

Faculty of Classics, University of Oxford



# Building an Interpretation Support System

Henriette Roued Olsen



eSAD workshop, 13th May 2009

**eSAD**  
e-Science and  
Ancient Documents  
[esad.classics.ox.ac.uk](http://esad.classics.ox.ac.uk)

# Building an Interpretation Support System

- What is an ISS and what can it do for the papyrologist?
- Building the ISS
- Knowledge Base
- Knowledge Base Web Service
- How does the Knowledge Base work with the ISS

# Interpretation Support System

# My D.Phil thesis

## “An Interpretation Support System (ISS) for the reading of Ancient Documents.”

- Not an expert system - won't take over the experts job!
  - Does the job that humans find difficult:
    - Remembering complicated reasoning
    - Searching huge datasets
    - Accessing other experts knowledge
    - Enable cooperation between experts on a single document
- Aiding tool to guide the expert through the steps:
  - Identifying and committing to possible:
    - Elements (Characters, Interpuncts, Indents and Spaces)
    - Characters (a, b or c)
    - Words (bovem or quem)
    - Phrases, sentences and paragraphs

# My D.Phil thesis

- Help put together an edition (final product)
  - Description
  - Transcription
  - Translation
  - Commentary
- Network of percepts (minor interpretations)
  - Low level: “these three line fragments are an incised stroke”
  - Higher level: “these five letters can make up the word *'legio'*”

# Stages of the ISS

'I will send(?) Publicus early in the morning so that I do not cause(?) Brocchus any delay. When he(?) comes to me will have breakfast.'

Publīcum p.rimo mane iam  
c.5 ] ne quam moram Broc-  
[ch- ....]am cum ad me uenerit  
].ē [[...]] `].detuiş' ian̄taturum  
].roţe[

Phrase where "primo" would fit in

+

primo

Stage 3: From words to phrases

"first / early" ←

primo

"believe" ←

credo

Stage 2: From characters to words

prelo

crilo

primo

predo

credo

p

c

r

i

e

m

l

d

o

Circularity

Stage 1: From elements to characters

Stage 0: Define elements (characters, spaces, interpuncts and indents)

# Building the ISS

- Browser based
- Different windows/views/divs
- Views built on an XML background

```
<lb n="1">
  <w id="1">
    <cb id="1" n="t">t</cb>
    <cb id="2" n="u">u</cb>
    <cb id="3" n="r">r</cb>
    <cb id="4" n="m">m</cb>
    <cb id="5"/>
    <gap reason="lost" extent="unknown" dim="right"/>
  </w>
</lb>
```

