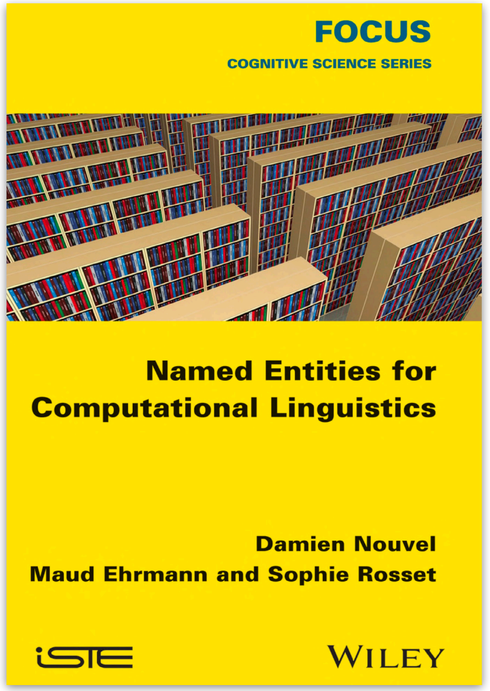
NER and Wikidata

February 22, 2024

SunoikisisDC Spring 2024
Digital Approaches to Cultural Heritage

Monica Berti
monica.berti@uni-leipzig.de



Named Entity Recognition (NER) is a task of information extraction that aims at finding mentions of named entities in the text and classify their types into categories corresponding to proper names and quantities of interest, such as people, places, organizations, time expressions, monetary amounts, percentages, etc.

Named-entity recognition

From Wikipedia, the free encyclopedia

"Named entities" redirects here. For HTML, XML, and SGML named entities, see List of XML and HTML character entity references.

Named-entity recognition (**NER**) (also known as **(named) entity identification**, **entity chunking**, and **entity extraction**) is a subtask of information extraction that seeks to locate and classify named entities mentioned in unstructured text into pre-defined categories such as person names, organizations, locations, medical codes, time expressions, quantities, monetary values, percentages, etc.

Most research on NER/NEE systems has been structured as taking an unannotated block of text, such as this one:

> Jim bought 300 shares of Acme Corp. in 2006.

And producing an annotated block of text that highlights the names of entities:

> $[\text{Jim}]_{\text{Person}}$ bought 300 shares of $[\text{Acme Corp.}]_{\text{Organization}}$ in $[\text{2006}]_{\text{Time}}$.

In this example, a person name consisting of one token, a two-token company name and a temporal expression have been detected and classified.

State-of-the-art NER systems for English produce near-human performance. For example, the best system entering MUC-7 scored 93.39% of F-measure while human annotators scored 97.60% and 96.95%.[1][2]

Named-entity recognition platforms [edit]

Notable NER platforms include:

- GATE supports NER across many languages and domains out of the box, usable via a graphical interface and a Java API.
- OpenNLP includes rule-based and statistical named-entity recognition.
- SpaCy features fast statistical NER as well as an open-source named-entity visualizer.
- Transformers features token classification using deep learning models.[3][4]

NER for Ancient Greek and Latin

State of the art

- **Herodotos Project** NER Annotation and Tagger (Latin)
  - https://github.com/alexerdmann/Herodotos-Project-Latin-NER-Tagger-Annotation
  - https://www.aclweb.org/anthology/W16-4012

- **Recogito** (beta support for Latin)
  - https://recogito.pelagios.org

- **Classical Language Toolkit** (CLTK) (Greek and Latin)
  - https://docs.cltk.org/en/latest/cltk.ner.html
  - "[…] a lack of annotated texts and robust language models underlies the problem." (P. J. Burns)

- **SunoikisisDC**
  - https://github.com/SunoikisisDC/SunoikisisDC-2016-2017/wiki/Named-Entity-Extraction-I
  - https://github.com/SunoikisisDC/SunoikisisDC-2016-2017/wiki/Named-Entity-Extraction-II
  - https://github.com/SunoikisisDC/SunoikisisDC-2022-2023/wiki/Using-and-Editing-Wikidata

- **Digital Athenaeus Project** (Greek)
  - BookStream (indexTOtext): http://www.digitalathenaeus.org/tools/Book-Stream/
  - Named Entities Digger: http://www.digitalathenaeus.org/tools/KaibelText/named_entities_digger.php
  - Named Entities Concordance: http://www.digitalathenaeus.org/tools/KaibelText/named_entities_concordance.php
  - NER Based Catalog: https://www.digitalathenaeus.org/tools/Catalog/

- **LAGL - Linked Ancient Greek and Latin** (Greek)
  - NER spaCy based: https://www.lagl.org/entity_recognizer.php
  - Digital Harpocration: https://www.lagl.org/tools/harpocration/

Recogito | AN INITIATIVE OF Pelagios commons

Named Entity Recognition (NER)

NER identifies places and persons in your text automatically. Depending on the length of your text, this may take a while.

