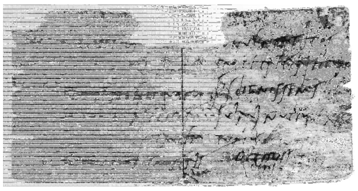
Aiding the Expert: Computers, Reading, and Ancient Texts



Dr Melissa Terras
Department of Information Studies
University College London
m.terras@ucl.ac.uk

The experiment

"We can try a little experiment. Let us resort to the fiction of programming an information transducer, a machine to read [ancient] texts. While so far only human beings have learned it, it is equally possible, and may one day be tried, to teach this skill to a machine …"

Erica Reiner (1973) "How We Read Cuneiform Texts." Journal of Cuneiform Studies 25: 3 -58, p.6.

Handwriting and Character Recognition

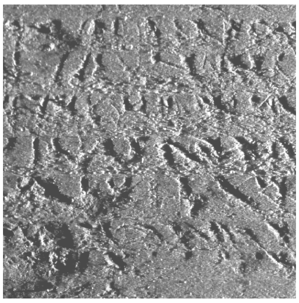
- Reading Handwriting is a primary aim of computing and engineering science
  - Vast research projects, various successes (OCR, etc)
  - Reading "difficult" texts beyond capacity of most computational approaches
    - Copperplate, dirty, noisy images, damaged, deteriorated
- What, if any, approaches can be used to assist papyrologists in reading damaged and abraded texts?
- How can you train computers to "read" ancient texts?
- Do we want them to "read" them?
- Case study regarding Vindolanda tablets

→ Henriette's current research on Interpretation Support Systems

Vindolanda Texts

- Roman Fort on Hadrian's Wall, England
- Texts from AD 92 onwards
- Two types
  - ink texts
    - Carbon ink on wood. 750 texts survive
  - stylus tablets
    - recessed centre filled with wax. 150 texts
- Only contemporary and immediate written evidence of Roman Army in Britain

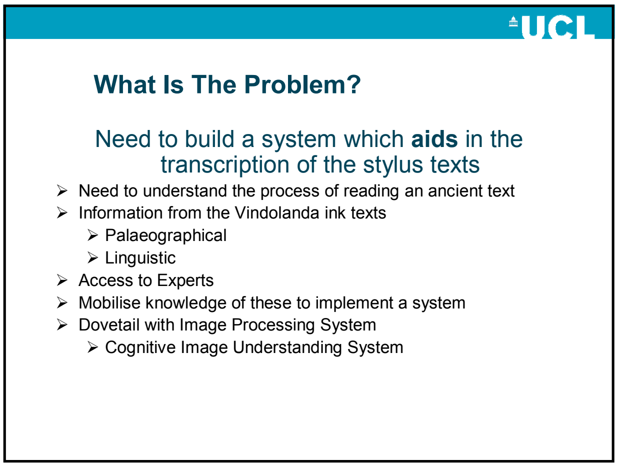
Close up - Tablet 1563



- Complex incisions
- Woodgrain
- Surface discolouration
- Warping
- Cracking
- Noisy image
- Palimpsest
- Long process

Vindolanda and Research – a reminder

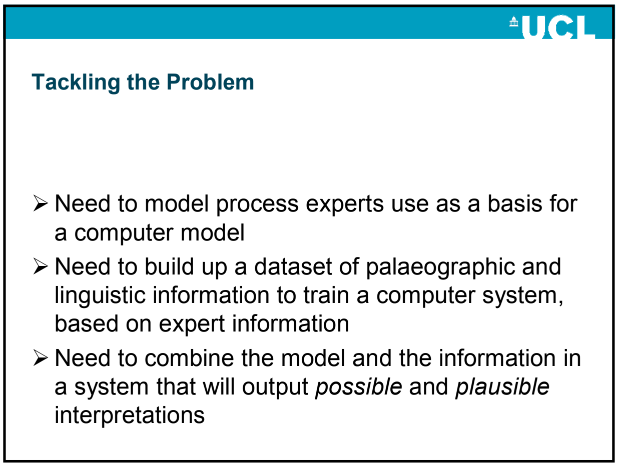
- Experts were observed reading ancient texts
  - Use raking light
- Digital Imaging Techniques were developed to analyse the surface of texts and to identify candidate strokes
  - "Phase Congruency"
- Professor Sir Mike Brady, Dr Veit Schenk, Dr Nick Molton, Dr Xiao-bo Pan (Engineering Science, University of Oxford)
- Professor Alan Bowman, Dr Charles Crowther (Centre for the Study of Ancient Documents, University of Oxford)
- Dr Segolene Tarte (e-Science Centre, University of Oxford)



What Is The Problem?

