Towards a decision support system for reading ancient documents

Henriette Roued-Cunliffe
Centre for the Study of Ancient Documents,
University of Oxford, UK

Abstract

Constructing readings of damaged and abraded ancient documents is a difficult, complex, and a time-consuming task. It frequently involves reference to a variety of linguistic and archaeological datasets and the integration of previous knowledge of similar documentary material. Due to the involved and lengthy reading process, it is often difficult to record and recall how the final interpretation of the document was reached and which competing hypotheses were presented, adopted, or discarded in the process of reading. This article discusses the development of the application called DUGA, which uses Decision Support System (DSS) technology to aid the day-to-day reading of damaged documents. Such an application will facilitate the process of transcribing texts by providing a framework in which scholars can record, track, and trace their progress. DUGA will include a word search facility of external resources such as the Vindolanda ink tablets through the knowledge base Web Service called APPELLO. This functionality will support the scholars through their reading process by suggesting words, which may confirm current interpretations or inspire new ones. Furthermore, DUGA will allow continuity between working sessions, and the complete documentation of the reading process, that has hitherto been implicit in published editions.

Correspondence:
Henriette Roued-Cunliffe, University of Oxford, UK.
E-mail:
henriette.roued@classics.ox.ac.uk

1 Introduction

Reading ancient documents is a skill developed by epigraphers, palaeographers, and papyrologists, who will be referred to as scholars in this article. They are experts in synthesising experience and scholarly resources into plausible interpretations of the meaning of texts that are often several thousand years old. This interpretation is commonly followed by the publication of editions in the form of transcripts with translations and commentaries. The texts that are being read could be written on a variety of materials such as papyrus, wood, wax, potsherds, leather, stone, or metal, some of which are more degradable than others. This results in a number of legibility issues ranging from broken pieces of text through to palimpsests.[1] Therefore, when faced with damaged and highly illegible documents such as the Vindolanda stylus tablets, the reading process becomes even more complicated (Bowman *et al.*, 1997). Furthermore, scholars reading ancient documents, often do so whilst undertaking other scholarly work. Consequently, they are rarely able to finish their interpretation in one session and must instead return to it again and again, hoping to be able to pick up their previous train of thought.

This research is concerned with the interpretation of an ancient document. It recognizes that a final interpretation of this document is made up of many minor interpretations. In this article, these

Literary and Linguistic Computing, Vol. 25, No. 4, 2010.
doi:10.1093/llc/fqq020 Advance Access published on 21 October 2010

will be known as percepts to avoid confusing them with the overall interpretation of the text.

This research examines how a typical reading is performed, the issues arising from this, and how Decision Support technology would be able to alleviate these issues. Vindolanda ink tablet 159 is used as an example of a damaged piece of text, which has been read and published as an *editio princeps* (i.e. the first edition of a text) (Bowman and Thomas, 1994, p. 102). This article will examine the process of reading this text. Tablet 159 will then be revisited in the Decision Support System (DSS) prototype, demonstrating how the DSS can aid in the transcription and reading of ancient texts. The article will subsequently present ideas behind using Decision Support technologies in the interpretation process. DSSs are very popular within medical engineering but the question is whether the same ideas can be transferred, without obstacles, to the work of classical scholars. This is one of the lines of enquiry of this research.

This article will demonstrate how to tackle the element of uncertainty that is naturally present during an interpretation process. This uncertainty cannot be meaningfully quantified and therefore this research looks towards an evidence-based approach.

By building a DSS for the reading of ancient documents (which we have called DUGA[2]) this research aims to guide and support the scholar through their interpretation process. It is important that this should not be considered an expert system built to make the scholars' expertise redundant. Rather it should be considered a tool which will guide the scholars through the process of reading while performing the task that they find difficult. In this case these tasks would mainly be capturing complicated reasoning processes, searching huge datasets, accessing other scholars' knowledge, and enabling co-operation between scholars working on a single document.

This article will demonstrate a prototype that has been built as a proof of concept for using a DSS within this textual environment. The prototype is connected to the knowledge base Web Service (which we have called APPELLO), which has also been developed as part of this research. This development has used the Vindolanda ink tablet volumes II and III as a case study. APPELLO will be further developed to function as a knowledge base Web Service for any textual corpus using EpiDoc standard XML.[3] This will provide a valuable resource of word frequencies, which a scholar can call upon to suggest new words for existing character patterns or as evidence to help cement a current interpretation. This article will explain how APPELLO has been developed for the Vindolanda tablets and how it will be further developed to create a range of different knowledge bases all able to support the scholar's reading, depending on the nature of the material. If the scholar is reading an Ancient Greek text, a knowledge base from the Vindolanda tablets would be of little use. However, a list of word frequencies from Greek lexical resources may prove very effective. This article will furthermore discuss the use of external corpora, how accessible they are, and how they could be made more accessible. Finally, the article will present the next steps in the development of DUGA. It will look at the lessons that have been learnt from the development of the prototype and how the research can proceed.

This research is conducted within the scope of the e-Science and Ancient Documents project (eSAD: http://esad.classics.ox.ac.uk/) which aims to use computing technologies to aid experts in reading ancient documents. The project is involved in developing image processing algorithms to provide illumination correction and wood grain removal. The aim of this is to make writing more legible on documents that may otherwise be difficult or impossible to read (Tarte *et al.*, 2008). Furthermore, the project is researching software applications such as character recognition and DSSs, which would prove valuable for the process of reading ancient documents (Tarte *et al.*, 2008).

2 Reading Ancient Documents

When scholars embark upon the task of reading an ancient document they may do so either with the aim of publishing the text for the first time (*editio princeps)* or for the purpose of adding their own

Table 1 The semantic mark-up conventions used by the Vindolanda Publications (Bowman and Thomas, 1994, p. 19)

i, ii	Designate separate columns of text following the original layout.
. . .	Indicates that the text is broken or incomplete at the top or bottom.
m^1, m^2	Distinguish different hands in the text.
[ ]	Indicates a lacuna in the text.
[*c.4*]	Estimate of the number of letters missing in a lacuna.
uacat	A space left by the scribe on the tablet.
[[abc]]	Letters crossed out or erased by the scribe.
'abc'	Letters or words added by the scribe above the line.
<abc>	Letters erroneously omitted by the scribe.
{abc}	Superfluous letters written by the scribe.
ạ□ç	Doubtful or partially preserved letters.
. . .. . .	Subscript dots represents traces of letters visible on the tablet, which have been left unread.
( )	Expansion or resolution of an abbreviation or symbol, e.g. *praef(ecto), (centurio)*.

interpretation of the text to the current editions available. An edition is the culmination of a long and cumbersome interpretation process and will focus around the 'pièce de résistance' (Youtie, 1963, p. 22), the transcription of the text. The act of reading an ancient document is in essence a process of copying down or, in other words, transcribing the text as seen by the reader (Turner, 1968). During this transcription process, most scholars will use a type of semantic mark-up to indicate where they find unclear letters or damaged text and where they have added or omitted letters. Modern scholars have long used semantic encoding with editorial symbols to mark-up their transcriptions. In 1931, at the 18th International Congress of Orientalists, van Groningen (1932) demonstrated the need for some form of standardisation in this editorial encoding. He consequently proposed a set of conventions which became known as the Leiden conventions. Thereafter, most scholars basically adhered to these conventions for semantic mark-up, clarifying and making explicit departures or variations where necessary, as in the editions of the Vindolanda tablets (see Table 1).

