Hugh A. Cayless

Sustaining Linked Ancient World Data

Abstract: May 31st, 2018 marked the sixth anniversary of the Linked Ancient World Data Institute (LAWDI), a workshop funded by the US National Endowment For the Humanities. This makes it a good time to take stock of the Ancient World Linked Data initiatives that have been around for some time, as well as some that have foundered and some that are new. What makes for sustainable Linked Open Data? Why do some initiatives thrive while others fail? What resources do successful LOD sites need, and how may *they* be obtained? The promise of LOD is that it frees our information from the silos in which it is housed, permitting cross-system interactions that improve the quality and usefulness of the information in any single system. This article will take the broader view of the definition of Linked Data suggested by Tim Berners-Lee's foundational "Linked Data – Design Issues" paper, as encompassing more types of data than simply RDF and other "Semantic Web" technologies. This view of LOD is pragmatic and leverages the strengths of semantic technologies while avoiding their weaknesses.

Introduction

The title of this paper will require some definition before discussion of its subject matter can proceed. What is "sustainable" data? What is "Linked Data"? What counts as "Ancient World" data? May 31st, 2018 marked the sixth anniversary of the first Linked Ancient World Data Institute (LAWDI), a program funded by the US National Endowment for the Humanities (NEH).[1] A number of projects represented at LAWDI's two events, at the NYU Institute for the Study of the Ancient World in 2012, and then the following year at Drew University are still up and running, meaning they have successfully passed the startup phase. This paper will examine five of these long-running projects in the field of Ancient Studies which may be considered Linked Open Data sites and discuss how they have managed to sustain themselves and what their prospects for the future are.

[1] See Elliott (2014) for follow-up articles by many of the participants.

Hugh A. Cayless, Duke University

https://doi.org/10.1515/9783110599572-004

Broadly speaking, data, and the applications that disseminate data, may be said to be sustainable when their maintenance costs do not exceed the resources available and are not likely to do so in the future. Moreover, the communities that use that data should find its continued availability important enough to contribute to its maintenance, whether monetarily or via their own labor. Data sets may be fairly static, e.g. reports of completed work, or may require periodic revision; they may grow steadily as new data are deposited and updated or remain relatively constant in size. Different types of curatorial intervention and expertise will be required depending on whether data sets change by addition or via editing, and both scholarly and technical expertise may be required in order to keep them going. Questions of survivability factor into the data sustainability question also. How hard would it be to migrate the data to a new dissemination platform? How hard are they to edit? Would they survive a period of neglect?

Sustainability boils down to questions about the nature of the data and the community's investment in its continued availability. Who is responsible for it? How available and discoverable is it? Is its maintenance funded or voluntary? What systems does it depend upon in order to remain available? What are the costs of maintaining it? As we will see, there are a number of possible answers to these questions, and making Linked Open Data sustainable requires a combination of strategies, including institutional support, collaboration agreements, keeping costs manageable, keeping user communities engaged, and keeping (or at least exporting) data in forms that can survive a loss or transition of support.

Turning to Linked Open Data, we find a similar set of questions. There is an inherent tension in the definition of Linked Data over how that data should be represented. Must it be modeled according to the Resource Description Framework (RDF)? Can Linked Data be in any format made discoverable via a set of encoded relationships? Berners-Lee's original notes on the subject in "Linked Data – Design Issues"[2] define five levels, of increasing quality:

1. Available on the web (whatever format) but with an open licence, to be Open Data.
2. Available as machine-readable structured data (e.g. excel instead of image scan of a table).
3. As (2) plus non-proprietary format (e.g. CSV instead of excel).
4. All the above plus: Use open standards from W3C (RDF and SPARQL) to identify things, so that people can point at your stuff.
5. All the above plus: Link your data to other people's data to provide context.

[2] (Berners-Lee 2006).

