Rainer Simon[*], Bernhard Haslhofer[**], Joachim Jung[*]

Annotations, Tags & Linked Data - Metadata Enrichment in Online Map Collections through Volunteer-Contributed Information

Keywords: Annotation, Linked Data, Crowdsourcing.

Summary: Cultural heritage institutions and private collections such as the Library of Congress or the David Rumsey Map Collection are increasingly providing free online access to high-resolution scans of old maps. With the YUMA Map Annotation Tool, we want to facilitate collaborative scholarly annotation for such online resources. A central feature of our tool is the integration of *semantic linking* into the annotation process: annotations are semi-automatically enriched with context information from sources on the *Linked Data* Web. We argue that this semantic contextualization is, on the one hand, relevant for scholarly collaboration. On the other hand, we believe that it can be exploited to improve search and retrieval in the collection by providing additional structured metadata and geo-referencing information. In this paper we present current work in which we aim to verify our assumptions based on maps provided by the Library of Congress, and annotations collected from volunteer users in an ongoing crowdsourcing experiment.

Introduction

Annotations are a fundamental scholarly practice common across disciplines (Unsworth 2000). They enable scholars to organize, share and exchange knowledge, and work collaboratively in the interpretation and analysis of source material. They provide additional explanations which may help others in the interpretation and understanding of the original document or item (Haslhofer et al. 2009) and provide context by supplementing the document or item with information that may better reflect a user's setting (Frisse 1987). Beyond that, annotations also act as a source of additional metadata, which can be exploited to improve search and retrieval in digital collections, in particular for non-expert users who may be unfamiliar with domain-specific terminology (Hunter et al. 2008).

Within the ongoing Europeana*Connect* project[1], we are concerned with the development of multimedia annotation components for Europeana[2], the European cultural heritage portal. Annotation of digitized old maps has been of particular interest to us from the early phases of the project (Simon et al. 2010). We are now in the process of making our annotation components independently available as open source software under the name *YUMA* (YUMA Universal Multimedia Annotator)[3]. For this reason we are focusing on more concrete use cases, application fields and target audiences, and are looking for feedback from interested parties in order to inform the future development of our toolset. The remainder of this paper is structured as follows: in the next section, we provide an overview of the YUMA Map Annotation Tool. We introduce the concept of *semantic linking* – the enrichment of annotations with links to relevant resources on the Web. We conclude

---

[*] AIT – Austrian Institute of Technology, [firstname.lastname@ait.ac.at]
[**] Cornell University, [bernhard.haslhofer@cornell.edu]

[1] http://europeanaconnect.eu
[2] http://europeana.eu
[3] An online showcase can be found at http://dme.ait.ac.at/annotation; source code is hosted at http://github.com/yuma-annotation

the paper with an overview of ongoing work in which we aim at testing our assumptions based on a map collection provided by the Library of Congress, and annotations collected from volunteer users in a crowdsourcing experiment.