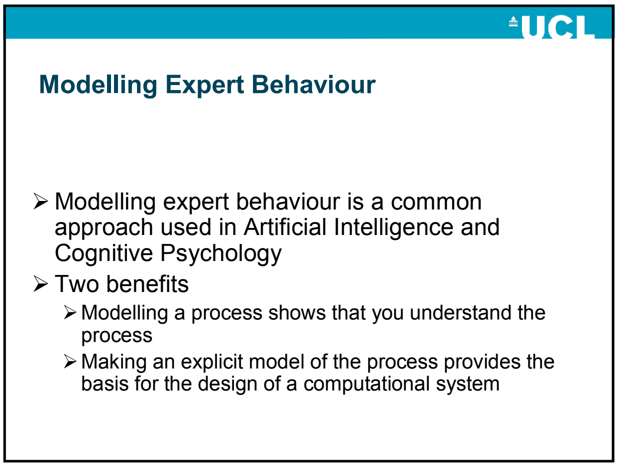
Need to build a system which **aids** in the transcription of the stylus texts

- Need to understand the process of reading an ancient text
- Information from the Vindolanda ink texts
  - Palaeographical
  - Linguistic
- Access to Experts
- Mobilise knowledge of these to implement a system
- Dovetail with Image Processing System
  - Cognitive Image Understanding System



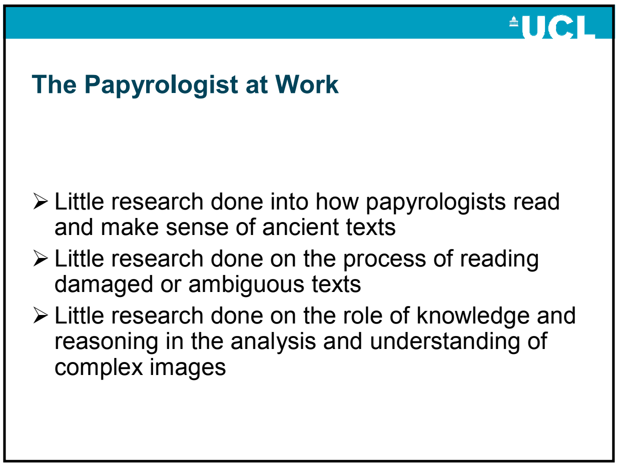
Tackling the Problem

- Need to model process experts use as a basis for a computer model
- Need to build up a dataset of palaeographic and linguistic information to train a computer system, based on expert information
- Need to combine the model and the information in a system that will output *possible* and *plausible* interpretations



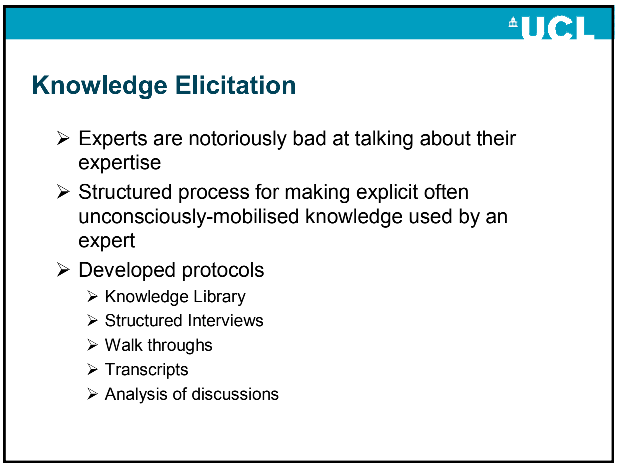
Modelling Expert Behaviour

- Modelling expert behaviour is a common approach used in Artificial Intelligence and Cognitive Psychology
- Two benefits
  - Modelling a process shows that you understand the process
  - Making an explicit model of the process provides the basis for the design of a computational system