Vindolanda ink tablet 159 is an example of how scholars would read a document while using semantic mark-up to indicate their perception of it (see Fig. 1). On this tablet, seven lines of writing have been identified. At the end of the text there is a line of dots separated by spaces (. . . . . .). This indicates that the editors believe that more lines of text have been there, which have broken off. To the right of each line, there is a square bracket facing outward of the text ([). These symbols indicate where the editors judged that these lines of text were incomplete. On the third line, a couple of letters have been read, although the editors still believe these letters to be unclear and have, therefore, indicated this by adding a dot under the letter. Between these letters the editors have added single dots, which represent traces of letters that have not been read.

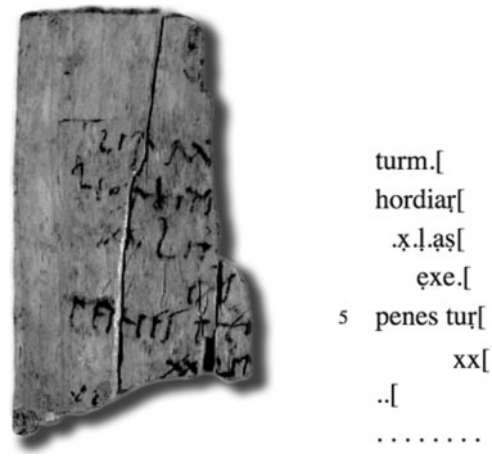

Fig. 1 Image of the front of the Vindolanda ink tablet 159 (left) and the transcription of the same text from the Vindolanda writing-tablets II (Bowman and Thomas, 1994, 102). Tablet is copyright of CSAD and The Trustees of the British Museum

The transcription of Tablet 159 provides an example of the uncertainty which is inherent in the process of reading ancient documents. Many factors

play a role in this uncertainty. The physical condition of the tablet with the broken edges, the cracks running down it, and the faded ink will have a huge impact on the legibility of the document. In this case the editors believe this text to be a fragment of the top left hand corner of a larger text or, in other words, the beginning of a document. This and other general information about this document, such as the size and the subject matter, has been published in a description,[4] which is another important part of the finished edition.

The editors will, if possible, follow up the transcription with a translation, but this may not yield connected sense as in the case of tablet 159. The commentary is another valuable component of the edition, which often follows the transcription line by line and gives further explanation for the lines of enquiry encountered during the reading of the text. This is where the editors will express doubts about certain characters and conviction about others.

'2. *hordiar[*: the letter at the break can comfortably be read as *r* but not as *t*; *hordiat[or(es)* (cf. *RMR* 47.ii.5) is therefore excluded and the most likely restoration is *hordiar[ia* or a cognate. This term occurs in *Doc.Masada* 722.6 and 13 where it means ''barley-money'' deducted from the *stipendium* of cavalrymen (see note *ad loc.*).' (Bowman and Thomas, 1994, p. 102)

This note relates to line 2 of tablet 159 and it explains the reasoning behind the reading of the last letter as *r*. The editors propose two different words, which are equally possible based on the legibility of the first six letters (i.e. *hordiator(es)* and *hordiaria*). However, the editors are convinced that the seventh letter cannot be read as *t* but on the other hand a reading of *r* is possible. This settles the case and the reading for these editors must be *hordiar[ia*. It is very important that this evaluation of the evidence for and against the different readings is conducted. However, the commentary only presents the conclusions of this exercise. It would be a great aid, both for editors as they go through this process, and also for future editors of the same text, if it were possible to present this evaluation for each character and word in a structured format. This article will discuss below the development of a DSS for this purpose.

This research builds upon the work of Dr Melissa Terras who began her research into constructing a computer system to aid papyrologists in reading stylus tablets by examining how a reader of ancient documents works (Terras, 2006). Terras used different knowledge elicitation techniques, such as Think Aloud Protocol, to collate explicit and quantitative data on three scholars' approach to the reading of a stylus tablet from Vindolanda. The conclusion drawn from this experience was mainly that reading ancient documents is not a process of transcribing the document letter-by-letter and line-by-line. Instead, it is a cyclic process of identifying visual features and building up evidence for and against continually developing hypotheses about characters, words and phrases. This is then checked against other information in an ongoing process until the editors are happy with the final interpretation (Terras, 2006).

Dr Ségolène Tarte has, within the framework of the eSAD project, been working on similar research. Tarte has examined the work of three scholars re-reading a Frisian stylus tablet (Bowman *et al.*, 2009) (Tarte, 2010). Through analysis of video recordings of the three scholars' reading process, she has concluded that each scholar uses his own combination of two different approaches to the task at hand. She also believes that these approaches are highly interconnected with the scholars' personal skills (de la Flor *et al.*, 2010).

Tarte has identified what she calls the palaeographical/kinaesthetic[5] approach. This makes use of tracing the characters and applying artistic skill to a reconstruction of the movements of the scribe and, through this, gain understanding of the text. The second approach which she calls the philological/cruciverbalistic[6] approach on the other hand sees the reading as analogous to a crossword-puzzle-solving task. A scholar using the latter approach will often begin by establishing the letters that are legible and use these as a foundation for a subsequent hypothesis (Tarte, 2010). The research presented in this article has up to this point focused on the cruciverbalistic approach. Nonetheless, it would be interesting to examine how the kinaesthetic approach could be incorporated into the DSS.

3 Decision Support for Interpretations

The rationale behind developing a DSS is to be able to support the decision-making process of scholars by recording not only the final decisions but also the evidence on which these decisions were based. The DSS will consequently act as a memory bank for the decision makers and thus save them time, reducing the risk of a series of incompatible decisions (Austin *et al.*, 2007). The inspiration for using DSS to aid the reading of ancient documents was research conducted at the Department of Engineering Science in Oxford into the development of an application called MDTSuite. MDTSuite has been developed to provide decision support for multidisciplinary decision-making meetings in a medical environment.

