8. The Pleiades Gazetteer and the Pelagios Project

Rainer Simon, Leif Isaksen, Elton Barker, Pau de Soto Cañamares

Introduction

Pelagios is a community-driven initiative that facilitates better linkage between online resources documenting the past, based on the places that 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" (Elliot and Gillies, 2009) – 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.

The Mantra of Pelagios: Connectivity through Common References

Online resources relevant to those with an interest in the past are multiplying rapidly. Large-scale digitization campaigns by major memory institutions and search engine giants, as well as big meta-aggregation portals like Europeana and the Digital Public Library of America are, of course, the most prominent examples in this trend. But there are countless smaller initiatives that, while receiving far less public visibility, often have depth and specialization. They may be the result of a research project or thesis, supported within an academic institution, or may be realized by the efforts of a committed group of likeminded individuals. Surfing the Web in search of such

resources is one thing; it is an entirely different matter, however, to be able to seamlessly navigate between data from different sources, to compare and combine it, and to re-use it in one's own tools and workflow, or serendipitously stumble on related material that may fill in gaps in one's knowledge or point to other salient resources.

Making digital data fit together like the pieces of a puzzle requires agreements on some of the basic principles of how we share 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 things described in it. We refer to this approach as *connectivity through common references*, and the benefit is twofold. First, the approach is applicable to any type of heterogeneous digital content – text, images, media, 3D objects, online databases, etc. Second, it minimizes the entry threshold for participants by placing no requirements concerning the adherence to specific metadata schemas, data models, vocabularies or technical implementations with regard to the data. It is important to note that Pelagios does not, of course, preclude the use of such standard models. In fact, it can serve as a perfect complement (and potentially even entry path) to shared reference models such as the CIDOC CRM (Crofts, *et al*, 2011).

The key to connectivity in Pelagios is the use of common gazetteer references when referring to places. Phrased in more technical terms, the first Pelagios convention states that whenever you refer to a place in your data, you should do so using a gazetteer URI. Such place references could express the find spot of a particular item in an archaeological database; mark up a piece of literature or a research article; identify a toponym in a digitized old map; or record the location

of a historic site depicted on a photograph. By expressing them in the form of URIs from a shared gazetteer, otherwise isolated datasets are implicitly joined up to an interconnected graph, with the gazetteer as its central backbone.

The second Pelagios convention is that the resulting place metadata must be published online as open data, according to a common technical serialization format. For this purpose we have chosen the Open Annotation Data Model (Sanderson, *et al*, 2013). The metaphor of *annotation* is not only appropriate for the act of identifying (or 'tagging') a place reference in arbitrary digital content. 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. It is thus an assertion of an interpretation. Open Annotation includes support for recording this kind of provenance information, making it a suitable publishing format not only for the place references themselves, but also for the metadata that needs to accompany them in a scholarly scenario.

Another significant aspect is that the place metadata can be stored separate from the source data. This approach is sometimes referred to as "standoff markup" (Thompson and McKelvie, 1997), and helps to avoid the 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). Additionally, for this approach only a single dump file (i.e. an extra static "file download" placed somewhere on the institutional Web

server) with the place metadata is needed, so no changes to existing Web presences or APIs are necessary.

A Brief History of Pelagios

In its initial phases, supported financially with two grants from the UK Joint Information Systems Committee (JISC) between 2011 and 2012, Pelagios established basic practices and developed prototypical applications to demonstrate how online resources documenting the past can be linked together via common place identifiers (Simon, Barker and Isaksen, 2012). The aggregated place metadata from all members was also made searchable via the *Pelagios API*,[1] a demonstration service that provides lookup functionality for place references in the Pelagios network, along with an *Application Programming Interface* that enables Web developers to build "Pelagios mashups" (i.e. their own Web applications that incorporate data from Pelagios, as well as other sources).

