Towards Linked Data: Some Consequences for Researchers in the Social Sciences and Humanities

Lyne Da Sylva
Université de Montréal, Canada. Lyne.Da.Sylva@UMontreal.CA

ABSTRACT

This paper addresses the introduction of Semantic Web and Linked Data technology in Social Science and Humanities research. On the basis of a sample of existing research projects, we examine the impact that the technology has on the research methodology. Three main points of impact were observed. The first is the epistemological foundations of the research, including a focus on individual entities in the research area (or "atomization"), the reification of research objects, and the favouring of analytical skills from researchers. Secondly, in the data analysis phase, research relies more heavily on technical skills during data discovery and processing, and the technology induces growing confusion between data and metadata. Thirdly, we show how the types of models that are built often imply a new encoding of already existing information.

KEYWORDS

Linked Data, Semantic Web, research methods, Social Science and Humanities

INTRODUCTION

Within the global movement of data-driven science (Borgman 2015; Kitchin 2014; Hey, Tansley, & Tolle 2009), there has been an increasing amount of research in the Social Sciences and Humanities (SSH) which makes use of Semantic Web technology – specifically, Linked Data (Berners-Lee, 2006). It has already been recognized that Big Data is causing an epistemological shift in the way research is carried out (Borgman, 2012; Boyd & Crawford, 2012). A similar case can be made about Linked Data. The goal of this paper is to discuss certain implications of the Semantic Web (specifically Linked Data) for researchers in the Social Sciences and Humanities (SSH), on the basis of selected case studies.

This paper proceeds as follows. We first give brief descriptions of Semantic Web and Linked Data (SW/LD) technology and a schematic representation of SSH methodology. We then report on related work. In the methodology section, we present our case study approach, in which we have selected a sample of research projects in the SSH which make use of SW/LD technology; against the backdrop of the methodology outlined for research in the SSH, we identify within the case studies steps or areas which are affected by the SW/LD methods and tools. The conclusion provides some additional thoughts on the study.

BASIC CONCEPTS – METHODOLOGICAL TOOLS

We first briefly define the Semantic Web and Linked Data and provide a description of SSH research methodology.

Semantic Web and Linked Open Data

The goal of the Semantic Web (Berners-Lee, Hendler, & Lassila 2001) is to make explicit the meaning of information on the Web, in order to allow complex automated processing of the data. The first such application would be semantic search: to perform information retrieval on the meaning of a word and not on a character string, e.g. "file" as in "computer file" and not "nail file"; or "Gandhi" as in "Mahatma Gandhi" and not "Indira Gandhi". This is possible only if the targeted documents have an explicit representation of the meaning of the "file" and "Gandhi" occurrences which they contain. Thus, the Semantic Web relies on the identification and naming of entities on the Web: unique entities such as people ("Gandhi"), places ("Paris") and dates, but also concepts such as "file" or "beauty". It also requires the explicit representation of relationships among entities, such as "is-used-for", "is-the-author-of" or "knows". Such explicit encoding allows computer agents (devoid of real understanding) to manipulate and process data appropriately, according to their algorithms. This is the meaning of "Semantic" in "Semantic Web".

Another example of complex automated processing of semantic data is the manipulation of different sets of data, to be combined in complex ways. For instance, one could create a personalized sightseeing tour for a city, using information from various sources: a list of tourist attractions; a database of museums with their opening hours and their intended audience; attendance statistics for each of these venues; a current weather forecast; routes and timetables for public transportation to these venues. Appropriate technology could combine this information (and perhaps more, notably user profiles) to produce daily suggestions for sightseeing. Note that this may only be possible if (i) the data sources are sufficiently explicit and disambiguated, (ii) the exchange of data between them is made easy by strict adherence to a shared standard, and (iii) appropriate rules are provided for reasoning on the data. This is the essence of the Semantic Web.

Transferring this example to a context of research in the SSH, one can envision a data-driven application which would provide background for an analysis of a historical event: given for example a list of locations, time periods, a database of people, their roles and the relationships among them, a number of time-situated events (natural phenomena such as earthquakes or human-created ones such as the 2018 Winter Olympics), one could imagine providing possible factors which may help contextualize a specific event or state of affairs. Alternatively, this data could be used to uncover heretofore undetected relationships among people, places or events. Some such research projects have been developed, as will be shown in the following sections. Again, to make this possible, the data must be explicit, easily accessible and uniformly represented. And so, an important part of the Semantic Web is the Linked Data, which we describe below.

Brief description of Linked Data

The basic notions of Linked Data (LD) (https://www.w3.org/wiki/LinkedData; http://linkeddata.org/) include: identifiers to represent distinct entities uniquely; relation names which will be used to make explicit links among these entities; RDF triples which encode the relations among pairs of entities; vocabularies and ontologies which encode regular, productive relationships among sets of data and can enforce constraints on their use; and triplestores, the databases used to store the sets of triples. Each is described below.

Identifiers are alphanumeric strings (such as numbers or « codes ») used to represent entities in a unique manner. For instance, ISBN (International Standard Book Numbers) are identifiers for books in the physical world; in the virtual world the work corresponding to a book can be unambiguously identified with an identifier that can be a URL which leads to a description of it. An example is found in the French national library catalogue (Catalogue de la Bibliothèque nationale de France), where the novel "War and Peace" has as its identifier "ark:/12148/cb31478402c", declared at the following URL: http://catalogue.bnf.fr/ark:/12148/cb31478402c.

