What is Linked Historical Data?

Albert Meroño-Peñuela[1,2] and Rinke Hoekstra[1,3]

[1] Department of Computer Science, VU University Amsterdam, NL
{albert.merono,rinke.hoekstra}@vu.nl
[2] Data Archiving and Networked Services, KNAW, NL
[3] Faculty of Law, University of Amsterdam, NL

Abstract. Datasets that represent historical sources are relative newcomers in the Linked Open Data (LOD) cloud. Following the standard LOD practices for publishing historical sources raises several questions: how can we distinguish between RDF graphs of primary and secondary sources? Should we treat archived and online RDF graphs differently in historical research? How do we deal with change and immutability of a triplified History? To answer these fundamental questions, we model *historical primary and secondary sources* using the OntoClean metaproperties and the theories of perdurance and endurance. We then use this model to give a definition of Linked Historical Data. We advocate a set of publishing practices for Linked Historical Data that preserve the ontological properties of historical sources.

Keywords: Semantic Web, Linked Data, Historical Data

1 Historical Sources as RDF Graphs

Historical sources have traditionally been encoded in different formats: from papyrus to digital images, through books, tapes and photographs. It is not difficult to see the benefits of publishing historical sources as Linked Open Data [7]. However, it is unclear whether standard Linked Data modeling and publication pipelines are suitable for historical sources. In this paper, we are interested in modeling the fundamental properties of historical sources in order to make explicit to what extent current Linked Data publication procedures are adequate.

Independence and *reliability* of sources are fundamental issues historians take into account in scholarly writing [5]. To address these, historians distinguish between *primary* and *secondary* sources. Primary sources are "original materials created at the time under study that have not been altered or distorted in any way" [2,1]. Secondary sources are "documents that relate or discuss information originally presented elsewhere, written after the fact with the benefit of hindsight" [1]. A fundamental difference between the two is that primary sources must be *immutable*: they cannot be altered once they are created. Traditionally, immutability of sources is achieved through *archiving* them, either as books (in a library or book archive), as physical objects (in a museum archive), or more recently as digital objects (in a digital archive). The archive is the authority that

protects the primary source from change, providing *independence* and *reliability*. As a consequence, primary sources are inevitably detached from their original context. Secondary sources are attempts from historians to recreate this context.

A strict requirement then is that RDF graphs of primary sources need to be immutable as well. But how does RDF deal with change over time [8]?

> Intuitively speaking, changes in the universe of discourse can be reflected in the following ways:
>
> 1. An IRI, once minted, should never change its intended referent.
> 2. Literals, by design, are constants and never change their value.
> 3. A relationship that holds between two resources at one time may not hold at another time.
> 4. RDF sources may change their state over time. That is, they may provide different RDF graphs at different times.
> 5. Some RDF sources may, however, be immutable snapshots of another RDF source, archiving its state at some point in time.

Statement 1 is problematic: a primary source that changes keeps its IRI although its identity is changed (see Section 2). In addition, statements 4 and 5 have important consequences for historical sources. First, it follows the *alive-dead* Linked Data dichotomy: on the one hand, there is a *living* LOD cloud that is constantly updated and changed; on the other hand, a *dead* and archived LOD cloud exists as old snapshots of what once was alive. This situation corresponds to the life cycle of primary and secondary sources. All sources are first ordinary living LOD data, but the fact that they are *archived to preserve their immutability* turns them into primary sources. The metaphor of the *alive-dead* LOD serves well the purpose of primary and secondary sources as RDF graphs. For a primary source to be represented as an RDF graph, it is necessary (and sufficient) to be archived and preserved from change. RDF graphs of secondary sources, on the other hand, live in the LOD cloud similar to other datasets.

2 An Ontological Framework for Historical Sources

What are the basic ontological properties that characterize historical sources? In order to come with appropriate proposals on how to publish historical primary and secondary sources as LOD, we first need to understand their fundamental characteristics. We apply the philosophical stances of *perdurance* and *endurance* and the OntoClean methodology of [4] to study ontological properties that apply to historical sources; we model these sources according to their properties.

