Semantics-based Indexing of Historical Entities

Julie Allinson

H00113866

August 2014

Computer Science

School of Mathematical and Computer Sciences

Dissertation submitted as part of the requirements for the award of the degree of MSc in Computing
Abstract

Indexing and researching historical entities, such as people, places and events, is common practice both for historical researchers who consult, analyse and annotate data, and also for cultural heritage institutions (libraries, archives, museums and galleries) who seek to describe their collections for discovery and re-use. This research project will investigate the potential offered by the semantic Web, Linked Data and allied technologies in supporting these endeavours. The project will prototype a semantic tool for the "researcher's workbench" to help research personal names. The tool will make use of Linked Data sources, such as the Virtual Name Authority File, DBPedia (the linked data representation of Wikipedia) and the British Museum collection catalogue, to allow the researcher to search across multiple sources, identify relevant entities and enrich the data found. Local data will be stored in a semantic RDF data store.

In developing the prototype tool, the project seeks to answer two research questions: would users benefit from such a tool? and can current semantic technologies and data sources support, and bring added value, to such a tool?

Acknowledgements

I would like to thank my supervisor, Albert Burger.

Jessica Stephens, Matthew Herring, Chris Webb, Sarah Rees Jones, Richard Stephens and Steph Taylor for their help with the requirements and evaluation.

The University of York for their support.

Finally, @louisemoody, @scrabblefscruff, @matildabscruff, @benosteen, @criticalsteph
# Table of Contents

1. Introduction : The Hypothesis -- page 5

2. Literature Review -- page 6
   2.1 Overview
   2.2 The World Wide Web, the Semantic Web and Linked Data
   2.3 Linked Data and the Semantic Web for the Humanities
   2.4 Semantics, Technology and Standards
   2.5 Summary

3. The Research -- page 20
   3.1 User Requirements
   3.2 Research Questions
   3.3 Technology and Design
   3.4 Building the application
   3.5 Summary

4. Results and Evaluation -- page 45
   4.1 User Evaluation
   4.2 Results and Technical Evaluation
   4.3 Summary and Future Directions

5. Conclusions -- page 53

References

Appendices

Appendix 1. User Group

Appendix 2. Data Sources

Appendix 3a. Requirements Survey Summary

Appendix 3b. User Research Questions

Appendix 4 : User Evaluation
List of Figures

Figure 1: The LOD Cloud

Figure 2: Use Case Diagram

Figure 3: Activity Diagram

Figure 4a: Class Diagram (Proposed)

Figure 4b: Class Diagram (Final)

Figure 5: Info level logging to show endpoint setup

Figure 6: Search Results

Figure 7a: These names are not the same

Figure 7b: These names are the same

Figure 8: sameAs link

Figure 9: Enhance with semantics
1. Introduction : The Hypothesis

"My first answer therefore to the question 'What is history?' is that it is a continuous process of interaction between the historian and his facts, an unending dialogue between the present and the past." (Carr, 1987, p. 30)

It is a familiar fact that historical research frequently involves consulting disparate and varied sources, often in physical form, often with the aim of tracking people, places, events and other 'things' over time. Whilst there are many digital and digitized sources available, the literature suggests these are underused. At the same time those responsible for cultural heritage collections, the librarians, archivists and museum curators, are tasked with describing collections to make them discoverable. They too need to describe people, places, events and other 'things', but the literature points to a mis-match between their practices and what the users' want.

Linked Data offers a solution to the 'data deluge' by outlining a paradigm for linking together exactly the kinds of entities of concern to the historians and information professionals above. Not only that but a solution that uses the very architecture of the Web, which is open and available to all.

The motivation for the project, then, is to bridge the gap between Linked Data, and the Semantic Web, as well as non-technical users, in order to allow the user to work more quickly, more efficiently and draw on a greater corpus of sources than would otherwise be possible whilst benefitting from both a platform they understand (viz. the Web) and the added value that semantic technology brings.

The project, Semantics-based Indexing of Historical Entities, aims to explore two research questions through building a prototype tool for researching personal names:

1) Would a tool, using semantic technologies to help with research into historical names, benefit it's target audience?

2) Can current standards, technologies and linked data services support, and add value to, the proposed service?

The following sections will provide a review of relevant current literature and technologies, an outline of the requirements, design and implementation of the tool, and, finally, a summary of the results and evaluation of the tool, along with a discussion of issues and potential for further work.

The prototype tool is available from http://geekscruff.me/peoplesparql.

The code is available from https://github.com/geekscruff/peoplesparql/releases/tag/final-submitted.
2. Literature Review

"If I am searching for "John Smith", it is of no interest whatever to learn that there was a different person called "Joan Smith" who happens to occupy the neighbouring place in the people authority file" (Anon., 2014)

2.1 Overview

The project aims to use the ubiquitous platform that is the World Wide Web to facilitate historical researchers in researching named entities, by exploiting what the Semantic Web promises - "a Web of data - of dates and titles and part numbers and chemical properties and any other data one might conceive of." (W3C, 2013a). The project focuses on the four and five stars of Berners-Lee's open data star system (Berners-Lee, 2010), aiming to explore what structured semantic information the Web has to offer historical research and the links to related entities and information that radiate from there.

As will be shown in the following survey of the literature, the Semantic Web and the Linked Data paradigm have been in development for over a decade yet are still in their relative infancy. There remain challenges to overcome in whether they will deliver on the desire to "open up the knowledge and workings of humankind to meaningful analysis by software agents" (Berners-Lee, Hendler and Lassila, 2001).

A second aspect of this work is exploring the interface between those "software agents" (Berners-Lee, Hendler and Lassila, 2001) and the end user, providing efficient and effective interfaces for users to perform digital scholarship whilst hiding the technical layer of URIs, RDF and ontologies that make the research possible. The domains of use that are the focus for the project are the humanities and the cultural heritage sector and in particular, the area of researching historical personal names. The literature will show much support, and some challenges, for a Linked Data approach for both cultural heritage cataloguing practice and humanities research.

2.2 The World Wide Web, the Semantic Web and Linked Data

In May 2001, Berners-Lee, Hendler and Lassila published an article in Scientific American entitled 'The Semantic Web'. In it they outline a direction for the Web, a movement away from traditionally document-centric Web 'pages' where machines have no way of understanding the meaning of the content (2001) to a semantically data-driven Web. They propose a vision for a semantic extension to the existing Web, where agents "roaming from page to page can readily carry out sophisticated tasks for users" (2001). XML and RDF are identified as the chief technologies, with URIs identifying every subject and object, and ontologies providing the "common meaning" needed to compare disparate data sources.

Five years later, in 2006 Berners-Lee, this time writing with Shadbolt and Hall, provides a 'state of the nation' for the Semantic Web, pointing to leaps forward in the development and use of the key building blocks - RDF, ontologies and URIs - noting particularly that "the argument in favour of ontologies has been won" (2006, p.96). The authors admit that "the data exposure revolution has not yet happened" (2006, p.99) but they remain hopeful and inspirational.
Talk of a "linked information space" (Shadbolt, Berners-Lee, White, 2006, p.100) had, by 2010, become established in the community as Linked Data and Berners-Lee, in a talk delivered at Gov 2.0 Expo, provides us with a framework for open and Linked Data that has become an unofficial benchmark:

"... one star for putting it up there, another if it's in machine-readable format ... an extra star if it's not a proprietary format, it's an open standard ... to get the fourth star you have to put it in Linked Data format ... you get the fifth star by doing the work of linking" (2010, 00:32-01:30)

At the same time he also gives the "bag of chips" example, breaking away from the traditional view of a single ontology whose use excludes other domains towards mix and match approaches where terms from different ontologies can co-exist and where it is perfectly reasonable that agents may understand only a part of a given dataset (Berners-Lee, 2010). As will be discussed later, the LOCAH project has taken this approach.

Heath and Bizer published, in 2011, *Linked Data: Evolving the Web into a Global Data Space*, a "conceptual and technical introduction" to Linked Data. Arguably the first 'manual' for Linked Data, as well as outlining technologies and approaches, the book lays down the Linked Data principles:

"1) Use URIs as names for things.

2) Use HTTP URIs, so that people can look up those names.

3) When someone looks up a URI, provide useful information, using the standards (RDF, SPARQL).

4) Include links to other URIs, so that they can discover more things." (Heath and Bizer, 2011)

In their conclusion, the authors look at the current movement towards linking data by companies like Facebook, Yahoo and Google and it's inherently closed nature, echoing comments in the introduction to the W3C's Semantic Web Activity that such data is "controlled by applications, and each application keeps it to itself" (W3C, 2013a). Like Berners-Lee and his co-authors before them, Heath and Bizer remind us that although publishing Linked Data takes effort, it's contribution to the open and global Web of data is much more in line with the founding principles of the Web and the open, free flow of information.

This notion of democratisation, of people, not companies, sharing information is not new. Allen (2013), writing about Web 2.0, reminds us that "usenet, bulletin boards, email lists, chat environments and multi-user domains (MUDs) had all previously demonstrated that, over the internet, many people wished to share information, create content and work with others" (2013, p.263). In short, individuals sharing information has *always* been at the heart of the Web:

"examples ... of online activity across many years, and the gradual nature of change, stand in contradiction to the current widely held belief that Web 2.0 made possible an online world of participatory, user-generated and open content and communication" (Allen, 2013, p.263)

The Web as a data sharing platform, developing as a gradual continuum from document-centricity to structured machine-understandable data, is a view echoed through all of the literature surveyed. Practical
details of application aside, the final word on this should go the original *Scientific American* paper whose authors, in conclusion, set out the benefits to human knowledge and humanity:

"The Semantic Web, in naming every concept simply by a URI, lets anyone express new concepts that they invent with minimal effort. Its unifying logical language will enable these concepts to be progressively linked into a universal Web. This structure will open up the knowledge and workings of humankind to meaningful analysis by software agents, providing a new class of tools by which we can live, work and learn together." (Berners-Lee, Hendler and Lassila, 2001)

### 2.3 Linked Data and the Semantic Web for the Humanities

How can the opportunity afforded by the Semantic Web and Linked Data be realised within the humanities?

It is widely acknowledged in the literature that there is a "proliferation of relevant digital resources in widely varying formats" across humanities disciplines (Burrows, 2011, p.178). This is in addition to the physical primary source materials that remain vital to humanities scholars. This wealth of resources bring several challenges. Talking specifically about studying medieval manuscripts, Burrows describes "this extraordinary - and continually growing - collection of riches brings its own set of complications and difficulties, and leaves us with something resembling the Tower of Babel" (2010, p.122). This is not limited to medieval manuscripts, as Burrows notes: "the data deluge is real and evident in the humanities" (2011, p.177). Yet evidence suggests that digital resources are underused. A study by Warwick found "the slow uptake of digital resources by humanities researchers has been recognized as a problem since the 1990s" with the cause attributed, perhaps, to the fact that "unused (or underused) digital resources may not have been designed in a way that can support the complex and different research methods used by humanities researchers" (2012). Burrows talks of "a lack of integration and interoperability between the many different sites" and the difficulty in finding out "systematically what research and digitization are being undertaken in collections around the world" (2011, p.127).

This paints a curious picture, of a landscape awash with primary and secondary source material where huge amounts of effort and funding have gone into making digital products available, but where researchers "have long chafed at the time necessary to work painstakingly through silos of bibliographic and digital materials" (Schaffner, Erway and OCLC Research, 2014, p.9). Resources seem to exist in silos that need to be independently discovered and queried.

The real challenge appears to be, not the availability of source material online, but of providing those resources in such a way as to make them useful to the very users who need them - "while the proliferation of Web resources is undoubtedly of great value to manuscript research, their sheer complexity and variety impose a significant barrier to our ability to ask large-scale research questions both about manuscripts as physical objects and about their content" (Burrows, 2010, p.122).

Cultural institutions, the keepers of primary source material and frequently the producers of digital resources, seek to describe and catalogue collections and resources, in order to provide access. As noted by Summers and Salo "for as long as libraries, museums and archives have built collections they have also created descriptions of the cultural artifacts in their collections" (2013). Whilst this is clearly a necessary...
activity and one involving a range of professional standards, some of the evidence mentioned above suggests that these practices aren't necessarily aligned with research methods (Warwick, 2012) - "institutions have their own ways of organising and describing source materials which may be quite different from the information produced by the research process" (Burrows, 2011, p.182). In addition, the interfaces to digital resources lack the "ability to encourage serendipitous discovery" (Burrows, 2013, p.578).

Yet there is a strongly aligned aspect of both humanities research and the curatorial practices of cultural institutions around the identification and description of 'entities' and the links between them that speaks directly to the subject of this project. Burrows identifies two basic components to humanities data. The first is around annotations and commentary, but it is the second that is of particular interest here: "the entities to which these annotations refer: concepts, persons, places and events, as well as creative works, artworks, publications, texts and other physical and digital objects" (2011, p.182) - the very thing that cultural institutions must describe: "controlled vocabularies for identifying people, topics, places and genre/form" (Summers and Salo, 2013), common across both domains.

The will to collaborate across these two professions seems to exist too. Schaffner and Erway suggest that digital humanities academics would like to "supplement, correct and add name authorities, and they imagine that librarians and archivists would be happy to have them do so" (Schaffner, Erway and OCLC Research, 2014, p.12) whilst VIAF 3, OCLC's international name authority, has actively sought contributions from scholars, for example by including names from the Perseus Project (Smith-Yoshimura, 2013a). Smith-Yoshimura from OCLC, in a separate post, identifies that "scholars and librarians share a mutual appreciation for each others' work on identifying names appearing in historical research" and evidences the learning that comes from scholars and librarians working together (2013). Burrows sees "bridging the gap between academic research and curatorial activities ... [as] ... a vital method of ensuring the future value of public cultural collections" (2011, p.189).

This leads neatly onto discussion of Linked Data and semantic techniques, seen by many as a possible solution to some of the issues and opportunities outlined above. For scholars, Linked Data has the potential to "represent a complex body of knowledge in a connected and interoperable way, and to provide a platform for applying sophisticated discovery and analytical tools to data across the Web" (Burrows, 2010, p.123). Linked Data could also offer greater opportunities for exposing the work of digital scholars, who, on the whole, want their projects to be discoverable (Schaffner, Erway and OCLC Research, 2014, p.10).

For cultural heritage institutions Linked Data is seen as an ideal mechanism due, in part, to their mission of facilitating access, and also because it brings them more into line with the wider Web architecture, something that could greatly increase the reach of digital resources. In the past, mechanisms for generating descriptions have been "wide ranging and often duplicative" (Summers and Salo, 2013) and Linked Data offers possibilities for aligning practice and sharing information.

Libraries, Archives and Museums have already begun to champion and make use of Linked Data. LODLAM, "an informal, borderless network of enthusiasts, technicians, professionals and any number of other people who are interested in or working with Linked Open Data pertaining to galleries, libraries, archives, and museums" 4 has been running since 2010 and helps promote Linked Open Data amongst the
relevant professionals. The Open Knowledge Foundation's OpenGLAM initiative \(^5\) goes further, providing community resources and workshops, and also publishing the OpenGLAM principles which set out to guide cultural heritage institutions in adopting standard open principles (Open Knowledge Foundation, [2013]):

"Galleries, libraries, archives and museums have a fundamental role in supporting the advance of humanity's knowledge. They are the custodians of our cultural heritage and in their collections they hold the record of humankind. The internet presents cultural heritage institutions with an unprecedented opportunity to engage global audiences and make their collections more discoverable and connected than ever, allowing users not only to enjoy the riches of the world's memory institutions, but also to contribute, participate and share. We believe that cultural institutions that take steps to open up their collections and metadata stand to benefit from these opportunities." (Open Knowledge Foundation, [2013])

There are now many examples of cultural institutions engaging with these principles, albeit at different levels of Berners-Lee's star system (Berners-Lee, 2010). Notable highlights include:

- The Library of Congress LC Linked Data service which provides 'access to commonly found standards and vocabularies promulgated by the Library of Congress' \(^6\)
- The British Museum Linked Data service which provides a Linked Data endpoint for it's entire collection catalogue \(^7\)
- The Getty Thesauri, long established as a central source, particularly for art scholars, are in the process of being released through a Linked Data service \(^8\)

An excellent example of Linked Data use in a project driven directly by scholarship is the Pelagios project \(^9\). Pelagios is actually a collection of projects "connected by a shared vision of a world" with each project contributing a different perspective on a common history. The aim of Pelagios is "to help introduce Linked Open Data goodness into online resources that refer to places in the historic past" (Pelagios Project, 2014).

