Mining Paper Catalogues

A Multilingual Solution to Reduce Verbose Fields to Consistent Terminology

Tim Evans[1], Felix Kumaul[2]

23rd Annual Meeting EAA, Maastricht

31 August 2017

[1]Archaeology Data Service, University of York

[2]Archaeological Institute, University of Cologne

MOTIVATION

Data Source

92

Form 23 Conical cup with smooth vertical rim
Konische Schale mit glattem Steilrand
Coppa troncoconica con orlo verticale
Coupe tronconique à rebord vertical lisse

Conical cup representing the further evolution of Form 22. The floor is now always flat or biconical (meeting the wall at a sharp angle on the inside), usually with a low foot.

23.1: Plain tapering rim, inclined slightly inwards, sometimes bearing applied decoration.
23.2: Rim with flat outer face bearing applied decoration bounded above and below by simple convex mouldings; inner face plain or with a groove at lip.

Production

Subform 23.1 is probably made in many parts of Italy; examples in Padana ware do not show applied decoration. Subform 23.2 is made in Italy but apparently not in the Padana region.

Date

Subform 23.2 belongs to the second and third quarters of the first century A. D.: it is common in the South Stoa deposits at Corinth and at Pompeii. Subform 23.1 is less readily datable as it may occur as a simplified version of Form 22 or Form 23: other features of the vessel (e. g. foot-profile, decoration) may provide a clearer indication of date than the shape of the rim.

Distribution

Subform 23.2 is very common throughout the Mediterranean region, with sporadic examples found in the North and in North-Italy; Subform 23.1 is relatively uncommon.

References

23.1.1 Karthago K 78/172a, unpublished. Stamp L.MA, O.-C. –. Italy.
23.1.2 Berenice B210.2. Anepigraphic stamp. Italy.
23.2.1 Corinth 1973 pl.84,70. Stamp CAMVRI, O.-C. 397. Arezzo.
23.2.2 Berenice B216.2. Italy.

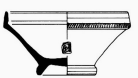
Other findspots

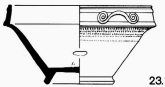
23.1 Asciburgium, Bologna, Bolsena, Conimbriga, Köln, Luni, Magdalensberg, Ordona, Pollentia, Roma.
23.2 Not separately listed.

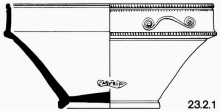
Concordance

23.1: Goudineau 20c; 25a; 37a. – Berenice B210.2.
23.2: Goudineau 40. – Barocelli 11. – Berenice B216. – Hayes 23.
Pieces described as Haltern 9 sometimes belong to this form.

P. M. K.


23.1.1


23.1.2


23.2.1

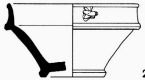

23.2.2

Figure 1: Sample from **consp**

Oh dear!

Problem

Running texts contain a lot of *irrelevant information* (for machine processing).

This makes database lookups without keywords **extremely inefficient**.

What we **have**:

Production

Subform 23.1 is probably made in many parts of Italy; examples in Padana ware do not show applied decoration. Subform 23.2 is made in Italy but apparently not in the Padana region.

Distribution

Subform 23.2 is very common throughout the Mediterranean region, with sporadic examples found in the North and in North-Italy; Subform 23.1 is relatively uncommon.

UNSTRUCTURED DATA

What we **want**:

```
{
   "form": "23.1",
   "origin": "Italy",
   "decoration": "none",
   "occurs": "uncommon"
},
{
   "form": "23.2",
   "origin": "Italy, not Padana",
   "occurs": "Mediterranean region;
              North-Italy"
}
```

STRUCTURED DATA

What we **have**:

Production

Subform 23.1 is probably made in many parts of Italy; examples in Padana ware do not show applied decoration. Subform 23.2 is made in Italy but apparently not in the Padana region.

Distribution

Subform 23.2 is very common throughout the Mediterranean region, with sporadic examples found in the North and in North-Italy; Subform 23.1 is relatively uncommon.

