MAPPING A DICTIONARY

Using Atlas ti and XML to analyse a late XVIIth century dictionary

Geoffrey WILLIAMS

ABSTRACT • The problem with humanities research is that it takes time, and this is all the more true when you are dealing with old primary documents that require thorough analysis before any serious conclusions can be drawn. the advantages of creating a Text Encoding Initiative conformant dictionary are enormous as not only does the act of encoding get one close to the text and its specificities but the fully digitalised version will allow the extraction and comparison of information in an exhaustive way that working on a paper edition can never do.
This text thus describes the process of digitising in XML TEI a late seventeenth century French dictionary, principally in an attempt to understand the terminology it contains whilst carrying out on-going analysis using a CAQDAS, namely the Atlas ti system.

KEYWORDS • Dictionary, Atlas it, TEI, CAQDAS.

1. Introduction

The problem with humanities research is that it takes time, and this is all the more true when you are dealing with old primary documents that require thorough analysis before any serious conclusions can be drawn. Whilst a novel or other literary work can be read and notes taken, a dictionary is much more difficult due often to its sheer size and the complex structure of the entries. The latter also mean that digitizing a dictionary is a major, time-consuming act, even if current optical character recognition (OCR) tools are increasingly efficient and if machine systems capable of doing basic mark-up are being developed (ref Mohamed https://github.com/MedKhem/grobid-dictionaries). Nevertheless, the advantages of creating a Text Encoding Initiative conformant dictionary are enormous as not only does the act of encoding get one close to the text and its specificities but the fully digitalised version will allow the extraction and comparison of information in an exhaustive way that working on a paper edition can never do. This in-depth rigorous approach is the key element that Digital Humanities brings to the Humanities, but the problem of time remains, as until a considerable amount of the work has been encoded, it is impossible to draw even tentative conclusions. One way around this is to adopt a mixed methodology using two different, but complementary tools: XML TEI encoding for full digitalisation and Computer-Assisted Qualitative Data Analysis Software (CAQDAS) for on-going qualitative analysis of data. This text thus describes the process of digitising in XML TEI a late seventeenth century French dictionary, principally in an attempt to understand the terminology it contains whilst carrying out on-going analysis using a CAQDAS, namely the Atlas ti[1] system. This research is being carried out by the LiCoRN research group[2] and in collaboration

[1] www.atlasti.com

[2] www.licorn-research.fr

with the Consortium CAHIER[3] and COST action European Network for e-Lexicography (ENeL)[4].

2. The Dictionary

The work that is being digitised is the second edition of Antoine Furetière's *Dictionnaire Universel* (Furetière, 1701), the first attempt to build a complete dictionary for the French language. The Dutch publisher Arnaud Leers published the first edition posthumously in in 1690 (Furetière, 1690), the author having died in 1688 much affected after his fight with the French academy (Rey, 2006). Following the publication of the Dictionnaire de l'Académie Française in 1694, Leers commissioned a French protestant émigré, Henri Basnage de Beauval[5], to create a new revised edition of the *Dictionnaire Universel*, an edition that was published in 1701. This second edition had tripled in size and many of the original entries had been revised, with Basnage calling on field specialists for areas about which he lacked knowledge himself. Although the first edition has been digitised, the much larger Basnage edition has not and can only be obtained in PDF (Williams et Galleron, 2016). Size is probably one of the reasons why no serious attempt has been made to digitise it, the second is probably because there is a tendency to associate a dictionary always with its first author and to consider subsequent editions as simply a revision. An added problem is that there is very little published about the new editor, Henri Basnage de Beauval. What there is is mostly in Dutch and concerns his work as editor of a scholarly journal, the Histoire des ouvrages des savans, which despite its name is not a history, but a quarterly journal concerning the contemporary publications in the arts and sciences (Bots et Van Lieshout, 1984; Bots, 1976). The great analysis of early French dictionaries (Quémada, 1967) barely mentions Basnage and there is little reference to his lexicographical work other than his own defence of his dictionary published in the Journal des savants in 1701, and in his obituary published in the *Mémoires de Trévoux* in 1710[6].

Henri Basnage de Beauval, henceforth simply Basnage, came a well-to-do French Calvinist family in Rouen. He studied law, married and settled in Rouen, until he was forced to flee to the Low Countries where he joined his brother and the French Huguenot community in 1687 after the revocation of the Edict of Nantes outlawed French Protestants. To all intents and purposes, he remained reasonably well off and was also able to practice law (Mercier-Faivre et Reynaud, 1999; Bots, 1976). From his arrival in the Low Countries he started work on his journal. In a letter dated 22 December 1695 he speaks of the publication of the dictionary of the French Academy in the Low Countries, and then on 23rd February 1696 notes that "Mrs. Leers me pressent fort de m'engager à la revision de leur dictionnaire de Furetiere. Je leur ay seulement promis d'en repasser quelques feuilles pour essayer mes forces"7 (Bots et Van Lieshout, 1984, 113‐114). Whatever his initial plans, he did take on the massive job of revision. As said earlier, direct references to his lexicographical aims and methods are few, the only way forward will be in studying his journal and the dictionary for the sources he cites. This 'web of knowledge' will be discussed later in this paper.

[3] http://cahier.hypotheses.org

[4] www.elexicography.eu

[5] Beauval is also found spelt Bauval. Throughout the text, I shall respect the different spelling used in the different publications as it is quite easy to relate them to a single person or work.

[6] Both these primary sources have been digitised and are available at : http://www.licorn-research.fr/Basnage.html

[7] Mrs. Leers are strongly pushing me to take on the revision of their dictionary of Furetière. I have only promised to look at some sheets to test myself.

The dictionary itself consists of three copious volumes organised as A-D, E-N and O-Z. All in all, this makes for some 3225 pages condensed into an A4 format, which obviously reduces its legibility. The dictionary text consists of two columns on a page with entries grouped by their first three letters. We do not know the exact number of headwords, and will not until the digitising process is complete, even if we could spend a happy few days counting them manually. The situation is complicated by the fact that Basnage groups different senses, and also different parts of speech when dealing with participles and run-on entries under one heading. In terms of content, this is an encyclopaedic dictionary, hailed by Alain Rey (Rey, 2006) as a precursor of the encyclopaedia of Diderot and Alembert. Unlike its perceived rival from the Academy, Furetière had sought to create a universal dictionary that would include terms of the arts, crafts and sciences. The entries can be very long and detailed and have copious references to the literature of the period, the latter is of great interest as it gives the web of knowledge within which Basnage was working as he revised the dictionary. Another intriguing aspect is the extent that Basnage saw the dictionary as a teaching tool for the French language. Apart from his technique of introducing updated spelling practice by giving the headword in the form used by Furetière, but putting the content in an updated spelling, for example colomne and colonne, he gave very detailed entries for verbs and included numerous collocations. Entries often have numerous examples of use as well as citations which are not present simply for prestige, as in later French dictionaries, but to illustrate usage or add content information. This makes this a very complete dictionary and one which merits detailed analysis, in academic terms, one might say that there are several potential theses around this one work. Given its size and breadth of content, adequate tools are need to handle the mass of data. Here I propose two, Text Encoding and CAQDAS.

3. Tools and methods

3.1 Digitalisation and TEI

What this paper seeks to do is to show the complementarity between two different approaches to legacy dictionaries. In both cases the problem is the same, the vast majority of legacy dictionaries are available only in image format, at best PDF, but often sill microfilm, if they are available at all. French scholars are fortunate in that most of the great dictionaries in French are available via the digital collections of the French national library – Gallica BNF. These are freely available for non-commercial use and so the individual researcher, or research team, is free to attempt their digitisation. This is done by either keying in or using optical character recognition. Although OCR have radically improved in recent years, legacy dictionaries with the close knit layout and specific characters, such as the s-long ſ, which is easily confused for an f, mean that keying in remains a valid strategy, particularly for very large publically funded projects which generally outsource such work to India or China. In the first stages of this project, the data was keyed in, hence the decision to only encode terms that had been found by a close reading of the dictionary using Atlas ti. Currently, we are using Abbyy Finereader[8] software, but this does entail considerable reworking of the output.

Once the data has been rendered machine readable, the next stage is mark-up following the TEI Guidelines for dictionaries[9]. The point of mark-up is to make certain features, such as structural elements as definitions, usage or etymology explicit so that they can later be called up, analysed and compared, and even given deeper mark-up within an analytical framework. I shall not detail the mark-up adopted except when looking at specific issues as this is not the purpose of

[8] https://www.abbyy.com/fr-fr/finereader/

[9] http://www.tei-c.org/release/doc/tei-p5-doc/en/html/DI.html

this paper, however, as a baseline the project is endeavouring to follow the guidelines being laid down in the TEI-Lex0 simplified framework being built within the European Network for e-Lexicography COST action[10]. The TEI markup is a long, time-consuming process, but is a fairly classic task, more unusual is the use of CAQDAS alongside the marking up process.

3.2 CAQDAS and Atlas ti

Humanities research is basically qualitative. It entails collecting, ordering and analysing large amounts of data from which an interpretation can be drawn. This means making numerous notes and cross-references, and, in the case of textual data, leaving a forest of page markers. If a truly rigorous and broad picture is to be drawn, this is far from optimal as a method and it is one area where digital tools can be precious for ordering data whilst not changing the basic research methodology.

The need for such tools was recognised a long time back by sociologist seeking to build models bottom up from the data. This approach was termed grounded theory (Glaser et Strauss, 1967). CAQDAS were developed so as to meet the needs of sociologists applying a grounded approach on the basis that they needed tools to analyse very large amounts of potentially heterogeneous data types, from textual to audio-visual sources, printed documents, images and transcriptions of interviews. The point of a CAQDAS is to be able to add codes and comments to all forms of data so that the researcher can gradually build a bottom up ontology of codes and their interpretations. The tool allows to link information across the different forms of data. Although such tools are now common in sociology and the political sciences, they have as yet to really penetrate the humanities[11].

4. The problem of the dictionary

The big problem is the sheer size of the dictionary coupled with the classic problem in lexicography of an alphabetical structure that disperses word families. In Western European Dictionaries, the alphabet is both a blessing and a curse. It is a blessing as we have a 26 letter alphabet and are used to it. We have been taught to think in terms of alphabetical lists and simply forget that it is not the same in all languages and that the first three letters of a word may not always be the best way to find the word we seek, all the moreso as we also expect the dictionary to group words by lemma. The curse is precisely the fact that neat lemmata are not what we necessarily encounter in text and also certain information, such as synonyms, collocations, terms are better grouped in some way. Modern electronic dictionaries should be able to handle data in different ways, although they rarely can as we are far from having truly electronic dictionaries. Here though, we have a PDF of a late seventeenth century dictionary. As a PDF we can either scroll it on screen, and possibly use a PDF reader to add notes and comments that we'll then have to scroll to find later, or we can print it and use the time honoured technique of highlighter and post-it, with comments in a notebook. It works. However, we can do the same and better with Atlas ti.

