Linked Data Annotation Without the Pointy Brackets:

Introducing Recogito 2

Abstract: Recogito 2 is an open source annotation tool currently under development by Pelagios, an international initiative aimed at facilitating better linkages between online resources documenting the past. With Recogito 2, we aim to provide an environment for efficient semantic annotation—i.e. the task of enriching content with references to controlled vocabularies—in order to facilitate links between online data. At the same time, we address a perceived gap in the performance of existing tools, by emphasizing the development of mechanisms for manual intervention and editorial control that support the curation of quality data. While Recogito 2 provides an online workspace for general-purpose document annotation, it is particularly well-suited for geo-annotation, i.e. annotating documents with references to gazetteers, and supports the annotation of both texts and images (e.g. digitized maps). Already available for testing at http://recogito.pelagios.org, its formal release to the public is scheduled for December 2016.

1. Introduction

Annotation as a fundamental scholarly practice common across disciplines is well recognized (Unsworth 2000). The idea of adding notes or marginalia to documents goes back at least as far as the medieval manuscript, but it is in a digital context that annotation is emerging as a key means of facilitating research, by enabling scholars to organize, share and exchange knowledge, while working collaboratively

in the analysis and interpretation of source material (Barker and Terras 2016). This additional information can take various forms. Annotations provide enriched context by supplementing the document with information about provenance, composition and authorship in ways that better reflect a user's setting (Frisse 1987), or that can be exploited to improve search and retrieval in digital collections, in particular for lay users unfamiliar with domain-specific terminology (Hunter et al. 2008); they make transparent the structure of a document (e.g. the section demarcations of a text, such as book, chapter, paragraph, etc.), which can aid in its identification and analysis; or they may supply further detail about certain aspects of the content of the document that might be of assistance in its interpretation and understanding (Haslhofer et al. 2009). One such aspect, for example, are the places referred to in a document.

Annotation of place names or other kinds of geographic entities (such as peoples, regions or natural features) can be an important first step in the analysis of many different kinds of historical documents, particularly travelogues, historiographical accounts and maps. It also plays a critical role in the *Linked Open Data* (Bizer et al. 2009)approach being developed by Pelagios,[1] an international initiative aimed at facilitating better linkage between online resources documenting the past, by using the places to which these documents refer. The ability to annotate the content of Web documents, however, has tended to be restricted to researchers with technical expertise, and to tools that offer little to no opportunity for interoperability and data exchange, no matter whether we are talking about place or some other common entity. At the same time, the ever-increasing importance of the Web as a medium for the publication, curation and exchange of research data and scholarly results, along with the growing adoption of computational tools and methods in the humanities (Bodard and Romanello 2016), demands the development of platforms for digital annotation that any researcher can use. Recogito 2 is a response precisely to these new requirements. Based on an earlier prototype that had been focused on the idea of

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

annotating place, Recogito 2 is a platform for making annotation per se easy—Linked Data annotation without the pointy brackets.

This article is organized as follows: Section 2 introduces the Pelagios initiative—its goals, history and current activity—and discusses the role that Linked Data annotation plays in its context. Section 3 surveys some related work in the field of Linked Data applications and semantic annotation tools. Section 4 describes Recogito 1, an earlier, prototypical version of our tool. It presents some of the results produced with Recogito 1, and charts how user participation inspired the development of Recogito 2, a fully revised version with a more comprehensive scope, aimed at a more general audience. Section 5 provides a guided tour of the functionality and features of Recogito 2 implemented at the time of writing, and the developments scheduled on the roadmap. Section 6 discusses internal architecture, interfaces, and, in particular, how different thesauri and authority lists—gazetteers in particular—can be integrated. Section 7 concludes by laying out our longer-term vision for turning Recogito into an extensible platform that can be adapted to operate in institutional environments, with a customized feature set meeting the demands of different use cases and users.

2. The Pelagios Project

Pelagios is a community-driven initiative that facilitates better linkages between online resources documenting the past, based on the places that they refer to. Since 2011, Pelagios has been developing practices, methods and tools for interlinking data as diverse as text corpora, image collections, inscription records, or archaeological and numismatic databases. By addressing the problems of discovery and reuse, Pelagios aims to help digital humanists in making their data more discoverable, and to empower real-world users—scholars as well as the general public—to find information about particular ancient places and visualize it in meaningful ways.

There are two cornerstones to connectivity in the Pelagios model. The first is the use of unique stable references in the form of Uniform Resource Identifiers (URIs). Since Pelagios links documents via the places that they refer to, in our case these URIs are supplied by shared online *gazetteers*—authoritative directories of places that assign each place a unique identifier, as well as provide a host of related information such as names, coordinates, place types, periodization, etc. Pelagios advocates the idea that whenever you refer to a place in your data, you should do so using a gazetteer URI. *How* the place relates to that data may vary, and will generally depend on the type of data. For example, the place could be the find spot of a coin or an item in an archaeological database; it could be mentioned in a piece of literature or a research article; it could be attested to in a digitized old map; or it could be the location of a historic site depicted on a photograph. By expressing the places through the use of shared gazetteer URIs, otherwise isolated datasets become implicitly joined up to an interconnected graph, with the gazetteers as their central backbone (Isaksen et al. 2014, Simon et al. 2016). As a result, it then becomes possible to ask questions like: *"what are all the items related to these places?"; "which places are most commonly referred to in this collection?"; "which documents are primarily about places in this region?"*; or to discover similarities or contextual relations between documents, based on their place statistics or spatial patterns within them. Pelagios is open to any type of content, as long as it is available on the Web and itself "linkable to" through a URI.

The relation between the two URIs (that of the online content, and that of the gazetteer record) is established—and this is the second fundamental basis for Pelagios connectivity—through an *annotation.* On the one hand, annotation works as a suitable conceptual metaphor, since it carries the connotation that the association being made ought not to be considered certain fact; rather somebody (a human editor or an automated geo-parsing script) is making a *claim* that there is some relation between content x and place y.