## EDITION

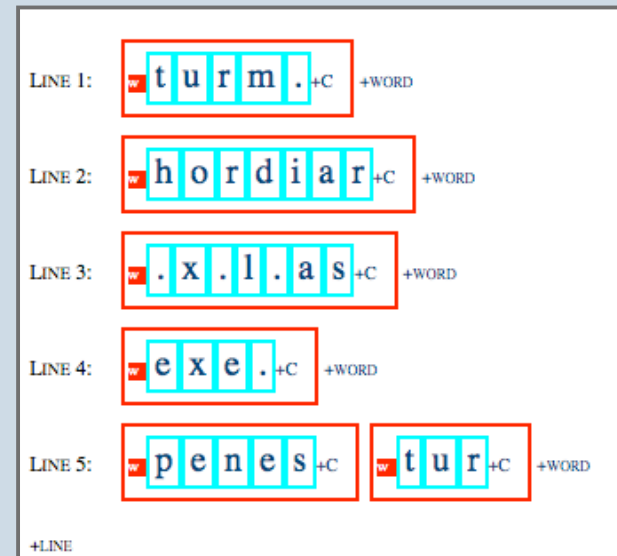
LINE 1: turm.[

LINE 2: hordiar[

LINE 3: .x.l.aş

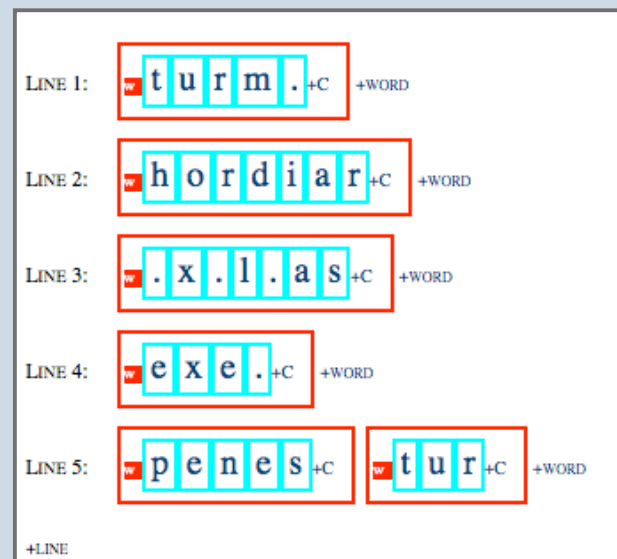
LINE 4: exe.[

LINE 5: penes[tur[

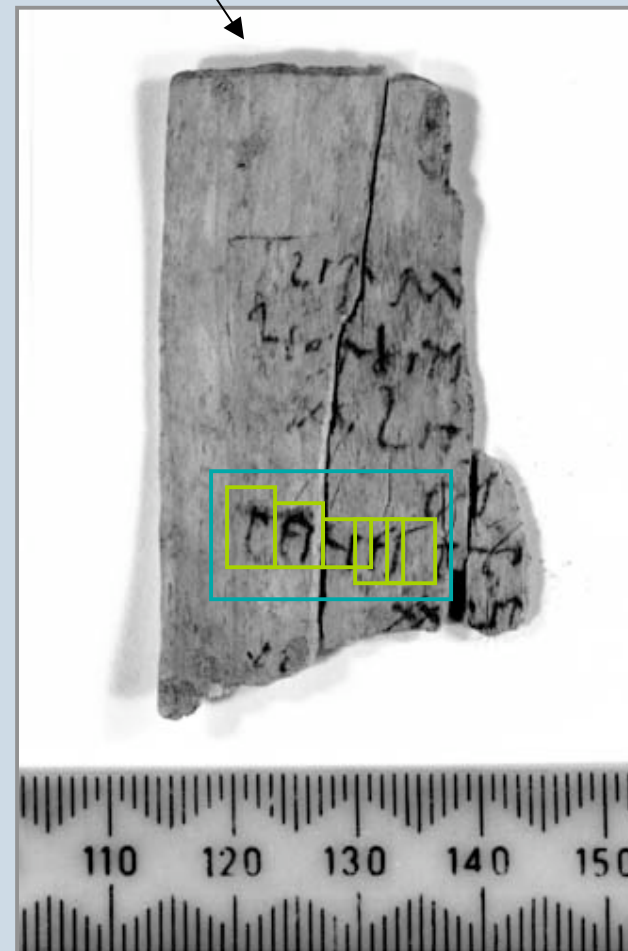


# Building the ISS

- Working with the BVREH project to use their annotation viewer get the effect of the box view on top of an image of a document.