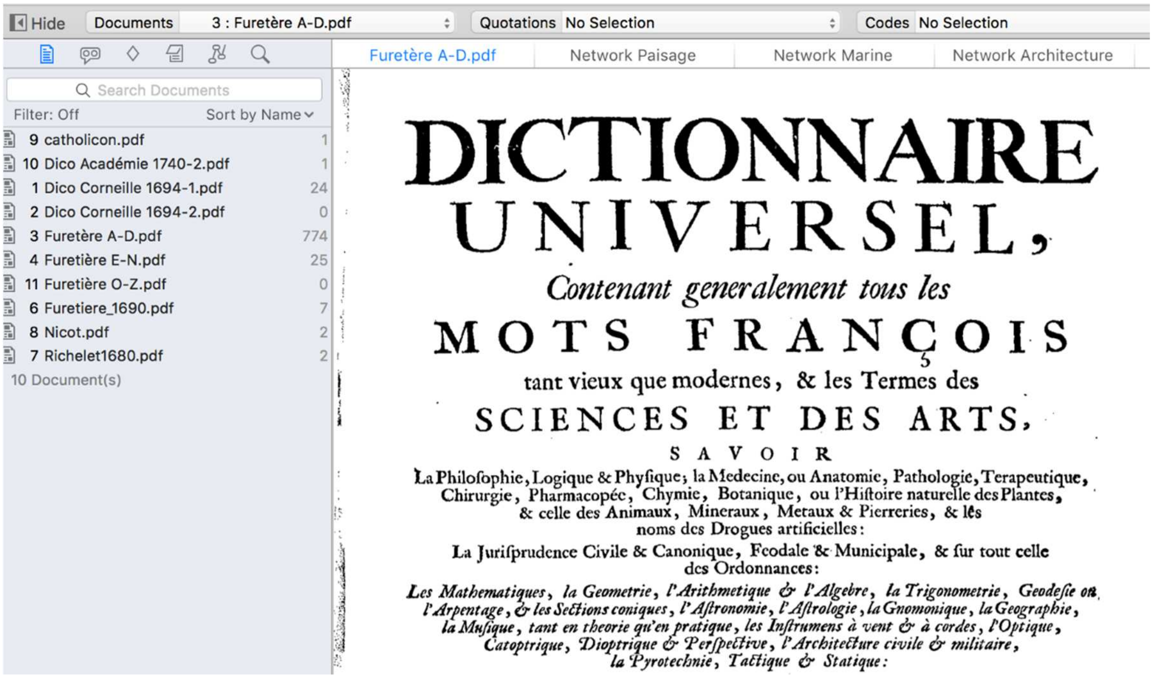
Let us start with three simple questions: what words are explicitly marked as being terms? What formulae does he use to mark out terms? And, what are the domains and crafts described.

[10] https://docs.google.com/document/d/1GPfXG3KtwApTSyAfyM3soVAiw2IyVXbnHGsFfAVM7N4/edit

[11] The LandLex Training school in Waterford, Ireland has a dedicated session on the use of Atlas ti and related tools. http://www.licorn-research.fr/LandLex_TS

4.1. Mapping terms

The advantage of Atlas ti is that a library of digital texts can be built with the texts activated when needed. This is the case here with a library consisting of the three volumes of Basnage, but also the single volume 1690 edition, and other dictionaries. In this way, a number of dictionaries can be called up so that the same word can be coded in different dictionaries, for instance between the two editions, or in the case of terms the Corneille dictionary of terms that the Academy had hastily commissioned to attempt to counter the success of Furetière. It is possible to have a split screen with several dictionaries displayed simultaneously. For the needs of the questions listed above, only the first volume of the second edition is open in Atlas ti (Fig 1.) as having such heavy digital images open uses a lot of machine memory and slows processing.



9 catholicon.pdf	1
10 Dico Académie 1740-2.pdf	1
1 Dico Corneille 1694-1.pdf	24
2 Dico Corneille 1694-2.pdf	0
3 Furetère A-D.pdf	774
4 Furetière E-N.pdf	25
11 Furetière O-Z.pdf	0
6 Furetiere_1690.pdf	7
8 Nicot.pdf	2
7 Richelet1680.pdf	2

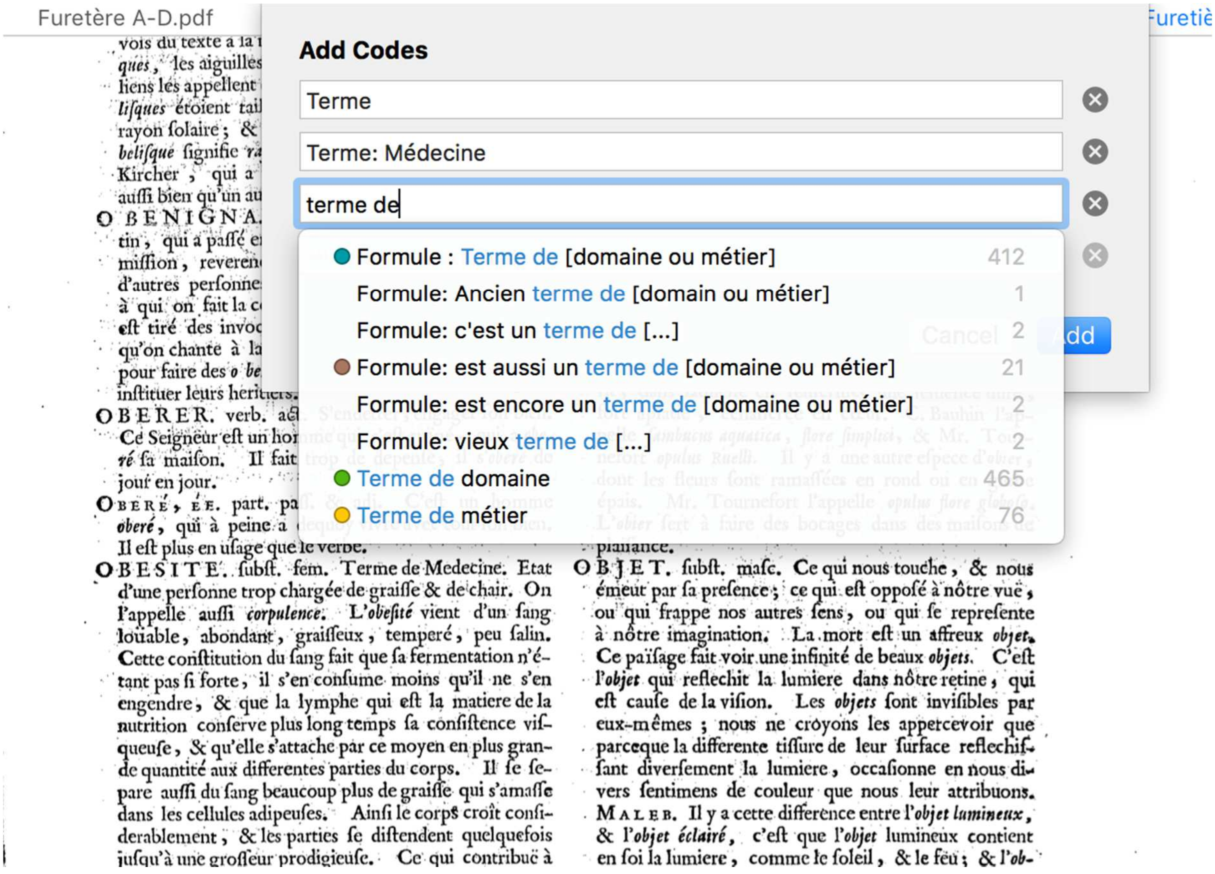
10 Document(s)

DICTIONNAIRE
UNIVERSEL,
Contenant generalement tous les
MOTS FRANÇOIS
tant vieux que modernes, & les Termes des
SCIENCES ET DES ARTS,
SAVOIR
La Philoſophie, Logique & Phyſique; la Medecine, ou Anatomie, Pathologie, Terapeutique,
Chirurgie, Pharmacopée, Chymie, Botanique, ou l'Hiſtoire naturelle des Plantes,
& celle des Animaux, Mineraux, Metaux & Pierreries, & lès
noms des Drogues artificielles:
La Juriſprudence Civile & Canonique, Feodale & Municipale, & ſur tout celle
des Ordonnances:
Les Mathematiques, la Geometrie, l'Arithmetique & l'Algebre, la Trigonometrie, Geodeſie ou
l'Arpentage, & les Sections coniques, l'Aſtronomie, l'Aſtrologie, la Gnomonique, la Geographie,
la Muſique, tant en theorie qu'en pratique, les Inſtrumens à vent & à cordes, l'Optique,
Catoptrique, Dioptrique & Perſpective, l'Architecture civile & militaire,
la Pyrotechnie, Tactique & Statique:

Fig. 1 Digital document library

The first task is coding, and that requires scrolling through the dictionary while resisting the temptation of reading everything[12]. The temptation is great as this is a window on another world. In order to answer our questions, we need to stop and code whenever an entry or sense is marked as a term. If you take the word *obesité* it is first marked as being a term, as we want to know how many terms there are in the dictionary, it is marked as being in the medical domain and having been introduced using the formula *terme de [domain ou metier]* (term of [field or craft]). It is important to stress the codes are created as they are need and do not reflect a pre-existing schema. The term found in the dictionary was from mythology, the second from architecture and the third from pastry cookery. Thus the list gradually grows so that the definitive list will only become clear towards the end of the dictionary. However, as codes are reused, we are beginning to have quantitative data as to how many terms, how many refer to broad domains or precise crafts and how many refer to any one domain or craft.