This scheme is, on the face of it, agnostic about what data should be represented in what format (with a bias towards non-proprietary formats), but most subsequent implementations and interpretations of “Linked Data” have focused on RDF and the suite of protocols around it as the delivery mechanism (not simply the means of identification) for information, and many LOD datasets have thus been published encoded entirely in RDF formats. In this guise, the Linked Data enterprise seems clearly to be a continuation of the original Semantic Web, an idea originally popularized by an article in *Scientific American*, also by Berners-Lee. Indeed, the definition given on the W3C’s site explicitly ties Linked Data to the Semantic Web.[3] For the purposes of this paper, however, I will consider sites that make an attempt to follow Berners-Lee’s general principles, but do not necessarily store, nor expose *all* of their data as RDF as “Linked Open Data” projects. Further, I will argue that to do so would incur the risk of exploding the costs of already-expensive projects. The LOD sites we will examine take a pragmatic view which leverages the strengths of Linked Data architectural styles and semantic technologies while avoiding their weaknesses.

Linked Ancient World Data sites

The projects which were represented at the LAWDI meetings and which this paper will examine are Pleiades, which serves as a digital gazetteer of ancient places, Papyri.info, which publishes texts and data relating to ancient handwritten documents on surfaces such as papyrus and ostraca, Trismegistos, which aggregates data about ancient documents, people, and places, Open Context, which collects archaeological reports, and Nomisma, which provides a thesaurus of numismatic concepts with links out to coin records in a variety of numismatic datasets.

Pleiades (https://pleiades.stoa.org/) is arguably the oldest of these, having originally been conceived in 2000, as a follow-on to the printed *Barrington Atlas of the Greek and Roman World*.[4] Formal work on the project did not begin

[3] W3C, *Linked Data*, passim. “Linked Data lies at the heart of what Semantic Web is all about”; “To achieve and create Linked Data, technologies should be available for a common format (RDF), to make either conversion or on-the-fly access to existing databases (relational, XML, HTML, etc)”.

[4] Ed. by Talbert (2000).

until 2006, however, after a successful funding bid to the National Endowment for the Humanities.

> Pleiades has received significant, periodic support from the National Endowment for the Humanities since 2006. Development hosting and other project incubation support was provided between 2000 and 2008 by Ross Scaife and the Stoa Consortium. Additional support, primarily in the form of in-kind content research and review, has been provided since 2000 by the Ancient World Mapping Center at the University of North Carolina at Chapel Hill. Web hosting and additional financial support (not least our annual hosting costs and my time as managing editor) has been provided since 2008 by the Institute for the Study of the Ancient World at New York University.[5]

Pleiades's internal data model does not rely on RDF, but it does publish its data in various forms, which include RDF (see https://pleiades.stoa.org/downloads). The system is built on top of Plone, a Content Management System based on the Zope application server, written in Python. It deals with entities in the form of Places, Locations, and Names. Places are abstractions which may be associated with zero or more Locations and Names. Each of these entities will have an HTTPS URI that identifies it. For example, https://pleiades.stoa.org/places/727070 (Alexandria) has an associated location (https://pleiades.stoa.org/places/727070/darmc-location-1090) and a set of names, e.g.

> Alexandreia ad Aegyptum: https://pleiades.stoa.org/places/727070/alexandreia-ad-aegyptum
> Alexandria: https://pleiades.stoa.org/places/727070/alexandria
> al-Iskandariya: https://pleiades.stoa.org/places/727070/al-iskandariya-1

All of these have variant spellings. Because Pleiades treats these as distinct "pages" a search for "al-Iskandarīya" on Google will turn up the name page listed above, which will in turn direct the searcher to the Place record for that Alexandria (there are many).

Parts of the Papyri.info data set began their existence much earlier.[6] The Duke Databank of Documentary Papyri (DDbDP) began work in 1982, and was issued on CD-ROM. The Advanced Papyrological Information System (APIS) and the Heidelberger Gesamtverzeichnis (HGV) began in the 1990s. The DDbDP reproduced the texts of published editions of papyrus documents; HGV holds expanded metadata about them, including bibliography, better provenance

[5] Elliott, personal communication, 2018-09-21.

[6] The author was the principal architect of the Papyrological Navigator – the browse and search portion of the Papyri.info site.

information, some translations, and links to images where available; APIS contains what are essentially catalog records, focusing on description of the artifact, along with images for some of the papyri and translations. Thus, the DDbDP and HGV are focused on editions, while APIS focuses on the document itself. Data from Trismegistos[7] was added on more recently.