Identifiers are also used for people, e.g. 0000-0001-2242-4494 for Leo Tolstoy and 0000-0001-2138-6043 for Mahatma Gandhi in the International Standard Name Identifier (ISNI) database; or https://orcid.org/0000-0003-2125-060X for Albert Einstein in the ORCID (Open Researcher and Contributor ID) registry. Identifiers may also designate places (e.g. http://id.dbpedia.org/page/Vancouver), and various other entities. They are a shorthand for any entity which can be described and mentioned on the Web. In the Semantic Web, identifier syntax is governed by the URI (Uniform Resource Identifier) standard (https://www.w3.org/wiki/URI). Very often, identifiers for LD are the by-now familiar URLs (Uniform Resource Locators). Identifiers are usually associated with names, labels that may vary across languages: Leo Tolstoy, Léon Tolstoï, Толстой, Лев Николаевич, etc.; or Mahatma Gandhi, Mohandas Karamchand Gandhi, મોહનદાસ કરમચંદ ગાંધી, etc.

To allow automatic processing, relevant **relations** among entities are encoded explicitly. For example: "is-the-father-of" or "has-written-to" may be relations between persons or corporate bodies; "is-situated-in" can relate two geographical entities such as a city and a country. Persons may also be linked to entities such as works or other concepts, with relations such as "is-the-author-of" (conversely, "has-author") or "is-the-president-of". Other types of relations are those which express properties of a given entity: "has-publication-date-of", or "is-the-family-name-of". Intuitive relations such as "is-the-author-of", "knows" or "is-the-family-name-of" are defined formally and publicly in so-called **vocabularies** or **ontologies**: "Vocabularies are the basic building blocks for inference techniques on the Semantic Web" (https://www.w3.org/standards/semanticweb/ontology). They allow uniform encoding of relations. Some examples are given in Table 1.

The sameAs relation plays a special role in LD. It is used to relate two identifiers, defined in different datasets, that actually denote the same entity.

As the vocabularies and ontologies are defined on the Web, the relations are assigned identifiers: URIs which uniquely define them. And so not only entities, but also relations, are encoded with identifiers in the Semantic Web: http://purl.org/dc/terms/creator for "creator" and http://xmlns.com/foaf/spec/knows for "knows".

All data in the Web of data are represented by binary relations between two entities. The two entities and their relation are referred to as a **triple**, consisting of a subject (the first entity), the predicate (i.e. the relation) and the object (the second entity). Thus "Leo Tolstoy is-the-author-of War and Peace", or "Leo Tolstoy has-written-to Mahatma Gandhi" are examples of triples, as well as: "War and Peace has-publication-date-of 1869". Librarians will recognize some of these as metadata, but relations are of a more general nature. In the Semantic Web, triples use a uniform encoding, provided by the **RDF** (Resource Description Framework -- https://www.w3.org/RDF/) standard: it expresses a binary relation between the subject and the object, where the subject and the relation are represented by URIs, and the object may be another URI or a literal (i.e. a string of characters); the latter option is the obvious one to represent titles or names. Thus, the triples above are represented as in Table 2, once with human-readable strings, and once using RDF-compliant URIs.

Vocabulary	Main use	Sample relations
Dublin Core	Documents	creator date rights
FOAF (Friend of a friend)	People and relationships	firstName familyName knows
SKOS (Simple Knowledge Organization System)	Thesauri	broader narrower prefLabel
OWL (Web Ontology Language)	Ontologies	sameAs

Table 1: Sample relations from vocabularies or ontologies used in RDF triples

Subject	Relation	Object
War and Peace http://catalogue.bnf.fr/ark:/12148/cb31478402c	has-author http://purl.org/dc/terms/creator	Leo Tolstoy urn:ISNI:0000-0001-2242-4494
Leo Tolstoy urn:ISNI:0000-0001-2242-4494	knows http://xmlns.com/foaf/spec/knows	Mahatma Gandhi urn:ISNI: 0000-0001-2138-6043

Table 2: Sample relations from vocabularies or ontologies used in RDF triples

Triplestores are databases used to store the triples (they are basically relational databases optimized to process triples). Ideally, these databases are "open", meaning that their data are freely accessible in reusable formats. In this case they hold not only Linked Data (LD), but Linked Open Data (LOD), which ensures accessibility and access.

Note that the Semantic Web involves more than simply Linked (Open) Data; but our focus on LD suffices for the purpose of our study. LD defines the Web of data, which is an extension of the Web of documents.

Web of data vs Web of documents

The Web of data and the Web of documents are complementary. Triplestores hold triples of data; this data may designate documents themselves and may also represent references to entities mentioned within documents. Some implicit relationships among documents may be made explicit given that they share occurrences of the same entities. Different technologies may be used to extract, analyze and abstract entities from documents (to add to the Web of data), or they may be used to help create new documents from the entities in the Web of data.

The Web itself is the product of contributions from all Internet users: corporate bodies, government, individuals, various organizations, etc. For the present discussion, however, it is important to note that some of the documents are authored by researchers, and some of the data constitute the output of research projects. This means that research contributes new information to the Web of documents and to the Web of data – and is doing increasingly so; it is this research output and this new research material which we consider in our study. In fact, a document-centric discipline such as History has seen several developments incorporating LD and ontologies to analyze, structure and display historical information.