Interestingly, although Pelagios is open to the world and thus can be used for a wide range of purposes, the project has a clear sense of what it will be used for, and this has driven the development:

"Why do we want to do that? Well, we think it will make all sorts of other things possible, including new modes of discovery and visualization for scholars and the general public. Pelagios also means 'of the sea', the superhighway of the pre-industrial world - a metaphor we consider appropriate for a digital resource that will connect references to ancient places."

(Pelagios Project, 2014)

Appendix 2 contains a list of Linked Data sources identified as potentially useful for this project.

Personal names, rather than places, are the focus of this research project. The literature has shown that describing and researching 'entities' and the connections between them are central to both humanities scholarship (Burrows, 2011, p.182) and cultural institutions' descriptive and cataloguing practices. Personal names are just one in a network of entities and in isolation, as static concepts, they are not of significant scholarly interest. But their relationship to topics, places, events, artworks, text and other physical or digital objects is where personal names become alive for scholars (Summers and Salo, 2013). Burrows has 'names' as one of his semantic categories, pointing to several existing sources seeing people, not as isolated
individuals, but in the context of their wider role: "people (authors, owners, editors, bibliographers, commentators, artists)" (Burrows, 2011). As a first step, though, researchers need to find information about the people 'entities' of interest, before they move onto the questions of what, when, where and how they are linked to other entities. The anonymous quote that opens this section is from one of the respondents to the project requirements survey and neatly captures for us the dislocation between traditional 'Library' approaches and what scholars want.

There are already many Linked Data sources of information for personal names, a selection of which will form the basis of this project. These include OCLC's Virtual International Authority File (VIAF) and the Getty's Union List of Artists' Names (ULAN) as sources of biographical and authority information, the Archives Hub's Linked Data service, Europeana, the British Library's Dictionary of National Bibliography and the British Museum Collection catalogue as sources of names and their links to other entities. Others will undoubtedly be discovered. But Burrows, in his piece on Medieval manuscript research, identifies a number of sources that are important but not available in any semantically queryable form. One such example is the Fasti Ecclesiae Anglicanae 1066-1857, 32 volumes of detailed information about the English clergy, freely available online but impossible to machine-query in a meaningful way.

Although this paints a slightly downbeat picture of a digital landscape that will remain "difficult to navigate and overwhelming in its richness and complexity" (Burrows, 2010, p.128), initiatives like VIAF, as noted above, do show how progress can be made (Smith-Yoshimura, 2013). Part of the challenge for this project, and mentioned in much of the literature is how to build an infrastructure that supports the release, analysis and use of new data sources, avoiding silos and supporting the needs of different domains.

Toby Burrows has been writing on the benefits of Semantic Web and Linked Data approaches in the humanities for a number of years. In a 2010 paper, he outlines the benefit in interconnectedness and linking and the need for a collaborative infrastructure for Medieval manuscript research (Burrows, 2010). In 2011 he talks about humanities more broadly, of semantic technologies offering "collaborative ways of building a shared body of semantically rich information" (Burrows, 2011,p.189). In his 2013 paper on HuNI, the Australian Humanities Networked Infrastructure, Burrows envisages a new kind of humanities research tool, which, unlike some existing digital resources, is based on "an understanding of humanities research processes, and of the nature of 'data' in the humanities" (Burrows, 2013, p.580) that can encourage "serendipitous discovery" (p.578). This 'data' is made up of "semantic entities and their relationships", readily supported in a Linked Open Data framework where the emphasis lies in both finding and analysing data, and in saving and sharing results (p.580). Progress is being made with HuNI, with a virtual lab prototype already available.

We have seen some evidence above that the digital resources and descriptive practices of cultural institutions are not always aligned with the needs of researchers (Burrows, 2011, p.189). This raises some important questions, which are worthy of further consideration. Michele Barbera provides a useful perspective, again expressing concern at the "lack of creative reuse of data both within the scholarly community and within the cultural industry" (Barbera, 2013, p.91) but seeing the solution, not solely in technical advances but firmly routed in "profound cultural shift" (Barbera, 2013, p.91).

There are two aspects to this cultural shift, one lies in breaking the tradition in cultural heritage and digital
humanities of "two-dimensional, paper-like thinking" (Barbera, 2013, p.91) and the widespread use of databases and spreadsheets where "people mentally model data in tabular structures" (Barbera, 2013, p.96) in favour of "Thinking in the graph' (Barbera, 2013, p.96, citing Berners-Lee, 2007). Shifting thinking from the tabular to the structured graph is vital if we are to make use of the "Giant Global Graph" (Berners-Lee, 2007). The other is around breaking down traditional notions of 'publication' and content production, where traditional 'consumers' of information are now "active producers of information" (Barbera, 2013, p.95). In this new world one "cannot know in advance how data will be used, combined, enriched and repurposed" (Barbera, 2013, p.95) which means data must be made available in universal formats, and optimised and contextualized at the point of consumption. For cultural institutions, this can be addressed more brutally: "don't waste resources trying to fix problems that don't exist" (Schaffner, Erway and OCLC Research, 2014, p.8).

With this democratisation of the production of information come issues of trust, provenance, quality and versioning (Barbera, 2013, p.94). But if those issues can be overcome, then there are riches to be gained. Maarten writes about linking historical entities with Wikipedia ('wikification') as a means of finding "common background information" for primary sources, increasing precision through disambiguation, improving recall by bringing together variant terms and "showing gaps in existing knowledge bases" (Maarten, 2014). DBpedia, the Linked Data version of Wikipedia, is a huge source of historical information, and semantically structured, but most scholars would acknowledge the need to verify and compare it's content.

This project aligns with Barbera's view that data production must become less centralized, due, not least, to the impossible task of creating a comprehensive dataset of the size needed, with ways of "slicing large datasets ... [to] reduce them down to a manageable size" (Barbera, 2013, p.92). We have also seen Burrows' discussion of a national infrastructure for the humanities (Burrows, 2013) yet he has also acknowledged the need for services to be "decentralized and distributed" (Burrows, 2010, p.127). All of this, in the context of Linked Data, should be happening on the Web. Crupi foresees libraries discovering that they can integrate information from other catalogues and third parties into their own structured information catalogues, with the future for the library catalogues as being "integrated into the web, queryable from it, able to speak and to understand the language of the web" (Crupi, 2013, p.35). Crupi also has something to say on the "radical transformation of the relationship between the user and the bibliographic universe":

"the integration of one's own data with those of other institutions not only increases their informative potential but renders them more complete, more usable and reusable, even in contexts very different from the original ... the explanatory clarity of the language used on the web makes the language of the library and the semantic tools it adopts for the classification and organization of knowledge less obscure and therefore more comprehensible to the user" (Crupi, 2013, p.36)

As we have seen, Linked Data is perceived to offer opportunities for bridging the gap between humanities scholars and cultural institutions, for democratising the production and consumption of information and for moving away from traditional silo's to a more distributed, collaborative web of data. As Barbera says: "in this novel scenario, individuals and organizations play at the same time the roles of information producers, gatekeepers, and consumers of information in an ever-reconfiguring ecosystem" (Barbera, 2013, p.95). This
echoes strongly the comments of Allen and his discussion of the long history of people as producers and consumers of information (2013).

This new future for access to digital resources is not without barriers and challenges. Barbera, again, notes that large investments in "research over semantic technologies ... has generated brilliant ideas ... unfortunately, the research community has not yet been able to leverage this potential within the industry to build production-ready tools easily usable by end-users" (Barbera, 2013, p.98). This lack of usable tools is important as the process of 'doing' Linked Data without such tools requires a a level of technical expertise that many humanities scholars, librarians, archivists and museum professionals simply do not have. Mia Ridge, Chair of the Museums Computer Group, provides a useful 2013 position piece on 'Where next for open cultural data in museums?' in which she draws out a number of potential explanations for the under-use of open cultural data: "possible reasons for the under-use of open cultural data include confusing or incompatible licenses, poor or inconsistent record quality within datasets, a lack of images or interesting descriptions, and undocumented or ambiguous vocabularies" (Ridge, 2013). In addition, Ridge discusses the tension between the need for more generic "easy-to-use datasets using common vocabularies for simple 'mash-up' style applications" and "more sophisticated data structures and specialised vocabularies to support internal uses, partnerships between museums, libraries and archives, or for use in research-led projects" (Ridge, 2013). This echoes Barbera, above, and others in the decentralisation of consumption, and stresses that we need approaches that meet a mixture of needs.

2.4 Semantics, Technology and Standards

Burrows makes an observation that there is a "tendency among commentators to assert that primary sources are the humanities researcher's data and therefore that primary sources ... and 'data' are one and the same thing" (Burrows, 2011, p.181, citing Borgman, 2007). This blurring of the distinction between 'data' and 'sources of data' makes establishing the nature of humanities data difficult and easily leads to semantically different and incompatible datasets. It is already difficult to define 'data' in the humanities, according to Burrows, because it is often made up of 'abstract entities' rather than physical data, incorporating annotation, commentary, transcription and more. This is important because to create online semantic services that support the researchers needs, we need machine-readable, compatible structured data. Burrows asks "how can the modelling which underpins scientific e-research environments be applied to the humanities?" (Burrows, 2011, p.182), pointing to work by Unsworth (2000) and others who have attempted to define such models. Such fundamental questions have not yet been fully examined in publishing Linked Data for the humanities and are by no means limited to the humanities. If there was a definitive definition of data, there are other semantic and structural questions to answer. One of these has been discussed by Summers and Salo (2013), Barbera (2013) and Crupi (2013). It is an issue that has plagued the Web architecture and Linked Data around "the use of URIs or IRIs to identify informational and non-informational resources, that is seldom accepted and understood, even within the experts community ... what mechanisms should be used to distinguish between statements about web pages and statements about the real world item or concepts the web page talks about" (Barbera, 2013, p.96). Put simply, there remains a lack of clarity in whether a URI refers to a real-world thing, or to a page of information about that thing. This iterates the importance of accurate, ontologically valid data.

As Summers and Salo acknowledge, such issues could be seen as small stumbling blocks, but they are, in
fact, significant. To take a simple example from Summers and Salo, if looking for a painting, we want to know the artist of the original and not the digital technician who created the scan. Yet both of these are 'real world things' (2013) and although a human user may be able to identify the inaccuracy, this becomes a problem when machines are trying to reason over data that contains such inconsistency and fuzziness. Summers and Salo propose "RDF vocabularies that recognize that representations, not resources, are transmitted between the client and server" (2013) but for the time being, this remains an issue for reasoning over datasets that may have taken disparate modelling decisions.

The use of different ontologies and modelling approaches has wider consequences too. It is the use of ontologies that, by providing 'semantic precision', allow inferential and deductive processing of data (Crupi, 2013), or expressed another way, allow us to ask research questions of data and expect valid answers across multiple sources. The Linked Open Vocabularies web site (Vandenbussche, n.d.) provides a useful summary of the Linked Data ontology space grouped by domain, although it is not exhaustive (for example, CIDOC-CRM 17, for cultural heritage collections, is missing). Looking at these groupings, where a sizeable group exists for science, there is no such equivalent for the arts and humanities. The Library space is well covered with a large grouping of relevant ontologies, this also includes archival and preservation ontologies.

Of the data sources surveyed for the current project, a small number of ontologies are regularly found: DBPedia, SKOS, FOAF and DC or DCTERMS, with CIDOC-CRM also significant, although less widely used. CIDOC-CRM's significance lies in the fact that it is the only one that is really domain specific, intending to provide "a model of the intellectual structure of cultural documentation in logical terms" (CIDOC CRM Special Interest Group, 2013, p. i). Work by Gergatsoulis et al. (2010) and Bountouri and Gergatsoulis (2013) has attempted to generate automatic mappings between other domain schemas and vocabularies and CIDOC-CRM - namely EAD, MODS and VRA - perhaps an acknowledgement that CIDOC-CRM is the most appropriate and appropriate ontology for cultural heritage 19.

In 2011, the Joint Information Systems Committee (JISC) funded two sets of projects around Linked Data 18. Amongst these was LOCAH, a project that released the entire Archives Hub catalogue as Linked Data. Writing about ontology selection for LOCAH, Pete Johnston (2011) selects some standard and widely used vocabularies, such as SKOS for knowledge organisation, FOAF for people and Dublin Core for describing 'resources' more generally. More specific to the historical domain, Johnston points to specific ontologies for describing events and biographical events and the project ultimately published a small set of archival description terms that could not be found in other ontologies. This 'mix and match' approach to modelling is efficient, re-using terms from existing schemes and it is the inferencing capability of RDF databases that allows for the use of multiple domain ontologies, as evidenced by LOCAH (Johnston, 2011). OpenART, again JISC funded 18 designed an ontology for the Artworld art historical dataset, a sample of which has been used in this project, taking what Mia Ridge would call a "sophisticated" (2013) domain-specific approach. The ontology for OpenART was based on Dolce Ultra Lite and Linked Open Events 20 with specialised terms to allow for rich description (Dow, 2012). This approach would allow the inferencing capability of RDF tools to infer, for example, that an artistic sale is also a more generic event, making use of the power of OWL. Whilst OpenART was an interesting demonstrator, it was time-consuming to create, and, with no ongoing support, has not been taken up or ratified by the community.
Other ontologies in the Library domain include MADS for name authorities, MODS, BIBFRAME and BIBO for bibliographic data and VRA for artworks (the latter only being experimentally available as RDF) yet these have not been found in the datasets used, and the author has found little coming directly out of humanities scholarship.

The most common ontology, by some distance, in those datasets surveyed, is the DBPedia Ontology, described as "a shallow, cross-domain ontology, which has been manually created based on the most commonly used infoboxes within Wikipedia" (The DBpedia Ontology (3.9), 2014). No literature has been found which discusses how to reconcile differences in modelling approaches, yet even a brief consideration of DBPedia and CIDOC-CRM highlight considerable variation. These issues will be considered in later sections. The ChartEx project, in it's ontology development phase, found the 'idiosyncratic' application of the XML standard for archival description, EAD, proved a barrier to effective modelling:

"ChartEx quickly encountered two problems with the English datasets. Of these the most significant was the lack of a standard approach to either the abridgment or the digitisation of the material. Even though both the Ward 2 and the Yarburgh collections used the standard markup language (EAD XML), there were radical differences in the design and provision of both metadata and data content." (ChartEx Project, 2014, p. 8)

The review, so far, has touched on technologies and standards for Linked Data, but it is worth providing a brief summary of this area. These technologies and standards are fairly well documented, in terms of the broad alignment around use of RDF, as the "grammar" of the Semantic Web and the "tripartite (triple) syntagmatic model" of subject - predicate - object (Crupi, 2013, p.27), URIs "as the names of things" (Summers and Salo, 2013 citing Berners-Lee, 2006), and ontologies as "a machine-readable conceptual map of a domain of knowledge" (Burrows, 2010, p.). For serialization, XML, Turtle and N-Triples are all widely supported, with the newer JSON-LD offering a more Web friendly alternative.

The other pieces of the broad infrastructure for Linked Data are data stores and query endpoints. It is beyond the scope of this project to compare solutions, but there is some useful comparative literature on native RDF databases (or 'triple stores') (see Bioontology Wiki, n.d., Clark, 2010, EUCLID Project, n.d., Garshol, 2012, Hashlofer et al., 2011), NoSQL databases (see Cudré-Mauroux et al., 2013, EUCLID Project, n.d.) or the map reduce approach of Apache Hadoop (see EUCLID Project, n.d.). This project will use a native RDF database, AllegroGraph, for the advantages expressed by the EUCLID Project:

"RDF graph based data model ...
Inference of implicit facts
Expressive query language (SPARQL)
Compliance to standards" (EUCLID Project, n.d.)

Looking at discovery, the ability to discover, access and query distributed datasets is central to Linked Data and to this project. SPARQL has become the de-facto standard for providing query endpoints. This
agreement on a standard approach is excellent for interoperability, but there are questions about the robustness of SPARQL endpoints. Dave Rogers, writing on his personal blog, sees SPARQL endpoints as "brilliant for hackdays, prototypes, experiments, toy projects etc. But I don't think anything 'real' could ever be built using one" (Rogers, 2013). Buil-Aranda et al., in their assessment of SPARQL paint "a mixed picture of the maturity of the SPARQL infrastructure: applications built on top of this infrastructure must cope with the intrinsic characteristic of imperfection and varying degrees of reliability that one often finds on the Web" (2013, pp.290-291).

