Pelagios and the Emerging Graph of Ancient World Data

Leif Isaksen
University of Southampton
Highfield, Southampton
SO17 1BF United Kingdom
l.isaksen@soton.ac.uk

Rainer Simon
AIT: Austrian Institute of Technology
1220 Vienna, Austria
Rainer.Simon@ait.ac.at

Elton T. E. Barker
The Open University
Walton Hall, Milton Keynes
MK7 6AA United Kingdom
e.t.e.barker@open.ac.uk

Pau de Soto Cañamares
University of Southampton
Highfield, Southampton
SO17 1BF United Kingdom
p.desotocanamares@soton.ac.uk

ABSTRACT

This paper discusses an emerging cloud of Linked Open Data in the humanities sometimes referred to as the Graph of Ancient World Data (GAWD). It provides historical background to the domain, before gong on to describe the open and decentralised characteristics which have partially characterised its development. This is done principally through the lens of Pelagios, a collaborative initiative led by the authors which connects online historical resources based on common references to places. The benefits and limitations of the approach are evaluated, in particular its low barrier to entry, open architecture and restricted scope. The paper concludes with a number of suggestion for encouraging the adoption of Linked Open Data within other humanities communities and beyond.

Categories and Subject Descriptors

J.2 [**Computer Applications**]: Physical Sciences and Engineering—*Archaeology*; J.5 [**Computer Applications**]: Arts and humanities

Keywords

Linked Open Data; Humanities; Geospatial

1. INTRODUCTION

As recently as 2011, one of the authors of this paper concluded their doctoral thesis with the following claim and query: '*it remains a moot point as to whether Berners-Lee's vision of an open and decentralized Knowledge Representation is possible. The question left for archaeologists to consider is: Could an open and decentralized archaeology be possible?*'[8] This paper argues that since then a quiet revolution has been taking place which suggests that open and decentralised Linked Data is not only possible in archaeology but in the humanities more widely. This is a different issue from whether the field is willing or able to adopt the *technologies* of the Semantic Web. RDF has been used in a variety of humanities projects since at least the early Noughties[1][7], and conferences and funding organisations continue to deem it an area of significance in the Digital Humanities. Rather, the change has been towards an ecology-driven approach in which a community of independent initiatives has gradually increased mutual connectivity by creating and using Linked Open Data.

Perhaps the most significant area of growth has been in what is sometimes dubbed the Graph of Ancient World Data (GAWD)[13], encompassing both philological and archaeological approaches to the study of antiquity. This trend was unquestionably facilitated by the National Endowment for the Humanities (NEH) funded Linked Ancient World Data Institute summer programme which ran in the summer of 2012 and 2013 and has recently published a series of short articles in a special issue of *ISAW Papers*[6]. Yet these two workshops are in themselves insufficient to explain why more than fifty widely divergent projects and organisations from the public, academic and private sectors have so quickly begun to establish connections through the Semantic Web. This is especially given these fields' historical tendencies to work in comparative isolation, and grant prestige principally to the work of individual scholars rather than collaborative ventures.

This paper attempts to explain: why Linked Open Data approaches have started to gain traction in this area; the current state of play; what it may mean both for the future of academic and public engagement with the ancient world; and subsequently draws wider conclusions for the humanities and Linked Open Data communities.

2. WHY THE ANCIENT WORLD?

Classics and archaeology are not fields which many associate with the bleeding edge of technology, let alone Web Science, but in fact they have formed the domain of several significant initiatives in the history of open digital data, including the Perseus Digital Library[4] and the work of Rahtz, Hall and Allen on hypertext for excavation reports [12]. Are there any particular features of these disciplines, or their union, that either makes them especially well-suited

WebSci'14, June 23–26, 2014, Bloomington, IN, USA.

ACM 978-1-4503-2622-3/14/06 ...$15.00.
http://dx.doi.org/10.1145/2615569.2615693.

to Linked Open Data or encourages its adoption? We argue that the following factors have all played a role:

A diverse but tractable domain The study of antiquity is divided into a wide array of individual subfields, both in terms of a traditional separation between philological and archeological evidence, but also specialisms which cross this divide, including prosopography (historical individuals and their relationships), numismatics (coins), epigraphy (inscriptions), geography, political, military and social history, and so on. Yet these often starkly differing approaches are united by a reasonably well-defined domain with a limited, if slowly growing, evidence base. A vast proportion of this information is hidden not beneath the soil or on dusty monastery shelves, but within libraries, museums and archives compiled in the nineteenth and twentieth centuries that are increasing available online. Furthermore, as Greece and Rome were literate cultures, at a basic level we are often able to identify and refer to emic concepts, i.e. those which originated in the languages of those periods. These are arguably more stable than the etic conceptual schemes developed by contemporary scholars to describe cultural phenomena for which no linguistic evidence survives.

