Peripleo: a Tool for Exploring Heterogeneous Data through the Dimensions of Space and Time

Rainer Simon, AIT Austrian Institute of Technology, rainer.simon@ait.ac.at
Leif Isaksen, University of Lancaster, l.isaksen@lancaster.ac.uk
Elton Barker, The Open University, elton.barker@open.ac.uk
Pau de Soto Cañamares, University of Southampton, p.desotocanamares@soton.ac.uk

Abstract: This article introduces Peripleo, a prototype spatiotemporal search and visualization tool. Peripleo enables users to explore the geographic, temporal and thematic composition of distributed digital collections in their entirety, and then to progressively filter and drill down to explore individual records. We provide an overview of Peripleo's features, and present the underlying technical architecture. Furthermore, we discuss how datasets that differ vastly in terms of size, content type and theme can be made uniformly accessible through a set of lightweight metadata conventions we term "connectivity through common references". Our current demo installation links approximately half a million records from 25 datasets. These datasets originate from a spectrum of sources, ranging from the small personal photo collection with 35 records, to the large institutional database with 134.000 objects. The product of research in the Andrew W. Mellon-funded Pelagios 3 project, Peripleo is Open Source software.

Introduction: the Pelagios Initiative

Pelagios [1] is a community-driven initiative that facilitates better linkages between online resources documenting the past, based on the places they refer to. Our member projects are connected by a shared vision of a world – most eloquently described in Tom Elliott's article "Digital Geography and Classics" [2] – in which the geography of the past is every bit as interconnected, interactive and interesting as the present. Each project represents a different perspective on our shared history, whether expressed through text, map or archaeological record. But as a group we believe passionately that the combination of all of our contributions is enormously more valuable than the sum of its parts.

Initially focusing on the classical worlds of Greece and Rome, Pelagios has been working with a growing network of partners towards connecting different types of online resources relevant to the study of the past more generally – the literature of different periods and languages, archaeological data and databases, maps and other images, the results of scholarly research, etc. – so that they become more easily discoverable and seamlessly navigable for users. The key to connectivity in Pelagios is the use of shared online gazetteers [3] – directories of places that assign each place a unique, stable identifier in the form of a Uniform Resource Identifier (URI). Pelagios advocates the idea that whenever you refer to a place in your data, you should do so using a gazetteer URI. This way, otherwise isolated datasets become implicitly joined up to an interconnected graph, with the gazetteers as their central backbone [4], [5].

Pelagios has been working on developing best practices and tools to support its vision: e.g. by establishing conventions for the publication of gazetteer metadata as *Linked Open Data* [6]; or through the development of the Recogito annotation platform, which aids the process of linking digital texts and maps to the places they refer to [7]. In this paper, we introduce another outcome of the Pelagios initiative: Peripleo, a tool to visualize and navigate the growing pool of interconnected open data, published collectively by the Pelagios community. Our test installation of Peripleo [8] presently holds approx. half a million records from 25 datasets. These datasets encompass gazetteers (7 datasets), as well as content linking to them (18 datasets). The smallest dataset is a personal Flickr photostream with 35 records; the largest dataset is provided by the American Numismatic Society and contains approx. 134.000 numismatic records.

How Peripleo Works

Peripleo is ancient Greek for "to sail (or swim) around" in the sense of exploration or discovery. The notion of being able to freely roam the "sea of open data" brought together by our partners (and discovering the treasures hidden in remote places and ancient times!) was exactly the metaphor we had in mind when starting development. A key design goal was to provide, on the one hand, a familiar Google-Maps-like interface, with full-text search and auto-completion; while allowing a more free-form mode of exploration on the other. In this "exploration mode", Peripleo conveys a sense of the scope, breadth and structure of the data as a whole, by representing geographic coverage on the map; by showing temporal spread as a histogram; and by graphically illustrating distribution across different thematic facets (such as document language or data source). This way, users can easily gain an overview first, and then filter and drill down according to their own interests. An additional goal was that Peripleo should function as a service, and provide an API that enables other sites to re-use Pelagios data, either by building their own mashups or by embedding and linking to Peripleo search results easily. A short example below will illustrate how Peripleo works.

Fig. 1: Peripleo search result example.