Dix et al. see the use of semantic web and Linked Data as a potential solution to the user having to act as their own service - "web activity is glued together by the user, often through crude cutting and pasting between web applications" (2010). If the number of semantic services does increase, though, Dix et al. foresee challenges in achieving the "rapid switching between multiple contexts that we do (relatively) easily as human beings" (2010). Discovery mechanisms will face a range of issues, is both service discovery and scalability in trying to search across vast numbers of services, whilst trying to make sense of the questions posed by the user. Already the Linked Open Data cloud is "a compendium of more than 300 datasets and more than 31 billions triples" (Shekarpour et al., 2014). This is illustrated in Figure 1, below.

![Figure 1: The LOD Cloud](http://data.dws.informatik.uni-mannheim.de/lodcloud/2014/ISWC-RDB/extendedLODCloud/extendedCloud.png)

For the issue of scalability, García, Ruiz and Ruiz-Cortés (2012) propose a SPARQL-based solution for pre-processing queries in order to filter our unrelated services: "in a general scenario, our proposed filtering stage discriminates service descriptions depending on whether concepts referenced within their terms are present in the user request or not." (2012, p. 15). They also point to standards for providing service
descriptions, such as WSMO and OWL-S, highlighting that there is work to address discovery. Since their article was published, the W3C itself has also published a standard for SPARQL service description. Working in the same space, there is current research into extracting structured queries from natural language strings, for example the SINA (Semantically INterpreting user query towards question-Answering) and FRED projects.

Shekarpour et al., discusses the success of SINA's processing of natural language strings to identify the correct properties and objects therein. But the research highlights some issues too, which would almost certainly affect the asking of complex research questions in the humanities. Two examples stand out: the failure of the question 'Who created Goofy?' demonstrates that by 'typing' Goofy as a Person and having only Work as the domain of Creation, no answer can semantically be provided. One could argue that the domain of Creation is too narrow, or that the type given to Goofy is incorrect for a fictional creation. But real world complexity will inevitably create ambiguity and mismatch. Another example from Shekarpour et al. notes that "join the EU" should be mapped to the property "dpb:accessioneudate" (2014), but, in this case, the natural language expression is not likely to mirror the property required. Again, the real world does not match the ontological construct provided. The active research in these areas does hint at future potential both for reducing the size and range of searches prior to querying and also in extracting structured information from human derived questions.

2.5 Summary

In the context of this project, which is settled on taking a Linked Data approach, the literature confirms that there is evidence that Linked Data can benefit both cultural institutions in describing their collections and humanities scholars in the pursuit of their research. Indeed, there is a huge opportunity to use the existing Web architecture and move away from silos of content. There is also support for a "decentralized and distributed" (Burrows, 2010) approach, a 'personal desktop' where small-scale Linked Data clouds can contribute to the greater whole whilst filtering semantic sources and processing user queries to target them appropriately is the subject of current research. It is well established that personal names are one of a set of 'entities' that humanities scholars and information professionals seek to research and describe, and that a range of semantic sources already exist in this space. It has also been acknowledged that there is a "lack of creative reuse of data" and that this is, at least partly, the result of a lack of available tools (Barbera, 2013) but also, particularly for scholars, a mismatch between their research needs and the digital resources available. It is into this landscape that this project hopes to propose some ways forward.

However, there are a variety of issues and barriers.

Entering into this way of working demands a culture shift away from 'tabular' modelling of data towards "thinking in the graph" (Berners Lee, 2007) and a type of social behaviour more in line with Web 2.0 notions of folksonomies and crowdsourcing than traditional scholarship and cataloguing. Asking users to freely share data might not sit comfortably with "digital humanists [who] are fiercely independent" (Schaffiner, Erway and OCLC Research, 2014, p.7) and there is work to do to establish trust, in terms of provenance (where did the data come from?) and in versioning (how do I trust this information is up to date?). Scholars and information professionals must be able to make their own reasoned judgements about data quality.
We have seen evidence of a disparate array of resources and it is vital that we avoid the risk that the "user becomes overwhelmed and lost within the data" (Blanke et al., 2012) and also that the domain ontologies achieve a level of consistency that prevent mis-matches and inaccuracies from drowning out, or missing completely, valuable data.

Technical challenges persist too, from endpoint discovery and range of query standards in use to reliability of services. Yet there are opportunities to make better use of existing Web infrastructure, for example by using "HTML's metadata facilities as well as the newer Microformats, Microdata, and RDFa specifications [that] provide the mechanics for expressing metadata in Web pages themselves" (Summers and Salo, 2013). These more lightweight standards may help with the easier production of tools. Web Data Commons, for example, has published a huge dataset of RDFa, Microdata, and Microformat Data scraped from across the web and Schema.org offers an ontology for the web albeit one that "is not sufficient for many library-specific information exchange use cases" (Cole et al., 2013).

Linked Data is not the only approach available, and much of the discussion above is heavily rooted in the involvement of users as creators, annotators, curators and publishers of data. Summers and Salo allude to the success of more RESTful, API-driven approaches, hinting that the Linked Data paradigm may have become rather too focussed on formats and semantics - "it is significant that Fielding's pluralistic notion of Web architecture as an architectural style has been quietly, much more successful in guiding the design of Web applications and services, than prescriptive and often dogmatic rules about what format to publish data in, and how to name entities with URLs" (Summers and Salo, 2013).

Entity recognition, text mining and natural language processing are all techniques that can provide researchers with large corpuses of processed data to work on. This intersection between researcher and machine, in how far a machine can be trained to work on behalf of researchers to ask research questions of large scale datasets is particularly interesting, as is seeing how far these different approaches can work together, but those must remain out of scope for this project. Schaffner, Erway and OCLC Research remind us that "the sum total of cultural heritage digitization programs is chaotic and ill-suited to serve scholars" (2014, citing McGann, 2011) and the question of how we bring these existing data sources into being 'suitable' is another major issue for future consideration.

1. If page numbers are omitted, there was no pagination. ↩
2. For example, in the UK the Joint Information Systems Committee (JISC) have funded many digitisation and digital resource projects http://jisc.ac.uk ↩
3. VIAF (Virtual International Authority File) http://viaf.org/ ↩
4. LODLAM http://lodlam.net/ ↩
5. OpenGLAM http://openglam.org/ ↩
6. LC Linked Data Service: Authorities and Vocabularies http://id.loc.gov/ ↩
8. Getty Vocabularies as Linked Open Data http://www.getty.edu/research/tools/vocabularies/lod/ ↩
9. Pelagios: Linking together the places of our past through the documents that refer to them [http://pelagios-project.blogspot.co.uk/](http://pelagios-project.blogspot.co.uk/)


14. For example, JSON for Linking Data [http://json-ld.org/](http://json-ld.org/)


22. See [http://www.loc.gov/ead/](http://www.loc.gov/ead/)

23. SPARQL [http://www.w3.org/TR/sparql11-query/](http://www.w3.org/TR/sparql11-query/)

24. See [http://lists.w3.org/Archives/Public/public-lod/2014Jul/0125.html](http://lists.w3.org/Archives/Public/public-lod/2014Jul/0125.html) - post to the public-lod mailing list and various ongoing discussions in the archives for July and August 2014 [http://lists.w3.org/Archives/Public/public-lod](http://lists.w3.org/Archives/Public/public-lod)

25. See [http://www.w3.org/Submission/WSMO/](http://www.w3.org/Submission/WSMO/) and [http://www.w3.org/Submission/OWL-S/](http://www.w3.org/Submission/OWL-S/)

26. See [http://www.w3.org/TR/sparql11-service-description/](http://www.w3.org/TR/sparql11-service-description/)

27. See [http://aksw.org/Projects/SINA.html](http://aksw.org/Projects/SINA.html) and [http://wit.istc.cnr.it/stlab-tools/fred](http://wit.istc.cnr.it/stlab-tools/fred)


29. See [http://schema.org](http://schema.org)

30. For example the ChartEx project has produced a historian's workbench based on entities extracted through machine processing [http://www.chartex.org/](http://www.chartex.org/)
3. The Research

"So far most of these web-like or web-based applications are separate, and web activity is glued together by the user, often through crude cutting and pasting between web applications."

(Dix et al., 2010)

Consider the historical researcher whose work requires them to consult a wide range of historical sources in order to build up a picture of their particular research question both geographically and over time. This scenario is common, but applications to support this in the digital world, are not.

The project, Semantics-based Indexing of Historical Entities, is an exploratory software development project that has built a prototype application for researchers to address two research questions:

1) Would a tool to help research and further describe historical names benefit its target audience of researchers and cultural heritage professionals?

2) Can current standards, technologies and linked data services support, and add value to, the proposed service?

The semantic technologies of chief interest to the project are: RDF as the underpinning data model, HTTP URIs linking 'things' together and ontologies to bring meaning and structure to the data. For querying, SPARQL is the de facto standard and an RDF store sits at the back offering stable storage for the data. Wherever possible, existing applications and libraries will be used rather than writing new functionality. These have been identified through the literature and technology research and will be explored in more detail later in this section.

The research has essentially all been carried out through building the application, and the steps of doing this will be explored in the sections below.

3.1 User Requirements

The first step was to find out what users wanted.

A user group of 8 people including researchers and information professionals were asked to help answer research question (1). Further information about the user group can be found in Appendix 1.

For the requirements, a descriptive survey was used to help prioritise features. This was a positivistic exercise to produce a ranked list. These were analysed into a list of essential and desirable features, shown below.

Further information on this survey can be found in Appendix 3.

<table>
<thead>
<tr>
<th>Requirement</th>
<th>Status</th>
<th>Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>See exact matches and likely matches</td>
<td>Essential</td>
<td>1</td>
</tr>
<tr>
<td>Query multiple sources</td>
<td>Essential</td>
<td>1</td>
</tr>
<tr>
<td>Decide which sources should be searched</td>
<td>Essential</td>
<td>2</td>
</tr>
</tbody>
</table>
Users were also asked about attitudes to sharing data.

5 of 8 responses selected 'I want to choose what I share'; only one did not want to share at all.

An option of 'I don't understand what this means' in the list of features helped point out issues around terminology and understanding that could affect the usability of any application.

Two questions received 'I do not understand' responses from two people. Those were 'I want to add local information elements, drawn from a list of options' and 'I want to add custom local information elements'. This was seen by the author as an important feature and thus identifies an area where the description used did not fully explain the concept to the user group. On reflection, this could have been expressed in a much clearer way as 'I want to add information I have gathered in my research on a selected person', perhaps with examples to aid understanding.

### 3.2 Research Questions

The phenomenological element were questions that elicit qualitative and subjective responses, such as 'what are your research questions' and 'why won't this work?'. Whilst the latter question did not offer much insight, arguably the most useful data in the user requirements survey is the set of research questions shown in full in Appendix 3b. The range of these, from such a small sample, demonstrates how varied research in the humanities is. The scope of the current project cannot seek to answer all of these in full, but it highlights some useful repeating themes. In particular, name disambiguation and comparison and the need to gather...
sufficient information about a person to make a reasoned judgement comes up several times. There are other themes around: tracing the origin and history of names, tracking people as they move, tracking name changes and tracing the evolution and movement of occupations. These common themes have informed the 'Explore' features described below.

3.3 Technology and Design

In order to explore the research questions, a prototype application has been built\(^1\), intending to deliver essential features 1-4 in the user requirements, and also to explore some further ideas drawn from the user-supplied research questions.

The essential user requirements (1-4) can be summarised thus:

- **Query**
  - Search across multiple sources
  - Select which sources to search
- **Results**
  - Search specific fields, e.g. forename, surname, occupation, birth date
  - See exact matches and likely matches
  - Filter search results, e.g. by date range
- **Explore**
  - View full information about each name
  - Follow links from people to other people, and to other entities
- **Create**
  - Local 'authority record' for names where no external authority is identified
- **Sharing**
  - Store sharing preferences

These are illustrated in the use cases diagram in figure 2.
Figure 2: Use Case Diagram

The following user activity diagram shows the flow of activity through the system (figure 3).
The class diagram in figure 4a provides the original outline structure for the application. It has been designed along a broadly MVC pattern, with views provided by the html templates, the Python Flask .py files being the controllers and data being handled by the model classes. The classes, in particular those handling queries and data will be designed to exist as standalone Python modules in order that they can be re-used by others. The intention is not to duplicate functionality available in other Python modules, but to create mechanisms to build up sophisticated queries for the given domain.
Figure 4a: Class Diagram (Proposed)

A second class diagram (figure 4b) shows the final application design, after re-factoring. Those items highlighted in red have not been implemented but still form part of the design.

Figure 4b: Class Diagram (Final)

Taking a simple user-entered query as an example the following is the flow of code through the system.

1. View: user enters query into query box.
2. Controller: the Flask Python blueprint takes entered data and any options and instantiates the necessary model object(s) with information about SPARQL services to be searched and query.
parameters.
3. Model: the query object performs the query and returns the results to the controller.
4. Controller: receives the results and processes them.
5. View: displays the results.

The core software is written in Python and uses the Flask Web development 'microframework' for the interface. Python was initially chosen above Java in the hope that it would be quicker to write the software. Flask was recommended as a lightweight approach to building web sites with Python, rather than the more fully-featured, but complex, Django. An initial prototype was built to ensure Python, and Flask, could support the needs of the project, in particular that their were libraries to support SPARQL and RDF. It has proved to have been a good choice, although it introduced a learning curve that threw in some challenges.

Test driven development (TDD) has been used to ensure all new classes can be tested prior to deployment. TDD is methodology by which tests are written prior to any code. The tests initially fail and then code is written to make them pass. This approach demands discipline as it requires the developer to code more thoughtfully, working out what the code is meant to achieve before writing it. By writing a suite of tests in advance, many bugs and errors should be avoided as small chunks of code must have passed their tests before before being committed into the larger application and makes testing part of development. TDD will be used in this project to achieve fully tested working code throughout the development lifecycle. Each development iteration will start with defining the purpose of that iteration and then writing tests. Tests are available for all new Python classes in a /tests directory.

Standalone Python classes have been written to handle querying and exploring data, and data storage. Exiting existing Python modules are used wherever possible. A design decision to create a layer of abstraction between code driving the interface and code dealing with data and querying means that at no point does the Flask blueprint code interact with any third party Python library, there is always a local class handling this interaction. This is to enable easy replacement of libraries as needed.

### 3.3.1 Python Libraries

Three key Python libraries have been used:

- **SPARQLWrapper** provides a wrapper around a SPARQL service and helps with creating the query and converting the result to a manageable format. It is part of the RDFLib group of python rdf utilities.

- **rdflib** is a library for working with RDF data. It offers support for loading and saving RDF and for merging and navigating RDF graphs. It is part of the RDFLib group of python rdf utilities.

- The AllegroGraph Python API is a Python client API to the AllegroGraph RDFStore. It supports connecting to, querying, creating, updating and deleting items in an AllegroGraph store.

### 3.3.2 Data Store

The data store is AllegroGraph 4.13. The software provides an unlimited term free license for use with up
to 5,000,000 triples. There is a freely available Python client that provides a range of functionality to interact with the data store.

It was not an aim of this project to assess data store options. AllegroGraph was selected for several reasons: because of the free license, the python client and the support for access controlled repositories (to support the sharing requirement noted above). Throughout the development it has been fast and responsive and easy to work with through the web GUI and client code.  

AllegroGraph is written in LISP. Although this could have been a barrier to more complex use of the store, it has not proved to be a problem for this project.

