Mining Paper Catalogues

A Multilingual Solution to Reduce Verbose Fields to Consistent Terminology

Tim Evans[1], Felix Kußmaul[2]

23rd Annual Meeting EAA, Maastricht

31 August 2017

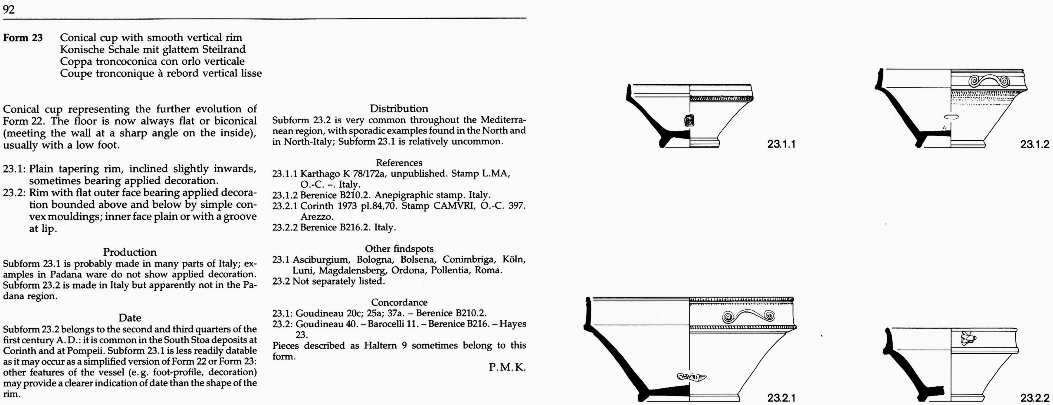
[1]Archaeology Data Service, University of York

[2]Archaeological Institute, University of Cologne

ARCHAIDE
ARCHAEOLOGICAL
AUTOMATIC INTERPRETATION
AND DOCUMENTATION
OF CERAMICS

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23.1.1 Karthago K 78/172a, unpublished.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.

Figure 1: Sample from Ettlinger, *Conspectus*.

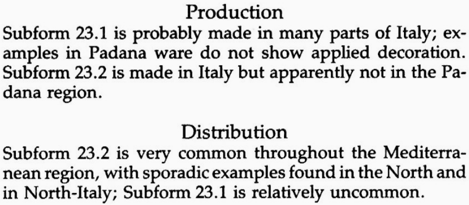
Oh dear!

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.

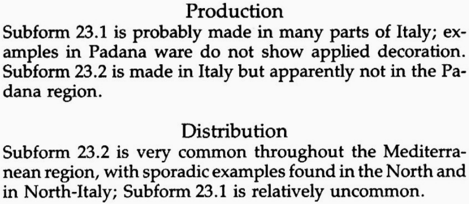
UNSTRUCTURED DATA

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.