Fig. 1 shows the result of a search for the term "tetradrachm" (an ancient Greek coin type). The search, in this case, produces a total of 21,576 results. The map shows a distribution of dots that indicate where those results are located. Peripleo also shows preview images for some of the results. Clicking on the "filters" button reveals additional information about how our result set is organized (Fig. 2). For example, we can see that most results come from the American Numismatic Society Collection [9]. A small fraction (hovering the mouse over the bars we can find out it is nine results) originates from the Fralin | UVa Art Museum [10]. Peripleo also shows how the results are distributed over time, between 545 BC and 280 AD, with a peak of finds dated to around 300 BC.
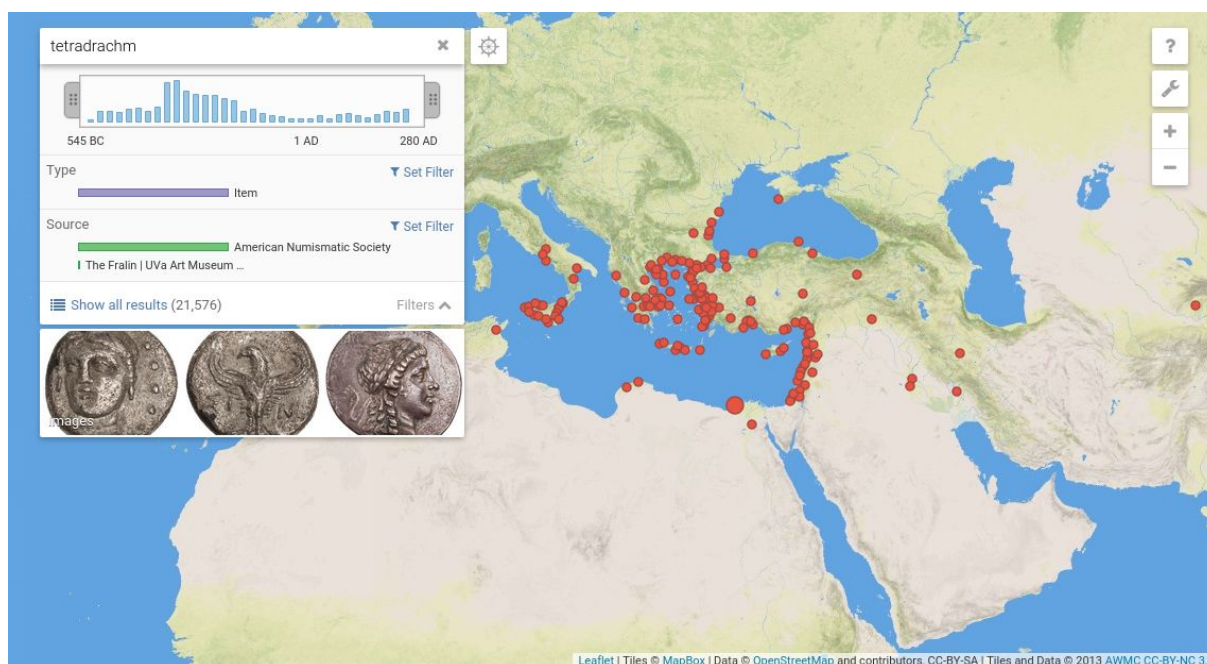

Fig. 2: Peripleo search result example, with open filters panel.

By zooming and panning the map, it is possible to explore the distribution of the search results in more detail. The information shown in the panels updates in real time. The total result count, temporal coverage, thumbnail previews, the sizing of the map markers which indicates the relative number of results at a place vs. others on the map – everything reflects the situation in the currently viewed map area. Fig. 3, for example, shows that the region around Sicily holds a subset of 898 results out of the total of 21,576. It also shows how the local temporal distribution differs from the global distribution, being limited to the time range from 500 to 200 BC (as opposed to 600 to 300 AD globally).