The YUMA Map Annotation Tool

The YUMA Map Annotation Tool (see Figure 1) is a browser-based rich Web application displaying a full-screen Google Maps-like drag- and zoomable representation of the digitized map. A floating window shows a list of all existing annotations for the map. The window also contains GUI elements that allow users to create new annotations, edit or reply to existing ones, or delete their own annotations. To prevent abuse, a basic moderation feature allows users to report inappropriate annotations to the system administrator. When creating an annotation, users may either annotate the whole map, or they have the option to draw a polygon shape on the map to identify a specific area to which the annotation pertains. Using the tool's integrated geo-referencing functionality (Simon et al. 2010), shapes can be automatically translated to their (approximate) geographical coordinates, and overlaid on top of a present-day map, shown in a separate floating window. To support collaborative work, YUMA provides flexible RSS feed support: users can follow the discussions around a particular map, a particular annotation, or the public activity of particular users. For the future, we also plan to enable following annotation activity pertaining to a particular geographical area, or historical period.
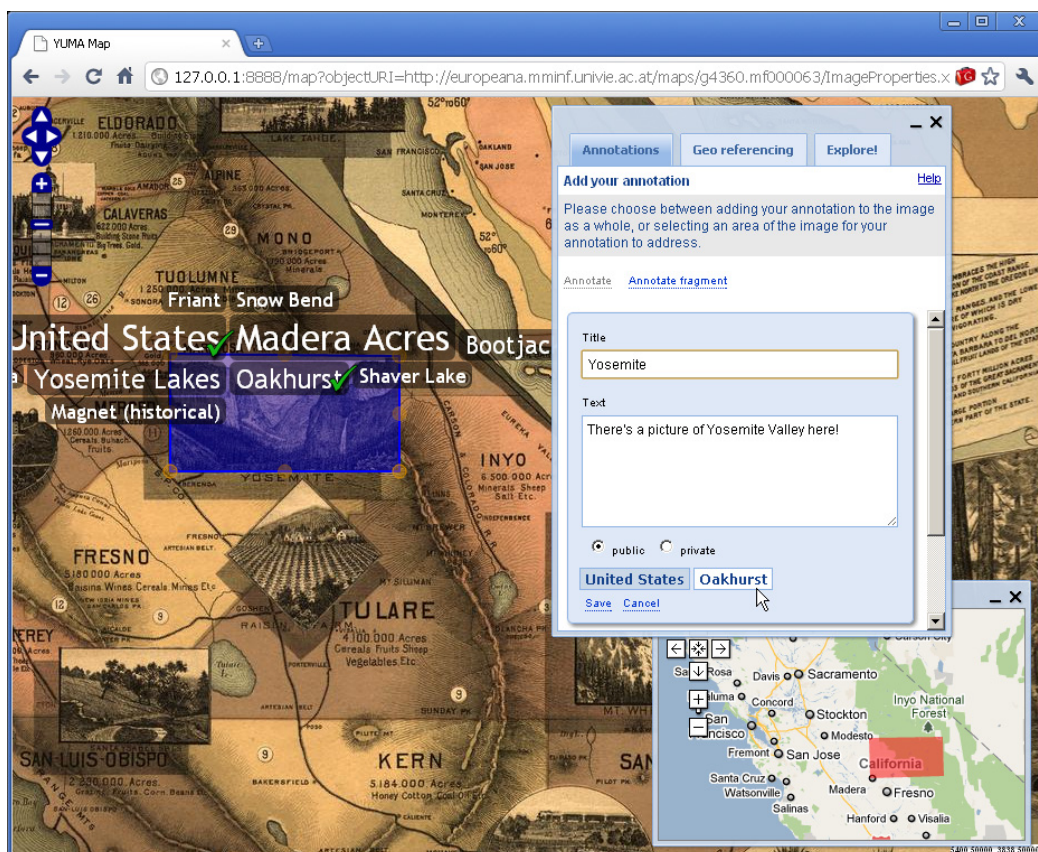

Figure 1: YUMA Map Annotation Tool Screenshot.

YUMA builds on a number of open source components such as the *OpenLayers*[4] Web mapping toolkit or the *Raphaël*[5] vector graphics library for the user interface, as well as several open source libraries to support framework functionalities: e.g. database persistence, RSS feed generation or the REST API used to store and retrieve annotations in different serialization formats. To display map images in the browser, the tool relies on zoomable Web image formats – in particular Zoomify[6] and TMS[7], which are natively supported by the *OpenLayers* viewer. However, the system is technically open for any type of map image and can perform on-the-fly conversion from a wide range of source image formats if no zoomable image is available. The integrated conversion functionality has meanwhile been made available as a stand-alone open source project named *MagickTiler*[8].

Semantic Linking

One central feature of YUMA is that it integrates s*emantic linking* into the annotation process (Haslhofer et al. 2010). When users create or edit their annotations, the system supports them in making them semantically more expressive by suggesting links to *semantic resources* which may be relevant to the context of the annotation. Using *Named Entity Recognition* (NER), the system attempts to identify mentions of e.g. place or person names in the annotation text, and will suggest appropriate links. Furthermore, if there is geo-referencing metadata available for the map, the system suggests links to relevant geographical entities pertaining to the annotated area. For example, if a user annotates the region of Yosemite National Park on a map of California (as shown in Figure 1), the system will suggest the country – United States – and cities in the area – such as Oakhurst or Yosemite Lakes. In the user interface, suggestions are presented in the form of a tag cloud. Hovering over a tag with the mouse pointer will display a short textual description for the suggestion, allowing the user to judge whether it is truly relevant to the annotation. If so, the user can accept the suggestion by clicking on the tag. The tag will then show a tick mark to indicate that it has now been selected, and that the link has been added to the annotation.