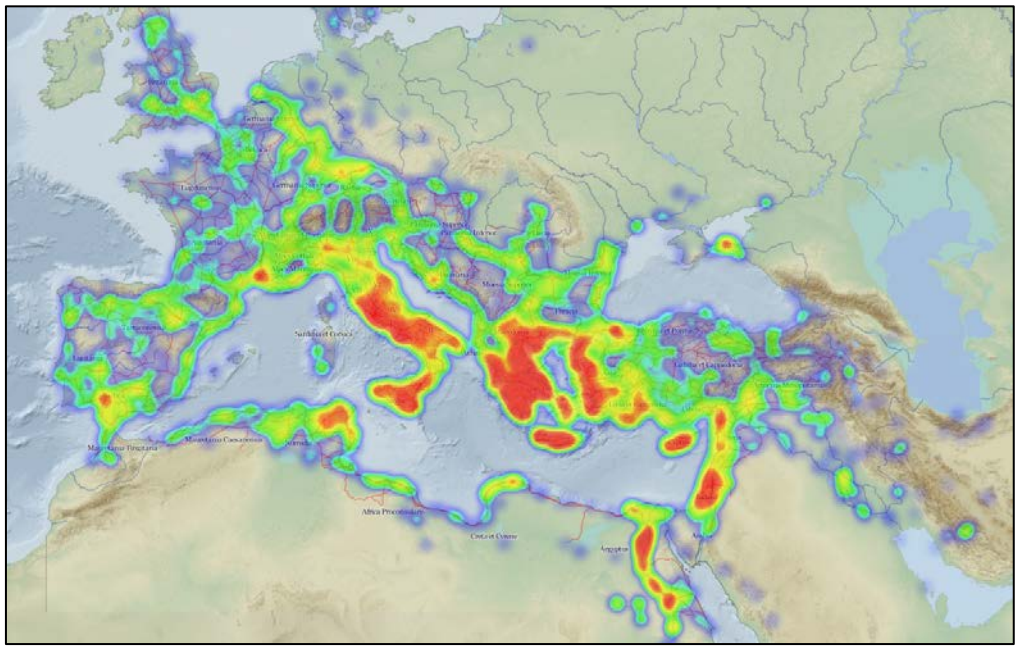

Figure 8.1 Heatmap of place references suggested in the initial two phases of Pelagios. Red areas represent regions with highest densities of place references. Background map: Digital Atlas of the Roman Empire (http://imperium.ahlfeldt.se/).

In these two initial phases, Pelagios had a specific thematic focus on classical antiquity. This was not least due to the fact that for this period of time and geographic area, a suitable, focused historical URI-based gazetteer exists already – and is widely acknowledged among the research community: the Pleiades Gazetteer of the Ancient World.[2] Pleiades provides URIs for more than 34,000 places in the Greco-Roman world, and is thus exactly the kind of shared referencing system to enable *connectivity through common references*. By promoting the use of Pleiades as a common vocabulary for place references, Pelagios has fostered a growing community of some 30 Ancient World projects who, at the time of writing, have collectively interlinked their resources through approx. 800,000 annotations. Fig.1 shows a map that

illustrates the geographic distribution and density of the place references aggregated during Pelagios's initial phases. An interactive version of this map is available online.[3]

Pelagios has recently entered a new phase of development ("Pelagios 3") with support from the Andrew W. Mellon Foundation. This phase, which started in September 2013 and will run until August 2015, significantly expands the scope of Pelagios. First, in space and time: Pelagios 3 addresses the corpus of *Early Geospatial Documents*, documents that use written or visual representation to describe geographic space prior to the European discovery of the Americas in 1492, an event which not only radically transformed beliefs about the globe, but triggered the development of several standardising global cartographic conventions in the following century – including the Werner, Bonne and Mercator projections. Early Geospatial Documents include ancient and medieval geographic descriptions (geographiae, chorographiae and itineraries), world maps (mappaemundi) and sea charts, and are products of Greek, Roman, Christian, Islamic and Chinese traditions.