The MDTSuite research used the management of the multi-speciality disease colorectal cancer as a case study. MDTSuite was able to aid the multidisciplinary team in their decision-making process by presenting, for each patient, a full decision history. The multidisciplinary team could then evaluate this and use it to make new decisions on the patient's further treatment. Furthermore, the MDTSuite could highlight the absence of certain data and demonstrate whether and how this would affect the decision (Austin, 2008).

The present research will not attempt to translate the MDTSuite development directly into a DSS for the reading of ancient documents. There is too vast a difference between the work process of a classical scholar and that of a medical professional. The MDTSuite uses a rulebase which combines national and international guidelines with local clinical knowledge. This could, for example, contain a rule stating that if the team wishes to send the patient on to the next stage of treatment, the patient must have come through the previous stage of treatment with a certain result. MDTSuite would then highlight whether this or any other rules were being broken. Where the medical professional can make use of this rulebase to guide them to make informed decisions, this would not work for classical scholars as they do not operate with clear guidelines on how to read and publish ancient documents. The effort involved in making a rulebase for the interpretation process of ancient documents would be enormous. Not only would it involve collating a huge amount of so far unwritten rules, but there would also be the issues of applying these to a network of percepts on different levels ranging from letters, to words and paragraphs. Instead classical scholars use previously interpreted documents from the same period to guide their understanding of the text they are working on. Therefore, this research has looked into another approach of using a set of knowledge bases, such as word lists and frequencies from relevant corpora. These could then be used to suggest different interpretations of words and letters as the reading progressed.

DSSs have been used since the early 1970s and are now considered an accepted tool in corporate areas from production to human resources, and in other areas such as urban planning, military, and government (Eom *et al.*, 1998). However, to the best of our knowledge, this is the first attempt to use DSSs in the humanities.

3.1 Stages of interpretation

Interpretation of textual material happens on many different levels. Terras identified ten such levels of reading in her research (see Table 2).

Terras (2005) defined these levels in order to encode and examine the reading process of the three scholars as mentioned above. This analysis concluded that the scholars operated mainly on

Table 2 Terras' original encoding scheme for identifying levels of reading

Reading level	Thematic subject
8	Meaning or sense of document as a whole
7	Meaning or sense of a group or phrase or words
6	Meaning or sense of word
5	Discussion of grammer
4	Identification of possible word or morphemic unit
3	Identification of sequence of characters
2	Identification of possible characters
1	Discussion of features of characters
0	Discussion of physical attributes of the document
−1	Archaelogical or historical context

Note that the first level is presented as −1 to explicitly indicate that at this level the scholar is referring to a resource other than the document being read. (Terras, 2005, p. 8).

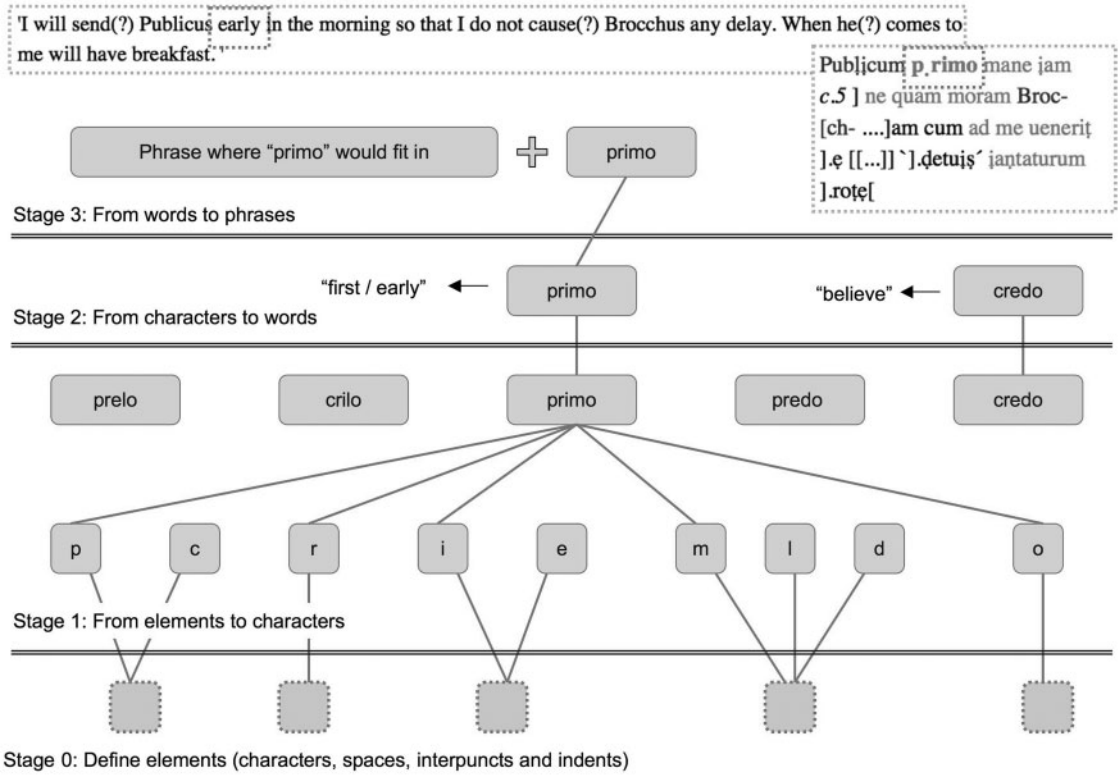

Fig. 2 This model presents the different stages between which the scholar can move in the network of percepts. Stage 0 is the move from having an un-interpreted image to identifying the image elements. Stage 1 is the jump from identifying these elements to reading the actual letters of the character element. Stage 2 moves from the identification of letters in the character elements to the interpretation of words or character sets. Finally, stage 3 represents the jump from single words to an interpretation of a phrase or a sentence.

three of the reading levels. There were features of characters (level 1), identification of characters (level 2), and identification of possible words (level 4). These reading levels reflect the domain covered by the DSS. However, in order to explain the idea behind the network of percepts and the structure of the DSS, this research is using a different system to understand the interpretation process (using the term 'stages' instead of 'levels' to avoid confusion).

Stage 0 (See Fig. 2) is the step from having an un-interpreted image to identifying on the image elements such as characters, spaces, interpuncts,[7] or indents. Stage 1 is the jump from identifying these elements to reading the actual letters of the character element (the latter is the equivalent of Terras' level 2). Stage 2 moves from the identification of letters in the character elements to the interpretation of words or character sets (the latter can be identified as Terras' level 4). Finally stage 3 represents the jump from single words to an interpretation of a phrase or a sentence.