The foundation on which this approach builds is *Linked Data*: Linked Data refers to a set of best practices for publishing structured data on the Web using a graph-based data model (the *Resource Description Framework RDF*), and connecting data from different sources using *typed links* (Bizer et al. 2009). The amount of data published as Linked Data has been growing steadily over the past years[9]. Meanwhile, a range of data sets relevant to the domain of geography and cartography exist online. Noteworthy sets include Geonames[10], an online Gazetteer; LinkedGeoData[11], an RDF representation of data collected in the OpenStreetMap project[12]; or DBpedia[13], an effort to extract Linked Data from the Wikipedia online encyclopedia. There are additional sources on

---

[4] http://openlayers.org
[5] http://raphaeljs.com
[6] http://www.zoomify.com
[7] http://wiki.osgeo.org/wiki/Tile_Map_Service_Specification
[8] http://code.google.com/p/magicktiler
[9] http://richard.cyganiak.de/2007/10/lod/
[10] http://www.geonames.org/ontology
[11] http://linkedgeodata.org/
[12] http://www.openstreetmap.org/
[13] http://dbpedia.org

the Web that expose relevant data, although not yet according to the principles of Linked Data. A noteworthy effort in this regard is e.g. the Pleiades[14] gazetteer of ancient world place names.

Complementing Existing Metadata with Linked Data

Adding semantic context to user annotations provides a number of benefits. Firstly, unlike a free-form tag, a link to a semantic resource is not ambiguous. By clicking the "Oakhurst" tag, the user makes it explicit to the system that the annotation relates to the city of Oakhurst, California – not Oakhurst, Oklahoma, Oakhurst in the United Kingdom, or Oakhurst in New South Wales, Australia. Secondly, semantic resources are interlinked with more data on the *Linked Data* Web. This means that users can easily obtain additional information – e.g. short explanatory text abstracts, related geographic or demographic data, or information about related persons. They can furthermore discover related online content such as relevant Web pages or images. Leveraging Linked Data, it is also a straightforward process to obtain synonymous name variants, alternative spellings, or names in different languages for a resource. Since such information can be included by the system when storing the annotation, it can be considered during the retrieval process alongside the maps' original metadata. This way, multilingual and synonym search become possible without the need for additional manual cataloguing effort. Moreover, since all links generated with YUMA are user-verified, it can be expected that the quality and correctness of the links is high, in particular if links have been confirmed by multiple users independently.
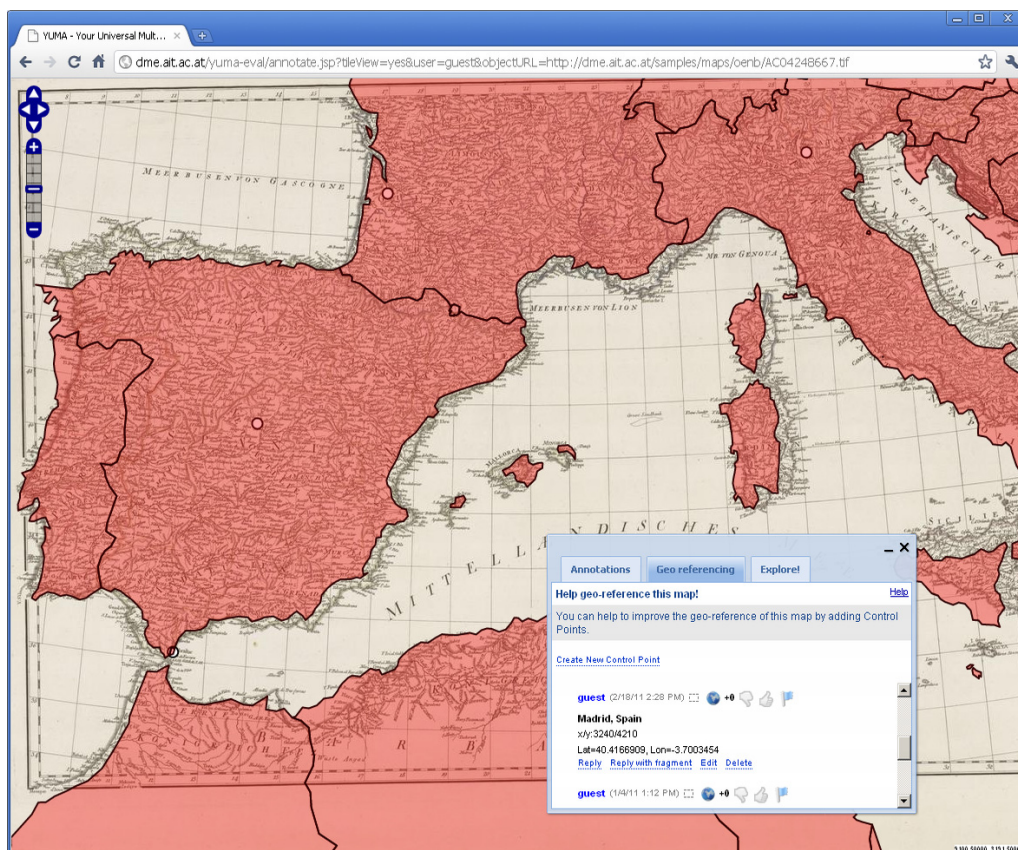