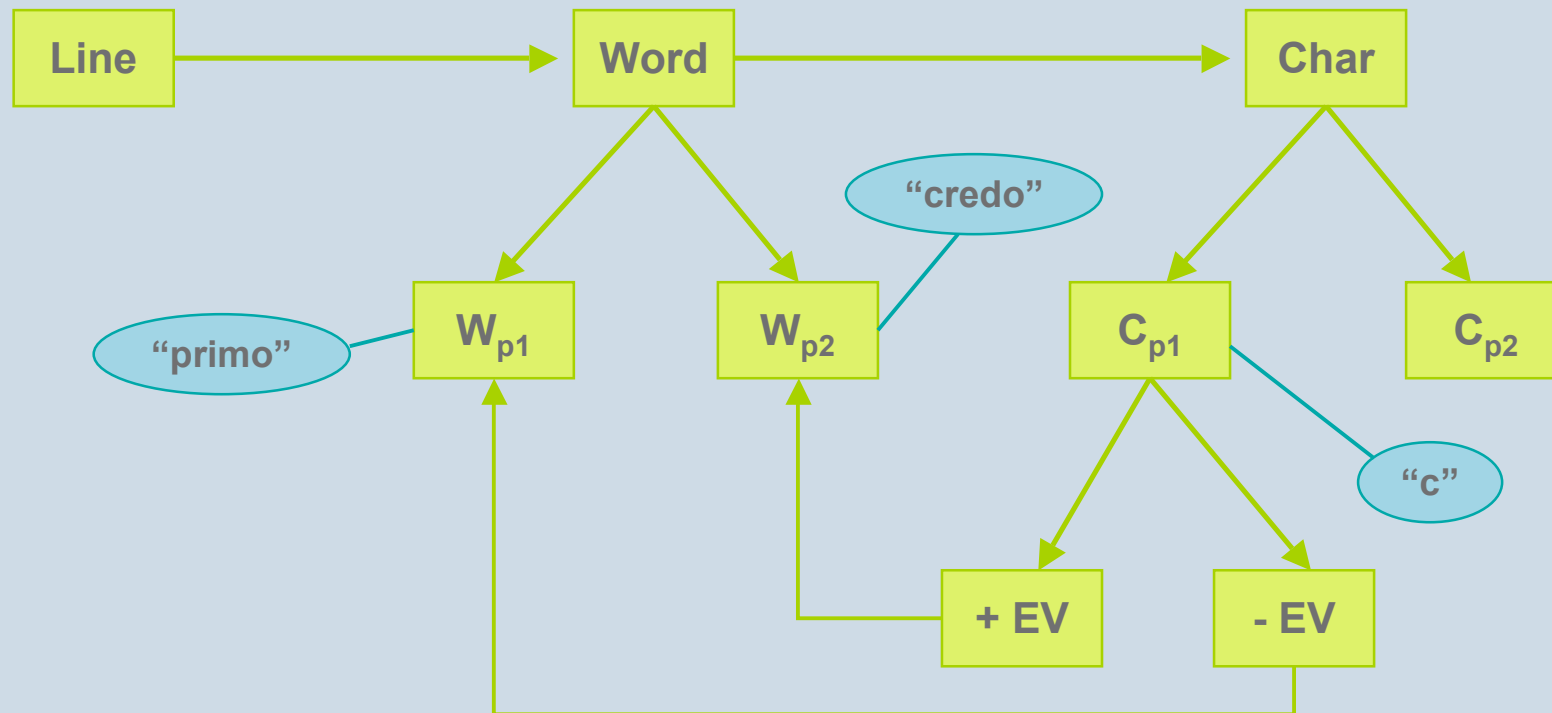


Vindolanda Tablet 159





# Ontology/model for the ISS



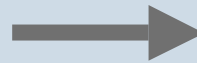
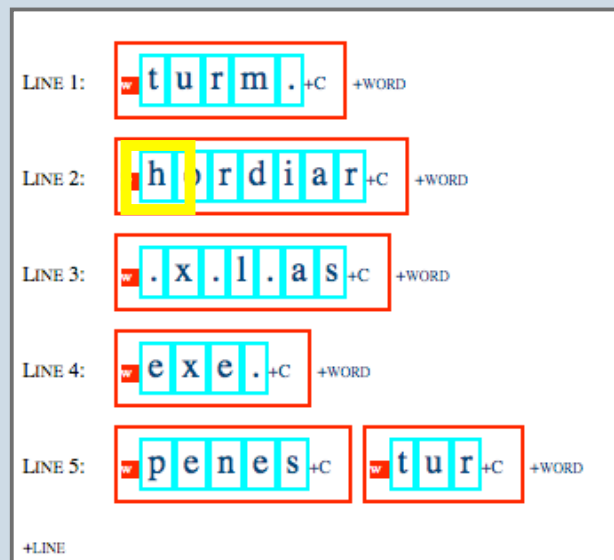
# Evidence based decision process