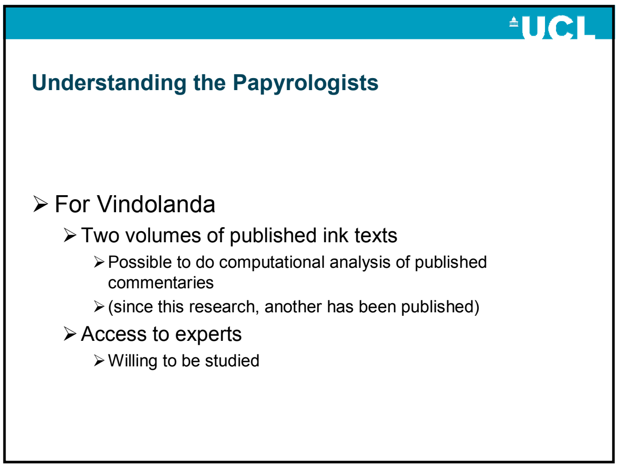
The Papyrologist at Work

- Little research done into how papyrologists read and make sense of ancient texts
- Little research done on the process of reading damaged or ambiguous texts
- Little research done on the role of knowledge and reasoning in the analysis and understanding of complex images



Knowledge Elicitation

- Experts are notoriously bad at talking about their expertise
- Structured process for making explicit often unconsciously-mobilised knowledge used by an expert
- Developed protocols
  - Knowledge Library
  - Structured Interviews
  - Walk throughs
  - Transcripts
  - Analysis of discussions



Understanding the Papyrologists

- For Vindolanda
  - Two volumes of published ink texts
    - Possible to do computational analysis of published commentaries
    - (since this research, another has been published)
  - Access to experts
    - Willing to be studied

A commentary – Stylus 836

banus bello suo salutem
(traces only)
acc__**erunt in in uecturas**
de_arios **oc**to **reliquos** solues
5 **rios nouem qua__r_r___**
sam dari debeb__
(interlinear addition?)
em libris
dus uale

‘Albanus to his Bellus greetings … they have received for transport costs 8 denarii. You will pay the remaining 9 denarii … ought to be given (?) … nine pounds (?) … Farewell.’

Notes:

1. There is a trace between the first and second **l** in **bello** which might or might not be a letter. The scratches on the wood show that this overlies an earlier text.
2. The correct reading is almost certainly **acceperunt**.
3. The word at the end of the line presents particular difficulty. Of the first three letters of **solues** only the **o** is certain. There is a clear high horizontal which has to be ignored if the first letter is read as **s**. The third letter might be **p**, and there is another apparent high horizontal which is discounted. The attraction of reading the word **solues** (from the verb **soluere** ‘to pay) is obvious if the word ‘**denarios**’ occurs twice in lines 4-5.

Understanding the Papyrologists (2)

- Knowledge Library
  - In depth knowledge about texts
- Analysis of Published Commentaries
  - Textual Analysis of contents
- Unstructured Interviews
  - Gaining broad insight into process
- Think Aloud Protocols
  - Setting experts tasks, and asking them to talk their way through it
    - Documenting and transcribing these sessions allows more textual information to analyse
      - – Content Analysis

Think Aloud Protocols- III 663

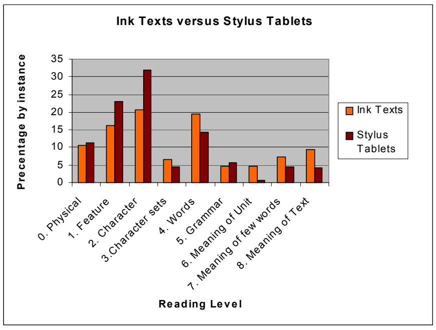
Basic Textual Analysis

- Using TACT and Wordsmith
- Allows analysis of the types of words used when discussing ancient texts
- Collocates
- Frequency
  - Ink Texts:
    - HORIZONTAL, BOLD, FORMAT, and DISCOLORATION, HYPOTHESIS, REASON
  - Stylus texts:
    - AFRAID, ASSUME, CONFUSING, CONVINCE, DECIDING, SURPRISED, and TRIED
- Analysis of the Latin itself
  - 10% of the characters in the published commentaries are marked as being uncertain

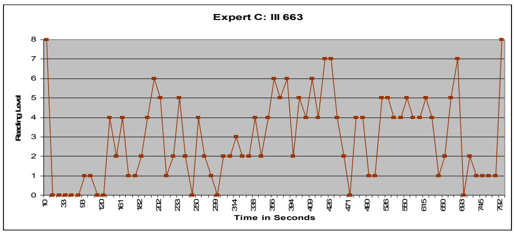
Content Analysis

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 a word
5	Discussion of grammar
4	Identification of possible word or morphemic unit
3	Identification of sequence of characters
2	Identification of possible character
1	Discussion of features of character
0	Discussion of physical attributes of the document
-1	Archaeological or historical context