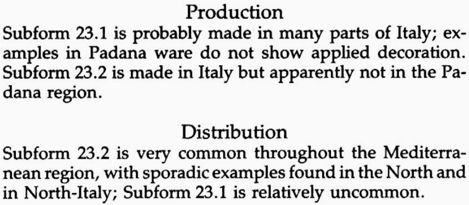
UNSTRUCTURED DATA

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

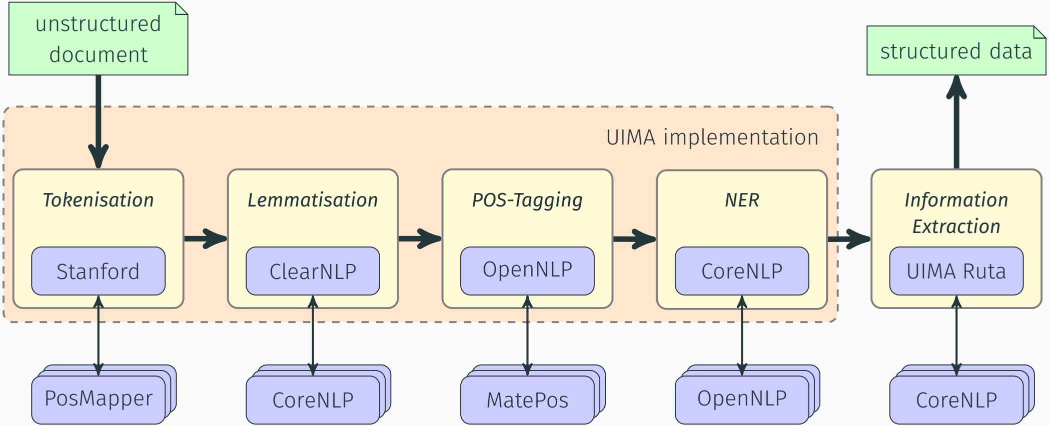
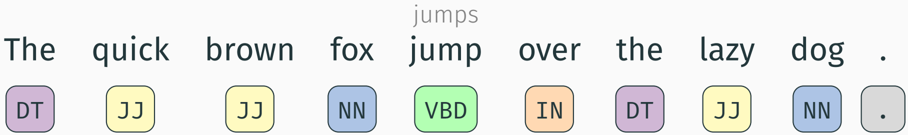


Figure 2: IE Process Pipeline.

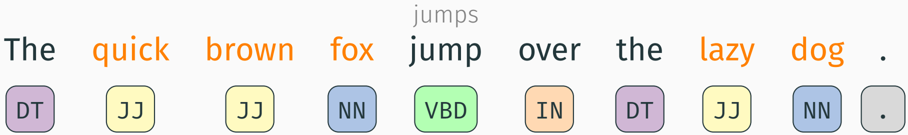
POS-Tagging



				jumps					
The	quick	brown	fox	jump	over	the	lazy	dog	.
DT	JJ	JJ	NN	VBD	IN	DT	JJ	NN	.

Figure 3: POS-tagging examples after lemmatisation.

POS-Tagging



				jumps					
The	quick	brown	fox	jump	over	the	lazy	dog	.
DT	JJ	JJ	NN	VBD	IN	DT	JJ	NN	.

Figure 3: POS-tagging examples after lemmatisation.

Adapting the NER

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

PERSON
ORGANIZATION
LOCATION
PERCENT
DATE
TIME
MONEY
MISC

So we have to add the custom entity type FORM.

Two approaches for NER

Rule-based approach

- High precision, but lower recall

  ⇒ Many many rules?!

Machine-learning approach

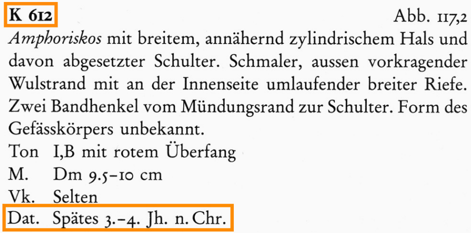
- Lower precision, but high recall
- Needs to be trained!

Two approaches for NER

Rule-based approach

- High precision, but lower recall

  ⇒ 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 Gempeler, *Elephantine X*.

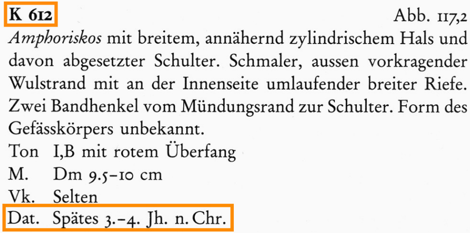
Machine-learning approach

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

Two approaches for NER

Rule-based approach

- High precision, but lower recall
  ⇒ 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 Gempeler, *Elephantine X*.

Machine-learning approach

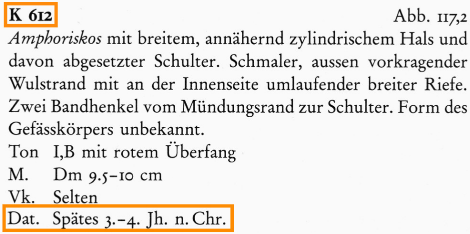
- Lower precision, but high recall
- Needs to be trained!

Two approaches for NER

Rule-based approach

- High precision, but lower recall

  ⇒ Many many rules?!



