CLAROS: experiences in integrating disparate art-historical resources using a linked data approach

Sebastian Rahtz,
Director (Research) of Academic IT,
Oxford University IT Services

DHOXSS July 2014

Data: the CLAROS philosophy

Data is not just for Christmas

The oldest approach Collect data, put it on index cards, read cards, write a book. END

The old approach Collect data, manipulate it using Access or Excel, write a book, lose floppy disks. END

The newer approach Collect the data, put it in a MySQL database, write a web front end with search boxes and browsing, archive database backup files. END.

The best approach Collect data, share it with others, look at their data, write a book, leave the data available for the next generation. NO END

The CLAROS programme

CLAROS is an Oxford-based international collaboration working on:

1. Development of a humanities data web combining leading classical art history and related databases
2. Demonstration interfaces to explore world art
3. Innovative searching based on shape analysis
4. Large-scale RDF database providing a testbed for performance research
5. Changing the approach to data discovery by development of a conversational Companion

CLAROS in the context of academic research data management

Our aim is to help our academic researchers:

- publish an index to resources across a wide range of the humanities by (at least) minimal mapping to a common standard
- create a neutral data format which can be archived
- see their work and data as addressable resources
- make use of off-the-shelf, easily-maintained, and powerful query systems and visualizations
- put their work in the same spectrum as the rest of the cultural heritage sector

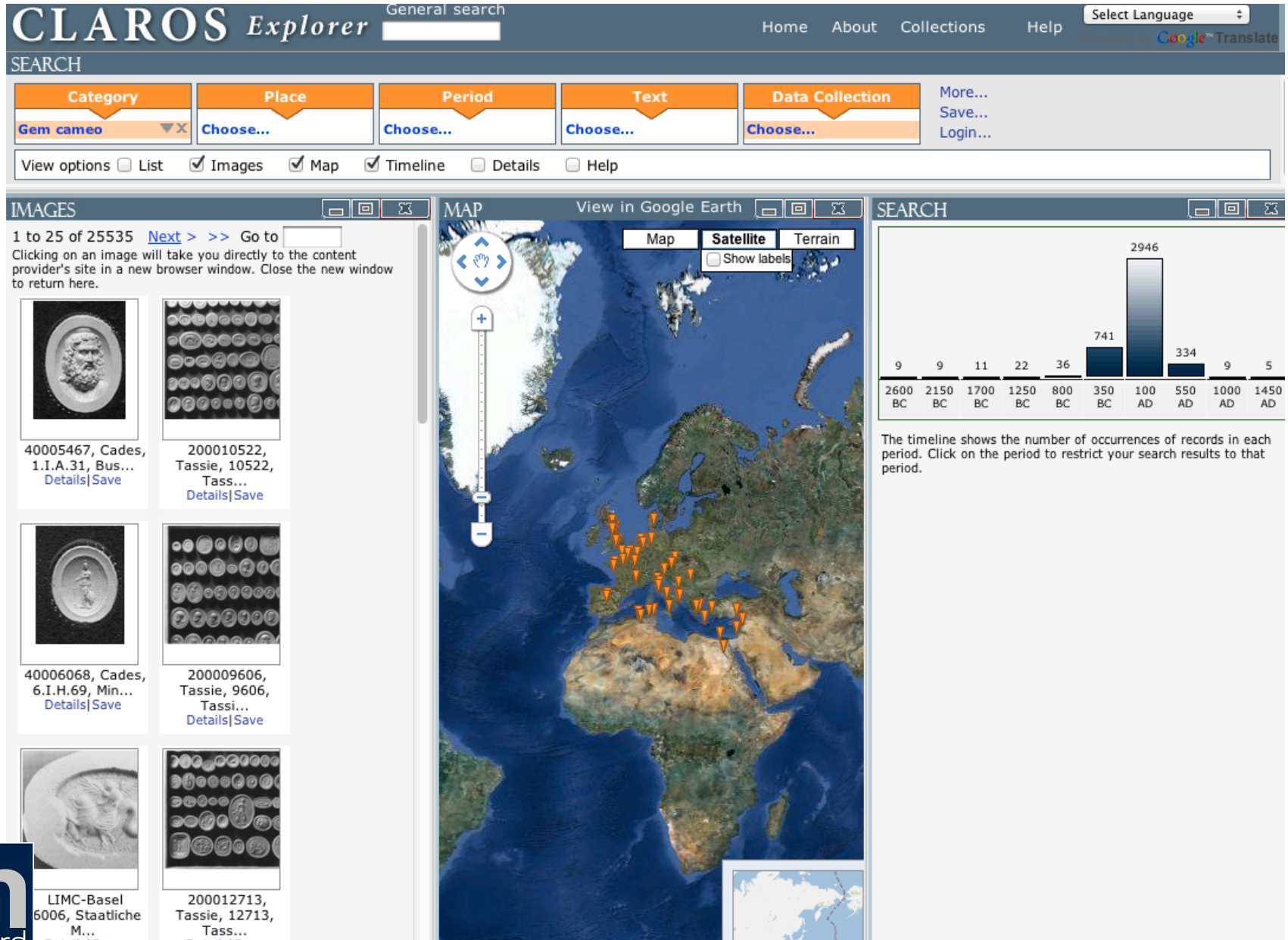
and thus meet the increasingly stringent requirements of research funders.

Our technical aspiration: 5 star data

- ★ Available on the web (whatever format), but with an open licence
- ★★ Available as machine-readable structured data (e.g. Excel instead of image scan of a table)
- ★★★ As (2) plus non-proprietary format (e.g. CSV instead of Excel)
- ★★★★ All the above, plus: Use open standards from W3C (RDF and SPARQL) to identify things, so that people can point at your stuff
- ★★★★★ All the above, plus: Link your data to other people's data to provide context

`http://www.w3.org/DesignIssues/LinkedData.html`

CLAROS web


CLAROS Explorer
General search
Home
About
Collections
Help
Select Language
Powered by Google Translate
SEARCH
Category
Gem cameo
Place
Choose...
Period
Choose...
Text
Choose...
Data Collection
Choose...
More...
Save...
Login...
View options
List
Images
Map
Timeline
Details
Help
IMAGES
1 to 25 of 25535
Next >
>> Go to
Clicking on an image will take you directly to the content provider's site in a new browser window. Close the new window to return here.
40005467, Cades, 1.I.A.31, Bus...
Details|Save
200010522, Tassie, 10522, Tass...
Details|Save
40006068, Cades, 6.I.H.69, Min...
Details|Save
200009606, Tassie, 9606, Tassi...
Details|Save
LIMC-Basel 6006, Staatliche M...
200012713, Tassie, 12713, Tass...
MAP
View in Google Earth
Map
Satellite
Terrain
Show labels
SEARCH
2946
741
334
9
9
11
22
36
9
5
2600 BC
2150 BC
1700 BC
1250 BC
800 BC
350 BC
100 AD
550 AD
1000 AD
1450 AD
The timeline shows the number of occurrences of records in each period. Click on the period to restrict your search results to that period.