It is thus an assertion of an interpretation. On the other hand, there exists a suitable technical mechanism for publishing annotations online as Linked Open Data (LOD): the W3C Web Annotation Data Model.[2] One core advantage of this model is that annotations can be published separately from the dataset they are annotating, as "standoff markup" (Thompson and McKelvie, 1997). We refer to this approach as *connectivity through common references*—as opposed to connectivity through a common schema—because it doesn't mandate a specific model for the data themselves, or otherwise put any constraints on how the data are being represented.

In its starting phase, 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 existed already, and was widely acknowledged among the scholarly community: the Pleiades Gazetteer of the Ancient World.[3] Pleiades provides URIs, names, and geographic data for more than 35,000 places in the Greco-Roman world, and was thus exactly the kind of shared referencing system for making annotations that would enable *connectivity through common references*. Pelagios has since expanded its scope significantly into periods and regions outside the realm of Pleiades, to encompass the early geographic documents of the pre-modern era, including early Christian, Islamic and Chinese traditions. To this end, Pelagios has been working with the respective gazetteer communities and initiated the development of LOD-based mechanisms that make it possible for different gazetteers (each serving a particular community) to create connections between each other. This in turn enables researchers to move more or less seamlessly between data from divergent traditions (Simon et al. 2016).[4] Key partners who have since made data from their gazetteers available for interlinking include: the *Digital Atlas of the*

[2] https://www.w3.org/TR/annotation-model/
[3] http://pleiades.stoa.org
[4] https://github.com/pelagios/pelagios-cookbook/wiki/Pelagios-Gazetteer-Interconnection-Format

Roman Empire (DARE),[5] the global historical gazetteer *PastPlace*,[6] the *China Historical GIS*,[7] the community-driven archaeological atlas *Vici.org*,[8] and the *Digital Index of North American Archaeology*.[9]

Pelagios has generated sustained and lively community interest that extends well beyond both its initial ancient world focus and its concern with place. Its distributed model of linking between independent datasets has been recognized (Mostern and Arksey, 2016) and adopted by similar LOD initiatives such as SNAP (Bodard et al. 2016) or PeriodO (Rabinowitz 2014), which are semantically annotating different reference types like people or time periods, respectively. Within this growing network of resources, the ability of LOD to promote the discovery of, and connections between, online documents of a highly varied nature has the potential to transform traditional scholarship (Elliot and Gillies 2009, Elliott et al. 2014, Bodard et al. 2016). By enabling new ways of analysis and "mutual contextualization"—the ability for Web resources to automatically draw on external content to enrich and help situate their own within an expanding ecosystem of independent online historical resources—it can have a broad and significant impact across disciplines as diverse as Archaeology, History, Classics, Cultural Studies, Mediaeval Studies, English, Modern Languages, Cartography and Geography.

3. Related Work

Tools for the annotation of online content take diverse shapes and forms. Social bookmarking tools like Delicious,[10] or social tagging features on content sharing sites like Flickr[11] are examples for basic annotation functionality that has entered the mainstream, as a means to add contextual information that

[5] http://dare.ht.lu.se/
[6] http://pastplace.org/
[7] http://www.fas.harvard.edu/~chgis/
[8] http://vici.org
[9] http://ux.opencontext.org/archaeology-site-data/
[10] http://del.icio.us/
[11] http://www.flickr.com/

aids re-use in a personal setting, or among a specific community of users; as well as a means to catalogue online materials according to personal preference and requirements. Among a scholarly audience, bibliography reference management software like JabRef,[12] or online services like Zotero[13] fulfil a similar purpose, in terms of providing the means to add metadata and organize materials. By and large, these tools are limited to annotating the item as a whole, rather than providing functionality to annotate *inside* the actual (text, image or media) content. Noteworthy exceptions that focus specifically on annotation of the content itself are Annotator[14], an open source library to add annotation functionality to any Web page; and Hypothesis[15], an online service and open source application[16] for Web annotation.

While the above tools have been providing easy-to-use interfaces for various forms of annotation, *semantic annotation*—and the integration of Linked Open Data specifically—has generally remained out of their scope. Oren et al. (2006) distinguish three types of annotations: *informal*, *formal* and *ontological*. Informal annotations are those that do not use a formal language, whereas formal annotations differ in that they use formally defined terms. (The tools listed above all fall into either of those two categories.) Ontological annotations, finally, are formal annotations where terminology has a commonly understood meaning according to a shared conceptualization. Oren et al. rightfully point out that whether a term is ontological or not is a purely social matter, not a technical, nor formal one. The benefit of the ontological—or semantic—annotation, and in fact of Linked Data as a whole, is therefore not so much the machine-readability of the data as such, but rather that it is represents a shared social understanding, expressed with a shared vocabulary. Andrews et al. (2012) provide an overview of different annotation systems, and how they pay attention to semantic annotation. A further noteworthy example is *Pundit*

[12] http://www.jabref.org/
[13] http://www.zotero.org
[14] http://annotatorjs.org/
[15] http://hypothes.is/
[16] http://github.com/hypothesis/h

(Grassi et al. 2013), a suite of online applications to support semantic annotation of arbitrary Web resources. Using graphical user interface components, users can select text or image elements on a Web page, and associate them with a mix of free commentary and references to Linked Data resources. However, Pundit confronts the user with the full breadth of technical terminology. At least a moderate understanding of the key concepts behind the Semantic Web, the RDF triple, made up of subject, predicate and object, and a familiarity with prominent Linked Data sources such as Freebase and DBpedia are required in order to make proper use of Pundit.

We are concerned that a tool where features and user interface metaphors are strongly guided by the underlying technology represents a severe threshold for adoption. Indeed, Grassi et al. (2012) state the main idea behind Pundit as being one of enriching the Linked Data Web. We agree entirely that Linked Data is absolutely essential in enabling and connecting digital scholarship. But we should exercise care that the mere production of new RDF triples on, and between, the established Linked Data hubs is not practiced as an end in itself. Instead, we argue, tools must first and foremost provide scholars with efficient workspaces in which they can engage with their materials. The role of Linked Data should be to support scholars in their aims (e.g. to help generate a map from a text with little effort); while new Linked Data should follow from scholarly results, almost as an accidental byproduct.