Perdurantism holds that ordinary things like animals, boats and planets have temporal parts (things persist by *perduring* through time). *Endurantism* is the stance that ordinary things do not have temporal parts; instead, things are wholly present whenever they exist (things persist by *enduring*) [6]. The DOLCE ontology [3] translates these stances to two types of entities: *endurants* and *perdurants*, which can be characterized by whether or not they can exhibit change in time. Endurants "can "genuinely" change in time, in the sense that the very

same endurant as a whole can have incompatible properties at different times; perdurants cannot change in this sense, since none of their parts maintain identity in time." Secondary sources, such as comments, notes, articles, annotations, are *endurants*; at any point in time they can be appreciated as a whole, while they still may undergo changes (e.g. working papers). Primary sources, on the other hand, have the same enduring properties, but *cannot* change: if any of their properties change, they lose identity. The accumulation over time of secondary sources that share a dependency on a primary source form our "body of knowledge" about the historical entity represented by the primary source. This accumulation is a **perdurant** (similar to item 4 in Section 1). On the other hand, the "state of our knowledge" at any point in time is a *slice* of that body of knowledge: a collection/set of **endurants**.

Distinguishing between perdurants and endurants is closely related to the question of identity: if sources can change over time, how can we guarantee that they are the same entity? To help answering this question, we use the OntoClean methodology [4]. Following OntoClean, some properties are essential to *all* their instances; we call these properties **rigid (+R)**. For instance, an entity having the property of *being a person* is guaranteed to preserve its identity even if other of his properties change, because *being a person* is rigid. Properties that are not essential for *some* of their instances are called **non-rigid (-R)**; of these, properties that are not essential to *all* of their instances (i.e. required to change) are called **anti-rigid (∼R)**. According to this, the property of a source being primary is **+R**, because being primary is essential to all sources (i.e. if it stops being a primary source, it no longer exists); the property of being secondary is **∼R**, since secondary sources may become primary sources through archiving.

Another way of looking at identity of historical sources consists of considering them as *sortals*. A **sortal (+I)** is a class all of whose instances are identified in the same way. The class of secondary sources does not carry any identity criteria (i.e. a secondary source cannot be identified by any predefined set of characteristics). On the other hand, a primary source is always **+I**: its identity criteria cover all of its properties (in order for something to be a primary source, none of its properties is allowed to change).

Unity (+U) is the metaproperty of classes all of whose individuals are *wholes* under the same relation. A *whole* is an instance that, in opposition of *mere sums*, does not create new instances of the same class it belongs to when an arbitrary subsection of such instance is considered. For instance, splitting a piece of *clay* in two, constitutes two pieces of *clay* (it is a *mere sum*), while this does not typically happen with e.g. instances of the class *person* (a *whole*). Primary sources are **anti-unity (∼U)**, since any part in which a primary source may decompose creates a new primary source. Think of historical objects for which only certain parts could be preserved; these parts constitute the genuine primary source. In case new parts of the object were found, these would constitute different independent primary sources. Secondary sources are **+U** because specific relations between their parts preserve their integrity as wholes, and arbitrary parts of them do not constitute secondary sources anymore.

Secondary sources	**Primary sources**
Endurant	(Strong) Endurant
Alive datasets	Dead datasets
Non-timestamped resources	Timestamped resources
Dereferenceable IRIs	Archive-only-dereferenceable IRIs
Anti-rigid $\sim$**R**	Rigid $+$**R**
Dependent (on the primary Source) $+$**D**	Independent **-D**
Not a sortal **-I**	A sortal $+$**I**
Unity $+$**U**	Anti-unity $\sim$**U**

Table 1: Ontological metaproperties of historical sources.

Finally, a property is **dependent ($+$D)** if each instance of it implies the existence of another entity [4]. Primary sources are **independent (-D)**, given that they can exist independently of other entities. However, secondary sources are $+$**D**: every secondary source is always *about* some existing primary source.

Table 1 shows the correspondence between the properties of primary and secondary sources. We model historical sources using the study of this Section and the considerations made in Section 1.

3 Linked Historical Data: From Modeling to Publishing

The model proposed in Section 2 conflicts with some of the basic LOD publishing principles, more concretely with the openness of the Web. The AAA rule (Anyone can say Anything about Any topic) is one of the essential principles of the Web, which also holds for RDF data. The IRI of a primary source can be used by anyone as a subject of an RDF statement; this changes the graph of the primary source, and breaks the basic principle of immutability of primary sources (see Section 1). In this Section we investigate mechanisms to publish Linked Historical Data as LOD according with the model of Section 2.