UNSTRUCTURED DATA

What we **want**:

```
{
   "form": "23.1",
   "origin": "Italy",
   "decoration": "none",
   "occurs": "uncommon"
},
{
   "form": "23.2",
   "origin": "Italy, not Padana",
   "occurs": "Mediterranean region;
              North-Italy"
}
```

STRUCTURED DATA

What we **have**:

Production

Subform 23.1 is probably made in many parts of Italy; examples in Padana ware do not show applied decoration. Subform 23.2 is made in Italy but apparently not in the Padana region.

Distribution

Subform 23.2 is very common throughout the Mediterranean region, with sporadic examples found in the North and in North-Italy; Subform 23.1 is relatively uncommon.

UNSTRUCTURED DATA

What we **want**:

```
{
   "form": "23.1",
   "origin": "Italy",
   "decoration": "none",
   "occurs": "uncommon"
},
{
   "form": "23.2",
   "origin": "Italy, not Padana",
   "occurs": "Mediterranean region;
              North-Italy"
}
```

STRUCTURED DATA

TEXT MINING

Classification

Definition: Text Mining

Text Mining is a general term covering several different ideas, e. g.:

- Information retrieval
- Statistical analysis
- Information extraction
- ...

Classification

Definition: Text Mining

Text Mining is a general term covering several different ideas, e. g.:

- Information retrieval
- Statistical analysis
- **Information extraction**
- ...

Classification

Definition: Text Mining

Text Mining is a general term covering several different ideas, e. g.:

- Information retrieval
- Statistical analysis
- **Information extraction**
- ...

Classification

Definition: Text Mining

Text Mining is a general term covering several different ideas, e. g.:

- Information retrieval
- Statistical analysis
- **Information extraction**
- . . .

Classification

Definition: Text Mining

Text Mining is a general term covering several different ideas, e. g.:

- Information retrieval
- Statistical analysis
- **Information extraction**
- ...

Information Extraction

Definition: Information Extraction (IE)

"*[***IE** *refers to] the* **identification** *and extraction of instances of a particular class of events or relationships in a* **natural language text** *and their* **transformation** *into a structured representation.*" – Grishman 1997, Eikvil 1999

Some other facts about Information Extraction:

- Computer scientists have a hard time with IE (for over 30 years now!)
- IE is really super difficult and often inaccurate.

Information Extraction

Definition: Information Extraction (IE)

"*[**IE** refers to] the* **identification** *and extraction of instances of a particular class of events or relationships in a* **natural language text** *and their* **transformation** *into a structured representation.*" – Grishman 1997, Eikvil 1999

Some other facts about Information Extraction:

- Computer scientists have a hard time with IE (for over 30 years now!)
- IE is **really super difficult** and **often inaccurate**.

Information Extraction

Definition: Information Extraction (IE)

"*[**IE** refers to] the* **identification** *and extraction of instances of a particular class of events or relationships in a* **natural language text** *and their* **transformation** *into a structured representation.*" – Grishman 1997, Eikvil 1999

Some other facts about Information Extraction:

- Computer scientists have a hard time with IE (for over 30 years now!)
- IE is **really super difficult** and **often inaccurate**.

Sorry!

DISCLAIMER

In this presentation, we show **preliminary** results, as this project is still work in progress.