Recognition Engines

◉ **Example Kima NER Plugin**	he	An attempt to use Kima with the Recogito NER plugin interface.
○ **Stanford CoreNLP**	en	The standard engine with the default English language model
○ **Stanford CoreNLP**	fr	The standard engine with the default French language model
○ **Stanford CoreNLP**	de	The standard engine with the default German language model
○ **Stanford CoreNLP**	es	The standard engine with the default Spanish language model
○ **Herodotus Latin NER**		An experimental Latin NER plugin by Alex Erdmann

Note: different engines can provide different results and are generally optimized for a specific language.

Authority Files

☑ identify entities against all available authority files

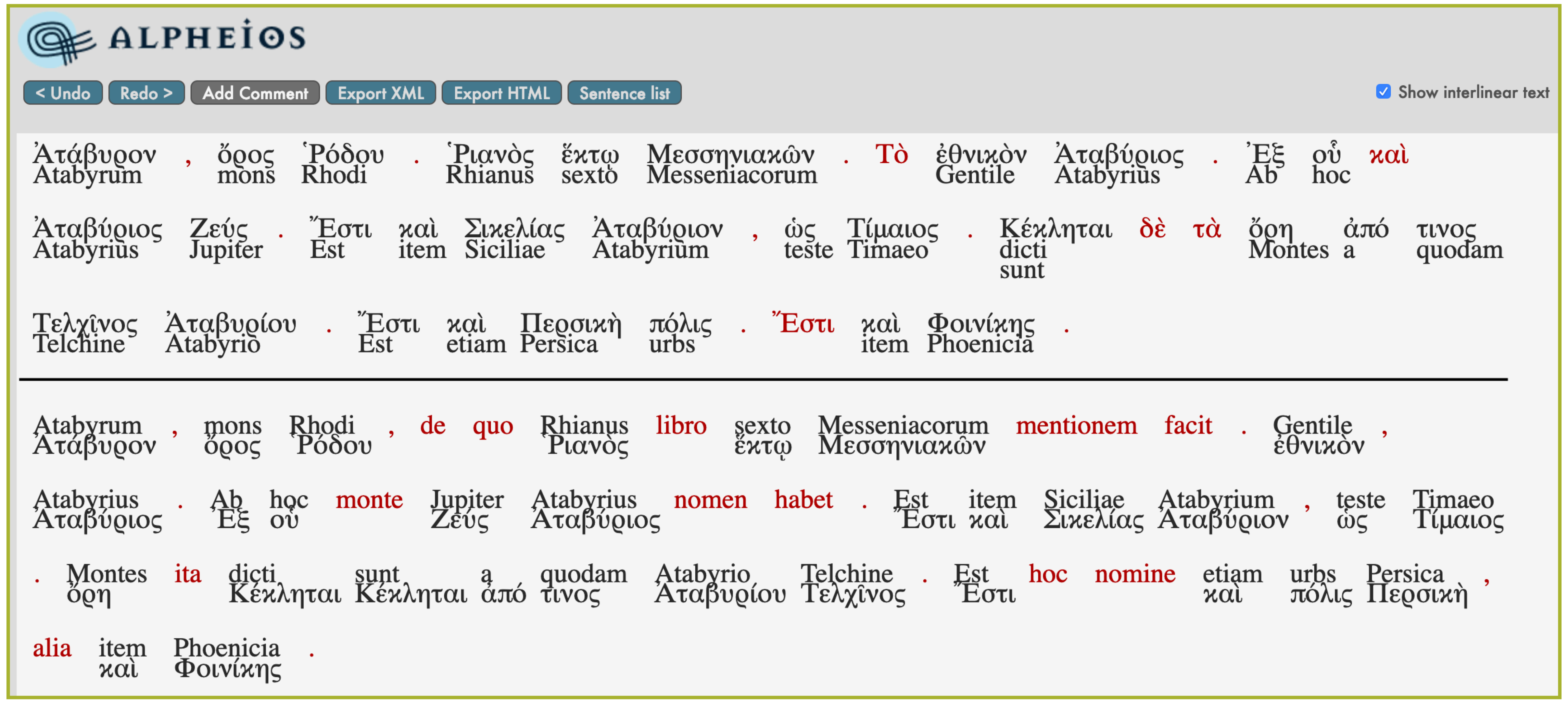
☑ Pleiades	Pleiades Gazetteer of the Ancient World
☑ CHGIS	China Historical GIS
☑ DPP Places	Places from the Digitizing Patterns of Power project
☑ DARE	Digital Atlas of the Roman Empire
☑ MoEML	Map of Early Modern London
☑ HGIS de las Indias	Historical-Geographic Information System for Spanish America (1701-1808)
☑ GeoNames	A subset of GeoNames populated places, countries and first-level administrative divisions
☑ Kima	Kima Historical Gazetteer - place names in the Hebrew script