The second major difference to prior phases of Pelagios is that the new phase is actively annotating new source documents, rather than purely facilitating interlinking between existing ones. This means that a significant part of the project is devoted to developing new annotation tools and infrastructure, e.g. for extracting place name data from digitized texts and maps semi-automatically. A further difference concerns a major architectural consequence arising out of the thematic expansion beyond antiquity – Pleiades alone is no longer sufficient as a shared referencing system. First of all, Pleiades does not provide the global geographical coverage that is needed for Pelagios 3; and even in those regions that Pleiades *does* cover, it may not always be

appropriate to use it. For example, it may be perfectly acceptable under some circumstances to equate the ancient city of *Gades* – as identified in Pleiades – with the medieval town of *Cádiz* and even modern-day *Cádiz*. However, there is clearly a significant semantic difference between the three of them, and it is necessary to retain this difference in the context of a project like Pelagios 3. A specialist gazetteer like Pleiades, which focuses on a specific temporal or cultural milieu, inherently captures this difference with explicit semantics. As a result, Pelagios 3 needs to work with multiple gazetteers in combination, and had been producing requirements and methods for linking and interoperating between them. Key partners of the Pelagios 3 consortium in this regard are *PastPlace* [4], who will provide the gazetteer for annotating medieval European materials, and the China Historical GIS. [5]

Growing the Graph of Linked Ancient Geo-Data

Pelagios 3 contributes to the growing Web of Linked Open Data for the Ancient World (Elliott, Heath, and Muccigrosso, 2014) by aiming to produce, first, a network of associations that connect **documents** to **places** relevant to them. Pelagios deliberately makes no strong assumptions about the nature of the "documents" in this regard: they could be literary texts, inscriptions, maps, even physical artefacts like archaeological finds. Nor does Pelagios place any particular requirements on their digital representation: a digital text or an image, a Web page, a 3D object – as long as there *is* a representation online, available under a stable URI, the principle of connectivity through common references is applicable. As explained above, we create and share the associations through *annotations*; and use *Open Annotation* as the syntax to encode them on a technical level. It should also be pointed out that the nature of the associations

themselves (i.e. the reason *why* a place is relevant to a document) can be diverse: a museum object may have been found at that place; a certain photograph may have been taken there. Sometimes – but not always – it may be that the document has relevance to the history of the place, too (and not just the other way round): e.g. in the case of an archaeological artefact that is inscribed with a written attestation to the place. Again, while Pelagios provides the syntactic means to include such extra metadata as part of the annotation, it does not place mandatory requirements in this regard.

Second, Pelagios aims to produce a network of associations that connect **places** in one gazetteer to corresponding **places** in (an)other gazetteer(s). This is a consequence of our need to annotate documents with URIs from different gazetteers, depending on the document's temporal and cultural setting. Of course, Pelagios could just treat its different gazetteers of choice as isolated datasets, each providing their own separate pool of URIs to tag with. But this way, the annotations would generally remain isolated along the same lines, and we would miss the opportunity to grow a globally navigable graph of documents and places. Furthermore, communities maintaining specialist gazetteers increasingly include mappings between their places and other gazetteers already: Pleiades, for example, includes links not only to the global gazetteer GeoNames[6] but also to Wikipedia which can be considered a "gazetteer-like dataset", too. It seems only sensible to make use of such existing links, and evolve this practice into a more general framework for gazetteer alignment. It is important to note that we do **not** use annotations to express this second type of associations. One could argue that they do share some of the characteristics which we attribute to Pelagios annotations: e.g. they are also statements

(about equivalence) made by a human or machine process, rather than a undisputable fact; and that the nature of the association may vary (i.e. truly identical places vs. just reasonably close matches, e.g. between an ancient place and the modern day city now located on its former foundations). However, we believe that the conceptual metaphor of *annotation* starts to break when one would essentially be annotating "one place with another"; and that the use case of aligning gazetteer records is sufficiently specific and well-defined to merit a native – and potentially less ambiguous and verbose – gazetteer-centric approach.