State of the LOD cloud

As a final introduction to SW/LD technology, we note the impressive growth of the LOD datasets available. In 2017 the LOD cloud was described as containing about 1139 datasets (http://lod-cloud.net/), divided among different categories, the main ones being geography, government, life sciences, linguistics, media, publications, and social networking. Many of these are RDF versions of highly-valued, dependable databases produced by recognized authorities. For instance, there are RDF versions of thesauri, subject headings and classification schemes from various national libraries (the USA, France, Germany, the UK, Canada, Australia, etc.). There are lists of geographical names from the world over, registries of authors, works, and artists; vocabularies for describing time, genes, drugs, etc. The most important perhaps, in terms of links to other datasets, is DBpedia (which encodes, in RDF format, triples of information extracted from Wikipedia). Given the size and growth of Wikipedia and the fact that it gradually contains more and more verified information, one can only expect DBpedia to grow as well; eventually, DBpedia, and consequently the LOD cloud, will encode large amounts of data deemed relevant for scholarly research.

A schematic representation of research methodology

Many publications, aimed at budding researchers or seasoned researchers alike, detail the steps for conducting research in the SSH (e.g. Bryman 2015; van Peer, Hakemulder, & Zyngier 2012). Abstracting from the particulars of each, we present the following outline for stages of a research project:

1. Defining the research problem: formulating research questions and hypotheses; performing the conceptual analysis; assessing feasibility. This is where the epistemological stance of the research becomes apparent, assuming a conceptual framework such as described in Baronov (2004), including positivism (or a variation thereof), structuralism and hermeneutics.

2. Designing the methodology: choosing an inquiry method, whether direct (using an "informant") or indirect (using documents or data), and whether quantitative or qualitative; and then, choosing a specific technique among several options (in direct methods: in situ observation, interviews, think-aloud protocols, questionnaires/surveys, diaries, focus groups, experimentation, etc.; in indirect methods: content analysis, statistics analysis, analysis of output from previous experiments, etc.).

3. Collecting data: choosing a sample and applying the data collection tools.

4. Analyzing and interpreting data: this stage begins with various types of qualitative, statistical or other quantitative processing of the data. This leads to interpretations and, it is hoped, conclusions which support the initial research hypotheses and answer the research questions. Included in this stage should be the creation of some type of model emerging from the research. Arsham (2015) identifies four types of models: verbal, physical, analytical and simulation models. The first type is usually developed, as a resulting thesis or article is a verbal model of the problem investigated, the data analyzed, and the conclusions drawn. It may be the case that other models are created as well: an analytical model in form of a graph or LD, or a simulation model (especially in fields related to Computer Science).

Unsurprisingly, we will see that the data collection and analysis phases of research are strongly impacted by SW/LD methodology; we will also see how the very definition of the research problem (or rather, the epistemological stance of the research) is also affected.

RELATED WORK

The use of SW/LD technology in SSH research is fairly recent; for instance, as of July 2018, few of the datasets that can be accessed via the datahub platform (https://old.datahub.io/), which lists some 11,300 datasets, are relevant to SSH. However, close to 1,000 of its datasets are tagged with "publications" and represent corpora of documents (digital libraries, national bibliographies, etc.), many of which are relevant for SSH scholars. Some (11) are identified with cultural heritage: e.g. LD from museums, such as the Rijksmuseum (https://old.datahub.io/), the British Museum Collection (https://old.datahub.io/) and Europeana (https://old.datahub.io/dataset/europeana-lod-v1). Also, 37 datasets are retrieved by a search using "Digital Humanities". Only two bear the "arts and humanities" tag or the "social science" tag. There is a much greater number of datasets tagged for biomedicine (249), science (133), earth science (68), biology (33) or physical sciences (13).

Among important ontologies for SSH, one can cite CIDOC-CRM, the Conceptual Reference Model from CIDOC (ICOM International Committee on Documentation). It "provides definitions and a formal structure for describing the implicit and explicit concepts and relationships used in cultural heritage documentation" (http://www.cidoc-crm.org/). The latter has been widely used already: Eide et al. (2008) describe the production of CIDOC-CRM compliant datasets from diverse collections of source data. Also, noteworthy is the Records in Contexts-Conceptual Model (RiC-CM) (https://www.ica.org/en/egad-ric-conceptual-model), "the new standard for the description of records based on archival principles" and the soon-to-be-released related RiC-O, Records in Context Ontology.

Some other notable work includes the following: in Literary Studies, Brando, Frontini, & Ganascia (2016) use LD (from DBpedia and *Bibliothèque nationale de France* Linked Data) to help disambiguate authors' mentions in a corpus of French literary criticism; in Linguistics, Chiarcos, Nordhoff, & Hellmann (2012) explore LD for language data and metadata; Hyvönen (2012) gives an overview on the motivations and the methods for using SW/LD technology for publishing cultural heritage collections and other content on the Web. Unsworth (2016) recognizes the role which the SW will play in Digital Humanities (DH) and the need for researchers within humanities to participate in the definition of ontologies, given their expertise. Additionally, a careful study of papers presented at the Digital Humanities conference in 2017 (Lewis, Raynor, Forest, Sinatra, & Sinclair, 2017) reveals a number of projects which use LD in the humanities, as shown in Table 3.