IE Process Pipeline with UIMA

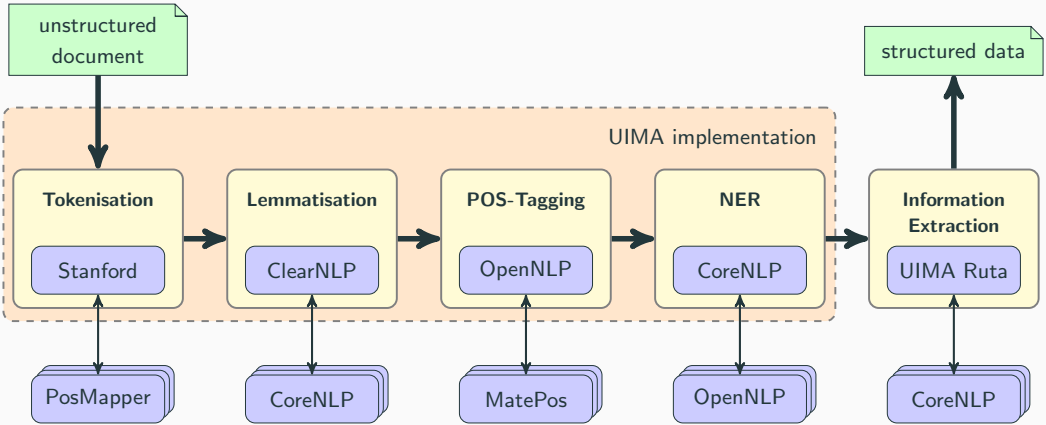


Figure 2: IE Process Pipeline.

POS-Tagging

Figure 3: POS-tagging examples after lemmatisation.

POS-Tagging

Figure 3: POS-tagging examples after lemmatisation.

Adapting the NER

Most NERs (e. g. **Stanford CoreNLP**) only recognise 8 entities types:

PERSON	DATE
ORGANIZATION	TIME
LOCATION	MONEY
PERCENT	MISC

So we have to add the custom entity type FORM.

Two approaches for NER

Rule-based approach

- High precision, but lower recall
  $\Rightarrow$ Many many rules?!

Machine-learning approach

- Lower precision, but high recall
- Needs to be trained!

Two approaches for NER

Rule-based approach

- High precision, but lower recall

  $\Rightarrow$ Many many rules?!

K 612 Abb. 117,2
Amphoriskos mit breitem, annähernd zylindrischem Hals und davon abgesetzter Schulter. Schmaler, aussen vorkragender Wulstrand mit an der Innenseite umlaufender breiter Riefe. Zwei Bandhenkel vom Mündungsrand zur Schulter. Form des Gefässkörpers unbekannt.
Ton I,B mit rotem Überfang
M. Dm 9.5–10 cm
Vk. Selten
Dat. Spätes 3.–4. Jh. n. Chr.

Figure 4: Excerpt from **eleph**

Machine-learning approach

- Lower precision, but high recall
- Needs to be trained!

Two approaches for NER

Rule-based approach

- High precision, but lower recall

  $\Rightarrow$ Many many rules?!

K 612 Abb. 117,2

Amphoriskos mit breitem, annähernd zylindrischem Hals und davon abgesetzter Schulter. Schmaler, aussen vorkragender Wulstrand mit an der Innenseite umlaufender breiter Riefe. Zwei Bandhenkel vom Mündungsrand zur Schulter. Form des Gefässkörpers unbekannt.

Ton I,B mit rotem Überfang

M. Dm 9.5–10 cm

Vk. Selten

Dat. Spätes 3.–4. Jh. n. Chr.

Figure 4: Excerpt from **eleph**

Machine-learning approach

- Lower precision, but high recall
- Needs to be trained!

Two approaches for NER

Rule-based approach

- High precision, but lower recall

  $\Rightarrow$ Many many rules?!

K 612 Abb. 117,2

Amphoriskos mit breitem, annähernd zylindrischem Hals und davon abgesetzter Schulter. Schmaler, aussen vorkragender Wulstrand mit an der Innenseite umlaufender breiter Riefe. Zwei Bandhenkel vom Mündungsrand zur Schulter. Form des Gefässkörpers unbekannt.

Ton I,B mit rotem Überfang

M. Dm 9.5–10 cm

Vk. Selten

Dat. Spätes 3.–4. Jh. n. Chr.

Figure 4: Excerpt from **eleph**

Machine-learning approach

- Lower precision, but high recall
- Needs to be trained!