Will probably be the most used because “Historians know too much”



- **Judgements**
  - The expert believes it to be so (can justify with own words)
  - “Because I say so!”
- **Word search**
  - Connect to the knowledge base (described later)
- **Character recognition**
  - Connect with Segolene and Mike’s work
- **Contextual**
  - The tablet might have been found in a kitchen, which could be used as evidence for specific kitchen related words.
  - Connected with the BVREH work
- **Physical characteristics**
  - Is it a list or letter, one narrow column or two columns

# Evidence based decision process



## CURRENT INTERPRETATIONS

**Character:** h

**Change character:**

+ Word: hordiaria

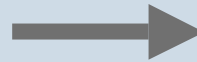
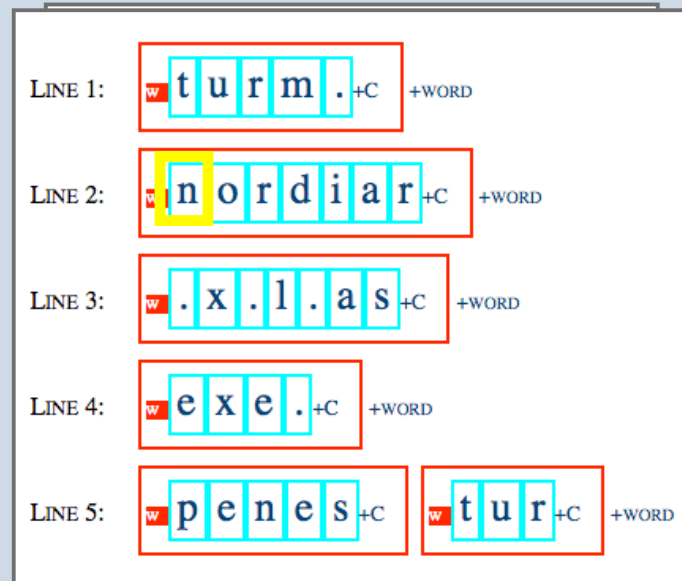
+ AKB: h, 'Because I say so'

+ Word search: hordiaria, hordiator(es)

- Character recognition: n, r

Evidence for the character H

# Evidence based decision process



## CURRENT INTERPRETATIONS

**Character: n**

**Change character:**

+ Character recognition: n, r

- Word: hordiaria

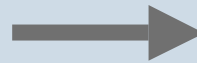
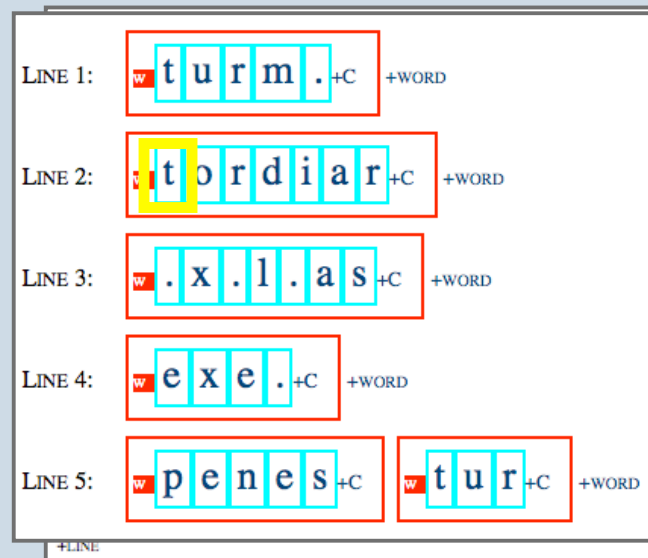
- AKB: h, 'Because I say so'

- Word search: hordiaria, hordiator(es)

Change character!