Discipline	Authors	Data encoded as LD
Art History	Bégnis & Mendes da Silva	Images and descriptive metadata (place names, building names)
Cultural Studies	Kameda & Hara	Documents and "words"
Digital Libraries	Dussault et al. Hardy et al	Document metadata from catalogue entries
	Page et al.	Document metadata, genre attributes
History	Beals	"Historical evidences", document provenance, rhetorical relationships
	Goto & Shibutani	Resources (i.e. documents), spatial information
	Grossner et al.	"Data about the movement of people, ideas, cultural practices, and commodities between places" (e.g. routes, people, commodities, information, places, paths, time)
Literary Studies	Bauer et al.	"Readings" represented by annotations
	Laiacona	Places, events, artwork, people, texts
	Malta et al.	Units of analysis in poetry
Musicology	Craig-McFeely et al.	Digital imaging of musical scores, cataloguing metadata, semantic notation (e.g. time, proportion, duration, roles of people/organizations)
	Nurmikko-Fuller et al.	Musicians and their properties, details of solos within performances (including pitch, key and chord changes), recordings
	Page, Lewis & Weigl	Audio recordings, musicological relationships
Philology	Lorenzini & Sanna	Document contents (e.g. events, activities)
Sociology	Brown et al.	Socio-demographic data related to authors or themes in texts

Table 3. LD projects at DH2017.

At the same conference, Brown et al. held a panel called "Advancing Linked Open Data in the Humanities", where questions were raised regarding the impact of LD-based research on scholars, and which tools would be most needed. A study similar to our own has addressed epistemological considerations of research using Big Data (Boyd & Crawford, 2012); indeed, Big Data entail important changes to the manner in which science is conducted (Hey et al., 2009) and in the very definition of knowledge (Boyd & Crawford 2012, 665). The observations encoded in the data suggest an objective character and a precision that actually obscure the subjectivity and inaccuracy inherent in the selection and preparation of data (Boyd & Crawford, 2012, pp. 666–668). This may very well apply to Linked Data as well. Epistemological aspects of the Semantic Web and of its Linked Data have been addressed by d'Aquin & Motta (2016, p. 53), who highlight among other things the illusion of uniformity or homogeneity among data, as suggested by the very regular format of RDF:

> Somehow, there is one aspect of scalability which is much harder to address by means of purely technical means. In d'Aquin et al. [2014a] we called it *diversity*: the fact that data and knowledge not only come in different formats and subscribe to different modeling principles, but also that they originate from different sources, might be of different scope and quality, and might be distributed under different constraints, with different regulations applying to them…

Baronov (2004) addresses the conceptual foundations underlying research methods, especially in the social sciences. Four methodological orientations are defined and contrasted: positivism (in three phases, embryonic positivism, logical positivism and postpositivism); structuralism; hermeneutics; and antifoundationalism. This discussion has been useful for framing some of the epistemological consequences presented below.

DATA: SSH RESEARCH USING SEMANTIC WEB AND LINKED DATA TECHNOLOGY

We focus now on a sample of research projects in SSH using Linked Data in disciplines well represented in Table 3: Literary Studies, Music and History. They were retained here for their variety, for the way their data or the ontologies used were clearly presented, and, for case #3, because it was well-documented. We briefly present the LD used in the research or resulting from it, as was gathered from the publications. The main focus of the descriptions is the use made of SW/LD technology.

Case #1: Literary Studies

Curado Malta, González-Blanco, Martínez Cantón, & Del Rio (2017) use LD in a literary project which explores poetry: "…our work is interested in a structural and formal approach that looks at poetry into discrete units, categories, and their relationships. Thus, we are involved in analyzing how metrical repertoires in digital form model those structures" (p. 210). It aims to bridge the gap between alternative, linguistically diverse databases called "repertoires of poetry metrics", i.e. catalogues that "give an account of the metrical and rhythmical schemes of either a poetical tradition, a period or school, gathering a corpus of poems that are defined and classified by their main characteristics" (p. 210). Already existing repertoires (currently in various formats: MySQL or XML databases and files, Perl scripts and Worksheet files) will be published as LD, allowing heretofore closed silos to become interoperable. Thus, the linking of all this data should make it possible to compare poetic traditions and to create new repertoires, among other benefits.

Case #2: Music

The second case studied involves Music Informatics ("the study of information related to, or [which] is a result of, musical activity" (Humphrey, Bello, & LeCun, 2013, p. 461)), an interdisciplinary research area dealing with the production, distribution, consumption, and analysis of music through technology (especially in digital formats). Cannam et al. (2010) present the Sonic Visualiser, an application to visualize, annotate, and automate the analysis of audio recordings. It is a "highly customisable play-back and visualisation environment that includes such features as variable-speed playback, looping, and the ability to annotate the recording, for instance to identify specific points of reference"[1]. This research project makes use of a variety of existing ontologies, including the Music Ontology[2]; the Similarity Ontology which provides concepts and properties for describing similarities between things in the RDF/OWL framework; the Audio Features Ontology, "a descriptive framework for expressing different conceptual-isations of and designing linked data formats for content-based audio features" (Allik, Fazekas, & Sandler, 2016, p. 73). More ontologies have been created within the project: The Chord Ontology, "intended to provide a common, versatile vocabulary for describing chords and chord sequences in RDF"[3] and the Timeline Ontology which "used with the Event ontology, can be used to annotate sections of a signal, a video, or any temporal object"[4]. They encode, in this new format, information which previously existed in another form. The heavy use of ontologies by the project makes it of particular interest for our study. Great emphasis is put on individual, low-level structures which participate in music production, analysis and visualisation. The paper claims that with their tool, the interchange of data, the analysis of experimental results, and resulting outputs are rendered relatively easy. The underlying requirement is the use of standardized, uniform data. This is a case where vocabularies and ontologies provide readily available, standardized descriptions for the building blocks of the application.