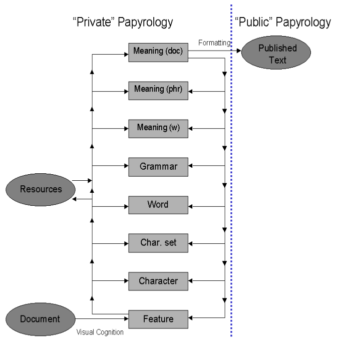
Content Analysis



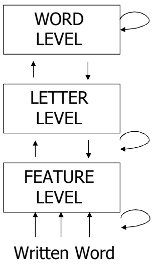
Content Analysis (2)



Model of the Papyrology Process

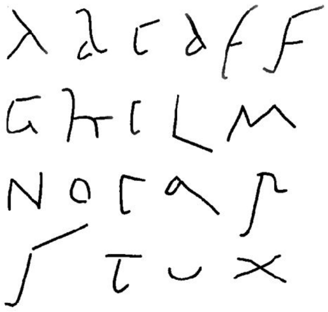


Word Superiority Effect



Rumelhart and McClelland's interactive-activation model of word recognition

Palaeographical Information

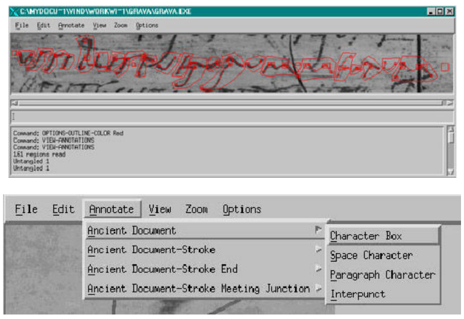


- Old Roman Cursive (ORC)
- Every day Roman Script
- Same used on ink and stylus?
- Forensic evidence
- => ink info can be used for stylus texts

Corpus Building

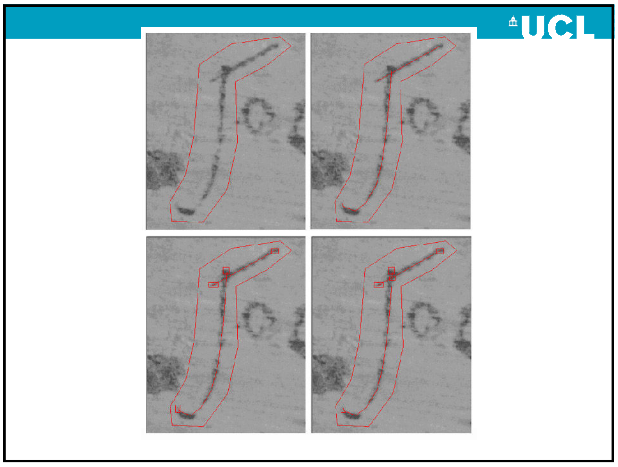
- Collect palaeographical information
  - Textual Sources
  - Knowledge Elicitation exercises
- Develop an encoding scheme
  - based on expert information
  - markup images -> XML text file
- Choose sample set and obtain Digital Images
  - Expert to provide data
  - British Museum
- Identify tool to markup images of text
- Mark up a corpus of images of large enough size to train a system

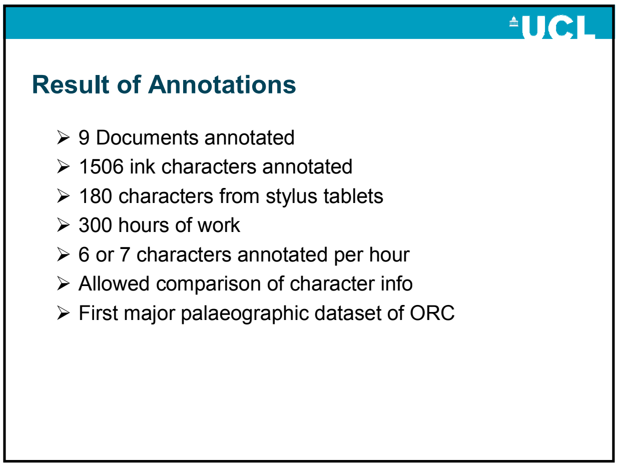
Annotating Program


C:\MYDOCU~1\VIND\WORKWI~1\GRAVA\GRAVA.EXE
File Edit Annotate View Zoom Options

```
Command: OPTIONS-OUTLINE-COLOR Red
Command: VIEW-ANNOTATIONS
Command: VIEW-ANNOTATIONS
161 regions read
Untangled 1
Untangled 1
```