In a similar fashion to how Peripleo facilitates the exploration of spatial sub-regions of the result set (by zooming and panning the map), it also enables temporal investigation. Using handle controls, a specific range can be selected in the timeline graphic. Dragging the selection interval across the timeline will, as before, update all the information in real time: result counts, geographic distribution, thumbnail images, etc. As an example, the sequence of images in Fig. 4 shows how the spatial "footprint" of the "tetradrachm" search results exhibits a distinct and changing pattern over the period for which we have evidence of the coins' find spots.
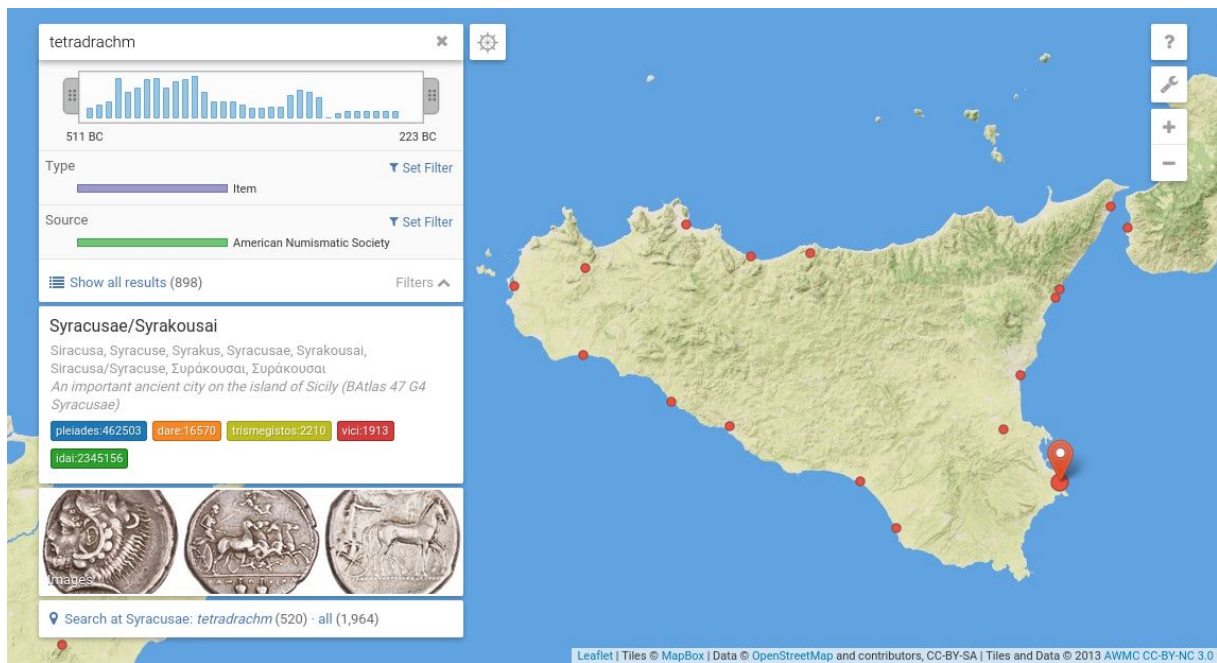
Fig. 3: Results for the area around Sicily, with Syracusae as local center (indicated by largest dot).

Peripleo features a range of additional capabilities. For example, it is not necessary that each object corresponds to exactly one point location on the map, as in the previous example. It is perfectly possible for a single object to be linked to many – even thousands of – locations. This is, in fact, a typical case for literary texts, where Peripleo can even provide full-text search within the document, and render maps based on the places occurring most closely to the search term (cf. Fig.5).
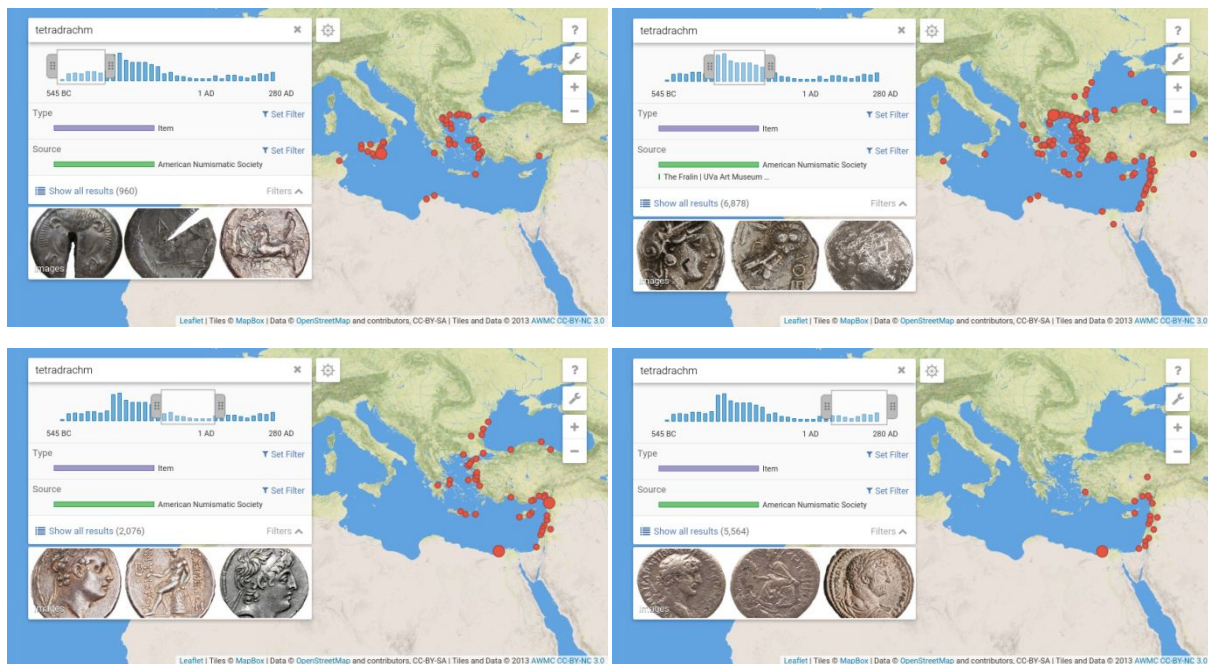