Start NER | Cancel

State of the art

- **Digital Fragmenta Historicorum Graecorum** (DFHG) (Greek and Latin)
  - http://www.dfhg-project.org

Ἀτάβυρον, ὄρος Ῥόδου. Ῥιανὸς ἕκτῳ Μεσσηνιακῶν. Τὸ ἐθνικὸν Ἀταβύριος. Ἐξ οὗ καὶ Ἀταβύριος Ζεύς. Ἔστι καὶ Σικελίας Ἀταβύριον, ὡς Τίμαιος. Κέκληται δὲ τὰ ὄρη ἀπό τινος Τελχῖνος Ἀταβυρίου. Ἔστι καὶ Περσικὴ πόλις. Ἔστι καὶ Φοινίκης.	Atabyrum, mons Rhodi, de quo Rhianus libro sexto Messeniacorum mentionem facit. Gentile, Atabyrius. Ab hoc monte Jupiter Atabyrius nomen habet. Est item Siciliae Atabyrium, teste Timaeo. Montes ita dicti sunt a quodam Atabyrio Telchine. Est hoc nomine etiam urbs Persica, alia item Phoenicia.



FHG 1, Timaeus fr. 3 (= Stephanus Byzantinus, s.v. Ἀτάβυρον)

Automatic Translation Alignment for Ancient Greek and Latin

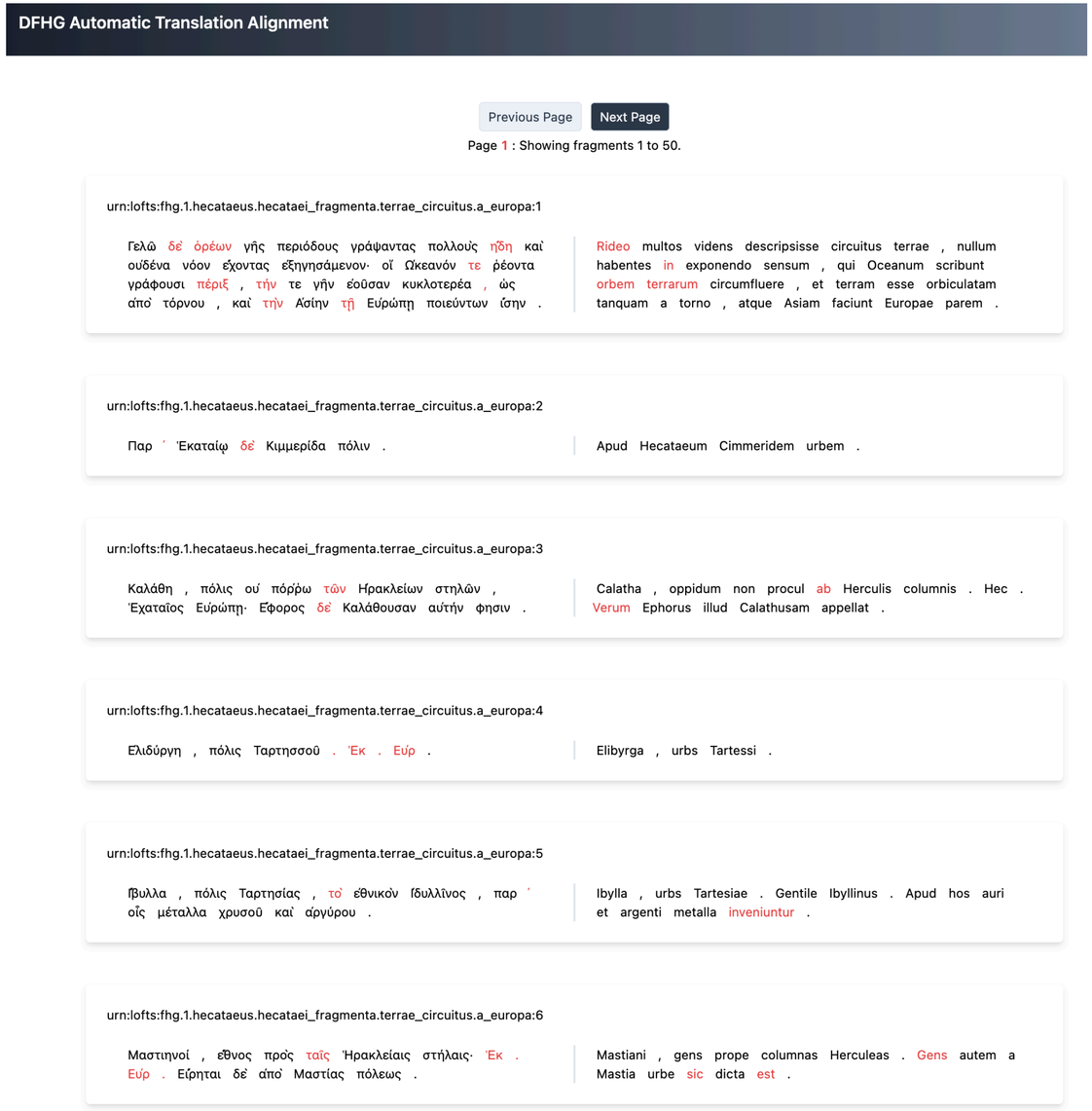
Tariq Yousef, Chiara Palladino, David J. Wright, Monica Berti