Evidence for the character N

# Evidence based decision process



Change character!

**CURRENT INTERPRETATIONS**

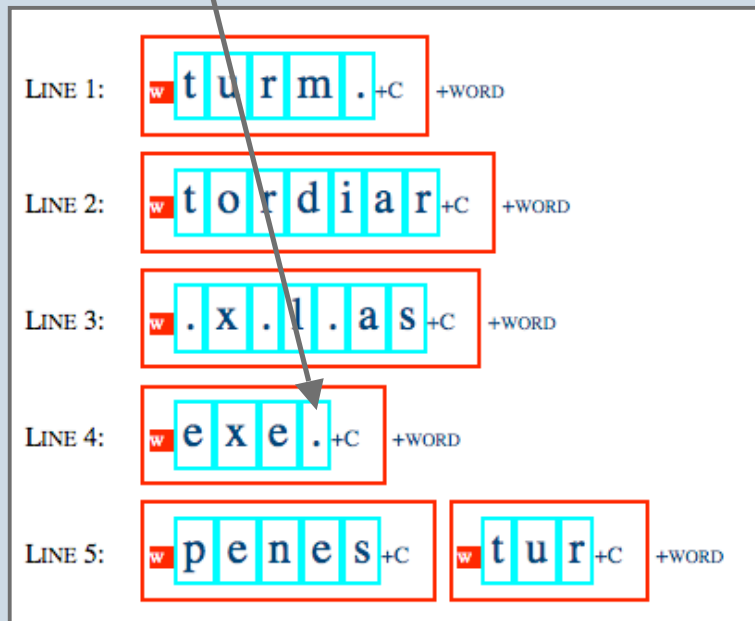
**Character: t**  
**Change character:**

- Word: hordiaria
- AKB: h, 'Because I say so'
- Word search: hordiaria, hordiator(es)
- Character recognition: n, r

Evidence for the character T

# Several suggestions

CLICK!



## CURRENT INTERPRETATIONS

### Character: m

+ AKB: m, 'Is the only character compatible with exe'

+ Word search: exemplum, exempti

+ Character recognition: m, r

### Character: r

+ Character recognition: m, r

- AKB: m, 'Is the only character compatible with exe'

- Word search: exemplum, exempti

Expert has not decided, which character is correct or does not think there are enough grounds for a decision.

Knowledge base

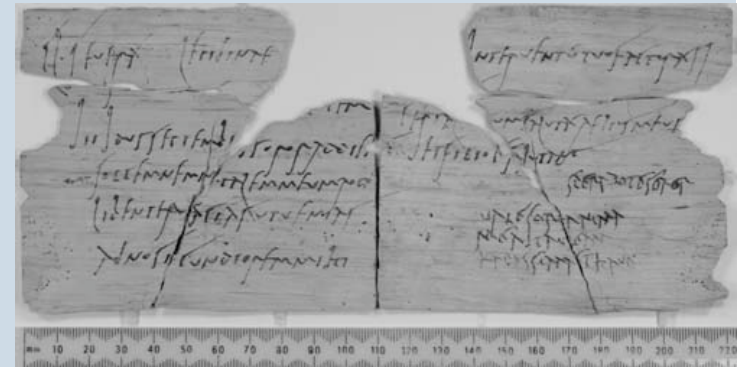
# External evidence used in stages

- Stage 0: Define characters and spaces (Elements)
  - Tessellation and identifying strokes (Segolene and Mike)
- Stage 1: From elements to characters:
  - Character Recognition (Segolene and Mike)
- Stage 2: From characters to words:
  - Words lists from **Knowledge Base**
- Stage 3: From words and letters to phrases and sentences:
  - Word combinations, known phrases
  - This could come from the knowledge base but has not been explored yet.