Planning to revive the DDbDP, which was no longer being actively edited, and whose data had been hosted by the Perseus Project since the mid-1990s, began in 2006. Thanks to grant funding from the Mellon Foundation and the NEH, Papyri.info was developed as an update and replacement for the discovery facilities provided by Perseus and as a means to crowdsource the editing of the data, which the DDbDP was no longer able to sustain at Duke. Papyri.info began by following the some of the principles Berners-Lee outlined: all data would be openly available and licensed for re-use, each document would have a stable URI that both identified it and served allowed its retrieval, but it did not initially use any RDF technologies. Because the system is an amalgamation of several datasets, which do not align perfectly, deciding how to assemble the information was quite tricky. HGV might treat as many what the DDbDP considered as a single document, for example. Or HGV might rely on a different publication as the "principal edition". APIS might treat documents differently than either of the other two because of its emphasis on the artifact. An edition might assemble multiple fragments (with different curatorial histories) into a single text, for example.

All of this meant unifying the display of information about a papyrus document was not straightforward. The datasets knew about each other, and referenced each other to an extent, and after a few false starts, the project settled on using RDF to describe the links between records in the different datasets. Relationships between records are extracted from the source documents and then used to generate an aggregate view of each document. A page in Papyri.info like http://papyri.info/ddbdp/p.fay;;110 pulls together data from HGV, Trismegistos, APIS, and the DDbDP. Exploration of the Linked Data section linked at the bottom of the page will reveal that the source for the page's data is http://papyri.info/ddbdp/p.fay;;110/source, which is related to:

the TM text, https://www.trismegistos.org/text/10775,
the HGV record, http://papyri.info/hgv/10775/source,
the APIS record, http://papyri.info/apis/columbia.apis.p387/source,
the APIS images, http://papyri.info/apis/columbia.apis.p387/images.

[7] Trismegistos (https://www.trismegistos.org: last access 2019.01.31) will be treated in more detail below.

These relations are stored in an RDF triple store referred to as the "Numbers Server". This keeps track of the relationships between content from the various collections, as well as information about superseded editions in the DDbDP. All of Papyri.info's textual data is also maintained in a GitHub repository.[8] An hourly sync process keeps the data current. The system uses a triple store to manage relations between documents, which are stored on disk. Text documents are stored as TEI EpiDoc files, versioned using Git. So while Papyri.info makes use of RDF, it makes no attempt to store nor expose all of its data in that form.

Trismegistos (TM) began development in 2005, when its director, Mark Depauw, received a Sofja Kovalevskaja Award from the Alexander von Humboldt-Stiftung. The project, 'Multilingualism and Multiculturalism in Graeco-Roman Egypt', was the foundation of Trismegistos, which has grown beyond its initial focus on Egypt to encompass ancient documents of all kinds. Trismegistos assigns unique URL identifiers to documents, which means it can serve as a "data hub" for identifying documents across projects, in much the same way as Pleiades functions for places. Trismegistos and Papyri.info have a close relationship, in which TM identifiers help serve to disambiguate documents for the PN, and the PN's data is used as a source for TM's research. The two sites collaborate and interlink their documents extensively. Data exchange from TM to Papyri.info remains somewhat informal, based on periodic data dumps, while TM relies on Papyri.info's GitHub repository. As we have already seen, TM URLs are in the form https://www.trismegistos.org/text/10775. Besides texts, TM collects data around Collections, Archives, (ancient) People, Places, (ancient) Authors, and (modern) Editors. TM manages its data using a FileMaker Pro database, which exports to a MySQL database that serves as the back end of the PHP-based TM website. It does not export nor expose any RDF.

Open Context began in December 2006. It provides a platform for the publication, archiving, and annotation of archeological data. The site has gone through several cycles of refactoring, from PHP and MySQL, to PHP-Zend Framework, MySQL and Solr, to its current state as a Python-Django, PostgreSQL, and Solr site. Open Context is organized around Projects, Subjects, and Media, each instance of which has its own stable URL in the following forms:

Projects: https://opencontext.org/projects/3DE4CD9C-259E-4C14-9B03-8B10454BA66E
Subjects: https://opencontext.org/subjects/0801DF9C-F9B2-4C76-0F34-93BE7123F373
Media: https://opencontext.org/media/48c1bdeb-ffb9-4fd3-84d2-20ba189a1f4a

[8] https://github.com/papyri/idp.data (last access 2019.01.31).