4. A Brief History of Recogito

Initial development of our annotation platform Recogito began in November 2013, as part of a research project funded by the Andrew W. Mellon Foundation named "Pelagios 3: Early Geospatial Documents". The aim of the project was to establish a comprehensive index of places referred to in *Early Geospatial Documents*—documents that use written or visual representation to describe geographic space prior to the year 1492, and make it accessible as Linked Open Data.

From the outset of the project, it was clear that in order to cover a reasonable breadth of material, we would need to work with the community and identify existing datasets on the one hand, but also do significant amounts of annotation in-house ourselves. (This marked a departure from the two earliest phases of Pelagios, where we had worked with partners and their pre-existing datasets to develop a standard way of referring to the place names in their documents—out of which Pelagios's "connectivity through common references" was born.) Part of this work entailed using and adapting automated annotation approaches, such as Named Entity Recognition (NER), and evaluating to what extent they improved productivity. What soon emerged, however, was the critical importance of maintaining the quality of data, leading to the conclusion that quality needed to be ensured through proper editorial control. While there exist many tools for *automatic* semantic annotation (i.e. the task of associating document fragments to terms in controlled vocabularies like gazetteers), none were felt sufficient in order to attain the tractability required for ensuring data quality. In short our research demonstrated that there was a significant gap in resources currently available vis-à-vis: (i) manual intervention (that is to say, human verification and correction); and (ii) simplicity and efficiency of use, especially for non-technical users. On this basis we decided to undertake the development of a tailor-made environment that would support our specific goals of annotating place names in historical texts and maps.

This initial prototype ("Recogito 1") was a Web-based tool featuring several work areas dedicated to different stages of the geo-annotation workflow (Simon et al. 2015): a text annotation area to demarcate place names in digital texts; an image annotation area to mark up and transcribe place names on map and manuscript scans; and a geo-resolution area, where place names identified (and transcribed) in the initial phase could then be mapped to gazetteer URIs. (It is by means of this second step that the documents being annotated would then be incorporated within Pelagios.) Recogito 1 also provided some basic means

for managing documents, and for recording and visualizing user activity, annotation progress, and document statistics.

By the end of the Pelagios 3 project in August 2015, Recogito helped us make significant progress. We tagged more than 120,000 references to places in about 200 documents from the Latin, Greek, European Medieval, Maritime, and Early Islamic and Chinese traditions; and aligned about half of them to historical gazetteers like Pleiades or PastPlace. For the duration of the project, Recogito had been used primarily by members of the core project team. However, a number of people outside the team had also expressed interest in using it, either to contribute to materials from the project, or to work on their own materials. By the end of Pelagios 3, the number of registered editors had grown to about 90.

This growing interest encouraged us to think about ways of opening Recogito to a wider audience. Supported by a grant from the Open Humanities Awards 2014,[17] we organized two geo-annotation workshops with students and academics from various disciplines (geography, history, engineering, and archaeology) in group sizes of 27 and 22, respectively (Simon et al. 2015). For each workshop, we presented participants with different geographic traditions on which to work, along with accompanying materials (Classical Latin texts and Medieval maps; Medieval travel writing, pilgrimage itineraries and medieval nautical charts). Beyond that, however, workshop participants were free to choose whichever document(s) and task(s) that they wanted (i.e. identifying place names in texts or maps, transcribing, mapping to gazetteers). The quantity of contributions made by our participants greatly exceeded our expectations: after annotation sessions of approx 2½ hours, a total of 6,620 contributions were recorded in the first workshop, and 7,511 contributions in the second. We also received highly positive feedback from

[17] http://dm2e.eu/open-humanities-awards-round-2-winners-announced/

participants on their overall experience of the tool, and a significant number of contributions were made even after the workshops had ended.

The results achieved during the Pelagios 3 project and the annotation workshops were also valuable in revealing some of Recogito's shortcomings. With regard to usability, for example, our own experience was that the time required for the task of geo-resolution significantly outweighed all other tasks (such as identifying place names in texts or transcribing from maps). This was confirmed by the contribution statistics from the workshops: in the first workshop, geo-resolution actions accounted only for 2% of the total number of user actions. A redesign of the user interface between the first and the second workshop yielded noticeable improvement, raising the percentage of geo-resolutions relative to other tasks to about 7%. But it is obvious that geo-resolution remained the productivity bottleneck in our workflow.

There were other limitations that arose from Recogito's original design goal, which was to meet specifically the aims of Pelagios 3, and the needs of the core project team members. For example, while Recogito had always been Web-based (and thus usable remotely by different people at the same time) it was never really collaborative. Working with it required assistance by a tool administrator—to set up user accounts, upload the documents and their metadata, and assign them to collections, etc. Provenance tracking (i.e. who contributed what to an annotation) was very basic. Forming teams and managing access to specific documents for specific users was not possible. Restoring annotated documents to a previous state in time (e.g. to quickly revert all additions made during a "demo session") involved additional maintenance work on the application database. Likewise, features or data model aspects that were not originally required to meet the goals of Pelagios 3 could not be implemented (even if they might have been frequently requested by peers outside our project team). Such unimplemented features included: the ability to create general commentary annotations; the creation of text annotations that overlap; a simple

point selection tool for maps (as opposed to Recogito's place name-specific box selection tool); the possibility to attach multiple alternative readings for a transcription; or functionality to make the annotated content visible to the public, so that it could be viewed without requiring a Recogito user account.

Nevertheless, the overwhelming feedback that we received suggested that Recogito 1 succeeded in addressing a range of unmet needs in the community more widely. Crucial to this approval were the fundamental design choices that we had identified initially, namely that: (i) every automated step—Named Entity Recognition, and automated matching of place names to gazetteer records—must always require human verification, and, whenever such verification was missing, this would be prominently displayed visually; and (ii) that users felt comfortable performing semantic tagging in an interface that put the emphasis on manual control, while offering support through automated suggestions. Above all, it was evident that not having to deal directly with the intricacies of either URIs or Linked Data more generally was perceived as a significant benefit. Indeed, the quantity of contributions that both the project team as well as our workshop participants were able to make—both without significant prior training—seemed to support our overall positive impression. Beginning in February 2016, and supported by renewed funding from the Andrew W. Mellon Foundation, we therefore started to develop a new version of Recogito, redesigned from the bottom up, with the aim of establishing the technical infrastructure for an open, collaborative, generally usable, and useful, work environment.