Controlled vocabularies An extremely important development has been the establishment of services providing stable URIs for shared categorical and instance thesauri. These include place gazetteers,[1] type classifications for coins[2] and canonical citations for classical literature.[3] Without them, earlier attempts at 'interoperability' were seriously hampered by the lack of common reference terms for analogous content despite the availability of ontologies that defined shared or equivalent properties.

Simple ontologies The CIDOC Conceptual Reference Model (CIDOC CRM) remains perhaps the most powerful ontology available for describing the creation, evolution and destruction of cultural heritage[5]. Nonetheless its complexity, in combination with unfamiliar Linked Data technologies such as RDF, has proven off-putting to newcomers. Ontologies such as Open Annotation[4] have offered an easier on-ramp, along with a variety of direct benefits to both contributors and users, without preventing the adoption of additional (and more sophisticated) ontologies later on.

Sufficient open data An enormous amount of information about the ancient world remains inaccessible to the general public and researchers at all but a handful of elite institutions[9]. However, the tightly-knitted nature of the field has meant that much interrelated material, especially ancient text, is increasingly available. The situation for material culture is more varied but pioneering work by organisations such as the German Archaeological Institute, the Alexandria Archive and the UK Archaeological Data Service may be stimulating progress on open archaeological data elsewhere.

[1]http://pleiades.stoa.org
[2]http://numismatics.org/ocre
[3]http://cts3.sourceforge.net
[4]http://www.openannotation.org

The most significant difference between GAWD and earlier 'Semantic Web' developments in the humanities is the increasing interconnection between heterogeneous, independently maintained resources through the common use of URIs. Whereas earlier initiatives were often characterised by intensive collaboration between small numbers of projects, often without persistent URIs or making data openly accessible[8], GAWD is an informal collective of independent participants treating Linked Data as as just one more means of making their data more accessible. So how does this 'ecosystem' work in practice?

3. CASE STUDY: PELAGIOS AND GEOGRAPHIC ANNOTATION IN GAWD

An example of this decentralised structure is the interconnection of resources based on common references to place. The foundations for this were laid by the Pleiades Gazetteer of the Ancient World, developed and hosted by the Institute for the Study of the Ancient World, New York University. Initially conceived of as an online and community-driven continuation of the *Barrington Atlas of the Greek and Roman World*[15], it was soon realised that providing a stable URI would be an essential for each entry in the gazetteer. This would not only allow other projects to derive information such as coordinate locations automatically, but furthermore that they could act as a point of intersection between projects otherwise unknown to one another. The Pelagios project, led by several of this paper's authors and supported by Jisc , a UK funding body, took on the task of formalising this process while seeking to maintain the twin principles of openness and decentralisation[2].

In consultation with a variety of stakeholders it proposed the use of the Open Annotation ontology which describes an annotation comprising a *target* URI representing a document (or fragment thereof) and a *body* representing its content[14]. While the general specification allows the latter to have any value, Pelagios compliance requires it to point to a URI representing a place defined in a digital gazetteer such as Pleiades or Geonames.[5] While such annotations are themselves extremely simple (essentially tripartite links), collectively they form a two-mode graph of associations between document[6] resources and places. This allows not only for the first-order querying of places associated with a document and vice versa, but questions of greater interest to humanists - which places are commonly referred to together? Which documents appear to cover similar geographic territory? Additionally they are language neutral, an important consideration for a field that operates across many modern European languages as well as Latin and Ancient Greek. The addition of further metadata to the annotations, such as the specific toponym used, as well as the date, time and author of the annotation, can collectively provide valuable information for understanding variation in the way geographic concepts are referred to, and help address issues of provenance and trust.