Fig. 4: Exploring changing search result distribution over time, by using the time range selection tool.

Furthermore, Peripleo is not restricted to representing places as points. Place geometries can be rendered as point, line, or polygon features. Peripleo makes use of this information for the purposes of not only visualization, but also spatial ranking. This way, when focusing the map on Crete, for example, the gazetteer record for the island will be ranked among the top relevant objects; whereas, when zooming out to view the entire Mediterranean and near-eastern area, a literary geographic work like Herodotus's *Histories* is much more likely to rank among the top hits, as its geographical coverage is a much better match to the currently viewed map area.
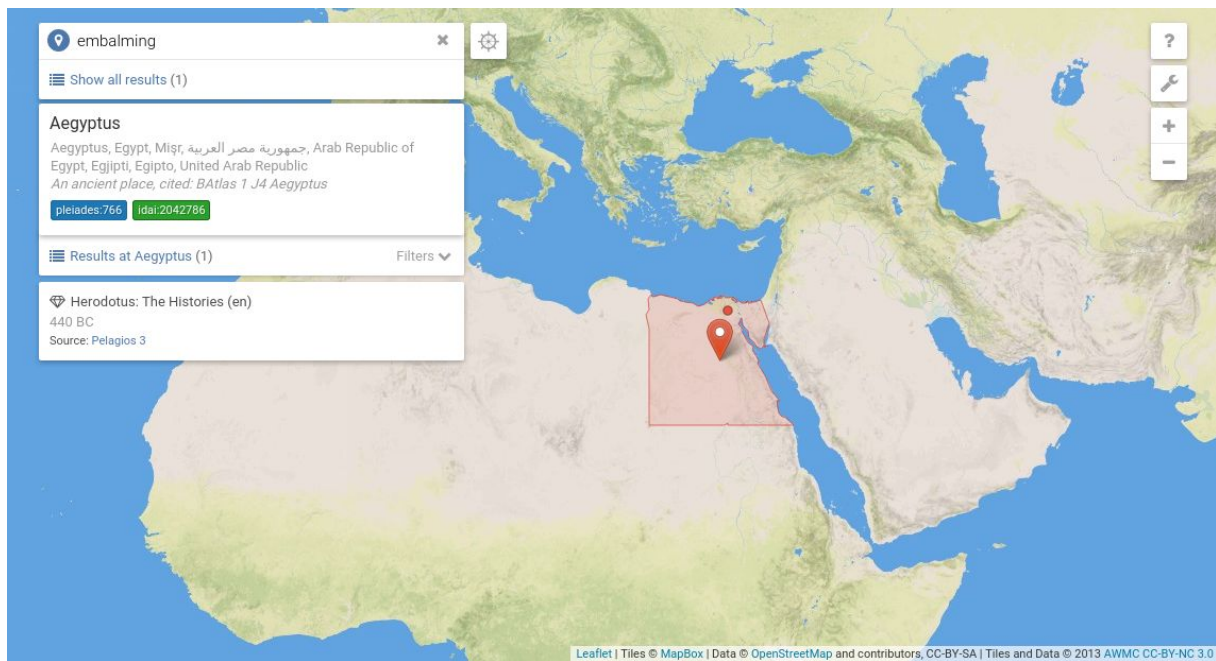

Fig. 5: A search for the term "embalming" returns a single result - *The Histories* by Herodotus.
The map shows places that appear close to the search term in the text.