While it does not use RDF internally, Open Context models its data in a PostgreSQL database in a graph-like fashion, and it only produces RDF for external services (e.g. Pelagios) to consume. Most consumers of its data prefer to receive it in tabular form. Eric Kansa reports that, while an internal RDF triple store is a desideratum, questions of data provenance and versioning, and the difficulties RDF has with these problems, make it a low priority.[9]

Nomisma is the youngest of the projects we will discuss, having first begun in 2010, and also adheres most closely to the standard definition of a Linked Open Data site, as it models and stores all of its data in RDF. The site provides "stable digital representations of numismatic concepts". These concepts serve as a backbone for browsing and querying across several numismatic datasets. Nomisma entities are drawn from concepts such as mints, coin types, and numismatic concepts, and these link out to datasets from sources including the American Numismatic Society (ANS), the Portable Antiquities Scheme, the British Museum, and the Staatliche Museen zu Berlin. Despite its offering the purest version of Linked Open Data that we have seen, Nomisma's RDF does not provide a complete representation of the scholarly space it represents. Data from the ANS is edited in XML form using the Numismatic Description Standard (NUDS) and then transformed to RDF for ingestion into Nomisma. Not all of the data represented in a NUDS file makes its way into Nomisma's triple store. The site plus its associated datasets thus serve as a kind of distributed database of coinage information.

All of the entities modeled by Nomisma are dealt with as Simple Knowledge Organization System (SKOS) Concepts,[10] meaning that they are essentially treated as subjects in a taxonomy. SKOS makes available several useful properties for relating Concepts to other entities. So the Nomisma identifier http://nomisma.org/id/ephesus represents the "idea" of the mint at Ephesus and http://nomisma.org/id/ephesus#this represents the "spatial location" Ephesus (which has, e.g. geocoordinates). Information about the provenance of this data is attached to the URI http://nomisma.org/id/ephesus#provenance. The Nomisma interface surfaces a list of links to the first 100 coins related to an entity from partner projects, with the opportunity to download the full set as CVS or to view and modify the SPARQL query that produced the list.

[9] Kansa, personal communication, 2018.

[10] SKOS develops "specifications and standards to support the use of knowledge organization systems (KOS) such as thesauri, classification schemes, subject heading lists and taxonomies within the framework of the Semantic Web".

Models for sustainability

The funding models for these five Linked Data resources all vary. Pleiades is led by Tom Elliott, the Associate Director for Digital Programs at the Institute for the Study of the Ancient World (ISAW). He, and occasionally other personnel at ISAW are responsible for its ongoing maintenance, while its development cycles have been funded by grants from the NEH with support from ISAW. Papyri.info was developed under the auspices of the Integrating Digital Papyrology project (IDP), led by Joshua Sosin and funded by grants from the Andrew W. Mellon Foundation, along with some funding from the NEH for APIS. Since the completion of IDP, Duke University Libraries has supported its ongoing development and maintenance. The Duke Collaboratory for Classics Computing (DC3) is the group responsible for technical maintenance and upgrades. Trismegistos is supported by Mark Depauw's position as a faculty member at Leiden, and Mark has been successful in obtaining funding from various sources to support its ongoing development. Open Context was begun and continues to be developed by Eric and Sarah Kansa, with its funding dependent on grants and consulting work. It recently received an NEH Challenge Grant, with which Open Context hopes to put its funding on more stable ground. Nomisma is a project of the American Numismatic Society (ANS). It was begun in 2010 by Andrew Meadows and Sebastian Heath. Ethan Gruber took over as lead developer in 2012 and has continued in that position since.