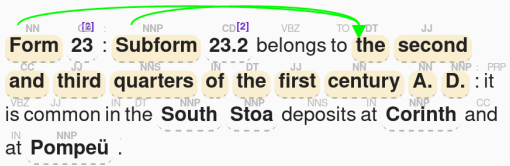


Figure 5: Manually annotated sentence from **consp** in `iepy`.

Temporal Expressions

With **HeidelTime** temporal expressions are mapped to TIMEX3 standard

around 140 B. C.	⟼	APPROX BC0140
Sptes 3.4. Jh. n. Chr.	⟼	END 02; 03
second quarter first century B. C.	⟼	XXXX-Q2 BC00
first half third century A. D.	⟼	XXXX-H1 02

HeidelTime supports many other languages, e. g. German, Italian, French, ...

Relation Extraction

Subject	Relation	Object
quick brown fox	jump over	lazy dog
K 612	dates	03[1]
Form 23	dates	`XXXX-Q2 00`[2]
Subform 23.2	dates	`XXXX-Q2 00`

$\Rightarrow$ e.g.

```
{  "form":   "23.2",
   "dating": "XXXX-Q2 00"  }
```

[1] "4th century A. D."

[2] "second and third quarters of the first century A. D."

Relation Extraction

Subject	Relation	Object
quick brown fox	jump over	lazy dog
K 612	dates	03[1]
Form 23	dates	`XXXX-Q2 00`[2]
Subform 23.2	dates	`XXXX-Q2 00`

$\Rightarrow$ **e. g.**

```
{  "form":   "23.2",
   "dating": "XXXX-Q2 00"  }
```

[1] "4th century A. D."

[2] "second and third quarters of the first century A. D."

Locations with HeidelPlace

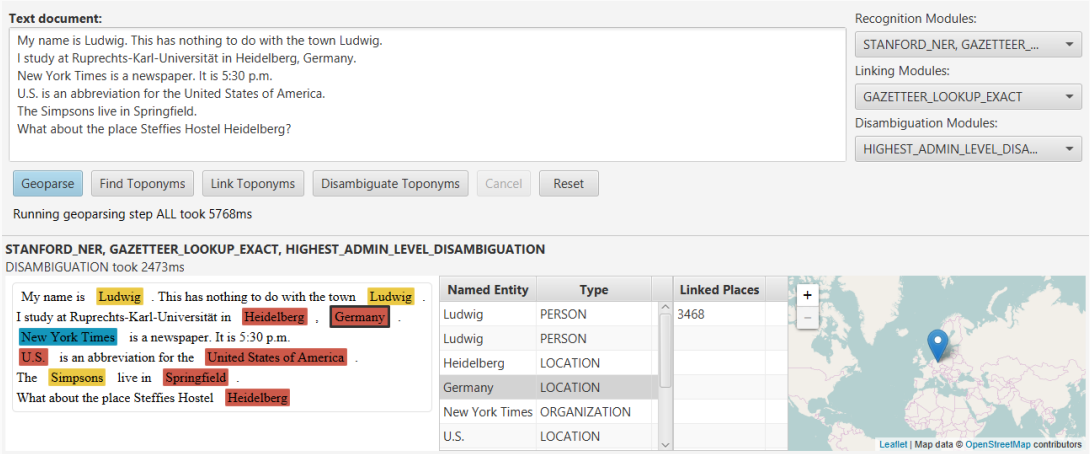


Figure 6: Screenshot of HeidelPlace.

MULTILINGUALISM

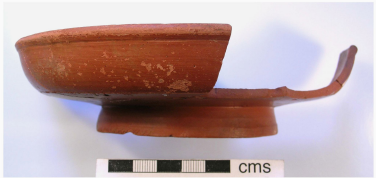
Background

Two problems:

- Linguistic
- Conceptual

Different languages

Different traditions



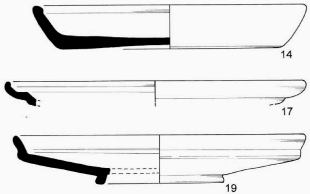


Figure 7: Plate, platter or dish?