Technical Architecture

Peripleo is Open Source software, available through our GitHub repository [11]. The technical architecture follows a state of the art 3-tier Web application model. It is implemented on a JVM (Java Virtual Machine) technology stack, using the Play application framework [12] and the Scala programming language as a basis for the "middle tier", which provides, primarily, a JSON API. The presentation tier is a client-side JavaScript/HTML5 single page application built on top of this API. At the same time, the API is also open to the public, and can be used freely by external developers to implement their own Peripleo-like interfaces, or re-use and integrate Peripleo data into their own applications.

In terms of storage, Peripleo implements a hybrid approach: a relational PostgreSQL database holds detailed record information, as well as auxiliary tables to speed up specific kinds of queries or drive special detail visualizations (e.g. network visualizations of places co-occurring in documents). The primary storage backend, however, is a set of indices (based on the Open Source Lucene indexing engine), providing fast access to: i) place information; ii) document and item metadata; iii) document full-text search; and iv) auto-completion hints. Peripleo makes particular use of spatial indexing functionality and different types of faceting approaches (taxonomy

faceting, temporal faceting, 2D spatial faceting) to enable real-time visualization of the thematic, temporal and spatial composition of large search result sets.

Integrating Peripleo

Making data that differ vastly in terms of size, content type and theme uniformly accessible under a single user interface requires agreements on some of the basic principles of how we express it. Pelagios's approach to this challenge is simple and pragmatic: rather than getting everyone to agree on how to *represent* the data, Pelagios provides a set of lightweight conventions for how to express *links* between the data and the things described in it. We refer to this approach as "connectivity through common references". Pelagios uses this approach specifically with regard to places. But the approach as such is, of course, equally applicable to other "ordering dimensions" such as people, time periods, events or classification schemes.

In principle, any object that is published online, under a stable URI – and which contains references to places – can be made discoverable through Peripleo. What is needed for integration is a *dataset summary*: a data file that lists all objects, along with basic Dublin Core metadata (title, description, date, provenance information, etc.), and information about the places each object is related to (and, optionally, how it is related to them). To encode the latter, we have chosen the Open Annotation Data Model [13]. The metaphor of annotation is not only appropriate for the act of identifying (or "tagging") a place reference in arbitrary digital content with a gazetteer URI. It also has the connotation that, in general, the identification (or "tag") is not to be considered certain fact, but rather that someone (a human editor, an automated geo-parsing script) is making a claim about some kind of relationship between part of the source document and the place. The RDF snippet below provides a "minimum dataset summary" example of a dataset with a single object, expressed in RDF/Turtle notation. (Full documentation is available at [14].)

```
@prefix cnt: <http://www.w3.org/2011/content#> .
@prefix dcterms: <http://purl.org/dc/terms/> .
@prefix foaf: <http://xmlns.com/foaf/0.1/> .
@prefix oa: <http://www.w3.org/ns/oa#> .
@prefix pelagios: <http://pelagios.github.io/vocab/terms#> .
@prefix relations: <http://pelagios.github.io/vocab/relations#> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema> .

# An object you want to link to Pelagios
<http://example.org/pelagios/dump.ttl#items/001> a pelagios:AnnotatedThing ;

  # Title and homepage URL are MANDATORY
  dcterms:title "Honorific inscription of Ostia" ;
  foaf:homepage <http://edh-www.adw.uni-heidelberg.de/...> ;

  # Everything else OPTIONAL (but highly encouraged)
  dcterms:description "Honorific inscription, findspot Ostia" ;

  # Use ISO 8601 (YYYY[-MM-DD) or time interval (<start>/<end>) for dates
  dcterms:temporal "366/402" ;

  # We encourage the use of PeriodO identifiers to denote time periods
```

```
    dcterms:temporal <http://n2t.net/ark:/99152/p03wskd389m> ; # Greco-Roman

    # Use RFC 5646 for the object's language (e.g. literature, inscriptions, etc.)
    dcterms:language "la" ;

    # Feel free to assign 'tags' to your data
    dcterms:subject "inscription" ;
    .

# Objects are 'annotated' with any number of gazetteer references
<http://example.org/pelagios/dump.ttl#items/00l/annotations/1> a oa:Annotation ;

    # MANDATORY: the 'annotation target' is the URI of your object;
    oa:hasTarget <http://example.org/pelagios/dump.ttl#items/00l> ;

    # MANDATORY: the 'annotation body' is the gazetteer reference
    oa:hasBody <http://pleiades.stoa.org/places/422995> ;

    # OPTIONAL: extra metadata about the nature of the place reference
    pelagios:relation relations:foundAt ;
    oa:hasBody [ cnt:chars "POINT (41.755740099 12.290938199)";
                 dcterms:format "application/wkt" ] ;
    oa:annotatedAt "2014-11-05T10:18:00Z"^^xsd:date ;
    .
```