[12] A detailed commentary on the initial stages of marking up terms is to be found in (Williams, 2017)



Furetère A-D.pdf

Add Codes

Terme

Terme: Médecine

terme de

● Formule : Terme de [domaine ou métier] 412

Formule: Ancien terme de [domain ou métier] 1

Formule: c'est un terme de [...] 2

● Formule: est aussi un terme de [domaine ou métier] 21

Formule: est encore un terme de [domaine ou métier] 2

Formule: vieux terme de [...] 2

● Terme de domaine 465

● Terme de métier 76

vois du texte a la
ques, les aiguilles
liens les appellent
liſques étoient tail
rayon ſolaire; &
beliſque ſignifie *ra*
Kircher, qui a
auſſi bien qu'un au
OBENIGNA.
tin, qui a paſſé e
miſſion, reveren
d'autres perſonne
à qui on fait la c
eſt tiré des invoc
qu'on chante à la
pour faire des *o be*
inſtituer leurs heritiers.
OBERER. verb. act
Ce Seigneur eſt un hor
ré ſa maiſon. Il fait
jour en jour.
OBERÉ, ÉE. part. pa
oberé, qui à peine a
Il eſt plus en uſage que le verbe.

OBESITE. ſubſt. fem. Terme de Medecine. Etat d'une perſonne trop chargée de graiſſe & de chair. On l'appelle auſſi *corpulence*. L'*obeſité* vient d'un ſang loüable, abondant, graiſſeux, temperé, peu ſalin. Cette conſtitution du ſang fait que ſa fermentation n'étant pas ſi forte, il s'en conſume moins qu'il ne s'en engendre, & que la lymphe qui eſt la matiere de la nutrition conſerve plus long temps ſa conſiſtence viſqueuſe, & qu'elle s'attache par ce moyen en plus grande quantité aux differentes parties du corps. Il ſe ſepare auſſi du ſang beaucoup plus de graiſſe qui s'amaſſe dans les cellules adipeuſes. Ainſi le corps croît conſiderablement, & les parties ſe diſtendent quelquefois juſqu'à une groſſeur prodigieuſe. Ce qui contribuë à

plaiſance.

OBJET. ſubſt. maſc. Ce qui nous touche, & nous émeut par ſa preſence; ce qui eſt oppoſé à nôtre vuë, ou qui frappe nos autres ſens, ou qui ſe repreſente à nôtre imagination. La mort eſt un affreux *objet*. Ce païſage fait voir une infinité de beaux *objets*. C'eſt l'*objet* qui reflechit la lumiere dans nôtre retine, qui eſt cauſe de la viſion. Les *objets* ſont inviſibles par eux-mêmes; nous ne croyons les appercevoir que parceque la differente tiſſure de leur ſurface reflechiſſant diverſement la lumiere, occaſionne en nous divers ſentimens de couleur que nous leur attribuons. MALEB. Il y a cette difference entre l'*objet lumineux*, & l'*objet éclairé*, c'eſt que l'*objet* lumineux contient en ſoi la lumiere, comme le ſoleil, & le feu; & l'*ob-*

Fig. 2. Coding

This coding and counting tells us that over 600 terms are recorded in letter A alone, and that, for instance, 29 concern architecture, but only one each, at this stage, for mythology and pastry cookery. On the other hand, there are 38 for medical practice and 95 for maritime activity. As we shall see later, neither of the last two fields are simple as there are a number of related activities with their own designation.

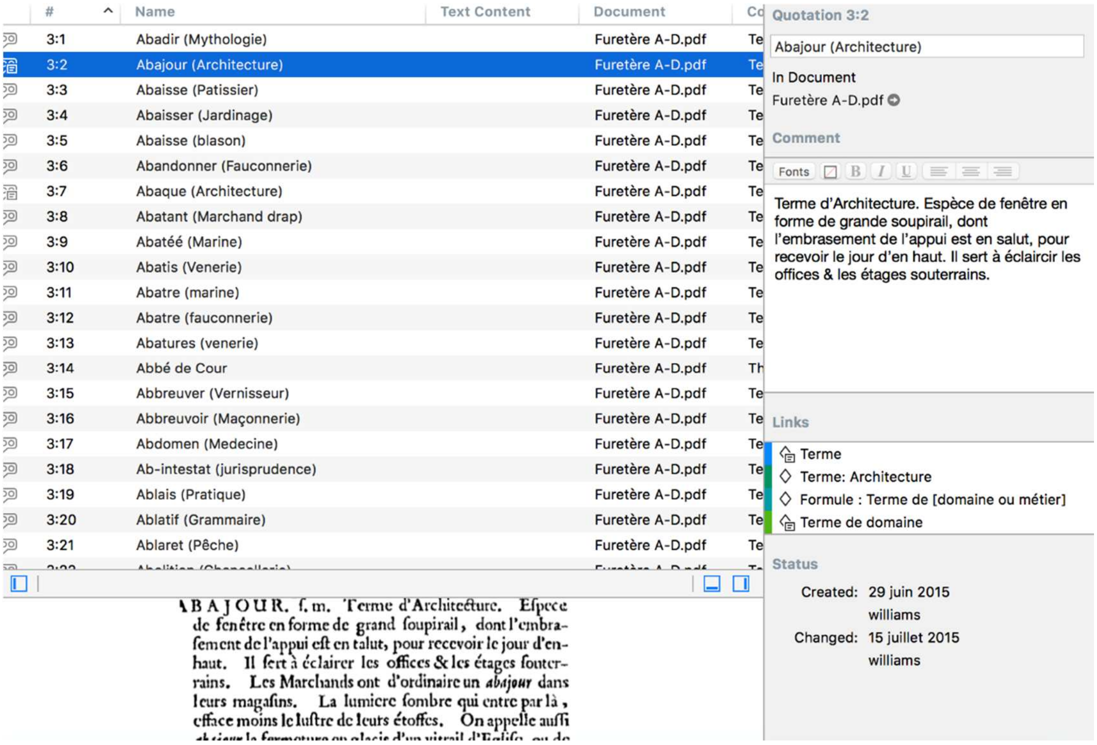
There are a number of formulae used to designate terms

- Terme de [domaine ou métier] (307)
- en termes de [domaine ou métier] (109)
- est aussi un terme de [domaine ou métier] (18)
- Terme [modifiant] (12)
- Ce terme est particulier au [utilisateur]
- Terme qui affirme
- Ancien terme de [domaine ou métier]
- C'est un terme de [...]
- Ce terme signifie la même chose que...

- On dit en termes de [domaine]
- Se dit en termes de [domaine]
- Vieux terme de [...]

The most used is a formal marker, but the others reflect the very discursive style used by Basnage. We are find the expression *terme populaire*, but this is not listed as it is using the word term simply as a synonym for ‘word’ rather than designating some precise field of use.

Whilst codes are reused, the section of data they are attached to is unique. These are termed quotations. As we work through the dictionary, we are both listing all the items that are, for example, medical terms, and also building an alphabetical list of all the words we have highlighted by the simple act of replacing the quotation number with the headword (Fig 3)



#	Name	Text Content	Document	Co
3:1	Abadir (Mythologie)		Furetère A-D.pdf	Te
3:2	Abajour (Architecture)		Furetère A-D.pdf	Te
3:3	Abaisse (Patissier)		Furetère A-D.pdf	Te
3:4	Abaisser (Jardinage)		Furetère A-D.pdf	Te
3:5	Abaisse (blason)		Furetère A-D.pdf	Te
3:6	Abandonner (Fauconnerie)		Furetère A-D.pdf	Te
3:7	Abaque (Architecture)		Furetère A-D.pdf	Te
3:8	Abatant (Marchand drap)		Furetère A-D.pdf	Te
3:9	Abatéé (Marine)		Furetère A-D.pdf	Te
3:10	Abatis (Venerie)		Furetère A-D.pdf	Te
3:11	Abatre (marine)		Furetère A-D.pdf	Te
3:12	Abatre (fauconnerie)		Furetère A-D.pdf	Te
3:13	Abatures (venerie)		Furetère A-D.pdf	Te
3:14	Abbé de Cour		Furetère A-D.pdf	Th
3:15	Abbreuver (Vernisseur)		Furetère A-D.pdf	Te
3:16	Abbreuvoir (Maçonnerie)		Furetère A-D.pdf	Te
3:17	Abdomen (Medecine)		Furetère A-D.pdf	Te
3:18	Ab-intestat (jurisprudence)		Furetère A-D.pdf	Te
3:19	Ablais (Pratique)		Furetère A-D.pdf	Te
3:20	Ablatif (Grammaire)		Furetère A-D.pdf	Te
3:21	Ablaret (Pêche)		Furetère A-D.pdf	Te

ABAJOUR. ſ. m. Terme d'Architecture. Eſpece de fenêtre en forme de grand ſoupirail, dont l'embraſement de l'appui eſt en talut, pour recevoir le jour d'en-haut. Il ſert à éclairer les offices & les étages ſouterrains. Les Marchands ont d'ordinaire un *abajour* dans leurs magaſins. La lumiere ſombre qui entre par là, efface moins le luſtre de leurs étoffes. On appelle auſſi

Quotation 3:2

Abajour (Architecture)

In Document

Furetère A-D.pdf

Comment

Terme d'Architecture. Espèce de fenêtre en forme de grande soupirail, dont l'embrasement de l'appui est en salut, pour recevoir le jour d'en haut. Il sert à éclaircir les offices & les étages souterrains.

Links

- Terme
- Terme: Architecture
- Formule : Terme de [domaine ou métier]
- Terme de domaine

Status

Created: 29 juin 2015
williams

Changed: 15 juillet 2015
williams

Fig 3. Quotations

As can be seen from figure 3, the quotation window contains links to the different codes that describe the word we have highlighted, but also has a window for comments. Insofar as we are here essentially concerned with terms, the comment section is being used hold the actual definition given in the dictionary. If the data being analysed is already machine readable, the quotation is automatically captured, but in the case of image data this is not possible and so the definitions have to be keyed in.

In this work, the quotation list gives the headwords in alphabetical order. As I am focussed on Basnage’s work, the headword is that found in his dictionary, so when I encode the same word in another work, I simply add the name of the dictionary after. This is rather advantageous as clicking on any word immediately takes me to the appropriate place in the dictionary, whichever volume it is in. If the word is in one of the other dictionaries in the library, then that dictionary is opened at the word located. In other words, this is very much like having hyper links as in any modern dictionary. The difference with most modern dictionaries, is that the encoder is free to add more codes and these are keys for searcher according to the users needs rather than be pre-ordained. As this is very much a discovery process, there are obviously points of interest that the researcher will need to note down. This is handled very easily by the memo function that allows note to be created and attached to documents and part of documents.