Figure 2: Approximate Geo-Referencing and Vector Data Overlay in the Current YUMA Implementation.

---

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

[4]

Exploiting Linked Data for Geo-Referencing

The current version of the YUMA Map Annotation Tool supports geo-referencing of maps by creating *control points*: users can place markers on locations they recognize, and provide geographical coordinates for those locations. Coordinates can be entered either manually, or by entering a place name which the system will then attempt to resolve automatically using a Gazetteer (Simon et al. 2010). The system uses these control points to perform piecewise interpolation, thus warping the geographical coordinate space into the coordinate space of the digitized image and vice versa (see Figure 2 for an example screenshot). Preliminary functional trials with test users, however, revealed that first time users, in particular, had difficulties to intuitively understand the principle or process of creating control points. At the same time, though, many users' initial impulse when trying out the system for the first time was that, unsurprisingly, they started to explore the map and intuitively added annotations to locations they knew. As annotation text, they would frequently simply add the name of the place, or the name and some extra anecdotal or personal context information.

We argue that this type of behavior can be exploited to generate structured location metadata for maps. A Gazetteer, or an appropriate Linked Data source such as Geonames or LinkedGeoData, can provide centroid, bounding box, or outline polygon coordinates, as well as geographical feature type information for places mentioned in user annotations. By linking this data to the user-created markers or polygons, control points can be created implicitly, as a by-product of annotations. To make this approach practically feasible, however, two pre-requisites must be met. First, although it can be expected that a place mentioned in an annotation is in some way related to the area annotated on the map, it is not necessarily equivalent to (or even located within) the area. To avoid "misinterpretation", the system needs to provide a suitable mechanism that allows users to specify the kind of relation between the mentioned place and the annotated area. This also requires an appropriate semantic vocabulary to express these relations. Second, similar to the tag cloud-based approach to Semantic Linking (where the system offers potential links, but the user chooses the appropriate ones), the system would need to provide a "human feedback loop" for control points: users should review control point suggestions first (e.g. by pre-viewing on a present-day map), and by explicitly accepting (or if necessary editing) them as a quality assurance measure.

Location metadata collected this way can certainly not replace accurate geo-referencing conducted by experts. But we argue that it will help to establish an approximate spatial footprint for the map with no additional cost for the collection holder. In the portal, this approximate information can then be exploited to improve geographical search (i.e. search by place name or through a map interface) or drive map-based result visualization (i.e. showing results to a keyword query on a map GUI). Furthermore, since the spatial metadata gets more complete and accurate as more users add to it, it can help to support or bootstrap the activities of domain experts and scholars working with specialized geo-referencing or analysis tools such as georeferencer.org[15] or MapAnalyst (Jenny 2006).

