From Linking Places to a Linked Pasts Network

Karl Grossner and Timothy Hill (*Linked Pasts Working Group co-coordinators*)
With contributions from Rainer Simon, Richard Light, Arno Bosse, Gabriel Bodard, Mia Ridge, and Wolfgang Schmidle

1. Introduction

This white paper presents results of a 2017 initiative undertaken by the Linked Pasts Working Group (LPWG) of Pelagios Commons. First, a compilation of requirements as expressed by our community of interest for what the Pelagios investigative team has called “a wider ecosystem of projects dedicated to interlinking online resources about the past.” Secondly, a draft road map to aid discussion about steps the Pelagios project and broader community can take towards fulfilling them.

The initial focus of Pelagios has been on linked open *geo*data: references to place found in early maps and texts, particularly for ancient periods in the Mediterranean region. Toward that end, the project has been instrumental in promoting a community of interest (Pelagios Commons) and has developed tools, systems, and best practices to assist in developing such data (Recogito), publishing it, linking it, and making the results accessible both programmatically (Pelagios API) and in a web application for search and spatial-temporal visualization (Peripleo)[1].

Pelagios' successes have inspired discussion about how its community, systems, models, and software might be extended and made sustainable beyond its immediate funding horizon. Two kinds of extensions to systems and models are contemplated: in spatial-temporal coverage, and in the types of data that are linked. In fact, experimental work has begun on both and will be reflected in the second version of Peripleo software, due in December 2017. This LPWG initiative begins a coordinated effort to outline further pragmatic next steps with some specificity, derived from realistic use cases and the community’s collective technical expertise.

2. Requirements

(See also, *Appendix A - User Stories*)

In July, 2017 eight LPWG members gathered for a 2-day working meeting in London, to brainstorm how we might extend coverage of Pelagios data-modelling capacities beyond geographical entities, and to begin developing what was to become the content of this paper: a set of requirements and next-step recommendations for an envisioned Linked Pasts "ecosystem." Several institutional and research domain perspectives were represented, including ancient prosopography, Early Modern epistolary networks, museum collections and aggregators thereof, archaeology, genealogy, temporal representation, and historical

[1] Peripleo 2 is due to be launched in December, 2017

gazetteers. The meeting was followed by email discussion and conference calls over the next several months.

We began by enumerating requirements drawn from our respective research domains, then distilling them into a small set of generic "user stories"—statements in the form, "as a {*user type*} I want {*some goal*} so that {*some purpose*}" (Appendix A, §1). It was universally agreed that persons (agents more broadly) were, after places, the most important historical entity type to be able to model. Others include time periods, artefacts (e.g. archaeological finds and museum objects), works (e.g. texts, art objects), events, and concepts. The July meeting included a "hack day" spent experimentally modeling people data for the Peripleo index and search interface. In follow-up discussions we identified "interconnection formats" for the various entity types as a core system requirement and those are taken up below in §4, "A Roadmap and Next Steps."

Community

Although the focus of the London meeting was on *technical* requirements, it became increasingly clear they are closely interwoven with issues of community and sustainability. This was brought home by the results of a pre-conference workshop, *Advancing Linked Open Data in the Humanities* at DH2017 in Montréal the following month. Participants were asked to describe in submitted position papers "gaps or opportunities with respect to Linked Open Data for the humanities". From those, workshop organizers derived a set of "pitches," which were delivered in the meeting session. We have taken the further step of distilling that compilation of needs and wishes into a list of requirements expressed as user stories. The authors' original wording has been lightly edited to fit the story format, and stories have been grouped thematically, rather than by paper (see Appendix A, §2).

