Aggregating Classical Data Sets with Linked Data

David Scott and Mike Jackson

Who is Involved

- Centre for eResearch, KCL.
- EPCC, The University of Edinburgh.
- Humboldt Technical University, Berlin.
- Funded by JISC as part of the Managing Research Data strand.

Duration: August 2010 to July 2011

Aims

- LongTerm: Looking for a way to
  - Facilitate searches
  - Enable new ways of exploring the data.
- In order to allow classicists/epigraphers to establish links between inscriptions more easily.

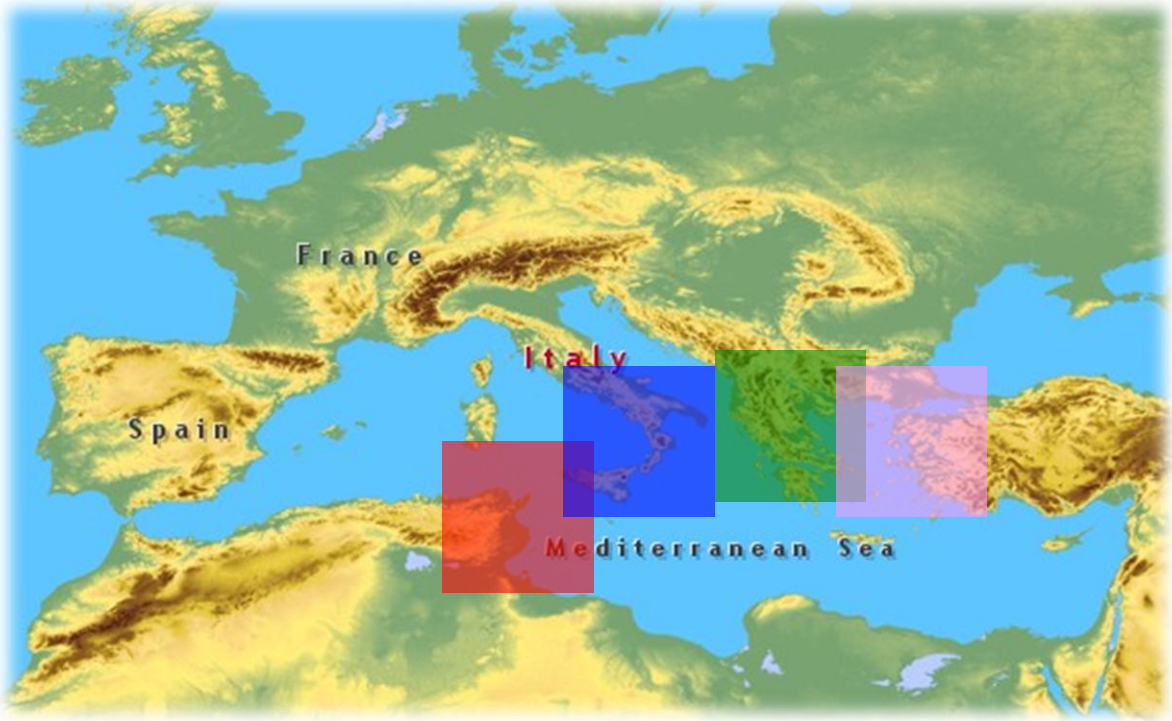
- Current: Conduct a preliminary evaluation of a linked-data approach.

Challenges

- Epigraphers are not computer scientists – they cannot be expected to program.
- Complex, highly-interactive workflow which is difficult to predict:
  - Depends on the specific researcher and his/herquestion.
  - Answer determines his/her direction and next questions.
- May be looking for links between objects such as common properties - searching/exploring.
- Provenance and peer assessment/opinions are important - annotation.