Towards an Evaluation of Collaborative Map Metadata Enrichment

At present, we have yet to evaluate our assumption that end-user annotations augmented with semantic links will (i) indeed provide an improvement over traditional metadata-only-based retriev-

---

[15] http://www.georeferencer.org/

al approaches in practice, and (ii) produce sufficiently accurate and correct location metadata. For verification, we are currently collecting real world data. For our research, a base-data set was kindly provided by the Library of Congress Geography and Map Division. The set consists of (a) 130.935 user *search queries* extracted from query logs collected in the map search portal over a period of two years, (b) 6.306 high-resolution *digitized map images* in the public domain, in JPEG 2000 and TIFF file formats, and (c) descriptive *metadata* for each of the maps. All maps were converted to *Zoomify* images using MagickTiler; the metadata was indexed using Apache Lucene.

Building the Ground Truth

In order to quantify the performance of different retrieval approaches – based on traditional meta-data only vs. including semantically linked annotations – the first step is to build a *ground truth* against which we can measure: a corpus of decisions as to whether a particular map is relevant to a search query or not. To build the ground truth, we initiated the *COMPASS Experiment*[16]. In this ongoing crowdsourcing effort we are collecting binary (yes/no) *relevance judgments* from invited volunteer users through a custom-built Web application. Volunteers were recruited from appropriate mailing lists in the map history and digital library domain (the MapHist list, and the IFLA DIGLIB and the INETBIB lists, respectively).
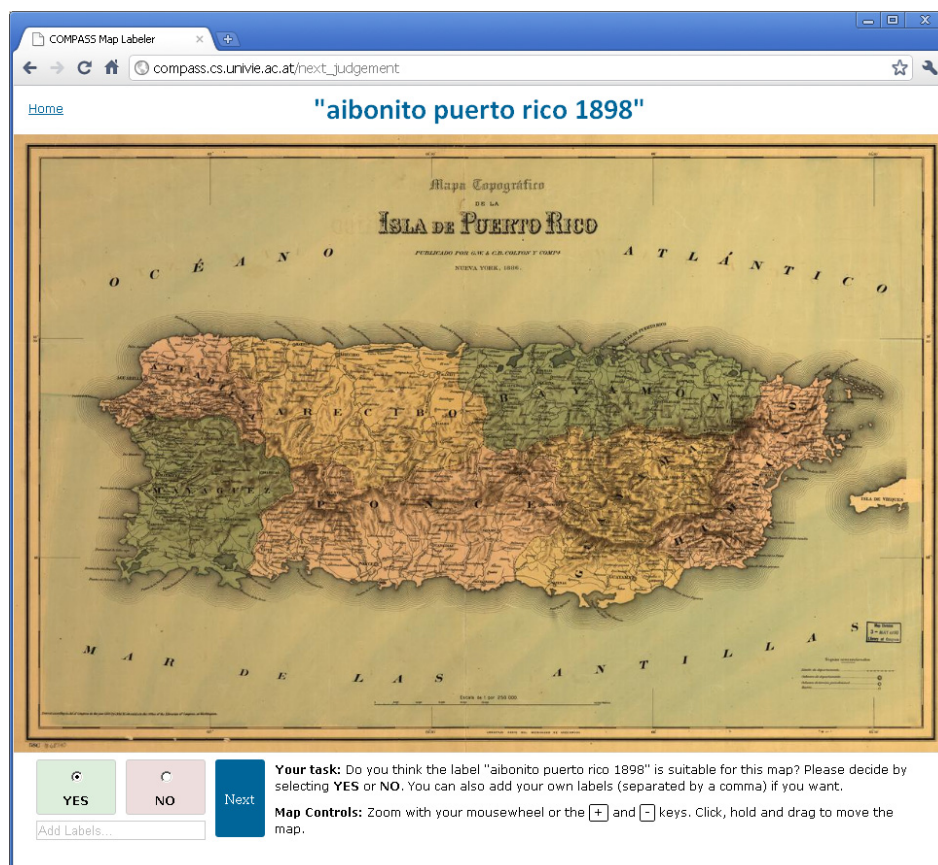

Figure 3: COMPASS Experiment Web Application.