The Pelagios experience and a close examination of user stories indicate strongly that Linked Pasts development requires a simultaneous bootstrapping at the technical and community levels. Both aspects are sufficiently complex to require the support of an institutional sponsor. We offer the following further observations:

- A Linked Pasts ecosystem will gain value as it grows—the more high-value datasets linked, the better. The value of LOD in historical research must be demonstrated well in order to promote contributions.
- Linked Open Data is a relatively new paradigm and set of methods, which will not be widely adopted without high quality instructional resources, including publicly shared experiences of the community.
- The aphorism, "software comes and goes but data is forever" is demonstrably true, but it is also the case that data without software to access, display and analyze it is not especially useful.
- Models and systems developed to work with a few exemplars will be continually tested by the arrival of new datasets. New issues, challenges, requirements, and wishes will surface continually. For these reasons, an engaged open-source software development community will be essential and will benefit by institutional support.

- Without a sustainability strategy, systems development might ultimately be wasted effort.
- Achieving widespread participation will depend in part on active critical discussion of the promise and pitfalls of sharing data in this way, including relevant ethical, philosophical, and societal considerations.

3. A shared vision of linking pasts

The users stories gathered in Appendix A reflect what many humanities researchers from a number of disciplines want from a linked data ecosystem. We have summarized them here as follows. Humanities researchers want to…

- Acquire and share data related to named historical entities as subjects of academic research, including places, people and groups, cultural objects, periods, and events
- Reference and/or link their data to existing authority records for named historical entities
- Annotate resources published by others, and track the provenance of all such annotations
- Use this emerging historical knowledge graph to track and study concepts reflecting dimensions of cultural practice, politics, and historical processes
- Locate and learn about resources and best practices for acquiring and publishing linked data, including
    - What systems, tools, online tutorials, training events, and workshops are available and well-regarded by the community
    - Data modeling theory and practice; ontology engineering
    - Developing and publishing specialized ontologies and authorities
    - Minting stable URIs
- Develop, and encourage development of, useful and usable interfaces to data
    - Web applications for searching, browsing, annotating, and downloading data from distributed sources
    - SPARQL-free access and/or a helpful functional layer above it[2]
    - Robust and intelligible APIs with content negotiation[3]
    - Reconciliation services for disambiguating and matching records of named historical entities
- Discuss the practice and implications of such digitally-aided research, including
    - Intersectionality, diversity and inclusion, and ethics

4. A roadmap and next steps

The successful efforts of Pelagios and its key partners (e.g. the Pleiades and PeriodO projects) offer models of practice as start points for designing and realizing a broader Linked Pasts ecosystem. In our view, the high-level steps before us are a) further extending an aspirational

[2] SPARQL is the recommended query language for accessing RDF data stores, but has a deserved reputation for being difficult to use

[3] "Content negotiation" refers to the ability to specify in a query the desired format of returned data

Linked Pasts architecture informed by community input, and b) developing a pragmatic strategy for funding and sustaining those activities and products.

Towards a Linked Pasts architecture

Feedback

This document presents a summary of user requirements, a few high-level questions and issues, and several specific recommendations. We actively encourage response to all of these from members of Pelagios Commons and beyond, at any level of detail.

Follow-up specialist meeting

Robust feedback to this document should be followed by a specialist working meeting in the near future to turn an expressed vision into a plan.

The products of that meeting would be (i) consensus on Linked Pasts system architecture; (ii) explicit strategies for building and sustaining software, systems, and the community whose research they support; (iii) a list of prospective institutional partners and supporters, or if a single consolidated Linked Pasts effort is ultimately deemed unrealistic; and (iv) a roadmap for further ad hoc "ground up" activity to achieve Linked Pasts goals. An open Request For Comments on a published report would follow.

Pre-work for its participants would include considering the following questions, as well as the related recommendations listed later in this section.

- Rainer Simon, Technical Director of Pelagios, recently said, "It doesn't make sense to build a system that just handles place, when people keep expecting to search/filter by everything else in one UI. At least that's my lesson learned from prototyping over the past years." However, a single interface to all linked historical data for all places and periods seems unrealistic. Is there in fact a best way to slice historical data?
- Pelagios, along with its data partners and other cognate projects, have demonstrated considerable progress on systems addressing the needs a community of interest defined by place and period—in this case, the Classical Mediterranean—and spanning multiple academic fields: history, archaeology, philology, historical geography. Should the Humanities LOD community encourage multiple similar systems for different place/period contexts[4]?
- Are there other, equally useful frames: by entity type: places, periods, persons and so on? By research community theme at various scales (e.g. history of science or South American literature)? Some other combination? Will a successful Linked Pasts architecture need to accommodate all of these? Can most or all variations be anticipated?