5. A Guided Tour of Recogito 2

Arguably, the main conceptual difference between Recogito 2 and its predecessor is the transformation from a "global project repository", which hosts all documents in a single space, to a personal working environment. After registering an account, Recogito 2 provides a "user space", which acts as the user's

personal start page. This page is the place where documents are managed, and new documents are uploaded. The page is also visible to the public on the Web, along the lines of a profile page in a social network, under a personal URL in the form *http://recogito.pelagios.org/{username}*. When opening a document from the user space, Recogito 2 offers different "views"—work areas dedicated to different aspects of the annotation workflow.

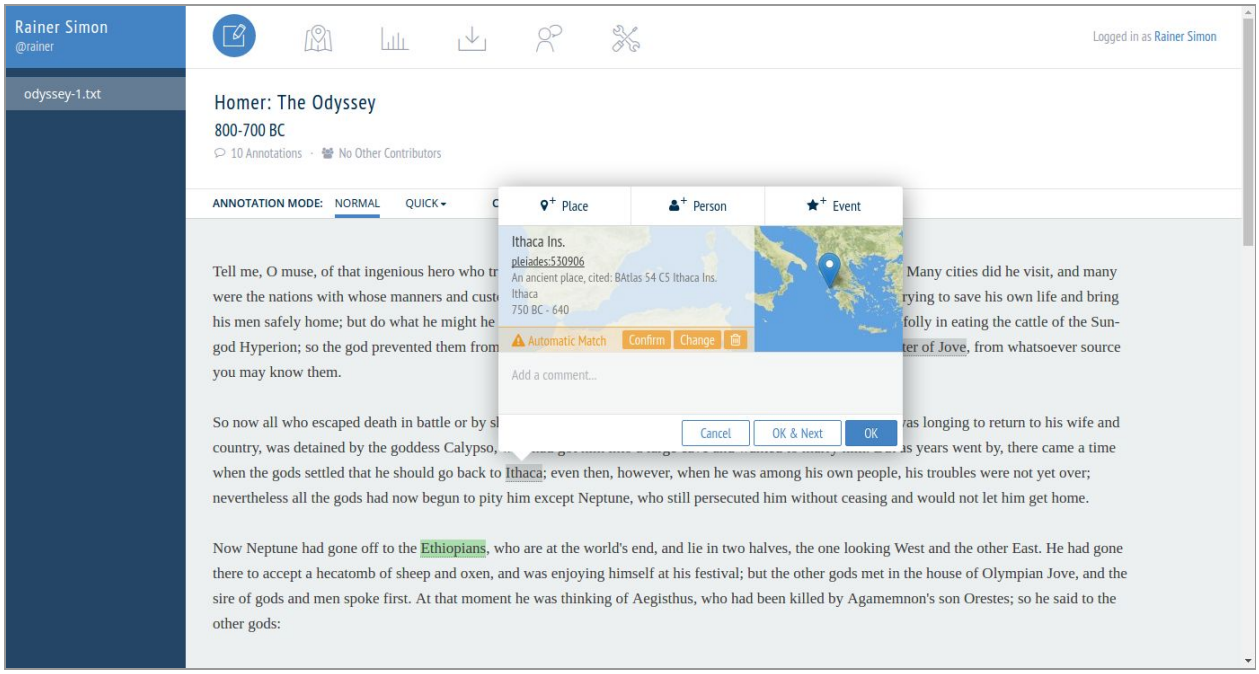

Fig. 1. Recogito 2 text annotation view.

5.1 Annotation

The text annotation view is a reading view which provides tools to select segments in the text and add annotations. Unlike Recogito 1, Recogito 2 now supports a combination of free-text commentary and semantic annotations (URI-based references to terms in controlled vocabularies). At the time of writing, gazetteers are the only type of controlled vocabulary available for semantic annotation. But support for person authority lists is scheduled. In addition, tagging with free keywords is also being implemented.

While Recogito 2 aims to provide as much automation as possible as an aid to annotation, human intervention remains primary. For example, when categorizing a selected phrase as a "Place", Recogito 2 will automatically perform a lookup across the gazetteers in the system, and provide a first match (Fig. 1). As a general policy, every automatic match remains marked as "unverified"—indicated using the colour grey—until a user either explicitly confirms its correctness, or manually changes it using the integrated gazetteer search (Fig.2)—in both instances turning it green. It is also possible to explicitly flag a place as "not identifiable" (yellow), when no suitable gazetteer match could be found; or to add multiple gazetteer matches, when, for instance, the reference is unclear, or when the annotated phrase does indeed refer to multiple places at once.