A fundamental premise of the Pelagios initiative was that it should avoid social and technical bottlenecks wherever possible. This was achieved by encouraging individual re-

[5]http://geonames.org
[6]'Document' is here used to denote any kind of human interpretable online resource, whether image or text, static or dynamically generated.

source providers to produce and host their own annotations. RDF proved to be a powerful format in this regard. Simple ontologies such as Open Annotation are reasonably comprehensible to the technically literate, can be templated easily, and expressed in a range of notations and technical solutions suited to any level of Web-based hosting. The capability (in terms of both knowledge and resources) to host Web content remained a prerequisite, but one which - almost by definition - any provider of online ancient world resource is likely to meet. On the other hand, allowing third parties to annotate content and host it in a decentralised fashion remains an open challenge in a field where few day-to-day practitioners have either experience or facilities for Web-hosting. In addition to hosting their annotations, resource providers were also encouraged to release them under the most liberal licensing terms possible, ideally CC0.[7] There remains an important question as to whether such annotations should be deemed to constitute data, and thus be licensed under an Open Data Licence such as PDDL.[8] The Pelagios stance is that such annotations usually constitute an interpretive rather than a factual assertion to the affect that 'reference x refers to place y' as the author's intention is inherently inaccessible to the annotator unless they are one and the same person.

Pelagios annotations thus form an interconnected and open set of RDF triples, dispersed across the Web. While this is an important outcome that meets the objectives of both openness and decentralisation, it is also not terribly easy for information consumers to make use of. The Pelagios project therefore established a demonstrator Webservice which harvests such annotations and makes them available through a human-readable search interface, a machine-readable API,[9] and a series of embeddable widgets.[10] The API provides a number of functions of direct benefit to to those who have annotated their own content, not least of which is the ability to access a growing cloud of related content hosted elsewhere, as well as the increased likelihood of discovery by consumers following the same API in the opposite direction. Indeed, it is precisely the utility of this interface that led to a rapid growth in both partners and content - from five institutions in the first phase of the project to almost forty at the time of writing and some 800,000 annotations.[11] It might be asked whether the claim for decentralisation is merely sleight-of-hand if both contributors and consumers are making use of this API rather than the source data? We share concerns that any system which relies too heavily on a single point of failure will ultimately prove unsustainable, and thus encourage the development of alternative APIs and web services, building upon the project's open source code base where this assists, ignoring it where it does not. The fact that every aspect of the project is open access, from content to code, to a cookbook of best practices, means that no aspect is not reproducible elsewhere should the need or desire arise.

Contextualisation and discovery are not the only benefits that large-scale, distributed, but structurally simple, graphs can bring. Traditional humanistic questions may also be approached with such data. For instance, the latest cycle of Pelagios, funded by the Andrew W. Mellon Foundation, is creating infrastructure and content for the annotation of Early Geospatial Documents (EGDs) extending up to the end of the Pre-Modern period (c. 1500). The ability to compare the places referred to in maps, itineraries and geographic descriptions across diverse linguistic and ethnic traditions is likely to transform our understanding of historical developments in geographic thought. It is also a clear indicator, were it needed, that Linked Open Data is not exclusively suited to classical resources but has much to offer the study of, and engagement with, other regions and periods as well.

Geography is only one of many dimensions across which Linked Open Data can interconnect online resources. Work on Canonical Text Services is creating Web-based infrastructure for uniquely identifying canonical citations in classical texts[3]. Such citations form the backbone of most scholarly research in this literature, providing a global reference system that transcends arbitrary page numbering divisions. The Standards for Networking Ancient Prosopographies (SNAP)[12] project is similarly defining both URI and annotation conventions for referencing ancient people. This introduces new challenges: while identifying a shared conception of a place can often be achieved 'intuitively' by means of geodetic, administrative or mereological relationships, people can be harder to denote, especially where the evidence is fragmentary. Should Aristotle be defined by his place of birth, his association with Athens (of which he was not a citizen), his contributions to philosophy (which?), his tutoring of Alexander the Great, or a combination of these and other 'facts'? What if such identifications are controversial or turn out to be wrong? Establishing best practices for this process will be an important contribution to GAWD. Finally, classificatory thesauri for fields such as numismatics and ceramics may over time greatly assist our ability to compare distribution networks from archaeological excavations which are currently very difficult to assimilate[10].