CLAROS data web

LIMC-Basel 17643, Musei Capitolini Roma

http://www.limc.ch/public/monument_view.aspx?id=17643

Nearby

Other formats

- HTML
- N-Triples
- Notation3
- RDF/XML
- Turtle

Other things of type: crm:E22_Man-Made_Object, oac:Target

What links here

View more detail at partner's website

claros:coordinates-current	41.89999961853027, 12.483333110809326
claros:coordinates-find	41.89999961853027, 12.483333110809326
crm:P102_has_title	LIMC-Basel 17643, Musei Capitolini Roma
crm:P108i_was_produced_by	Production of LIMC-Basel 17643, Musei Capitolini Roma
crm:P111i_was_added_by	http://www.limc.ch/public/monument_view.aspx?id=17643/part_addition/3 http://www.limc.ch/public/monument_view.aspx?id=17643/part_addition/4 http://www.limc.ch/public/monument_view.aspx?id=17643/part_addition/2 http://www.limc.ch/public/monument_view.aspx?id=17643/part_addition/5
crm:P138i_has_representation	Image of LIMC-Basel 17643, Musei Capitolini Roma Image of LIMC-Basel 17643, Musei Capitolini Roma
crm:P16i_was_used_for	Found at Rome
crm:P2_has_type	Mosaic mosaic
crm:P48_has_preferred_identifier	LIMC-Basel 17643
rm:P53_has_former_or_current_location	Musei Capitolini Roma
crm:P70i_is_documented_in	LIMC-Basel 17643, Musei Capitolini Roma
rdf:type	crm:E22_Man-Made_Object oac:Target
sameAs	LIMC-Basel 17643, Musei Capitolini Roma

CLAROS: (some of the) data resources

- University of Oxford – Beazley Archive of pottery and gems; Lexicon of Greek Personal Names
- University of Cologne – Research Sculpture Archive
- German Archaeological Institute - photographs
- University of Paris X - Lexicon Iconographicum Mythologiae Classicae
- University of Grenoble - Lexicon Iconographicum Mythologiae Classicae
- Ashmolean Museum - Jameel Islamic Collection; Creswell Photographic Collection
- British School at Rome - antiquarian photographs and prints
- Cycladic Museum, Athens - Cycladic art

The minimal entry criteria are openly-licensed data, and persistent URIs for records.

CLAROS geographic coverage

Places

The CLAROS data web approach

- No changes to the databases of the individual sources
- Semantic differences between data sources are resolved by mapping selected metadata from each source to CIDOC-CRM
- Syntactic differences between data sources are resolved by converting the selected metadata to RDF
- Complete records are pulled and stored, not just annotations (cf Pelagios)

CLAROS is simply a cacheing resource discovery service — the user is ultimately directed back to the original data publisher's site for full information about an event, object, place or person of interest.

CIDOC Conceptual Reference Model CRM

The CRM provides definitions and a formal structure for describing the implicit and explicit concepts and relationships used in cultural heritage documentation.

- Actors (people)
- Conceptual objects
- Physical things
- Events
- Time spans
- Places

and relationships between them, e.g.:

- participate in
- refer to
- have location
- within

The CRM concepts

Modification Acquisition Activity Actor Actor_Appellation Address Appellation Attribute_Assignment Authority_Document Beginning_of_Existence Biological_Object Birth CRM_Entity Collection Conceptual_Object Conceptual_Object_Appellation Condition_Assessment Condition_State Contact_Point Creation Curation_Activity Date Death Design_or_Procedure Destruction Dimension Dissolution Document End_of_Existence Event Formation Group Identifier Identifier_Assignment Image Information_Carrier Information_Object Inscription Joining Language Leaving Legal_Body Legal_Object Linguistic_Object Man-Made_Feature Man-Made_Object Man-Made_Thing Mark Material Measurement Measurement_Unit Move Part_Addition Part_Removal Period Persistent_Item Person Physical_Feature Physical_Man-Made_Thing Physical_Object Physical_Thing Place Place_Appellation Place_Name Production Propositional_Object ight Section_Definition Site Spatial_Coordinates Symbolic_Object emporal_Entity Thing Time-Span Time_Appellation Title

The CRM properties

added assigned assigned_attribute_to at_some_time_within
augmented bears_feature beginning_is_qualified_by borders_with
brought_into_existence brought_into_life by_mother
carried_out_by carries classified concerned consists_of continued
created_type curated custody_received_by
custody_surrendered_by deassigned depicts destroyed diminished
dissolved documents employed end_is_qualified_by exemplifies
falls_within finishes foresees_use_of from_father
had_as_general_use had_at_least_duration had_at_most_duration
had_general_purpose had_participant had_specific_purpose
has_alternative_form has_broader_term has_component
has_condition has_contact_point has_created has_current_keeper
has_current_location has_current_or_former_curator
has_current_or_former_member
has_current_or_former_residence has_current_owner
has_current_permanent_location has_dimension has_formed
has_former_or_current_keeper has_former_or_current_location

CLAROS and CIDOC CRM

We have found CIDOC CRM to be pretty well suited for CLAROS data

- The CRM choices are documented at `http://www.clarosnet.org/wiki/index.php`
- There is the useful OWL implementation of CRM by Erlangen University
- We focused initially on the CRM Core terms, and employed additional terms as necessary
- CRM Core covers our needs for complex provenance of artefacts and their relationships with key events, people, places and times. We extended a little to cover geographic coordinates, and dating.
- The CIDOC CRM "E55_Type" system is useable to permit faceted/drill-down queries, e.g. restricting results by the shape of a pot, but we probably abuse it.

What is the point of all this?

- Standardized schema to allow interoperability
- Permanent identifiers, accessible via the web
- Linking to other peoples data
- Access to toolkits (eg Research Space

CLAROS: getting the data ready

For each contributor:

1. make sure every object has a unique, open, URI
2. decide which data categories are licensed for open access
3. map local schema to CRM
4. write database export/data wrangler from local form to RDF/XML

then we

- mangle RDF/XML to
  1. link place names where possible to CLAROS gazetteer and thence to Pleiades and Geonames (cf Pelagios)
  2. add consistent typology

Examples of practicalties

- Lexicon of Greek Personal Names delivered entirely by transformation of TEI XML to RDF
- Beazley Archive generated RDF by ASP scripts run on SQL database
- Creswell archive emailed an XML dump from MuseumPlus which we mangled using XSLT

most of the problems arise from mapping local schema to CRM, or licensing, or unstructured data, or issues of quality, or concerns over uncertainty. Actually making RDF is generally easy.