Vision: From Islands of Data …

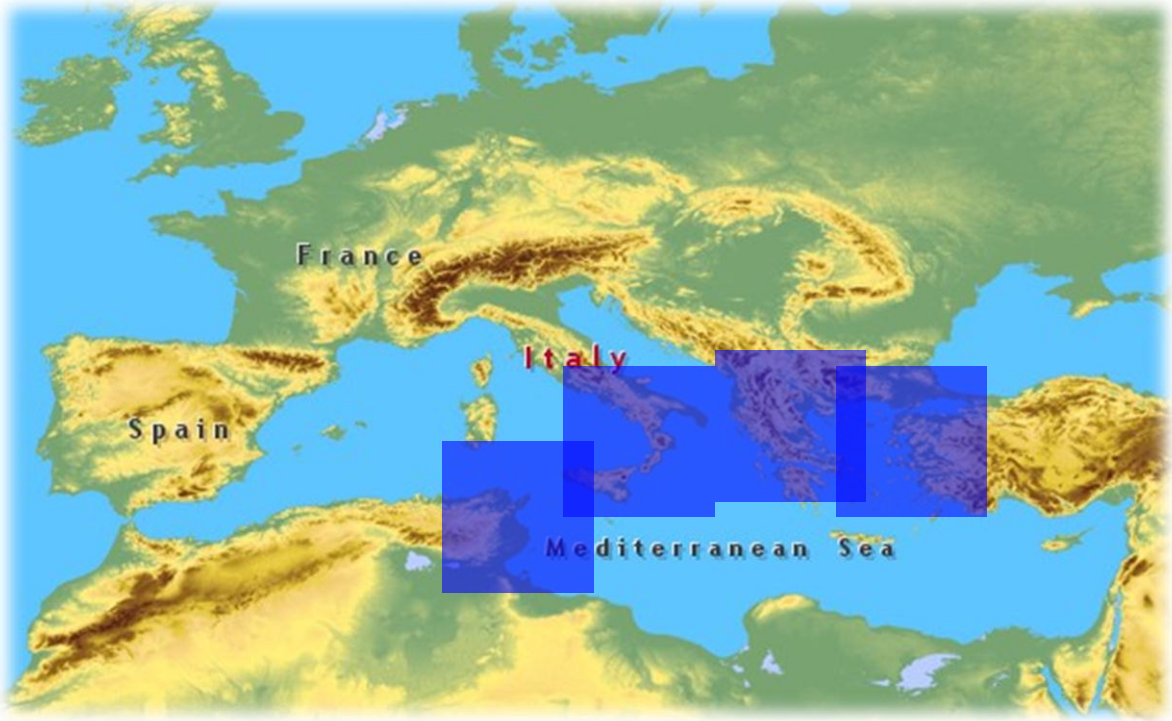
- Data is distributed.
- Diverse representations by distributed individuals and groups e.g. XML, Word, Access,…
- Different ways of access e.g. bulk download, online web form, download text file,…
- But these data sets are conceptually related, overlap in place, or time or people.



Map from http://www.free-extras.com/images/mediterranean_sea_map-12019.htm

… to a Unified Data Landscape

- Want to view data as part of a single data landscape e.g. a single virtual data source.
- So that it can be searched/browsed as a whole.
- Whole is greater than sum of parts.



Map from http://www.free-extras.com/images/mediterranean_sea_map-12019.htm

Obstacles

- Data are complex
- Same thing may be named in different ways
- Some data may be fuzzy or incomplete
- Some data may be uncertain
- Some data may be or implicit or open to interpretation
- Some data may be erroneous

Scope

- Grand vision but must narrow the scope for this small project.
- Not building a system but examining and evaluating some possibilities.
- Restrict attention to data encoded in EpiDoc.
- EpiDoc – flexible but rigorous standards/tools for digital encoding and interchange of ancient texts.