Abstract

This paper presents the results of automatic translation alignment experiments on a corpus of texts in Ancient Greek translated into Latin. We used a state-of-the-art alignment workflow based on a contextualized multilingual language model that is fine-tuned on the alignment task for Ancient Greek and Latin. The performance of the alignment model is evaluated on an alignment gold standard consisting of 100 parallel fragments aligned manually by two domain experts, with a 90.5% Inter-Annotator-Agreement (IAA). An interactive online interface is provided to enable users to explore the aligned fragments collection and examine the alignment model's output.

PDF | Cite | Search | Code

Anthology ID: 2022.lt4hala-1.14
Volume: Proceedings of the Second Workshop on Language Technologies for Historical and Ancient Languages
Month: June
Year: 2022
Address: Marseille, France
Editors: Rachele Sprugnoli, Marco Passarotti
Venue: LT4HALA
SIG: –
Publisher: European Language Resources Association
Note: –
Pages: 101–107
Language: –
URL: https://aclanthology.org/2022.lt4hala-1.14
DOI: –
Bibkey: `yousef-etal-2022-automatic-translation`
Cite (ACL): Tariq Yousef, Chiara Palladino, David J. Wright, and Monica Berti. 2022. Automatic Translation Alignment for Ancient Greek and Latin. In *Proceedings of the Second Workshop on Language Technologies for Historical and Ancient Languages*, pages 101–107, Marseille, France. European Language Resources Association.
Cite (Informal): Automatic Translation Alignment for Ancient Greek and Latin (Yousef et al., LT4HALA 2022)
Copy Citation: BibTeX | Markdown | MODS XML | Endnote | More options...
PDF: https://aclanthology.org/2022.lt4hala-1.14.pdf
Code ugaritalignment/alignment-gold-standards

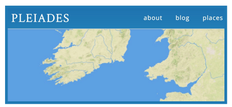


https://ugarit.ialigner.com/dfhg/

https://github.com/ugaritalignment/alignment-gold-standards

Authority Lists and Knowledge Bases

- **Lexicon of Greek Personal Names** (LGPN): http://www.lgpn.ox.ac.uk



- **Pleiades** Gazetteer: https://pleiades.stoa.org

- **Trismegistos People**: https://www.trismegistos.org/ref/

- **Standards for Networking Ancient Prosopographies**: https://snapdrgn.net

- **VIAF**: https://viaf.org

- **DBpedia**: https://wiki.dbpedia.org

- **Wikidata**: https://www.wikidata.org

VIAF
Virtual International Authority File

Search

Select Field: All Headings | Select Index: All VIAF | Search Terms: Αρχέστρατος, Αρχαίος Έλληνας ποιητής | Search

Αρχέστρατος, Αρχαίος Έλληνας ποιητής

1 heading found for *Αρχέστρατος, Αρχαίος Έλληνας ποιητής*

	Heading	Type	Sample Title
1	Archestratus, of Gela	Personal	Archestrato di Gela, 1983- :
	Archestratus.		Archestrato di Gela, 1983- :
	Archestrate		
	Archestrate, 03..-03.. av. J.-C.		Archestrato di Gela, 1983- :
	Archestratus, Gelensis		Archestrato di Gela, 1983- :
	Arquèstrat		Hēdypatheia
	Archestratus Gelous ca. v370		I frammenti della Gastronomia / Archestrato ; raccolti e volgarizzati da Domenic ...
	Archestratus Gelous, sec. IV a.C.		Archestrato di Gela, 1983- :

https://viaf.org/viaf/search?query=local.names all "Αρχέστρατος

https://www.wikidata.org/

Archestratus (Q210507)

Item Discussion
Read View history

ancient Greek poet (4th century BCE)
Archestratos | Archestratus of Gela

edit

In more languages
Configure

Language	Label	Description	Also known as
English	Archestratus	ancient Greek poet (4th century BCE)	Archestratos Archestratus of Gela
American English	No label defined	No description defined	
Italian	Archestrato di Gela	poeta greco antico	Archestrato
German	Archestratos von Gela	griechischer Dichter	Archestratos