Summary of data

arachne	Arachne	185119 objects
ashmol	Jameel Collection, Ashmolean	2316 objects
beazley	Beazley Archive	130960 objects
bsa	British School at Athens	(pending)
bsr	British School at Rome, photographs and plans	16043 objects
creswell	Creswell Photographic Archive, Ashmolean	6521 objects
cycladic	Cycladic Museum, Athens	348 objects
lgpn	Lexicon of Greek Personal Names	251821 people
limc	LIMC Paris	4724 objects
limcbasel	LIMC Basel	55852 objects
metamorphoses	Gazetteer	9396 places (6325 geolocated)
oxrep	Oxford Roman Economy project	(pending)
waa	World of Ancient Art	406 places

Partner data provision

arachne	Arachne	OAI feed
ashmol	Jameel Collection, Ashmolean	static file
beazley	Beazley Archive	REST retrieval
bsa	British School at Athens	(pending)
bsr	British School at Rome, photographs and plans	static file
creswell	Creswell Photographic Archive, Ashmolean	static file
cycladic	Cycladic Museum, Athens	static file
lgpn	Lexicon of Greek Personal Names	REST retrieval
limc	LIMC Paris	database dump, conversion to file
limcbasel	LIMC Basel	database dump, conversion to file
metamorphoses	Gazetteer	(natively managed)
oxrep	Oxford Roman Economy project	(pending)
trendall	Trendall Archive	(unclear)
waa	World of Ancient Art	database dump, conversion

Example: aerial photograph from British School at Rome

BACK TO LIST VIEW ZOOMABLE IMAGE PRINT XML RECORD

Forms part of: Ward-Perkins Collection. Photographs. South Etruria Series.

COPYRIGHT
Publication restricted (BSR copyright)

REPRODUCTION NO.
wpset-00728 (b&w digital file from original neg.)

ACCESS
Closed access material (Archive)

CALL NO.
WP[PHP]-SEtD05-047b
WP[PHN]-A-SEt00728

TITLE
Rusellae

NAME
Ward-Perkins, J. B. (John Bryan), 1912-1981 Photographer

RELATED NAMES
British School at Rome. 1954-1968 South Etruria Survey.

DATE
[between 1954 and 1968]

ITEM
1 photographic print glued on card : gelatin silver, b&w ; 15 x 11 cm.
1 negative : safety film ; 6 x 9 cm.

DESCRIPTION
Aerial photograph of the archaeological site of Roselle.

NOTE
Title from card.
Note on card: (Photos after D.A.I.)

GEOGRAPHICAL SUBJECT
Roselle (Extinct city)

TGN
Italy -- Toscana -- Rusellae

GENRE/FORM
Gelatin silver prints -- 1950-1970
Safety film negatives -- 1950-1970
Aerial photographs -- 1950-1970

As RDF XML (1)

```
<E22_Man-Made_Object
    rdf:about="http://id.clarosnet.org/BSR/0006317">
  <P53_has_former_or_current_location
      rdf:resource="http://id.clarosnet.org/places/place/rome-bsr"/>
  <P138_represents
      rdf:resource="http://id.clarosnet.org/places/place/rusellae"/></E22_Man-Made_Object>
<E53_Place
    rdf:about="http://id.clarosnet.org/places/place/rusellae">
  <rdfs:label>[IT] Rovine di Roselle</rdfs:label>
  <P87_is_identified_by>
    <E48_Place_Name
        rdf:about="http://id.clarosnet.org/places/placename/rovine_di_roselle">
      <value>Rovine di Roselle</value>
    </E48_Place_Name>
  </P87_is_identified_by>
  <P87_is_identified_by>
    <E47_Place_Spatial_Coordinates
        rdf:about="http://id.clarosnet.org/places/place/rusellae/coordinates">
      <claros:has_geoObject>
        <geo:Point>
          <geo:lat>42.83333</geo:lat>
          <geo:long>11.16667</geo:long></geo:Point></claros:has_geoObject>
    </E47_Place_Spatial_Coordinates>
  </P87_is_identified_by>
  <skos:closeMatch
      rdf:resource="http://pleiades.stoa.org/places/413288#this"/>
  <skos:closeMatch
      rdf:resource="http://sws.geonames.org/3168944/"/>
  <P89_falls_within
      rdf:resource="http://id.clarosnet.org/places/country/IT"/>
E53_Place>
```

As RDF XML (2)

```
<P108i_was_produced_by>
  <E12_Production
      rdf:about="http://id.clarosnet.org/BSR/0006317/production">
    <P14_carried_out_by>
      <E21_Person
          rdf:about="http://id.clarosnet.org/BSR/person/Ward-Perkins-J.-B-(John-Bryan)-1912-1981-
British-School-at-Rome">
        <P131_is_identified_by>
          <E82_Actor_Appellation
              rdf:about="http://id.clarosnet.org/BSR/name/Ward-Perkins-J.-B-(John-Bryan)-1912-
1981-British-School-at-Rome">
            <value>Ward-Perkins, J. B (John Bryan) 1912-1981 British School at Rome</value>
          </E82_Actor_Appellation>
        </P131_is_identified_by>
      </E21_Person>
    </P14_carried_out_by>
    <P4_has_time-span>
      <E52_Time_Span>
        <P82_at_some_time_within>
          <claros:Period>
            <claros:period_begin
                rdf:datatype="http://www.w3.org/2001/XMLSchema#gYear">1954</claros:period_begin>
            <claros:period_end
                rdf:datatype="http://www.w3.org/2001/XMLSchema#gYear">1968</claros:period_end></claros:Perio
        </P82_at_some_time_within>
      </E52_Time_Span>
    </P4_has_time-span>
  </E12_Production>
</P108i_was_produced_by>
```

Viewed in CLAROS data explorer

claros:coordinates-current	http://id.clarosnet.org/places/metamorphoses/place/rome/coordinates
claros:coordinates-find	42.83333, 11.16667
crm:P102_has_title	Rusellae
crm:P108i_was_produced_by	http://id.clarosnet.org/BSR/0006317/production
crm:P138_represents	Rovine di Roselle
crm:P138i_has_representation	Image of Rusellae
crm:P2_has_type	http://id.clarosnet.org/type/object/graphic Gelatin silver prints 1950-1970 Aerial photographs 1950-1970 graphic Aerial Photograph Safety film negatives 1950-1970 Photograph
crm:P3_has_note	Note on card: (Photos after D.A.I.) Aerial photograph of the archaeological site of Roselle Title from card
crm:P53_has_former_or_current_location	British School at Rome
crm:P67i_is_referred_to_by	http://www.bsrdigitalcollections.it/details.aspx?ID=0006317
rdf:type	crm:E22_Man-Made_Object
label	Rusellae

OTHER FORMATS

- HTML
- JSON-LD
- N-Triples
- Notation3
- RDF/JSON
- RDF/XML
- Turtle

Other things of type:
crm:E22_Man-Made_Object

Example: a Greek

An inscription published in Inscriptiones Graecae volume XI (4), p. 1256 documents a man called Παράμονος, attested at Delos in the 3rd or 2nd century BC. He is noted as being the father of someone called Δημήτριος.