### 3.3.3 Code Overview

Although it is not the intention of this report to discuss, in detail, the python code, the following provides a brief summary of the structure of the application.

```plaintext
peoplesparql.py
/templates/ - html templates
  500.html - template for the HTTP error code 500 page
  base.html, used by all other templates
/static/ - static files used across the application, including images, javascript and css files
/home/ - a flask blueprint for site home and HTTP error code 404 page (page not found)
/templates/
  index.html
  404.html
/project/ - a flask blueprint for project pages
/research/ - a flask blueprint for the 'research' element of the site
/templates/ - templates for each different page
  create.html
  explore.html
  query.html
  research.py
  buildresults.py
  sameperson.py
/userinteraction/ - unused, intended to provide a html template and python code to show a users own submitted data
/datawrangler/
  /tests/ - tests for each of the following classes
    connect.py - connect to the AllegroGraph store
    add_triple.py - add a single triple
    delete_triple.py - delete all triples from a named graph
/queryandexplore/
  /tests/ - tests for each of the following classes
    SPARQL service.py - initiates the setup a new SPARQL service, or retrieval of information for an existing one in the datasources repository
    SPARQL service_creator.py - does the grunt work of setting up the SPARQL service
    SPARQL service_list.py - returns a list of SPARQL services in the datasources repo
    opensearchquery.py - unused, experiment with opensearch queries
    process_rdf.py - given a URI get the RDF from it and return the graph
    sparql_ask.py - perform a SPARQL ASK query using SPARQLWrapper
    sparql_select.py - perform a SPARQL SELECT query using SPARQLWrapper
    sparql_query_specialised.py - build the query to pass to the above ask or select classes
    sparql_query_private.py - perform a query on a local access, controlled data store (using the AllegroGraph Python API)
/franz/ - the allegrograph python client (third party library)
/flaskext/
  kvsession.py - third party library used to store session variables that are larger than those normally supported by browsers
other third party libraries included in the python path
  rdflib
  SPARQLWrapper
```

Julie Allinson: Semantics-based Indexing of Historical Entities 27
The full code for the project is available in github.

### 3.3.4 Useful Tools

Two tools proved particularly useful during development:

Olivier Cuenot's Json Parser Online was used for checking JSON output, and Laurens Rietveld's YASGUI (Yet Another SPARQL GUI) was incredibly useful to testing SPARQL queries against different SPARQL services.

In addition, SPIN is a way of representing SPARQL rules and constraints. It has several specifications, including the SPIN SPARQL Syntax, a W3C Member Submission. Although SPIN has not been directly used in the project, reading up on SPIN and it's high-level syntax, and experimenting with SPIN in the free version of TopBraid composer helped build an underspanding of SPARQL's inferencing capabilities.

### 3.4 Building the application

This section illustrates the design and functionality of the application. The focus of the application is on querying for personal names. To limit SPARQL queries to search for personal names only, it was necessary to establish two pieces of information: 1) the class used to indicate that a resource is a Person, and 2) the property that would contain a human-readable label for the personal name and therefore be the target of the search.

Given that RDF is the core semantic data model underpinning the semantic Web, the RDF Schema is the most appropriate source of this information, and it provides us the property whose object would be the class for person:

```
 rdf:type is an instance of rdf:Property that is used to state that a resource is an instance of a class [^15]
```

To find the actual class, we need to query the RDF dataset or SPARQL service directly. For example, the following query would provide a list of all classes used in a dataset, for example the British National Bibliography (BNB) dataset.

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
SELECT DISTINCT ?o WHERE
{ ?s rdf:type ?o }
```

If this query is performed against the BNB's SPARQL service, a list of 28 classes is returned, including one for Person: [http://xmlns.com/foaf/0.1/Person](http://xmlns.com/foaf/0.1/Person). The next section will show how this information is gathered and used by the application.

RDF also suggests a standard property for "a human-readable version of a resource's name" (Brickley and
rdfs:label is an instance of rdf:Property that may be used to provide a human-readable version of a resource's name [^15]

For datasets where Simple Knowledge Organization System (SKOS)\(^{17}\) is used, the skos:preferredLabel, a subProperty of rdfs:label, might be used.

These label properties for the target of the personal name search.

### 3.4.1 SPARQL Service Setup

SPARQL \(^{18}\) is the query language for RDF. Many RDF datasets offer a SPARQL 'endpoint' or SPARQL service to enable external querying of the data. SPARQL is the central query technology for the application and in order to make the right queries the application needs to know something about the SPARQL service and, indeed, whether it supports SPARQL. The key information required is as follows:

1. SPARQL version supported (1.0 or 1.1) - to ensure valid queries are used whilst allowing the code to use the enhancements in 1.1 where appropriate \(^1\).
2. What descriptive label is used by the SPARQL service (as noted above). In most cases this will be rdfs:label, but there may be cases where this is not used.
3. What rdf:type value is used to identify Person entities (as noted above). So far, eleven different possibilities have been identified.

There are recommended standards for providing information about SPARQL services \(^{19}\) and for describing datasets \(^{20}\), but after some testing it became clear that the availability of this information cannot be guaranteed. This led to the following design decision.

**DESIGN DECISION** Provide code that will analyse a SPARQL service and store details in the local datasources repository \(^{21}\). If the service is already listed in the repository, the code will return information about it.

For ontological correctness, it is necessary to store information about the SPARQL service (the SPARQL service address and SPARQL version) and also about the dataset itself (the types and labels).

A small data modelling exercise identified the properties and objects needed for the project. Terms were selected from existing schemes, namely the Vocabulary of Interlinked Datasets (VoID) \(^{20}\) and the SPARQL 1.1 Service Description \(^{19}\). Two properties have been newly defined.

#### 3.4.1.1 Properties from existing ontologies

To describe the dataset:

- rdf:type void:Dataset
- void:sparqlendpoint endpoint-uri # links together dataset and endpoint

To describe the SPARQL service:
3.4.1.2 New properties

Property to describe the rdf:type used for Personal Names

```plaintext
Property: dataset:typeForPersonalName
Description: dataset:typeForPersonalName is an instance of rdf:Property that may be used to indicate to class for 'persons' used in a given dataset
URI: http://geekscruff.me/ns/dataset#typeForPersonalName
Domain: void:Dataset
Range: rdfs:Class
Superproperty: rdf:Property
```

Property to describe the standard label used across all resources

```plaintext
Property: dataset:labelForPersonalName
Description: dataset:labelForPersonalName is an instance of rdf:Property that may be used identify the rdf:Property used to provide a human-readable version of a resource's name in a given dataset
URI: http://geekscruff.me/ns/dataset#labelForPersonalName
Domain: void:Dataset
Range: rdfs:Property
Superproperty: rdfs:Property
```

3.4.1.3 Example

The following example describes DBPedia. It is presented in the Turtle (ttl) serialisation format for RDF.

```turtle
@prefix void: <http://rdfs.org/ns/void#> .
@prefix dataset: <http://geekscruff.me/ns/dataset#> .
@prefix sd: <http://www.w3.org/TR/2013/REC-sparql11-service-description-20130321/#> .
@prefix foaf: <http://xmlns.com/foaf/0.1/> .
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .

<http://dbpedia.org/sparql> a sd:Service ;
    sd:endpoint <http://dbpedia.org/sparql> ;
    sd:supportedLanguage sd:SPARQL11Query .

<http://geekscruff/ns/datasetservicedby#http://dbpedia.org/sparql> a void:Dataset ;
    void:sparqlSPARQL service <http://dbpedia.org/sparql> ;
    dataset:typeForPersonalName foaf:Person ;
    dataset:labelForPersonalName rdfs:Label .
```

The dataset URI is stored in this localised form:

```plaintext
http://geekscruff/ns/datasetservicedby#http://dbpedia.org/sparql
```

This is used partly for the convenience of a predictable format, but also because it is not always possible to get the a formal identifier for the dataset itself and this format avoids storing incorrect or inaccurate data.

3.4.1.4 SPARQL services

The following SPARQL services have been setup in the application. Each has a SPARQL endpoint. Appendix 2 lists further endpoints, datasets and services that, for a variety of reasons, were not added.

- Artworld (http://artworld.york.ac.uk/): sample person data from a History of Art research dataset
- British Museum Collection Catalogue: the full collection catalogue of the British Museum
- DBPedia: a very generic resource, the Linked Data version of Wikipedia
3.4.1.5 Adding a new SPARQL service

Each SPARQL service needs to be set up in the AllegroGraph datasources repository. In order to set up an endpoint, the code first checks that the SPARQL service isn't already in the datasources repository by issuing an ASK query. If it is not setup, the next step is to gather the necessary information. For this, the first step is to issue the following ASK request to the external SPARQL service url. If the endpoint URI is not a SPARQL service at all, the ensuing exception is caught by the application and no further steps are taken.

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX foaf: <http://xmlns.com/foaf/0.1/>
ASK { ?s rdf:type foaf:Person }
```

As ASK is not supported in SPARQL 1.0, the application uses this to determine whether the service supports SPARQL version 1.1. If an error is returned, the application assumes the SPARQL service supports only SPARQL 1.0.

At this point, the properties described earlier can be stored.

The next step is to check what rdf:type value is used for Persons and what label is used. The code sends several further queries to determine these properties and stores the information as:

- dataset:typeForPersonalName
- dataset:labelForPersonalName

The list of potential types and labels checked has been manually assembled through assessing a range of endpoints. A limitation, then, is that the code would not pick up a new person type or label.

When a query is made later to that particular SPARQL service, this stored information will be used.

The fragment from the full INFO level logging from the application in figure 5 gives a sense of the different steps. The QueryBadFormed error is the one that indicates the ASK query is not supported. A true of false value is expected.
Figure 5: Info level logging to show SPARQL service setup

The datasources repository is publicly available and can be viewed and queried via the allegrograph web interface.

3.4.2 Query

To query, the user simply has to enter a personal name into the search box and select the SPARQL services to search.

This simple approach was taken based on the understanding that the user will enter personal names only. Other terms would, of course, be searched for but limited results would be returned.

**DESIGN DECISION** Keep querying simple to allow more time for the more semantic 'Explore' features.

Based upon what the application knows about each SPARQL service, the search will differ as shown in the examples below.

3.4.2.1 Examples
This query would be used for DBPedia. The SPARQL service setup has identified three Person classes in use in the dataset and has confirmed that rdfs:label is used. Therefore the query has three parts to query for the term in conjunction with each of  the three Person classes. These are then combined with UNION and DISTINCT to ensure that only the unique results are returned.

Please note 'a', in the following query is a valid shorthand for rdf:type.

```sparql
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema>
PREFIX foaf: <http://xmlns.com/foaf/0.1/>
PREFIX dbo: <http://dbpedia.org/ontology/>
PREFIX sc: <http://schema.org>

SELECT DISTINCT ?s ?o WHERE
{
  { ?s rdfs:label ?o .
    ?s a dbo:Person .
    FILTER ( (regex(?o, " t","i") || regex(?o, "^t","i")) &&
            (regex(?o, " e","i") || regex(?o, "^e","i")) &&
            (regex(?o, "lawrence","i")) ) FILTER (langMatches(lang(?o),"en" ))
  } UNION
  {
    ?s rdfs:label ?o .
    ?s a foaf:Person .
    FILTER ( (regex(?o, " t","i") || regex(?o, "^t","i")) &&
            (regex(?o, " e","i") || regex(?o, "^e","i")) &&
            (regex(?o, "lawrence","i")) ) FILTER (langMatches(lang(?o),"en" ))
  } UNION
  {
    ?s rdfs:label ?o .
    ?s a sc:Person .
    FILTER ( (regex(?o, " t","i") || regex(?o, "^t","i")) &&
            (regex(?o, " e","i") || regex(?o, "^e","i")) &&
            (regex(?o, "lawrence","i")) ) FILTER (langMatches(lang(?o),"en" ))
  }
}
ORDER BY ?o
```

For SPARQL services that support inferencing the UNION of results shown above may not be necessary, but the current code does not assess this.

These examples are based on a search for 'T.E. Lawrence'. The fragment above shows how this term has been split up so that the search looks for instances where T is the first or second initial (or letter of the name) in combination with E being the first or second initial (or letter of the name). Another example from the Archives Hub could be a search for DH Lawrence - this returns results for DH Lawrence and George Heneage Lawrence Dundas. The current implementation attempts to strike a balance between not excluding relevant results and not overwhelming the user.

The following example query to the British Museum SPARQL service shows that the application has identified only one Person type: http://erlangen-crm.org/current/E21_Person.

```sparql
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema>
PREFIX ecrm: <http://erlangen-crm.org/current/>

SELECT DISTINCT ?s ?o WHERE
{
  ?s rdfs:label ?o .
  ?s a ecrm:E21_Person .
  FILTER ( (regex(?o, " t","i") || regex(?o, "^t","i")) &&
            (regex(?o, " e","i") || regex(?o, "^e","i")) &&
            (regex(?o, "lawrence","i")) )
}
ORDER BY ?o
```
Results look like this in the application and are grouped by endpoint:

![Query interface](image)

Users can then tidy up the results set by discarding any unnecessary results before moving to the explore step.

The application makes widespread use of sessions to allow for passing sets of results around. The kvsession Python library was used to get around the issue of long results lists being too large for the standard session size used by browsers.

### 3.4.3 Explore

At the outset of development, this area was not fully defined, but had the broad intention of allowing the user to 'explore' results in a meaningful way.

---

© 2014 by Julie Allinson.

Figure 6: Search Results
Given the project focus is on semantic technologies, it seemed best to dig deeper into the added value these bring. To do this, three experimental features were created and will be discussed in more detail below.

Given that AllegroGraph was proving to be a fast and reliable data store, that code was readily available to add, update and query data and that it had in-built support for inferencing, it was decided to create a temporary store that would hold the results of a user's query and exploration steps. To facilitate this, the user needs to be logged in, something that they would have to do if storing data and deciding on sharing settings. Login was achieved very simply by using Mozilla's Persona 'login with email' service via the Flask persona example made readily available in the Flask github account. This provides the application with the user's email address which is used as a key for the user throughout.

**DESIGN DECISION** Store a local copy of all queried and explored data in a temporary repository within a named graph for the logged in user.

### 3.4.3.1 Are these the same person?

Within the user research questions, there were a number around name disambiguation and comparison - are these the same person? This feature attempts to experiment in helping to answer this question by comparing two selected names.

The application tries to establish whether two names are the same person, or not, and with what degree of confidence. If there are more than two results, only the first two will be checked, a limit imposed for simllicity - extending this to cover any number of results is relatively trivial, although there would be potential performance issues to consider with large numbers of results. For example, 2 results requires one comparison, of result 1 to result 2, but 6 results would require 23 comparisons to cover, rising exponentially. If the application cannot determine the similarity due to lack of information it will try and extract related information.

In the Semantic Web space, OWL, the Web ontology language, provides a mechanism for explicitly asserting that two resources are the same.

```
resource A <http://www.w3.org/2002/07/owl#sameAs> resource B
```

The first step is to look for explicit owl:sameAs links between the two resources. The application also looks for http://dbpedia.org/ontology/wikiPageRedirects links as this is a DBPedia specific mechanism for redirecting to the canonical DBPedia resource, and serves a broadly similar function as owl:sameAs.

If a link is found, then we can say for certain that they are the same person.

If not, a second step looks at the names themselves and using a small python library called name_tools matches the names and retrieves a score.

For example:

```
D H Lawrence compared to George Heneage Lawrence Dundas: score = 0.0
D H Lawrence compared to D H Lawrence: score = 1.0
D H Lawrence compared to David Herbert Lawrence: score = 0.8
```
The score is used as an indicator, but even an exact match (score 1.0) does not really confirm they are the same, so the next step looks for birth and death dates.

