

Textual Analysis using XML: Understanding Ancient Textual Corpora

Henriette Roued

e-Science and Ancient Documents project (eSAD)

Centre for the Study of Ancient Documents, University of Oxford

henriette.roued@classics.ox.ac.uk

Abstract

This paper describes a project concerned with the highly granulated XML encoding of Vindolanda ink tablets using the EpiDoc Schema and performed as part of an ongoing project between the e-Research Centre, Centre for the Study of Ancient Documents and Engineering Science at University of Oxford. It aims to provide a large-scale corpus as a knowledge base for scholars reading ancient documents.

This paper will examine to what extent this rigorous encoding of the Vindolanda ink tablets can be used to create a reusable word and character corpus for a networked e-Science system and other e-Science applications.

Furthermore, it will suggest other benefits of XML encoding of ancient documents, such as enabling textual analysis of uncertainty and supplied characters in transcriptions, or the study of grammatical patterns and subject division (e.g. people, dates, military terms) of the tablets.

1. Introduction

The idea of the EpiDoc guidelines (Epigraphic Documents in TEI XML, <http://epidoc.sourceforge.net/>) have been around since 1999, when Tom Elliot began encoding epigraphic material using TEI (Text Encoding Initiative, <http://www.tei-c.org/>) and found that a more specific set of guidelines for epigraphic material was needed.

The development of such a set of new guidelines, first drafted in 2001, have since been promoted and developed by the Inscriptions of Aphrodisias project (IAph2007, <http://insaph.kcl.ac.uk/iaph2007/>), who were assessing new technologies to aid the publication of their inscriptions [1].

At the same time, the Centre for the Study of Ancient Documents (CSAD: <http://www.csad.ox.ac.uk/>)

began development of the Vindolanda Tablets Online website (<http://vindolanda.csad.ox.ac.uk/>) utilizing EpiDoc XML encoding as a way of publishing the Vindolanda publications (Tab.Vindol.I [2] and Tab.Vindol.II [3]) on the web.

The motivation of the epigraphic community to develop the EpiDoc guidelines and to use XML encoding for ancient documents was to create an “*on-line, free and unrestricted database of all surviving Greek and Latin epigraphical texts produced down to the end of Antiquity.*” [1].

Thus, digitization projects of inscription corpora are usually undertaken with the aim of publishing material online, which would otherwise only have been published in printed format and at considerable cost. In some cases (e.g. Vindolanda Tablets Online), the digitization projects complement a printed corpus and typically aims to provide additional or enhanced capabilities, such as interactive search facilities [7].

It is generally agreed that digitizing epigraphic documents using EpiDoc XML has numerous benefits for the publication of epigraphic material and enables more rapid and interactive search facilities. Nevertheless, the field of e-Science could benefit from further research into the added advantages of XML encoding of ancient document corpora. This paper illustrates the benefits of a more rigorous XML encoding of such documents, using as an example the Vindolanda ink tablets as a part the e-Science and Ancient Documents project (eSAD: <http://esad.classics.ox.ac.uk>).

The eSAD project is an ongoing project between the e-Research Centre, Centre for the Study of Ancient Documents and Engineering Science at University of Oxford. It aims to develop both advanced e-Science tools to aid the interpretation of damaged texts and novel image analysis algorithms for application to digitized images of ancient documents [9].

Our work aims at the development of an e-Science application using decision support principles to aid readers of ancient documents [8]. The encoded XML

will add value to this Interpretation Support System as a knowledge base of Latin words.

The encoding began as an attempt to add the Vindolanda tablets publication III [4] to the Vindolanda Tablets Online website in order to have the full corpus of published documents from Vindolanda on the web. We soon discovered that the first two volumes, currently online, were not encoded with great detail. This motivated us to encode the third volume with a greater granularity, which then lead us to identify a number of ideas for further uses of the encoded material other than online publishing of the corpus.

2. Encoding Leiden Mark-up

The Leiden Conventions are a type of semantic encoding, which consists of various brackets, under-dots and other markings relating to missing or broken characters, uncertainty, additions and corrections made by the editor of an ancient text. First agreed upon in 1931 at the 18th International Congress of Orientalists in Leiden [11], the conventions are now used by almost all scholars who transcribe ancient documents.