The Data

- Inscriptions
  - Aphrodisias (Inscriptions of Aphrodisias), Turkey
    - ~1,500
  - Tripolitana (Inscriptions of Roman Triploitania), Libya
    - ~1,000
  - HGV (Heidelberger Gesamtverzeichnis der Griechischen Papyrusurkunden Ägyptens), Egypt
    - ~55,000
- Encoded in a form of XML called EpiDoc.
- One XML document per inscription.

Information In the EpiDoc

- Provenance
- Description
- Date
- Language
- Edited text (with annotations)
- Translation
- Findspot (ancient and modern)
- Material from which constructed

This is not a complete list and not everything need be present.

Examples of Descriptions of Materials

<rs type="material">White marbl</rs>

May have to cope with mis-spellings.

<material>stuccoed sandstone</material>

Note different formats.

When we re-encode the data we strive for uniformity.

Example of a Description

Upper right corner of a <rs type="material">white marble</rs> <rs type="objectType">block</rs> (<measure dim="width" type="length" unit="metre">0.36</measure> x <measure dim="height" type="length" unit="metre">0.24</measure> x <measure dim="depth" type="length" unit="metre">0.34</measure>).

Note embedded description of the material.

Examples of Dates

```
<date notAfter="-0001" notBefore="-0033" exact="none">Late
    first century B.C.</date>

<origDate notBefore="0101" notAfter="0188" precision="low"
    evidence="lettering">Second to early third centuries A.D.
    (lettering)</origDate>
```

Note different formats.

Example of an Edited Text

```
<persName type="aphrodisian" full="yes"><name reg="Γάϊος">Γάϊος</name>  <name
    reg="Ἰούλιος">Ἰούλιος</name>  <name reg="Ζωΐλος" type="       ">Ζώ<supplied
    reason="lost">ι</supplied>λο<unclear
    reason="damage">ς</unclear></name></persName> <w lemma="ὁ">ὁ</w> <w
    lemma="ἱερεύς">ἱερεὺς</w> <w lemma="θεός">θ<unclear
    reason="damage">εο</unclear>ῦ  </w><persName type="divine" full="yes"><name
    reg="Ἀφροδίτη">Ἀφροδείτη<supplied
    reason="lost">ς</supplied></name></persName><lb n="2"/><space unit="character"
    dim="horizontal"/> <w lemma="σωτήρ">σωτὴρ</w> <w lemma="καί">καὶ</w> <w
    lemma="εὐεργέτης">εὐεργέτης</w> <w lemma="ὁ">τῆς</w> <w
    lemma="πατρίς">πατρίδος</w> <space extent="1" unit="character" dim="horizontal"/>
    <w lemma="ὁ">τὸ</w> <w lemma="ἱερός">ἰ<unclear
    reason="damage">ερ</unclear>ὸ<unclear reason="damage">ν</unclear></w>
    <persName type="divine" full="yes"><name reg="Ἀφροδίτη">Ἀφροδ<unclear
    reason="damage">εί</unclear>τῃ</name></persName>
```

Some Annotations in the Text

Text often includes names of people and/or places.

<persName type="divine" full="yes"><name reg="Ἀφροδίτη">Ἀφροδ<unclear reason="damage">εί</unclear>τῃ</name></persName>

Translation of the Text

C. Julius Zoilos, priest of the god Aphrodite, saviour and benefactor of his country <supplied cert="low" reason="subaudible">gave</supplied> the sanctuary to Aphrodite.

Findspots

IAph

<placeName type="ancientFindspot"
key="Aphrodisias">Aphrodisias</placeName>

<placeName type="modernFindspot" key="Geyre">Geyre</placeName>

IRT

<placeName type="ancientFindspot” ref="http://atlantides.org/batlas/abrotonum-sabratha-35-e2"
key="db659">Sabratha</placeName>

<placeName type="modernFindspot"
key="http://www.geonames.org/2208578/marsa-zawaghah.html">Marsa Zawaghah</placeName>

Search Problems

Suppose we want to identify all inscriptions that either

- mention a particular emperor, or
- mention a particular place, or
- were found in a particular spot?