[4] The Cultures of Knowledge project at the University of Oxford is preparing this for the early modern period with EM People, EM Places, and EM Dates

- More concretely, should Pelagios seek to grow in scope, indefinitely and in all directions—linking data about places, people, cultural objects, events etc. for an ever-expanding spatial-temporal extent in a single set of indexes and software exposing them? If not, how should it be constrained?
- Assuming Pelagios won't tackle everything, what should be its relationship to other, similar, enterprises? For instance, the Data for History project appears to have goals quite close to those of Pelagios, as does the World-Historical Gazetteer project. How should Pelagios/Linked Pasts attempt to align with these, if at all?

Interim Recommendations

Given that Pelagios' successes have come about incrementally, driven "ground up" by a community of interest with an expressed need, a further expansion of the Pelagios linked data graph to include entities other than places, e.g. persons, groups, cultural objects, and events should also proceed incrementally, driven by the requirements of active partners. In fact, that expansion has already begun on an experimental basis.

Pelagios is a place-centered project and should remain such. It aggregates (conflates) multiple gazetteer records about given places, and gathers annotations concerning relations of places to contributors' records of non-place entities. The result is a set of "union indexes" that return rich responses to queries for places in Peripleo, or via the Pelagios API. Those responses could be made richer by extending and reconfiguring the existing standard formats for contributed annotations (see ***Enhanced Annotation*** below). Such enhancements would aid development of a reconciliation service for named places (see *Appendix B - Reconciliation*).

The requirements listed in Appendix A and summarized above suggest our community would like to see the sort of capabilities Pelagios provides for places made available for data about other named entities, especially people, events, periods, and cultural objects. In addition to place-centric "hubs" like Pelagios (and soon, the University of Pittsburgh's World-Historical Gazetteer), there can and should be other classes of hubs, all of them linked. In fact, several projects with potential to fulfill those requirements exist or have been proposed[5]. For further discussion along those lines see ***A Linked Pasts Network*** below.

Enhanced Annotation

As presently configured, the Pelagios platform accepts and indexes contributions of a) place data from gazetteers (e.g. Pleiades) and individual research projects (e.g. Nomisma, FASTI), and b) annotation records linking entities tracked by research projects (e.g. coins, hoards, archaeological sites, persons) with places. There is no practical limit to the kinds of entities (termed "Items") that can be annotated with places; the system as presently constituted has evolved according to the interests of motivated data partners.

[5] Cultures of Knowledge and the Huygens submitted an EU H2020 proposal, 'CommonPlace' on these lines earlier this year.

We propose that the current formats for contributed places and annotations be experimentally unified to enable richer descriptions of typed Items and make it possible/easier for contributors to publish mixed datasets.

The Pelagios Gazetteer Interconnection Format (PGIF) was developed particularly for data about places and attestations of place names. The Pelagios RDF standard for contributed annotations, which we will refer to here as the Pelagios Item Annotation Format (PIAF), is derived from the Open Annotation standard (OA). Currently, a typical contributed annotation file in PIAF format includes a minimal record for each Item (pelagios:AnnotatedThing), and any number of annotation records (oa:Annotation) referencing those Items and external records in authority gazetteers.

A new, unified format would incorporate:

1. Items
    a. Require that Items be typed (e.g. person, artefact, event, etc.)
    b. Encourage contributors to add Item attributes (currently only "Title" and "HomePage" (uri) are mandatory).
    c. Develop a list of core type-specific Item attributes. These could potentially serve as facets in interfaces; any additional attributes included could be returned in API responses.
2. Annotations
    a. Specify relatively small sets of allowed type-specific relations between Items and Places (pelagios:relation). For example, Items of type Person might have a *hasBirthplace* relation to a place.
3. Events
    a. The system should permit contributions of named historical events (e.g. battles, treaties, social movements, natural catastrophes, journeys of exploration), considerations for which resemble other named entity Item types.
    b. The system should also accommodate data modeled as events having participant entities, in addition to the more usual flattened or shorthand approach. For example, the fact that PlaceA was the birthplace of Jane Doe could be derived from an Event Item (titled "birth of Jane Doe" having participants, date, and a locale) or as a Person Item with attribute "Birthplace."
    c. See *More about events* below.

