Curating and Archiving Linked Data Datasets from the Humanities - From Data of the Present to Data of the Future

Overview

Digital Humanities are inseparable from digital (digitized) collections. In research infrastructures, be it on national or international level, we see networks between established and new repositories, digital libraries, archives and other content-providers and research institutions. This panel addresses new challenges for the alliance of research communities and service institutions which emerge from new data formats, vocabulary standardization efforts, and collaborative practices.

We focus on semantic web technologies and the adoption of Linked (Open) Data principles in humanities research, and more particular, data curation practices around them. The appeal of Linked Data to humanities can be explained by the inherent capacity of this new technology to link heterogeneous resources using shared and standardized schemas and vocabularies. This resonates with epistemic cultures in the humanities where scholars are used to work with different sources, viewpoints, and interpretations. Although, still early in the adoption process, the past decades have seen a number of projects which incorporate semantic web technologies into digital humanities methods. [e.g. 1,2,3]

At the same time, there are intriguing pending questions about the long-term preservation of Linked Data datasets as complex digital objects. Examples of challenges are 1) the versioning aspects of the vocabularies and classifications, 2) the provenance of the metadata, 3) the attribution of authority for independently and heterogeneously created (meta)content and 4) the selection process of vocabularies and schemas, especially when there are multiple options.

This panel enables a discourse on higher level principles of working with and preserving Linked Data in the humanities based on reports of concrete experiences on the workfloor. We discuss differences in data curation as conceptualised and executed during research and when preserving research data for the long-term. We further discuss how to align those different curation practices with the ultimate goal to making Linked Data used in digital humanities FAIR.

Questions to be addressed:

- How to take care of Linked Data curation and management during research, and how to organise effective collaboration between semantic web technology pioneers and pioneers in Digital Humanities application of this technology?

- How to exploit the possibilities for efficiency and synergy of Linked (Open) Data in distributed networks between research, collections and overarching service providers?
- How to best bridge conflicting priorities between enabling research (building, enriching, analysing) based on Linked (Open) Data technology and the long-term preservation of Linked Data datasets (data of the present and data of the future)?

Papers

1. When to store what and how? Data curation challenges during research projects

2. Indexing Cultural Heritage Resources for Research and Education

3. Automated Vocabulary Suggestion, Domain-Specific Linked Entity Recognition and Visually Faceted Verification

4. 404 Error - Resource Not Found: Why we Need to Rescue Endangered Knowledge Organization Systems

5. Archiving Linked Data Datasets - Experiences from a Humanities Research Infrastructure

6. Dutch Historical Censuses - Preserving Data for Research, the Wider Public and Future Generations

References

[1] Hyvönen, E. (2012). Publishing and using cultural heritage: Linked data on the semantic web. S.l.: Morgan & Claypool.
[2] Merono-Penuela, A., Ashkpour, A., van Erp, M., Mandemakers, K., Breure, L., Scharnhorst, A., Schlohbach, S., van Harmelen, F. (2014). Semantic technologies for historical research: A survey. Semantic Web Journal, 6(6), 539-564. DOI: 10.3233/SW-140158
[3] Idrissou, A, Zamborlini, V, Latronico, C, van Harmelen, F & van den Heuvel, CMJM 2018, 'Amsterdamers from the Golden Age to the Information Age via Lenticular Lenses: Short paper'.

Papers

Paper 1

When to store what and how? Data curation challenges during research projects

Many funding agencies require that data is archived after the end of the project and made available to the research community for further reuse [1]. There is also a lively discussion about making experimental data available in various research communities, for example in connection to publications [2]. In short, data curation challenges appear at different stages of the research and data life cycle and can create various tensions within research projects when they try to comply with this. This holds in particular for humanities projects that are not explicitly aimed at creating one or more datasets.

In this paper, I provide examples from the intersection of semantic web and the humanities and illustrate data curation issues at three stages of the research process: 1) knowledge modelling, 2) exploratory experiments, and 3) userfriendliness of semantic web technology. In each step, data curation is a prerequisite to data sharing and presents different challenges.

Knowledge Modelling

One of the main difficulties in digital humanities research is the gap between the 'digital' and the 'humanities'. Simply put, computers need clearly defined concepts and boundaries between concepts whereas in the humanities concepts are more fluid and the world is less black and white, but more a variety of shades of grey. Many computer science datasets are simplifications of the world, for example GeoNames [3] focuses on current location names and country boundaries, whereas countries and cities have grown, shrunk, merged and split and borders have changed through time. Concepts and entities also change over time. Present-day England is not the same as the England of 200 years ago (think of the kingdom's borders, population, customs and how others perceive this concept). Current LOD resources fail to capture this complexity. Making the essence and limits of the data context explicit both for machines and scholars is a core challenge.