This project creates and uses technology possibly applicable to the data alluded to in the Musicology examples in Table 3.

Case #3: History

The use of Linked Data in History research has been well documented in the work of Michon (2016); the author's rich descriptions of existing projects as well as the insightful analysis of their consequences provided valuable material for this case study. The specific case used here stems from the work of historian Tim Sherratt, who uses LD in various projects. One of them presents findings on a specific character, Inigo Jones, who was a meteorologist and farmer in Queensland, Australia. The research project resulted in not only a book (Sherratt 2007), but also an enriched document on the Web, which told the story of Inigo Jones while making use of various datasets from the LOD cloud to provide outgoing links to material related to the document: source documents, as well as links to people, places, events, etc. External datasets used included archives of newspaper articles from Trove, "Australia's national discovery service", which "provides access to more than 300 million resources managed by more than 1000 Australian and international organisations, and by members of the public" (Ayres, 2012). Data from Wikipedia and DBpedia were also used. Data encoding used Dublin Core and FOAF vocabularies. In the author's own words, he proceeded to "create narratives that embed Linked Open Data" (Sherratt, 2015). Thus, in his research methodology, data collection included LD sources as well as traditional print materials (the mainstay of History research). The research output combined not only the addition of LD to a document (a purely formal result), but also additional links made between existing data, based on the researcher's analysis of existing data and documents (novel associations which embody a new model of the phenomenon under study).

This is an illustration of how technology (LD) can modify the way research is conducted (data gathering, collection and analysis) as well as the product of the project.

Case #4: Classics

Classics studies are akin to History in relying heavily on documents (ancient ones in fact) and so might be deemed quite similar to our latest case. The SPQR ('Supporting Productive Queries") project (Blanke et al., 2012) was retained nonetheless as highly relevant to our study, as it presents itself as "Linked data for humanities research".

[1] http://www.charm.rhul.ac.uk/analysing/p9_0_1.html
[2] http://musicontology.com/
[3] http://motools.sourceforge.net/chord_draft_1/chord.html
[4] http://motools.sourceforge.net/timeline/timeline.html

The project investigated the potential of a SW/LD approach "for linking and integrating datasets related to classical antiquity, focusing on certain targeted datasets as test cases". In the field of Classics research, ancient texts are an important source of information. Even though a large amount of digital material has been created, it can be difficult to access due to its being distributed across locations, stored in diverse and incompatible formats, and either not available online or made available only in isolation (Blanke et al., 2012). The SPQR project uses LD principles and technology to link the existing datasets, "to deliver an integrated data landscape through which researchers can explore and so seek to understand this data" (Blanke et al., 2012, p. 1). Specifically, three datasets were transformed into RDF: the Heidelberger Gesamtverzeichnis der griechischen Papyrusurkunden Aegyptens (HGV), a collection of metadata records for 65,000 Greek papyri from Egypt (300 BC to 700 AD); the Inscriptions of Aphrodisias (InsAph), a corpus of 2,000 ancient Greek inscriptions from the Roman city of Aphrodisias in modern Turkey (200 BC to 700 AD); the Inscriptions of Tripolitania (IRT), a corpus of over 1,000 inscriptions from modern Libya (100 BC to 700 AD). Also, a SPQR-specific ontology was created: "The SPQR ontology consisted of URIs for a range of epigraphic concepts. The subject of a triple can be an epigraph, a date of origin, a person or a location." (Blanke et al., 2012, p. 3). Entities in their ontology were linked to other external sources, using the OWL 'sameAs' relationship described above. The result was deemed to offer "an intuitive and usable means of exploring and understanding the data, exceeding the capabilities offered by current online portals to classics data" (Blanke et al., 2012, p. 1). The goal of the project was thus the very creation of this LD resource.

The projects we have described claim that access to new datasets in a standardized form and in easily accessible places allow for new discoveries and new opportunities. Our goal here is not to question these efforts. It is simply to observe what implications this research agenda may have on research in the SSH.

DATA ANALYSIS: IMPLICATIONS OF SW/LD TECHNOLOGY FOR RESEARCH IN SSH

We have found that three aspects of the research methodology are especially impacted by the introduction of LD and SW technology: first, the epistemological stance of SW/LD-driven research; secondly, data collection and analysis; and thirdly, the nature of models created by this type of research.

Impact on epistemological stance

The adoption and use of LD and SW technology prove to have important consequences in the very outlook on research and on how they determine the orientation of a research project. We explain here: (i) how they change the focus of research (how they tend to "atomize" it), (ii) how they reify the research objects themselves, and (iii) how they may affect the terminology of all concerned disciplines.