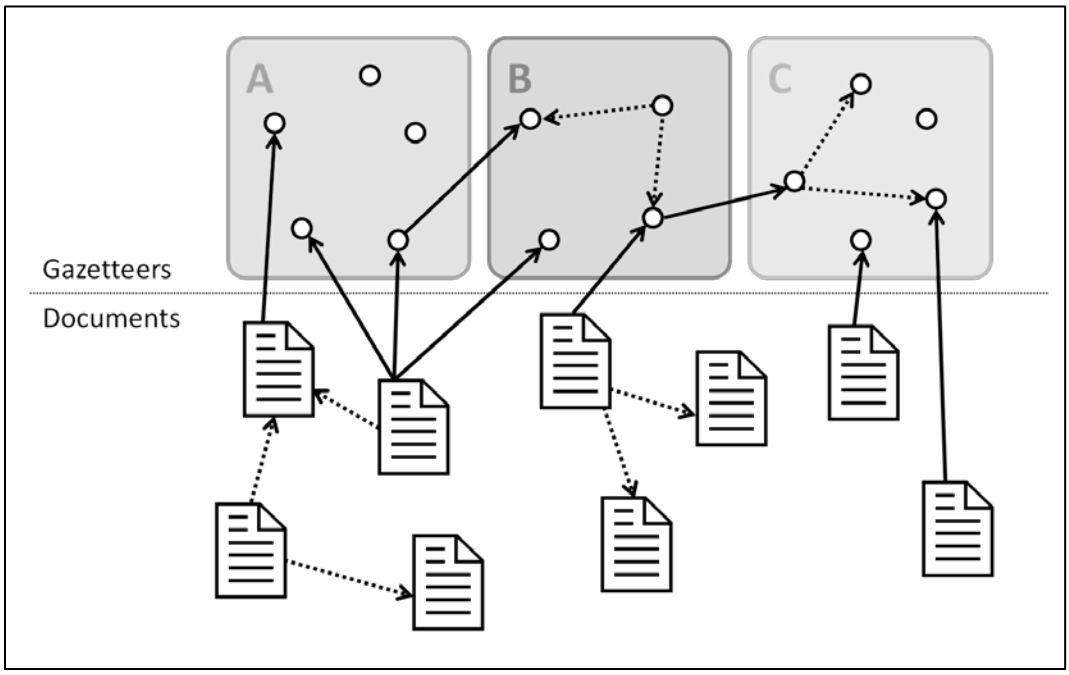

Figure 8.2 Pelagios linking documents and gazetteers

Figure 8.2 illustrates the two types of links graphically, and shows how these links collectively help to grow a graph that globally spans multiple datasets and gazetteers. The connections which already exist *within* each dataset are shown as dotted lines. Pelagios adds new connections,

shown as solid lines, which either link *documents* to *places* (crossing the Documents/Gazetteers boundary) or link *places* to *places* (crossing the boundary between gazetteers A, B or C).

It should be pointed out that the strict distinction between gazetteers and documents as illustrated in Figure 8.2 has its own pitfalls. Experience from the first two phases of Pelagios has shown that it can be surprisingly difficult to decide which category a certain dataset falls into. Prime examples for this are "encyclopedic", Wiki-like datasets which consist, for example, of Web pages dedicated to specific ancient places. Should these be considered to be "documents" about places? Or should they rather be treated as gazetteers in their own right? Some datasets also exhibit a hybrid structure internally, with one part modeling entities like places or persons – and thus functioning as a gazetteer – and another part representing documents or items, linked to that gazetteer. Ultimately, one must acknowledge that the lines we draw in Pelagios are blurry, and that the appropriate strategy for a particular dataset must be decided on a case-by-case basis. Nonetheless, the above distinction has so far provided us with a reasonable division of our "problem space", and served as a useful abstraction for the purposes of system implementation and user interface development. The following two subsections will go into more technical detail as to how we currently implement the two types of associations, and how they are expressed according to the Resource Description Framework (RDF) data model.

Linking Documents to Places: Pelagios Annotations

Figure 8.3 shows a screenshot from *Recogito*, [7] a tool we have been developing in Pelagios 3 specifically for the task of annotating toponyms in documents. Presently, *Recogito* allows us to

work with texts only, but it will be extended to support images as well, once the project reaches phases that deal with map imagery. Our general workflow in *Recogito* starts with a plain text document. In cases where the text is an English translation, we also perform automatic geo-parsing using an open-source Named Entity Recognition engine.[8] To each toponym identified we assign an initial gazetteer match (based on string similarity and basic geographical disambiguation rules such as distance to the previous identified toponym in the document). After this, we can work with the document in two kinds of user interfaces: in a text-based view (see Figure 8.3), which is designed for annotation of the text (or verifying and correcting the automatic results produced earlier); and in a map-based view in which we can more easily assign gazetteer URIs to the toponyms annotated earlier.