Since, the Leiden Conventions were only agreed upon in 1931, any edition published before this date tends to use a similar but slightly different system. Furthermore, editors still differ in their employment of the Leiden conventions.

An example of this is under-dots (i.e. $\dot{a}\dot{b}\dot{c}$), which some editors apply to indicate partially preserved characters (e.g. I Aph2007), while others use it to indicate doubtful characters not resolved by the editor, and yet others again use it to signify both (e.g. Tab.Vindol.III). Hunt [6] discusses another tendency amongst some of his colleagues in the 1930s, of underlining a broken character thought to be certain, which he does not agree with.

This example illustrates the primary advantage of encoding the editions in XML. If editors wish to differ between uncertain characters and broken characters they can encode them with different tags. They can then transform both tags into under-dots if they still wish to present both instances as such or they can decide to visualize one instance, underlined and the other under-dotted to distinguish between them.

However, the semantic mark-up, which editors have performed with the Leiden Conventions (and earlier), means that the step to XML encoding is not a substantial conceptual leap. It is simply a matter of understanding the structure of tags and getting used

to the idea of using the *supplied* tag instead of [pullus]:

```
<supplied reason="lost"> pullus </supplied>
```

2.1. Encoding the Vindolanda ink tablets

XML encoding using the EpiDoc guidelines was performed on the first two Vindolanda Tablets publications [2], [3] for the Vindolanda Tablets Online website in 2003. When we undertook to add the third publication [4] to this web site we soon discovered that the encoding of the earlier tablets was not performed to a great detail and that the current web site was not built to utilize the granularity of the encoding that followed.

The detail of which a project decides to encode text depends on the future use of the texts and also to some extent the technology applied by the project. Vindolanda Tablets Online did not need to encode the Leiden Conventions as they were adding the transcripts to a relational database. The new encoding also concerns words and terms in the transcription. This proved useful for an interactive search functionality for the new Vindolanda Tablets website, but also adds value to the eSAD project as a knowledge base. This is the motivation for the more rigorous round of encoding presented in this paper.

2.2. Analysis of uncertainty

Encoding instances of uncertainty, added characters and abbreviations enables us to extract these instances from their respective texts and analyze them. We can, for example, count how many characters in the text or texts are deemed to be uncertain. Similarly, we can look at the type of characters that are most likely to be supplied. These illustrate the many new possibilities for analyzing the reading of ancient documents.

Terras [10 p.71] performed an analysis of the Vindolanda ink tablets for her thesis, which showed that 5.3% of the characters and 11.7% of the words were supplied by the editors and 9.9% of the characters were either uncertain or broken (i.e. under-dotted). This was in essence a quick count of certain aspects of the text. Counts of this sort are such a simple procedure that it alone does not justify the effort of encoding the Vindolanda tablets. However, once they have been encoded, it is possible to perform similar counts of and extractions from the texts.

As an example, the supplied characters in the Vindolanda tablets could be extracted and would

allow an analysis of the type of characters most likely to be supplied or the placement of the supplied characters (i.e. beginning, middle or end of word or sentence).

3. Contextual Encoding

An investigation was undertaken in 2005 into improving the contextual encoding of the Vindolanda tablets [5]. It recommended encoding the instances in the transcription of words, people, dates and military terms. In other words, those items, which tended to be found in the indices in the back of the publications.

Where the primary encoding of the tablets (i.e. sections, header and Leiden Convention markings) could be encoded automatically using PHP scripts and pattern recognition the contextual encoding had to be performed manually.

In the case of the words, the index contained a series of lemmas with a reference to the place in each text where a word corresponding to this lemma could be found. In the case below, the word found in the transcription is *pulli* (i.e. chicken). From the index we know that the equivalent lemma is *pullus* and we also know that this is the first case of *pullus* in this text (Tablet 581).

```
<w lemma="pullus" n="1">pulli</w>
```

The encoding of this text enables us first, to extract the information that there are 15 instances of this lemma in this particular text. It also enables us to link to and from this particular instance by using XSLT transformations of the XML in various ways.