The idea behind the stages is that an interpretation consists of a network of percepts ranging from low level percepts, such as 'these three line fragments are a character' (stage 0), to higher level percepts, such as 'these five letters can make up the word *primo*' (stage 2). However, there is a large amount of circularity inherent in this process. The scholars may begin by reading a couple of characters

of which they are quite certain (stage 1) and from there decide on the word '*primo*' based on the knowledge that this word would fit into the phrase (stage 3). They could then use this interpretation to decide on the other characters of the word. DUGA would have to enable the scholars to operate within these different stages while building and visualising the network of percepts which will allow the scholars to keep track of their interpretative process.

3.2 Reading with uncertainty

The process of reading ancient documents comes with an inherent uncertainty due to the nature of interpretation. Classical scholars do not traditionally justify their interpretations, for example, by claiming to be 85% sure of a character or word. Therefore, there would be no point in trying to quantify their perceptions by expressing a percentage of certainty for a given percept. Instead, this research is working on a model of evidence for (+) and against (−) each percept. The network of percepts would furthermore enable each percept to act as evidence for or against other percepts (see Fig. 3). The types of evidence can be anything from physical characteristics of the document to character recognition software, word search of a knowledge base, or the scholar's own judgements. While character recognition and word searches can provide valuable suggestions towards each percept, it is important to note that they are not meant to stand alone without the support of the scholar's own experience and knowledge. This research looks towards capturing this scholarly expertise and adding it to the system as pieces of evidence under the heading 'Scholarly Judgements'. This article will now explain how this idea has been used within the prototype system.

4 Developing a DSS

This section will first present the prototype that has been developed as a proof of concept of how a DSS can be used to aid scholars through a complicated reading process and overcome the issue of reasoning under uncertainty.

A part of this concept is to enable the use of knowledge bases to support the interpretation by suggesting possible words throughout the reading. This research has developed a knowledge base of words using the Vindolanda ink tablets as a case study. This has then been incorporated into the prototype and will be developed further for use in DUGA.

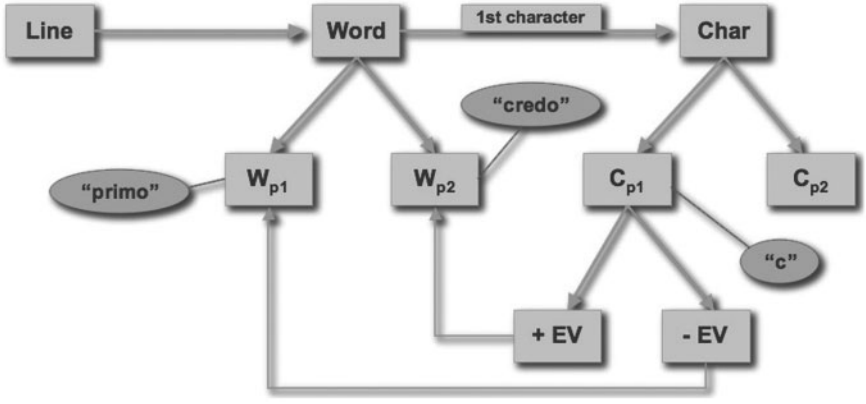

Fig. 3 Model of the use of evidence-based percepts, where the two different percepts for a single word can each act as evidence for and against the percept of a character. In this case, the perception of the first character of the word is 'c'. The perception that the word is 'credo' is evidence for the 'c' percept, while the perception that the word is 'primo' is evidence against.

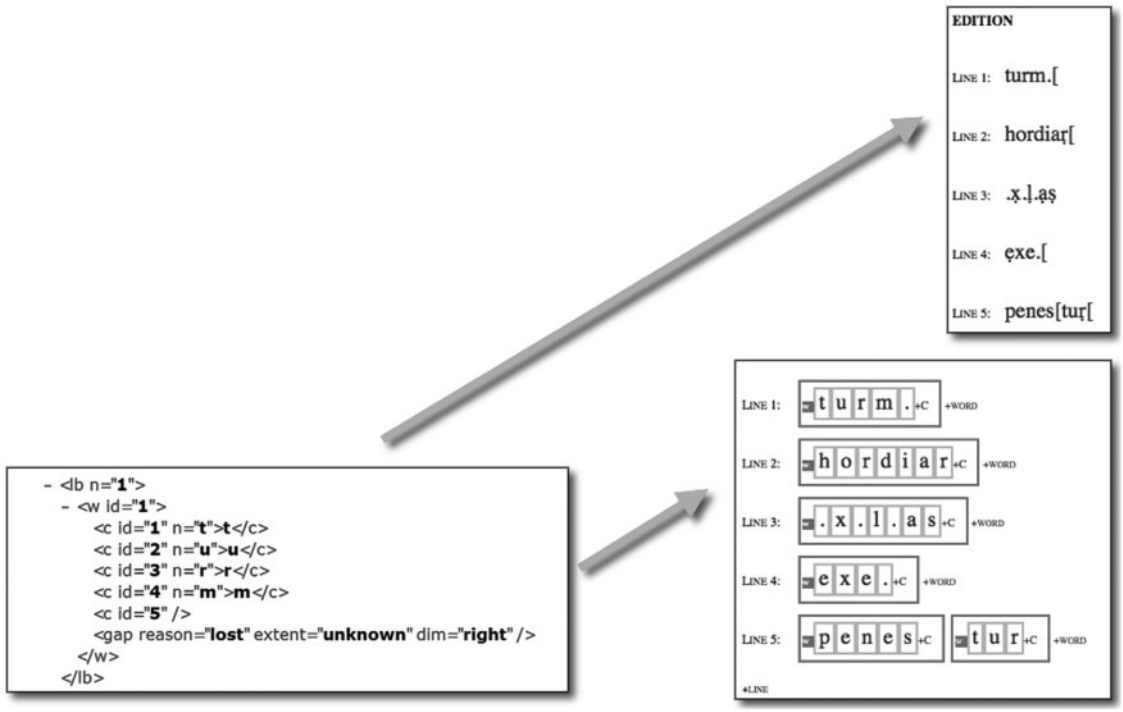

Fig. 4 Example of how the network of percepts for Vindolanda ink tablet 159 is stored as XML (left). This figure also demonstrates how this XML is being transformed into two different views (right). The top view presents the current interpretation as a Vindolanda style transcript. Here the XML is replaced by Leiden conventions (see section 2). The bottom view is the so-called box view, which visualizes each character and word as boxes.

