USING RDF TO DESCRIBE AND LINK SOCIAL SCIENCE DATA TO RELATED RESOURCES ON THE WEB

By Stefan Kramer, Amber Leahey, Humphrey Southall, Johanna Vompras, and Joachim Wackerow

2012-08-10

DDI Working Paper Series – Semantic Web, No. 1

This paper is part of a series that focuses on DDI usage and how the metadata specification should be applied in a variety of settings by a variety of organizations and individuals. Support for this working paper series was provided by the authors' home institutions; by GESIS - Leibniz Institute for the Social Sciences; by Schloss Dagstuhl - Leibniz Center for Informatics; and by the DDI Alliance.

Using RDF to Describe and Link Social Science Data to Related Resources on the Web

LEVERAGING THE DATA DOCUMENTATION INITIATIVE (DDI) MODEL

INTRODUCTION

This document focuses on how best to relate Resource Description Framework (RDF)-described datasets to other related resources and objects (publications, geographies, organizations, people, etc.) in the Semantic Web. This includes a description of what would be needed to make these types of relationships most useful, including which RDF vocabularies should be used, potential link predicates, and possible data sources. RDF provides a good model for describing social science data because it supports formal semantics that provide a dependable basis for reasoning about the meaning of an RDF expression. In particular, it supports defined notions of entailment which provide a basis for defining reliable rules of inference in RDF data[1].

Our findings are discussed in the context of social science data and more specifically, how to leverage existing metadata models to use alongside linked data. We provide a case for leveraging the Data Documentation Initiative (DDI) to enable semantic linking of social science data to other data and related resources on the Web. This document is organized into five use cases, which we consider in turn. Use cases include: linking related publications to data, linking data about people and organizations to research data, linking geography, linking to related studies, and linking data to licenses. We briefly discuss emerging or known issues surrounding the potential use of linked data within each of the defined use cases. Following these, we list more topics that could develop into additional use cases. Appendix A lists elements from the DDI-Codebook and DDI-Lifecycle specifications that are relevant to each use case.

Semantic/Linked Data Web

The Linked Data Web holds great promise for users and information professionals alike, especially as we begin to expose the meaning and define relationships between and among digital objects. The vision of the Semantic/Linked Data Web is an attempt to facilitate translation and interoperability in and among digital objects in the Web though connecting, sharing, and defining linkages between data and information using unique identifiers (URIs) and RDF. From this, users can discover resources based on more implicit knowledge and interconnections of linked data. This vision of the Web offers many lofty goals, and revitalizes problems that have traditionally challenged information and data professionals, such as those relating to the discovery, classification, and organization of information.

[1] http://www.w3.org/standards/techs/rdf

Semantic Web technologies can offer many possibilities for enhancing the discovery and use of digital objects, including social science data and metadata. We can envision the potential of the Linked Data Web in a use case like the following:

A researcher is interested in finding the research outputs described in a particular publication. The researcher discovers that a publication links to a data table and additional documentation. This data table is part of a larger dataset, based on a multi-wave study. The researcher then goes on to discover other similar datasets that relate to a particular theme, topic, or variable found within the original data table first discovered. The researcher can also discover a profile for the principal investigator or contributing authors, and find other publications they may have authored or co-authored, or been interested in. This may lead to further contact and collaboration between the two (or more) researchers. (Scenario adapted from: Gregory, A. & Vardigan, M. (2010). The Web of Linked Data: Realizing the Potential for the Social Sciences.)

In effect, enabling semantic linking improves the discovery of resources, accuracy of searches, and potential collaboration between and among people, and digital objects.

The use of semantic technologies like XML schemas, including the Resource Description Framework (RDF), goes beyond the limits of simple meta-tagging, to introduce further domain specificity and semantic understanding. The use of unique identifiers or URIs, such as Digital Object Identifiers (DOIs), facilitates the exposure and connectivity described by RDF-data on the Web. In effect, searching and discovery of information and knowledge can be more advanced, complex, and meaningful. Furthermore, through semantic technologies (RDF) and query languages (such as SPARQL), there can be more meaningful user-driven sharing on the Web. From this, we may think about the future of the Web in terms of information that is well-defined, predictive, and providing enhanced communication between people and computers that is more productive and meaningful.

Use Case 1: Linking related publications

Publications that make up the scholarly literature, including publications based on statistical datasets that may be documented with DDI, are captured in bibliographic and full-text databases and research-oriented Web search engines[2] that are already widely used by researchers. Enabling two-way linking – that is, from data to related publications and from publications to underlying data – is ideal because it facilitates literature reviews, demonstrates impact of the data, and provides recognition to data creators and to authors.