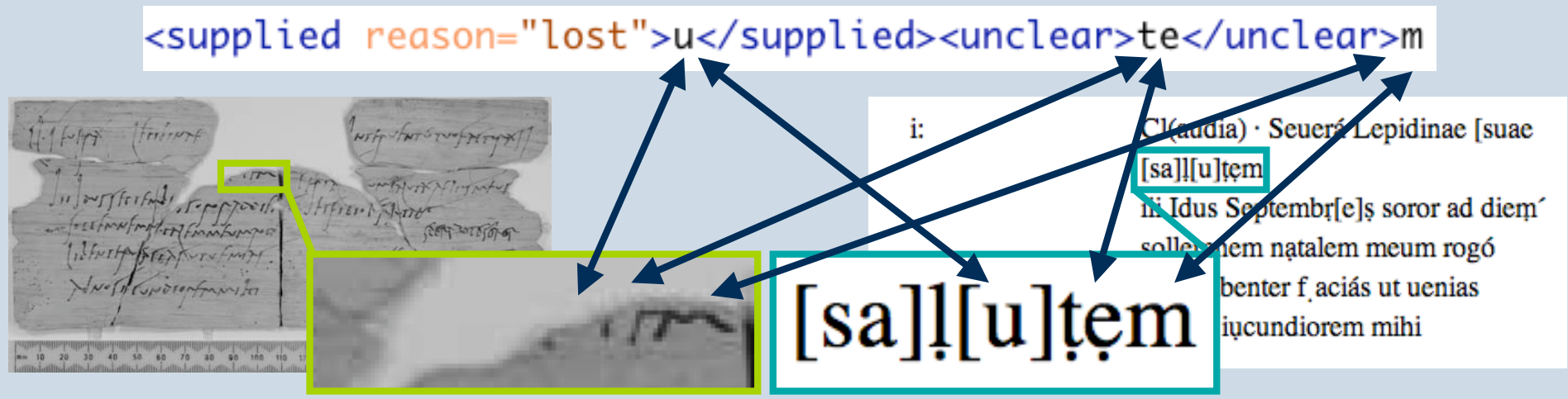
# Vindolanda Tablets - the lab rat

- c. 750 ink tablets (published)
- c. 150 stylus tablets
- Found at the Roman Fort of Vindolanda, Hadrians Wall
- Tablets are dated AD 92 onwards.
- Themes:
  - Letters
  - Military Documents
  - Accounts
  - Lists
- The tablets were marked up with EpiDoc XML to provide a Knowledge Base of words.



# Leiden Convention mark-up - EpiDoc

- Is a set of symbols used to show where an interpretation is not clear, where text is added or deleted.
  - Only the last character 'm' is perfectly legible.
  - .te are two characters found but unclear in the text.
  - [u] is a gap that has been filled in with what the expert supplied.



# Using XML - outputs

- Same XML allows different outputs

Traces of letters visible but unread  
Space left by scribe on tablet  
uacat vs. vac.

Vindolanda Style Leiden transcript

[diem] interuentú tuo facturá si  
[ ] [ c.3 ] § uacat  
Cerial[em t]uum salutá Aelius  
meus .  
10 et filioꝛuꝝ salutant uacat  
m<sup>2</sup> uacat sperabo te soror  
uale soror anima  
k

EpiDoc Style Leiden transcript

[ ] interuentú tuo facturá si  
[· ? ·] [·] § vac.  
Cerial[ ] uum salutá Aelius meus ·[· ? ·]  
10 et filioꝛuꝝ salutant vac.  
vac. sperabo te soror  
uale soror anima  
mea ita ualeam

<gap reason="space" dim="horizontal" /> <space dim="horizontal" /> <="character" />