Vocabularies for new Item types will ideally incorporate and/or build out from the Linked Ancient World Data (LAWD) ontology. The effect of these enhancements will be to make Peripleo and similar place-centered systems increasingly like the *enhanced descriptive gazetteers* alluded to in several chapters of the recent edited volume, "Linking Places" (Berman, Mostern & Southall 2016). That is, given a place name, they will a) offer a reconciliation service as required, and b) provide a summary description of the place as attested in historical sources, including links to

web-accessible resources from which one can build a rich description--navigating the emerging scholarly historical "knowledge graph[6]" referenced earlier.

A Linked Pasts Network

In addition to these enhancements to Peripleo, we recommend that members of the Pelagios Commons, along with other interested individuals and organizations, seek to establish a **Linked Pasts Network**. As it stands, the Pelagios group and Pitt's World History Center are coordinating their development plans for the Peripleo and World-Historical Gazetteer (WHG) projects respectively, effectively as distinctive place-centered *Linked Pasts Hubs*. The Standards for Networking Ancient Prosopographies (SNAP) project has aimed to provide Peripleo-like services for persons and though currently suspended, has made considerable progress. If its development were renewed it could become the first person-centered Linked Pasts Hub. The PeriodO project is effectively *the* Linked Pasts Hub for named periods.

In this framework, a *Linked Pasts Hub* is defined as a system that

- aggregates and indexes records for a particular class of named historical entity, such as Place, Person, Period, or Event
- provides access to merged index records via an API, preferably including a reconciliation service. (See *Appendix B*)
- coordinates with like hubs the development of interconnection formats and vocabularies specific to their focus

Linked Pasts Hubs, by linking the growing number of Nodes as defined below, will be resources in their own right, providing key services such as reconciliation and visualization. Ideally, annotations contributed to a Linked Pasts Hub will make their way back to the source for the annotated record (to date, an unaddressed challenge). Linked Pasts Hubs should make no demands on contributors for how they organize their data internally; rather, they should get community buy-in for the abbreviated interconnection formats and vocabularies to be used by their contributors.

A *Linked Pasts Node* is defined as a project/system that in some combination

- publishes Linked Open Data relevant to historical research such that Linked Pasts Hubs can index it, AND/OR
- consumes Hub data and/or uses Hub links to acquire data, and/or replicates Hub links

Large research projects like Syriaca.org and Trismegistos could be significant contributing and consuming nodes in the network. Major publishers of LOD, such as the Getty Research Institute and Library of Congress could also become key contributors.

Agreement on a division of labor amongst Linked Pasts Network members should produce efficiencies that benefit all.

[6] This term has been appropriated by Google, Inc. but was in common use by AI and semantic web researchers previously