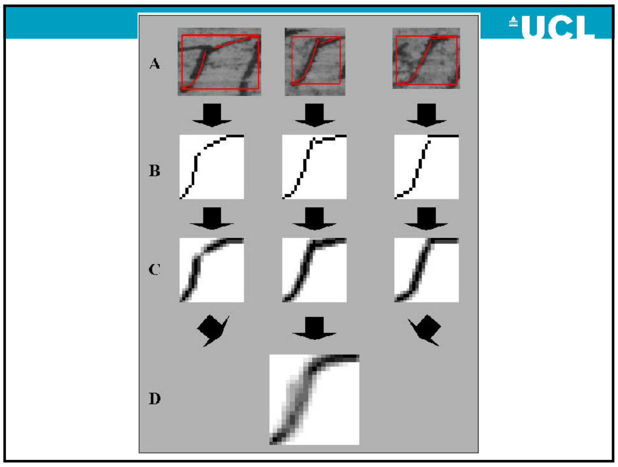
File Edit Annotate View Zoom Options
Ancient Document
Ancient Document-Stroke
Ancient Document-Stroke End
Ancient Document-Stroke Meeting Junction
Character Box
Space Character
Paragraph Character
Interpunct

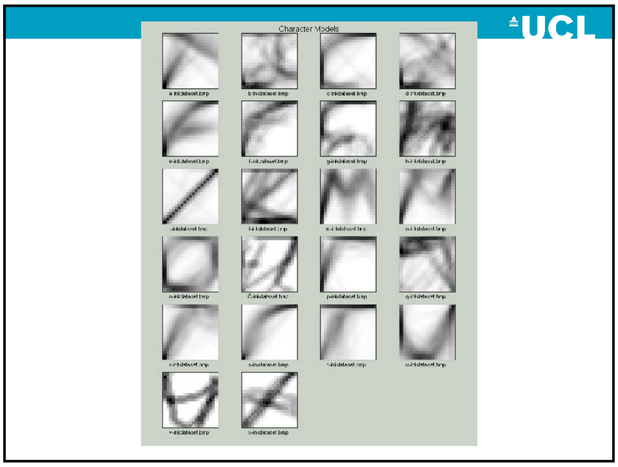




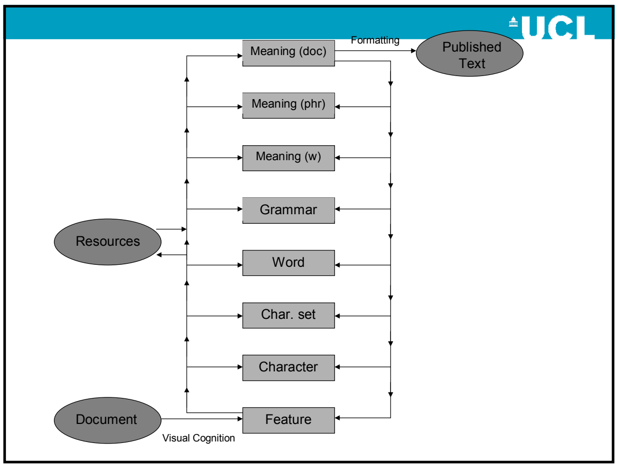
Result of Annotations

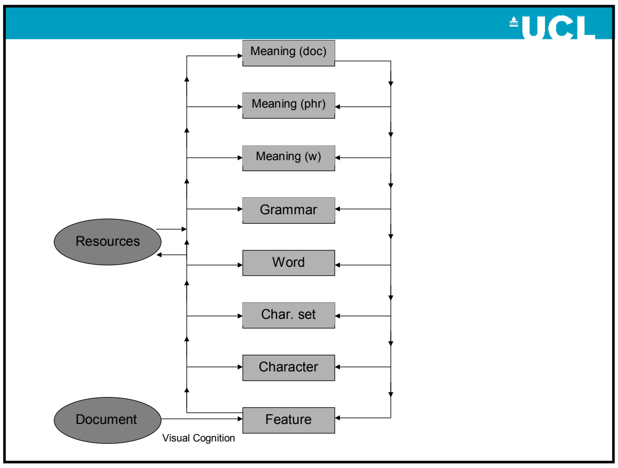
- ➢ 9 Documents annotated
- ➢ 1506 ink characters annotated
- ➢ 180 characters from stylus tablets
- ➢ 300 hours of work
- ➢ 6 or 7 characters annotated per hour
- ➢ Allowed comparison of character info
- ➢ First major palaeographic dataset of ORC