# Contextual Encoding mark-up

- The encoding of all the textual features of an XML document.
- For the Vindolanda tablets this is based on the indices in the publications and is done manually.
  - People
    - Key = 'Crescens'
  - Words
    - Lemma = 'Pullus'
  - Dates
    - Calender = 'Maius'
    - Value = 'vi Kalendas'
  - Consuls
    - Key = 'AD 103 Traianus V'
  - Military and official terms
    - Key = 'ueterano'
  - Abbreviations
    - Abbr = 'pr'
    - Supplied = 'idie'

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

# Index searcher

Vindolanda Tablets use of knowledge base and new XML mark-up

## Vindolanda Index Searcher

[About Tablets](#)[Building Indices](#)[View Tablets](#)[Search Indices](#)

Choose an index to search

### SEARCH THROUGH THE INDEX OF LATIN WORDS

apud 581 (2)

caput 613 (1)

pullus 581 (15)

pullum 679 (1)

.....

.[

mīn[

pullum [

.opa.[

a.c.u[

.ati.[

oliua[

N

N

N

N

N

N

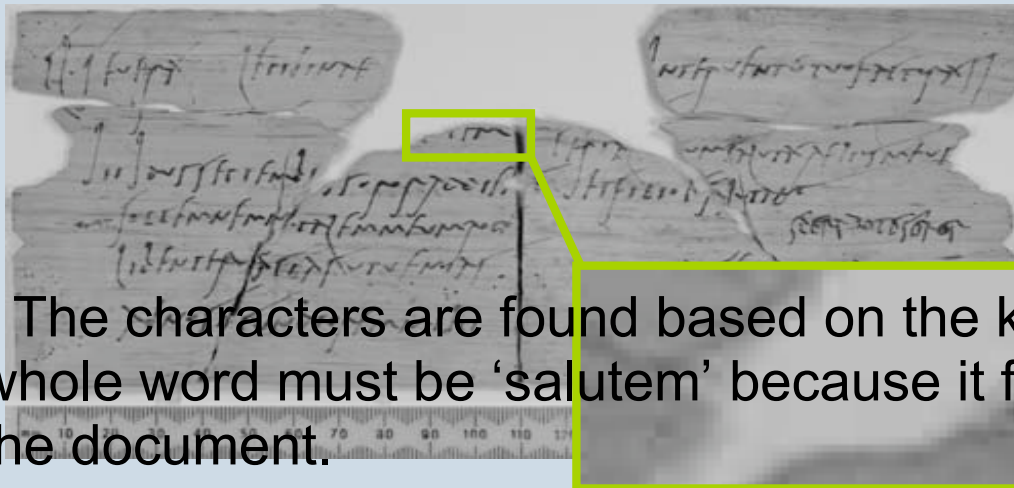
N

Try it out on: [www.roued.com/vindtab](http://www.roued.com/vindtab), will soon be improved using web service

# Example of non knowledge base use

- Example of how we can use circumstantial knowledge to find the characters of a word
- we only have a certain *m*

□ □ □ □ □ □ m



- The characters are found based on the knowledge that the whole word must be 'salutem' because it fits into the context of the document.

s a l u t e m

# Example of knowledge base use 2

- Example of how we can use the word list to find the missing letters in a word like a cross-word puzzle.
- *Salutem* - we only have a certain *m*

□ □ □ □ □ □ m

- With the character recognition tools we may be able to interpret the unclear letters as:

□ □ l □ t e m

- There is no trace of the last 3 letters. In fact we don't even know that the 2 letters at the beginning are present. If we compare this to a word list of words used at Vindolanda we may get:

s a l u t e m



# Knowledge base web service

- RESTful Web Service = URL with parameters

- The web service will search through the knowledge base and return lists of XML depending on the method used

- Method:

- get\_word - returns a list of words (currently only 'idem' and 'eodem')
  - get\_tablet - returns a list of tablets

[http://localhost/vindoWebService/tablet.php?method=get\\_word&pattern=e\\*d](http://localhost/vindoWebService/tablet.php?method=get_word&pattern=e*d)

