to encode subject indicators.

The best candidate found so far was `<usg/>` from the *dictionaries* module but it is not available directly under a `<head/>` element. All inclusion paths from the latter to the former of length less than or equal to 3 contain irrelevant elements (`<cit/>`, `<figure/>`, `<castList/>` and

<nym/>) so it must be discarded. The next best elements appear to be <term/> (not very accurate) and <rs/> (“referring string”, quite a general semantics but a possible match — subject indicators refer to a given domain of knowledge — although all the examples in the documentation refer to concrete persons, places or object, not to the abstract objects that mathematics or poetry are).

For this reason, no particular encoding of the subject indicator is recommended and it is left open to each particular context: they are often abbreviated so an <abbr/> may apply, in *LGE*, biographies are not labeled by a knowledge domain but usually include the first name of the person when it is known so in that case an element like <persName/> is still appropriate. This choice applied to the same article “Cathète” produces Figure 8.

```
<div xml:id="cathète-0">
  <head>CATHÈTE (<abbr>Archit.</abbr>)</head>
</div>
```

Figure 8: Encoding the head word of article “Cathète”

Each different meaning could then be wrapped in a separate with the type attribute set to sense to refer to the <sense/> element that would have been used within the *dictionaries* module. The s should be numbered according to the order they appear in with the n attribute starting from 0 as shown in Figure 9.

```
<div xml:id="cathète-0">
  <head>CATHÈTE (<abbr>Archit.</abbr>)</head>
  <div type="sense" n="0"></div>
</div>
```

Figure 9: The empty structure for the only meaning of the word “Cathète”

In addition, each line within the article must start with a <lb/> to mark its beginning including before the <head/> element as demonstrated by Figure 10, which, although a surprising setup, underlines the fact that in the dense layout of encyclopedias, the carriage return separating two articles is meaningful. Stating each new line explicitly keeps enough information to reconstruct a faithful facsimile but it also has the advantage of highlighting the fact than even though the definition is cut from the headword by being in a separate XML element, they still occur on the same line, which is a typographic choice usually made both in encyclopedias and dictionaries where space is at a premium.

To complete the structure, the various sections and subsections occurring within the article body may be nested as usual with and sub- s, filled with <p/> for paragraphs which can each be titled with <head/> elements local to each .

Some articles such as “Boumerang” have figures with captions, as illustrated by Figure 11, which should be encoded the standard way by <figure/> and <figDesc/> as in Figure 12.

Another issue arising from giving up on <entry/> is the unavailability of the <xr/> element, not allowed under any of the *core* elements used but which is useful to represent cross-references occurring in encyclopedias as well as in dictionaries, for example in article “Gelocus” (see Figure 13). It is prefered to use the <ref/> element instead which is available in the context of a <p/>. Its target attribute should be set to the xml:id of the article it points to,

```
<div xml:id="cathète-0">
  <lb/><head>CATHÈTE (<abbr>Archit.</abbr>).</head>
  <div type="sense" n="0">
    <p>
      On désigne ainsi la ligne
      <lb/>d'aplomb qui passe par l'œil de la volute du chapiteau
      <lb/>ionique et qui sert, comme point fixe, au tracé de cette
      <lb/>volute.
    </p>
  </div>
</div>
```

Figure 10: A complete encoding of article “Cathète”

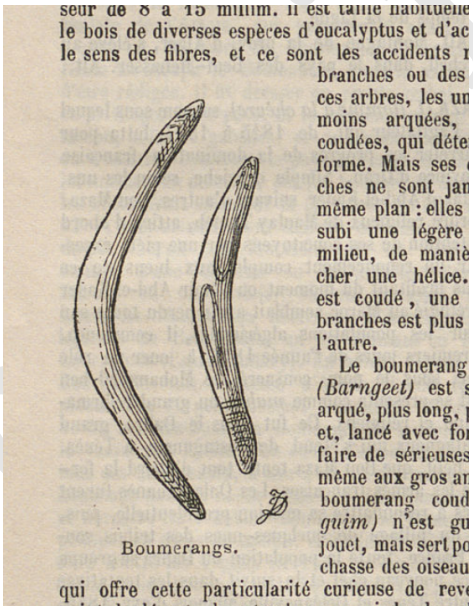


Figure 11: La Grande Encyclopédie, tome 7, article “Boumerang” (BnF - Gallica)

```
<figure>
  <graphic url="t7p725_1.png"/>
  <figDesc>Boumerangs.</figDesc>
</figure>
```

Figure 12: Encoding the figure in article “Boumerang” and its captions

prefixed with a `'#'` as shown in Figure 14. Another solution would have been to introduce a `<dictScrap/>` element for the sole purpose of placing an `<xr/>` but this would add unwanted verbosity to the encoding and implicitly suggest that the previous context was not the one of a dictionary which is rather problematic.