The Greek in data source

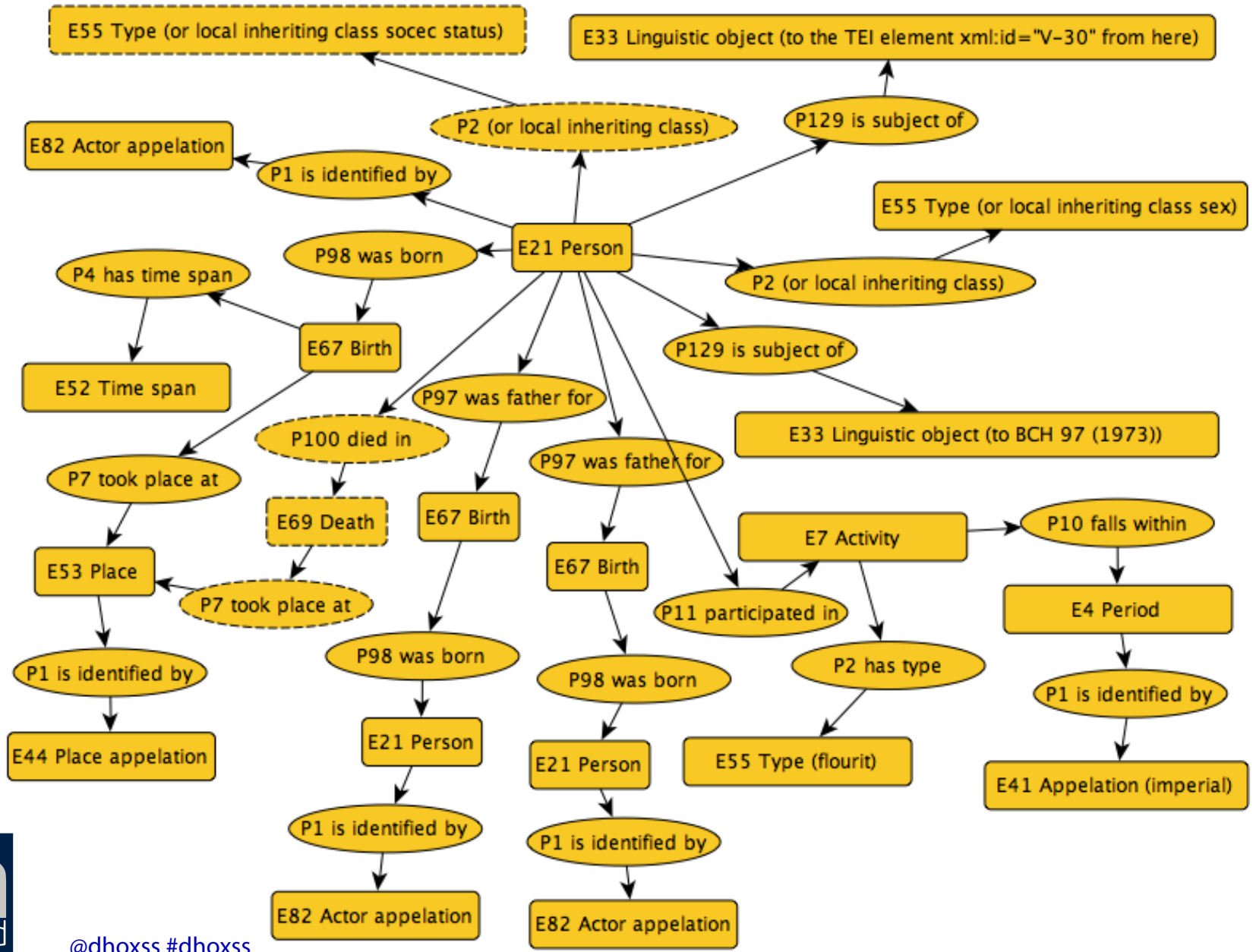
Relational DB:

```
id | name | floruit | sex | status | settlement
----------+-----------+------------+-----+----------+----------
-------
V1-43005 | Παράμονος | hell.-imp.| 1 | | Mytilene
V1-76555 | Παράμονος | ii-i BC   | 1 | | Eretria Vathia
V1-76557 | Παράμονος | c.100BC   | 1 | | Histiaia-Oreos
V1-78877 | Παράμονος | iv/iii BC | 1 | | Eretria
V1-85238 | Παράμονος | iii/ii BC | 1 | |
```

... or in XML

```
<person n="1-7" xml:id="V1-85238">
<sex value="1"/>
<persName type="main" nymRef="#nPara1monos">Παράμονος</persName>
<birth notAfter="-0175" notBefore="-0225">
<placeName key="LGPN_11270" evidence="attested">Delos</placeName>
</birth>
<floruit>iii/ii BC</floruit>
<state key="#relationship">
<label>f.
<persName type="relationship" xml:lang="el-grc" nymRef="#nDhmh1trios">Δημήτριος</persName>
</label>
</state>
<bibl>
<title>IG</title> XI (4) 1256</bibl>
</person>
```