Simple text search (using, e.g., grep) does not work well as it picks out whole lines, but we can try other, less direct, approaches which may involve reformatting the data.

LaQuAT

- SPQR arose out of the JISC-funded LaQuAT project – linking and querying of ancient texts.
- LaQuAT used a relational representation.
- Data stored as tables.
  - One row per inscription.
  - Columns contain fields or facts about the data.

LaQuAT and Relational Data

ID	Object	NotAfter	NotBefore	Ancient	…	…
154	block	14	-50	Aphrodisias		
155	lintel	-1	-33	Antiochia		
156	column	1000	1	Aphrodisias		
157	column	1	33	Antiochia		

- If want to find related facts e.g. which inscriptions name the same people or places then need to visually inspect whole table or run a query.
- More complex if data in multiple tables. And in multiple databases where column names are different.
- OGSA-DAI allowed tables in multiple databases to appear as if they're in the same table and to rename columns to have common names.
- But it was still difficult to query the data or even just to explore it in an intuitive way.

LaQuAT – Queries

SELECT ID, Place, Date, Title FROM Inscriptions WHERE Place LIKE “%Antiochia%”;

SELECT ID, Place, Date, Title FROM Inscriptions WHERE Place LIKE “%Antiochia%”

UNION

SELECT ID, Place, Date, Title FROM MoreInscriptions WHERE Place LIKE “%Antiochia%”;

Not appealing to the non-programmer.

Linked Data

Very simple data representation.

- Subject: URI
- Relation: URI
- Object: URI or literal.

http://insaph.kcl.ac.uk/iaph2007/iAph010002/

http://spqr.epcc.ed.ac.uk/material

WHITE MARBLE

As we shall see, there is a simple graphical representation.

Principles of Linked Data

- Use URIs as names for things
- Use HTTP URIs so that people can look up those names
- When someone looks up a URI provide useful information
- Include links to other URIs. so that they can discover more things

http://www.geonames.org/2208578/ refers to a place.