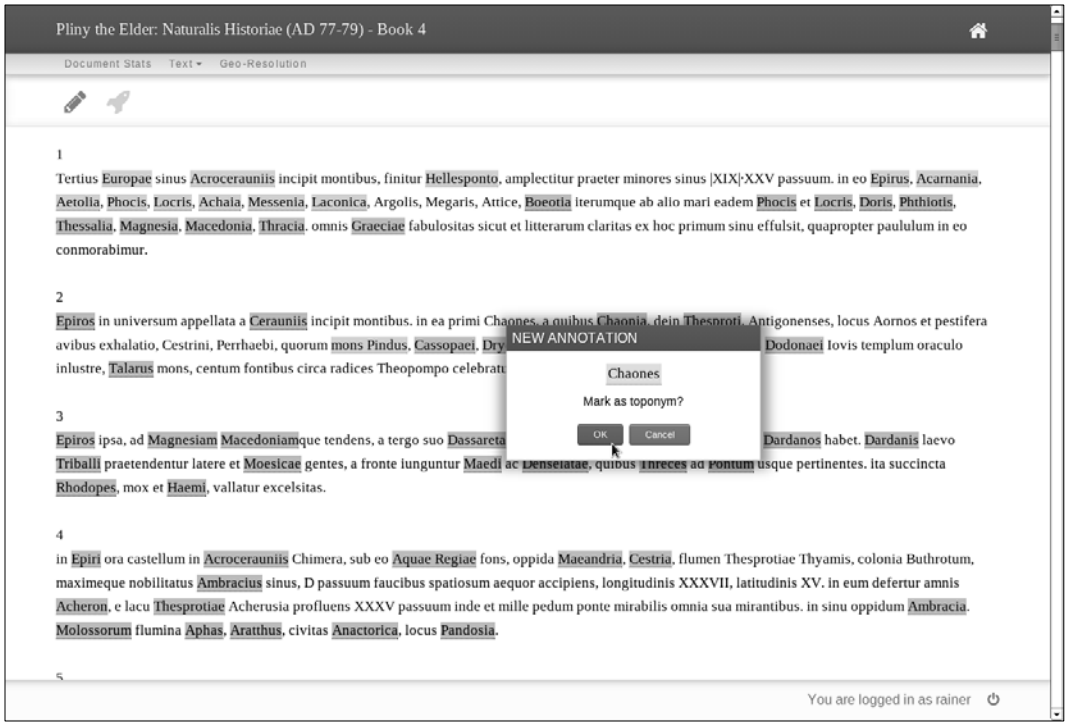

Figure 8.3 Pelagios' *Recogito* geo-annotation tool (text annotation view)

Recogito represents just one, project-specific example of how annotations can be prepared. Irrespective of tools and workflows, however, the key requirement for inclusion in Pelagios is for the production of annotations that can be published in Open Annotation format, as described above. Open Annotation is RDF-based; and Fig. 4 provides an example of how a single toponym from our source text in Figure 8.3 can be expressed in Pelagios-compliant Open Annotation format.

```
@prefix cnt: <http://www.w3.org/2011/content#> .
@prefix dcterms: <http://purl.org/dc/terms/> .
@prefix dctype: <http://purl.org/dc/dcmitype/> .
@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 relation: <http://pelagios.github.io/vocab/relations#> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema> .

# Annotations may include details about the annotator (person or automated process)
<http://pelagios.org/downloads/latin/egd01.ttl#agents/rainer> a foaf:Person ;
 foaf:name "Rainer Simon" ;
 foaf:workplaceHomepage <http://www.ait.ac.at> ;
 .

# First, we need to describe the annotated document
<http://pelagios.org/recogito/api/documents/22>
 a pelagios:AnnotatedThing ;      # We need to type the document as an "annotated thing"
 a dctype:Text ;                  # Additionally (and optionally) we can be more specific

 dcterms:title "Naturalis Historiae" ;   # A dcterms:title is mandatory

 dcterms:author "Pliny the Elder" ;      # All other metadata is optional & unrestricted!
 dcterms:issued "start=77; end=79;" ;
 dcterms:identifier <http://www.wikidata.org/wiki/Q442> ;
 dcterms:description "An encyclopedia published circa AD 77–79 by Pliny the Elder." ;
 foaf:primaryTopicOf <http://en.wikipedia.org/wiki/Natural_History_(Pliny)> ;
 dcterms:language "lat" ;
 .

# Then, we list all annotations on this document (only one in this example)
<http://pelagios.org/recogito/api/annotations/26b1c3d4-caee-4222-9625-c3f6fe52c7e5>
 # Type, target (= document) and body (= gazetteer URI) are mandatory
 a oa:Annotation ;
 oa:hasTarget <http://pelagios.org/recogito/api/documents/22> ;
 oa:hasBody <http://pleiades.stoa.org/places/481787> ;

 # Optionally, we can add extra annotation metadata
 oa:hasBody [ cnt:chars " Chaones" ; a pelagios:Toponym ] ;
 pelagios:relation relation:attestsTo;
 oa:annotatedBy <http://pelagios.org/downloads/latin/egd01.ttl#agents/rainer> ;
 oa:annotatedAt "2014-04-03T10:18:00Z"^^xsd:date ;
 .
```