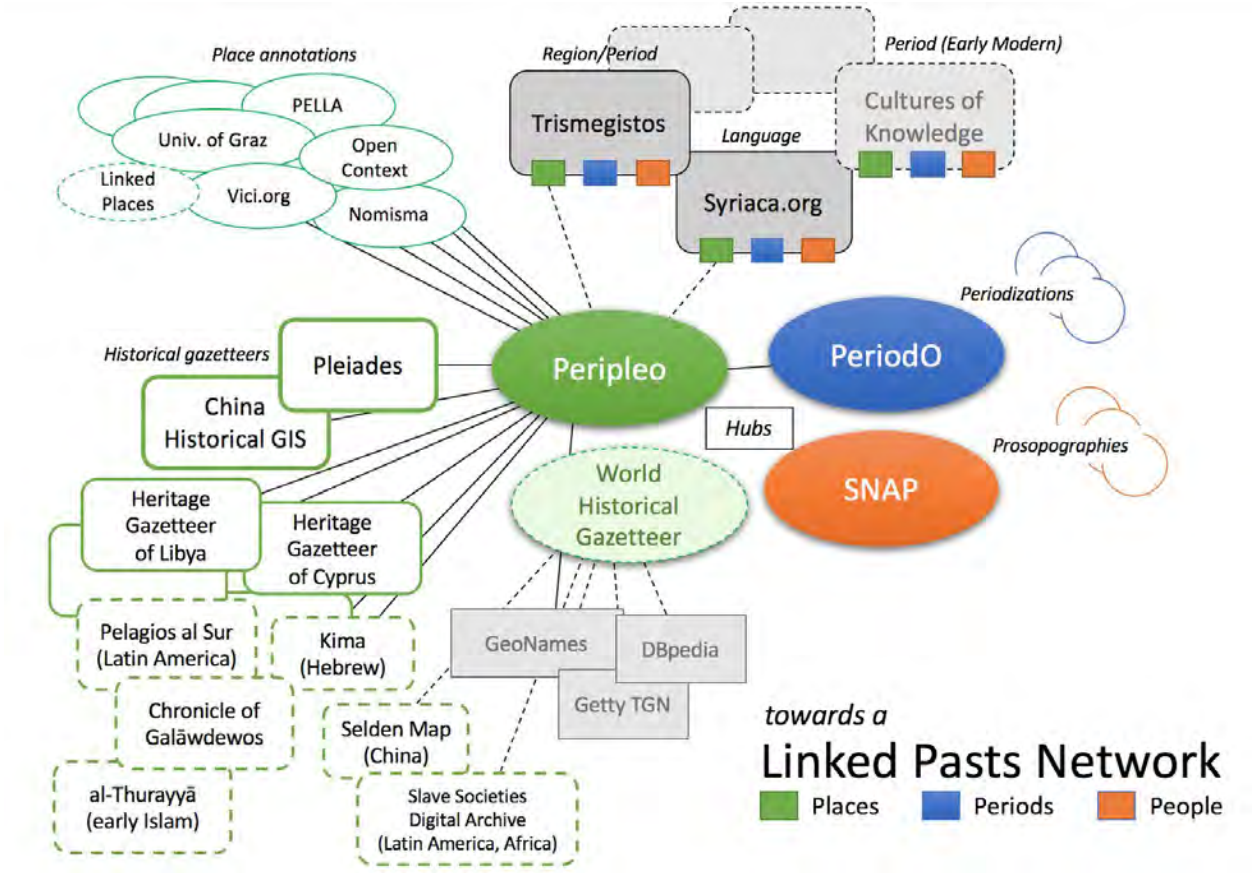

Fig 1 *Partial view of the emerging Linked Pasts Network. Many nodes not shown due to space constraint*

More about events

We have stated events should be supported by a Linked Pasts Hub ingest process in two ways:

- In the sense of discrete entities (Items) having participants that were present in some role
- As inferred from attributes of another type of Item, e.g. from the birthDate and birthPlace of a Person record

Linked Pasts Hubs' interconnection formats will have to account for event types. Note that some types are generic; others a function of Item type, e.g.

- Person: Birth, Death, Marriage, Residence, Matriculation, Visit, Burial, Meeting, Correspondence, Creation, Journey[7]
- Artefact: Production, Deposit, Find, Exhibition and Acquisition/Provenance
- Work: Creation, Publication
- Named Event: Battle, Treaty, Journey/Expedition

[7] cf. Lineage-Linked GEDCOM (5.5) Tag Definitions for a range of possibilities

The two supported event item models should be normalized in a Linked Pasts Hub's software, either at the indexing step or in GUIs and APIs. The use cases driving normalization of event format include, for example:

- In a web page summarizing a Place, display of links to related people should be consistently formatted, regardless of contributors' data models
- Alternatively, an API Place index record could return whatever attributes a contributor chose to publish in their Item/Annotation dump file