The article will also examine how other corpora can be integrated into DUGA as knowledge bases. This would be particularly important if DUGA is to be used for interpretation of documents that are not written in Latin and not from the same context as the Vindolanda ink tablets.

Finally this section will present ideas and issues for the forthcoming development of DUGA.

4.1 The prototype

This DSS prototype was built as a way of demonstrating the usability of the concept of a network of percepts and to develop a suitable system for managing the process of reasoning under uncertainty. The prototype is using Vindolanda ink tablet 159 in order to visualize, through the example of the last letter of the word on line 2 (i.e. *hordiar*), how the evidence for and against each percept would work.

For the purpose of the prototype, the ongoing interpretation is stored in an XML document. Figure 4 presents an example of this XML encoding for line one of tablet 159. In this example, the empty character tag with the id 5 (i.e. <*c id*='5'/>) represents stage 0 from the stages of interpretation above. Here the presence of a character has been perceived but no identification of the letter in this character space is available at this point. The other characters, in this first line, have reached stage 1. At this stage, actual letters have been read in the spaces that were previously identified as containing characters. Stage 2 is visualized through the word tag (i.e. <*w*>) that encloses the character tags. Finally the line tag (i.e. <*lb*>) represents stage 3. For the purpose of this prototype, storing and updating the network of percepts as an XML document has been sufficient. However, for the purpose of DUGA, the research would need to include storage

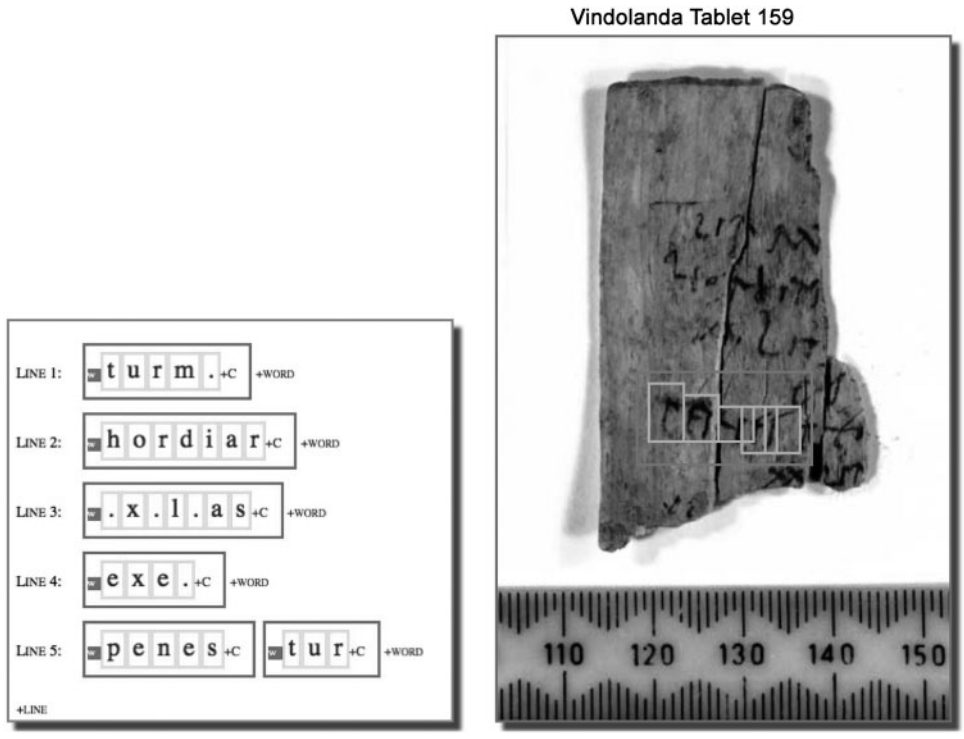

Fig. 5 Visualisation of how an annotation viewer (right) would enable the scholar to add percepts of lines, words, and characters to the interpretation through the drawing of boxes on the image instead of using the box view (left). Tablet is copyright of CSAD and The Trustees of the British Museum.

of not only the current percepts but also the evidence for and against each percept.

The prototype is divided into a set of views, the use of which will be demonstrated below. The transcript and box view are populated on-the-fly by two different XSLT transformations of a XML document (see Fig. 4).

The development of the prototype has enabled this research to visualize how a scholar could use a DSS application. The first step would be to view the image of the document that is being read. The scholar would begin by interpreting the image through the identification of paragraphs, lines, words, and characters (see Fig. 5). For this task, the research would include annotation software such as AXE developed by the TILE project (Porter *et al.*, 2009) or the annotation viewer developed by the BVREH project (Bowman *et al.*, 2010). This would give DUGA the information needed to make a start on building the box view with line, word, and character boxes.

The box view was first intended to be a place in which the scholars could add lines, words, and characters. However, using annotation software in a separate viewer would be a more flexible approach. In the prototype, the box view only enables the scholars to add lines at the end of a paragraph, words at the end of a line, and characters at the end of a word. Using annotation software would give the opportunity to draw character, word, or line boxes anywhere on the image at any time. This would then be translated into XML, which would then be transformed into the transcription and box view as above.

The next step for the scholars would be to use the box view to access and edit each word or character. The box view is designed especially with the cruciverbalistic approach (see Section 2) in mind.