All entered languages

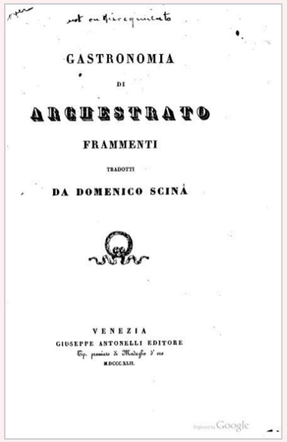
Statements

instance of — human — edit

1 reference

+ add value

image — edit



Gastronomia.djvu
1,781 × 2,800, 20 pages; 431 KB

Main menu hide
Main page
Community portal
Project chat
Create a new Item
Recent changes
Random Item
Query Service
Nearby
Help
Donate
Switch to old look
Lexicographical data
Create a new Lexeme
Recent changes
Random Lexeme

Tools hide
Actions
Merge with...
Select for merging
General
What links here
Related changes
Special pages
Permanent link
Page information
Concept URI
Cite this page
Get shortened URL
Download QR code
Reasonator

https://www.wikidata.org/wiki/Q210507

Suda Lexicon | 1_alpha_03916.json

1-6 / 6 sentences [doc 3 / 14]

1 urn:cts:greekLit:tlg1188 | Aristocles of Messene
PER Ἀριστοκλῆς | LOCderiv Μεσσήνιος | LOC Ἰταλία | ORGderiv Περιπατητικός
Ἀριστοκλῆς, Μεσσήνιος τῆς Ἰταλίας, φιλόσοφος Περιπατητικός.

2 urn:cts:greekLit:tlg1188.suda001
συνέταξε Περὶ φιλοσοφίας βιβλία ι'·

3 urn:cts:greekLit:tlg0012 | Homer PER Ὅμηρος | urn:cts:greekLit:tlg0059 | Plato PER Πλάτων
Πότερος σπουδαιότερος Ὅμηρος ἢ Πλάτων .

4 καταλέγει δὲ ἐν τούτοις πάντας φιλοσόφους καὶ δόξας αὐτῶν.

5 urn:cts:greekLit:tlg1188.suda002 OTH τέχνη | urn:cts:greekLit:tlg1188.suda003 PER Σάραπις
ἔγραψε δὲ καὶ Τέχνας ῥητορικὰς , Περὶ Σαράπιδος .

6 urn:cts:greekLit:tlg1188.suda003 OTH ἠθικός
Ἠθικὰ βιβλία θ ' .

Layer: Catalog

Span — Delete | Clear

Layer: Catalog

Text: Ἀριστοκλῆς, Μεσσήνιος τῆς Ἰταλίας

No links or relations connect to this annotation.

Author: urn:cts:greekLit:tlg1188

author_identifier: Aristocles o...

1st-century AD Greek philosopher

Work:

work_identifier:

IDEA

International (Digital) Dura-Europos Archive

About History Gazetteer Digital Accessibility

About

IDEA is a major Digital Humanities initiative funded by the US's National Endowment for the Humanities, and aimed at digitally re-integrating dispersed collections and discipline-specific knowledge related to the important cultural heritage site of Dura-Europos (Syria).

The project's re-assembly and re-contextualization efforts are driven by Linked Open Data (LOD) methods and critical archival studies. Central to IDEA's approach is an exploration of how new technologies can be put to use for ethical collaborative interventions related to legacy archeological archives.

Imbalances in access and inclusion are among the most pressing problems facing the cultural heritage and archival sectors today. Blockbuster 'Big Digs' of the early 20th century (like Dura-Europos) have filled collections and textbooks in the West.