Focus of research: "atomization" of research objects

Research based on LD favours the analysis of piecemeal information, as opposed to whole documents or overviews of a given situation. First and foremost are the individual entities which are assigned identifiers. Each and every entity participating in the research problem must be represented as a URI. In the case of the Inigo Jones project, this includes not only authors of documents and important historical figures, but also weather events, meteorology itself, different place names, etc. Note that there is no hierarchy *per se* in linked datasets (apart from the importance that can be attached to an entity which is linked to very many others); all are represented by identical-looking identifiers, which all become equally-small singular entities in the research space.

Another example of piecemeal information of major importance in the Semantic Web is the relations defined by individual triples. These are very specific, linking exactly two entities with a precise, named relation. Note that non-binary relations cannot be represented by triples. And yet most interesting phenomena are not binary. For example, an act of buying involves a seller, a buyer, some goods and some type of currency. This simple example requires many triples to represent it completely. One can only imagine what is required to describe the most basic historical event or musical form. In all cases, however, individual triples are the units that are stored in triplestores. Although triples related to the same entity can be grouped when displayed for a user, they are individually encoded, stored and processed.

LD-driven research favours a methodological orientation in line with a form of positivism – i.e. assuming that the entities defined in the Web of data represent objective, shared knowledge. A structuralist approach is not excluded: if one considers the totality of datasets and their links, together with the relations defined within ontologies, one may arrive at a description of the system as a whole and each piece's role within it. It is hoped that future work will allow a more holistic approach, but in these rather initial stages of the Semantic Web endeavour, the focus of research has been to define the URIs, to declare the ontologies and to create the datasets. And so, to focus on individual entities.

This is contrary to the usual synthetical approach of some disciplines. In disciplines such as History, Literature, Linguistics or Anthropology, the traditional approach is for the researcher to study the primary data, which include documents, phenomena (e.g. speech production) and observations (in field studies, for example). The researcher's important contribution is his or her analysis of this complex corpus of data; the relationships that are identified and the information that is abstracted allow a better understanding of the research question. The research results take the form of documents which provide a synthesis or a narrative of the data and phenomena under study. In a SW/LD approach, in addition to traditional sources, a researcher uses datasets that have been encoded by others. These datasets are typically isolated silos of (piecemeal) information. Then, the researcher's contribution must include the selection of relevant datasets (thus the evaluation of existing datasets), the encoding of any new data uncovered by his or her analysis of the entire corpus of documents, data and observations, and finally the process of adding any relevant new links in the final data. Among the research results, then, must figure the addition of more Linked Data. In a word: whereas the traditional approach requires more synthesis and abstraction from the researcher, a SW/LD-driven approach includes a strong analytical component, to decompose observations into individual data and triples – what we call here the "atomization" of research objects. "Big data is made up of many small acts of living" (Sherratt, 2016).

Reification of research objects

Interestingly, the creation of identifiers and triples entails the reification of research objects, i.e. raising to the status of some type of concrete object something that may in fact be abstract. All entities are represented by identifiers: weather events, musical chords, time, languages such as Greek, etc. It actually may not seem unnatural for researchers to reify their research objects (this may well be a natural process in a researcher's work); but it may be surprising to treat relations as reified entities. How does one interpret "is-the-author-of" as an entity? Or "has-date-of-birth"? And yet this is what is done in a sense when these relations are assigned identifiers. This may raise interesting new research questions.

One that has already received attention is the identity relation, through the definition of the "sameAs" relation in the OWL ontology. Different issues related to the (abusive) use of the sameAs relationship have been identified. Davis (2009) has noted the influence of time on the validity of the identity relationship (for example, three different datasets may give different populations for the same city, which may be due to the time when the population was counted). Halpin, Herman, & Hayes (2010) specifically looked at different notions of "is the same as," distinguishing four subtypes (valid but opaque references, context-variable similarity, representation, and quasi-identity).

Levelling of terminology

While terminology and concepts are more stable in the so-called Exact or Pure Sciences, they show less consensus in SSH (this has been widely documented in the construction of documentary thesauri). One can expect a greater number of competing terminological databases, in other words, of competing ontologies. This may have a number of possible consequences. One of them can be the proliferation of sameAs-based triples: a growing number of entities in different ontologies may be recognized as actually being the same entity – or at least similar enough to warrant a sameAs link. Eventually this can lead to more and more links between entities that are increasingly dissimilar. A different consequence of competing ontologies may actually be a tendency towards standardization: the fact that the data is linked allows interconnections (and a potential for standardization) that are not observed or required when competing research teams work independently. And in fact, work on ontologies may turn out to have a significant effect on the definition of the basic concepts of the disciplines and the attempts at bridging across disciplines. Work in terminology and thesaurus design could have a major impact in cross-disciplinary (as well as basic disciplinary) research.

Impact on data collection and analysis

In our study of SW/LD-based research, we have identified two main impacts on data collection and analysis. First, there are technical implications associated with the efforts required to gather data, and the skills necessary to process them. Secondly, a SW/LD-driven research method leads to a blurred distinction between data and metadata.

Technical implications

The efforts required to gather data in the SSH is much greater than it is the Pure Sciences. For SSH, where large amounts of data still lie in historical documents, in music partitions, in dictionaries, in field manuals, in legal documents (let alone the tacit, undocumented data in social practice), data must be gathered manually, by intellectual methods: reading, interpretation, observation. This is in sharp contrast to the Pure and Applied Sciences, where large amounts of data are being gathered today by automatic means: by probes or sensors, by extraction from databases, by computer generation. So the data-driven science craze has very different implications and may suffer more resistance in the SSH.