To solve this conflict, we propose *dereferenceability* as a mechanism to preserve the fundamental properties of primary sources and their interplay with secondary sources. Concretely, we propose a dereferencing service for authoritative digital archives hosting RDF graphs with two essential characteristics: *reliability* and *independence* (the overlap with the requirements for historical sources in Section 1 is no coincidence). First, dereferenceability is the *only* mechanism by which users may know that they are talking about a primary source. Hence, it is necessary that, when asked, the authoritative archive provides information on whether it *knows* something about such primary source or not (e.g. via SPARQL ASK queries). This way we achieve *reliability*: the only reliable primary source triples are those for which the archive returns valid descriptions. This will not happen with triples about the primary source issued by anybody else.

Second, when users dereference triples of a primary source, they get back *copies* of it, where all IRIs are replaced by new ones, but refer back to the original primary source IRI (or IRIs inside the primary source graph) through `prov:wasDerivedFrom` relations. This way we achieve *independence*: new statements (i.e. secondary sources) refer to this qualified copy independently on the

original contents of the in-archive primary source graph, preserving its independence and immutability. Alternatively, the resolution creates a new FRBR expression of the same FRBR work, but then the primary source is also an expression (the first). URN-like tricks (cf. DOIs) could also be used, such that users have to use a trusted dereferencing service at the archive location to obtain the primary source data (e.g. crossref dereferences DOIs at `http://dx.doi.org/` to the printer's page, or to an RDF representation of the source).

With our proposed model and publishing study, we can now answer the question: what is Linked Historical Data? Historical data is the union of *primary sources* P and *secondary sources* S: (a) a *primary source* P is an accumulation of strong-endurant, dead, timestamped, only-archive-dereferenceable, rigid, independent, sortal, and anti-unity resources; (b) a *secondary source* S is an accumulation of endurant, alive, non-timestamped, dereferenceable-by-anyone, anti-rigid, dependent, non-sortal, and unity resources; (c) statements in S contain links to statements in P; and (d) for any time t in which a statement of P is made, and for any time t' in which a statement of S is made, $t' > t$.

In this paper we argue for an ontological model and a consequent adequate publication of historical sources as RDF graphs in the LOD cloud. We advocate the use of the OntoClean methodology and DOLCE to give characterizations of *primary* and *secondary sources*. We propose the implementation of specific IRI dereferencing services in digital archives to preserve these fundamental properties of historical sources in the form of *independence* and *reliability*.

Acknowledgements This work was supported by the Computational Humanities Programme of the KNAW (see `http://ehumanities.nl`) and the Dutch national program COMMIT. We acknowledge suggestions contributed by colleagues, especially Christophe Guéret.

References

1. Australia, J.C.U.: Primary, secondary and tertiary sources. `http://libguides.jcu.edu.au/primary`
2. Benjamin, J.R.: A Student's Guide to History. Bedfors/St. Martin's, Boston (2004)
3. Gangemi, A., Guarino, N., Masolo, C., Oltramari, A., Schneider, L.: Sweetening ontologies with dolce. In: EKAW. pp. 166–181 (2002)
4. Guarino, N., Welty, C.: A Formal Ontology of Properties. In: Dieng, R., Corby, O. (eds.) Knowledge Engineering and Knowledge Management Methods, Models, and Tools, Lecture Notes in Computer Science, vol. 1937, pp. 97–112. Springer Berlin Heidelberg (2000), `http://dx.doi.org/10.1007/3-540-39967-4_8`
5. Helge S. Kragh: An Introduction to the Historiography of Science. Cambridge University Press (1989)
6. Katherine Hawley: "Temporal Parts". The Stanford Encyclopedia of Philosophy (Winter 2010 Edition). `http://plato.stanford.edu/archives/win2010/entries/temporal-parts/`
7. Meroño-Peñuela, A., Ashkpour, A., van Erp, M., Mandemakers, K., Breure, L., Scharnhorst, A., Schlobach, S., van Harmelen, F.: Semantic Technologies for Historical Research: A Survey. Semantic Web Journal (2012), to appear
8. World Wide Web Consortium: RDF 1.1 Concepts and Abstract Syntax. `http://www.w3.org/TR/rdf11-concepts/`