This encoding is however limited to the original indexing of the publication, which covers: calendar terms such as consuls and dates, personal names, geography, military and official terms, abbreviations, symbols and Latin words. During the mark-up we found that the indices were not always correct and we have had to give this additional attention. However, we have as far as possible stayed true to the indices so that the contextual encoding reflects the extent of these in the publication.

3.1. Vindolanda Tablets Online 2.0

As mentioned earlier, simply adding the newly encoded documents to the current website (Vindolanda Tablets Online) would not enable us to properly utilize this more detailed encoding. For this reason, we have begun development of a new web site, which will be launched in 2010. This uses AJAX

LiveSearch, JavaScript and PHP for a more interactive searching experience of the indices.

LiveSearch, best known from the Google search engine, gives the user feedback while typing in a search. If the user typed 'pu' LiveSearch would feedback the words found, which contained the characters 'pu' (figure 1). If the user then added 'l' it would then narrow the search down to words containing the character pattern 'pul' (figure 2).

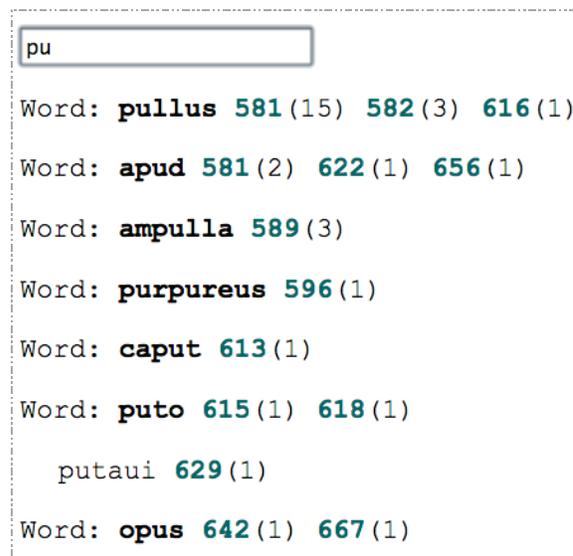


Figure 1: LiveSearch of characters 'pu'.

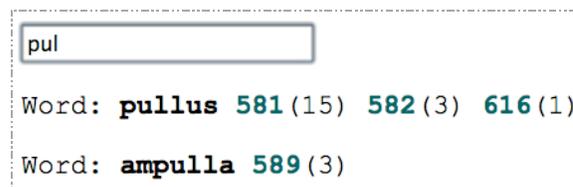


Figure 2: LiveSearch of characters 'pul'.

The search result is a list of the words, terms, names and dates, which contain the pattern searched for. Each word is followed by a list of the tablets that contain this word and the number (in brackets) of instances of this word in the tablet. Thus, there are 15 instances of *pullus* in tablet 581 (Figure 3), 3 instances in tablet 582 and only one instance in tablet 616.

The tablet numbers also serve as links to the first instance of the word in the HTML view of the tablet, highlighting the word (bold font) upon arrival (figure 4).



Figure 3: Tablet 582 with the word *pulli* highlighted.

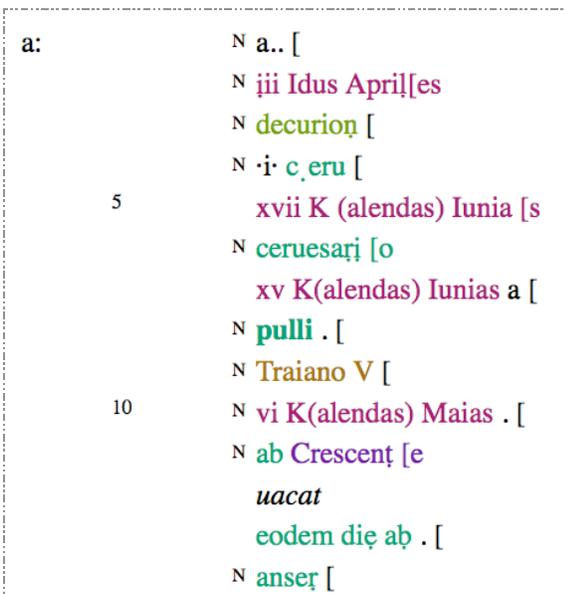


Figure 4: HTML view of tablet 581 with the word *pulli* highlighted (with bold).