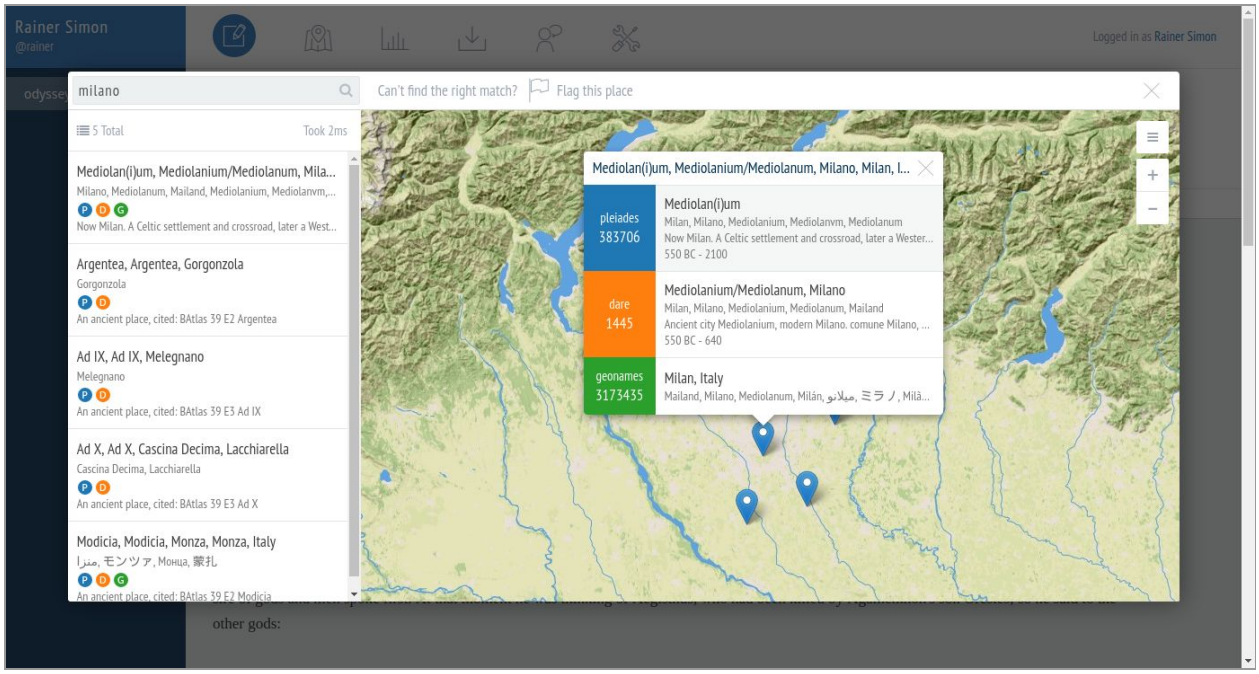

Fig. 2. Recogito 2 gazetteer search.

A fundamental departure for the new version of Recogito is its aim to support annotation as a collaborative process. Each contribution to an annotation is associated with the user who contributed it,

and the time it was made. Since annotations generally consist of a sequence of different contributions and interactions (comments, replies, transcriptions, gazetteer matches, etc.), they can essentially function as "micro discussion threads" with multiple participants. Also, because Recogito keeps an audit trail of additions and changes, it is possible to inspect the version history of the annotations, and revert the document to a previous state in time.

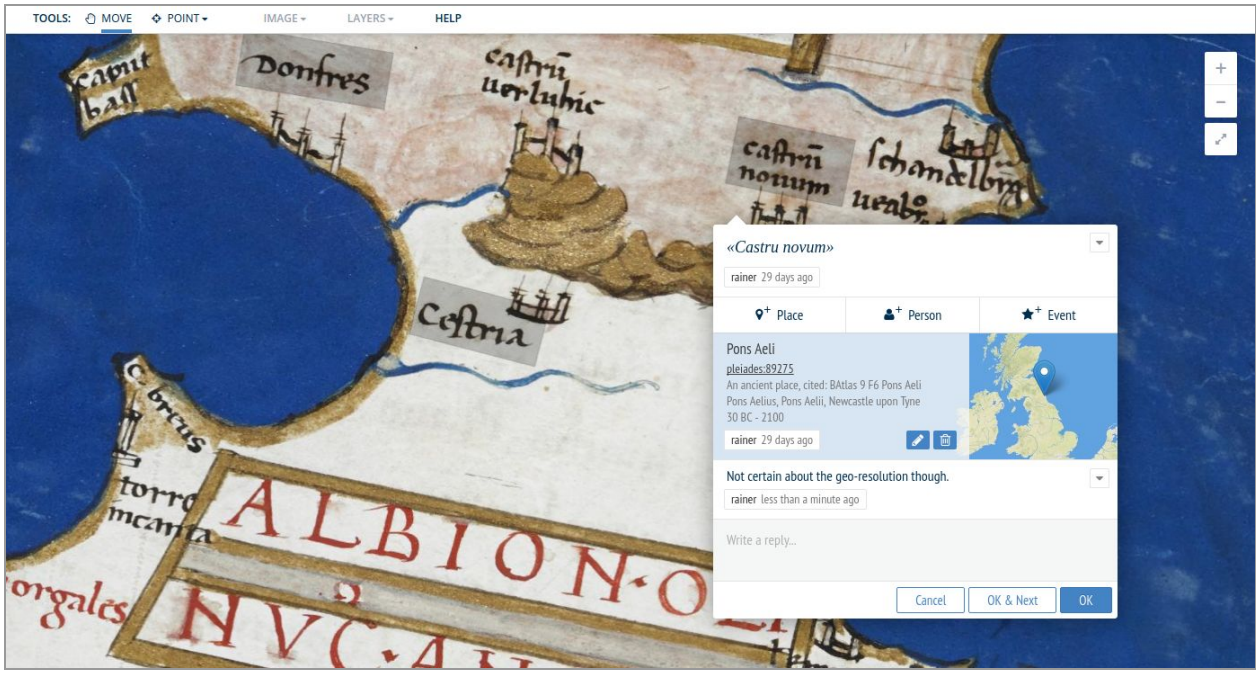

Fig. 3. Recogito 2 image annotation view.

The image annotation view serves the same purpose for images, as the text annotation view does for texts. It provides a zoom- and pan-able view for navigating high resolution images, along with drawing tools for marking points and regions, and attaching annotations. Since the interface has been designed specifically with digitized maps in mind, the view allows the user to rotate the image freely. There is also a unique drawing tool for selecting a tilted box (which retains orientation information, i.e. which side is "up" and which "down"), specifically for the purpose of annotating and transcribing place names (Fig. 3).

5.2 Using Named Entity Recognition and Automatic Geo-Resolution

When uploading a new text document to the personal space, users can choose to perform Named Entity Recognition, i.e. attempt to pre-annotate references to places and persons in the document automatically. At the time of writing, Recogito implements this functionality by integrating the Stanford NLP toolkit (Manning et al 2014), and provides support for English-language texts only. After NER, identified places are resolved against Recogito's gazetteer index to provide a first match. Following the general principle of mandatory human verification, automatic matches are again categorized as "unverified" (and identified accordingly in the user interface) until a user has confirmed or corrected the match. As part of our near-term roadmap, we are currently defining an extension mechanism that allows "plugging in" of additional or alternative languages and NER engines (such as the Classical Language Toolkit).[18] This way we hope to provide a flexible platform that can be customized to meet demands of diverse communities.