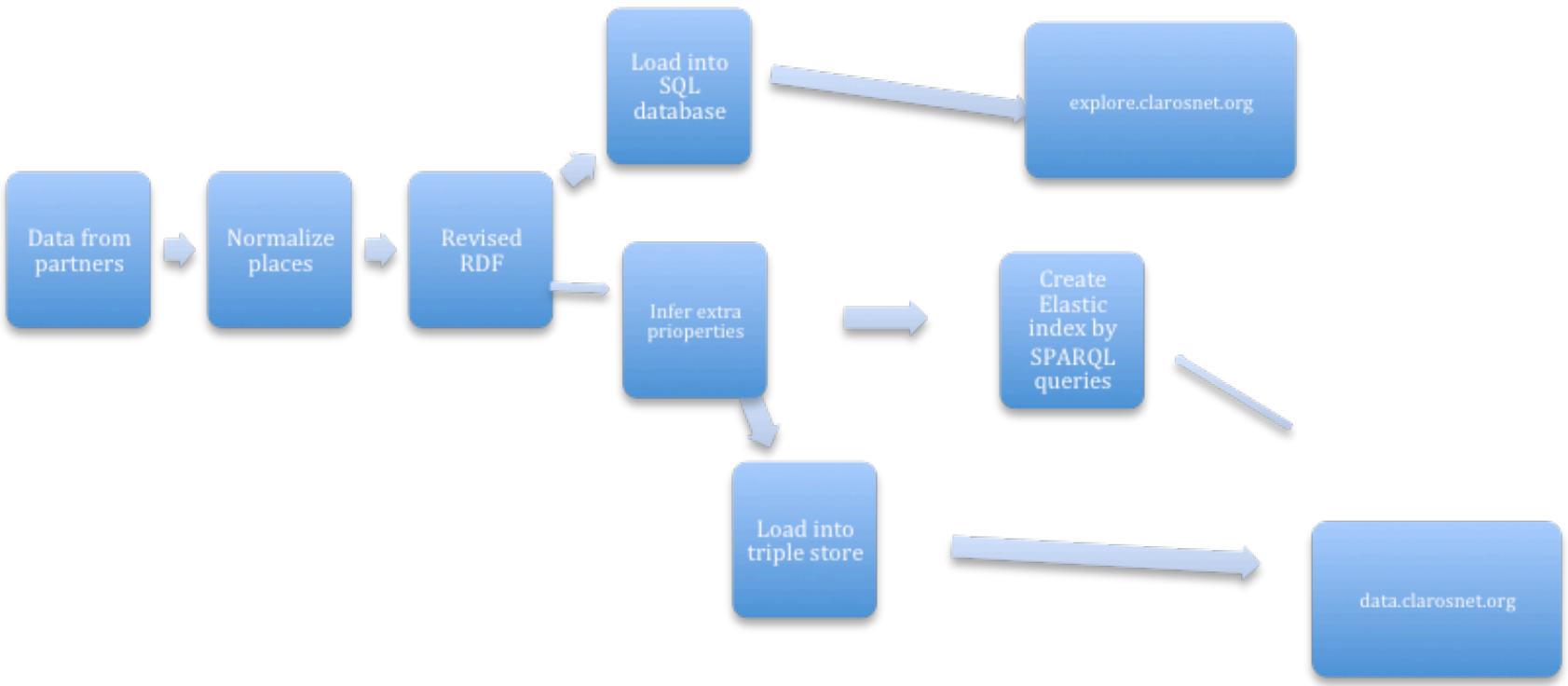
CRM structure for a person like this



The Greek expressed in RDF XML

```
<E21.Person
    rdf:about="http://clas-lgpn2.classics.ox.ac.uk/id/V1-85238">
  <P131.is_identified_by  xml:lang="el-grc">
    <E82.Actor_Appellation>
      <value>Παράμονος</value></E82.Actor_Appellation></P131.is_identified_by>
  <P131.is_identified_by  xml:lang="el-grc-x-lgpn">
    <E82.Actor_Appellation>
      <value>Paramonos</value></E82.Actor_Appellation></P131.is_identified_by>
  <P98.was_born>
    <E67.Birth>
      <P4.has_time-span>
        <E52.Time-Span>
          <P79.at_some_time_within>
            <claros:Period>
              <claros:period_begin
                  rdf:datatype="http://www.w3.org/2001/XMLSchema#gYear">-
0225</claros:period_begin>
              <claros:period_end
                  rdf:datatype="http://www.w3.org/2001/XMLSchema#gYear">-
0175</claros:period_end></claros:Period></P79.at_some_time_within></E52.Time-Span></P4.has_time-
span>
      <P7.took_place_at
          rdf:resource="http://clas-
lgpn2.classics.ox.ac.uk/placeid/LGPN_11270"/></E67.Birth></P98.was_born></E21.Person>
```

Practicalities



Name mapping procedure

- Does <E53_Place> have a geolocation? OK
- Normalize place name. Translate space to -, lower-case, normalize accents, etc
- Does name of place match a known place? link to that place
- Does name of place *partially* match a place? create an <E53_Place> which has a <P89_falls_within> linking to the half-match. Example 'athens-kerameikos'
- Does <E53_Place> have a geonames link? get lat/long from www.geonames.org

Primitive Typology

architecture	Architectural Sculpture
bound-volume	Bound volume
mirror	Mirror
coin	Coins
drawing	Eastern Drawing
eastern-ceramic	Eastern Ceramic
eastern-bronze	Eastern Bronze
eastern-painting	Eastern Painting
engraving	Engraved print
gem-cameo	Gems and Cameos
inscription	Inscription
jewellery	Jewellery
menhir	Menhir
mosaic	Mosaic
painting	Wall Painting
papyrus	Papyrus
aerialphotograph	Aerial Photograph
photograph	Photograph
etching	Etching
plan	Site plan
map	Map
portrait	Portrait
print	Eastern Print
sarcophagus	Sarcophagus

Mapping of types

```
<type
    url="http://arachne.uni-koeln.de/vocabulary/objectType#-lebewesen"
    type="other"/>
<type
    url="http://arachne.uni-koeln.de/vocabulary/objectType#anthropomorpher"
    type="other"/>
<type
    url="http://arachne.uni-koeln.de/vocabulary/objectType#architektur"
    type="architecture"/>
<type
    url="http://arachne.uni-koeln.de/vocabulary/objectType#attischer"
    type="western-ceramic"/>
<type
    url="http://arachne.uni-koeln.de/vocabulary/objectType#bauornamentik"
    type="architecture"/>
<type
    url="http://id.clarosnet.org/type/Man-Made_Object/cartographic"
    type="map"/>
<type
    url="http://id.clarosnet.org/type/Man-Made_Object/graphic"
    type="graphic"/>
<type
    url="http://id.clarosnet.org/type/Man-Made_Object/map"
    type="map"/>
<type
    url="http://purl.org/NET/Claros/vocab#Ashmolean/Category/bound_volume"
    type="bound-volume"/>
<type
    url="http://purl.org/NET/Claros/vocab#Ashmolean/Category/ceramic"
    type="western-ceramic"/>
ype
    url="http://purl.org/NET/Claros/vocab#Ashmolean/Category/drawing"
```

Technologies?

Servers	1 Linux Ubuntu, 1 Windows
Normalize phase	XSLT manipulation of RDF/XML
Data inference	SPARQL queries in Python wrapper
Triple store and SPARQL endpoint	Jena, in Fuseki packaging
Public web site	MS SQL server and ASP pages
Data web site	Humfrey (local open source), Elasticsearch (Lucene) and extra Python

What does work?

- Mapping to CIDOC CRM RDF
- Conversion to SQL database to drive user-friendly web site
- Loading into triplestore with SPARQL endpoint
- Map-based display and textual searching
- Export to Pelagios

What works, but not as well we would like?

- Provision of data by automated means
- Joining up places internally
- Mapping to common taxonomy
- Mapping places to Pleiades

What does not work yet?

- Updating of individual datasets by partners
- Location-based searching
- Managing periods intelligently

What can we do with the places component of CLAROS (1)?