Each encoded word in the tablet is displayed with a different color in the HTML transformation depending on the type (e.g. green for Latin words or purple for personal names). The words themselves also serve as links to a pop-up window sharing the information gathered from the encoding tag. For example, in the case of tablet 610, where we have the text *..rialia* the pop-up window informs us that this is thought to be *Flavius Cerialis*, who is known to be a *praefect*.

Using detailed XML encoding has also allowed us to encode each bibliographical instance. The abbreviation and reference lists were then encoded as separate XML documents, thus allowing links from the descriptions and commentaries to these lists.

3.2. Vindolanda tablets web service

The aim of this new encoding has been to ensure that all information about the text is contained in each XML document. For this reason, metadata is fully contained in the header and the reference number of photographs relating to each tablet has also been added to each XML documents.

With all this information contained in each document we needed to develop a method to extract only the information that is relevant. This was first done with a MySQL database and PHP scripts extracting data and depositing it in a series of tables. However, we soon found that each time the encoding changed for any reason we also had to perform an additional extraction.

As a result, we built a RESTful web service using the Zend Framework (<http://framework.zend.com/>) and PHP. The definition of a web service is in this case a service, which receives a URL with certain parameters and returns the answer as XML. This allows any e-Science project (including the authors) to reuse the encoded material.

The web service has a parameter called 'method' and a range of methods that can be called. It will return a list of the tablets available or the XML for a single tablet if the tablet ID is sent as a parameter. It is also possible to use the web service to access the knowledge base of words used in the LiveSearch above. There is a method that can be called for each index (e.g. Latin words or personal names) and there is a method to search for all words regardless of their category. These methods can return the entire list of words in each category but it is also possible to add a pattern forcing the method to return only words, which fit this pattern. The pattern used * as a wildcard character and square brackets for more than

one possible character in one characters place (Figure 5).

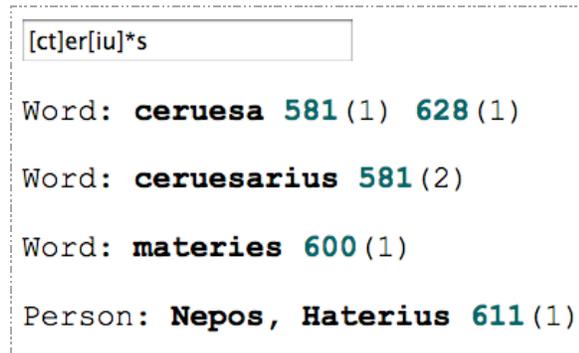


Figure 5: LiveSearch using the web service with the pattern.

This web service is presently being used in the LiveSearch of the new web site. However, we also hope to make it publicly available to any other e-Science project that may wish to search through the Vindolanda material or access any of the data in XML format.

The web service is also used in related work, which aims to develop an e-Science based Interpretation Support System for readers of ancient documents. In the prototype there is a word search, which takes the partially interpreted characters of a word and attaches them to the web service URL as a pattern, thus receiving suggestions for the word using the Vindolanda tablets as a knowledge base [8].

3.3. Textual analysis

The encoding can supply a list of all the words present (or recognizable) in the Vindolanda ink tablets. This has already been used in several different ways.

The next step could be to use these lists for textual analysis of words. It would be possible to find the most popular lemmas and most common use of them, and to analyze different spellings of each word. It could also be useful to study the placement of each word in the text and the category (e.g. letters, accounts or military documents) where certain words where more likely to be found.

4. Extracting character information

Information about the characters of the Vindolanda tablets could potentially be useful. However, we have not marked up each character in this current encoding project. Every encoding project involves deciding to which degree of granularity it is

judged necessary to encode. It was not in this case deemed to be worth the time and effort to encode on a character level.

Nevertheless, if one wanted to analyze the character use and frequency it would still be possible to extract each character from the transcription despite the lack of character level encoding.

5. Grammatical patterns

Terras [10 p.161] mentions how a grammatical labeling of each word in a corpus could be used to study grammatical patterns in a text. As an aid for reading ancient documents it could potentially predict the type of word needed next based on the type of word that came before.