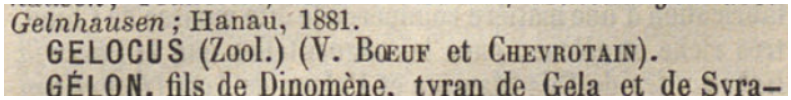

Gelnhausen ; Hanau, 1881.
GELOCUS (Zool.) (V. Bœuf et Chevrotain).
GÉLON, fils de Dinomène, tyran de Gela et de Syra-

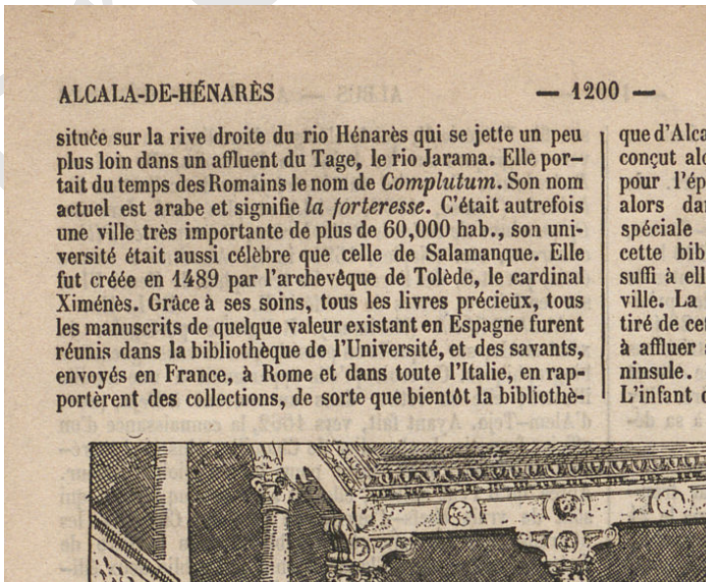
Figure 13: La Grande Encyclopédie, tome 18, article “Gelocus” (BnF - Gallica)

```
<p>
  (V. <ref target="#boeuf-0">Boeuf</ref> et
  <ref target="#chevrotain-0">Chevrotain</ref>).
</p>
```

Figure 14: Encoding the cross-references in article “Gelocus”

A typical page of an encyclopedia also features peritext elements, giving information to the reader about the current page number along with the headwords of the first and last articles appearing on the page. Those can be encoded by `<fw/>` elements (“forme work”) which `place` and `type` attributes should be set to position them on the page and identify their function if it has been recognised (those short elements on the border of pages are the ones typically prone to suffer damages or be misread by the OCR).

Finally there are other TEI elements useful to represent “events” in the flow of the text, like the beginning of a new column of text or of a new page. Figure 15 shows the top left of the last page of the first tome of *LGE* which features peritext elements while marking the beginning of a new page. The usual appropriate elements (`<pb/>` for page beginning, `<cb/>` for column beginning) may and should be used with this encoding scheme as demonstrated by Figure 16.


ALCALA-DE-HÉNARÈS — 1200 —

située sur la rive droite du rio Hénarès qui se jette un peu plus loin dans un affluent du Tage, le rio Jarama. Elle portait du temps des Romains le nom de *Complutum*. Son nom actuel est arabe et signifie *la forteresse*. C'était autrefois une ville très importante de plus de 60,000 hab., son université était aussi célèbre que celle de Salamanque. Elle fut créée en 1489 par l'archevêque de Tolède, le cardinal Ximénès. Grâce à ses soins, tous les livres précieux, tous les manuscrits de quelque valeur existant en Espagne furent réunis dans la bibliothèque de l'Université, et des savants, envoyés en France, à Rome et dans toute l'Italie, en rapportèrent des collections, de sorte que bientôt la bibliothè-

que d'Alca
conçut al
pour l'ép
alors da
spéciale
cette bib
suffi à ell
ville. La
tiré de ce
à affluer
ninsule.
L'infant