Event-participation models are well-established, and range from the relatively complex CIDOC-CRM ontology to the lightweight and ubiquitous Schema.org vocabulary. Their use is at times controversial, chiefly because of the perceived complexity of such models: CIDOC-CRM in particular has attracted criticism on the grounds of its modelling overheads. However, event-based patterns allow for more comprehensive representations of historical entities and demonstrably avoid many of the most commonplace and severe data-modelling errors. While the question of how to find a workable middle-ground between excessive complexity and over-simplification is an open one (as is the question of how to screen such complexity from users where possible), the advocacy of an event-centred approach to data modeling is well-considered, and it is recommended that Linked Pasts prioritise event-related issues in the immediate future.

Sustainability

Ensuring that a Linked Pasts Network is self-sustaining over a long period is of the utmost importance. We believe this can be achieved by the combination of a motivated multi-disciplinary research community and some institutional support. Researchers who publish data to be aggregated and indexed by Linked Pasts Hubs should have an assurance of permanence. Pelagios has relied on grants that will sunset within a few years. Some possibilities for ensuring for its future and that of future Linked Pasts Hubs includes sponsorship by:

- Universities: e.g. the University of Pittsburgh's World History Center has committed to hosting and maintaining the World-Historical Gazetteer indefinitely
- An existing foundation or consortium: Examples include Europeana, Online Computer Library Center (OCLC), the U.S. Institute of Museum and Library Services (IMLS)
- A new "Linked Pasts" consortium of institutions who can plausibly commit to shared long-term support of basic infrastructure (hosting, etc), developer time, and occasional community meetings. University libraries would seem to be a natural home for this kind of activity, and we recommend pitching this to them wherever and whenever possible. Consortium members might also include large museums and national libraries, such as the British Museum and British Library[8].

[8] These British institutions are mentioned only because existing Linked Pasts community members have ties to them. Membership by others, situated elsewhere, should be explicitly sought out.

In the meantime, it would be very useful if a new, person-centered Hub initiative were to undertake an exemplar of significant size that may serve as a further proof-of-concept, and would help tell a story of LOD furthering analytical possibilities and knowledge creation.

Appendix A - User Stories

During a July meeting several Linked Pasts Working Group members developed several "generic" user stories, and subsequently added a couple dozen more derived from materials submitted to the DH 2017 workshop, "Advancing Linked Open Data in the Humanities." Both are listed below the following Summary. The generic stories suggest a particular distributed system and interfaces to it. The workshop-derived stories encompass a future shared environment for Digital Humanities research more broadly (i.e. not exclusively historical). Importantly, those stories emphasize not only software systems and tools but aspects of community.

Generic User Stories

- As {a user}, I want to {disambiguate and precisely identify name-strings for an entity} so that {I can discover content elsewhere that is relevant to or associated with it}
- As {a cataloguer} I want {to enrich my data with external authorities' identifiers} so that {my data is more precise, accurate, and useful for others}
- As {a first-time user}, I want {a sense of who (organisations, individuals) created a given linked data system and why} so that {I can assess its credibility before using it}
- As {the owner of a dataset}, I want {to contribute it to a shared historical resource} so that {it will be discoverable and may be enhanced}
- As {a curator of a dataset}, I want {a persistent record of my contributions to it} so that {I can receive credit for this work}
- As {an end-user of the shared historical resource}, I want {full provenance information attached to each statement in the resource} so that {I can evaluate it}

DH Workshop User Stories

The following list was developed from the position papers and "pitches" of the pre-conference workshop, ***Advancing Linked Open Data in the Humanities*** at DH2017 in Montréal in August, 2017. Authors' verbiage has been lightly edited to fit the story format, and stories have been grouped thematically, rather than by paper/author.
We omit for the time being specific "as a {___}" user identities, so *As a DH researcher,*