It would be possible to go through each tablet once more and manually encode each word with their grammatical structure. However, this would be needlessly time consuming. A more efficient method would be to use the words lists through the web service to create a new XML document adding the grammatical structure to each word in the list.

However, this would need an expert in Latin to manually go through the list and determine the grammar of each word.

6. Subject search

As we noted earlier, the categorization of words in the Vindolanda ink tablets is limited by the publication indices. Therefore, if we wanted to be able to extract words on the subject of food we would have to encode this explicitly and manually into the XML documents. An alternative possibility would be to extract a list of all words found in the Vindolanda tablets and create a separate XML list of these words divided into different subjects. If the list included the lemma or other reference to the word and the tablet number it would always be possible to link this word back to the tablets.

However, the Vindolanda XML is already encoded subject-wise to a certain degree. Special military and official terms have been noted, as have dates and people. The tablets are also divided into larger categories (i.e. military texts, accounts, letters and minor texts).

The current Vindolanda Tablets Online website also contains a subject division (e.g. animals, health, weather). This has been added to the database, which runs this web site and is only put in place for tablets from publications I and II. However, someone with a good knowledge of the tablets could divide

publication III like this and these subjects could be added to the new encoding as described above.

7. Conclusion

By encoding the Vindolanda tablets with a high degree of granularity we are able to extract more information from the tablets. We are still able to transform the tablets into HTML and use this for online publishing. However the encoding of the Leiden Convention markings allows us to present the transcriptions in the users preferred way and not, as now, using purely the editors preferred markings.

The contextual encoding of words, personal names and dates gives us the opportunity to extract this information by means of a web service in such a way that it can be used simultaneously in several different applications within the e-Science community. The new Vindolanda tablets web site can use the word lists for a more interactive search functionality, while we can use the same web service other aspects of the project. Furthermore, anyone will be able to use the web service to search the Vindolanda tablets through their own e-Science application and could analyze the words or rearrange them and add new information by creating new XML documents.

8. Acknowledgements

Many thanks to Dr. C. Crowther from the Centre for the Study of Ancient Documents for his support with the encoding and MA students J. Gillespie and J. Hiller for their work on the manual contextual encoding.

Many of the ideas in this paper build on the work of Dr M. Terras. The author wishes to thank her and the other members of the e-Science and Ancient Documents group: Prof. A. K. Bowman, Prof. M. Brady and Dr S. M. Tarte.

9. References

[1] *EpiDoc: Epigraphic Documents in TEI XML*, <http://epidoc.sourceforge.net/>, Last checked: 04.08.09

[2] A.K. Bowman and J.D. Thomas, *Vindolanda: The Latin Writing Tablets*, Society for Promotion of Roman Studies, London, 1983.

[3] A.K. Bowman and J.D. Thomas, *The Vindolanda writing-tablets: (Tabulae Vindolandenses II)*, British Museum Press, London, 1994.

[4] A.K. Bowman and J.D. Thomas, *The Vindolanda writing-tablets (Tabulae Vindolandenses III)*, British Museum Press, London, 2003.

[5] D. Hippiusley, *Encoding the Vindolanda Tablets: An Investigation in Contextual Encoding using XML and the EpiDoc Standards*, Masters in Electronic Communication and Publishing, University College London, 2005.

[6] A.S. Hunt, "A note on the transliteration of papyri", *Chronique d'Égypte*, 1932, pp. 272-274.

[7] *Electronic Textual Editing: Epigraphy*, Last checked: 04.08.09

[8] H. Roued Olsen, S. Tarte, M. Terras, M. Brady and A.K. Bowman, "Towards an Interpretation Support System for Reading Ancient Documents", *Digital Humanities Conference*, Jun 2009,

[9] S. Tarte, M. Brady, H. Roued Olsen, M. Terras and A.K. Bowman, "Image acquisition & analysis to enhance the legibility of ancient texts", *UK e-Science All Hands Meeting*, Sep 2008.

[10] M. Terras, *Image to Interpretation. An Intelligent System to Aid Historians in Reading the Vindolanda Texts*, Oxford University Press, 2006.

[11] B.A. van Groningen, "Projet d'unification des systemes de signes critiques", *Chronique d'Égypte*, 1932, pp. 262-269.