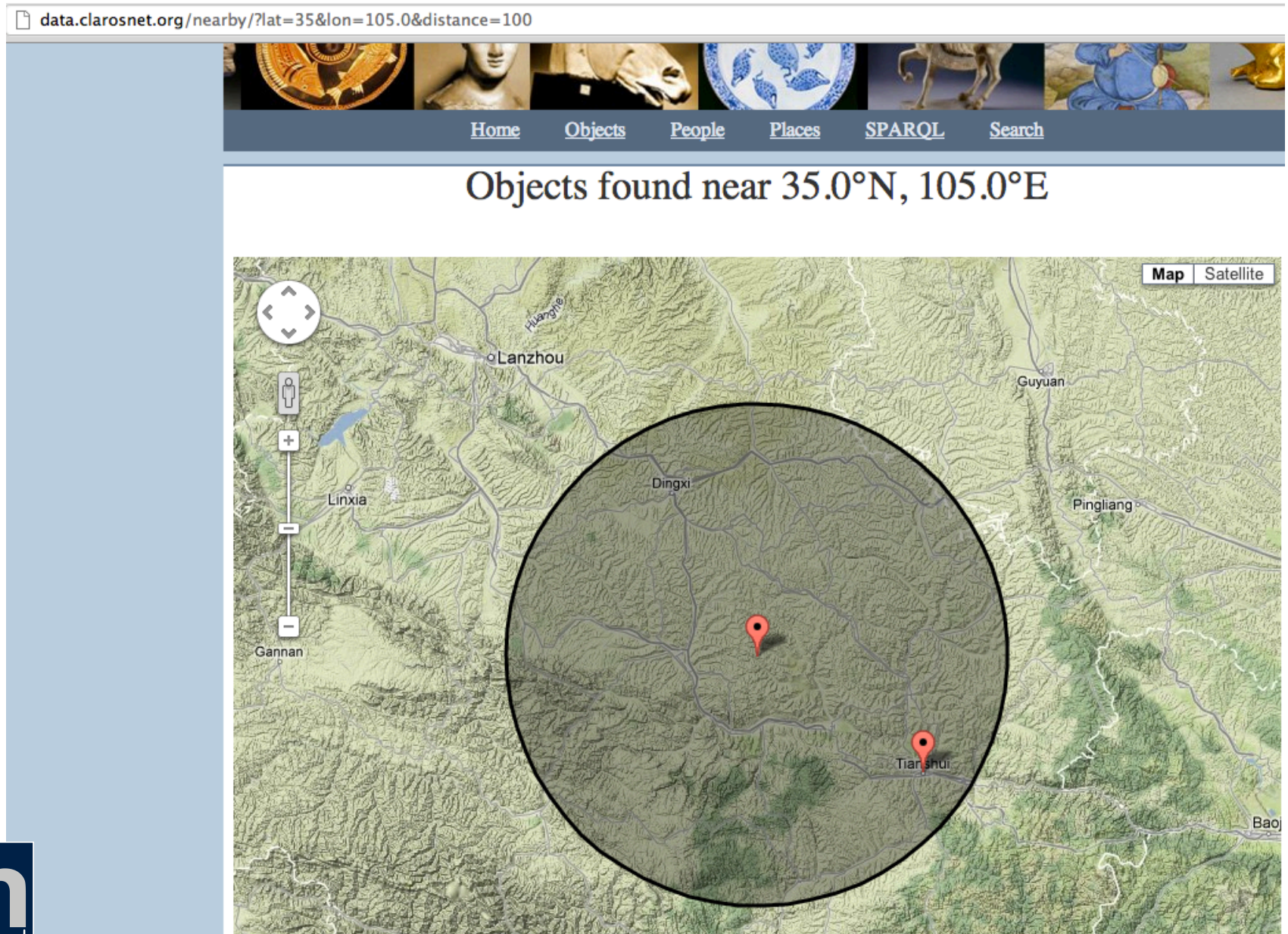
- Map the majority of commonly-occurring find spots to a geolocation (at the city level)
- Map some current location places to a geolocation
- Access c.125000 objects via find spot (out of c.400000)
- Access c.161000 people via a birth place (out of c.250000)

What can we do with the places component of CLAROS (2)?

- Show results of search on Google/Open StreetMap maps
- Select places on Google/Open StreetMap maps
- Find places by name browse
- Find places by free text search combined with material/type/title/name
- Find objects nearby (by radius) current object
- Find places nearby current place

Searches can be accessed by REST url

Searching nearby a coordinate



Known places

- c.9300 places known
- c.6200 places geolocated
- c.1500 places linked to Pleiades
- c.4330 places linked to geonames.org

What remains to be done at the data level?

- Resolving duplicates
  - same geonames ID? done.
  - same pleiades ID? done.
  - same name? done.
  - similar geolocation. TODO.
  - similar name. TODO.
- Finding new geolocations
  - simple name, waiting to be found. eg Dayton / Ohio/
  - obscure, but known, name. eg Maiori Nuraghe, Sardinia
  - obscure place. eg Yukarõ Dodurga
  - confusing names. eg Romische Stadt von Ampurias / Gerona (P) / Cataluna
  - sites within known places. eg 'Rome, In der Ecke eines Hauses nahe S. Maria in Pace'

All this assumes that the partners do not provide geonames or Pleiades links

What remains to be done at the interface level?

- Browsing by place hierarchy
- Click on map to find places
- Filter places by date range of objects/people
- Add periodization to place names

What else?