Figure 8.4 . A document with one added Pelagios annotation: RDF (Turtle) example

Linking Places to Places: Pelagios Gazetteer Interconnection Framework

Moving from a single gazetteer to an ecosystem of multiple gazetteers has significant consequences. Generally gazetteers can vary widely in how they represent places conceptually and syntactically – with different abstractions, relations and hierarchy models and different approaches to expressing changes over time or to recording the source or bibliographic references that lead to the inclusion of a particular place. In fact, even the definition of what a place *is* can radically differ from one gazetteer to the next. This is especially true for the specialist gazetteers that we are dealing with in the humanities. The goal of our "gazetteer interconnection" activities in Pelagios 3 is to bridge some of these gaps, and to create a framework for interlinking our member gazetteers into a coherent whole.

Naturally, one can (and will) never find *the one* generic data model that fits the needs of everyone, and to which every gazetteer should adhere from now on. Apart from practical issues of implementation and migration, such a model would inevitably end up being either hugely complex (because it would need to subsume all the complexities and subtleties of each gazetteer known at the time of design); or it would be overly simplistic (because it would force all gazetteers into a rigid, trimmed-down schema, sacrificing the richness and specialization of the original custom models). For this reason, Pelagios 3 is not aiming to create a common data model in the first place. Instead, our approach to linking gazetteers is based – just like our approach to linking documents – on the strategy of *connectivity through common references.* The principal design ingredients are as follows:

- **URIs.** First and foremost we require every participating gazetteer to provide URIs that identify its primary data model entities. In the context of Pelagios, we refer to these primary entities commonly as “place records”, although it is important to acknowledge the fact that the abstractions and definitions which underlie them may differ significantly across the different gazetteers (e.g. in one gazetteer place records may be based on a more geo-location-centric perception, in another they may be based on administrative units).
- **Links.** The purpose of the place records is to form anchor points for cross-linking between different gazetteer datasets (no more, no less) in order to indicate similarity. For the links, we use the semantics of *skos:closeMatch*,[9] which is defined as a relation *“[...] used to link two concepts that are sufficiently similar that they can be used interchangeably in some information retrieval applications”*.
- **Shared backbone.** While it is perfectly valid practice for every gazetteer to include links to any other gazetteer, a more pragmatic and immediate approach is to choose a single “reference gazetteer” (or a small number of them) to which every specialist gazetteer should strive to link. Suitable reference gazetteers are those that provide global coverage, have open licensing, high adoption among the community, and are available as Linked Data. Issues such as the quality of name or geometry data are – although desirable – in fact less of a concern. The primary value of the reference gazetteer, in this case, is in its URIs and their utility as an overarching referencing system that can serve as a backbone to the whole network. GeoNames is one obvious candidate for such a reference gazetteer, and indeed already enjoys considerable community adoption. (Pleiades, for example, has already established mappings to it, as noted above.) Another highly promising candidate in our view

is Wikidata,[10] a collaborative free knowledge base maintained by the Wikimedia Foundation. While not strictly a gazetteer, it has been seeded with Wikipedia links and infobox content, and thus provides a similarly broad geographic coverage, as well as coordinate information and multi-language name labels gathered from different language versions of Wikipedia.