To make such two-way linking possible, we might add unique, persistent identifiers of the type that are already widely used in scholarly publishing, such as Digital Object Identifiers (DOIs), into the DDI-based metadata for datasets, so that these datasets become easily and unambiguously citable *in* research output publications, and thereby linkable and discoverable *from* research publications. Similarly, adding citations and links (widely available already in the form of DOIs) from DDI metadata for datasets to known research publications based on these datasets can lead the researcher to subsequently written relevant publications. These can be publications emanating directly from primary research by the principal investigator who

[2] Examples of bibliographic and full-text databases include Medline, JSTOR, Google Scholar, Web of Knowledge/Web of Science, etc.

collected the data; subsequent publications based on data reuse and secondary analysis; or publications describing study methods, design of the study, theories behind it, etc.

An example of connecting datasets and research literature is already evidenced by the ICPSR Bibliography of Data-Related Literature. Additionally, the growth of the scientific communications landscape and a view towards data provenance has led to a need to archive the research data underlying research publications for reuse. Development in the area of enhanced publications[3] could benefit from an open Linked Data Web in which datasets are semantically linked and related to all predecessor and subsequent data. RDF-described data could enable the linkage of data to other resources data on the Web in a fairly de-centralized and non-linear fashion. DDI can support the semantic description of those linkages by providing the context for which linkages can be made and understood on the Web. This may lead us to envision a research environment in which researchers can actively discover, find, and access datasets related to publications including journal articles and other research outputs, and vice versa.

Relevant entities in DDI

The exposure, sharing, and connecting of pieces of data and information described in DDI records to information and linked data on the Semantic Web requires consideration of the relevant entities and elements in DDI. Two main branches or development lines of the DDI specification – DDI Codebook (DDI-C) and DDI Lifecycle (DDI-L) – exist in practice and identifying these entities within versions of DDI is useful for this discussion.

In DDI-Codebook, OtherStudyMaterials and its sub-element RelatedMaterials describe materials related to the study that are primarily related to the content and use of the study, such as appendices, sampling information, weighting details, methodological and technical details, publications based upon the study content, related studies or collections of studies, etc. These are associated with a URI and given a description. Often, this can take the form of bibliographic citations and can include a single URI or a series of URIs comprising multiple citations /references to external materials, which can be objects as a whole (journal articles) or parts of objects (chapters or appendices in articles or documents). OtherStudyMaterials maps to the Dublin Core Relation element.

This ability to define relations and assign URIs to any identifiable element is improved in the DDI-Lifecycle specification. Primarily, this use case is concerned with study unit linkages to related publications and other materials related to the study being described. This can include publications that present the study's main findings or other related findings, information on a study or related design, sampling methods, instruments, procedures, etc. It is also possible to relate publications at the variable/data specification level. In DDI-Lifecycle, relating information and knowledge at various levels of the DDI modules can be enabled through use of the repeatable OtherMaterial element. Types are defined by free-text, e.g., "publication," "report," "results." Referencing from an OtherMaterial as an external resource to a DDI item includes using the Relationship/ RelatedToReference element. The definition of RelationshipType is described by a relationship to other items and the item within the DDI instance to which it is related. Assigning URIs using OtherMaterials (DDI-L) is quite granular in fact, in that you can point to/from any identifiable element within DDI (variables,

[3] See, for example: www.surffoundation.nl/enhancedpublications

values, catagories, methods, etc.) using OtherMaterials. Attaching DDI elements with a predictable URL enables all elements of the metadata to be linked to RDF-data.

Target data on the Web

The following resources may be applicable to this use case.

- Academic publications on the Web, perhaps from a publication database that has metadata in RDF
- DOIs (found in DOI registries such as DataCite, Web of Science, etc.) used for the assignment of predictive URIs for citing data, linking publications, etc.
- Vocabularies used to describe the target such as Dublin Core, Bibliographic Ontology (BIBO), and other such bibliographic ontologies
- Colibrary (see: http://collab.di.uniba.it/Colibrary/books/)
- PloS One (see:http://www.plosone.org/home.action)
- W3C report "Library Linked Data Incubator Group: Datasets, Value Vocabularies, and Metadata Element Sets": http://www.w3.org/2005/Incubator/lld/XGR-lld-vocabdataset-20111025/

Possible link predicates

The "subject-predicate-object" structure is at the core of the Resource Description Framework data model underlying Linked Data. For this use case, we might create link predicates such as:

- ddilink:backgroundPublication: Theoretical background of the study
- ddilink:methodologyPublication: Methodical background of the study
- ddilink:resultsPublication: Presentation of main results; publication based on study

Possibly one would want more a detailed description of the relationship between the DDI entity and the publication, for example: "This publication describes the sampling method that was used." In this case a simple predicate would be insufficient, and one would like to have an intermediate resource that describes in more detail the relationship between the DDI entity and publication.

Example target resources

These resources offer some examples of linking data and publications or related functionality.

- RKB Explorer (http://www.rkbexplorer.com/explorer/)
- data.semanticWeb.com (metadata repository for computer science publications)
- Many eprints servers produce RDF metadata, for example: http://repec.org/docs/eprints2redif.pl

Issues with this use case

Various models that reflect and support relations between objects in the world of scholarly publishing exist and are not well defined. Relations that may seem obvious to some are often not shown or supported by others. Additionally, the linkage to publications at various levels relies on a multitude of factors that may exist

outside the realm of responsibility of the data producer, manager, or depositor. These factors include issues surrounding quality, legal and copyright issues, and integrity related to the intended use, among others.