Exploratory Experiments

Linked data practices originate from computer science, where the research process differs from the humanities. A historian may start digging into an archive with a general research question, that will be further specified during the search in the archive. This may mean that documents selected at the start of the process as relevant may be discarded later on, and the type of information extracted from the documents to base analyses on may evolve over time too. A provenance trace of the

search queries and resulting selections can only capture part of this information. It is therefore still an open question what kind of provenance trail is needed to be shared in the actual research process and in the long-term.

Userfriendliness of semantic web technology
Great strides have been made in making semantic web technology more accessible for non-experts (e.g. [4]). However, this technology is still not well integrated with the daily practices of humanists, barring the reuse and re-sharing of data according to linked data principles. The difference in knowledge between developers of semantic web research infrastructure and those conducting humanities research, even in interdisciplinary projects hampers discourse on data curation.

[1] http://ec.europa.eu/research/participants/docs/h2020-funding-guide/cross-cutting-issues/open-access-data-management/data-management_en.htm
[2] http://lremap.elra.info/
[3] http://geonames.org
[6] http://grlc.io/

Paper 2
Indexing Cultural Heritage Resources for Research and Education
In an increasingly digital world, e-infrastructure has become a key component of the daily life of researchers, underpinning a variety of research and academic activities. Over the period 2014-2020 the development of these infrastructures is supported by 850 M€ of funding from the European Commission. Those systems at the core of data-driven science are expected to not only contribute to the scientific discoveries but to also play a larger role in society. This paper reports on the results of a specific project [1], which engaged with cultural heritage content in the UK based on Linked Data principles. Its core is an open platform which indexes and organises the digital collections of libraries, museums, broadcasters and galleries to make their content more discoverable, accessible and usable to those in UK education and research. Among them are images, TV and radio programmes, documents and text from world class organisations such as the British Museum, British Library, National Archives, Europeana, Wellcome Trust and the BBC.

Next to linking content, the project also supported developers to create digital educational products that will inspire learners, teachers and researchers by using applications powered by the [*project, anonymised*] platform. This paper discusses the challenges faced by the project, the architecture developed for tackling them, and the lessons learned. In particular, we address challenges for consuming Web data; the problem of co-referencing (how to deal with the fact that several URI's can be created to refer to the same thing); and most prominently the problem of

licensing. In particular, we discuss how the lack of unambiguous declarations of copyright and license as metadata are hampering the re-use of existing published data, and which methods have been tested so far to circumvent this problem. The paper closes with an inspection of other existing collections and platforms and a discussion on how they solve the above listed problems.

[1] Removed for anonymisation purposes

Paper 3

Automated Vocabulary Suggestion, Domain-Specific Linked Entity Recognition and Visually Faceted Verification

Attempts to order our knowledge in a way that we can easily find what we are looking for is as old as the body of (written) knowledge grew. But, despite of sophisticated Knowledge Organisation Systems (as classifications used in libraries), or new network-based principles (as the page rank algorithm), we all share painstaking experiences when trying to find back something via Google, knowing it is *there*, but simply not being able to find it via the keywords coming into mind. Homonyms, synonyms, generalizations, specializations, spelling variations and mistakes, language versions etc. are all complicating the keyword-based approach. Linked Data is one step forward (or actually backward in the right direction) on solving some of the problems with keyword search. By having shared vocabularies created and maintained by experts on various domains, the digital items can be annotated with them and, in principle, easily retrieved by other experts from the same domain without being a skilled librarian as well. Unfortunately, there are still two main problems:

- First, there is no single authority that for each possible domain defines a satisfying and indisputable annotation vocabulary. What we observe instead is a multitude of attempts, often tailored for momentus and local needs, and characterised by conflicting paradigms, missing subjects, and logical inconsistencies. In other words, one *still* needs to know which vocabulary is good enough and shared by a critical mass, if available.
- Second, and even a more prevalent problem is annotating content is time consuming, requires specific skills and is not something with which one can boost one's career. Data curation as natural part of the research cycle is often invisible and undervalued. Where a researcher submitting a paper to a journal might still try her best to annotate it with some keywords, learning about Linked Data and then finding the proper URIs for a machine readable annotation is a bridge too far.