For example, image-based searching

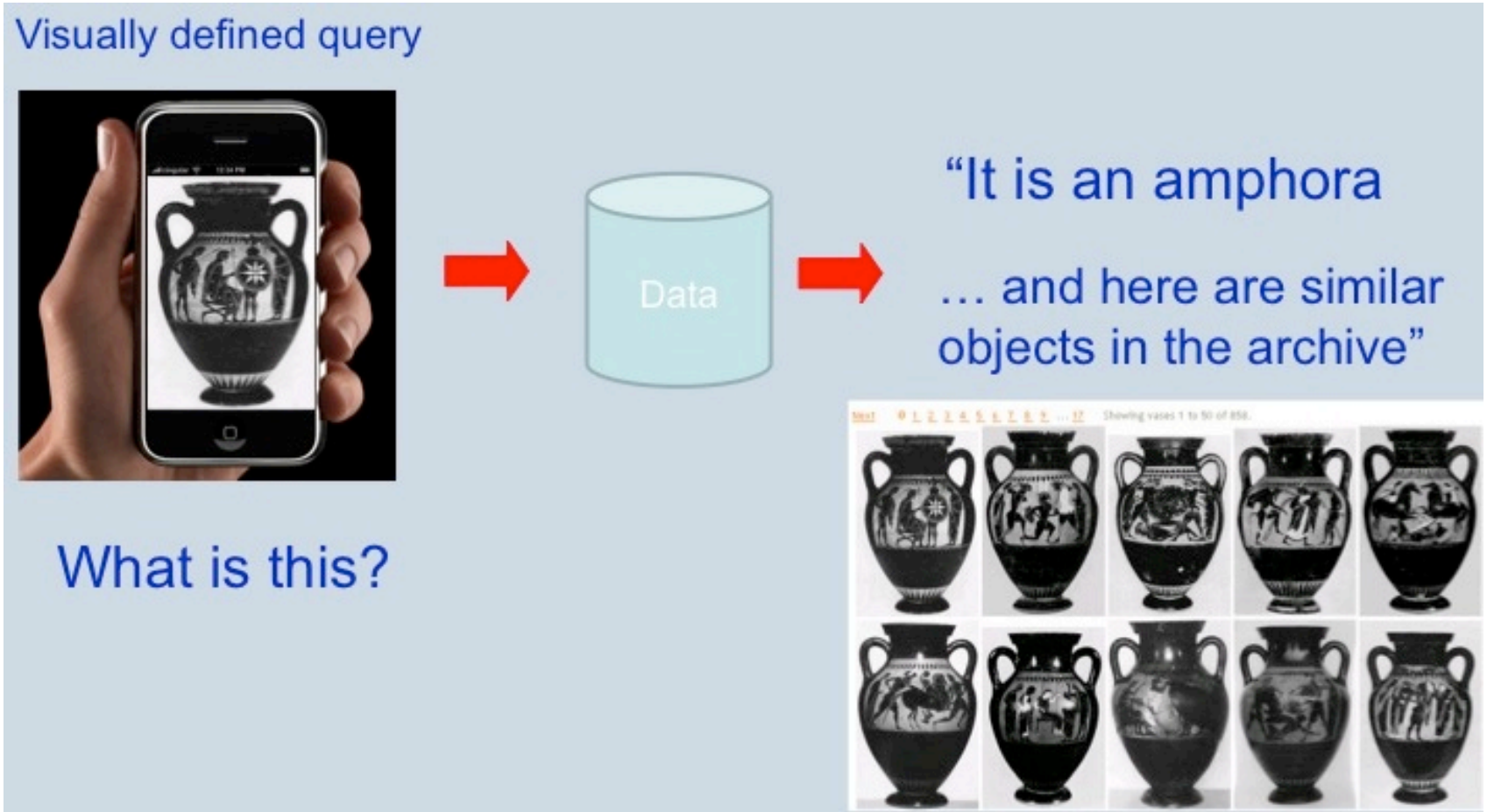


Image search

Image Search Results

Your original image

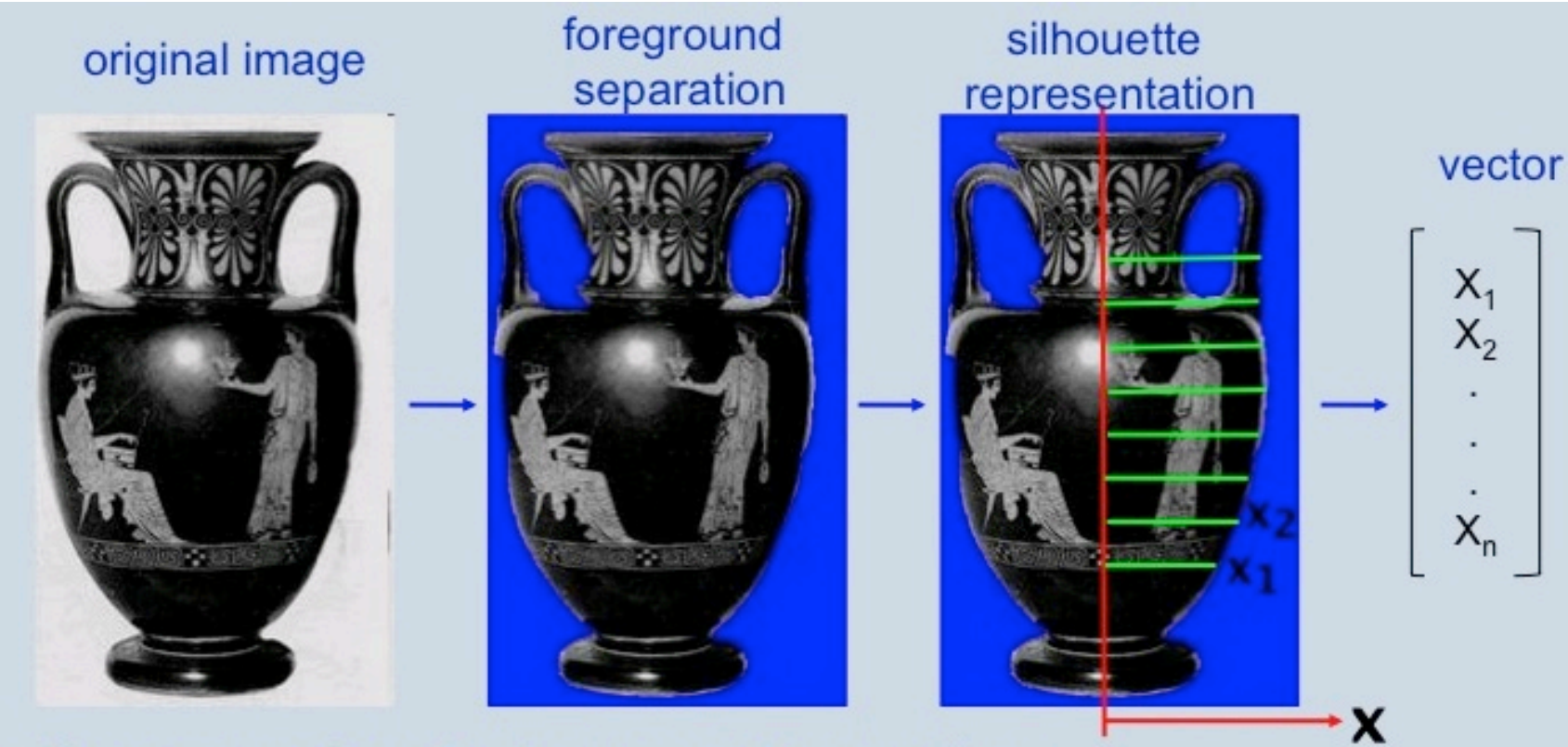
Listed below are pottery images with similar decorations and shape to the image you uploaded.
Click on the links or images to open details about each record.
Or click here to view the distribution of pottery of shape AMPHORA, NECK
in all the CLAROS partners' databases.