Along with the dataset summary, Peripleo also requires a small RDF file that describes the dataset as a whole (what's inside, who the publisher is, under what license the metadata is published, etc.) and which acts as machine-readable entrypoint to the - potentially large - summary file. We use the RDF *Vocabulary of Interlinked Datasets* (VoID) [15] to encode this information. A minimum example is shown below.

```
@prefix : <http://my-domain.org/my-data/> .
@prefix dcterms: <http://purl.org/dc/terms/> .
@prefix foaf: <http://xmlns.com/foaf/0.1/> .
@prefix void: <http://rdfs.org/ns/void#> .

:my-dataset a void:Dataset;
  dcterms:title "My Archaeological Dataset";
  dcterms:publisher "My Institution or Project";
  foaf:homepage <https://my-domain.org/>;
  dcterms:description "A dataset of archaeological items.";
  dcterms:license <http://opendatacommons.org/licenses/by/>;

  # This is VERY important - location of the dataset summary file
  void:dataDump <http://my-domain.org/downloads/pelagios.ttl> ;
  .
```

Depending on the nature of the data, different gazetteers may be suitable as annotation vocabulary. In our current demo installation of Peripleo we have, so far, included seven gazetteers from our partners: the Pleiades Gazetteer of the Ancient World [16], the Digital Atlas of the Roman Empire [17], Vici.org [18], iDAI.gazetteer [19], the China Historical GIS [20],