NATIONAL ENDOWMENT FOR THE HUMANITIES

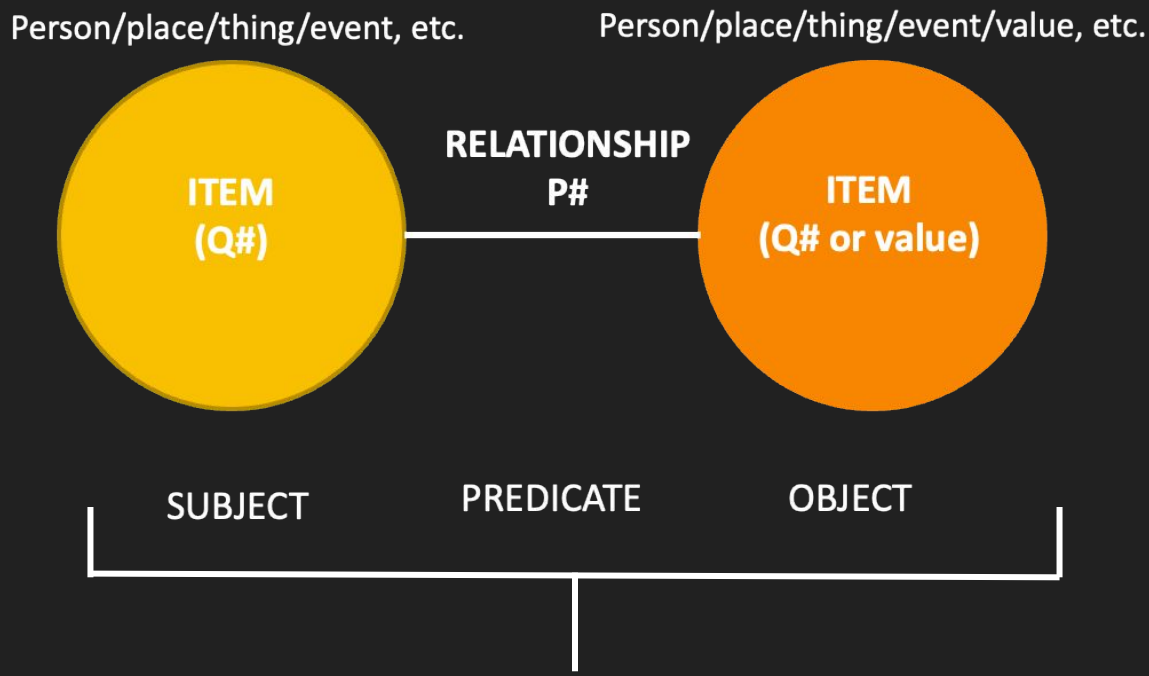
1. digitally reassemble dispersed collections
2. provide context
3. improve inequalities in access