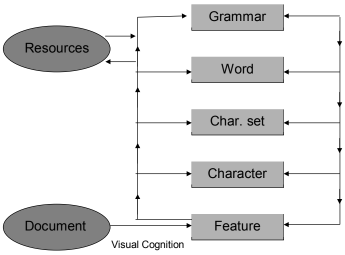


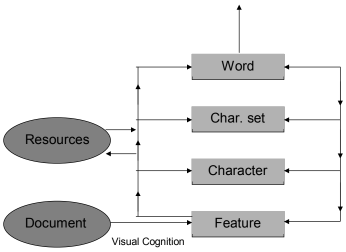


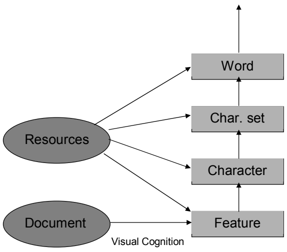
Character Models



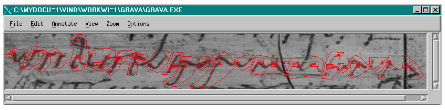


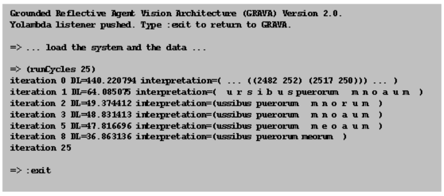






System in Action





```
Grounded Reflective Agent Vision Architecture (GRAVA) Version 2.0.
Yolambda listener pushed. Type :exit to return to GRAVA.

=> ... load the system and the data ...

=> (runCycles 25)
iteration 0 DL=440.220794 interpretation=( ... ((2482 252) (2517 250))) ... )
iteration 1 DL=64.085075 interpretation=(  u r s i b u s puerorum   m n o a u m  )
iteration 2 DL=49.374412 interpretation=(ussibus puerorum   m n o r u m  )
iteration 3 DL=48.831413 interpretation=(ussibus puerorum   m n o a u m  )
iteration 5 DL=47.816696 interpretation=(ussibus puerorum   m e o a u m  )
iteration 8 DL=36.863136 interpretation=(ussibus puerorum meorum  )
iteration 25

=> :exit
```

Can computers ever read ancient texts?

- Well, they can provide suggestions, based on known evidence
- They can keep a record of hypotheses encountered, discounted, and followed

Outcomes

- Built a prototype computer system that takes in unknown text and provides readings of that text based on known probabilities
- To speed up functioning of papyrologist, not replace them!
- Built for a specific audience and problem
- Proof of concept to show strength of architecture
- Indicate possibilities of a “Signal to Symbol” system
- No reason why this couldn’t be expanded across various types of text
  - Or individual tools – image markup- developed for the individual humanities scholar.

Outcomes (2)

- Computational techniques used to drive the system far from standard
  - Allowed real world application to test computational theory in AI
  - Benefited Engineering Science audience as well as Humanities scholars
- Research continues…
- Experimenting with truth maintenance systems
  - Online tools to aid in transcription
  - Record hypothesis and decisions

To conclude

- Can Computers ever read ancient texts?
  - Maybe
- Wrong question to ask:
- Can Computers ever be used to *aid* in reading ancient texts
  - Yes
  - Developing an understanding of how we can use technology to aid papyrologists brings an understanding of papyrology itself.

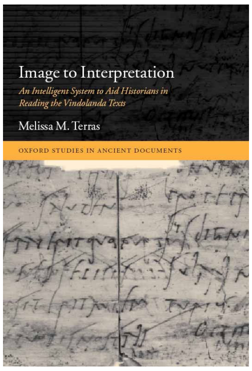


Image to Interpretation
An Intelligent System to Aid Historians in Reading the Vindolanda Texts
Melissa M. Terras
OXFORD STUDIES IN ANCIENT DOCUMENTS