None of the sites employ what might be called a "lightweight" digital infrastructure. All use backend databases of different types. Papyri.info and Nomisma both use Apache Jena and Fuseki, a Java-based RDF triple store. Papyri.info and Open Context use Apache Solr, a Java-based search engine. Papyri.info and Trismegistos both employ MySQL as a database, Open Context uses PostgreSQL, and Pleiades the Zope Object Database. Most of them have a dynamic front-end, where pages are assembled upon request from data in the database. Without taking a deep dive into the technologies involved, we can still say with confidence that all of the resources under discussion have both infrastructural and maintenance requirements that demand a significant allocation of server storage, memory, and CPU to host them. Moreover, they are of sufficient complexity and scale that experienced people are needed to maintain them. If we were to place them in Vinopal and McCormick's model for levels of support in Digital Scholarship Services, they would all be at the highest tier (4, Applied R&D), and deployed at tier 3 (Enhanced Research Services).[11] None

[11] See Vinopal (2013, 32, fig. 1).

of them could be simply moved into the care of, e.g. a typical university research library without additional funding (probably including additional staff) for their maintenance.

Most of the sites under discussion mitigate the risks involved in running a resource-intensive service by publishing their data in static forms and at multiple venues. Pleiades exports its data daily in a variety of formats, including JSON, KML, CSV, and RDF. Papyri.info exposes its RDF and TEI XML data alongside its web pages, and also provides a public repository on GitHub containing all of its source data. Nomisma provides downloads of its data in JSON-LD, Turtle, and RDF/XML. Open Context permits the download of project data or search results in tabular (CSV) or Geo-JSON form. Only Trismegistos does not currently provide a data export feature, but it does share its data with Papyri.info in the form of periodic database dumps. Papyri.info's data in particular provide a salutary lesson in the value of static data exports. Both the DDbDP and APIS data contained by the site were converted from older forms from previous projects. The DDbDP data is on its third iteration, having begun life as Beta Code,[12] created for the PHI CD-ROMs, then converted to TEI SGML + Beta Code for ingestion into the Perseus Project, and finally to EpiDoc XML and Unicode for import into Papyri.info. The open formats used by PHI and Perseus made these migrations an achievable, if not always simple exercise.

To varying degrees, all of these resources rely on the involvement and commitment of particular individuals. Pleiades would not exist without Tom Elliott, nor Trismegistos without Mark Depauw, nor Open Context without Eric Kansa. Were they to cease being involved, the futures of these projects might be in doubt. Pleiades is less vulnerable, as it has an institutional home at ISAW, which one hopes would decide to continue it without him. Papyri.info certainly would not exist in its current form without the director of DC3, Joshua Sosin, and its major components owe their architecture to and are still maintained by Ryan Baumann and myself, but it would likely survive the departure of any of its key personnel. It would take a withdrawal of support by its home institution to threaten it. Although Nomisma as it exists is largely the creation of Ethan Gruber, the ANS supports it, and so it would also be likely to continue if Ethan departed. All of the services under discussion have been significantly shaped by their developers, and many of these developers have been present since the inception of the project.

[12] (TLG 2016).

Of course, reliance on individual contributors is a double-edged sword: they are hard to replace, and there is some increased risk because of their importance to the project. On the other hand, maintenance costs may be cheaper because the people with the most intimate knowledge of the services are the ones who run them. These costs might go up significantly if service maintenance were handed off to less-expert teams and the continuance of the projects themselves might be at risk. The institutions which support these projects have chosen to do so by supporting individual developers in ways that bear more similarity to faculty than technical staff. Duke University Libraries created a new Digital Classics research unit, DC3, and hired Baumann and myself to staff it. ISAW has its own Digital Programs department which Tom Elliott heads. The Curatorial Department at the ANS employs Ethan Gruber as their Director of Data Science. All of us publish, and present at conferences both in our home fields and in Digital Humanities venues, with the support of our institutions. All of us are involved in initiatives that reach well beyond the walls of those institutions.

For institutions that wish to support "Tier 4" type projects, it may be beneficial to have the ability to hire project personnel in association with those projects. Acquiring successful or promising projects along with their personnel may be a better way to grow an institution's digital portfolio than attempting to grow it from scratch. The creation of DC3 certainly followed this model. Despite being the institutional leader of the Integrating Digital Papyrology grant that produced Papyri.info, Duke University was not able to field the personnel to actually develop it. The work was contracted out to King's College London, NYU, and the University of Kentucky Center for Visualization & Virtual Environments. At the conclusion of the grant, Duke University Libraries established DC3 to maintain and continue the project, and the Papyri.info site was transferred there from NYU in 2013. Pleiades similarly followed Tom Elliott to ISAW in 2008, and one might wonder whether Open Context might achieve long-term support via a similar route.