This paper presents research which addresses these two problems by means of a automated and visually attractive support system.

To tackle the first problem we will investigate various quality aspects of the *existing* vocabularies by analyzing their expressiveness, relations and popularity via various metrics on the currently largest aggregation of Linked Data available: the LOD

Laundromat. When successful, a domain expert can be confident that for her goals the best possible set of vocabularies and schemas are proposed to annotate her content.
For the second problem we develop an integrated visual interface for a selection of Linked Entity Recognition tools developed in [*a research infrastructure, anonymised*], and apply the most appropriate combination for the best matching vocabularies and schemas according to the results from the first step.
In other words, when a person wants to annotate a document, the first step will suggest the best vocabularies and the second step automatically detects entities in the document specified by these vocabularies.
To summarize, by applying the expertise on library classification systems and integrating results from various projects, the task is to investigate and develop automated support for vocabulary selection and its domain specific entity recognition approach accompanied with intuitive domain specific visual verification support. The paper reports about first experiences with the system, and discusses further requirements for a wider implementation.

Paper 4

404 Error - Resource Not Found: Why we Need to Rescue Endangered Knowledge Organization Systems

‘The name is the thing, and the true name is the true thing. To speak the name is to control the thing.’ - Ursula K. Le Guin, *The Rule of Names*

Reducing ambiguity in research and improving interoperability of data and results are increasingly important themes within the field of Digital Humanities. According to the FAIR Data Principles in order to achieve such interoperability, it is necessary to use ‘a formal, accessible, shared and broadly applicable language for knowledge representation’ and ‘vocabularies that follow the FAIR principles’ [1]. As suggested by American science-fiction and fantasy novelist Ursula K. Le Guin in the above quote, we can only know a thing is a thing, and that it is the same thing that we are both speaking about, once we know and agree upon its name.

Knowledge organization systems (KOS) include a variety of schemas that organize, manage, and retrieve information and knowledge. They range from ontologies, to classification schemes, thesauri, taxonomies, controlled vocabularies and semantic networks [2]. Given their role in facilitating access to and retrieval of information, they play an important role in vocabulary standardization. Since KOS are constantly changing to reflect dynamic growth in knowledge, they present unique challenges in terms of maintenance and tracking the provenance of their changes. We can

consider KOS as ‘bridges’ allowing for shared meaning, since they connect terms with the appropriate concepts and knowledge. But what happens when these ‘bridges’ are closed, broken or simply...lost?

Over the course of a yearlong study, using a method of humanistic empirical association, this project has studied KOS ‘in the wild’, exploring their attributes, histories and evolutions. During the study, the ‘404 Error’ became a relatively familiar sight, highlighting the fragility and lack of stability of (some of) these systems, leading to questions about how to identify KOS which are at risk of disappearing. A lost KOS is a lost ‘bridge’ to the meaning behind a term in a given context, place or moment in time.

This paper aims to make recommendations for ensuring the long-term preservation of KOS, for example, ensuring that systems used in scholarly research today remain FAIR, and can be guarded against loss of meaning through semantic drift or the disappearance of the resource. A number of case studies will be presented in order to highlight the importance of these resources within the field of Digital Humanities and also to raise awareness about the benefits of their application. Finally a number of recommendations will be proposed related to common tools and mechanisms which need to be developed for their proper archival and maintenance, in order to make them FAIR.

[1] https://www.force11.org/group/fairgroup/fairprinciples
[2] Mazzocchi, F. (2018). “Knowledge organization system (KOS): an introductory critical account”. *Knowledge Organization* 45, no. 1: 54-78.

Paper 5

Archiving Linked Data Datasets - Experiences from a Humanities Research Infrastructure

Archiving RDF (Resource Description Framework) datasets may seem trivial. The notorious problem of format-outdating is non-existent, RDF can be expressed as essentially plain text in various specialized formats. No need to update this format to rapidly changing format specifications and ditto applications. The challenge of placing data and datasets in a proper context in order for them to be understandable and interpretable in time seems no greater than that for datasets with files of a more traditional format. In fact, less so because RDF data is self-describing in the sense it places itself in context by linking to a network of concepts and meaning - as long as its links are resolvable. It is this condition - the resolvability of links which in essence are references to other knowledge graphs - which makes archiving RDF still a challenge. In this paper, we look at this problem more carefully for a concrete case. We start with the question: What are RDF data and what makes RDF datasets

different? We continue to ask: Do the nature of RDF data (sets) imply that we must reconsider the cycle of producing, archiving and reusing data?