4. SIGNIFICANCE FOR WEB SCIENCE

Collectively, these developments suggest a number of lessons for those seeking to introduce Linked Open Data practices to the humanities. Despite longstanding concerns as to the the humanities' tendency towards individualism and technical illiteracy, the case is now clear that decentralised Linked Open Data is not only possible but can flourish in this field. Nonetheless, much work remains to be done. Few if any humanities domains have embraced this model to the same degree as the ancient world, although that need not deter us unduly given how rapidly the situation can change. Furthermore, we are at the tip of the iceberg even in this case as the overwhelming majority of classicists and classical archaeologists have never heard of Linked Open Data, let alone contributed to it. It may be an unrealistic, perhaps even undesirable, goal to believe that they should, but we must certainly aim for a scenario in which they can benefit, and ideally offer their own content, regardless of whether they care for the terminology or grasp all of its underlying principles. How might this be achieved? The following suggestions reflect the authors' experiences contributing to GAWD:

[7] https://creativecommons.org/publicdomain/zero/1.0
[8] http://opendatacommons.org/licenses/pddl
[9] https://github.com/pelagios/pelagios-cookbook/wiki/Using-the-Pelagios-API
[10] http://pelagios-project.blogspot.co.uk/p/pelagios-in-use.html
[11] http://pelagios.dme.ait.ac.at/api/datasets
[12] http://snapdrgn.net/about

1. Simplicity is essential if we are to attract contributors with little prior investment in Semantic Web technologies. A huge advantage of the Linked Data approach over even relational database technologies is that almost everyone is now familiar with URLs, and the conceptual leap to URIs is neither difficult, nor ultimately essential for day-to-day users. Likewise, much of the apparent complexity faced by those acquainting themselves with Linked Data for the first time lies with the ontologies, rather than RDF *per se*. We should seek to attract broader communities by proposing Linked Data activities with immediate benefit, only introducing greater complexity as required. Complex ontologies remain useful however, and initiatives to facilitate their use, such as SENESCHAL[13] and ARCHES[11] are important.

2. Quickly establishing a critical mass of open and related content is a tremendous motivator, as the benefits of Linked Open Data largely derive from the ability to associate the contributor's content with external content. Without it, there are few technical advantages that RDF encoding provides which cannot be achieved by means with which the contributor is likely to be more familiar. Fortunately, once this critical mass is established, there is often a snowballing effect by which the benefits of connecting to the data cloud continually increase against the (stable) cost of contributing to it. Historically, much production of Linked Data has been happenstance, often producing semantic silos of conceptually unrelated content. As a community we should seek to target groups of related datasets, then continuously build around them, bridging between clusters where we can. Wide-application concepts such as geographic location are especially good for this.

3. Linked Data is sometimes discussed as though it exists as its own parallel Web, unpolluted by the Web of Documents. This is highly detrimental to its adoption. Linked Open Data approaches should be used in a 'mixed economy' of multiple technologies, each used for the task to which it is best suited. For instance, the dimensional aspect of geospatial information is poorly suited to expression and visual representation as Linked Data. Likewise, GIS and web mapping technologies require a geometric primitive to locate every entity and handle conceptual associations between places poorly. Attempting to reduce geographic knowledge solely to either is neither necessary nor helpful. The same holds true of web services, relational data formats, statistical datasets and mathematical functions.

4. Just like the Web, the wider graph of humanities data (and indeed all Linked Open Data) will grow organically, not be built according to a software architect's plan. With this in mind it makes sense for us to concentrate on small, individual steps which offer immediate benefits, whilst remaining aware of their limitations. There has been perhaps been too much emphasis by funders on Virtual Research Environments or domain ontologies which are expected to cater to every conceivable humanities question. This is not only a failure to understand the nature of humanistic inquiry (which seeks to challenge conceptual models, rather than defer to them), but can prevent us from focussing on goals more readily within our grasp. Growth can occur in two dimensions - both through the expression of data according to established URI schemes, or through the addition of new infrastructural components such as controlled vocabularies and ontologies. We should expect the former to be far more frequent than the latter and require a far lower level of technical capability. We should also maintain the principle that contribution can be at any level - that as the possibilities for contribution grow more complex, the requirements for doing so do not. A recent positive development in this area has been the Getty Research Institute's publication of its thesauri as Linked Open Data.[14] These widely used datasets could see yet greater adoption as online resources align content with their URIs without the earlier impediment of licensing fees.