Another important aspect of sustainability that all of these projects exemplify is community engagement. Nomisma and Papyri.info have made themselves indispensable tools for the small scholarly communities they represent (Numismatics and Papyrology). Pleiades, Trismegistos, and Open Context all have a larger purview, but they too have made themselves indispensable to the point where, if they ceased to exist, something would have to be created to replace them.

Linked Data and complexity

We have so far spent some time discussing the five projects' relationship to RDF and Semantic Web technologies without relating them to the definitions of Linked Data and its relationship to RDF. RDF works by encoding data as triples, in the form Subject, Predicate, Object, where the Subject and Predicate parts of each statement are URIs, and the Object is either a URI or a string (a "literal"). Modern triple stores further refine this scheme by adding a Graph URI, making each statement a quad. RDF data can be queried using the SPARQL query language, so once data has been structured as RDF, there is a ready-made way to extract information from it, or even to generate new information from existing statements. This makes for a powerful tool for scholarly inquiry, provided sufficient information has been encoded as RDF. Since statements can be linked (e.g. the Subject of one statement may be the Object of another), the information in a triple store may be said to form a graph. The foundation of Linked Data is the use of real, dereferenceable web URLs in RDF data sets, meaning that links to web resources are embedded in the semantic graph.

RDF is hard to criticize as a data format, because it is technically able to represent almost any more-complex data structure. But certain data formats have properties and affordances that may make them easier to work with and more suitable for representing certain types of data. XML and JSON Arrays, for example, both have intrinsic order, which RDF lacks.[13] In order to represent ordered data in RDF, it is typically necessary either to emulate a Linked List or to use a custom ontology for the purpose. RDF also has a hard time with qualified relationships. Recording the circumstances under which an assertion was made, for example, which would mean attaching extra metadata to a triple, requires rather extensive workarounds. All of this means that, while RDF can be devised that would represent something like a Text Encoding Initiative (TEI) XML document, the actual implementation might not provide any benefits over the original document beyond the ability to query it with SPARQL, and would be considerably harder to edit or even display in a usable fashion. Because RDF atomizes any data it represents into triples or quads, presenting or editing it means (re)assembling those atomic facts into a larger structure, in the correct order. Because it is a graph, the "records" therein are unbounded (i.e. the connections between

[13] RDF does have a built in Seq container type, which defines an order to its members based on their property names, but this order must be imposed by a client reading the RDF, which is itself an un-ordered set of triples (or quads). RDF Lists are analogous to lists in various programming languages, e.g. LISP. The first item has a property linking to the content (the first) and a property linking to the next node in the list (the rest).

pieces of data may extend to any length in any "direction"), so technology has to be applied to retrieving only the sensible pieces of data for the intended purpose.

One might consider TEI documents to be an edge case where RDF is an unsatisfactory representation, but in fact the data modeling around any scholarly project is likely to be esoteric. This sets up an inherent tension, as the explicit goal of LOD is interoperability. Even when (apparently) well-defined standards are adopted for the description of a project's data model, the local interpretation of those standards and the "gray areas" they inevitably contain will make the definition of mappings between datasets a necessary precondition for interoperation. If the goal of LOD is the same as the Semantic Web, where purely machine-mediated domain exploration is possible, then it is only likely to be achievable in cases where the semantics of the data are lightweight.

The TEI has struggled over the years with questions of interoperability, for precisely the same reasons.[14] Data modeling is an interpretive act, and because of that, the more complex and extensive it is, the more individualized it necessarily becomes. It follows that there is an inverse relationship between comprehensiveness and interoperability. Since the latter is the entire goal of LOD, concentrating on simplicity in the Linked Data one exposes would seem to be a better investment than working on fully encoding one's data in a semantic format. Recent developments, notably the introduction of the JSON-LD format, would seem to represent a turn towards such simplicity. JSON-LD is the basis for Linked.art, for example, which aims to develop a more usable profile of CIDOC-CRM, one of the more complex cultural heritage RDF vocabularies. Linked.art's analysis of CIDOC-CRM classes provides an interesting insight into the ways in which attempts to be comprehensive may result in unhelpful complexity or even failure to fulfill an obvious need. For example, the discussion of E30 Right, states:

> The basic problem with E30 Right is that it is a Conceptual Object, and Conceptual Objects cannot be destroyed. While there is any carrier of the object, including the CIDOC-CRM description of it or even within someone's memory, then the concept still exists somewhere. As it cannot be written down without persisting it, it cannot be destroyed and instead it can simply pass out of all knowledge. This means that the existence of the Right is not the same as the validity of the Right: the concept of slavery in America still exists, but it is no longer legally valid. There are no terms within the CRM to express the effective dates, and the CRM-SIG clarified that the right's effectiveness would be a different sort of resource. In particular that an E30 Right "is the formulation of the right, the terms", and not whether the right had any legal standing in any jurisdiction at any point in time.[15]

[14] (Bauman 2011).

[15] From https://linked.art/model/profile/class_analysis.html#ineffective-classes (last access 2019.01.31).

That a reasonable design decision might make it hard to do something practical, like express rights that are limited in time or space doesn't invalidate the whole enterprise by any means, but it is a signal that efforts to be complete and correct in a specification may come at the expense of usability. One should not need a Ph.D. in the philosophy of law to implement a small part of a data model.

Any sufficiently expressive data model runs the risk of provoking what we might term "over-encoding" by analogy to the idea of overengineering in software development. Specifications (like CIDOC-CRM or TEI) have a tendency to address problems that don't exist yet, but plausibly might, in their quest for completeness. Users of those specifications, especially new users, may tend to encode information without thinking about whether doing so provides any benefit, responding to a theoretical imperative rather than a real-world need. Doing so may, like overengineering, incur little immediate obvious harm but may also divert resources that might be used elsewhere and make processing and interoperability more complicated, thus having a net negative effect on project usability and sustainability.[16]

Simplicity is the hallmark of one of the more successful efforts at building a cultural heritage LOD network, Pelagios,[17] which aggregates data around places published by a variety of projects. Pleiades serves as the "hub" for these datasets, which use Open Annotation (OA) RDF to associate Pleiades place URIs with whatever information the project publishes. OA merely associates the annotation body with the URI being annotated (the target) without necessarily doing anything to characterize the nature of the link. Pelagios aggregates annotation datasets published by partner projects and provides tooling to research these. Pleiades, meanwhile, can use Pelagios's API to query what projects are referring to a particular Pleiades place. This means there is a straightforward way for pages in Pleiades to provide links out to associated material via Pelagios without having to maintain those linkages itself.

In this way, on a basic and practical level, the publication of stable resources and linkages with some (even if weak) semantics promises to be a huge boon for discoverability. This is likely to matter much more in the long run than whether a particular piece of data is in a particular format because it answers a basic scholarly need: "Can I find a piece of information and get from it to potentially useful related information?" Search engines use links for purposes of

[16] Sporny's (2014) discussion of the relationship between JSON-LD and the Semantic Web refers to the tendency of Semantic Web specification developers to focus on the wrong things. "Too much time is spent assuming a future that's not going to unfold in the way that we expect it to".

[17] (Simon 2014).

discovery and ranking and HTML links in the browser are only weakly (if at all) characterized. Google and its competitors employ machine learning algorithms to rank their search results with a great deal of success. The real strength of LOD may then be its architectural style, which by insisting on resolvable URLs for identifiers, exposes the components of a data set and the links between them to the web instead of hiding them behind a query interface.[18]

The five LOD projects under discussion all check at least some of the boxes in Berners-Lee's 5-star scheme, and all identify the important entities in their datasets using resolvable URIs and link to related data, both internally and externally. Most of them, however, put RDF somewhat at arm's length, using it as only one of several export formats (Pleaides, Nomisma) or structuring their data as nodes in a graph without attempting to encode the data using RDF (Papyri.info, Open Context, Trismegistos). Only Nomisma fully embraces RDF as a first-class data structure, and notably, it is only part of a broader infrastructure, the external nodes in which do not encode their data directly in RDF. Arguably, it performs, in a distributed way, the same function as the "Numbers Server" in Papyri.info. As we have seen, all of these are complex projects, requiring expert maintenance and support. It is notable that none of them, with the possible exception of Nomisma, embrace the Semantic Web interpretation of LOD.