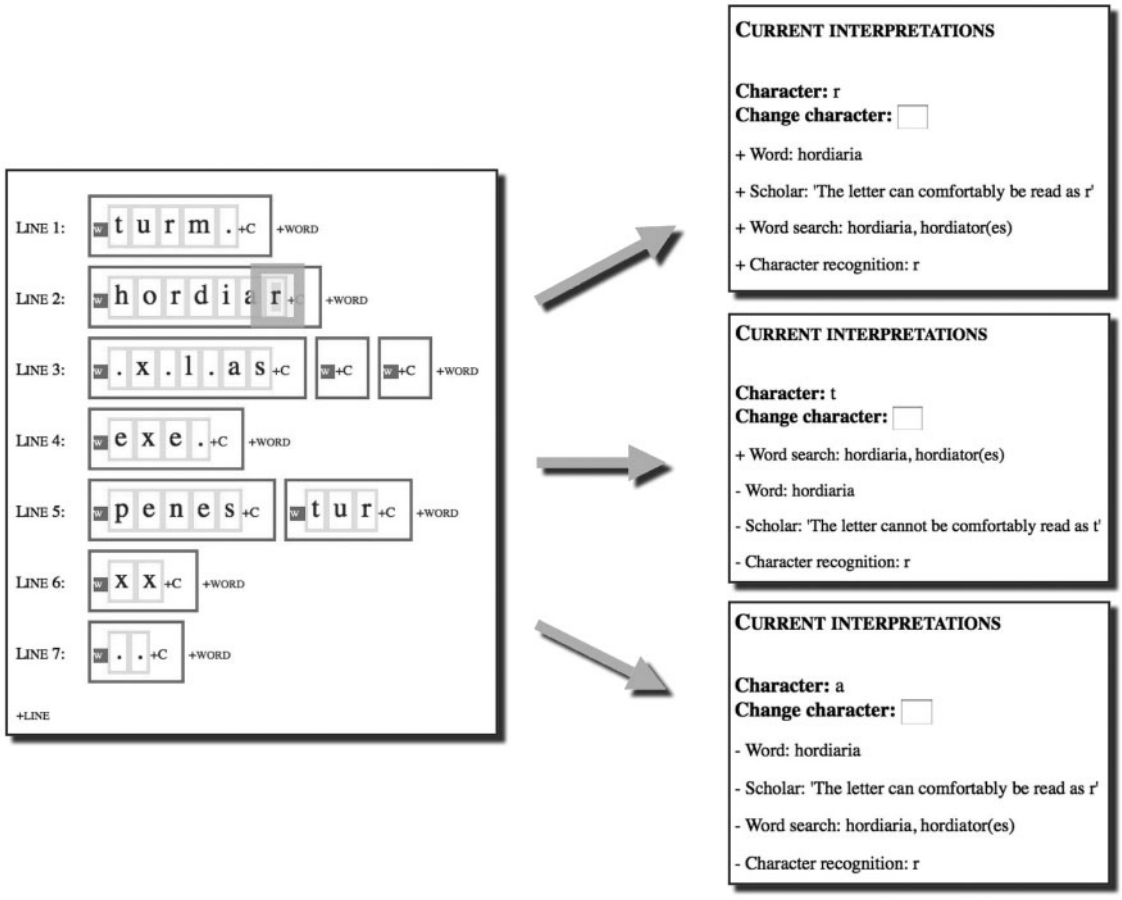

Fig. 6 The box view (left) enables the scholar to change their percept of the letters and words by clicking on this. This action will present the current percepts in the edit view (right) as well as the evidence for and against each. In the example above, there are three different edit views, one for each percept of the last character in the word on line 2 (i.e. *r*, *t*, and *a*).

It enables the scholars to begin by adding percepts of legible characters and at the same time use these characters to move through stage 2 to the word level. The scholars will be able to move back and forth between the interpretation of words and characters through the box view. It is also an important part of the design that they are able to move back and forth between the box view and the annotation view should they need to add or edit character, word, or line boxes. Enabling a circular interpretation process is a vital design feature for this DSS.

When scholars choose to edit their percept of a character or word in the box view, they will be presented with any current percepts and evidence for and against these in the edit view (Fig. 6).

The prototype uses the last character in the word on line 2 of tablet 159 as an example of how the evidence for and against may change depending on the percept. This is an example of a hypothetical work in progress.[8] In this example, the scholars have already gathered or created four pieces of evidence for the percept *r*. Before the scholars turned to the interpretation of the last letter in the word, they had already read the legible previous six letters as *hordia*. The letter combination *hordia* together

with a seventh letter, not determined yet, was then run through the word search (see Section 4.2). This returned the words *hordiaria* and *hordiator(es)*. The word search thus works as evidence for the seventh letter being *r* or *t*.

By using their scholarly experience and training, they found that the letter could comfortably be read as *r*. The strokes of the character have been run through character recognition software,[9] which has returned the result that the letter is indeed *r*. The scholars now return to the word level and decide that the current percept for this word is *hordiaria* (meaning barley-money). Back at the character level, this is now used as the fourth piece of evidence for the letter being *r*.

The next edit box in Fig. 6 is the evidence for and against the letter being *t*. When the word search returns the words *hordiaria* and *hordiator(es)* the scholars have to consider if the seventh letter could instead be *t*. However, the scholars are convinced that the letter cannot be comfortably read as *t*. This means that there is only one piece of evidence for the percept of the seventh letter being *t* and three pieces of evidence against it.

The last edit view in Fig. 6 represents all the other letters of the alphabet that the seventh letter could be interpreted as (in this case *a*). There is no evidence for this percept.

The evidence for and against each percept is not conclusive, nor is it quantitative. Even though the letter *r* includes most evidence for it and none against it, it is not a certainty that *r* is the correct reading of this character. It is important to remember that both the reading of *r* and the evidence for *r* are still interpretations made by the scholars. The word search and character recognition results should not be seen as conclusive evidence but as suggestions that may either confirm the scholars' current percept or inspire a new one. It is entirely up to the scholars to decide how they value each piece of evidence. If one piece of evidence is valued higher than any others, it may be enough to sway the interpretation even if it is the only evidence for a certain percept. Moreover, the scholars may come back to a new session of reading and find that *r*, which could be so comfortably read before, now looks more like *t*. The idea with a DSS, is to support this process by reminding the scholars of the evidence on which they based their percept and the consequences for the entire network of percepts if they change it. If, for example, the scholars changed the seventh letter to *t*, the system would highlight the fact that this percept is now incompatible with the percept of the word as *hordiaria*.

4.2 Knowledge base—Vindolanda case study

A word search through a knowledge base of relevant material is a valuable addition to DUGA. It can provide the scholars with suggestions for the word they are reading that will either confirm the percept or inspire a new interpretation of the word. In the above example, the prototype ran the pattern *hordia.[*[10] through the word search, which returned the words *hordiaria* and *hordiator(es)*. To run this word search, DUGA must be connected to a knowledge base of relevant words. In the example of reading tablet 159, a knowledge base of Latin words used in other Vindolanda ink tablets would be especially relevant. However, if the scholars were instead reading a Greek inscription, DUGA would need to be connected to a different knowledge base of Greek words. The idea is that the scholars, when working on a new text, can choose the knowledge bases they find relevant for their particular reading. However, there are not many knowledge bases of ancient document available that provides word search facilities. Therefore, this research has developed the knowledge base Web Service which we called APPELLO,[11] using the Vindolanda ink tablets as a case study.

The World Wide Web Consortium (W3C) defines a Web Service as 'a software system designed to support interoperable machine-to-machine interaction over a network' (Haas & Brown, 2004). In other words Web Services allow heterogeneous systems to communicate, solving the need for interoperability between these systems. Often XML will be used to format messages sent between systems because XML is a computer-readable text format.