This rather complex query looks for a range of common birth and death date properties, treating death as optional, in case the person is still alive and also checking whether the birth or death objects are a literal or a URI.

```
PREFIX hub: <http://data.archiveshub.ac.uk/def/> 
PREFIX dbp: <http://dbpedia.org/property/> 
PREFIX dbo: <http://dbpedia.org/ontology/> 
PREFIX ecrm: <http://erlangen-crm.org/current/> 
PREFIX yago: <http://yago-knowledge.org/resource/> 
PREFIX rd: <http://rdvocab.info/ElementsGr2/> 

SELECT ?b ?d ?u {
    { 
        OPTIONAL { <http://collection.britishmuseum.org/id/person-institution/144185> <ecrm:P100_died_in> ?d . } 
        <http://collection.britishmuseum.org/id/person-institution/144185> <ecrm:P98i_was_born> ?b . 
        BIND (isURI(?b) as ?u) 
    } 
    UNION 
    { 
        OPTIONAL { <http://collection.britishmuseum.org/id/person-institution/144185> <hub:dateDeath> ?d . } 
        BIND (isURI(?b) as ?u) 
    } 
    UNION 
    { 
        OPTIONAL { <http://collection.britishmuseum.org/id/person-institution/144185> <dbp:dateOfDeath> ?d . } 
        BIND (isURI(?b) as ?u) 
    } 
    UNION 
    { 
        OPTIONAL { <http://collection.britishmuseum.org/id/person-institution/144185> <dbo:deathDate> ?d . } 
        BIND (isURI(?b) as ?u) 
    } 
    UNION 
    { 
        OPTIONAL { <http://collection.britishmuseum.org/id/person-institution/144185> <yago:diedOnDate> ?d . } 
        <http://collection.britishmuseum.org/id/person-institution/144185> <yago:wasBornOnDate> ?b . 
        BIND (isURI(?b) as ?u) 
    } 
    UNION 
    { 
        OPTIONAL { <http://collection.britishmuseum.org/id/person-institution/144185> <rd:dateOfDeath> ?d . } 
        BIND (isURI(?b) as ?u) 
    }
}
```

If the dates are provided as literals, the code assembles a date span and this is used for the comparison. if the dates are URIs, the code will request the RDF from the URI and look for an rdfs:label in the returned graph, repeating the process if only a URI is returned, until a literal value is found. In the case of the British Museum, as per the following example, this process happens twice.

```
@prefix ecrm: <http://erlangen-crm.org/current/> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .

<http://collection.britishmuseum.org/id/person-institution/144185> 
<ecrm:P98i_was_born> <http://collection.britishmuseum.org/id/person-institution/144185/birth> .

<http://collection.britishmuseum.org/id/person-institution/144185/birth> 

<http://collection.britishmuseum.org/id/person-institution/144185/birth/date> 
rdfs:label 1948 .
```

Fetching the RDF from a given URI uses the rdflib python library. When given a URI this will retrieve the RDF, or return an exception if it cannot, and then provides an in-memory graph for further querying. In the case above, for dates, the application takes the graph and looks for an rdfs:label. In other cases, the
application will take the RDF graph and add it to the AllegroGraph temporary repository.

The following fragment of python code shows a simple example of processrdf in use. Here the application uses rdflib, via a local class called process_rdf, to retrieve a graph from a URI (s) and then adds the results of a SPARQL query against that graph to a list object. The query itself is asking for the URI and, if available, the rdfs:label for the given URI. This fragment would be wrapped inside a try/catch block to handle any errors throw, for example by an invalid URI.

```python
g = process_rdf.ProcessRdf().fromuri(s)
resultslist = g.query(''
    select distinct ?s ?l
    {
        <' + s + '> ?p ?o .
    OPTIONAL { <' + s + '> rdfs:label ?l }
    }
)''
```

The following figures show two examples, one where the application is confident the names are not the same and the other where it believes they are. The layout here is designed to provide a summary of the process, rather than be truly user friendly.

**Are these the same person?**

Please note, this is a proof of concept feature and currently only looks at the first two names.

**Fairly sure NO**

**Details**


match-sameas
There is no explicit sameas relationship between Marie Curie (http://collection.britishmuseum.org/id/person-institution/146156) and Marie Curie Cancer Care (http://collection.britishmuseum.org/id/person-institution/144185). Investigate further ...

match-name
confidence=0
The names are different, indicating that these are not likely to be the same person. Checking dates to be sure ...

match-dates
confidence=0
The dates do not match ([u'1867-1934'] and [u'1948-Active']). These are not likely to be the same person.

**Figure 7a: These names are not the same**
Are these the same person?

Please note, this is a proof of concept feature and currently only looks at the first two names.

Less sure YES

<table>
<thead>
<tr>
<th>Details</th>
</tr>
</thead>
<tbody>
<tr>
<td>match= <a href="http://collection.britishmuseum.org/id/person-institution/35418">http://collection.britishmuseum.org/id/person-institution/35418</a> and <a href="http://dlib.york.ac.uk/id/person/34385">http://dlib.york.ac.uk/id/person/34385</a></td>
</tr>
<tr>
<td>match-sameas</td>
</tr>
<tr>
<td>There is no explicit sameAs relationship between Sir Peter Lely (<a href="http://collection.britishmuseum.org/id/person-institution/35418">http://collection.britishmuseum.org/id/person-institution/35418</a>) and Lely, Sir Peter (1618-80); owner of prints/drawings; painter; seller of pictures; Nationality: Dutch (<a href="http://dlib.york.ac.uk/id/person/34385">http://dlib.york.ac.uk/id/person/34385</a>). Investigate further ...</td>
</tr>
<tr>
<td>match-name</td>
</tr>
<tr>
<td>confidence=0.75</td>
</tr>
<tr>
<td>The names are alike. Checking for dates ...</td>
</tr>
<tr>
<td>match-dates</td>
</tr>
<tr>
<td>Neither of the names have associated dates. Checking for sameas links to other sources ...</td>
</tr>
<tr>
<td>match-dates</td>
</tr>
<tr>
<td>No dates were found for Sir Peter Lely (<a href="http://collection.britishmuseum.org/id/person-institution/35418">http://collection.britishmuseum.org/id/person-institution/35418</a>)</td>
</tr>
<tr>
<td>sameas-external</td>
</tr>
<tr>
<td>There are sameas links in <a href="http://dlib.york.ac.uk/id/person/34385">http://dlib.york.ac.uk/id/person/34385</a></td>
</tr>
<tr>
<td>match-dates-external</td>
</tr>
<tr>
<td>dates found in <a href="http://viaf.org/viaf/47033515">http://viaf.org/viaf/47033515</a>. Check if they match.</td>
</tr>
<tr>
<td>match-dates</td>
</tr>
<tr>
<td>These dates were found: 1618-1680 but there is nothing to compare them to</td>
</tr>
</tbody>
</table>

Figure 7b: These names are the same

Dates are not always present and so the process often fails to give a satisfactory answer. Further steps could be added, for example to look for linked people who have links to both of the selected people, or similarly for linked names. Matches in occupations or roles, works created, awards received and so on would also be useful to analyse. The potential for these approaches is explored in the 'enhance with semantics' feature below.

Enhance from external sources

The next feature is rather simpler. When querying for information about each of the selected names, the application looks for any owl:sameAs links. If it finds any, these are listed separately and a button called 'Enhance from external sources' appears. The Artworld sample dataset, for example, contains 'sameAs' links to person records in the Virtual International Authority File for Names (VIAF). Once selected, this feature will retrieve RDF from the 'sameAs' link, add it to the temporary data and list it on the explore page itself.
This feature offers a simple demonstration that by using owl:sameAs, the process of linking resources becomes much easier and opens up more information about a person, or other class.

### 3.4.3.2 Enhance with semantics

The final option is called 'Enhance with semantics'. This uses a basic level of inferencing to draw together entities of a similar type. Selecting this option will process all of the links from the chosen resources to other resources (e.g. Charles Darwin is linked to a resource describing his wife, Emma), determine whether the linked resource's rdf:type fits into a set of pre-defined categories (Agent, Place, Event, Subject or Thing) and group them together in a display.

Although, ostensibly, this feature does nothing more than re-organise the presentation, it does demonstrate some of the potential offered by semantic approaches.
3.4.3.3 Inferencing setup

For the purposes of the project a design decision was taken to add the data being presented to the user into a local temporary data store. This store uses contexts (named graphs) to create a distinct graph for each user session. On performing a new query, any existing named graph for that user is cleared.

The benefit of this approach is that by having all the data in one store it is then possible to perform inferencing, with a known approach. From analysis it is clear that some of the SPARQL services being queried DO support inferencing - for example the British Museum and DBPedia - but others do not. In addition, some of the data is retrieved as stand alone RDF documents and thus cannot be efficiently queried without being stored locally.

It is perhaps worth stepping back at this point to understand why inferencing is valuable. According to Bob DuCharme:

"When you do RDF inferencing, your existing triples are the 'something known,' and your inference tools will infer new triples from them. (These new triples may let you come to some sort of conclusion or decision, depending on what your application does with them.)"

(DuCharme, 2013, ch. 9)

RDFS supports inferencing by enabling the expression of relationships between 'parent' and 'child' classes or properties.

"The property rdfs:subClassOf is an instance of rdf:Property that is used to state that all the
instances of one class are instances of another" (Brickley and Guha, 2014, rdfs:subClassOf)

"The property rdfs:subPropertyOf is an instance of rdf:Property that is used to state that all
resources related by one property are also related by another." (Brickley and Guha, 2014,
 rdfs:subPropertyOf)

Inferencing in the project has used only these two forms. To support inferencing a named graph was created
in the tmp repository for permanent data (http://geekscruff.me/tmp#permanentdata). Into this graph were
added, using a series of INSERT statements, sets of statements to express rdfs:subclassOf and
rdfs:subPropertyOf relationships.

For example:

```
PREFIX ecrm: <http://erlangen-crm.org/current/>
PREFIX dul: <http://www.ontologydesignpatterns.org/ont/dul/DUL.owl#>
PREFIX dbo: <http://dbpedia.org/ontology/>
PREFIX sc: <http://schema.org/>
PREFIX lmdb: <http://data.linkedmdb.org/resource/movie/>
PREFIX rdv: <http://rdvocab.info/ElementsGr2/>

INSERT DATA
{
  GRAPH <http://geekscruff.me/tmp#permanentdata>
  {
    ecrm:E74_Group rdfs:subClassOf dul:Agent # People, Orgs, Groups
    dbo:City rdfs:subClassOf dul:Place # Places
    sc:Book rdfs:subClassOf owl:Thing # Works and objects
    lmdb:performance rdfs:subClassOf dul:Event #Events
    rdv:dateOfDeath rdfs:subPropertyOf dbp:deathDate . # death date
  }
}
```

These have been principally used to facilitate grouping results into categories for places,
people/organisations/groups, events and works/objects as shown in figures 9a and 9b. But this demonstrates
much greater potential, in that with more fine-grained inferencing, the application could identify
comparable statements about roles and occupations or classify the types of relationship between people to
distinguish family members from co-workers. Although not explored here, inferencing can generate new
data, such as a person's age from their birth and death date, or suggestions for who a person may have
interacted with, by analysing places, dates and events associated with difference people.

The insertion of subclass and subproperty information was done manually, but one could envisage using an
automated or semi-automated process to analyse ontologies and extract this information.

The following example query relies on rdfs:subClassOf statements:

```
PREFIX rdv: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX dul: <http://www.ontologydesignpatterns.org/ont/dul/DUL.owl#>

SELECT DISTINCT ?v ?p ?l
FROM <http://geekscruff.me/tmp#permanentdata>
FROM <http://geekscruff.me/tmp#j.allinson@gmail.com>
WHERE
{
  FILTER ((isURI(?v))) .
  ?v a ?t .
  ?t rdv:subclassOf dul:Agent .
}
```

The query looks for Agents (people, organisations and groups) associated with the subject. The underlying
data may use a range of classes as the rdf:type for the given resource but the rdfs:subClassOf statements in the permanentdata graph enable results to be returned irrespective of this. So added value is demonstrated.

### 3.4.4 Create

Storing data, as has been noted above, has not been implemented in the application.

An outline of a basic model for collecting information about a person is outlined below.

```xml
PREFIX dbo: <http://dbpedia.org/ontology/>
PREFIX ecrm: <http://erlangen-crm.org/current/>
PREFIX foaf: <http://xmlns.com/foaf/0.1/>

# create a new URI for the person
<http://www.myapp.me/person/1234>
a <http://xmlns.com/foaf/0.1/Person>
# taken from existing information if available
rdfs:label "name of the person"
owl:sameAs <uri> # uris of any existing graphs

# if this information is not available elsewhere
# FOAF Core provides terms for a basic set of information, such as
foaf:familyName "name"
foaf:givenName "name"

# Options for recording birth and death dates could be
dbo:deathDate
dbo:birthDate

# or
ecrm:P100i_died_in <http://www.myapp.me/person/1234/death>
ecrm:P98i_was_born <http://www.myapp.me/person/birth>
```

There is a distinction to make between describing a person, and creating an authority record for a person. The latter would be a resource in it's own right that is used to collect authoritative information about a person. If a formal authority were being created, it would have a separate URI to the one created for the person.

```xml
PREFIX local: <http://www.myapp.me/ns/auth#>

<http://www.myapp.me/person/1234>
local:hasAuthority <http://www.myapp.me/person/1234/auth>
```

Standards do exist for name authorities in RDF, most obviously the Library of Congress' MADS format[26].

To go beyond these basic descriptive elements, the application could offer a list of properties, generated by examining other ontologies for properties in the domain of Person, such as ecrm:P96i_gave_birth.

Alternatively the application could allow the user to determine their own predicates, although that runs the risk of ill-judged and incorrect modelling decisions from inexperienced ontologists.

In technical terms, implementing a form and storing the data in the AllegroGraph data store is relatively trivial. It is the data modelling that is more intellectually challenging.

### 3.5 Summary

There were many directions open for the project once the basic search was in place, such as:

- supporting more SPARQL services and troubleshooting those that had failed in the setup or querying
stage (for example the British National Bibliography).

- supporting more types of data to go beyond simply people as Persons, to include people as creators and works about people
- supporting APIs and other standards for querying, eg. Europeana and opensearch
- modelling some important resources that are not available in semantic format, to demonstrate the added value of doing so
- adding more functionality to search, widening the properties searched to go beyond rdfs:label, adding filtering options as mentioned in the user requirements
- creating and storing data, as noted above

In the end, the approach shifted from building a fully performant user-focused tool, to experimenting with semantic technologies. The next section will discuss and evaluate the results described above, comment on some of the key issues and draw out possibilities for future work.

1. The prototype application can be found at [http://geekscruff.me/peoplesparql](http://geekscruff.me/peoplesparql)
2. See [https://www.python.org/](https://www.python.org/)
4. Flask recommended by @benoosteen; see [https://www.djangoproject.com/](https://www.djangoproject.com/)
6. See [https://github.com/RDFLib/rdflib](https://github.com/RDFLib/rdflib)
10. See [https://github.com/geekscruff/peoplesparql](https://github.com/geekscruff/peoplesparql)
14. See [http://www.w3.org/TR/rdf-schema](http://www.w3.org/TR/rdf-schema)
15. See [http://www.w3.org/2004/02/skos/core](http://www.w3.org/2004/02/skos/core)
17. See [http://www.w3.org/TR/sparql11-service-description/](http://www.w3.org/TR/sparql11-service-description/)
18. See [http://dbpedia.org/sparql](http://dbpedia.org/sparql), The Archives Hub (http://data.archiveshub.ac.uk/sparql), Artwork.york.ac.uk sample data, [http://geekscruff.me:10035/catalogs/public-catalog/repositories/artworld-people](http://geekscruff.me:10035/catalogs/public-catalog/repositories/artworld-people), Nobel Prizes ([http://data.nobelprize.org/sparql](http://data.nobelprize.org/sparql)) and Yago ([http://lod2.openlinksw.com/sparql](http://lod2.openlinksw.com/sparql)). The BNB ([http://bnb.data.bl.uk/sparql](http://bnb.data.bl.uk/sparql)) was originally included but a problem with using SPARQL FILTER meant queries were failing. Please note that in some cases these URIs are only accessible by
machine, visiting the URI may display an error.

22. See http://geekscruff.me:10035/catalogs/public-catalog/repositories/datasources

23. See https://pypi.python.org/pypi/Flask-KVSession


25. At the time of writing, I cannot find any references online to name-tools, nor is there any description provided in my python package manager (PyCharm). I have version 0.1.3 of the library installed and it is working. I discovered it only a couple of months ago so it is rather unexpected to find all references having disappeared.

26. See http://www.loc.gov/standards/mads/rdf
4. Results and Evaluation

"But what can SPARQL endpoints be used for? They are brilliant for hackdays, prototypes, experiments, toy projects etc. But I don't think anything 'real' could ever be built using one." (Rogers, 2013)

Section 3 has described the prototype application in detail. This section will summarise what the prototype can tell us in support of the research questions. The section will look at the strengths and weaknesses of the approaches and technologies used, draw on user feedback to evaluate the user-view of the application, and finally suggest areas of future work.

The 'results' for the project are the prototype application. This is a fully working application delivering two user-facing components (Query and Explore).

The application, as described in section 3, delivers a set of features in support of testing the hypothesis, but it does not meet all of the essential requirements it set out to (those numbered 1 to 4). This is not because the technologies could not deliver the features, nor because the users would not find them useful - they indicated in the requirements phase that they would - but, rather, because the delivery of these features would add little to the consideration of semantic technologies.