Figure 15: La Grande Encyclopédie, tome 1, article “Alcala-de-Hénarès” (BnF - Gallica)

The reference implementation for this encoding scheme is the program soprano (https://gitlab.huma-num.fr/disco-lge/soprano) developed within the scope of project DISCO-LGE

```
<lb/>nerie générale de Madrid et du diocèse de Tolède. Elle est
<pb n="1247" />
<fw type="header">ALCALA-DE-HÉNARÈS</fw>
<fw type="pageNum">- 1200 -</fw>
<lb/>située sur la rive droite du rio Hénarès qui se jette un peu
<lb/>plus loin dans un affluent du Tage, le rio Jarama. Elle por¬
<lb/>tait du temps des Romains le nom de Complutum. Son nom
```

Figure 16: Encoding the beginning of a page in article “Alcala-de-Hénarès”

to automatically identify individual articles in the flow of raw text from the columns and to encode them into XML-TEI files. Though this software has already been used to produce the first TEI version of *LGE*, it does not follow perfectly yet the specification described in this chapter. Figure 17 shows the encoded version of article “Cathète” it currently produces:

```
<div xml:id="cathète-0">
  <lb/><head>CATHÈTE</head> (Archit.). On désigne ainsi la ligne
  <lb/>d’aplomb qui passe par l’œil de la volute du chapiteau
  <lb/>ionique et qui sert, comme point fixe, au tracé de cette
  <lb/>volute.
</div>
```

Figure 17: The current encoding of article “Cathète” produced by `soprano`

The headword detection system is not able to capture the subject indicators yet so it appears outside of the `<head/>` element. No work is performed either to expand abbreviations and encode them as such, or to distinguish between domain and people names.

Likewise, since the detection of titles at the beginning of each section is not complete, no structure analysis can be performed at the moment on the textual development inside the article and it is left unstructured, directly under the entry’s `<div/>` element instead of under a set of nested `<div/>` elements. The paragraphs are not yet identified and for this reason not encoded.

However, the figures and their captions are already handled correctly when they occur. The encoder also keeps track of the current lines, pages, and columns to insert the corresponding empty elements (`<lb/>`, `<pb/>` or `<cb/>`) and number pages according to the order of the physical pages in the book, as compared to the “high-level” pages numbers inserted by the editors, which start with an offset because the first, blank or almost empty pages at the beginning of each book do not have a number and which sometimes have gaps when a full-page geographical map is inserted since those are printed separately on a different folio which remains outside of the textual numbering system. The place at which these layout-related elements occur is determined by the place where the OCR software detected them and by the reordering performed by `soprano` when inferring the reading order before segmenting the articles.

4.3 The constraints of automated processing

Encyclopedias are particularly long books, spanning numerous tomes and containing several tenths of thousands of articles. The *EDdA* comprises over 74k articles and *LGE* certainly more

than 100k (the latest version produced by `soprano` created 160k articles, but their segmentation is still not perfect).

XML-TEI is a very broad tool useful for very different applications. Some elements like `<unclear/>` or `<factuality/>` can encode subtle semantics information (for the second one, adjacent to a notion as elusive as truth) which requires a very deep understanding of a text in its entirety and about which even some human experts may disagree.

For these reasons, a central concern in the design of an encoding scheme was to remain within the boundaries of information that can be described objectively and extracted automatically by an algorithm. Most of the tags presented in section 4.2 contain information about the positions of the elements or their relation to one another. Those with an additional semantics implication like `<head/>` can be inferred simply from their position and the frequent use of a special typography like bold or upper-case characters.

The case of cross-references is particular and may appear as a counter-example to the main principle on which this scheme is based. Actually, the process of linking from an article to another one is so frequent (in dictionaries as well as in encyclopedias) that it generally escapes the scope of regular discourse to take a special and often fixed form, inside parenthesis and after a special token which invites the reader to perform the redirection. In *LGE*, virtually all the redirections appear within parenthesis (at least no counter-example has been found within the scope of the project), and start with the verb "voir" abbreviated as a single, capital "V." as illustrated in the article "Gelocus" (see again Figure 13).