---

[16] http://compass.cs.univie.ac.at

A screenshot of the application is shown in Figure 3. The title section displays a single search query, selected randomly from a subset of the queries. The main area of the screen shows the map, which is also chosen randomly out of the top ten ranked maps returned when querying the metadata index. The map is displayed with the open source OpenZoom[17] viewer, which allows the user to inspect the map in full detail, at the original resolution. At the bottom of the screen, YES/NO buttons allow the user to submit a relevance judgement for this map/query pair.

Collecting Annotations

In the next step of the experiment we will invite users to annotate maps using the YUMA Map Annotation Tool. We will index the collected annotations along with certain properties of the linked semantic resources, such as labels, descriptive abstracts, alternative names in different languages, etc. Based on the ground truth, we can then analyze the effects of including these properties on retrieval performance, using measurements of precision and recall.

Preliminary Results

Currently, we are still in the process of collecting relevance judgments to build a ground truth of reasonable size. At the time of writing, more than 75 users from at least 12 countries worldwide have joined the COMPASS experiment. In total, we have collected more than 1.800 judgments, which is approximately 45% of the judgments needed for a ground truth of 400 sample queries. Using the data collected so far, we conducted a first analysis of the effectiveness of pure metadata-driven retrieval. When disregarding all map/query pairs that have not yet received judgments, we have calculated a Mean Average Precision value of 41%. Despite the as yet incomplete data, this low value suggests that the performance of metadata-only based retrieval is clearly limited.

Summary and Outlook

In this paper we presented the YUMA Map Annotation tool, a prototype application for scholarly annotation of digitised old maps. We introduced the concept of *Semantic Linking*, i.e. the enrichment of unstructured annotations with structured semantics based on sources on the Linked Data Web. We discussed how Semantic Linking can complement traditional metadata and potentially enable advanced search and retrieval functionalities such as multilingual or synonym search. Furthermore, we argue that collaborative annotation combined with Semantic Linking can be exploited to obtain location metadata, which in turn enables geographic search in map collections that have not been geo-referenced yet. Professional geo-referencing or cartometric analysis tasks conducted by domain experts and scholars could be supported or bootstrapped. We presented an outlook on upcoming work in which we aim to verify our assumptions using test data provided by the Library of Congress, and collaborative input provided by volunteers in an ongoing crowdsourcing experiment.

As future work, we plan to focus on two areas. First and foremost, we will continue the COMPASS Experiment: we will collect annotations from volunteers in order to test our assumptions based on measurements of precision and recall. Recruiting a sufficiently large user base by disseminating information about the experiment more widely will be crucial in this regard. In reaction to early user

---

[17] http://openzoom.org/

feedback we also plan improvements to various aspects of the application's user interface and the introductory instructions for first time users. Second, we will continue ongoing work to align the YUMA framework more closely with relevant standards in the field. In particular this concerns the work of the Open Annotation Collaboration (OAC)[18], whose goal is to define a data model and ontology for describing scholarly annotations of Web-accessible information resources; and existing relevant geospatial standards governed by the Open Geospatial Consortium (OGC)[19], most importantly with regard to exposing location metadata, as well as the geographical footprint of annotated map areas in appropriate formats.

Acknowledgements

References

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

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

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

Haslhofer, B., Momeni, E., Gay, M., Simon, R. (2010) Augmenting Europeana Content with Linked Data Resources. In *6th International Conference on Semantic Systems*, Graz, Austria. ACM.

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.

Jenny, B. (2006) MapAnalyst – A digital tool for the analysis of the planimetric accuracy of historical maps. *e-Perimetron*, 1 (3): 239–245.

Simon, R., Korb, J., Sadilek, C., Baldauf, M. (2010) Explorative User Interfaces for Browsing Historical Maps on the Web. *e-Perimetron*, 5 (3): 132-143.

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. Available at: *http://www3.isrl.illinois.edu/ ~unsworth/Kings.5-00/primitives.html*.

---

[18] http://www.openannotation.org/
[19] http://www.opengeospatial.org/