The features that were not delivered are as follows:

- Filter search results, e.g. by date range (basic filtering to reduce results down by discarding unwanted results was in place)
- Search specific fields, e.g. forename, surname, occupation, birth date
- Create a local 'authority record' for names where no external authority is identified

4.1 User Evaluation

The timing of the user evaluation, sent out during July, has led to only one of the 8 original user group invitees responding. The survey they were requested to complete can be seen in appendix 4. The one response is telling:

"Not 100% sure whether the tool is meant for human researchers or automated tools. I guess the challenge for making it really useful is what you can do with the results you find. I can see it being really useful to augment metadata. Some of the results are not presented in a very friendly way for human users (rdf, xml etc)." (Anon., project user evaluation, 2014)

On discussion with the respondent, there had been a problem in testing the three 'Explore' features. These were demonstrated in person and a positive discussion about the potential for these approaches in cultural heritage ensued. The essence of the comment above remained, though - who is this tool for? and what is it trying to deliver?

It is important to note that the direction of work during the course of the project moved away from the user experience side of the application to the underlying technologies and the comments above confirm that the user experience was lacking.
Whilst end user evaluation is light, the literature and the one response received does support that more user tools are needed in this domain.

### 4.2 Results and Technical Evaluation

#### 4.2.1 SPARQL services

The most challenging issue in using SPARQL services is that endpoint behaviour differs with valid queries, irrespective of SPARQL version.

Take the following set two queries, both equivalent, but using a slightly different syntax. In the first there is a single FILTER statement with it's different parts combined with '&&&'. In the second two FILTER statements are provided.

Early in the development work, DBPedia would time out with the following (valid) query, although latterly this style of query has started to work:

```sparql
SELECT DISTINCT * WHERE {
    ?s a <http://xmlns.com/foaf/0.1/Person> .
    FILTER ( (regex(?o, "Beatrice","i")) && (regex(?o, "Webb","i")) )
  }
UNION
    FILTER ( (regex(?o, "Beatrice","i")) && (regex(?o, "Webb","i")) )
  }
}
```

DBPedia would respond very quickly to the same query in this format:

```sparql
SELECT DISTINCT * WHERE {
    ?s a <http://xmlns.com/foaf/0.1/Person> .
    FILTER (regex(?o, "Beatrice","i"))
    FILTER (regex(?o, "Webb","i"))
  }
UNION
    FILTER (regex(?o, "Beatrice","i"))
    FILTER (regex(?o, "Webb","i"))
  }
}
```

The Archives Hub handles two FILTER statements by returning results for either word, rather than both as is expected. It returns results as expected (ie. those containing both terms) when using the '&&&' convention used in the first example above.

The queries above can be tested in the SPARQL interfaces of the services discussed, or by using the YASGUI service, which has the benefit of validating the query and automatically adding PREFIXES. Please note that the YASGUI Request Method must be configured to use GET when querying the Archives Hub.
The British National Bibliography SPARQL endpoint returns a HTTP 500 error if the query includes any FILTER statements, despite FILTER being supported in SPARQL 1.0 AND 1.1. This clearly needs further troubleshooting.

Overall, SPARQL endpoints proved a particular challenge, with varying levels of functionality, robustness and availability. Varied problems encountered include: queries on large datasets that time out, or even appear to crash the SPARQL service; servers reporting a lack of space; and endpoints failing on valid queries. In addition, whilst the application always requests a JSON response (supported in SPARQL 1.1), sometimes XML is returned, requiring the application to handle specific error messages. There is often little in the returned error message to explain what went wrong, particularly where valid queries return QueryBadFormed errors.

A lot of work went into the endpoint setup code in the application and revealed a range of issues, as discussed above. Perhaps Rogers is right and the SPARQL endpoint is at best "an ephemeral, unstable method to share your data" (2013), and it would be useful to review and improve the code written now that the application is complete and also ask whether there are viable alternative approaches, such as pre-filtering datasets and undertaking local querying.

The author's personal view is that robust queryable service endpoints are a must for building services that need to query across varied datasets, but the technology to deliver those may not be quite fit for purpose yet.

### 4.2.3 SPARQL

Refreshingly, SPARQL itself has not been problematic to use for the project and has proved able to deliver on all features required. The application has really only scratched the surface of what SPARQL can do, as a cursory look at DuCharme's *Learning SPARQL* will demonstrate.

In the application itself, use of SPARQL has been restricted to ASK and SELECT queries, with query results returned as JSON and in some cases results limited to a specified NAMED GRAPH (a discreet set of data that can be queried).

ASK will perform the query but return only a boolean to indicate whether the query will return results (true) or not (false).

```
ASK { <http://data.archiveshub.ac.uk/id/person/nra/burtonrobert1577-1640author> ?p ?o }
```

SELECT performs the query and returns a results set, in the format requested.

```
SELECT ?o { <http://data.archiveshub.ac.uk/id/person/nra/burtonrobert1577-1640author> ?p ?o }
```

In addition, the following types of query have not been used:

CONSTRUCT allows for copying, creating and converting data whilst leaving the original data unchanged.
DESCRIBE asks for triples that describe a particular resource, and the query processor is free to decide what triples are returned.

Outside of the application, INSERT statements have been used to add data to the AllegroGraph store.

Other features of SPARQL that have been used include FILTER, to filter results, and regex, to match patterns. These can be seen in the query examples in section 4.2.

BIND offers a mechanism to bind the results of a particular check or sum to a result. The following example shows how it has been used during the search for birth and death dates to indicate whether the result is a URI. MINUS is also demonstrated in this query, a mechanism for excluding results.

```sparql
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>

SELECT DISTINCT ?s ?p ?o ?u WHERE {
     BIND (isURI(?o) as ?u) }
   MINUS { ?s rdf:type ?o }
   MINUS { ?s rdfs:seeAlso ?o }
} ORDER BY ?o
```

SPARQL has served the project well, and in the evaluation of the project is a technology that could have been used more widely, both in inferencing and in manipulating and inserting data.

### 4.2.4 Ontologies and Inferencing

Ontologies, "the third basic component of the Semantic Web" (Berners-Lee, Hendler and Lassila, 2001), provide the structure and meaning needed to understand, compare and combine data. Without ontologies we cannot say that a property in one dataset is equivalent to a property in another and, therefore, this prevents querying across different datasets with any real meaning. Ontologies have been central to building the application and from the outset the need to identify the different properties and Classes used across the different datasets being queried.

In the rather simplistic way the application works, being fed a manually-generated list of classes for Person and properties used for a human-readable label, the application can construct appropriate queries across different SPARQL services to enable drawing together these results into one results set. In the current version, this works, albeit with the limitation that new classes or properties cannot be accommodated without manual intervention.

There isn't a simple query to generate a list of all ontologies used by a SPARQL service, but following is a non-exhaustive list:

RDF Syntax <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
RDF Schema <http://www.w3.org/2000/01/rdf-schema#>
Dublin Core <http://purl.org/dc/terms/>
That there are 20 plus ontologies in use across 7 datasets shows how varied the space is and demonstrates clearly that this variety needs to be accommodated in any application. It is highly unlikely that there will ever be a time when all applications standardise on a single ontology.

Two of the ontologies illustrate neatly how different modelling approaches can make it more difficult to reconcile data structured according to their base ontologies' different rules. The DBPedia ontology is a very shallow, resource-centric, ontology. Whereas CIDOC-CRM is more event-driven, modelling things to a number of levels in order to more robustly describe it's domain.

An easy example of this is birth and death dates. In the DBPedia ontology there is a direct property that links together a person and their birth or death date, which are literal values, for example "1974". With CIDOC-CRM, as shown in the British Museum collection catalogue, the data model links a person to a birth event object and then the birth event to a timespan object, which has an rdfs:label providing the date.

Such modelling differences are inevitable and the application deals with this difference, by checking whether the object of the birth or death property is a URI, and if it is, then following that URI and repeating the check. But it is useful to a consider for a moment, what benefits the shallow and deep approaches bring. The shallow approach is easier to implement, easier to query and would fulfil many use cases. The deeper approach, however, provides greater flexibility, enabling other information about the birth event to be supplied, such as place, or people involved. For scholarship, where complex questions arise, the more flexible approach will almost certainly be needed.

Ontological mis-matches are less easy to deal with, though, and a good example is nationality, modelled by DBPpedia, arguably wrongly, as a type of Place and in CIDOC-CRM, more logically, as a Group, a Type and a Concept.

### 4.2.5 Inferencing

Inferencing, or reasoning, can have a range of purposes, from acting as a 'rules engine' to ensure that instance data modelled using an ontology or several ontologies are ontologically correct. For example, to ensure the data does not classify a Person as a Place.

The 'Explore' features of the application makes use of a basic level of inferencing to return results across a set of equivalent classes from different ontologies. There are a range of tools that can help with this, such as Protege, TopBraid Composer and reasoning engines like fuxi for python 4. Allemang's *Semantic Web*
for the working ontologist is an excellent book on ontology development and the semantic Web, with several chapters on inferencing.\(^5\)

Inferencing is one of the greatest value added aspects of using semantic technology in it's ability to create new data. Take the following, often used example, of calculating the age of a person based on their birth and death dates.

```prefix
PREFIX archiveshub: <http://data.archiveshub.ac.uk/def/>
PREFIX mine: <http://geekscruff.me/tmp/>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
CONSTRUCT { ?person mine:age ?age }
WHERE
  OPTIONAL { ?person archiveshub:dateDeath ?dyear }
  BIND ((xsd:integer(?dyear)) - (xsd:integer(?byear)) AS ?age)
} LIMIT 10
```

Five results are listed below (in a condensed form):

- [http://data.archiveshub.ac.uk/id/person/ncarules/chamberlainville1869-1940statesman] = 71
- [http://data.archiveshub.ac.uk/id/person/ncarules/georgev1865-1936kingofgreatbritainandireland] = 71
- [http://data.archiveshub.ac.uk/id/person/ncarules/shelleybyshe1731-1815sirfirstbaronet] = 84
- [http://data.archiveshub.ac.uk/id/person/nra/martinbasilkingsley1897-1969journalist] = 72
- [http://data.archiveshub.ac.uk/id/person/ncarules/gibsonwilfrid1878-1962authorandpoet] = 84

Before this query the data did not exist, now it can be further used or displayed. A rather fuller version of this, which handles those who don't have a death date, is provided in the TopQuadrant blog\(^4\). More complex examples could be inferring that two people may have worked together based on their both working in the same Place at the same time without that information being explicitly stated, or estimating the worth of a painting by inferring an estimation from data about comparable works sold during the same period. Such examples have clear uses to the types of research questions submitted by the user group.

In so far as it has been used, inferencing has proved a successful mechanism, but the project could have gone much further by delving into the research questions provided had time allowed.

### 4.2.5 Data Store

AllegroGraph and it's Python client proved adequate to deliver a back end RDF data store. Adding, deleting and querying data were fast and relatively straightforward to implement.

Other approaches could have been explored, such as much simpler RDF documents stored as files and queried using the rdflib libraries.

### 4.2.6 Python, Flask, Python Libraries and the Code

Python, Flask, rdflib and SPARQLWrapper suppported the application well. Login was achieved simply with the Flask persona example. rdflib, in particular, offers much more functionality than was used, in particular mechanisms for preparing queries, storing data and also the transitive module that can return a set of nodes by walking through a set of subjects, or objects via their properties.
As written, the 'enhance with semantics' feature, where RDF is pulled in from each object linked to a selected subject, is very slow. As a user-focused tool, this would not be acceptable and as a minimum, some indicator of progress and the ability to stop would be needed. Re-factoring is needed to find faster, or alternative ways of delivering this feature, perhaps by pre-processing ontologies to refine queries or requests before they are sent. This clearly illustrates the significant issue of performance.

4.2.7 The Datasets

The SPARQL services chosen for this project represented a range of datasets, from the relatively small Archives Hub, to the large but domain-specific British Museum Catalogue, to the massive and broad ranging DBPedia. One issue that rather plagued the development were the large numbers of results from DBPedia, almost all of which were different language versions of the same resource. Filtering these out proved difficult and there are a range of places in the code where there is functionality to deal specifically with DBPedia resources. Resource specific code like this far from idea in an application that should be abstract enough to deal with new SPARQL services. For the end user, duplicate results are a frustration and can undermine the quality of the resource delivered. This remains a problem without a satisfactory solution.

4.3 Summary and Future Directions

RDF is a universal data model and framework for the rich and structured description of data as graphs, which can then be serialized into a number of formats (RDF/XML, N3, TURTLE and RDF/A). It is supported well by existing libraries, data stores and query technology. It's model of describing data in graphs of triples is ideal for the kinds of questions historians want answered, like who did X, where did X happen, when did X happen and so forth. And for the 'OpenGLAM' community, cultural heritage institutions, with their remit of providing access to their collections, should be at the forefront of making information available in ways that a range of users can utilise.

But questions do arise, such as (i) is RDF too complex, demanding an understanding of ontologies and data modelling? (ii) what about the opportunities afforded by microformats and schema.org? and (iii) what if the big Web players force a different direction? The scope of this project is too small to address these large questions, but it is clear that RDF and SPARQL are not the only approach to open data. APIs, bespoke tailored resources and relational database-driven approaches are all in wide use and it can be difficult to argue against traditional databases when they are so performant and well-established. The key arguments in favour of semantic approaches are, first, the possibilities opened up by inferencing and, second, the use of the Web as infrastructure, doing away with silos of content, accessible only through proprietary or bespoke tools and interfaces.

For reasons noted above, the application built for this project clearly failed as a user-focused application. Nevertheless, it demonstrates the potential offered and starts to unpick how research questions can be answered using semantic techniques. Inevitably a key issue is that any tool can only be as good as the available data, and very many resources are simply not available in semantic form, or indeed, in any machine-readable form. The work of modelling large scale humanities resources into semantic datasets is vast, and with the growth of available sources come concerns about scale and performance. From another
perspective, why model data when you can apply natural language processing and data mining algorithms? Perhaps the answer lies in a combined approach, and the literature has shown that research is happening into using natural language techniques to extract structured queries. Quepy, for example, produces MQL, for freebase, and SPARQL for other services from natural language queries 2. Entity recognition and extraction techniques too could help with finding classes in a large dataset. As a side note, Freebase links were found during the project, but on retrieving them with rdflib, they could not be queried due to formatting within the documents - a known issue. Apache stanbol, for example, has done work in this area 8.

Future directions for the project, then, could be many. The project has certainly demonstrated success in using semantic techniques for searching for historical names. Effort could be well spent in using the work done here to inform an assessment of existing tools and approaches. For example any23 can scrape a range of RDF formats a variety of Web documents, pundit promises to turn Web pages into a semantic knowledge network, Open refine helps clean up messy data in spreadsheets and can export RDF and the Linked Media Framework aims to help publish Linked Data 9. But whether these can truly support the historical researcher remains to be tested. Going beyond historical people, too, would be very valuable, with projects and services like Geonames, pelagios and the Getty's TGN are contributing data about historical places. Finally, this tiny project is part of a much larger, and very interesting, space with projects like HuNI (Humanities Networked Infrastructure) building a 'virtual laboratory' for researchers to work across Australian cultural datasets and the recently funded CLARIAH, which aims to "offer humanities scholars a 'Common Lab' that provides them access to large collections of digital resources and innovative user-friendly processing tools, thus enabling them to carry out ground-breaking research to discover the nature of human culture" (CLARIAH, 2014) 11.

1. See http://yasgui.laurensrietveld.nl/ ↩
3. See http://json.org/ ↩
6. See http://topquadrantblog.blogspot.co.uk/2010/05/how-to-publish-your-linked-data-with.html ↩
7. See http://quepy.machinalis.com/ ↩
8. See http://stanbol.apache.org/docs/trunk/components/enhancer/engines/opennlpner ↩
10. See http://www.geonames.org/, http://pelagios-project.blogspot.co.uk/ and http://www.getty.edu/research/tools/vocabularies/lod/ ↩
5. Conclusions

"The Semantic Web we aspire to makes substantial reuse of existing ontologies and data. It's a linked information space in which data is being enriched and added. It lets users engage in the sort of serendipitous reuse and discovery of related information that's been a hallmark of viral Web uptake." (2006, p.100)

As has been described, the project has built a tool to help historical name research by using semantic technologies. The surveyed literature provides considerable evidence in support of this approach, both of the opportunities for supporting humanities researchers and cultural heritage professionals and, especially, the opportunities offered by Linked Data. The ability to describe and research 'entities' and to link them together with related entities, to consult multiple sources and to compare and validate information are common and real needs across both domains being studied: those of humanities research (historical research in particular) and cultural heritage cataloguing and description practice. The literature has also pointed to the array of underused digital source material currently available, none of which is in semantic form.