What we are building is a very convenient alphabetical electronic dictionary, but as said earlier, the problem is that knowledge is not alphabetical and we may need to relate words in some ways. To illustrate the best way to manage this is through networks. If we take the example from a study on architecture (Williams, « Architecture and a late 17th century dictionary: Terminology in the Basnage de Beauval 1701 edition of the Dictionnaire Universel») we can see how the relationship between terms can be visualised.

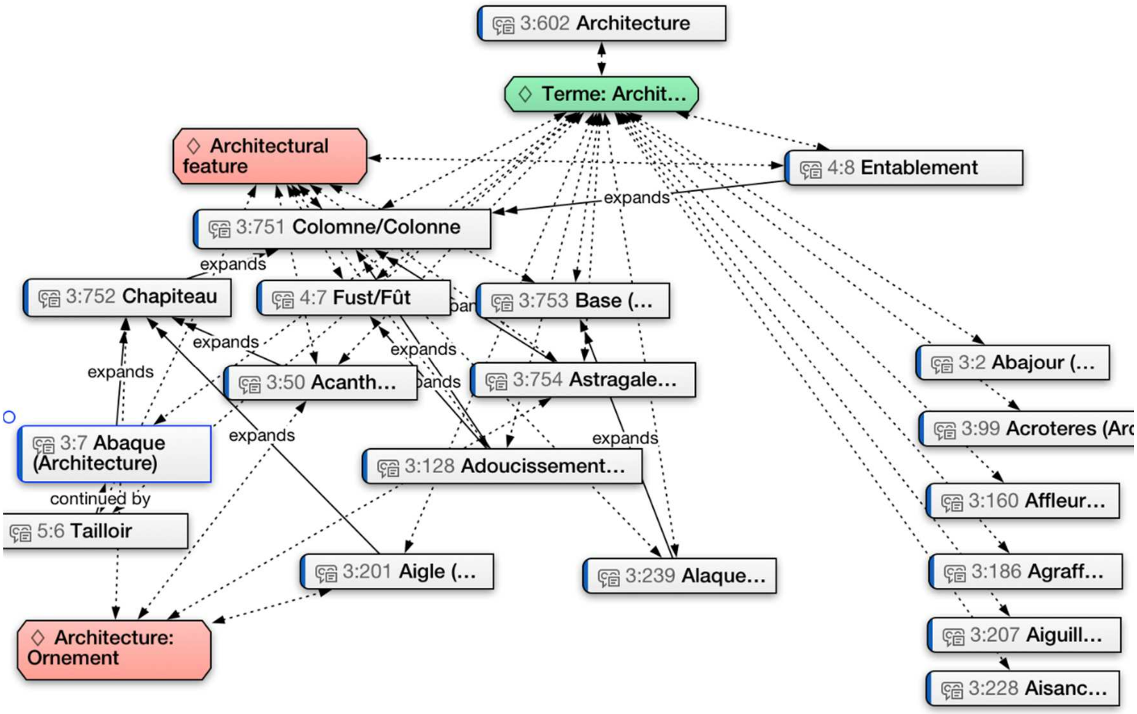
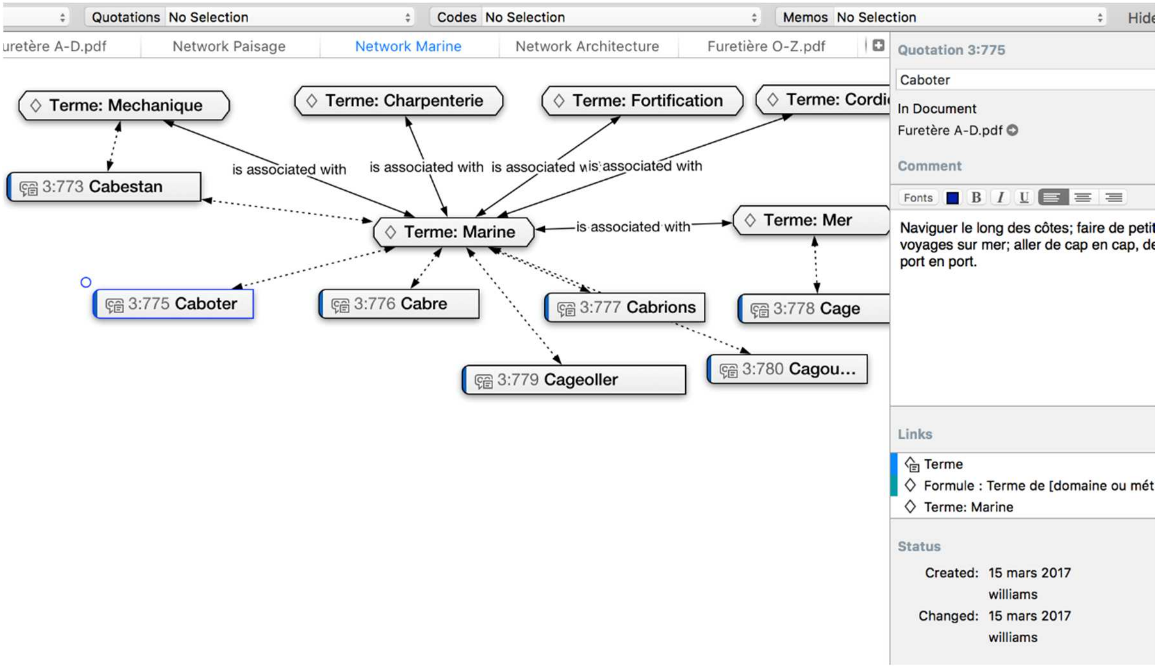


Fig. 4 Architectural network

Such a network is built by dragging entries from the code or quotation list. As this is done, the links already established through coding become apparent and other links can be added to join the elements which are moved around at will. In this case, the term data is shown in relation with the entry for the domain architecture with terms related to columns being brought together. This is not simply a convenient representation as each word in the network is clickable and will lead to the connected dictionary entry. This can be seen in another network, that built for the maritime domain (Fig. 5).



Quotation 3:775

Caboter

In Document

Furetère A-D.pdf

Comment

Naviguer le long des côtes; faire de petit voyages sur mer; aller de cap en cap, de port en port.

Links

Terme

Formule : Terme de [domaine ou mét

Terme: Marine

Status

Created: 15 mars 2017 williams

Changed: 15 mars 2017 williams

Fig 5. Maritime network

As each word is clicked on, the related comment window opens with the entry itself accompanied by the list of codes, and if needed any memos that have been attached to the word. Better than this, it is possible to go to the actual entry within the dictionary and have that displayed on screen.

However, this second network illustrates another important feature as it holds codes from several different domains

A third type of network illustrates collocational networks

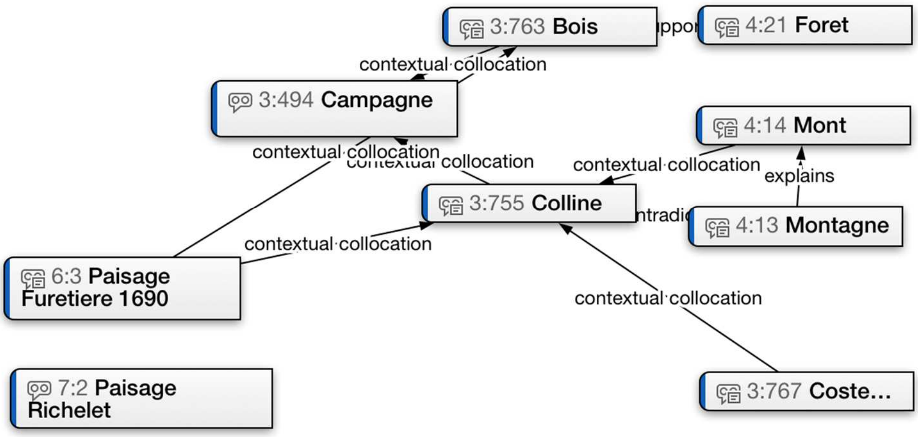


Fig 6. Paisage network

5. Interacting Atlas ti and TEI

The advantage of Atlas ti is that is allows us to work relatively quickly giving us an interesting overview of the dictionary and its contents. However, although the data can be partially shared and also generated in an XML format, it does not allow us to create a shareable version of the dictionary or allow us to go into detail. Thus, the Basnage project is also about creating an XML-TEI conformant fully open source dictionary. The fact that the data is freely available is vital as it is only in this way that researchers can appropriate the data for their own needs and add deeper layers of mark-up in line with their needs. Being able to access data via a commercial interface is no solution as the interface will only allow us to work at the level of mark-up of the system and research is not about visualisation, it is about getting into the data. This is also part of our European heritage and should be shared. As the data is standardised, it is being made available for comment, correction and non-commercial reuse on GitHub[13]

Currently, the letter C is being encoded in its entirety. As part of an educational crowdsourcing project, students scanned pages using an online OCR and then marked up the text in line with a series of recommendations. Following this, work is continuing using the Abbyy Finereader OCR, considered by the ENeL group on OCR as amongst the more reliable commercial systems[14]. At the moment, the task is to bring the data in line with the ENeL TEI-Lex0 recommendations as well as harmonising the data in line with the needs of the Basnage project. An example term can best illustrate the process:

```
<entry xml:lang="fr"
xml:id="Caillebotis"><form><orth>CAILLEBOTIS</orth></form><gramGrp><pos
ana="subst">s.</pos><gen ana="masc">m.</gen></gramGrp>
            <usg type="term" ana="#marine">Terme de Marine</usg>
             <def>Espece de treillis, ou tillac à jonc fait de menu bois, &amp; placé entre deux hiloire,
ou bordures pour servir à évaporer la fumée du canon quand on le decharge &amp; pour donner du
jour entre les ponts, quand les sabords sont fermez durant l'agitation de la mer.</def>
             <note>L'espace qui reste des ponts est couvert de bordage de pareil échantillon que celui
qui est attaché sur les membres, ou côtes du navire.</note>
          </entry>
```

The grammatical information has been noted, in this case we have a masculin (masc) noun (subst). This information has yet to be formalised beyond the simple fact of the part of speech and gender. It is apparent that Basnage also tried to classify verbs, but to analyse tis more data is needed and then the assistance of a specialist in grammars of the period. As this is clearly marked as a maritime term, this is linked to the list that has been created in Atlas ti and which has been reused in the TEI header so that the extraction of all the terms of a given field becomes possible[15]. As the project advances, both lists grow as more domains are cited.

```
<fDecl name="terminological_category">
                    <vRange>
                       <vAlt>
                          <symbol value="marine" xml:id="marine"/>
                          <symbol value="blason" xml:id="blason"/>
                          <symbol value="charpenterie" xml:id="charpenterie"/>
                          <symbol value="mythologie" xml:id="mythologie"/>
```

