The semantics of cyberpast
Modelling historical data for LOD research environments

Sebastiaan Derks (Huygens ING)
Menno den Engelse (Hic sunt leones)
Antske Fokkens (Vrije Universiteit)
Gerard Kuys (Ordina / Nationaal Archief)
Harm Nijboer (Huygens ING)
Lodewijk Petram (Huygens ING)
Veruska Zamborlini (Universiteit van Amsterdam)
Ivo Zandhuis (Adamnet)

Introduction

In recent years the Linked Open Data paradigm has given rise to several initiatives to aggregate large amounts of historical data in digital research environments. The technological standards to connect distributed resources are by now well established, but the semantic and conceptual challenges faced by these projects are not fully recognized and remain underdiscussed. One major problem is that the LOD paradigm starts with the premise that all data providers use more or less the same lingo, but in practice even (seemingly) simple concepts like place and time might be treated quite differently across separate domains and resources. This has repeatedly been referred to as the ‘identity crisis of the semantic web’ [1]. From an engineering point of view this might seem a minor issue remaining to be solved, but for historians it is at the heart of their data and historical knowledge.

The semantics of data aggregations in the field of Arts and Humanities form an issue that is set to become even more pressing in the coming years. Recent technical advances have brought the field of digital humanities to the point where giant text corpora and archival collections can be automatically mined for a multitude of concepts relevant to historical research. As a result, the amount of (semi-)structured historical data will multiply in the coming years. From a technical viewpoint, it will be possible to connect large part of these data observations to LOD digital research environments, e.g. using advanced automated record linkage methods. However, to really make the data usable for relevant and important research questions, and to turn mere data production into knowledge production, it is indispensable to get the semantics of the connections between data observations right. After all, nuances in the meaning of observations of the past are crucial to almost any research in the Arts and Humanities; if LOD research environments fail to convey these nuances, they will remain of limited use.

Objectives

The objectives of this round table are to raise (renewed) awareness for the need of an interoperable academic linked open data structure that accommodates the complexities of Arts and Humanities datasets, propose (preliminary) solutions to common problems, and set up a working group (composed of round table participants and interested members of the audience) that will continue the discussion and see to the practical implementation of solutions. To these ends, we bring together a group of LOD specialists from universities, research institutions and heritage institutions, who have complementary experiences on working with LOD research environments in various domains within the Arts and Humanities. Each of the invited speakers will give a brief introduction on the conceptual issues they came across and the solutions they came up with. Thereafter, we will discuss on a general level to what extent the proposed solutions are interoperable. Of course, there will be ample room for the audience to participate in the discussion.

(Non-exhaustive) list of issues to be addressed

- *Conceptual (dis)unity over time*
  Most modellers of historical data will sooner or later need to address the issue of conceptual unity over time. Should we for instance consider Octavian and August to be the same conceptual entity? And is New York the same entity as Nieuw Amsterdam? Thinking of these entities as a human being and a location, respectively, the answer to both questions would probably be ‘yes’. But from a political perspective there might in both cases be good reasons to consider them as separate entities. Moreover, the answer to these questions also depends on whether one defines identities as static or in flux. If the latter is the case it would even be impossible to step in the same river twice, as Heraclit pointed out already long ago.
- *Reification (history)*
  In the Arts and Humanities, and the social sciences, reification is defined as attributing agency to abstractions, for instance in a statement like ‘The economic crisis made many people lose their jobs’. Most historians will agree that there is no such actor as an economic crisis that can make people lose their jobs and that statements like these should be understood metaphorically. Nevertheless, automatically generated representations of events are likely to contain such reifications. Reifications also appear in the world of art and literature, often more subtly, for instance in the titles of works that appear in multiple forms. Is there an entity ‘Hamlet’ that exists independently of its concrete manifestations in books, texts and performances? Furthermore, in art history so called ‘notnames’ are often used for unidentified artists to whom one or multiple works of art can be attributed with some certainty. However, such invented artists might be equal to one (or more) of the painters described in other resources (e.g. guild records) and of whom no works are known.
- *Ambiguous observations*
  Different observers may interpret data differently. How do we deal with conflicting attributions of works of art to certain artists? And can we define best practices for data connections made by automatic disambiguation software? A text in which ‘Michiel de Ruyter’ occurs in the context of a movie and actor Frank Lammers would for example be difficult to process. ‘Michiel de Ruyter’ refers to the admiral, but to the real or the fictional one? A connection between this data observation and an entity that describes the admiral should somehow make clear what the connection with the admiral entails.
- *Ambiguous sources*
  Scholars in the Arts and Humanities work with sources made by humans. These sources often contain ambiguous information or downright errors. For example,

historical literature often names the Duke of Parma as the leader of the successful siege of Antwerp (1584-1585). However, Alexander Farnese was only later to become the Duke of Parma; during the siege, his father Ottavio still held the ducal title. How do we deal with this kind of ambiguous data observations in LOD digital research environments?

References

[1] Halpin, Harry, and Patrick J. Hayes (2010). When owl:sameAs isn’t the Same: An Analysis of Identity Links on the Semantic Web. LDOW 2010.