AMPHORA, NECK
302250, Munich, Loeb, SL458
Score: 26
Confidence: High

AMPHORA, NECK
302250, Munich, Loeb, SL458
Score: 22
Confidence: High

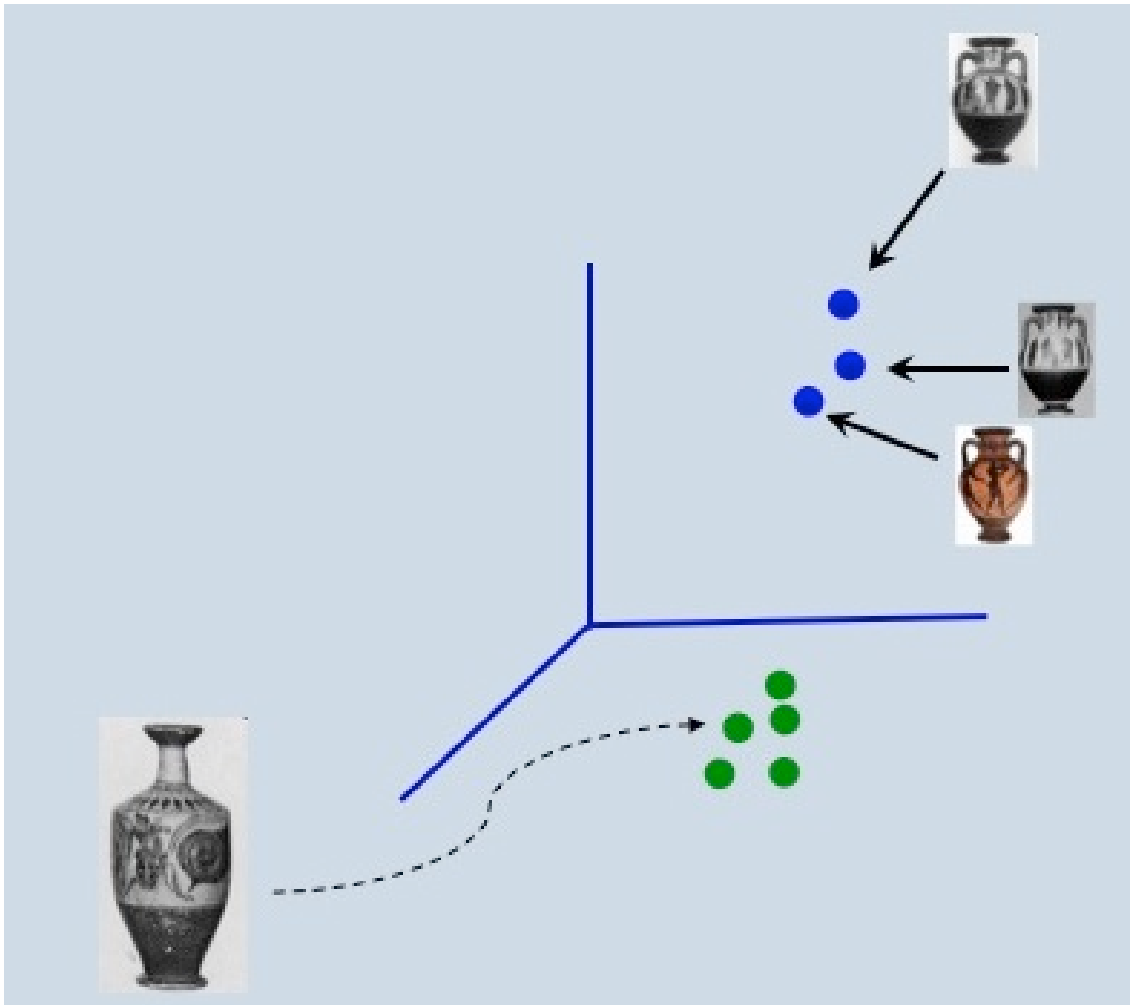
AMPHORA, NECK
320264, San Francisco (CA), M.H. de Young Mememorial Museum, 1925.368

Shape representation

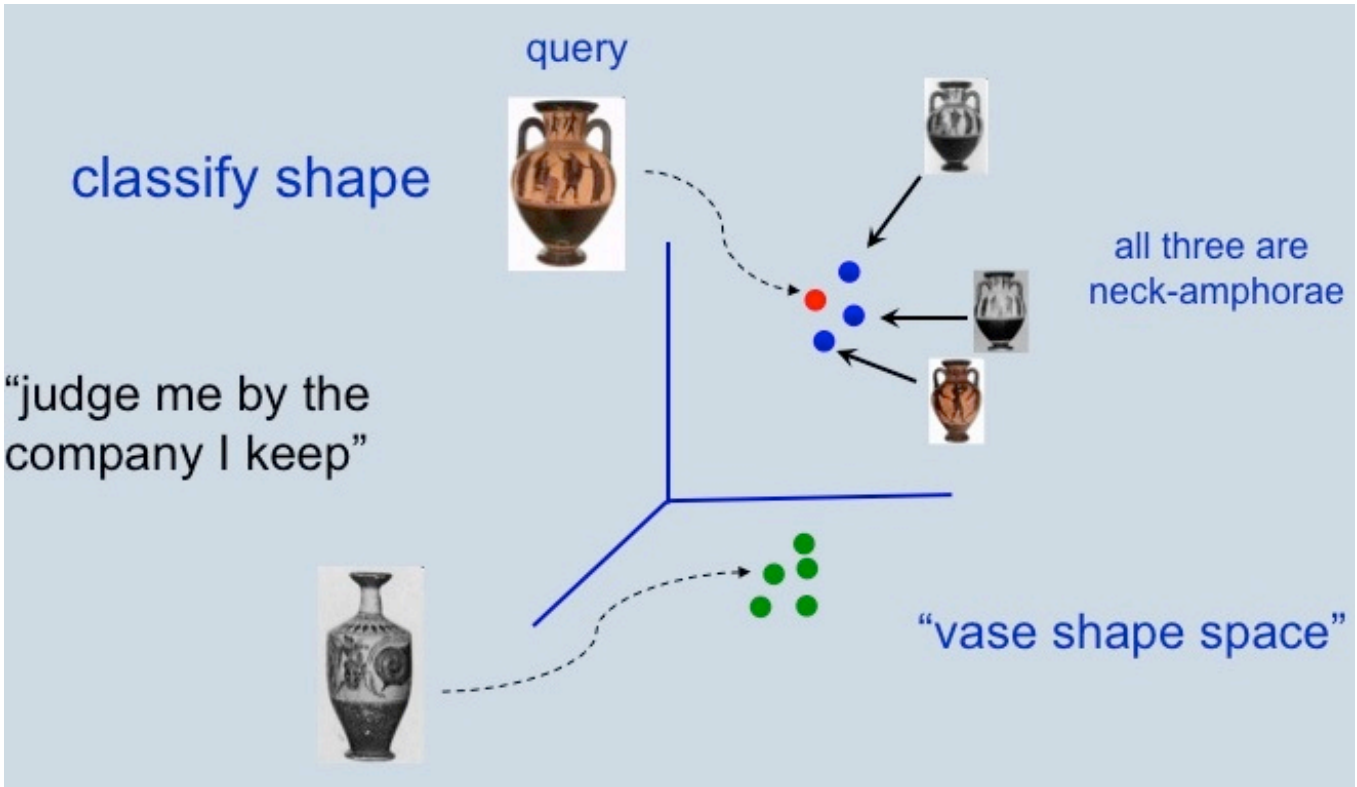


- No representation of patterns or surface markings
- 100-dimensional "vase shape space"

Vase shape space



Compute three nearest neighbours for each vase



What conclusions can one draw from CLAROS?

1. working with RDF and CIDOC CRM is not so very scary
2. aligning data to use the same taxonomies is harder
3. exporting data in RDF as a one-off is easy; making it harvestable is harder
4. this is only a start. we still need to find the research questions it answers

`http://www.clarosnet.org/`