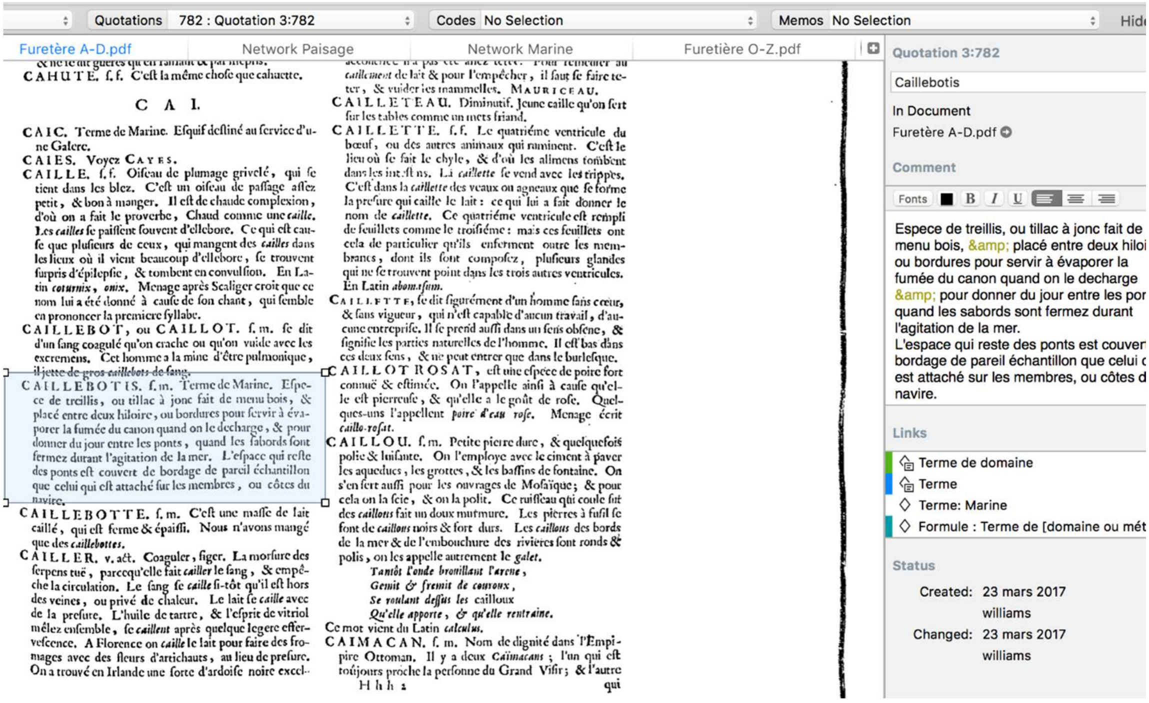
[13] https://github.com/WGBS2/Basnage

[14] https://digilex.hypotheses.org/153

[15] This list was originally created as shown as a feature declaration. In future editions it will be transformed into a taxonomy as has already been done for the list of cited authors.

```
<symbol value="architecture" xml:id="architecture"/>
<symbol value="morale" xml:id="morale"/>
```

In so far as the mark-up is for the entire page of the dictionary, whenever a term is found, Atlas ti is updated and the definition copied into the comments section as shown in figure 7.



& ne le dit gueres qu'en raillant & par mépris.

CAHUTE. ſ.f. C'eſt la même choſe que cahuette.

C A I.

CAIC. Terme de Marine. Eſquif deſtiné au ſervice d'une Galere.

CAIES. Voyez CAYES.

CAILLE. ſ.f. Oiſeau de plumage grivelé, qui ſe tient dans les blez. C'eſt un oiſeau de paſſage aſſez petit, & bon à manger. Il eſt de chaude complexion, d'où on a fait le proverbe, Chaud comme une *caille*. Les *cailles* ſe paiſſent ſouvent d'ellebore. Ce qui eſt cauſe que pluſieurs de ceux, qui mangent des *cailles* dans les lieux où il vient beaucoup d'ellebore, ſe trouvent ſurpris d'épilepſie, & tombent en convulſion. En Latin *coturnix*, *onix*. Menage après Scaliger croit que ce nom lui a été donné à cauſe de ſon chant, qui ſemble en prononcer la premiere ſyllabe.

CAILLEBOT, ou CAILLOT. ſ.m. ſe dit d'un ſang coagulé qu'on crache ou qu'on vuide avec les excremens. Cet homme a la mine d'être pulmonique, il jette de gros *caillebots* de ſang.

CAILLEBOTIS. ſ.m. Terme de Marine. Eſpece de treillis, ou tillac à jonc fait de menu bois, & placé entre deux hiloire, ou bordures pour ſervir à évaporer la fumée du canon quand on le decharge, & pour donner du jour entre les ponts, quand les ſabords ſont fermez durant l'agitation de la mer. L'eſpace qui reſte des ponts eſt couvert de bordage de pareil échantillon que celui qui eſt attaché ſur les membres, ou côtes du navire.

CAILLEBOTTE. ſ.m. C'eſt une maſſe de lait caillé, qui eſt ferme & épaiſſi. Nous n'avons mangé que des *caillebottes*.

CAILLER. v.act. Coaguler, figer. La morſure des ſerpens tuë, parcequ'elle fait *cailler* le ſang, & empêche la circulation. Le ſang ſe *caille* ſi-tôt qu'il eſt hors des veines, ou privé de chaleur. Le lait ſe *caille* avec de la preſure. L'huile de tartre, & l'eſprit de vitriol mêlez enſemble, ſe *caillent* après quelque legere efferveſcence. A Florence on *caille* le lait pour faire des fromages avec des fleurs d'artichauts, au lieu de preſure. On a trouvé en Irlande une ſorte d'ardoiſe noire excel-

accouchée n'a pas été aſſez tetée. Pour remedier au *caillement* de lait & pour l'empêcher, il faut ſe faire teter, & vuider les mammelles. MAURICEAU.

CAILLETEAU. Diminutif. Jeune caille qu'on ſert ſur les tables comme un mets friand.

CAILLETTE. ſ.f. Le quatriéme ventricule du bœuf, ou des autres animaux qui ruminent. C'eſt le lieu où ſe fait le chyle, & d'où les alimens tombent dans les inteſtins. La *caillette* ſe vend avec les trippes. C'eſt dans la *caillette* des veaux ou agneaux que ſe forme la preſure qui caille le lait: ce qui lui a fait donner le nom de *caillette*. Ce quatriéme ventricule eſt rempli de feuillets comme le troiſiéme: mais ces feuillets ont cela de particulier qu'ils enferment outre les membranes, dont ils ſont compoſez, pluſieurs glandes qui ne ſe trouvent point dans les trois autres ventricules. En Latin *abomaſum*.

CAILLETTE, ſe dit figurément d'un homme ſans cœur, & ſans vigueur, qui n'eſt capable d'aucun travail, d'aucune entrepriſe. Il ſe prend auſſi dans un ſens obſene, & ſignifie les parties naturelles de l'homme. Il eſt bas dans ces deux ſens, & ne peut entrer que dans le burleſque.

CAILLOT ROSAT, eſt une eſpece de poire fort connuë & eſtimée. On l'appelle ainſi à cauſe qu'elle eſt pierreuſe, & qu'elle a le goût de roſe. Quelques-uns l'appellent *poire d'eau roſe*. Menage écrit *caillo-roſat*.

CAILLOU. ſ.m. Petite pierre dure, & quelquefois polie & luiſante. On l'employe avec le ciment à paver les aqueducs, les grottes, & les baſſins de fontaine. On s'en ſert auſſi pour les ouvrages de Moſaïque; & pour cela on la ſcie, & on la polit. Ce ruiſſeau qui coule ſur des *caillous* fait un doux murmure. Les pierres à fuſil ſe font de *caillous* noirs & fort durs. Les *caillous* des bords de la mer & de l'embouchure des rivieres ſont ronds & polis, on les appelle autrement le *galet*.

Tantôt l'onde brouillant l'arene,
Gemit & fremit de couroux,
Se roulant deſſus les cailloux
Qu'elle apporte, & qu'elle rentraine.

Ce mot vient du Latin *calculus*.

CAIMACAN. ſ.m. Nom de dignité dans l'Empipire Ottoman. Il y a deux *Caïmacans*; l'un qui eſt toûjours proche la perſonne du Grand Viſir; & l'autre

Hhh 2 qui

Quotation 3:782

Caillebotis

In Document

Furetère A-D.pdf

Comment

Espece de treillis, ou tillac à jonc fait de menu bois, &amp; placé entre deux hiloi
ou bordures pour servir à évaporer la fumée du canon quand on le decharge &amp; pour donner du jour entre les por
quand les sabords sont fermez durant l'agitation de la mer.
L'espace qui reste des ponts est couver
bordage de pareil échantillon que celui c
est attaché sur les membres, ou côtes d
navire.