It is redirected to http://www.geonames.org/2208578/marsa-zawaghah.html

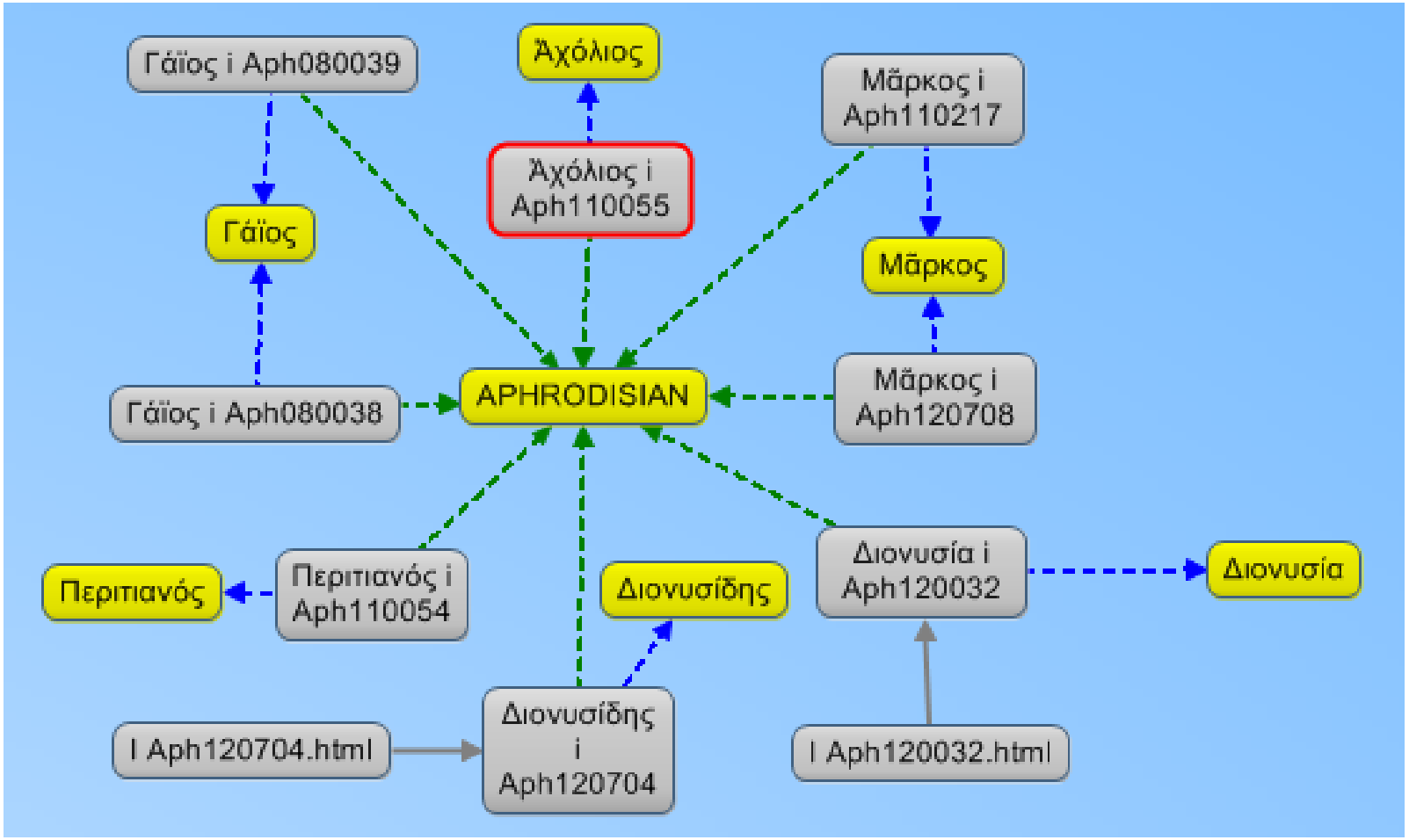
which describes the place.

From EpiDoc to Linked Data

- Scripts have been written to transform (a subset of) the EpiDoc into Linked data.
- Some human intervention is required.
- The scripts are written in Clojure which is a dialect of Lisp underpinned by Java.
- There are various formats for describing linked data. We have chosen N-Triples (RDF/XML is another popular choice).