Creating controlled vocabularies

Creating wordlists that project team would be most useful to describe the key features of a vessel or sherd

- Sherd type (e.g. rim or handle)
- Form (e.g. plate or bowl)
- Decoration form (e.g. burnished)
- Decoration color (e.g. yellow)
- Fabric (e.g. Dressel 28 fabric)

Lessons from ARIADNE

Used tools and methodology developed for the ARIADNE project by the Hypermedia Research Group at the University of South Wales

- Created a neutral spine based on the Getty Institute's Art and Architecture Thesaurus (AAT)
- This spine was poulated by memebers from partner organisations, identifying common terms and concepts within it
- Project partners then mapped terms in their language to this neutral spine
- French terms supplied courtesy of a 2001 Masters thesis by Caroline SOURZAT (thanks to Eleni Schindler Kaudelka for identifying this on the ArchAIDE blog!)

Mapping terms and concepts (part 1)

Often this was very straightforward, for example:

- The Italian terms *graffita*, *graffita a punta*, *graffita a stecca* = "sgraffito" (http://vocab.getty.edu/aat/300266416)
- The Spanish term *Cntaro* = "jars" (http://vocab.getty.edu/aat/300195348)
- The German terms *gebogener Henkel*, *Ohrfrmiger Henkel*, *langer Vertikalhenkel* = "handles" (http://vocab.getty.edu/aat/300266416)

Mapping terms and concepts (part 2)

Often this was more complicated, with partners having differing perceptions on what to call something (e.g. "plate versus platter)

In truth, this confusion may also be reflected by what has come out of the ground!

An advantage of using the AAT (a "SKOS'd" thesaurus), is that ambiguity or difference in nomenclature can be resolved by a broader term or concept, so for example...

Mapping terms and concepts (part 3)

Looking at the hierarchies for plate and platter in the AAT we can see that both are "dishes (vessels for food)", or even broader "culinary containers". So whole we can retain our original classifications (and this is essential for text mining), we can agree at a fundamental level *what these fundamentally are*

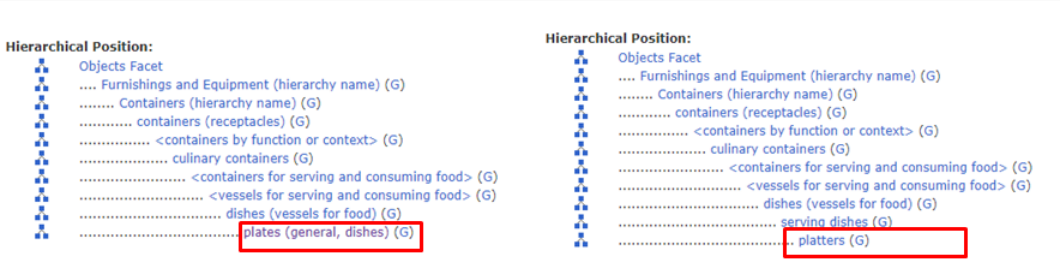


Figure 8: AAT Hierarchies for Plate and Platter

OUTLOOK

Challenges to meet

challenges:

choice of tools, coreferences in text, eloquence of archaeologists, maybe calculating F-value?

HeidelTime:

second and **third quarter** of the **first century A. D.** $\longmapsto$ `XXXX-Q3; 00`

References

Thank you very much for your attention!

Questions?

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement 693548

Mining Paper Catalogues

A Multilingual Solution to Reduce Verbose Fields to Consistent Terminology

Tim Evans[1], Felix Kumaul[2]

23rd Annual Meeting EAA, Maastricht

31 August 2017

[1]Archaeology Data Service, University of York

[2]Archaeological Institute, University of Cologne