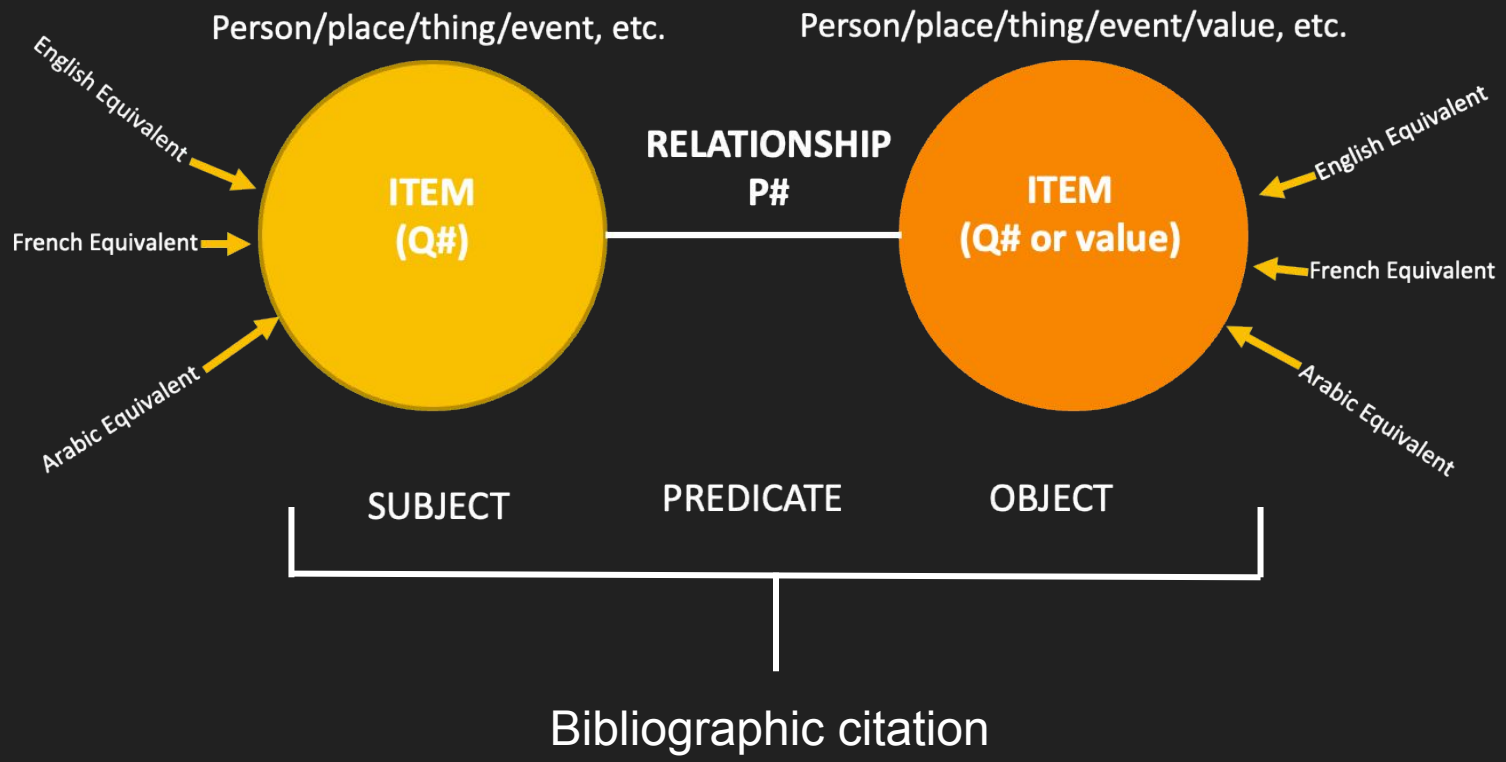
Duraeuroposarchive.org

- Modern Syria; West bank of the Euphrates;
- Seleucid foundation c. 300 BCE, besieged + conquered by the Sasanians c. 256 CE

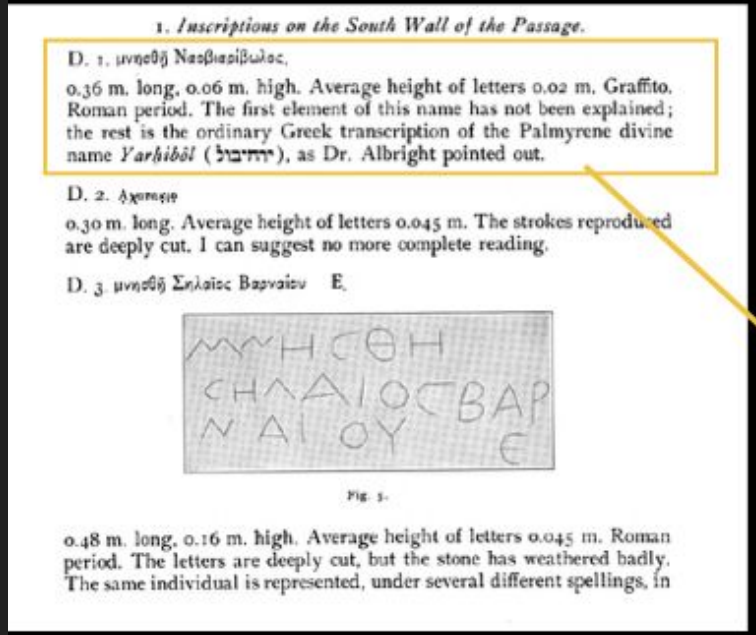
Distribution of Dura-Europos Artifacts

What kinds of content is IDEA dealing with?


Person/place/thing/event, etc.
Person/place/thing/event/value, etc.
ITEM
(Q#)
RELATIONSHIP
P#
ITEM
(Q# or value)
SUBJECT
PREDICATE
OBJECT
Bibliographic citation


Person/place/thing/event, etc.
Person/place/thing/event/value, etc.
English Equivalent
French Equivalent
Arabic Equivalent
ITEM
(Q#)
RELATIONSHIP
P#
ITEM
(Q# or value)
English Equivalent
French Equivalent
Arabic Equivalent
SUBJECT
PREDICATE
OBJECT
Bibliographic citation

DATA EMBEDDED IN TEXT
(old publications, field notebooks, etc.)



1. *Inscriptions on the South Wall of the Passage.*

D. 1. μνησθῆ Νασβιαρίβωλος.

0.36 m. long, 0.06 m. high. Average height of letters 0.02 m. Graffito. Roman period. The first element of this name has not been explained; the rest is the ordinary Greek transcription of the Palmyrene divine name *Yarḥibôl* (ירחיבול), as Dr. Albright pointed out.

D. 2. Ἀχαιρσιο

0.30 m. long. Average height of letters 0.045 m. The strokes reproduced are deeply cut. I can suggest no more complete reading.

D. 3. μνησθῆ Σηλαίοις Βαρναίου E.

Fig. 5.

0.48 m. long, 0.16 m. high. Average height of letters 0.045 m. Roman period. The letters are deeply cut, but the stone has weathered badly. The same individual is represented, under several different spellings, in

Plain language pseudo-*statement*
SUBJECT – PREDICATE – OBJECT
(ie. ITEM – RELATIONSHIP – ITEM/VALUE)

Example pseudo-statements:
Inscription D1 — was discovered — on the Palmyrene Gate