5.3 Map

The map view provides an overview of all places that were identified in the document. Marker size indicates the relative frequency of mentions of the place in the document. Clicking a marker provides additional information about the place, the gazetteer record(s) it was mapped to, and how it appears in the context of the document—that is to say, not only the number of times any given place is referred to in the document but also where those references occur—as well as a direct link back to the annotation(s) in the text or image view (Fig. 4). Different colour-coding and symbolization options are planned for the future, so as to visualize the distribution of places in different parts of the document; or how different places

[18] http://cltk.org/

were, for example, associated with different keywords, using the tagging functionality in the annotation view.

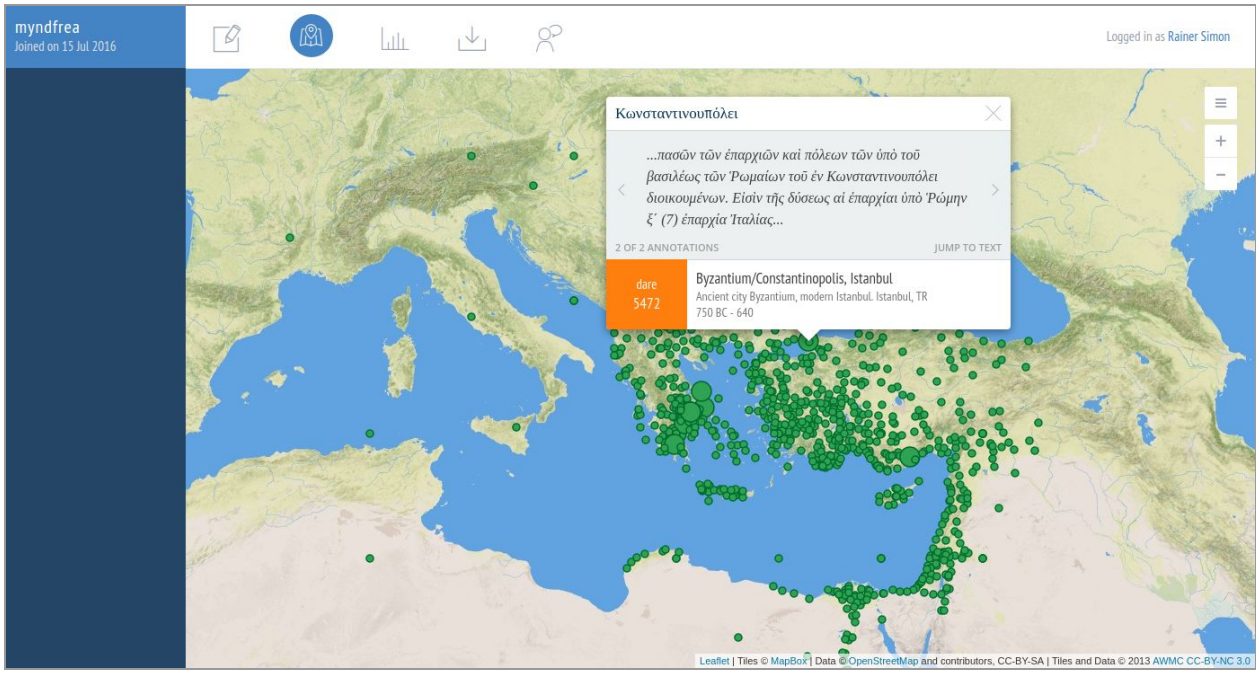

Fig. 4. Recogito 2 map view.

5.4 Annotation Statistics

Recogito 1 included a basic "statistics dashboard" displaying different metrics and aggregated information derived from the annotations in a document. An equivalent feature will be implemented in Recogito 2. It will provide an overview of document properties such as: the entity verification rate, i.e. how many entity matches are in an "unconfirmed" state vs. how many have been confirmed (or manually flagged as "not identifiable") by users; the contribution history, broken down into the total number of contributions by users, the type of contribution, and the contributions over time; the relative distribution of annotation types, i.e. the amount of commentary versus entity annotations (place, person); lists of tag

usage, of places and persons referenced in the document, and place and person names flagged as "not identifiable", ordered by frequency; outliers that may indicate potential issues, for example if the same place name has been assigned to different gazetteer records within the document. In the longer term, we also hope to provide visitor statistics, and functionality to list "similar" documents, which resemble the current one in terms of annotation patterns. This, we hope, will facilitate the discovery of other users' work that may relate to one's own, and encourage new collaborations among users across the Recogito network.

5.5 Exporting Data

Recogito 2 offers a range of options for exporting data to different formats. At the time of writing, it is possible to export annotations to comma-separated values (CSV), a tabular data format for use in, e.g., spreadsheets. Places can be exported to GeoJSON, a map-centric format compatible with Web mapping toolkits or GIS systems. In the case of text documents, there is the possibiltiy to export content and annotations as TEI,[19] an XML document encoding scheme widely used in the Digital Humanities community. Additional options currently under development include: KML, another map-centric format to export places, for use with virtual globes such as Google Earth; and, last but not least, document metadata and annotations in RDF (in XML and Turtle serialization), using a combination of Dublin Core properties and the W3C Web Annotation data model, in accordance with the general Pelagios conventions.

5.6 Collaboration, Sharing and Discussion

Enabling collaborative annotation and fostering open discussion around documents has been an important design goal for Recogito 2. Users are able to share documents that reside within their own personal space

[19] http://www.tei-c.org/index.xml

with others, using different levels of access permissions: granting other users read-access allows them to view the document and the annotations, but not make any contributions themselves; with write-access, others can create annotations, and make additions to existing ones (such as replying to a comment, editing or changing an entity association, etc.); admin-level access, in addition, grants permissions similar to those of the document owner, i.e. to edit the document's metadata, to invite other collaborators, to perform document backup and restore, or to revert the editing history to a previous state.

It is also possible to grant general read access to a document. This way, the document will be visible to the public on the user's profile page; the link to the document can be shared; and anyone on the Web can view the document and the annotations without the need to register an account. Once general access is enabled, it is also possible for the public to view map and annotation statistics pages, and access all the data download options.