The project has successfully demonstrated querying across multiple data sources, using diverging ontologies, and the usefulness, both real and potential, of SPARQL and semantic inferencing for answering complex research questions. It has also exposed a range of issues, from SPARQL service standardisation and reliability, through mis-matched ontological modelling, to the complexity level needed to truly scale querying across different 'entities' to answer serious research questions. Even a simple question of "is this the same person?" could involve a wide range of queries and could be de-railed by mis-matched data or semantic imprecision.

Nevertheless, SPARQL, Linked Data, ontologies and the whole gamut of semantic technologies explored in this project are ideal for precisely the kinds of things humanities scholars want to do, viz. to follow links between the who, what, where and when of history. Ontologies like CIDOC-CRM enable rich structured description, building a huge network of interrelated data to query. Yes, there are issues around scalability, of how to make vast amounts of data efficiently querable, but several possible directions have been illustrated through the literature and evaluation, and the answer is almost certainly a combination of different approaches. There are cultural barriers to overcome too, in order to build the global web of data, researchers must be happy to share and able to trust the data they consume.

A browse of the mailing list archives of the W3C's public-lod mailing list archives shows how much activity is happening in this space right now. Clearly a vibrant and active community is in place. But there is word of caution for services that present raw URIs in complex and unattractive interfaces, these will not meet the needs of humanities scholars. Nor is expecting scholars to understand complex mathematical models and even programming languages in order to make progress going to be possible for many who could genuinely benefit from the technologies and tools on offer.

Services like HuNI and CLARIAH, mentioned in the last section, offer researchers the promise of the virtual 'lab' and projects like ChartEx, with it's historian's 'workbench' offer alternate and complimentary
approaches drawing on technologies like machine learning and data mining to extract entities from huge corpuses of data. It is hoped that this small project has added something to the knowledge of the possibilities and issues in this domain.

In conclusion, it is difficult to argue against the principles laid out by Berners-Lee, that sharing data and making connections is a true and right aim for humanity and scholarship, and the author hopes that Linked Data will deliver rich services for the humanities and cultural heritage sectors.

   It is about getting excited about connections, rather than nervous.” (Berners-Lee, 2007)

1. See http://lists.w3.org/Archives/Public/public-lod/
References


Anon. (2014) 'MSc Project Requirements Questionnaire'.


Burrows, T. (2011) 'Sharing humanities data for e-research: conceptual and technical issues', in Sustainable Data from Digital Research. Melbourne, pp. 177–192. Available at:
Berlin Heidelberg (Lecture Notes in Computer Science), pp. 321–326. Available at: http://dx.doi.org/10.1007/978-3-642-12842-4_37.


Knowledge and Media Technologies group at Salzburg Research (n.d.) *Linked Media Framework*. Available at: https://code.google.com/p/lmf/.


Pelagios Project (2014) *Pelagios : Linking together the places of our past through the documents that refer to them*. Available at: http://pelagios-project.blogspot.co.uk/.


The DBpedia Ontology (3.9) (2014) DBPedia Blog.


W3C (2013a) Semantic Web, W3C. Available at: http://www.w3.org/standards/semanticweb/.

W3C (2013b) W3C Semantic Web Activity, W3C. Available at: http://www.w3.org/2001/sw/.


Williams, G. T. (2013) SPARQL 1.1 Service Description, W3C. Available at: http://www.w3.org/TR/sparql11-service-description/.
Appendices
Appendix 1. User Group

Demographic of User Group

Role of User Group

The following email was sent to all potential members of the user group:

For the past two years I have been undertaking a MSc in Computing with Heriot Watt and am just starting work on my dissertation project. I am writing to ask if you would consider being part of my ‘user group’.

The aim of the project is to build a tool to aid research into historical names. The tool would be web-based and would allow a researcher to do the following:

- search multiple sources from a single interface
- identify the correct name
- identify any information to reference locally, eg. name, dates, biographical notes
- add local information not found in external sources, eg. context of research, historical sources consulted
- *add a name authority record where the name does not appear in any external source

The result would be a kind of researchers’ online notebook. I want to assemble a group of around 8 people to help evaluate and validate the tool.

The motivation for the project is that I have seen many examples where researchers or cataloguers are having to manually enter, or copy/paste information into local databases or spreadsheets. On the flip side, there is a growing set of query-able data sources available online but integrating them into local practice demands a level of technical expertise that most historians, librarians or archivists would not have.

In my project I want to bring those two things together. I have chosen historical names to limit the scope of the project, but this approach could also apply to places, events and more.

If you are able to help, I will be asking you to do two things:

1) Complete a short requirements questionnaire in the next few weeks.

2) Carry out an evaluation of the prototype system in June / July.

Please let me know if you would be willing to help.
Appendix 2. Data Sources

Data Source Discovery

The following sites have been the main sources for identifying relevant data sources.

- [http://www.w3.org/wiki/SparqlEndpoints](http://www.w3.org/wiki/SparqlEndpoints)
- [http://sparqles.okfn.org/](http://sparqles.okfn.org/)
- [http://datahub.io/](http://datahub.io/)

W3C provide a recommendation for SPARQL service descriptions. This includes the following recommendation about how these should be discovered:

"SPARQL services made available via the SPARQL Protocol SHOULD return a service description document at the service endpoint when dereferenced using the HTTP GET operation without any query parameter strings provided. This service description MUST be made available in an RDF serialization, MAY be embedded in (X)HTML by way of RDFa [RDFA], and SHOULD use content negotiation [CONNEG] if available in other RDF representations." (Williams, 2013)

In order to check for the existence of a service description, a Firefox plugin called 'Poster' was used to issue a GET request with an accept header of application/rdf+xml. The results of this check is indicated in the column called 'RDF Service Description' below. 'SETUP' refers to whether the SPARQL service could be successfully added to the project application, and 'QUERY' notes if a query could subsequently be performed successfully.

### Data Sources

<table>
<thead>
<tr>
<th>Data Source</th>
<th>SPARQL Service</th>
<th>Notes</th>
<th>RDF Service Description</th>
<th>Person Identifier(s)</th>
<th>SETUP</th>
<th>QUERY</th>
</tr>
</thead>
<tbody>
<tr>
<td>Archives Hub</td>
<td><a href="http://data.archiveshub.ac.uk/sparql">http://data.archiveshub.ac.uk/sparql</a></td>
<td>Linked Data version of the UK National Archives catalogue</td>
<td>NO</td>
<td><a href="http://erlangen-crm.org/current/E21_Person">http://erlangen-crm.org/current/E21_Person</a></td>
<td>YES</td>
<td>YES</td>
</tr>
<tr>
<td>British Library British National Bibliography</td>
<td><a href="http://www.bl.uk/bibliographic/datafree.html#lod">http://www.bl.uk/bibliographic/datafree.html#lod</a></td>
<td>Linked Data version of the BNB. GET request for the service description returns RDF about the dataset.</td>
<td>PARTIAL</td>
<td><a href="http://xmlns.com/foaf/0.1/Person">http://xmlns.com/foaf/0.1/Person</a></td>
<td>YES</td>
<td>FAILS; 500 ERROR</td>
</tr>
<tr>
<td>Nobel Prizes</td>
<td><a href="http://data.nobelprize.org/">http://data.nobelprize.org/</a></td>
<td>The Nobel Prizewinners available as Linked Data. Service Description request responds with a&amp;neError 404: Service Description: nobel/query&amp;labs.</td>
<td>NO</td>
<td><a href="http://xmlns.com/foaf/0.1/Person">http://xmlns.com/foaf/0.1/Person</a></td>
<td>YES</td>
<td>YES</td>
</tr>
<tr>
<td>Data Source</td>
<td>SPARQL Service</td>
<td>Notes</td>
<td>RDF Service Description</td>
<td>Person Identifier(s)</td>
<td>SETUP QUERIES</td>
<td></td>
</tr>
<tr>
<td>-----------------------------------</td>
<td>----------------</td>
<td>----------------------------------------------------------------------</td>
<td>-------------------------</td>
<td>----------------------</td>
<td>---------------</td>
<td></td>
</tr>
<tr>
<td>Getty Vocabularies</td>
<td><a href="http://vocabulary.getty.edu/sparql">http://vocabulary.getty.edu/sparql</a></td>
<td>Linked Data versions of these art historical vocabularies are in development. AAT is the only one available at present, and doesn't include people. ULAN is on its way.</td>
<td>NO</td>
<td>N/A</td>
<td>N/A (YET) N/A</td>
<td></td>
</tr>
<tr>
<td>Pelagios</td>
<td>API</td>
<td>Linked Data for historical places</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A N/A</td>
<td></td>
</tr>
<tr>
<td>Europeana</td>
<td><a href="http://europeana.ontotext.com/sparql">http://europeana.ontotext.com/sparql</a></td>
<td>Europeana dataset.</td>
<td>NO</td>
<td><a href="http://purl.org/dc/elements/1.1/creator">http://purl.org/dc/elements/1.1/creator</a></td>
<td>NO; ERROR N/A</td>
<td></td>
</tr>
<tr>
<td>RKBexplore</td>
<td><a href="http://biolit.rkbexplorer.com/sparql/">http://biolit.rkbexplorer.com/sparql/</a></td>
<td>A Short Biographical Dictionary of English Literature by John W. Cousin: &quot;quick and dirty job of processing the data provided by A Project Gutenberg from A Short Biographical Dictionary of English Literature by John W. Cousin&quot;)</td>
<td>NO</td>
<td><a href="http://www.w3.org/2000/01/rdf-schema#Resource">http://www.w3.org/2000/01/rdf-schema#Resource</a> (all people)</td>
<td>N/A N/A</td>
<td></td>
</tr>
<tr>
<td>Linked Movie Database</td>
<td><a href="http://data.linkedmdb.com/sparql">http://data.linkedmdb.com/sparql</a></td>
<td>Linked Data about movies</td>
<td>NO</td>
<td><a href="http://xmlns.com/foaf/0.1/Person">http://xmlns.com/foaf/0.1/Person</a></td>
<td>YES YES</td>
<td></td>
</tr>
<tr>
<td>Europeana</td>
<td><a href="http://data.archaeologydataservice.ac.uk/query">http://data.archaeologydataservice.ac.uk/query</a></td>
<td>Linked Data catalogue for the collections held by the ADS</td>
<td>NO</td>
<td><a href="http://purl.org/dc/elements/1.1/creator">http://purl.org/dc/elements/1.1/creator</a> (creators of datasets)</td>
<td>NO; ERROR N/A</td>
<td></td>
</tr>
<tr>
<td>Freebase</td>
<td>API</td>
<td>&quot;Freebase is an open, Creative Commons licensed repository of structured data of almost 23 million entities&quot;. Can be searched with Fact Forge.</td>
<td>N/A</td>
<td><a href="http://rdf.freebase.com/ns/people.person">http://rdf.freebase.com/ns/people.person</a></td>
<td>N/A N/A</td>
<td></td>
</tr>
<tr>
<td>DBTune</td>
<td><a href="http://dbtune.org/">http://dbtune.org/</a></td>
<td>Music-related structured data. Includes the MusicBrainz catalogue.</td>
<td>NO</td>
<td><a href="http://purl.org/ontology/foaf/MusicArtist">http://purl.org/ontology/foaf/MusicArtist</a></td>
<td>YES NO; ERROR</td>
<td></td>
</tr>
<tr>
<td>VIAF</td>
<td>API</td>
<td>Authority files from many sources available in RDF.ttp://www.oclc.org/developer/develop/web-services/virtual-international-authority-file-viaf.en.html</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A N/A</td>
<td></td>
</tr>
<tr>
<td>LC Linked Data Service</td>
<td><a href="http://id.loc.gov/">http://id.loc.gov/</a></td>
<td>Supports Read, SRUsearch and AutoSuggest</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A N/A</td>
<td></td>
</tr>
<tr>
<td>Artworld</td>
<td>API</td>
<td>An art historical dataset. Partial dataset available.</td>
<td>NO</td>
<td>owl:NamedIndividual</td>
<td>YES YES</td>
<td></td>
</tr>
<tr>
<td>Sindice</td>
<td>API</td>
<td>A semantic search engine. The Sindice service has now been closed.</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A N/A</td>
<td></td>
</tr>
<tr>
<td>Nature</td>
<td>API</td>
<td>Authors publishing in Nature as Linked Data</td>
<td>NO</td>
<td><a href="http://ms.nature.com/terms/Contributor">http://ms.nature.com/terms/Contributor</a></td>
<td>NO; ERROR N/A</td>
<td></td>
</tr>
<tr>
<td>Austrian Ski Team</td>
<td>API</td>
<td>Authors publishing in Nature as Linked Data</td>
<td>NO</td>
<td><a href="http://xmlns.com/foaf/0.1/Person">http://xmlns.com/foaf/0.1/Person</a></td>
<td>YES YES</td>
<td></td>
</tr>
</tbody>
</table>

Julie Allinson: Semantics-based Indexing of Historical Entities 63
<table>
<thead>
<tr>
<th>Data Source</th>
<th>SPARQL Service</th>
<th>Notes</th>
<th>RDF Service Description</th>
<th>Person Identifier(s)</th>
<th>SETUP QUERY</th>
</tr>
</thead>
<tbody>
<tr>
<td>FreeBMD <a href="http://datahub.io/dataset/freebmd">http://datahub.io/dataset/freebmd</a></td>
<td>N/A</td>
<td>Manual access only to Births, Marriages and Deaths. &quot;Access to the data held by FreeBMD is only permitted manually via the search page. The use of front end programs or sites to enter search parameters is strictly forbidden&quot;</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>Project Gutenberg <a href="http://www.gutenberg.org/">http://www.gutenberg.org/</a></td>
<td><a href="http://wifo5-03.informatik.uni-mannheim.de/gutendata/sparql">http://wifo5-03.informatik.uni-mannheim.de/gutendata/sparql</a></td>
<td>Catalogue of books in the Project Gutenberg free library, as Linked Data</td>
<td>NO</td>
<td><a href="http://xmlns.com/foaf/0.1/Person">http://xmlns.com/foaf/0.1/Person</a></td>
<td>YES</td>
</tr>
<tr>
<td>Web Data Commons <a href="http://webdatacommons.org/">http://webdatacommons.org/</a></td>
<td>Dataset</td>
<td></td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>Yago</td>
<td><a href="http://lod2.openlinksw.com/sparql">http://lod2.openlinksw.com/sparql</a></td>
<td>â€œYAGO (Yet Another Great Ontology) is a knowledge base developed at the Max Planck Institute for Computer Science in Saarbrücken. It is automatically extracted from Wikipedia and other sources.â€œ (Wikipedia)</td>
<td>YES</td>
<td><a href="http://yago-knowledge.org/resource/wordnet_person_100007846">http://yago-knowledge.org/resource/wordnet_person_100007846</a></td>
<td>YES</td>
</tr>
</tbody>
</table>

Appendix 3a. Requirements Survey Summary

The tool needs to be web-based [Q1. On Functionality]

- Essential: 5 (100%)
- Desirable: 0 (0%)
- I have no opinion: 0 (0%)
- I don't understand this: 0 (0%)

I want to query multiple sources [Q1. On Functionality]

- Essential: 7 (100%)
- Desirable: 0 (0%)
- I have no opinion: 0 (0%)
- I don't understand this: 0 (0%)

I want to decide which sources should be searched [Q1. On Functionality]

- Essential: 4 (67%)
- Desirable: 0 (0%)
- I have no opinion: 0 (0%)
- I don't understand this: 0 (0%)
I want to be able to add new data sources to query [Q1. On Functionality]

- Essential: 6 (100%)
- Desirable: 0 (0%)
- I have no opinion: 0 (0%)
- I don't understand this: 0 (0%)

I want a single search box [Q1. On Functionality]

- Essential: 3 (75%)
- Desirable: 0 (0%)
- I have no opinion: 1 (25%)
- I don't understand this: 0 (0%)

I want the search to support boolean searching [Q1. On Functionality]

- Essential: 0 (0%)
- Desirable: 0 (0%)
- I have no opinion: 5 (100%)
- I don't understand this: 0 (0%)
I want to filter search results, e.g. by date range [Q1. continued]

I want to search specific fields, e.g. forename, surname, occupation, birth date [Q1. continued]

I want to see exact matches and likely matches [Q1. continued]
I want search suggestions as I type [Q1. continued]

I want suggested alternative terms (‘did you mean?’) [Q1. continued]

I want to view full information about each name [Q1. continued]
I want to select preferred information elements, e.g. name, birth/death dates, biographical description [Q1. continued]

I want to add local information elements, drawn from a list of options [Q1. continued]

I want to add custom local information elements [Q1. continued]
I want to create a local 'authority record' for names where no external authority is identified [Q1. continued]

I want to be able to follow links from people to other people, and to other entities [Q1. continued]

I want to add links between entities [Q1. continued]
I want a browser plug-in for researching a name from anywhere on the web [Q1. continued]

I want to be able to visualise connections in a tree or graph visualisation [Q1. continued]
I want a browser plug-in so that I can research a name from any web page without visiting the tool [Q1. continued]

I would like to download my data in different formats, e.g. a spreadsheet or PDF [Q1. continued]

I want an easy way to include information from this tool in other web pages [Q1. continued]
Q2. On Research Questions

I have found a letter with details of a shop (early 20th century), plus the name of someone who works there. I want to find out who the person was and if they, or someone else owned the shop. I want to find further information via the person's name about where they lived and worked. I also want to find details of who owned the shop and where they lived and how they worked. Referring to 'person X' from the set of letters I was researching, I would like to know 1. Who they knew and in what capacity (person and relationship) 2. Is person a named in the letters the same as person b named in the letters? 3. What time period was the person active/working in? 4. In the Music Preserved Marsh Collection there is a recording of 'Concerto in Eb major for harpsichord, two violins and cello' by JC Bach. Which JC Bach was this (not easy to answer from wikipedia or a quick Google search)? There are at least two JC Bachs and works vary in how they are described. 5. Is Thomas goldsmith of York the same person as Thomas son of Josce goldsmith of York? 6. The main question I would expect this system to answer is simply: is this person/place/entity already recorded in 'the literature' - ie have they been noticed already? - or not.