- **Minimal place metadata published as Linked Open Data.** Finally, we need a common syntax by which gazetteers can share place records and links online. To this end, we are in the process of drafting a recommendation for a possible lightweight RDF-based serialization format. The goals of this format – in addition to enabling simple harvesting of the interconnection links – is to expose enough additional place metadata that external applications (such as our own Pelagios API) can build their own search indexes. In turn, this will allow the implementation of independent cross-gazetteer search without the need to resort to *federated search* (i.e. where each search query must be distributed to all member gazetteers' own search endpoints, and where the response is then re-assembled based on the individual query results). Most importantly, the format will include conventions for the expression of name and geometry data. But in addition, it will also allow incorporating temporal characteristics and provenance. A minimal example for how a place record is represented in our draft format (without time or provenance) is shown in Figure 8.5. Full details of the format under development are being maintained at the Pelagios website. [11]

```
@prefix dcterms: <http://purl.org/dc/terms/> .
@prefix geo: <http://www.w3.org/2003/01/geo/wgs84_pos#> .
@prefix pelagios: <http://pelagios.github.io/vocab/terms#> .
@prefix pleiades: <http://pleiades.stoa.org/places/vocab#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix skos: <http://www.w3.org/2004/02/skos/core#> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .

# We need to type the resource as a "place record"
# Gazetteers are free to state additional types
<http://www.my-gazetteer.org/places/000001> a pelagios:PlaceRecord ;

 # A dcterms:title is mandatory – this should NOT be considered a "primary name",
 # but simply an identifier for the place, usually corresponding to a primary
 # key in the original gazetteer
 dcterms:title "Athens" ;

 # Optional metadata (all strings may generally provide an optional language tag)
 dcterms:description "A major Greek city-state"@en ;

 # Cross-gazetteer links
 skos:closeMatch <http://www.wikidata.org/wiki/Q1524> ;
 skos:closeMatch <http://pleiades.stoa.org/places/579885> ;

 # List of names
 pleiades:hasName [ rdfs:label "Athens"@en ] ;
 pleiades:hasName [ rdfs:label "Αθήνα"@gr ] ;

 # Geometry
 pleiades:hasLocation [
   geo:lat "37.97945"^^xsd:double ; geo:long "23.71622"^^xsd:double
 ] ;
 .
```

Figure 8.5 Minimal example for a single place record expressed in Pelagios' proposed draft RDF "gazetteer interconnection format" (Turtle notation)

Our gazetteer interconnection format, along with an implementation and cross-gazetteer search facilities, will – out of necessity – be one of the central outcomes of Pelagios 3. It should be stressed that we do not claim that this format is either exhaustive or sufficient as an

interchange medium for gazetteers in general. But, once again, we consider it the most pragmatic solution to address the challenges specific to Pelagios, and believe that it may be a further helpful stepping stone towards the goal of establishing a global network of historical gazetteers.

---

[1] http://pelagios.org/api

[2] http://pleiades.stoa.org/

[3] http://pelagios.github.io/pelagios-heatmap

[4] http://www.pastplace.org/

[5] http://www.fas.harvard.edu/~chgis/

[6] Pleiades Plus. An experimental machine alignment between Pleiades place resources and content in the GeoNames gazetteer, by Leif Isaksen and Ryan Baumann http://atlantides.org/downloads/pleiades/plus/README.txt See also GeoNames: http://www.GeoNames.org/

[7] An introduction to Recogito is found on the Pelagios project blog at http://pelagios-project.blogspot.co.at/2014/01/from-bordeaux-to-jerusalem-and-back.html and http://pelagios-project.blogspot.co.at/2014/01/theres-pliny-of-room-at-bottom-1.html. The software itself is available as Open Source and can be obtained from http://github.com/pelagios/recogito

[8] http://nlp.stanford.edu/software/corenlp.shtml

[9] http://www.w3.org/2004/02/skos/

[10] https://www.wikidata.org

[11] Pelagios Gazetteer Interconnection Format (draft). http://github.com/pelagios/pelagios-cookbook/wiki/Pelagios-Gazetteer-Interconnection-Format