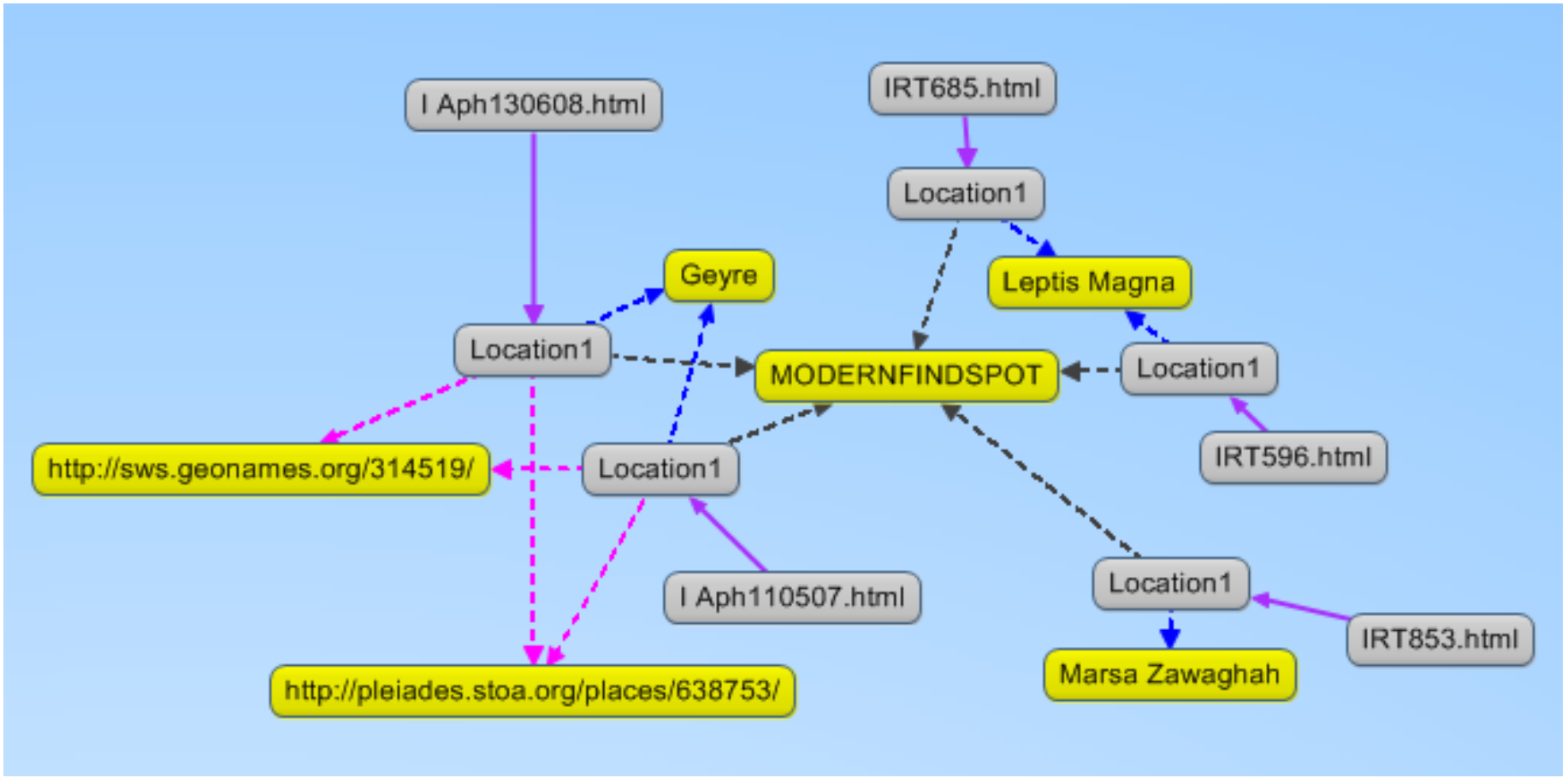
Some Linked Data about People

This is a screenshot of Gruff.



Note the colour coding.

Some Linked Data about Places



Can open up new links and browse the data.
Note the external links.

SPARQL

Using this interface one can identify all of those inscriptions that mention Aphrodite, say.

One can explore from this point but one might want to look for combinations of relationships, or there might be too many data to examine conveniently through this GUI.

This is where the query language SPARQL comes in.

Example of SPARQL

```
# Places named after Charles Darwin (in dbpedia).
PREFIX rdfs: http://www.w3.org/2000/01/rdf-schema#
PREFIX dbprop: http://bdpedia.org/property/
SELECT ?location
WHERE {
        ?person rdfs:label “Charles Darwin”@en
        ?location dbprop:nameFor ?person
        }
```

This is similar to SQL.

Canned Queries

- SPARQL, like SQL, is not very appealing to non-programmers.
- We need to identify query patterns that are useful to epigraphers.
- An epigrapher will then be able to select a template and fill in the details rather than constructing a query from scratch.
- KCL are looking at this.

Evaluation

- KCL have conducted a preliminary evaluation using Gruff.
- The GUI is thought to be useful.
- Some of the data need to be encoded differently or (with the help of epigraphers) changed to make searching and establishing links easier.
- Looking at what further data need to be encoded.
- This is work in progress.

Conclusion

- The approach looks promising.
- There is still a long way to go.