Learning/Best Practices

- I want {training resources for learning how to link my data with other data} so that {we can improve the quality of LOD, interfaces to it, and the reuse of data generally}
- I want to {learn about existing vocabularies, data standards, etc in use} so that {I can avoid duplication of effort}
- I want to {learn what LOD tools are available and being used successfully} so that {I can most efficiently give them a try; best practices are promoted and gaps highlighted}
- I want to {access good documentation on existing LOD projects} so that {I can learn best practices}

Modeling data

- I want to {develop a new integratable ontology specific to my local project} so that {I can make my dataset make linkable and worth linking from (and to)}
- I want to {version ontologies} so that {we can mitigate the problem of sustaining a huge number of URIs for similar versions of entities, classes, and predicates}
- I want to {capture the degree of certainty with which an assertion has been made} so that {we can account for those that are either probable or approximate}
- I want to {use a simple, upwardly-compatible event data standard} so that {I can more completely or efficiently integrate people, object and place datasets}
- I want to {maintain the heterogeneity of biographical data while implementing an ontology that overlaps as much as possible with other linked open data structures} so that {we can document persons without eliding the complexities and contradictions of their recorded lives}
- I want to {develop a machine- and human-readable Biblio/Print Culture Ontology focussed on people, rather than imprints} so that {we can bridge gaps between extant ontologies and vocabularies for the printing and publishing trades}

Producing Data

- I want to {create new, specialized local authorities (e.g. for Proust)} so that {they complement more broad-based ones (e.g. VIAF, BnF)}
- I want to {encourage creation of precise and consistent metadata} so that {we can understand nature of published LOD}
- I want to {use software tools to create LOD from richly encoded TEI resources} so that {I can make those resources more widely available}
- I want to {study whether every TEI tag should necessarily map to LOD} so that {someone can specify and develop useful conversion software}

Publishing Data

- I want to {share my data} so that {anyone can link to it}
- I want to {promote practices for generating stable, persistent URIs} so that {we mitigate the general fragility of URL type links.}
- I want to {use/promote Wikidata as a reliable data store for persistent, common URIs} so that {its very low threshold for the creation of entities may permit a productive space outside, or alongside, the more formal bounds of academic institutions and cultural heritage institutions}

Scholarly use and reuse: consumption, integration, refinement, enhancement

- I want to {access fairly standardized, well-documented APIs} so that {I can more easily query, harvest, mash-up, and hack stores of cultural linked data}

- I want {tools incorporating a vetted workflow for annotating, identifying, correcting, transcribing, or linking collection materials} so that {we can successfully enlist contributions from "the crowd" and/or a research domain community}
- I want to {access LOD at multiple stages of the research process} so that {I can a) think about the relationships between the archival evidence and LOD during research activity; b) imagine new publication environments and future combinations with other data}
- I want to {annotate and semantically enrich datasets (incl. imagery)} so that {we better engage our scholarly communities with our documentary collections}
- I want to {search for associated triples/objects, ordered by priority in a configurable list} so that {we allow for greater interaction with the content such as inviting contributions or sending it to tools}
- I want to {integrate data about cultural objects and people with the place data made accessible by Pelagios} so that {it is discoverable indexed by place and period}
- I want to {adopt Wikidata as a source of authority data} so that {I can dynamically integrate its rich representations of people, places, and events back into my own collections and works}
- I want to {identify the physical location of papyrological and epigraphic texts in museum and private collections} so that {I can understand their distributions and to learn where they may be accessed}
- I want to {link linguistic LOD with artifacts that are already present on the web in whatever form} so that {we can enable categorization with textual semantics; cross-relate artifacts with others and with people; prepare for applying data science algorithms to increase knowledge gain}

Useful and usable interfaces

- I want to {promote development of useful interfaces to LOD} so that {we ensure sustainability of LOD resources by facilitating access for reuse of their material}
- I want to {enhance existing LOD resource usability} so that {its reach is extended}
- I want {tools for exploring and visualizing linked data (incl. smarter reasoners) that don't require learning SPARQL} so that {I can find connections across degrees of separation; drive wider adoption of these technologies}
- I want a {natural language sparql query interface} so that {I can more easily access data}