K 612 Abb. 117,2

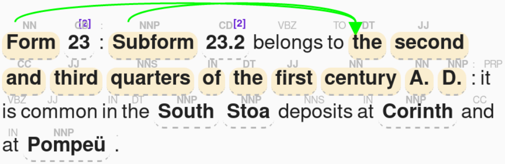
Amphoriskos mit breitem, annähernd zylindrischem Hals und davon abgesetzter Schulter. Schmaler, aussen vorkragender Wulstrand mit an der Innenseite umlaufender breiter Riefe. Zwei Bandhenkel vom Mündungsrand zur Schulter. Form des Gefässkörpers unbekannt.

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 Gempeler, *Elephantine X*.

Machine-learning approach

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



Form 23 : 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 Pompeü .

Figure 5: Manually annotated sentence from Ettlinger, *Conspectus* in `iepy`.

Temporal Expressions

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

around 140 B.C.	⟼	APPROX BC0140
Spätes 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

⇒ 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`

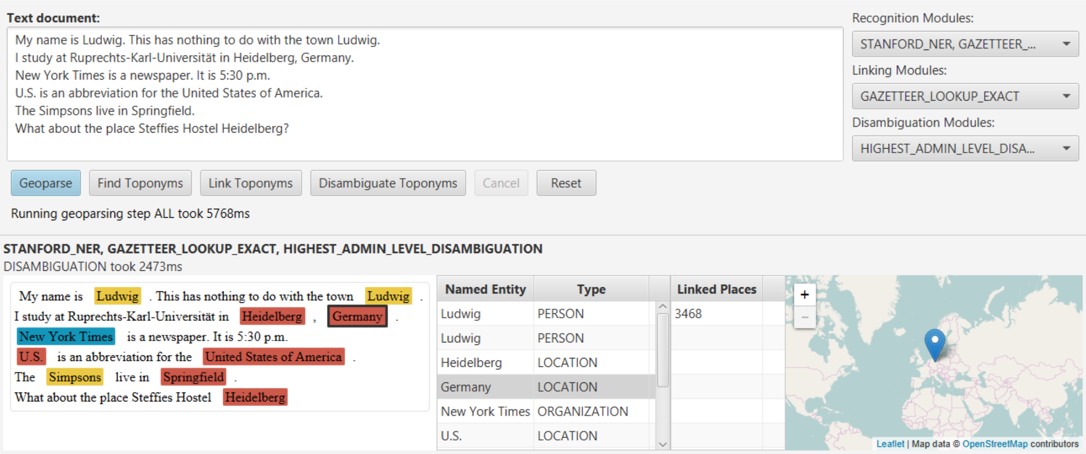
⇒ e.g.

```
{  "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



Text document:

My name is Ludwig. This has nothing to do with the town Ludwig.
I study at Ruprechts-Karl-Universität in Heidelberg, Germany.
New York Times is a newspaper. It is 5:30 p.m.
U.S. is an abbreviation for the United States of America.
The Simpsons live in Springfield.
What about the place Steffies Hostel Heidelberg?

Recognition Modules: STANFORD_NER, GAZETTEER_...

Linking Modules: GAZETTEER_LOOKUP_EXACT

Disambiguation Modules: HIGHEST_ADMIN_LEVEL_DISA...

Running geoparsing step ALL took 5768ms

STANFORD_NER, GAZETTEER_LOOKUP_EXACT, HIGHEST_ADMIN_LEVEL_DISAMBIGUATION
DISAMBIGUATION took 2473ms

My name is Ludwig . This has nothing to do with the town Ludwig .
I study at Ruprechts-Karl-Universität in Heidelberg , Germany .
New York Times is a newspaper. It is 5:30 p.m.
U.S. is an abbreviation for the United States of America .