Although this has not been implemented yet either, being able to detect and exploit those patterns to correctly encode cross-references does not pose any fundamental theoretical problem and should be achievable. Getting the `target` attributes right is certainly more difficult to achieve and may require processing the articles in several steps, to first discover all the existing headwords — and hence article IDs — before trying to match the words following "V." with them. Since the automated encoder implemented in the project handles tomes separately and since references may cross the boundaries of tomes, it cannot wait for the target of a cross-reference to be discovered by keeping the articles in memory before outputting them.

This is in line with the last important aspect of the encoder. If many lexicographers may deem this encoding too shallow, it has the advantage of not requiring to keep too complex datastructures in memory for a long time. The algorithm implementing it in `soprano` outputs elements as soon as it can. This is immediate for simple elements such as `<pb/>` or `<fw/>`; for articles, it pushes lines onto a stack and flushes it each time it encounters the beginning of the following article. This allows the amount of memory required to remain reasonable and even lets them be parallelised on most modern machines. Thus, even taking over three minutes per tome, the total processing time can be lowered to around forty minutes on a machine with 16Go of RAM for the whole of *LGE* instead of over one hour and a half.

4.4 Comparison to other approaches

The previous section about the structure of the *dictionaries* module and the features found in encyclopedias follows reflects the issues which have arised along the course of the project while trying to encode first manually and then by automatic means the articles of its corpus. This back and forth between trying to find patterns in the graph which reflects the patterns found in the text and questioning the relevance of the results explains the choice advocated in this chapter but also the alternatives considered.

Several elements exhibited some interesting properties, having for instance some interesting inclusion path corresponding to the structure needed to represent the nested structure of articles. This is the case for instance of the `<sense/>` and `<note/>` elements. It is very tempting to bend their documented semantics or to consider that their inclusion properties is part of what defines them, and hence justifies their ways in creative ways not directly recommended by the TEI specifications.

This is the approach followed by project BASNUM (see section 3.1). In the articles encoded for this project, `<note/>` elements are nested and used to structure the encyclopedic developments that occur in the articles.

For the sake of the FAIR principles, this was not the path chosen by project DISCO-LGE, in order to avoid the emergence of a custom usage differing from the one documented in the official guidelines.

The other major reason behind the choice that was ultimately made was the existing TEI rules governing element inclusions which prompted the search for different combinations. Another valid approach would have consisted in changing the structure of the inclusion graph itself, that is to say modify the rules. If `<entry/>` is the perfect element to encode article themselves, all that is really missing is the ability to accomodate nested structures with the `<div/>` element. This would also have the advantage of recovering the `<usg/>` and `<xr/>` elements which appear useful and which are lost as part of the tradeoff to get nested sections. Generating customised TEI schemas is made really easy with tools like ROMA (https://roma.tei-c.org/), which was used to preview this change and suggest it to the TEI community.

Despite it not getting a wide adhesion, some suggested it could be used locally within the scope of project DISCO-LGE. However it was preferred not to do so, partially for the same reasons of interoperability as the previous scenario, but also for reasons of sturdiness in front of future evolutions. Making sure the alternative schema would remain useful entails to maintain it, regenerating it should the schema format evolve, with the risk that the tools to edit it might stop being maintained or that some conflicts between this change and future modifications of the official guidelines might arise.

5 Conclusion

Though they are very close genres and share a common history, key differences between dictionaries and encyclopedias have been evidenced. Not only do entries tend to be longer in encyclopedias, they often have a deeper structure too. Their purpose also departs from the purpose of dictionaries from their inception, and, as anticipated by their pioneers, results in a different form of discourse.

The structure of the XML-TEI *dictionaries* module reflects the assumptions made by the eponymous genre and does not appear to be flexible enough to accomodate encyclopedias, despite the colossal effort which has gone into making it expressive enough for the wide variety of existng dictionaries. Forcing its use to some encyclopedic articles would breach the semantics of some elements or require the encoder to break the rules of the consortium's schema which would result in a less reusable encoding in opposition to the FAIR principles.

An encoding scheme which fully complies with XML-TEI while being able to represent the content of encyclopedias in all their complexity has been provided and demonstrated on con-

crete examples. The tool `soprano`, partially implementing this set of conventions demonstrates their practical usefulness.