Is person a named in the letters the same as person b named in an external source? Quite often there can be confusion over the identification of places, as either different places can have the same name, or else the same place can have different names/be described in different ways (esp at different historical time periods). Although with people's names there are some quite advanced authority files, which gives the 'best' name and a list of alternatives, no such authorities exist for places. Could this system, however, draw on a comprehensive list of place names (both UK-wide and street names in London) to start to give some shape and order to searching for places? Let's say I am searching for 'Fountain Tavern, Strand', it could give me - or link to - a list of all the addresses in the Strand. That would be a very useful way of seeing the data because even though most of the results will not be the Fountain Tavern (which actually will be quite easy to identify), with places it is often very useful to see other places nearby - other places that I wasn't searching for - in a way that isn't true of people's names. If I am searching for "John Smith", it is of no interest whatever to learn that there was a different person called "Joan Smith" who happens to occupy the neighbouring place in the people authority file. But if I am searching for "two doors up from the corner of Chancery Lane" it is of interest to know that there is also an entry for the place described as "three doors up from the corner of Chancery Lane." Sorry I know this isn't a very well-formed research question! Is there another person with the same name? Can it give enough information to allow the researcher to choose the correct one? I want to identify personal names in Bede's "Ecclesiastical History" and then see how many of them are of Anglo-Saxon origin and how many are of native British origin. This would include a search if the book itself, plus research and comparisons with a number of databases and online scholarly works. Is Sarah Croney the same person as Sarah Rees Jones? 2. Key events in their life

Were goldsmiths in York more likely to have Jewish names than goldsmiths in London? I want to know how Irish immigrants to the uk came into the country in the late 19th century and if they
changed their names. I want to start by looking at people and their names as they arrive in the UK, tracking their movements, employment and court records to see how many Irish names survive compared to how many people with Irish names entered the country over a specific twenty year period. I would need to search across census records, court records, employment records, workhouse records, immigration records etc and to compare names and calculate instances if a name used. Can person be located in an online archival source, such as the census, parish register, civil registration or probate? Where other archival sources are that may provide me with more information about them.

Archives Hub [Q3. On Data Sources]

<p>| | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>25%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>13%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>0%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>2</td>
<td>25%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>3</td>
<td>38%</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

British National Bibliography [Q3. On Data Sources]

<p>| | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>29%</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>29%</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>14%</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>0%</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>2</td>
<td>29%</td>
<td></td>
</tr>
</tbody>
</table>

Freebase ("a community-curated database of well-known people, places and things") [Q3. On Data Sources]

<p>| | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>4</td>
<td>50%</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>13%</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>38%</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>0%</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>0</td>
<td>0%</td>
<td></td>
</tr>
</tbody>
</table>
I know that I use Wikipedia - it has proved to be v. useful to me. The BNB is a great biographical source. I have looked at Freebase but I don't use it in an everyday sense... maybe this is more habit than anything else because it may now be more comprehensive than when I last looked at it. Getty Artists' Names is more niche. The Archives Hub is a good service for bringing archives together, although it is not as comprehensive as it ideally could be. I do not know the British National Biography (unless you mean the Dictionary of National Biography, which would rank as 5). Freebase I wasn't familiar with. I don't think it would be useful for my work as it is not clear where its information comes from (although there are lots of references to wikipedia). There is an issue about the quality of data, and about data whose sourcing is unclear just being recirculated endlessly around these huge databases. Wikipedia I find useful for checking basic biographical information about quite little-known people, such as Lord this or that. The archives hub seems to feature such a broad array of archives
that most of its results are irrelevant to my searches. Hard to choose, as I'm not actually doing research. Mostly too modern in their broad focus I've never used 2 or 5; 3 I've never heard of; 4 is useful only for specific topics; 6, while problematic for all sorts of reasons, is accessible and available (though I would always wish to evaluate the information provided).

Other sources

art world in britain 1660-1735…. of course! a2a - Access2Archives, part of the national archives website. history of parliament online Dictionary of National Biography PCC Wills - a database of wills, part of the national archives website (part of documents online, I think it's called) There is a very good list of artists and others maintained by the RKD, the Rijksbureau voor Kunsthistorische something-or-other (www.rkd.nl) The British Museum database also has a biographical database within it which is very good archive.org - not sure if this would work, but it would be great to be able to search this online library google books ESTC (http://estc.bl.uk) an amazing list of all the books printed in the UK before about 1800. British Library manuscripts catalogue DNB if it was open! British census records and court records would be useful. I think many are online now and historians are interrogating them. Dictionary of National Biography Clergy of the Church of England database The resources in Find My Past and Ancestry The resources in the HRI Sheffield (including cause papers and Old Bailey) Free BMD British history online. Connected Communities and medieval equivalent at Sheffield HRI Grove Music Online

Q4. On Sharing

I want to make data available for others to query 2 25%
I want to choose what I share 5 63%
I do not want to share my local data 1 13%
Other 0 0%

Q5. On technical expertise
This project really needs to consider ...

Making sure that users are aware of how to discover the data deficiencies in each source — Usability. Hooking up to some of the 19th C info available online. Modern historians into that period are very keen on using tools and would love this one. The issue of how authoritative is the data it's displaying. The internet is awash with second- and third-hand data. How can you present your data in a way that makes it clear to users where it is coming from & how authoritative it is. Clearly explaining why this method of bringing data together is more useful than simply going to sources researchers are familiar with. Why the tool won't give them everything (managing expectations).

This project won't work because ...

Possibly because of technical issues and challenges with data integration. Also challenges with relevance ranking of data sources. Are most places happy for you to use their data? Hopefully they are, as it sounds really useful! No reason why it shouldn't 'work' as long as you make sure you define 'work'! What is the unique benefit of using the system as against, say, doing a google search (which brings up all kinds of unexpected results, such as if someone has been doing a genealogical research project and just by chance has found something about the person I am searching for)?
Appendix 3b. User Research Questions

In the Music Preserved Marsh Collection there is a recording of 'Concerto in Eb major for harpsichord, two violins and cello' by JC Bach. Which JC Bach was this (not easy to answer from wikipedia or a quick Google search)? There are at least two JC Bachs and works vary in how they are described.

The main question I would expect this system to answer is simply: is this person/place/entity already recorded in 'the literature' - ie have they been noticed already? - or not.

What time period was the person active/working in?

Referring to 'person X' from the set of letters I was researching, I would like to know:

1. Who they knew and in what capacity (person and relationship)
2. Key events in their life
3. Where other archival sources are that may provide me with more information about them

Is person a named in the letters the same as person b named in the letters?

Is Thomas goldsmith of York the same person as Thomas son of Josce goldsmith of York

I have found a letter with details of a shop (early 20th century), plus the name of someone who works there. I want to find out who the person was and if they, or someone else owned the shop. I want to find further information via the person's name about where they lived and worked. I also want to find details of who owned the shop and where they lived and how they worked.

Quite often there can be confusion over the identification of places, as either different places can have the same name, or else the same place can have different names/be described in different ways (esp at different historical time periods). Although with people's names there are some quite advanced authority files, which gives the 'best' name and a list of alternatives, no such authorities exist for places. Could this system, however, draw on a comprehensive list of place names (both UK-wide and street names in London) to start to give some shape and order to searching for places? Let's say I am searching for 'Fountain Tavern, Strand', it could give me - or link to - a list of all the addresses in the Strand. That would be a very useful way of seeing the data because even though most of the results will not be the Fountain Tavern (which actually will be quite easy to identify), with places it is often very useful to see other places nearby - other places that I wasn't searching for - in a way that isn't true of people's names. If I am searching for "John Smith"", it is of no interest whatever to learn that there was a different person called "Joan Smith" who happens to occupy the neighbouring place in the people authority file. But if I am searching for "two doors up from the corner of Chancery Lane" it is of interest to know that there is also an entry for the place described as "three doors up from the corner of Chancery Lane."
Is there another person with the same name? Can it give enough information to allow the researcher to choose the correct one?

Is person a named in the letters the same as person b named in an external source?

Is Sarah Croney the same person as Sarah Rees Jones

I want to identify personal names in Bede's 'Ecclesiastical History" and then see how many of them are of Anglo-Saxon origin and how many are of native British origin. This would include a search if the book itself, plus research and comparisons with a number of databases and online scholarly works.

Can person a be located in an online archival source, such as the census, parish register, civil registration or probate?

Were goldsmiths in York more likely to have Jewish names than goldsmiths in London?

I want to know how Irish immigrants to the uk came into the country in the late 19th century and if they changed their names. I want to start by looking at people and their names as they arrive in the UK, tracking their movements, employment and court records to see how many Irish names survive compared to how many people with Irish names entered the country over a specific twenty year period. I would need to search across census records, court records, employment records, workhouse records, immigration records etc and to compare names and calculate instances if as name used.
Appendix 4 : User Evaluation
Semantics-based Indexing of Historical Entities

http://geekscruff.me/peoplesparql

The task is to search for a person or persons of interest to you. Working through the steps below, see whether you are able to find out more about the person than you expected to, and whether research using this tool seems quicker than searching individual sources.

Please bear in mind that this is a prototype and that the research focus for my project is on the application of semantic technologies rather than user interface design. Please comment freely on any aspect of the tool and process, but please be aware of the health warnings at the bottom of the document which point to some of the limitations of the tool.

If your mind is blank, I have provided some search suggestions at the end of the document too. These help demonstrate strengths (and weaknesses).

<table>
<thead>
<tr>
<th>Overall Thoughts and Comments</th>
</tr>
</thead>
</table>

**Step One: Login**

Please login first as some of the later steps require a logged in user.

This should be relatively simple using a Google account (York Google accounts will work fine). The application has access to your email address and no further information.

I am not asking for comments on this as it uses third party code.
**Step Two: Search**

Choose one or more endpoints from the list and enter a search term.

Don't worry about lower- or upper-case, the following are all fine:

LELY / peter lely / te lawrence / T.E. Lawrence / d. h. lawrence
phil (will find phil, phillip, philip, phillipe etc.)

<table>
<thead>
<tr>
<th>Requirements</th>
<th>Satisfaction Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>See exact matches and likely matches</td>
<td>1 (low) - 5 (high)</td>
</tr>
<tr>
<td>Query multiple sources</td>
<td></td>
</tr>
<tr>
<td>Choose sources to search</td>
<td></td>
</tr>
</tbody>
</table>

**Comments**

If you know of a sparql endpoint that you’d like to add, please try out the add endpoint feature. The application analyses the endpoint to see if it is able add it. If not, it will let you know.

<table>
<thead>
<tr>
<th>Requirements</th>
<th>Satisfaction Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Add new data source to search</td>
<td>1 (low) - 5 (high)</td>
</tr>
</tbody>
</table>
Step Three : Refine Results

Once you have some results, you can use the ‘discards’ checkbox and button to refine these down to a set to explore. Ideally, for the purposes of the prototype, you should just leave one or two results to explore but you can explore more than that if you want to.

<table>
<thead>
<tr>
<th>Requirements</th>
<th>Satisfaction Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Filter search results</td>
<td>1 (low) – 5 (high)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
</tr>
</tbody>
</table>
Step Four: Explore

Once you have a set of results to explore, click ‘Explore’. This will show a new page with a list of your selected results.

The ‘More details’ button will show all information retrieved from the endpoint for each result.

<table>
<thead>
<tr>
<th>Requirements</th>
<th>Satisfaction Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>View full information about each name</td>
<td>1 (low) - 5 (high)</td>
</tr>
</tbody>
</table>

Comments

This might be enough for your research, but three further experimental features demonstrate the potential offered by semantic technology.
Are these the same person?

If you have two results, you can ask the application to try and establish whether they are the same person. If there are more than two results, only the first two will be checked (this is a limit I imposed for simplicity, it would be relatively easy to extend to comparing more than two names). If the application cannot determine the similarity due to lack of information it will try and extract related information. Please try this feature.

<table>
<thead>
<tr>
<th>Requirements</th>
<th>Satisfaction Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Research question: name comparison</td>
<td>1 (low) - 5 (high)</td>
</tr>
</tbody>
</table>

Comments

Enhance from external sources

In the Semantic Web space, there is a mechanism for explicitly asserting that two resources are the same. It uses a term from the OWL ontology called ‘sameas’. When querying for information about each of the selected names, the application looks for any sameas links. If it finds any, then these are listed separately and a button called ‘Enhance from external sources’ appears. The Artworld sample dataset, for example, contains sameas links to the Virtual International Authority File for Names (VIAF). If the button appears, please try this feature.

<table>
<thead>
<tr>
<th>Requirements</th>
<th>Satisfaction Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mixture of viewing full information and helping in comparison.</td>
<td>1 (low) - 5 (high)</td>
</tr>
</tbody>
</table>
Enhance with semantics

The final button is called ‘Enhance with semantics’. This uses a basic level of what the Semantic Web community call ‘inferencing’ to draw together entities of a similar type. To demonstrate this, selecting this option will process all of the links from the chosen resources to other resources (e.g. Charles Darwin is linked to a separate resource describing his wife Emma), determine their ‘type’ (Agent, Place, Event, Subject or other Thing) and group them together. Please try this feature, but be warned, it is resource intensive and make take a minute or two.

<table>
<thead>
<tr>
<th>Requirements</th>
<th>Satisfaction Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Follow links from people to other entities</td>
<td>1 (low) – 5 (high)</td>
</tr>
</tbody>
</table>

Comments
Health Warnings

This is a prototype intended to explore and demonstrate the use of semantic technologies. As such please be aware that:

- it will be slow, particularly
  - more endpoints searched
  - more results explored
  - more information that needs retrieving about each result
- you may get unexpected errors (please let me know if you do)
- it is running on a lowish spec server, this may not cope well with a number of users at one time
- the discards feature has some bugs that I am still working on so may not work fully as intended

Please try to avoid using ‘back’, if you’ve made an error or want to refine your query, please start over.

Please avoid very general queries (eg. bob), particularly on DBPedia and the British Museum Catalogue as these are very large data sources.

Storing data has not been implemented but the underlying data store can support access control and thus can allow a choice of what is and what isn’t made public.

Examples to try:

- Peter Lely in DBPedia and Artworld (good for sameas links)
- Phil Mercier in British Museum and Artworld (show word stemming and range of similar results)
- Marie Curie in British Museum and Nobel Prizes (good range of related ‘entities’, also will include a sameas link to pull in data from DBPedia)
- Charles Darwin in DBPedia (first two results are good for ‘is this the same person’)
- Will Smith in Linked Movie Database and British Museum (demonstrates a few limitations)