The Simpsons live in Springfield .
What about the place Steffies Hostel Heidelberg

Named Entity	Type	Linked Places
Ludwig	PERSON	3468
Ludwig	PERSON	
Heidelberg	LOCATION	
Germany	LOCATION	
New York Times	ORGANIZATION	
U.S.	LOCATION	

Figure 6: Screenshot of HeidelPlace.

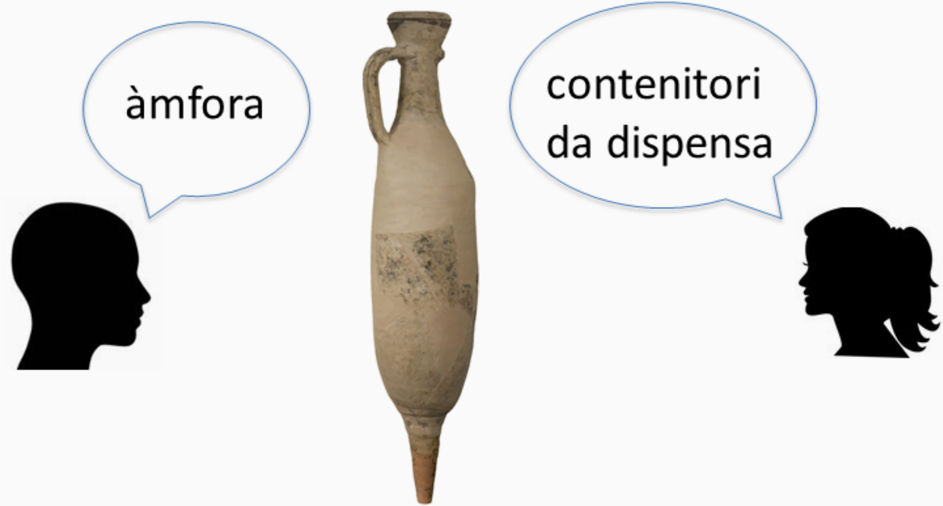
MULTILINGUALISM

Background

Two problems:

- Linguistic
- Conceptual

Different languages



Different traditions

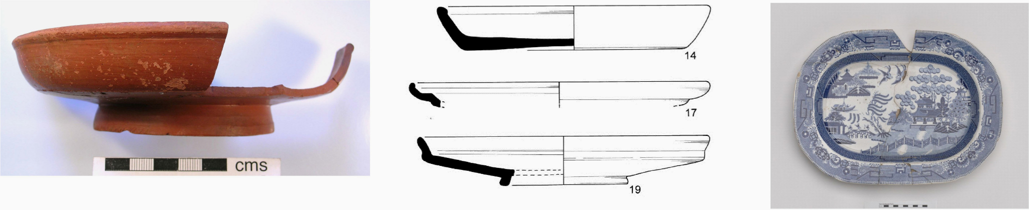


Figure 7: Plate, platter or dish?

Creating controlled vocabularies

- Sherd type (e.g. rim)
- Form (e.g. plate)
- Decoration form (e.g. burnished)
- Decoration color (e.g. yellow)
- Fabric (e.g. bla)

Lessons from ARAIDNE

Using tools developed for the ARIADNE project by the Hypermedia Research Group at the University of South Wales

Creation of a neutral spine based on the Getty Institute’s Art and Architecture Thesaurus (AAT)

more bla

more bla

more bla

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.** ⟼ `XXXX-Q3`; `00`

References

- Ettlinger, Elisabeth. *Conspectus formarum terrae sigillatae Italico modo confectae.* Ed. by Deutsches Archäologisches Institut zu Frankfurt and Römisch-Germanische Kommission. Materialien zur römisch-germanischen Keramik. Bonn: Habelt, 1990.
- Gempeler, Robert D. *Elephantine X. Die Keramik römischer bis früharabischer Zeit.* Mainz: Von Zabern, 1992.
- Indurkhya, Nitin and Fred J. Damerau. *Handbook of Natural Language Processing.* 2nd. Chapman & Hall/CRC, 2010.

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 Kußmaul[2]

23rd Annual Meeting EAA, Maastricht

31 August 2017

[1]Archaeology Data Service, University of York

[2]Archaeological Institute, University of Cologne

ARCHAIDE
ARCHAEOLOGICAL
AUTOMATIC INTERPRETATION
AND DOCUMENTATION
OF CERAMICS