The technical computer skills needed to tackle a research project where LD plays an important role are not mastered by all researchers. It requires a formal and analytical approach which may be at odds with some research methods or indeed with some entire fields of study. Specific training must be provided regarding the basics of Semantic Web technologies. New skills will be needed – on "elements of mathematics, logic, engineering, and computer science" (Unsworth, 2016, p. 46). This may be disconcerting. For instance, the historian will need to concern himself with the structure of the data and no longer simply with their delivery (Michon, 2016, p. 17).

Blurred distinction between data and metadata

As suggested by the examples above, some triples are clearly metadata (e.g. Tolstoy is-the-author-of War and Peace). Others introduce true relationships between two entities with an equivalent status (e.g. Tolstoy has-written-to Ghandi). Yet they are treated in the same manner by the RDF triples. If triples are created by extracting data from primary documents, then in a sense all such encoding becomes metadata. But if all is metadata, what is the data? Do the data disappear? Or is it rather the concept of metadata which disappears?

Impact on resulting model

Our observations lead to what French researchers (Pédauque, 2007) have described as « Redocumenting » (« Redocumentarisation »). This refers to the process, in the late 20th and early 21st century, of large-scale digitization of existing print documents: "redocumenting" is performing anew the process of documenting human endeavours. A similar process took place with the invention of the printing press, when handwritten documents were gradually typeset and printed. And so the encoding of triples in RDF is, in fact, another process of taking existing information (contained in narrative or unstructured documents already on the Web) and transcoding it into a new form. The output of research then includes an analytical model consisting of Linked Data. This new form is deemed necessary to allow the sophisticated automated semantic processing by software agents. It is thus seen as somewhat unavoidable. And, given the simple binary nature of RDF triples, this redocumenting leads to the atomization of research described above.

CONCLUSION

We have examined some research projects which take advantage of possibilities afforded by SW/LD technology. Our goal was to pinpoint aspects of the research methodology which are impacted by the technology. Our examination of these cases has revealed three main points of impact. The first is the epistemological foundations of the research; the focus is on individual entities in the research area, leading to what we have called the "atomization" of research. This in turn favours analytical skills in the researcher rather than his or her ability to synthesize and abstract away from specific phenomena. In addition, all research phenomena are reified – raised to a concrete status in their representation by URIs, regardless of their original characteristics. Secondly, and unsurprisingly, it is the data analysis phase which is affected the most; it relies more heavily on technical skills during data discovery and processing. It obscures the distinction between data and metadata, thus requiring even more analytical skills from the researcher to diligently differentiate between the two as necessary. Thirdly, we have shown how the model-building aspect of research is affected by SW/LD-driven approaches, in the sense that existing models of knowledge (i.e. existing documents) require a transcoding into this new format in order to facilitate further research.

The scale of the efforts to produce LD datasets is such that it cannot be ignored, and research in SSH must take it into account, as so much useful data are now available. Transformations to the conduct of research is inevitable. But it is important to understand how this shift may affect research methodology. What is needed then is new solutions for new challenges.

Some additional thoughts: the implications vary for different disciplines. Data-oriented disciplines, such as Economics or Demography may see little change. For metadata-oriented disciplines (Library Science, Museum Studies, etc.), the shift entails fairly straightforward transformations (with the caveat that data and metadata may be difficult to distinguish). It is the document-oriented disciplines that should see the greatest change – History, Literature or Sociology, for instance. And yet it can be argued that the basis of data analysis is at the heart of such disciplines. In the words of (Sherratt 2015):

> *Let's think for a moment about the work of a historian — identifying actors, defining relationships, documenting the complex networks that bring together people, places and events over time. It's painstaking, exhilarating and potentially soul-destroying work. It's also an exercise in data modelling. Whether the results are preserved in a triplestore, a spreadsheet, or on a drawer full of index cards — it's nodes and edges, it's entities and relationships, it's data.*

REFERENCES

Allik, A., Fazekas, G., & Sandler, M. (2016). An Ontology for Audio Features. In *Proceedings of the 17th ISMIR Conference* (pp. 73–79). New York (NY).

Arsham, H. (2015). *Applied Management Science: Making Good Strategic Decisions*. Retrieved from http://home.ubalt.edu/ntsbarsh/opre640/opre640.htm

Ayres, M.-L. (2012). *Digging deep in Trove: Success, challenge and uncertainty*. National Library of Australia. Retrieved 29 March 2018, from https://www.nla.gov.au/our-publications/staff-papers/digging-deep-in-trove-success-challenge-and-uncertainty

Baronov, D. (2004). *Conceptual foundations of social research methods*. Boulder; London: Paradigm Publishers.

Berners-Lee, T. (2006, July 27). *Linked Data - Design Issues*. Retrieved 3 April 2018, from https://www.w3.org/DesignIssues/LinkedData.html

Berners-Lee, T., Hendler, J., & Lassila, O. (2001). The Semantic Web. *Scientific American*, 28–37.

Blanke, T., Bodard, G., Bryant, M., Dunn, S., Hedges, M., Jackson, M., & Scott, D. (2012). Linked data for humanities research: The SPQR experiment. In *6th IEEE International Conference on Digital Ecosystems and Technologies (DEST)* (pp. 1–6). https://doi.org/10.1109/DEST.2012.6227932

Borgman, C. L. (2012). The conundrum of sharing research data. *Journal of the American Society for Information Science and Technology*, 63(6), 1059–1078. https://doi.org/10.1002/asi.22634

Borgman, C. L. (2015). *Big Data, Little, Data, No Data. Scholarship in the Networked World*. Cambridge, MA: MIT Press.