Links

- Terme de domaine
- Terme
- Terme: Marine
- Formule : Terme de [domaine ou mét

Status

Created: 23 mars 2017
williams
Changed: 23 mars 2017
williams

Fig. 7 Caillebottis

Caillebottis is a fairly simple entry without examples, entries for verbs can be much more complex, as here with the case of *abatre*:

```
<sense n="2"><usg type="dom" ana="#marine"><oRef corresp="#abatre"></oRef>ABATRE en
termes de Marine,</usg>
              <def>signifie,Decheoir, deriver, s&apos;écarter de la vraye route: ce qui se fait parla
force des courants,ou des marées, ou par les erreurs du pointage, ou du mauvais gouvernement du
timonier</def>.
                 <cit type="example"><quote>On dit aussi, qu&apos;<seg type="collocation">un
Pilote <hi rend="italic"> abat</hi> son vaisseau d&apos;un quart de
                Rumb ou d&apos;un autre aire de vent</seg>,</quote></cit>
                 <def>quand il vire ou change sa course, &amp; gouverne sur une autre Rumb de
celui de sa route.</def>
                 <cit type="example"><quote>On dit, <seg type="collocation"><hi
rend="italic">Abatre</hi> un navire</seg></quote></cit>
                 <def>pour dire, le faire obeïr au vent lors qu&apos;il est sur les voiles, ou
qu&apos;il presente trop l&apos;avant au lieu d&apos;où vient le vent.</def>
                 <cit type="example"><quote>On dit, <seg type="collocation">Le navire <hi
rend="italic">abat</hi></seg>,</quote></cit>
                 <def>lors que l&apos;ancre a quitté le fond, &amp; que le vaisseau obeït au vent
pour arriver.</def>
                 <cit type="example"><quote><seg type="collocation"><hi rend="italic">Aller à la
derive</hi></seg> s&apos;apelle aussi <hi rend="italic">abatre</hi>:</quote></cit>
```

```
                <def>c&apos;est quand on va de côté au gré du vent &amp; de la marée, au lieu
d&apos;aller en droiture.</def>
                <cit type="example"><quote>On dit aussi, <seg type="collocation"><hi
rend="italic">Abatre</hi> un vaisseau sur le côté</seg>,</quote></cit>
                <def>lors qu&apos;on veut travailler à la carene, ou en quelque endroit des oeuvres
vives.</def></sense>
```

Abattre is a highly polysemic verb with 7 different senses declared, of which 2 are noted as being terms relating to maritime activity, as shown above, and falconry. However, as so often with verb, the definition is in fact a list of synonyms so that even the terminological sense is polysemic. What Basnage does is to provide a series of examples that are in fact collocations which to the knowing illustrate the different interpretations of the word from steering a boat, allowing it to drift, or heaving it down for maintenance. It is clear impossible to get a true picture of what Basnage is describing by simply accumulating a list of what is mainly verb – noun collocations. By marking up the collocations as segments, it will be possible to extract all the collocations from the dictionary so as to carry out a linguistic analysis, but also to relate them to domains of use and above all to classify them according to the different actions described, for instance, in maritime terms, steering a vessel, attacking another vessel, carrying out maintenance. Deeper mark-up will ultimately allow to link synonyms to precise sub-senses through close analysis of examples and citations. The outcome will be far more than simply an encyclopaedic description of a word, but an insight into naval manoeuvres in the late seventeenth century.

As these entries or senses are added to the Atlas ti database, so they can be added to the network in an attempt to show the relations between the different terms. Looking back to figure 5, we can now see how the interrelationship between domains comes into play. The term *cabestan* (capstan) is marked as being mechanical in nature, which it is, but, as the encyclopaedic note makes clear, it is primarily on ships to raise sails and so must necessarily be linked to maritime activity.

```
<note>C'est en virant  les <hi>cabestants</hi>qu'on remonte les bateaux, qu'on tire sur terre les
vaisseaux pour les calfeutrer, qu'on les decharge des plus grosses marchandises, qu'on lleve les
ancres &amp; les voiles, &amp;c Il y a deux <hi>cabestans</hi> sur les vaisseaux. Le grand
<hi>cabestan</hi> est posé sur le premier pont, &amp; s'éleve jusqu'a quatre ou cinq pieds de
hauteur au dessus du deuxième. On le nomme <seg type="term_variant">cabestan double,</seg> à
cause qu'il sert à deux étages pour lever les ancres, &amp; qu'on peut en doubler les forces, en
mettant du monde sur les deux pons pour le virer, étant garni de barres et d'autres pièces, comme
taquets, entremises, &amp; languettes, pour le tourner, &amp; arrêter. Le <seg
type="term_variant">petit <hi>cabestan</hi></seg>, ou <seg rend="italique">cabestan
simple</seg>, estposé sur le second pont entre le grand mât,&amp; Le mât de misaine, qui sert à
faire isser les mâts de hunes &amp; les grandes voiles, où il faut moind de force qu'à élever les
ancres. On appellee <seg type="term_variant" rend="italique"> cabestan à l'Angloise</seg>, celui
où l'on employe que des demies barres, &amp; qui à cause de cela n'est percé qu'à moitié. Il est plus
renflé que les <hi rend="italique">cabestans</hi> ordinaires. Il y a aussi un <seg
type="term_variant" rend="italique">cabestan volant.</seg> C'est celui qu'on peut transporter d'un
lieu à un autre. </note>
```

Looking at the network, it can be seen that I have added in *charpenterie* (carpentry). This is not so surprising in a world where ships were built in wood and where the best carpenters were employed by the navy, it is sufficient to look at many 17th century country houses to see own the roof structure had clearly been based on naval practice. Other related domains are fortification as France was fortifying its coast at the time, not the interior. The advantage of building networks is precisely to see the interaction between domains so as to build a wider picture of how the terms were used and in what contexts.

Using the various facilities and tools provided by Atlas ti, it is possible to maintain an ongoing analysis of the data at a global level. However, for finer analyses, the key tool is XML TEI and deeper mark-up.

6. Deep mark-up: beyond Atlas ti

The dictionary entry can be very long and is generally structured as a single or series of senses. These do not follow a standard pattern and so mark-up requires a lot of interpretation. Generally however there is a definition, and encyclopaedic entry, tagged as <note> and often an etymology section, <etym>. These do not necessarily follow in a standardised order and may be interspersed with citation and examples. Etymology is not marked up as such but is found through phrases as ‘*Ce mot vient de …* (This word comes from [language], ‘*Menage le fait venir de ...*’ ((Ménage, 1650) says it comes from…) or simply ‘*En …* (In [language]). He does use a variety of sources, but the principal one is Gilles Ménage whose influential work on the French language, *Origines de la langue françoise,* was published in 1650 (Ménage, 1650). As with the introductory formulae for terms listed earlier, it may also be possible to list the indicators for etymology.

Basnage illustrates through example and citation as illustrated in the first sense for *cabane* (hut):

```
<entry xml:lang="fr" xml:id="Cabane"><form><orth>CABANE</orth><gramGrp><pos
ana="subst">s.</pos><gen ana="masc">m</gen></gramGrp></form>
            <sense n="1"><def>Petit toit ou maisonette bâtie de bauge &amp;amp; couverte de
chaume, où logent les pauvres gens, &amp;amp; sur tout à la campagne. </def>
               <cit type="example"><quote>Les solitaires meprisoient le sejour des villes, pour aller
dans les deserts habiter <pb/> des <hi>cabanes</hi></quote><bibl>DU PIN</bibl></cit>
               <cit type="citation"><quote><bibl
ana="#Malherbe_ISN0000000110606733">Malherbe</bibl> a dit en parlant de la
mort</quote><cit><quote><lg><l>Le pauvre en ca cabane où le chanoine le couvre,</l>
                  <l>Est sujet à ses loix, &amp;c</l></lg></quote></cit></cit>
               <etym>Ce mot vient de l'Italien <foreign xml:lang="it">cappana</foreign> qui
signifie petite maison de chaume, qui a été fait du Grec <foreign xml:lang="gr">kabané</foreign>,
signifiant creche. <name ana="#Men_ISN0000000080815971"></name>MEN. <name
ana="#Isidore_ISN0000000122756296">Isidore</name> dit que le mot de <foreign
xml:lang="it">capanna</foreign> vient <foreign xml:lang="lat">ex co qued unum tantium
hominen capiat.</foreign>. Les Espagnols disent aussi cabana.</etym>
```

This entry brings in an important aspect of Basnage’s work, his knowledge network. As editor of a literary and scientific journal he followed closely intellectual activities of his time. Citations are never there simply to add prestige as having been used by a ‘best author’, to quote Johnson’s term, but to illustrate usage and bring in supplementary knowledge. He also gives, where possible, etymological information, but rather than giving a single source will accept that there may be several interpretations and gives his sources. At the head of the dictionary, he lists abbreviations and the full name of the author or source text. These have all been listed in the TEI header, but leave us with a series of problem and a challenge. The problem is that the information is minimal as these persons were all presumed to be known at the time, they are not now. He other difficulty is that the printer did not always respect the list and a number abbreviations do not correspond. Finally, a number of persons cited are not in the list, which leaves us with only the citation and an elliptical abbreviation to trace them. The challenge is thus to use the list of authors so as to link to the people and work he actually cited. This is done by the use of a detailed list in the <SourceDesc> of the TEI header.

```
    <listPerson>
            <head>Personnes indexés par Basnage comme source de citations</head>
            <person>
             <persName ref="ISN:0000000083391828"
xml:base=https://fr.wikipedia.org/wiki/Jacques_Abbadie
xml:id="Aba_ISN0000000083391828"><forename>Jacques</forename><surname>Abadie</s
urname><abbr>ab ou abs</abbr></persName>
          <birth when="1654">1654</birth>
          <death when="1727-09-25"> 25 septembre 1727</death>
          <nationality>française</nationality>
          <event when="1699">
            <label>Dean of Killaloe and Clonfert</label>
          </event>
          <note>pasteur et théologien protestant français</note>
        </person>

        <person>
          <persName
            xml:base="https://fr.wikipedia.org/wiki/Nicolas_Perrot_d'Ablancourt"
            xml:id="Abl_ISN0000000107866286"><name>Mr.
d'Ablancourt</name><abbr>Abl</abbr></persName>
          <birth when="1606-04-06">6 April 1606</birth>
          <death when="1664-11-17">17 novembre 1664</death>
          <nationality>française</nationality>
          <event type="Acad_Fr" when="1637">
            <label>entre à l'Académie Française</label>
          </event>
          <note>Protestant de temps à autre. Traducteur contesté - "la Belle
infidèle"</note>
        </person>
        <personGrp>
          <bibl xml:base="https://fr.wikipedia.org/wiki/Académie_française"
            xml:id="Acad_Fr_ISN0000000404717682"><orgName>Académie
Française</orgName><name>Messrs de l'Academie Françoise</name><abbr>Ac. ou Mrs de
l'Ac.</abbr></bibl>
        </personGrp>
```

For each person listed, we have the abbreviation or abbreviations found in the dictionary as well as the full name of the person cited. Also given is the link to the Wikipedia page when that exists. Each person receives a unique identifier composed of the name or an abbreviation of the name with their International Standard Name Identifier (ISNI)[16] identifier. An ISNI entry lists all the known name and spelling variations for a given person and as such allows us to remove any ambiguity. For example, Basnage cites two persons with the name Régnier, using their ISNI number it is possible to remove any confusion.

```
    <person>
            <persName
              xml:base="https://fr.wikipedia.org/wiki/François-Séraphin_Régnier-Desmarais"
    xml:id="Ab_Regn_ISN0000000117402354"><name>François-Séraphin Régnier-
```