The Web Service design for APPELLO does this by enabling a software application such as DUGA to

interact with the Vindolanda ink tablets dataset. More specifically, it enables any user or application to send an URL to a server. The server will then return the answer formatted in XML, depending on the parameters which were sent in the URL. APPELLO has been developed using RESTful Web Service architecture and the Zend Framework (http://framework.zend.com/) as this is currently one of the simplest and most lightweight methods of developing a Web Service to date.[12]

The first step in building APPELLO was to reformat the Vindolanda ink tablets dataset into EpiDoc standard XML. EpiDoc is a schema for XML encoding, developed specifically for epigraphic documents (http://epidoc.sourceforge.net/).

The tablets found at the Roman fort of Vindolanda near Hadrian's Wall contain about 750 published ink tablets covering a variety of themes. The editions for these documents have already been published in three volumes (Bowman and Thomas 1983, 1994, 2003). The second published volume is available as a HTML-based searchable website (http://vindolanda.csad.ox.ac.uk/). Apart from encoding volumes II and III as well-formed EpiDoc standard XML, this research has also employed a new approach called contextual encoding (Hippisley, 2005). This consists of encoding words, personal names, geographical place names, calendar references, and abbreviations. For example, any instance of the word *pulli* (Latin for *chicken*) in a document will be encoded <w lemma='*pullus*'>*pulli*</w>. The lemma is the root of the word which enables APPELLO to find different inflections and variations of the same word. This particular encoding provides us with the information that the word *pulli* has the lemma *pullus* under which we can index this instance of the word. This encoding can then be used to extract and analyse all the words from each document. This functionality enables the APPELLO Web Service to return a collective list of words from the Vindolanda ink tablets.

APPELLO can also generate lists based on a pattern and this has been used in the prototype for the word search of the pattern *hordia.[*. APPELLO will return all words that begin with the letters *hordia* and have a seventh character.

4.3 Knowledge bases through web services

APPELLO has been developed using the Vindolanda ink tablets as a case study. As a part of this development, the Vindolanda ink tablets have been transformed into well-formed EpiDoc standard XML. Therefore, APPELLO would in theory work as a Web Service for any dataset that was encoded as EpiDoc standard XML with added contextual encoding.

This would be a great advantage to DUGA since, as mentioned earlier, there are not many linguistic datasets available that include a Web Service which DUGA could use for word search.

The next step of this research is to test whether APPELLO can be comfortably integrated within other linguistic dataset projects. This research plans to use the Monumenta Asiae Minoris Antiqua project (MAMA: http://mama.csad.ox.ac.uk/) as a test subject for this. The MAMA project is involved with the reading and publishing of 600 unpublished inscriptions and other ancient monuments. They have been recorded by Sir William Calder and Dr Michael Ballance in the course of annual expeditions to Asia Minor between 1954–57. MAMA uses EpiDoc standard XML together with contextual encoding as a means of publishing these texts on the Internet as well as in book form. Integrating APPELLO with this dataset has advantages for the MAMA project as well. APPELLO can aid the project by generating an on-the-fly index of words and their lemmas based on the contextual encoding. This could save the project time that would otherwise have been spent building indices for an ever-growing dataset.

This article does not want to give the impression that there are no other linguistic datasets available on the Internet, as this is not the case. The field of classics has been very active at digitising linguistic material and in the later years, at making them available online (Bagnall, 1997). Among the larger resources are the Perseus Digital Library (http://www.perseus.tufts.edu/), which includes Classical as well as Arabic, Germanic, Renaissance, and 19th century materials; the Papyrus Portal (http://www.papyrusportal.de/), which searches through all online German papyrus collections; and the

Advanced Papyrological Information System (APIS: http://www.columbia.edu/cu/lweb/projects/digital/apis/), which does the same for American papyrological material. These applications all function as portals enabling the user to perform one search and receive results from across many decentralized digital datasets. This is a great improvement, as now the user does not have to visit every single digital datasets' own website to find the information they are looking for. But if the user wishes to reuse the data they have found in their own application or database they really need to be able to access it through an interoperable format such as XML (Roued Olsen, 2007). This is the case with DUGA. In order for DUGA to send a pattern to one of these huge datasets and return a list of words fitting this pattern, it needs to be able to connect to a Web Service of some sort.

4.4 Next step with DUGA

The next step for this research is to transform the prototype into the working application DUGA. The DUGA interface will use the idea of the prototype views but with more interaction between them. Apart from the interface the biggest issue with DUGA will be to find an appropriate storage method for the network of percepts and the evidence for and against each percept. This storage method must incorporate XML output functionality of the current interpretation in order to build the transcript and box views. Published editions use a slightly different set of conventions for encoding damage and doubts, even though most of them subscribe to using the Leiden Conventions. Being able to output the current interpretation as XML is therefore especially important for the transcript view, as DUGA must enable the scholar to view the transcript of their reading using the conventions style of their choice.

DUGA will concentrate on aiding the interpretative processes of defining the line, word and character outline and content of a text. In order to define the outline of each category, this research will need to include annotation software. Therefore, another step in the development of DUGA will be to add a type of annotation software as mentioned in Section 4.1.

DUGA would also benefit from the use of more knowledge bases to perform word search on other than the Vindolanda ink tablets knowledge base. This article has already discussed how DUGA could gain access to more material of this type. Nevertheless, there is also the possibility of using readings created through DUGA as knowledge bases for future readings of similar material. This brings the article back to the subject of XML output. It is important that DUGA enables the scholar to extract a well-formed EpiDoc standard XML document from the concluded interpretation. This can then be used both as another knowledge base and also as a means of publishing the reading as an edition online or in book form.

5 Conclusion

One of the main issues when reading a damaged ancient document is remembering the minor percepts that lead the way for the final interpretation as well as the knowledge and experience, which supported each percept. The development of the DSS web application DUGA aims to provide this service in order to support the scholars through their reading.

The development of DUGA is based on the idea of using qualitative data as well as previous percepts as evidence for and against each new percept in the interpretation. This will allow the scholars to work on this circular process by beginning with the more legible characters and words and using these and other scholarly resources to work towards the interpretation of more illegible characters and words.

The views and functions of the prototype have been explained through a hypothetical example of a reasoning and evidence-gathering process for the percept of the seventh letter of the word in the second line of Vindolanda tablet 159. Together with the development of a knowledge base Web Service called APPELLO, the prototype has brought this research within reach of this highly ambitious development that is DUGA. The next step towards DUGA is a synthesis of the ideas from Section 3.4 with the hypothetical example from the prototype in Section 3.1 in order to build an interface with

appropriate storage methods and output functionality in XML. Further testing of APPELLO on new literary datasets (i.e. the MAMA project) will also contribute to DUGA by incorporating more knowledge bases that can be used for word searches.