Boyd, D., & Crawford, K. (2012). Critical Questions for Big Data. Information, *Communication & Society*, 15(5), 662–679. https://doi.org/10.1080/1369118X.2012.678878

Brando, C., Frontini, F., & Ganascia, J.-G. (2016). REDEN: Named Entity Linking in Digital Literary Editions Using Linked Data Sets. *Complex Systems Informatics and Modeling Quarterly*, 60–80. https://doi.org/10.7250/csimq.2016-7.04

Bryman, A. (2015). *Social Research Methods*. Oxford University Press.

Cannam, C., Sandler, M., Jewell, M. O., Rhodes, C., & d'Inverno, M. (2010). Linked Data and You: Bringing Music Research Software into the Semantic Web. *Journal of New Music Research*, 39(4), 313–325. https://doi.org/10.1080/09298215.2010.522715

Chiarcos, C., Nordhoff, S., & Hellmann, S. (Eds.). (2012). *Linked data in linguistics: representing and connecting language data and language metadata*. Berlin ; New York: Springer.

Curado Malta, M., González-Blanco, E., Martínez Cantón, C., & Del Rio, G. (2017). A Common Conceptual Model for the Study of Poetry in the Digital Humanities. In *Digital Humanities 2017 Conference Abstracts* (pp. 210–213). Montréal: ADHO.

d'Aquin, M., & Motta, E. (2016). *The epistemology of intelligent semantic web systems*. San Rafael, California: Morgan & Claypool. Retrieved from http://www.morganclaypool.com/doi/pdf/10.2200/S00708ED1V01Y201603WBE014

Davis, I. (2009). *Representing Time in RDF Part 1*. Retrieved 1 February 2017, from http://blog.iandavis.com/2009/08/representing-time-in-rdf-part-1/

Eide, O., Felicetti, A., Ore, C. E., D'Andrea, A., & Holmen, J. (2008). Encoding Cultural Heritage Information for the Semantic Web. Procedures for Data Integration through CIDOC-CRM Mapping. In D. Arnold, F. Niccolucci, D. Pletinckx, & L. Van Gool (Eds.), *Open Digital Cultural Heritage Systems Conference* (pp. 47–53). Budapest: Archeolingua. Retrieved from http://culturalinformatics.org.uk/sites/culturalinformatics.org.uk/files/Rome_Pro_20111012.pdf#page=47

Halpin, H., Herman, I., & Hayes, P. J. (2010). When owl:sameAs isn't the Same: An Analysis of Identity Links on the Semantic Web. In *RDF Next Steps Workshop*. Palo Alto, CA. Retrieved from https://www.w3.org/2009/12/rdf-ws/papers/ws21

Hey, T., Tansley, S., & Tolle, K. (2009). *The Fourth Paradigm: Data-Intensive Scientific Discovery*. Redmond, VA: Microsoft Research.

Humphrey, E. J., Bello, J. P., & LeCun, Y. (2013). Feature learning and deep architectures: New directions for music informatics. *Journal of Intelligent Information Systems*, 41(3), 461–481.

Hyvönen, E. (2012). Publishing and Using Cultural Heritage Linked Data on the Semantic Web. *Synthesis Lectures on the Semantic Web: Theory and Technology*, 2(1), 1–159. https://doi.org/10.2200/S00452ED1V01Y201210WBE003

Kitchin, R. (2014). *The Data Revolution: Big Data, Open Data, Data Infrastructures and Their Consequences*. SAGE Publications.

Lewis, R., Raynor, C., Forest, D., Sinatra, M., & Sinclair, S. (Eds.). (2017). *Digital Humanities 2017 Conference Abstracts*. McGill; University; Université de Montréal: ADHO. Retrieved from https://dh2017.adho.org/abstracts/DH2017-abstracts.pdf

Michon, P. (2016). *Vers une nouvelle architecture de l'information historique : L'impact du Web sémantique sur l'organisation du Répertoire du patrimoine culturel du Québec*. Master's thesis. Université de Sherbrooke, Sherbrooke. Retrieved from http://hdl.handle.net/11143/8776

Pédauque, R. T. (2007). *La Redocumentarisation du monde*. Paris: Éditions Cépadues.

Sherratt, T. (2007). *Inigo Jones : the weather prophet*. [Melbourne] : Bureau of Meteorology. Retrieved from https://trove.nla.gov.au/version/38963121

Sherratt, T. (2015, April 10). *Stories for Machines - Data for Humans*. Retrieved from http://discontents.com.au/stories-for-machines-data-for-humans/

Sherratt, T. (2016, August 25). *Telling stories with data – discontents*. Retrieved 29 March 2018, from http://discontents.com.au/telling-stories-with-data/

Unsworth, J. (2016). What is Humanities Computing and What is Not? In M. Terras, J. Nyhan, & E. Vanhoutte, *Defining Digital Humanities. A Reader* (pp. 35–47). London: Routledge.

van Peer, W., Hakemulder, F., & Zyngier, S. (2012). *Scientific methods for the humanities*. Amsterdam: John Benjamins Publishing Company.