- Pattern:

- \* is a wildcard - any letter big or small
  - {abc} means either the letter a, b or c
  - P\*\*{il}u\* will return words like "pullu"

```
<response>
- <element index="word">
  <lemma>idem</lemma>
- <types>
  - <type>
    <typeLemma>eodem</typeLemma>
  - <tablets>
    - <tablet>
      <tabletNumber>581</tabletNumber>
      <number>8</number>
    </tablet>
  </tablets>
  </type>
</types>
</element>
- <element index="word">
  <lemma>reddo</lemma>
```

Will soon be available online to use!



# Using web service in index searcher

- Using the knowledge base web service in the index searcher means that we can search with the wildcards.
- But the index searcher uses LiveSearch (AJAX) and this gives problems.

## SEARCH THROUGH THE INDEX OF LATIN WORDS

**pullus** 581(15) 616(1)

pullum 679(1)

**propitius** 628(1)

**hospitium** 632(1)

**poculum**

poculum 677(1)

**cuppedium**

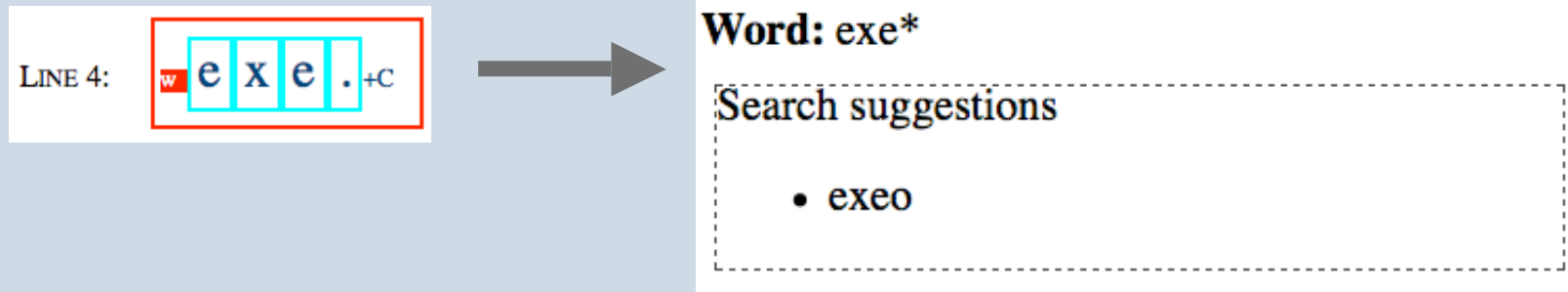
copadia 679(1)

## SEARCH THROUGH THE INDEX OF LATIN WORDS

**Warning:** DOMDocument::load() [**domdocument.load**]: Extra content at the end of the document in http%3A//localhost/vindtab/tablet.php%3Fmethod=get\_word&pattern=%7Bas, line: 2 in /Users/klas0252/Sites/vindtab/manage.php on line 401

# Using web service in ISS

- One of the aims of the web service is that it can be used in the ISS.
  - For the word on line 4 we have decided that the first three characters are “exe”. The last character is not visible but we are sure there is a character so this is a wildcard.



# Conclusion

- Building the ISS
  - Expert driven
  - Different views (transcription, box, edit view)
  - Evidence based (Judgements, physical, contextual, word, character)
  - XML outputs
- Vindolanda Knowledge Base Web Service
  - Mark-up of the Vindolanda Tablets
  - RESTful Web Service connecting to the tablets
  - Index Searcher Online
  - Work with other projects to get more knowledge base web services

# Further work

- Incorporate the BVREH annotation viewer
- Find a better storage functionality for the percepts
  - Triple store
- Turn the prototype into a working web application
- Ontology of the area in the ISS and make the ISS ontology driven.
- Find a way to combine the ISS with work of the rest of the eSAD group and enable it in the VRE.