To facilitate discussion about the document as a whole, and to provide a space for more general interaction between collaborators, we further plan to provide a dedicated "discussion board" page for each document. We envision this discussion board to function along the same lines as the comment thread at the bottom of a blog post. Document owners will be able to choose whether to restrict commenting to collaborators only, or open it out to the public as a whole.

5.7 TEI, IIIF and Tabular Data

At the time of writing, plaintext files are the only type of text content supported for import to Recogito 2. However, additional text formats are scheduled on the roadmap. Our primary focus in this regard is to support TEI (which, at the time of writing, is supported as a download format, but not for upload). The official release for this feature is scheduled for fall 2017. However, a first proof of concept has been

developed already, based on CETEIcean,[20] an open source library for displaying TEI documents natively in the browser.

With regard to images, Recogito supports upload of the most widely used file formats such as JPEG, TIFF or PNG. During upload, images are automatically converted to a format, that later allows them to be displayed as zoomable images in the annotation view. There have been recent there have been efforts to standardize the access to such zoomable images over the Web. The *International Image Interoperability Forum* (IIIF),[21] a growing community of research libraries, image repositories, and cultural heritage institutions, has issued specifications for interoperable delivery of images and their metadata. Support for consuming images via the IIIF standard is also scheduled on the Recogito roadmap; and first integration trials have been successfully implemented at the time of writing.

Another type of content we are planning to support is tabular data. A request that we frequently received over the course of the Pelagios 3 project was for an easy way to enrich a spreadsheet of place names with gazetteer URIs; typically this was needed by researchers as a preparatory step towards building their own interlinked gazetteer. Since this was such a common request, we already implemented a rudimentary kind of support during the development of Recogito 1. Recogito 2 will expand on this feature, by offering a dedicated "table view" with the standard manual annotation and verification features, along with a mechanism to pre-annotate table rows automatically during upload, based on the contents of selected table columns.

6. Architecture and Interfaces

[20] https://github.com/TEIC/CETEIcean
[21] http://iiif.io/

In terms of the technical architecture, Recogito 2 is a Web application based on a standard 3-tier application model. It is implemented on a JVM (Java Virtual Machine) technology stack, using the open source *Play* Web framework, and the Scala programming language for the "middle tier" (application server components).

The data tier is implemented with a combination of a PostgreSQL relational database and an ElasticSearch document store (both open source technologies as well). issue in a more transparent way.) There are several reason for this division into two storage technologies. First and foremost, it is a matter of minimizing the amount of data transformation—and thus lines of code—needed in the application. Data that is mostly tabular in nature, such as document metadata records or user account data, is relatively effortless to handle in a relational database (like PostgreSQL), whereas more flexibly nested data structures—such as annotations, which consist of a hierarchy of different elements—are naturally represented in so-called document-oriented databases (like ElasticSearch). Second, the types of data Recogito stores in ElasticSearch (annotations and their versions, gazetteer records, contribution events) are by far more numerous than document metadata and user records. Furthermore, they are frequently retrieved via fulltext searches—two more aspects that favour the use of ElasticSearch, which has been designed originally as a search index, and is easily scalable across multiple servers to support large volumes of data. (In fact, the need for fulltext search is a reason why the combination of relational database and search index is a frequently encountered Web application architecture pattern.) Third, Recogito requires rich analytics functionality in order to drive document and user statistics visualizations, in particular for annotations and contribution event records. This is another requirement that led us to decide in favour of ElasticSearch, which has a particularly strong focus on data analytics.

The presentation tier is implemented in JavaScript, making use of a range of general-purpose open source utility libraries (such as *RequireJS*[22] and *jQuery*[23]) and JavaScript user interface component frameworks for specific purposes, including the *Leaflet*[24] Web mapping library, or *OpenLayers*[25] for display of high-resolution zoomable imagery.

Our official installation of Recogito 2—while still technically in an alpha stage—is already available for testing at http://recogito.pelagios.org, and scheduled for formal release to the public in December 2016. The source code is available as open source software, under the terms of the Apache 2 license.[26] This means that everyone is free to set up their own installation, on their own server, e.g. for personal use, or within a research team or institution. Code and accompanying setup information can be found at the Pelagios GitHub site at http://github.com/pelagios/recogito2.

After setting up an installation of Recogito, an additional step that is required is to import vocabularies or thesauri—sources of URIs that can be used for semantic tagging. As mentioned above, gazetteers (which one uses for the semantic tagging of places) are presently the only type of tagging vocabulary implemented; directories of persons are already scheduled in for a later date. A number of (Creative Commons licensed) gazetteers in a form compatible with Recogito are now available. A "starting package" consisting of Pleiades, the Digital Atlas of the Roman Empire, and a subset of GeoNames,[27] is available through the GitHub repository, though extending Recogito with alternative or additional gazetteers is possible and will in all probability be highly desirable, as interest in linking online historical resources beyond the ancient world (and thence beyond the scope of Pleiades or DARE) continues to

[22] http://requirejs.org/
[23] http://jquery.com/
[24] http://leafletjs.com/
[25] http://openlayers.org/
[26] https://www.apache.org/licenses/LICENSE-2.0
[27] http://download.geonames.org/export/dump/

grow. Prior to using a new gazetteer for tagging, a full data export must first be obtained and imported into Recogito. Here a challenge presents itself, since there is currently no standard data format for exchanging gazetteer data. In order to meet the objectives of the Pelagios 3 project (which aimed at linking historical resources from across different geographical traditions, not only ancient Greek and Roman sources), we initiated a part-solution that relied on the development of an RDF-based format.[28] Since Recogito implements this format, any gazetteer made available in a dump file that adheres to it will be importable. At the time of writing, however, an alternative approach is gaining traction in the community, based on GeoJSON.[29] Pleiades, for example, has recently adopted it by making its nightly dumps available in this format, and, since we endeavour to stay alert to the needs of our community and are keen to be as responsive as we can, we have enabled Recogito to start supporting GeoJSON as well. A possible next step for the Pelagios initiative is, as a collective, to begin to document common practices and articulate a common vocabulary for publishing GeoJSON gazetteer dumps, so that over time, more and more gazetteers will become available in a standard format.