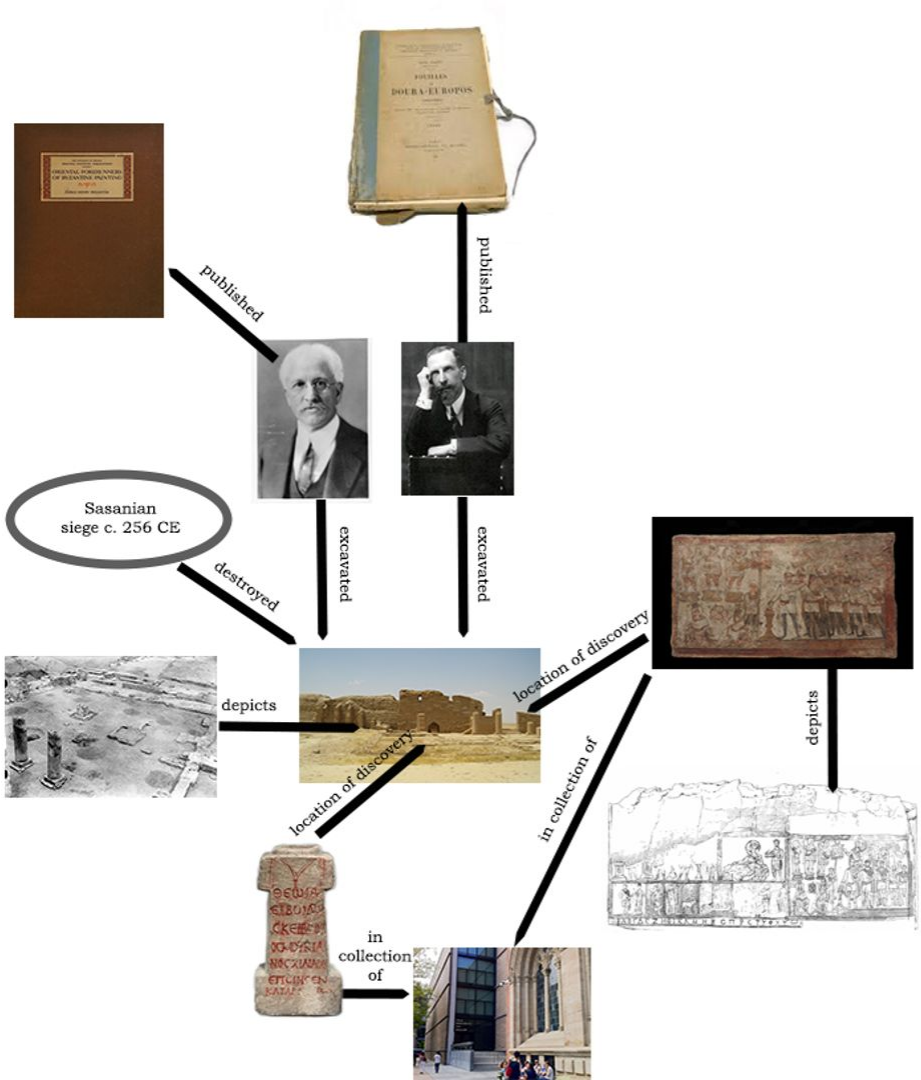
Inscription D1 — has a measurement — 0.36 m long

Inscription D1 — has a measurement — 0.06 m high

Inscription D1 — is written in — ancient Greek

Creating and networking WD records for:

People
Places
Things (artifacts, archival docs)
Events
Bibliography



Archaeological knowledge as a network of relationships among:

People
Places
Things (artifacts, archival docs)
Events
Bibliography

WIKIDATA

Hexagonal altar with inscription to the Gad of Dura, Yale University Art Gallery, inv. 1938.5999.3084 and 1929.372 (Q100348705)

Main page
Community portal
Project chat
Create a new Item
Recent changes
Random Item
Query Service
Nearby
Help
Donate

Lexicographical data
Create a new Lexeme
Recent changes
Random Lexeme

Tools

YUAG 89446 and YUAG 3507. Two fragments of a hexagonal altar excavated in Dura-Europos by the Yale-French team, 1928-1937, Syria edit

Most relevant properties which are absent [Help with translations]

In more languages

Configure

Language	Label	Description	Also known as
British English	No label defined	No description defined	
English	Hexagonal altar with inscription to the Gad of Dura, Yale University Art Gallery, inv. 1938.5999.3084 and 1929.372	YUAG 89446 and YUAG 3507. Two fragments of a hexagonal altar excavated in Dura-Europos by the Yale-French team, 1928-1937, Syria	PR1, 45 no. 2 PAT 1079 Doura 13 SEG 7, 664
Arabic	مذبح سداسي منقوش عليه نقش للإله جاد دورا، معرض فنون جامعة ييل، رقم الجرد 1938.5999.3084 و1929.372	معرض فنون جامعة ييل، رقم 89446 و3507. قطعة أثرية اكتشفت في دورا أوروبوس من قبل بعثة جامعة ييل - الفرنسية المشتركة، 1928-1937، سوريا	