nomisma.org [21], and place records from the Trismegistos project [22], along with all links to GeoNames, Wikipedia and Wikidata that these gazetteers have included in their data exports. Importing gazetteers works in a similar fashion to adding data; that is to say, through a lightweight "gazetteer summary" data format. It is therefore perfectly possible to set up an installation of Peripleo with a different combination of gazetteers, or a single institutional one. A minimum example for a gazetteer summary with a single place is shown in the RDF snippet below. (Full documentation is available at [23].)

```
@prefix cito: <http://purl.org/spar/cito> .
@prefix cnt: <http://www.w3.org/2011/content#> .
@prefix dcterms: <http://purl.org/dc/terms/> .
@prefix foaf: <http://xmlns.com/foaf/0.1/> .
@prefix geo: <http://www.w3.org/2003/01/geo/wgs84_pos#> .
@prefix geosparql: <http://www.opengis.net/ont/geosparql#> .
@prefix gn: <http://www.geonames.org/ontology#> .
@prefix lawd: <http://lawd.info/ontology/> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix skos: <http://www.w3.org/2004/02/skos/core#> .

<http://www.mygazetteer.org/place/Athens> a lawd:Place ;

  # Don't think of label and description in terms of a 'primary name' or
  # detailed abstract. Think of it in terms of UI: what do you want users
  # to see about your place in a list of search results?
  rdfs:label "Athens"@en ;
  dcterms:description "A major Greek city-state"@en ;

  # Optional: a present-day (ISO-3166 alpha2) country code
  gn:countryCode "GR" ;

  # Don't think of dates in terms of 'how long the place existed'. Use dates
  # to specify the period your gazetteer is concerned with the place and/or
  # provides attestations for it.
  # Use ISO 8601 (YYYY[-MM-DD]) or time interval (<start>/<end>).
  dcterms:temporal "-750/640" ;

  # We encourage the use of PeriodO identifiers for time periods
  dcterms:temporal <http://n2t.net/ark:/99152/p03wskd389m> ; # Greco-Roman

  # Use skos:closeMatch to express 'vague' matches, e.g. to link to a
  # modern-day town now located there
  skos:closeMatch <http://sws.geonames.org/264371/> ;

  # Use skos:exactMatch to express (geographical, temporal, cultural) identity
  skos:exactMatch <http://pleiades.stoa.org/places/579885> ;

  # Express names for the place using lawd:variantForm. For language encoding,
  # use RFC 5646.
  lawd:hasName [ lawd:primaryForm "Athenae" ] ;
  lawd:hasName [ lawd:primaryForm "Athens"@en ];

  # You can provide attestions (e.g. bibliographic reference) for individual
  # names as in the example below).
  lawd:hasName [
    lawd:primaryForm "Aθήνα"@el ;
    lawd:hasAttestation <http://www.mygazetteer.org/att/0001>
```

```
  ] ;

  # OPTIONAL: a representative point coordinate
  geo:location [ geo:lat 5.16 ;  geo:long 52.05 ] ;

  # OPTIONAL: detail geometry as WKT string.
  # Alternatively, use osgeo:asGeoJSON for a GeoJSON string
  geosparql:hasGeometry [
    geosparql:asWKT "LINESTRING (5.16 52.05, 5.17 52.05, 5.16 52.06)" ;
  ] ;

  foaf:primaryTopicOf
    <http://www.mygazetteer.org/place/Athens.html> ;

  # OPTIONAL: express hierarchical relations, if they exist in your gazetteer
  dcterms:isPartOf <http://www.mygazetteer.org/place/Greece> ;
  .

# Example: metadata of an attestation
<http://www.mygazetteer.org/att/0001> a lawd:Attestation ;
  dcterms:publisher <http://www.mygazetteer.org/> ;
  cito:citesAsEvidence <http://www.mygazetteer.org/documents/01234> ;
  cnt:chars "Α θ ή ν α "
  .
```

Dataset and gazetteer summaries can be published separately from the source data. This approach is sometimes referred to as "standoff markup" [24], and helps to avoid data management problems which can arise when annotations have to be natively incorporated into an existing data model (potentially introducing changes to internal metadata schemas and database implementations). In addition, it is possible to publish a summary as a single dump file (i.e. an extra static "file download" placed somewhere on the institutional Web server), with the result that no changes to existing Web presences or APIs are necessary.

Outlook

Peripleo is prototypical rather than finished software. But we are convinced that an interface which empowers users to tap into and navigate through highly heterogeneous online collections is crucial in order to demonstrate the utility of lightweight linking approaches, and to make their benefits more tangible to end-users and non-technical specialists. On a technical level, we feel that the loose coupling that exists between Peripleo and the resources it makes searchable is very much in the spirit of Linked Open Data, in contrast to more tightly coupled traditional search frontends which are typically bound to a specific repository.

As regards future steps, we plan to extend Peripleo to make use of other Linked Open Data initiatives as and when they emerge. Priority will be given to supporting temporal classification schemes like PeriodO [25], and person authority schemes such as SNAP [26], as well as customizable type classifications. Another future work item pertains to scalability in terms of the amount of data. We are confident that the present architecture can support several times the number of records we are presently hosting in our demo installation (approx. half a million). But it

remains to be tested how considerably larger datasets (coming, for example, from museum collections) would affect query response times and visualization performance. Inevitably, the single-index backend that drives the bulk of the Peripleo frontend interaction will have to be replaced. To this end a solution that supports distribution across a cluster is being prepared.