[16] http://www.isni.org

```
Desmarais</name><abbr>Ab. Reg</abbr></persName>
            <birth when="1632-08-13">13 août 1632</birth>
            <death when="1713-09-06">6 septembre 1713</death>
            <nationality>Française</nationality>
            <event when="1670">
              <label>entre à l'Académie Française</label>
            </event>
            <event when="1667">
              <label>entre à l'Accademia della Crusca</label>
            </event>
            <note>Linguiste, poète, diplomate, traducteur, écrivain</note>
          </person>
          <person>
            <persName xml:base="https://fr.wikipedia.org/wiki/Mathurin_Régnier"
              xml:id="Regn_ISN0000000108814744"><name>Mathurin
Regnier</name><abbr>Regn</abbr></persName>
            <birth when="1573-12-21">21 décembre 1573</birth>
            <death when="1613-10-22">22 octobre 1613 </death>
            <nationality>Française</nationality>
            <note>poète satirique</note>
          </person>
```

The same process is followed for authors who are not in the list and will ultimately be done for all persons cited. A similar process is done when a work is cited as this happens with works that were particularly known at the time, as were the '*Œuvres mêlées de Saint-Évremond*'. In this case, the task is to link the work to an author and also to an extant edition of the work in a public library. This is illustrated below:

```
 <taxonomy n="Ouvrages_indexés">
<listBibl>
          <bibl xml:base="https://fr.wikipedia.org/wiki/Le_Comte_de_Gabalis"
corresp="http://catalogue.bnf.fr/ark:/12148/cb37280623n" xml:id="Comte_de_Gabalis"><title>Le
Comte de Gabalis ou Entretiens sur les sciences secrètes,</title> <author
ana="#Villars_ISN0000000121185233">Abbé de Villars</author><date>1670</date></bibl>
          <bibl xml:base="https://fr.wikisource.org/wiki/Œuvres_mêlées_de_Saint-Évremond"
corresp="Gallica-ark:/12148/bpt6k57754j" xml:id="Oueuvres_melees"><title>Œuvres mêlées de
Saint-Évremond</title><author ana="#St_Ev_ISN0000000356744298">St
Evremond</author></bibl>
```

In the case of the two works cited above, the texts are freely available for download in PDF format from the French national library digital archive, which raises the intriguing possibility of exploring the texts with Atlas ti to see how the words to which they are linked are used in the wider context of the published work, and ultimately to create an electronic corpus for the period, working from dictionary to corpus, rather than the reverse which is the modern situation.

6.1 Extracting information

Any research tool is a means and not the end. It is all to easy to forget this as ever more exciting digital humanities tools for visualisation come available. An XML TEI legacy dictionary can be transformed into many formats making consultation as a traditional dictionary possible easier, but the point of making a dictionary available in XML is to be able to interrogate the dictionary for the purposes of research. The idea behind a fully open source resource is that users can add their own deeper mark-up based on their own needs.

An electronic text can be analysed as part of a corpus, but concordancer is not a viable tool for dictionary analysis. Insofar as a dictionary resembles a database, then a too that can analyse in this way is required. BaseX[17] is one such tool in that it uses XQuery, the XML query system. It is a powerful tool, but not easy to use. However, in mastering its capacities, it is possible to build a dedicated and evolving interface so that the user does not need o worry about programming a complex XQuery. Two simple examples can illustrate the use of BaseX: a closer look at idioms and etymology.

As has already been underlined, the aim in the Basnage project is to describe the data and not impose a pre-existing system. This is why the term extraction follows the list of arts, crafts and sciences established by Basnage himself. Using Atlas ti, these domains are listed as they appear, and tentative groupings are made using the network tool as shown above (Figs. 4, 5 & 6). The same lists are them used when marking up the terms in TEI. Although much work has to be done later extracting, grouping and analysing terms, these have been classified into categories by the author. The case of idiomatic input is different and requires another approach as such formulae appear in the text, especially in the examples and citations, but are not signalled as such. To handle these elements, the <seg>, segment, element used with four broad categories of collocation, idiom, proverb and term-variant. Collocation is taken in its broadest sense of being a relationship between two words, here generally verb - noun or modifier – noun. Idioms are longer conventional units and proverbs are when Basnage describes units as being proverbs as '*chaud comme une caille'* (warm as a quail) under the entry for '*Caille'* (Quail). Thus, to go further in creating a classification we need to extract the segments and see whether an order can be imposed using an accepted classification, or building a new one from the data.

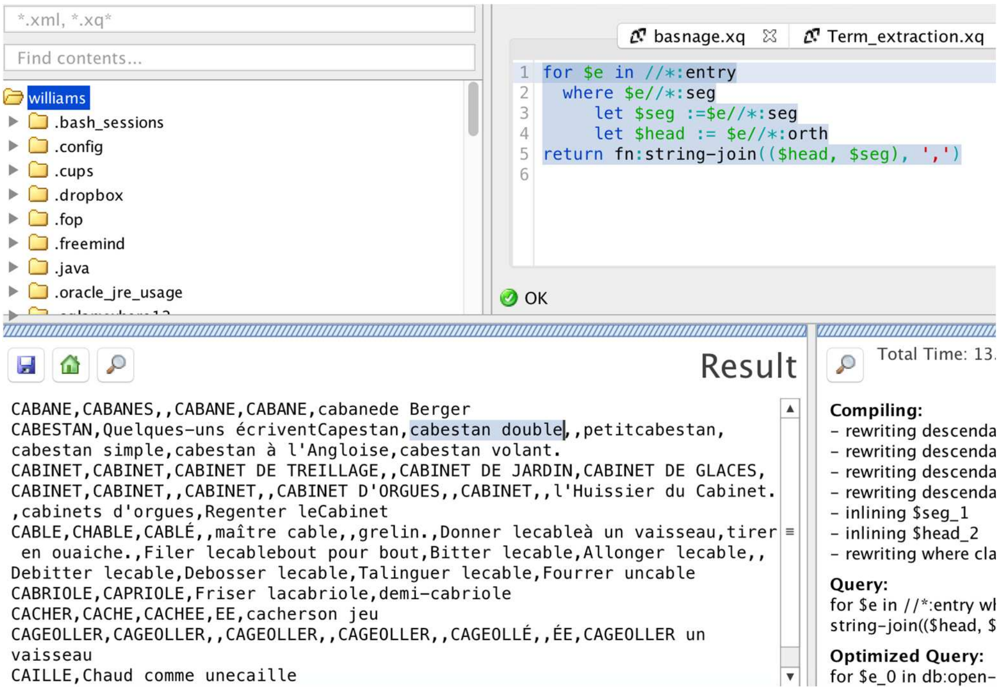
Simply extracting the element segment, or even segment with its attribute as <seg type="collocation"> is not sufficient as we would not know the headword to which it is related, we therefore need a query that extracts the segments at the same time as the related headword as in the example below.

```
for $e in //*:entry
  where $e//*:seg
     let $seg :=$e//*:seg
     let $head := $e//*:orth
return fn:string-join(($head, $seg), ',')[18]
```

This results in the output shown in Figure 8 where we see, for example, '*cabane de berger*' (shepherd's hut), a noun-noun form, 'cabestan double' (double capstan) and other term variations and under '*cable'* (rope) verbal collocations as '*donner un cable à un vaisseau'* (throw a rope to a vessel).

[17] https://basex.org. The tool is multiformat running on both Mac and PC, under Windows or Linux.

[18] The author thanks Dr. Ioana Galleron for her knowledge of BaseX and her precious help in building queries.



```
for $e in //*:entry
  where $e//*:seg
      let $seg :=$e//*:seg
      let $head := $e//*:orth
return fn:string-join(($head, $seg), ',')
```

```
CABANE,CABANES,,CABANE,CABANE,cabanede Berger
CABESTAN,Quelques-uns écriventCapestan,cabestan double,,petitcabestan,
cabestan simple,cabestan à l'Angloise,cabestan volant.
CABINET,CABINET,CABINET DE TREILLAGE,,CABINET DE JARDIN,CABINET DE GLACES,
CABINET,CABINET,,CABINET,,CABINET D'ORGUES,,CABINET,,l'Huissier du Cabinet.
,cabinets d'orgues,Regenter leCabinet
CABLE,CHABLE,CABLÉ,,maître cable,,grelin.,Donner lecableà un vaisseau,tirer
 en ouaiche.,Filer lecablebout pour bout,Bitter lecable,Allonger lecable,,
Debitter lecable,Debosser lecable,Talinguer lecable,Fourrer uncable
CABRIOLE,CAPRIOLE,Friser lacabriole,demi-cabriole
CACHER,CACHE,CACHEE,EE,cacherson jeu
CAGEOLLER,CAGEOLLER,,CAGEOLLER,,CAGEOLLER,,CAGEOLLÉ,,ÉE,CAGEOLLER un
vaisseau
CAILLE,Chaud comme unecaille
```

Compiling:
- rewriting descenda
- rewriting descenda
- rewriting descenda
- rewriting descenda
- inlining $seg_1
- inlining $head_2
- rewriting where cla

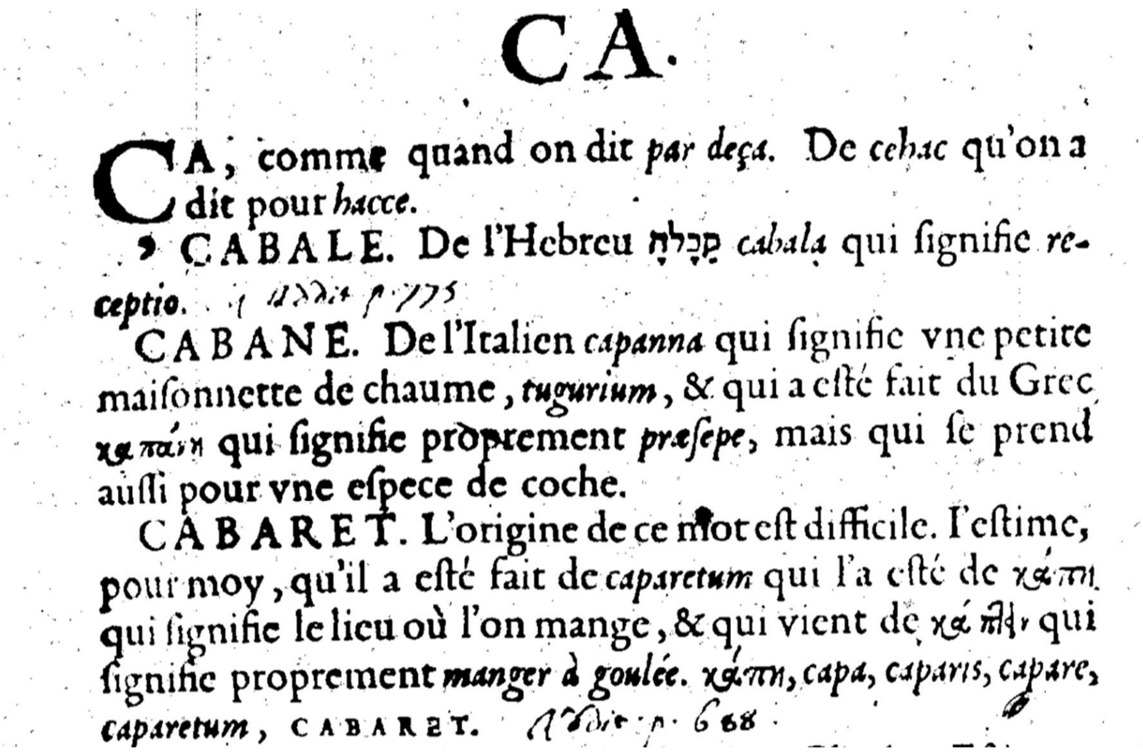
Query:
for $e in //*:entry wh
string-join(($head, $