Community

- I want to {join a big community of LOD users} so that {I can share common difficulties and possible solutions}
- I want to {promote collaborative relationships between researchers and developers} so that {there are lower technical barriers for teams and improve technical quality}
- I want to {assemble requirements for technical infrastructure useful to the historical LOD scholarly community} so that {we can promote development of fundable, widely useful, and pragmatic plans for system and tool development}

Reflexive practice

- I want to {cultivate an intersectional approach to building technologies for curating, preserving, and linking the objects and records of cultural history} so that {we are advancing an approach to data curation that considers context and history as first principles}
- I want to {encourage discussion of questions of diversity and inclusion} so that {we better address issues of preservation, persistence and control impact for traditionally underrepresented communities}
- I want to {encourage discussion of the ethics of LOD (e.g. giving back; acknowledging contributors)} so that {we might reframe what “collaborative” research means in the humanities}
- I want to {integrate the cultural critical lenses developed by communities} so that {LOD creators may take an intersectional, decolonizing ethics as the basis of its future development in general}

Appendix B - Entity reconciliation service

This document outlines the functionality of a service to take a ‘seed’ RDF description of an entity and match it against one or more authorities. This service offers two modes: ‘search’ and ‘add’:

- Search mode simply returns information about one or more existing entities which match the ‘seed’ description
- Add mode creates and returns a new authority record based on the ‘seed’ record, assigning it a unique, persistent identity

Interconnection formats

In Linked Pasts we define Interconnection Formats for the key entities about which we want to share information. Initially, these entities are People, Places and Events. We encourage all developers of authorities to include Interconnection Format data where it is known, and then to provide some form of indexed access to their resources via this Interconnection data.

The default expression of Interconnection data will be the RDF patterns which we define, and the default mechanism for searching authorities will be via a SPARQL query. However, neither of these is essential, so long as each authority has an efficient HTTP-based search facility which returns machine-processible results.

Seed description

Each ‘seed’ description must include an rdf:type statement whose object is the URI of a supported entity type (“person”, “place” or “event”). It must also include at least one statement which matches the Linked Pasts Interconnection Format for that entity type. These Interconnection Format statements (and only these statements) will be used to match the ‘seed’ record to existing authority records.

In Add mode, the ‘seed’ description can contain an arbitrary number of RDF statements (of arbitrary complexity) in addition to the Interconnection Format statements. These will be retained ‘as found’ in the resulting Linked Data resource. In Search mode any additional statements will simply be ignored.

Authority metadata

The service makes reference to an explicitly defined set of external authorities. Metadata for each authority specifies:

- The entity type(s) about which it contains data
- A base URL for searching the authority
- Optionally, for each entity type, the mapping from each Interconnection Format concept to a search syntax
- Optionally, a mapping to convert search results to RDF (e.g. XSLT to process an XML response)

By default, it is assumed that the authority contains data in the relevant Interconnection Format, and that the base URL specifies a SPARQL end-point.

Search options

There are two options that control the nature of the search:

- Precision: by default the search looks for an exact match on each ‘seed’ concept that is specified. Alternatively, the user can request a ‘fuzzy’ search
- Recall: by default the only first matching entity is returned. Alternatively, the service can be asked to return a list of all matching entities. Thirdly, the service can merge all the data from all matching entries into a single graph

Add options

When a new entry is added, it will contain skos:exactMatch links to all entities which are returned by an ‘exact match’ search. If precision is set to ‘fuzzy’, it will additionally contain skos:closeMatch links to those entities which are returned by a ‘fuzzy’ search.

Search operation

A ‘precise’ search is carried out in all cases.

When the search returns a URI, the properties of that URI are checked for owl:sameAs and skos:exactMatch statements. [9]

Add operation

As the result of an Add operation, a new entity will be created in the ‘home’ triple store supported by the reconciliation service. The response will simply be the URI of this new entity.

[9] … how far do we want the service to go as regards resolving matches across multiple entities? For example, one record might specify date of birth to the day; another might have dob to the year, but additionally specify the place of birth. There is an argument for keeping the service as simple (and so efficient) as possible, while delivering useful results