7. Future and Outlook: Recogito as an Extensible Platform

In this article we have introduced Recogito 2, an open source tool for semantic annotation, currently under development by the Pelagios initiative. Having provided the background and context to its origins and need, we have set out the currently implemented feature set and technical architecture. At the time of writing, Recogito 2 is under active development, with a first release scheduled for December 2016, and a defined roadmap until the end of 2017. Throughout this remaining time, we intend to work closely with the community and respond to new ideas and requests that emerge.

[28] https://github.com/pelagios/pelagios-cookbook/wiki/Pelagios-Gazetteer-Interconnection-Format
[29] http://geojson.org/

Ultimately, however, we hope that Recogito will begin to stand on its own feet as an open source project. Indeed, our vision is to gradually evolve it into an extensible framework—that is, a platform that takes care of the mundane formalities of annotation, such as: storage, versioning, recording provenance and activity metrics; managing documents and access rights; handling data transformation, import and export. At the same time, we envisage that it will provide the necessary hooks to plug in new functionality as and when it is needed, so as to provide a tailor-made work environment for different domains and use cases. We are already making some small first steps towards this evolution, with the manufacture of a prototype for an alternative Named Entity Recognition engine "plugin", currently under development by another project. Further examples of domain-specific extensions we imagine for the future are: plugins that add additional fields to the annotation editor user interface component, or connectors that integrate Recogito directly with existing document repositories, rather than importing documents through upload. We are only just starting to identify what will make useful extension points, and how to best design them. The next steps, we anticipate, will be driven by the wider community, much more than through Pelagios itself. We believe that developments will show how a single tool with a specific purpose—that of semantic annotation—can play a beneficial role for, and make a better claim to contribute to, scholarship in the wider ecosystem of digital research.

Acknowledgements

The authors wish to thank the Andrew W. Mellon Foundation for funding this work.

References

Andrews, P., Zaihrayeu, I., Pane, J. (2012) A Classification of Semantic Annotation Systems. In *Semantic Web*, vol. 3, no. 3, pp. 223-248.

Barker, E. and Terras, M. (2016) Greek literature, the digital humanities, and shifting technologies of reading. *Oxford Handbooks Online*, Oxford. DOI: 10.1093/oxfordhb/9780199935390.013.45

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.

Bodard, G., Romanello M. (2016) *Digital Classics outside the Echo-Chamber: Teaching, Knowledge Exchange and Public Engagement*. London: Ubiquity Press, 2016. ISBN 9781909188617.

Bodard, G., Gheldof, T., Lawrence, F. K., Stoyanova, S., Tupman (2016) Networking Ancient Person-data: community building and user studies around the SNAP:DRGN project. In *Digital Humanities 2016 Conference Abstracts*, pp. 44-45.

Elliott, Tom and Sean Gillies (2009). "Digital Geography and Classics," In *Digital Humanities Quarterly 3(1)*. http://www.digitalhumanities.org/dhq/vol/3/1/000031/000031.html

Elliott, T., Heath, S. and Muccigrosso, J. (2014). Current Practice in Linked Open Data for the Ancient World. ISAW Papers 7. http://dlib.nyu.edu/awdl/isaw/isaw-papers/7/

Frisse, M. E. (1987) Searching for Information in a Hypertext Medical Handbook. In *Proceedings of the ACM Conference on Hypertext* (HYPERTEXT '87), pages 57–66, Chapel Hill, North Carolina, 1987. ACM.

Grassi, M., Morbidoni, C., Nucci, M., Fonda, Ledda, G. (2012) Pundit: Semantically Structured Annotations for Web Contents and Digital Libraries. In *Proceedings of the 2nd International Workshop on Semantic Digital Archives (SDA 2012).*

Grassi, M., Morbidoni, C., Nucci, M., Fonda, S., Piazza, F. (2013) Pundit: Augmenting Web Contents with Semantics. In *Literary & Linguisting Computing*, 2013.

Haslhofer, B., Jochum, W., King, R., Sadilek, C., Schellner, K. (2009) The LEMO Annotation Framework: Weaving Multimedia Annotations with the Web. In *International Journal on Digital Libraries*, 10(1): 15-32.

Haslhofer, B., Sanderson, R., Simon, R. and Van de Sompel, H. (2012) Open Annotations on Multimedia Web Resources. In *Multimedia Tools and Applications*.

Hunter, J., Khan, I., Gerber, A. (2008) HarvANA – Harvesting Community Tags to Enrich Collection Metadata. In *Proceedings of the 8th ACM/IEEE-CS Joint Conference on Digital Libraries* (JCDL 2008), Pittsburgh, Pennsylvania, United States, June 16–20, 2008: 147–156.

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.

Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S. J., McClosky, D (2014) The Stanford CoreNLP Natural Language Processing Toolkit. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pp. 55-60.

Mostern, R., Arksey, M. (2016) Don't Just Build It, They Probably Won't Come: Data Sharing and the Social Life of Data in the Historical Quantitative Social Sciences. In *International Journal of Humanities and Arts Computing*, vol. 10, issue 2, pp. 205-224, ISSN 1753-8548. Available at: http://www.euppublishing.com/doi/full/10.3366/ijhac.2016.0170

Oren, E., Möller, K., Scerri, S., Handschuh, S., Sintek, M. (2006) What are Semantic Annotations? In *Relatória técnico*, DERI Galway (2006).

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/

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.

Simon, R., Isaksen, L., Barker, E. and de Soto Cañamares , P. (2016) 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, 2016, ISBN: 978-0-253-02244-8.

Thompson, H.S., McKelvie, D. (1997) Hyperlink semantics for standoff markup of read-only documents. In *Proceedings of SGML Europe 97*. p. 227-229.

Unsworth, J. (2000) Scholarly Primitives: what methods do humanities researchers have in common, and how might our tools reflect this? In *Humanities Computing: formal methods, experimental practice*. King's College, London, May 2000. Available at: http://www3.isrl.illinois.edu/unsworth/Kings.5-00/primitives.html