Acknowledgements

References

[1]   Pelagios Project blog. http://pelagios-project.blogspot.co.uk (last accessed December 2015)

[2]   Elliott, T. and Gillies, S. 2009. Digital Geography and Classics. In Digital Humanities Quarterly. Vol 3. Number 1. http://www.digitalhumanities.org/dhq/vol/3/1/000031.html (last accessed December 2015)

[3]   To excerpt from the introductory chapters of *Berman, Mostern & Southall, Placing Names: Enriching and Integrating Gazetteers, Indiana University Press* (in print): *"Most simply, a gazetteer is a list of places." [...] "In recent decades, and especially in a computing context, gazetteers have often been defined as simple pairings of place names and coordinates, sometimes including feature types as well." "[...] gazetteers are the basis for much of the spatial search and visualization that specialists and the public have come to take for granted." [...] "The future of [gazetteers] is online and sharing-centric, following a general development trend that is generally known as Linked Open Data." "[...] all sharing-centric repositories must be online, and have a way to uniquely identify every data record [by means of a Uniform Resource Identifier (URI)]."*

[4]   Isaksen, L., Simon, R., Barker, E. and de Soto Cañamares, P. 2014. Pelagios and the Emerging Graph of Ancient World Data. In Proceedings of the 2014 ACM Conference on Web Science, pp. 197-201.

[5]   Simon, R., Isaksen, L., Barker, E. and de Soto Cañamares, P. 2015. The Pleiades Gazetteer and the Pelagios Project. In Placing Names: Enriching and Integrating Gazetteers. Berman, M. L., Mostern, R. and Southall, H. (Eds.) Indiana University Press (in print).

[6]   Bizer, C., Heath, T., Berners-Lee, T. 2009. Linked Data – The Story So Far. In International Journal on Semantic Web and Information Systems, 5(3): 1-22.

[7]   Simon, R., Barker, E., Isaksen, L. and de Soto Cañamares, P. 2015. Linking Early Geospatial Documents, One Place at a Time: Annotation of Geographic Documents with Recogito. In e-Perimetron. Vol.10, No.2 (2015), pp. 49-59. ISSN 1790-3769.

[8]   Peripleo public demo. http://pelagios.org/peripleo/map (last accessed December 2015)

[9]   American Numismatic Society Collection Database. http://numismatics.org/search (last accessed December 2015)

[10]   The Fralin | UVa Art Museum Numismatic Collection. http://coins.lib.virginia.edu/ (last accessed December 2015)

[11]   Pelagios GitHub repository: Peripleo. http://github.com/pelagios/peripleo (last accessed December 2015)

[12]   *Play* Web application framework http://www.playframework.com (last accessed December 2015)

[13]   Sanderson, R., Ciccarese, P. and Van de Sompel, H. (eds.). 2013. Open Annotation Data Model. Community Draft, 08 February 2013. http://www.openannotation.org/spec/core/ (last accessed December 2015)

[14]   Joining Pelagios - documentation of Pelagios dataset and metadata description RDF profile. https://github.com/pelagios/pelagios-cookbook/wiki/Joining-Pelagios (last accessed December 2015)

[15]   Describing Linked Datasets with the VoID Vocabulary. W3C Interest Group Note 03 March 2011. http://www.w3.org/TR/void/ (last accessed December 2015)

[16]   Pleiades Gazetteer of the Ancient World. http://pleiades.stoa.org/ (last accessed December 2015)

[17]   Digital Atlas of the Roman Empire. http://dare.ht.lu.se/ (last accessed December 2015)

[18]   Vici.org. Archaeological Atlas of Antiquity. http://vici.org (last accessed December 2015)

[19]   iDAI.gazetteer, German Archaeological Institute. http://gazetteer.dainst.org (last accessed December 2015)

[20]   China Historical GIS http://www.fas.harvard.edu/~chgis/ (last accessed December 2015)

[21]   nomisma.org. http://nomisma.org (last accessed December 2015)

[22]   Trismegistos project. http://www.trismegistos.org (last accessed December 2015)

[23]   Pelagios Gazetteer Interconnection Format. https://github.com/pelagios/pelagios-cookbook/wiki/Pelagios-Gazetteer-Interconnection-Format (last accessded December 2015)

[24]   Thompson, H.S. and McKelvie, D. 1997. Hyperlink semantics for standoff markup of read-only documents. In Proceedings of SGML Europe 97. p. 227-229.

[25]   Rabinowitz, A. 2014, It's about time: historical periodization and Linked Ancient World Data. In ISAW Papers 7. http://dlib.nyu.edu/awdl/isaw/isaw-papers/7/rabinowitz/ (last accessed December 2015)

[26] Standards for Networking Ancient Prosopographies (SNAP) project Website. http://snapdrgn.net/ (last accessed December 2015)