According to well accepted definitions, RDF datasets package up zero or more named RDF graphs along with a single unnamed, default RDF graph[1] and a RDF graph is a set of RDF triples[2]. A triple or quad is, by its nature, an atomic statement; a RDF dataset can be fragmented into smaller chunks without compromising the validity or understandability of these smaller sets of triples. Does the fact that RDF datasets can be fragmented make special demands on the findability of these fragments and therefore on our search engines? Does it make demands on the way we must be able to retrieve these fragments and reuse them?

The role of digital archives can be described as guarding the technical and cognitive interpretability of its assets over time. Technical interpretability can be secured by scanning and converting the formats of files in the archive at regular intervals. Cognitive interpretability, from the side of the archive, is usually covered by a one-time effort that comprises securing and placing the dataset in its due context (metadata, code books). In this traditional scheme, the role of cognitive interpretation is carried out by humans. With RDF datasets, for the first time, the role of cognitive interpretation can (in fact should, because no one is going to read an RDF file) be carried out by machines. Does this inevitable machine interaction change conditions of archival interfaces and the datasets themselves?

In this paper, we report our findings to build a pipeline from a tool developed for academic research, which offers the ability to access and edit data as RDF graph to a long-term archive.

Paper 6

Dutch Historical Censuses - Preserving Data for Research, the Wider Public and Future Generations

This paper describes research data curation steps which put the Dutch historical censuses (1795-1971) in a cultural and historical context by using semantic web technologies. We reflect in particular about data curation and archiving after the project has ended.

The fact that the resulting RDF is an integrated dataset means that we can study the landscape of the historical censuses, the evolution of demography, labour and

[1] https://www.w3.org/TR/rdf11-mt/#rdf-datasets

[2] https://www.w3.org/TR/2014/REC-rdf11-concepts-20140225/#section-rdf-graph

housing in the Netherlands for a period of two centuries. With the transformation into RDF, for the first time ever, we can ask questions on the Dutch historical census data as a whole. Due to its connections to external resources in the LOD cloud (DBpedia, *Gemeentegeschiedenis* [7], HISCO, ICONCLASS, and Dutch Ships and Sailors [8]) questions that can only be addressed by combining all these data sources together can be answered. To give an example, we can now ask how many houses have been under construction across specific years at municipality level across the country.

During the project, a pipeline was developed [1] from the 'raw data' to an enriched RDF graph. Data have been harmonised along certain dimensions and perspectives, creating an iterative cycle between the domain specific analysis, annotation, enriching of the original data and its ingestion into new versions of the RDF graph (in the image the harmonised dataset).

At the end of the project we applied a multi-tiered strategy to preserve data both for present research as well as for the future. Concerning the first, data is deposited in online repositories (GitHub) to stimulate further reuse. Additionally, the RDF data graph is available online [2] as SPARQL endpoint. A website has been created [3] to enable human interaction. Moreover, we followed a similar approach to store, manage, and publish queries using the GitHub infrastructure [4]. As a result, these can be collaboratively edited by humans, but also reused by machines for different purposes. Importantly, this includes the automatic generation of APIs [5], through which the whole RDF dataset can be easily accessed by users and standard HTTP clients. The enrichment of the RDF, through updates in all these resources, is an ongoing effort.

Complementary, two coupled datasets (1) A first version of the Linked Data publication of the Dutch Historic Census (http://dx.doi.org/10.17026/dans-xpk-wj5w), and (2) a related dataset Geharmoniseerde census data, 1859-1920 which details the harmonisation for the tables Plaatselijke Indeling as part of Volkstellingen years 1859, 1869, 1879, 1889, 1899, 1909 and 1920 (http://dx.doi.org/10.17026/dans-z9j-x2vy) have been deposited in the long-term, certified, data archive EASY (hosted by DANS). To address the problem of long-term preservation of Linked Data which by nature come with references to other resources, the deposit (1) includes snapshots of all vocabularies, in their specific versions, on which the RDF database depends.

[1] Removed for anonymisation purposes
[2] Removed for anonymisation purposes

[3] Removed for anonymisation purposes
[4] Removed for anonymisation purposes
[5] Removed for anonymisation purposes
[6] Removed for anonymisation purposes
[7] https://www.gemeentegeschiedenis.nl
[8] http://dutchshipsandsailors.nl