Acknowledgement

The author would like to thank the CollEx-Persée group for supporting the DISCO-LGE project and is also grateful to the ASLAN project (ANR-10-LABX-0081) of the Université de Lyon, for its financial support within the French program "Investments for the Future" operated by the National Research Agency (ANR).

Bibliography

Bański, P., Bowers, J., & Erjavec, T. (2017, September). *TEI-Lex0 guidelines for the encoding of dictionary information on written and spoken forms*. https://inria.hal.science/hal-01757108

Blanchard, G., & Olsen, M. (2002). Le système de renvois dans l'Encyclopédie : Une cartographie des structures de connaissances au XVIIIe siècle. *Recherches Sur Diderot Et Sur l'Encyclopédie*, *31-32*, 45. https://doi.org/10.4000/rde.122

Bohbot, H., Frontini, F., Luxardo, G., Khemakhem, M., & Romary, L. (2018). Presenting the Nénufar Project: a Diachronic Digital Edition of the Petit Larousse Illustré. *GLOBALEX 2018 - Globalex workshop at LREC2018*, 1–6. https://hal.archives-ouvertes.fr/hal-01728328

Bowers, J., & Romary, L. (2018). Bridging the Gaps between Digital Humanities, Lexicography, and Linguistics: A TEI Dictionary for the Documentation of Mixtepec-Mixtec. *Dictionaries: Journal of the Dictionary Society of North America*, *39*(2), 79. https://inria.hal.science/hal-01968871

d'Alembert. (n.d.). DICTIONNAIRE. In R. Morrissey & G. Roe (Eds.), *Encyclopédie, ou dictionnaire raisonné des sciences, des arts et des métiers, etc.* (Vol. 4). University of Chicago: ARTFL Encyclopédie Project. Retrieved May 31, 2023, from https://artflsrv04.uchicago.edu/philologic4.7/encyclopedie0922/navigate/4/4871

Dijkstra, Edsger. W. (1959). A note on two problems in connexion with graphs. *Numerische Mathematik*, *1*, 269–271. https://ir.cwi.nl/pub/9256/9256D.pdf

Haiman, J. (1980). Dictionaries and encyclopedias. *Lingua*, *50*(4), 329–357. https://doi.org/10.1016/0024-3841(80)90089-3

Heiden, S. (2010). The TXM Platform : Building Open-Source Textual Analysis Software Compatible with the TEI Encoding Scheme. In R. Otoguro, K. Yoshimoto, K. Ishikawa, H. Umemoto, & & Y. Harada (Eds.), *24th Pacific Asia Conference on Language, Information and Computation* (Vol. 2, p. 389-398). Institute for Digital Enhancement of Cognitive Development, Waseda University. https://halshs.archives-ouvertes.fr/halshs-00549764

Ide, N., & Véronis, J. (1995). Encoding dictionaries. *Computers and the Humanities*, *29*(2), 167–179. https://doi.org/10.1007/BF01830710

Ide, N., & Véronis, J. (1998). *Background and context for the development of a corpus encoding standard*.

Jacquet-Pfau, C. (2015). Élaboration et destinée d'une encyclopédie la fin du XIXe siècle : Les trente-et-un volumes de *la grande encyclopédie*, inventaire raisonné des sciences, des lettres et des arts par une société de savants et de gens de lett. *Éla. Études de Linguistique Appliquée*, *177*, 85–100. https://doi.org/10.3917/ela.177.0085

Lojkine, S. (2013)." Et l'auteur anonyme n'est pas un lâche … " Diderot, l'engagement sans le nom. *Littératures Classiques*, *2013/1*(80), 249–263. https://hal.archives-ouvertes.fr/hal-

01081454

Moureau, F. (2001). *Le Roman vrai de l'Encyclopédie* (pp. 124–129). Gallimard. https://catalogue.bnf.fr/ark:/12148/cb36204966x

Previous drafts of the Guidelines. (n.d.). [Text]. Retrieved May 31, 2023, from https://tei-c.org/Vault/Vault-GL.html

TEI Consortium. (2023). *TEI P5: Guidelines for Electronic Text Encoding and Interchange*. https://doi.org/10.5281/ZENODO.3413524

Williams, G. (2017). MAPPING A DICTIONARY Using Atlas ti and XML to analyse a late XVII th century dictionary. *RiCognizioni*, *7*(4), 161–180. https://doi.org/10.13135/2384-8987/2104