5. Such organic growth in turn requires multiple stakeholders to take responsibility for clearly defined aspects of the wider graph. The nature of this responsibility will vary. Those maintaining controlled vocabularies or ontologies will need to guarantee at least a moderate level of stability and documentation, along with clearly defined contingency plans should the service fail or be discontinued. In contrast, the requirements for those aligning otherwise 'non-semantic' content with such services might be requested simply to provide identifying information for the purpose of provenancing and other such metadata.

6. This last point touches on perhaps the most important challenge of all: trust. GAWD has grown largely due to the establishment of trust at a range of levels among the parties involved. This is partly a matter of sustainability. When investing time producing RDF that aligns content with URIs offered by an external body it is important to believe that they will remain online for the foreseeable future, and have a plan for what happens if they do not. There is an additional social component to this trust network, however. If content is accessible through third parties, it is essential that all parties receive appropriate levels of attribution, regardless of licensing stipulations, and provenance be transparent. Obscuring the source of content will disenchant consumers and contributors alike.

5. CONCLUSIONS

Linked Open Data—conceived of as an ecology of independent online resources interconnected by RDF—has finally taken significant hold in an area of the humanities. The factors leading up to this development are manifold and it is not yet clear how easily repeatable they are. Nonetheless, this provides strong evidence that RDF-based approaches are applicable outside the laboratory and in scenarios that do not require extensive financial resources or support for complex technical solutions. Most importantly, they can also work across wide networks of stakeholders and show

[13] http://www.heritagedata.org/blog/about-heritage-data/seneschal

[14] http://www.getty.edu/research/tools/vocabularies/lod/

potential for growth. By fostering open but collaborative initiatives, with institutions taking responsibility for clearly defined roles, a nascent Semantic Web for the humanities is starting to emerge.

6. ACKNOWLEDGMENTS

The authors would like to thank Jisc, the Andrew W. Mellon Foundation, the organisers of the Linked Ancient World Data Institute, and all contributors to the Graph of Ancient World Data.

7. REFERENCES

[1] M. Addis, M. Boniface, S. Goodall, P. Grimwood, S. Kim, P. Lewis, K. Martinez, and A. Stevenson. SCULPTEUR: Towards a New Paradigm for Multimedia Museum Information Handling. pages 582–596. 2003.

[2] E. Barker. Welcome to PELAGIOS, Feb. 2011.

[3] C. Blackwell and N. Smith. A Brief Guide to the Canonical Text Service. Technical report, Homer Multitext Project, May 2013.

[4] G. Crane. The Perseus Project and Beyond: How Building a Digital Library Challenges the Humanities and Technology. *D-Lib Magazine*, Jan. 1998.

[5] N. Crofts, M. Doerr, T. Gill, S. Stead, and M. Stiff. Definition of the CIDOC Conceptual Reference Model. Version 5. Technical report, ICOM, Jan. 2011.

[6] T. Elliott, S. Heath, and J. Muccigrosso, editors. *Current Practice in Linked Open Data for the Ancient World*, volume 7 of *ISAW Papers*. Institute for the Study of the Ancient World, New York, 2014.

[7] B. Fuchs, L. Isaksen, and A. C. Smith. The Virtual Lightbox for Museums and Archives: A Portlet Solution for Structured Data Reuse Across Distributed Visual Resources. In *Museums and the Web 2005*, Victoria, 2005. Archives & Museum Informatics.

[8] L. Isaksen. *Archaeology and the Semantic Web*. PhD thesis, University of Southampton, 2011.

[9] E. Kansa. Openness and Archaeology's Information Ecosystem. *World Archaeology*, 44(4):498–520, 2012.

[10] A. Meadows and E. Gruber. Coinage and Numismatic Methods. A Case Study of Linking a Discipline. *ISAW Papers*, (7), 2014.

[11] D. Myers, Y. Avramides, and A. Dalgity. Changing the Heritage Inventory Paradigm: The Arches Open Source System. *Conservation Perspectives: the GCI Newsletter*, pages 4–9, 2013.

[12] S. Rahtz, W. Hall, and T. Allen. The Development of Dynamic Archaeological Publications. In S. Rahtz and P. Reilly, editors, *Archaeology in the Information Age*. Routledge, 1992.

[13] R. Robineau. Graph of Ancient World Data, June 2012.

[14] R. Sanderson, P. Ciccarese, and V. d. S. Herbert. Open Annotation Data Model. Technical report, W3C, Feb. 2013.

[15] R. J. A. Talbert. *Barrington Atlas of the Greek and Roman World*. Princeton University Press, Princeton, 2000.