Optimized Query:
for $e_0 in db:open-

Figure 8 Extraction of segments using BaseX

Doing such fine coding in Atlas ti would not be ideal as far too much data would be generated. It is also not ideal to carry out an on-going analysis with such data as it is only understandable when a series of formulae have been located. To analyse idioms a lot of data is required, and these must in turn be related to the entry and entry type. In this way, an ontology of terms and their variants can be built, and then these can be demonstrated in Atlas ti. More complex queries would be needed to link segments to headwork and their domain of usage so as to start a more elaborate analysis of collocations by domain, and by the function within a domain.

The situation is the same with etymologies and a very similar extraction code would be needed. As with idiomatic usage, it is the encoder who has code the parts of the text that contain etymological information. This is not necessarily neatly contained in one place and there is more or less detail. Thus, as with segments, it is necessary to extract the etymologies using a very similar Xquery to that for <seg> and then carry out an analysis. In this case, the sources have been coded where possible, which means that it still be possible to find the information from different authors and in many cases even link to the source document as illustrated below (Figure 9) with an example of entries from the Dictionary of Gilles Ménage.



CA.

CA, comme quand on dit *par deça*. De *cehac* qu'on a dit pour *hacce*.

CABALE. De l'Hebreu קַבָּלָה *cabala* qui ſignifie *receptio*. (addit p. 775

CABANE. De l'Italien *capanna* qui ſignifie vne petite maiſonnette de chaume, *tugurium*, & qui a eſté fait du Grec καπάνη qui ſignifie proprement *præſepe*, mais qui ſe prend auſſi pour vne eſpece de coche.

CABARET. L'origine de ce mot eſt difficile. I'eſtime, pour moy, qu'il a eſté fait de *caparetum* qui l'a eſté de κάπη qui ſignifie le lieu où l'on mange, & qui vient de κάπτειν qui ſignifie proprement *manger à goulée*. κάπη, *capa*, *caparis*, *capare*, *caparetum*, CABARET. addit p. 668

Figure 9 '*Cabane*' in Ménage

Like with idioms, the etymological information is rich and Basnage is not afraid of giving different opinions on the roots of a word. Whether the information is accepted in modern etymologies, and much of Ménage is, it illustrates the development of the discipline of etymology and also how a given unit of language has been perceived in the past. In both cases illustrated, BaseX has provided a large amount of selected data, and as with all humanities research, that requires human knowledge and wisdom to interpret. However, to interpret such a mass of data, it would make sense to generate text that can then be coded with Atlas ti, thus closing the circle.

7. Conclusion

This text has dealt with two issues: methodological issues of digital humanities and lexicographical issues relating to the analysis of legacy dictionaries. To conclude, it will be necessary to look at each of the aspects in turn.

To sum up, in technical terms, we have two data types and two tools for their analysis. Much humanities data is in non-machine readable format, and will remain so for a long time. Much has been micro-filmed, and some of that rendered into online PDF, but it is the tip of an enormous archival iceberg of text, without talking of the manuscripts which can be scanned, but not handled by OCR. This data can best be handled by CAQDAS as heterogeneous in format and quality. As has ben shown, there is no better tool for coding and analysing such dispersed textual data, and even adding in other image data, including paintings. This is highly relevant to the work on Basnage as his journal, the *Histoire des ouvrages des savans*, is a major unexploited source for understanding his dictionary, as well as his other contemporary sources. In addition, there are numerous engravings illustrating the crafts of his period and linking these to his dictionary entries could be very instructive, and only a CAQDQS as Atlas ti can do this. Thus, we can say that a CAQDAS system is an idea tool for carrying out humanities research on textual data without having to change in any way normal humanities research paradigms. The disadvantage is that when dealing with great masses of data as in dictionary analysis its is limited to assembling and obtaining an overview of disparate data, to go deeper, then encoding the document in XML TEI is essential.

Digitising an entire dictionary is a daunting task, but as has been shown, the complementarity between a CAQDAS analysis and TEI encoding enables the researcher to see where priorities lie in encoding all or part of a dictionary. XML mark-up gives the potential for deep analysis and the automated extraction of data, but above all it means that data can be shared so that other researchers can build on what has been done before. Surface mark-up can simple make content available, but this is not sufficient as the researcher will want to differentiate definitions and encyclopaedic data, isolate idiomatic forms, analyse examples and citations, and link to external sources. This is best done through encoding and hyperlinks in a fully machine-readable document. Insofar as dictionaries form part of our common European heritage, it is normal that for research purposes and public consultation they should be available as both open access and open source.

The second aspect of this conclusion concerns the *Dictionnaire Universel* in its 1701 edition by Henri Basnage de Beauval. This is obviously the subject of on-going research as we get deeper into the three volumes of this massive work. To date, what has been revealed is the great range of terminology for arts, crafts and sciences that is included in the work. The work demonstrates the ontology of Basnage, but also how other ontologies can be built that link domains, such as the many crafts and sciences that surround maritime activities. This networking of the dictionary should make access to the dictionary easier for the modern user by effectively making a legacy dictionary into an e-dictionary and e-terminology.

One interesting aspect of Basnage's description of terms is his use of examples and citations to illustrate usage and introduce collocations. These collocations can be very instructive as they do not simply describe, but give access to, for example naval manoeuvres and the maintenance of vessels. This is very much in line with Basnage's obvious intention to create some form of learner's dictionary as can be seen through his declining of verbs in his entries, with again use of examples and citations to illustrate use and also to disambiguate synonyms. There is much work to be done on this.

A final aspect that is quite clear in Basnage's knowledge network. His sources are rich and contemporary. They are used not for prestige but to illustrate and give greater encyclopaedic information. Furetière has already started this process of an encyclopaedic dictionary, but with Basnage we have a far greater wealth of sources and information.

Using digital humanities tools to link dictionary to source and source to dictionary will open up a window on the state of the art in seventeenth century arts, crafts and sciences. Digital humanities opens great perspectives for historical lexicography. Digital tools should not be an end in themselves, by applying appropriate tools and seeking complementarity of approaches we can build better tools while furthering knowledge. That is the aim if the Basnage project.

REFERENCES

Bots, Hans (1976), *Henri Basnage de Beauval en de Histoire des Ouvrages des savans 1687-1709*. 2 vol. Amsterdam, Holland University Press.

Bots, Hans, et Lenie Van Lieshout (1984), *Contribution à la Connaissance des Réseaux d'information au début du XVIIIe Siècle: Henri Basnage de Beauval et sa correspondance à propos de l'« Histoire des Ouvrages des Savans » (1687-1709). Lettres & Index.* Amsterdam & Maarssen, Holland University Press.

Glaser, Barney G., et Anselm L. Strauss (1967), *The Discovery of Grounded Theory: Strategies for Qualitative Research.* Chicago, Aldine de Gruyter.

Mercier-Faivre, Anne-Marie, et Reynaud (1999), éd. *Dictionnaire des journalistes*. Voltaire Foundation.

Quémada, Bernard (1967), *Les Dictionnaires du Français Moderne 1539 - 1863*. Paris, Didier.

Rey, Alain (2006), *Antoine Furetière : Un précurseur des Lumières sous Louis XIV*. Paris, Fayard.

Williams, Geoffrey (2017), *Architecture and a late 17th century dictionary: Terminology in the Basnage de Beauval 1701 edition of the Dictionnaire Universel*. in *Past in present: The language of heritage*. Firenze, Firenze University Press.

Williams, Geoffrey (2017), *Le temps des termes : les termes et la phraséologie dans les dictionnaires du 17 siècle*, in Cosimo De Giovanni (a cura di), *Fraseologia e paremilogia: Passato, presente e futuro*. Milano, FrancoAngeli.

Williams, Geoffrey, et Ioana Galleron (2016), *Digitizing the second edition of Furetière's Dictionnaire Universel: challenges of representing complex historical dictionary data using the TEI*. In Proceedings of the XVII EURALEX International Congress. Tbilisi, Georgia.

Dictionaries

Furetière, Antoine. *Dictionnaire universel, contenant généralement tous les mots françois tant vieux que modernes, & les termes des sciences et des arts. Tome 1 /,... par feu messire Antoine Furetière,... 2e édition revue, corrigée et augmentée par M. Basnage de Bauval*. La Haye et Rotterdam: Arnoud et Reinier Leers, 1701. ark:/12148/bpt6k57951269. Web.

---. *Dictionnaire Universel, contenant généralement tous les mots françois tant vieux que modernes et les termes des sciences et des arts*. La Haye et Rotterdam: N.p., 1690. Web.

Ménage, Gilles. *Les origines de la langue françoise*. Paris: Augustin Courbé, 1650. Print.

GEOFFREY WILLIAMS • is a corpus linguist, lexicographer and digital humanist. He is co-founder and President of the EvalHum Initiative, a European association seeking to promote the Social Sciences and Humanities through improved evaluation procedures and impact studies. As linguist and lexicographer, he is a former president of the European Association for Lexicography – EURALEX. He is a member of EADH and numerous other scholarly societies. He is currently director of the Department for Document Management in UBS, and a member of the Digital Humanities group of the Litt & Arts research unit of the Université Grenoble Alpes.

E-MAIL • williams@licorn-research.fr

SeGNALI