Conclusion

Having explored some of the more successful Linked Ancient World Data systems and the ecosystems around them, we can summarize the characteristics that have enabled these projects to continue for years, well past the startup phase. All of them have provided long-term support for key personnel. None of them have attempted to build a resource and then hand it off to some other entity to maintain. All of them either have institutional or other long-term support, or are actively working on developing a support framework. All of them have become an indispensable resource for their communities, so that support or pressure might be brought to bear should they become threatened. All of them have embraced LOD as a means to connect their data to the wider digital cultural heritage infrastructure, but have at the same time avoided the complexity of attempting to represent their full range of data as RDF.

[18] Cf. Ogbuji (2016) on the beneficial effects for visibility on the web of recasting public library catalogs as Linked Data.

If we can attempt to derive a recipe for long-term success in cultural heritage LOD from the examples in this essay then, we might say the following:

1. Involve and provide long-term support for technical specialists who also have content expertise and interest if possible.
2. Obtain Institutional commitments to ensure #1.
3. Prioritize focus on the needs of the community or audience and the practicalities of meeting those needs over following rubrics for LOD.
4. Expose or export data in reusable formats as both a means of attracting partners and as a hedge against disaster.
5. Intentionally engage partner projects and share data with them to ensure that links endure.

It should surprise no one that there are no "silver bullets" here. LOD opens up many interesting possibilities for cross-project data reuse and for building a true ecosystem of online cultural heritage resources, but the technology does not obviate the need for human collaboration and community engagement to make these possibilities real.

Bibliography

Bauman, S. (2001): "Interchange vs. Interoperability". In: Proceedings of Balisage: The Markup Conference 2011. Balisage Series on Markup Technologies. Volume 7. Mulberry Technologies, Inc. https://doi.org/10.4242/BalisageVol7.Bauman01.

Berners-Lee, T. (2006): Linked Data. https://www.w3.org/DesignIssues/LinkedData.html (last access 2019.01.31).

Berners-Lee, T.; Hendler, J.; Lassila, O. (2001): "The Semantic Web". Scientific American, May 2001, 29–37.

Elliott, T.; Heath, S.; Muccigrosso, J. (eds.) (2014): "Current Practice in Linked Open Data for the Ancient World". ISAW Papers 7. http://doi.org/2333.1/gxd256w7.

Gruber, E. (2018): "Linked Open Data for Numismatic Library, Archive and Museum Integration". In: M. Matsumoto; E. Uleberg (eds.): CAA2016: Oceans of Data. Proceedings of the 44th Conference on Computer Applications and Quantitative Methods in Archaeology. Oxford: Archaeopress, 35–40.

Linked.art. https://linked.art/index.html (last access 2019.01.31).

Ogbuji, U.; Baker, M. (2015): "Data Transforms, Patterns and Profiles for 21st century Cultural Heritage". In: Proceedings of the Symposium on Cultural Heritage Markup. Balisage Series on Markup Technologies. Mulberry Technologies, Inc. Volume 16. https://doi.org/10.4242/BalisageVol16.Ogbuji01.

Simon, R.; Barker, E.; de Soto, P.; Isaksen, L. (2014): "Pelagios". ISAW Papers 7. http://doi.org/2333.1/gxd256w7.

Sporny, M. (2014): "JSON-LD and Why I Hate the Semantic Web". http://manu.sporny.org/2014/json-ld-origins-2/ (last access 2019.01.31).

Talbert, R.J.A. (2000): Barrington Atlas of the Greek and Roman World. Princeton, NJ: Princeton University Press.

Thesaurus Linguae Graecae (TLG) (2016): The TLG® Beta Code Manual. http://www.tlg.uci.edu/encoding/BCM.pdf (last access 2019.01.31).

Vinopal, J.; McCormick, M. (2013): “Supporting Digital Scholarship in Research Libraries: Scalability and Sustainability”. Journal of Library Administration 53, 27–42.

World Wide Web Consortium: Linked Data. https://www.w3.org/standards/semanticweb/data (last access 2019.01.31).