Acknowledgements

This article presents research in progress for the author's D.Phil thesis 'Using a Decision Support System to aid the reading of ancient documents'. This research is conducted as a part of the e-Science and Ancient Documents project (eSAD: http://esad.classics.ox.ac.uk/) and the author would like to thank fellow project members Prof. Alan Bowman (Centre for the Study of Ancient Documents (CSAD), University of Oxford), Prof. Sir Michael Brady (Department of Engineering, University of Oxford), Dr Melissa Terras (Department of Information Studies, University College London), and Dr Ségolène Tarte (Oxford e-Research Centre, University of Oxford) for their input and support.

References

Abeysinghe, S. (2008). *RESTful PHP Web Services*. Birmingham: PACKT Publishing.

Austin, M. (2008). *Information Integration and Decision Support for Multidisciplinary Team Meetings on Colorectal Cancer*. Ph.D. thesis, University of Oxford.

Austin, M., Kelly, M., and Brady, M. (2008). The benefits of an ontological patient model in clinical decision-support. *Twenty-Third AAAI Conference on Artificial Intelligence* 1774–75. http://www.aaai.org/Papers/AAAI/2008/AAAI08-325.pdf (accessed 2 October 2009).

Bagnall, R. S. (1997). Imaging of Papyri: A strategic view. *Literary and Linguistic Computing*, 12: 153–54.

Bowman, A. K., Brady, J. M., and Tomlin, R. S. O. (1997). Imaging incised documents. *Literary and Linguistic Computing*, 12: 169–76.

Bowman, A. K., Crowther, C. V., Kirkham, R., and Pybus, J. (2010). A virtual research environment for the study of documents and manuscripts. In Bodard, G. and Mahoney, S. (eds), *Digital Research in the Study of Classical Antiquity*. Farnham: Ashgate, pp. 87–103.

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

Bowman, A. K. and Thomas, J. D. (1994). *The Vindolanda Writing Tablets : (Tabulae Vindolandenses II)*. London: British Museum Press.

Bowman, A. K. and Thomas, J. D. (2003). *The Vindolanda Writing Tablets (Tabulae Vindolandenses III)*. London: British Museum Press.

Bowman, A. K., Tomlin, R. S. O., and Worp, K. A. (2009). Emptio Bovis Frisica: the 'Frisian Ox Sale' reconsidered. *Journal of Roman Studies*, 99: 156–70.

de la Flor, G., Luff, P., Jirotka, M., Pybus, J., Kirkham, R., and Carusi, A. (2010). The case of the disappearing ox: Seeing through digital images to an analysis of ancient texts. *Proceedings of the 28th International Conference on Human Factors in Computing Systems, CHI:* 473–482.

Eom, S. B., Kim, E. B., Lee, S. M., and Somarajan, C. (1998). A survey of decision support system applications (1988–1994). *Journal of the Operational Research Society*, 49(2): 109–20.

Haas, H. and Brown, A. (2004). *Web Services Glossary*, http://www.w3.org/TR/ws-gloss/ (accessed 2 October 2009).

Hippisley, D. (2005). *Encoding the Vindolanda Tablets: An Investigation in Contextual Encoding using XML and the EpiDoc Standards.* Masters in Electronic Communication and Publishing, University College London.

Porter, D. C., Reside, D., and Walsh, J. (2009). Text-Image Linking Environment (TILE). *Digital Humanities* 388–390. http://www.mith2.umd.edu/dh09/?page_id=99 (accessed 2 October 2009).

Richardson, L. and Ruby, S. (2007). *RESTful Web Services*. Cambridge: O'Reilly.

Roued Olsen, H. (2007). *Heritage Portals and Cross-border Data Interoperability* MSc in Archaeological Computing, University of Southampton.

Tarte, S. (2010). Papyrological investigations: transferring perception and interpretation into the digital world. *Literary and Linguistics Computing* (forthcoming).

Tarte, S., Brady, M., Roued Olsen, H., Terras, M., and Bowman, A. K. (2008). Image acquisition & analysis to enhance the legibility of ancient texts. *E-Science All*

Hands Meeting http://www.allhands.org.uk/2008/talks/1030.pdf (accessed 2 October 2009).

Terras, M. (2005). Reading the readers: modelling complex humanities processes to build cognitive systems. *Literary and Linguistic Computing*, 20: 41–59.

Terras, M. (2006). *Image to Interpretation. An Intelligent System to Aid Historians in Reading the Vindolanda Texts*. Oxford: Oxford University Press.

Turner, E. G. (1968). *Greek papyri: An Introduction*. Oxford: Clarendon.

Youtie, H. C. (1963). The Papyrologist: artificer of fact. *Greek, Roman, Byzantine Studies*, 4: 19–32.

van Groningen, B. A. (1932). Projet d'unification des Systemes de Signes Critiques. *Chronique d'Egypte*, 7: 262–9.

Notes

1 A Palimpsest document is one where the previous text has been scraped off and the document has been reused.

2 DUGA means to help, aid, or support in Old Norse which is exactly what the DSS web application aims to do for the scholar.

3 EpiDoc (http://epidoc.sourceforge.net/) is a schema for encoding epigraphic documents in TEI (Text Encoding Initiative: http://www.tei-c.org/index.xml) XML.

4 Description from the edition of tablet 159 (Bowman & Thomas, 1994, p. 102): Inv.no.85.048. 30 ×53 mm. A fragment of the top left-hand corner of a leaf, complete only at the top and left. The space between the top edge and the first line of writing suggests that we might have the beginning of the document. The text is difficult to classify, not least because so little of it survives. It might be part of an account but it does not look like the other accounts which seem to relate to military units and it has no date. It is noteworthy that lines 4 and 6 are indented by a considerable amount. The fact that it refers to a *turma* indicates the presence of cavalry at Vindolanda.

5 The kinaesthetic approach can be loosely translated to 'learning by doing' as in the scholar is understanding the text by drawing or otherwise reproducing what they see.

6 Cruciverbalism is the construction of crossword puzzles.

7 An interpunct (·) is a centred dot used in classical texts instead of a space between words.

8 The example of the last letter of the second line is taken from the notes in the published edition of tablet 159 (Bowman & Thomas, 1994, p. 102). However, the reading process, which is demonstrated here, is purely hypothetical.

9 The character recognition software is under development by the eSAD project.

10 The pattern is six letters that are certainly *hordia* at the beginning of the word, then one character space where the letter is unknown. This is followed by an open end because the tablet was damaged on the right side.

11 Appello is Latin for *to call*.

12 See (Abeysinghe, 2008) and (Richardson & Ruby, 2007) for an overview of RESTful Web Services.