It should also be noted that when data are cited properly with persistent identifiers like DOIs in the References section of publications, harvesters like Google Scholar organically make the connections and linkages between data and related publications through the identifiers. Similarly, the decentralized nature of RDF lessens the burden of maintenance as compared to other mechanismism for defining and delivering data on the Web.

Use Case 2: Linking people and organizations

Researchers and contributors to statistical datasets (e.g., data analysts) often maintain online profiles about themselves in commercial social networking services (e.g., LinkedIn, XING) or on the Web sites of their institution or department (e.g., VIVO). By adding links from these profiles to the data they have generated in their research, ideally using persistent identifiers (such as DOIs) that data repository managers have assigned to the data, the data become discoverable for reuse, validation, etc., by other researchers via the originating researchers' profiles. If data repository managers add links to the profiles of the researchers and contributors involved in generating a dataset housed in the repository, and also to any existing Open Researcher and Contributor ID (ORCID: http://orcid.org/) records, not only will this lead users of the dataset to information about the researcher, but indirectly also to the researcher's publications and datasets.

Relevant entities in DDI

Identifying individuals, their affiliates, and organizations is captured in all versions of the DDI. Particularly, in DDI-Lifecycle, Organization (which might be identified as (a) person(s), or agency) is associated with each step in the data lifecycle. The element "Relation" (in Individual) describes relations between two actors (organizations and/or individuals). This includes the identification of individual researchers and their relationship to other researchers, and organizations.

Target data on the Web

- Detailed description of the organization or person (agent) on the Web of data, which would be particularly interesting if further information were associated with the agents (e.g., who do they work for, who funds them, what else have they done)
- Vocabularies: Friend of a Friend, or FOAF, Epimorphics Organization Ontology, "vCard in RDF" vocabulary

Possible link predicates

- owl:sameAs. This assumes that we assign a URI to the agent in the DDI-RDF and then would have an owl:sameAs link connecting that URI to another URI for the agent in the other dataset.

Example target datasets

- FOAF profiles

- University metadata, e.g., data.southampton.ac.uk, http://thedatahub.org/dataset/vivo-cornell-university and other VIVO deployments
- research.data.gov.uk -- see http://ckan.net/dataset/research-data-gov-uk
- CORDIS database as linked data: http://thedatahub.org/dataset/fu-berlin-cordis

Issues with this use case

Linking any data that relate to identifiable individuals or specific organizations may lead to issues surrounding authority and privacy. On the Web, it is very easy to have multiple identities or information that is not authoritative. As the famous caption from *The New Yorker* cartoon reads, "On the Internet, no one knows you're a dog." A digital footprint often lacks context, comprehensiveness, timeliness, and authority. While there are many approaches to tackling the issue of authority, including ranking lists and credibility scores, the fact of the matter is that control over information sources when it comes to people and social relationships is still relatively unstable on the Web.

Linking people and information can sometimes lead to information privacy concerns. Current legislation around privacy on the Web remains somewhat unclear; however, in recent court cases it has been shown that in linking personal information to an individual, that information may reveal private attributes that were not authorized for release or even for collection, potentially constituting an invasion of privacy.

Use Case 3: Linking geography

In some senses, "geography" is simply an instance of a controlled vocabulary: lists of US states, European countries, etc., appearing as the possible answers to a survey question or a category in a tabulation. However, geography is especially important as (a) a key basis on which statistical users wish to subset data, and (b) a system of categories which are very rarely designed by statisticians. In most cases, statistical agencies have to follow the system of areal units defined by their governments, and formally delineated by national mapping agencies. Although non-government research projects are in principle free to define their own system of regions, they rarely do so. This is at least partly because defining a detailed geographical partitioning of a country needs a large body of data incorporating geographical coordinates, which can be manipulated only using specialized GIS software. This is therefore a context where creators of statistical datasets need to link to external data resources, or are even required by law to do so (such is the case in the UK).

Although an external definition of "geography" might simply be a set of identifiers, a number of gazetteers have already been published as RDF with embedded point coordinates, including Geonames and the geo-referenced entities within dbpedia. However, statistics are usually recorded for areas, not points. To date, the main example of reporting **areas** being defined as RDF is the Ordnance Survey's listings of UK administrative units.

For example, these data taken from a presentation on the Datacube Vocabulary list life expectancy within Welsh Unitary Authorities:

	2004-6		2005-7		2006-8	
	Male	Female	Male	Female	Male	Female
Newport	76.7	80.7	77.1	80.9	77.0	81.5
Cardiff	78.7	83.3	78.6	83.7	78.7	83.4
Monmouthshire	76.6	81.3	76.5	81.5	76.6	81.7
Merthyr Tydfil	75.5	79.1	75.5	79.4	74.9	79.6

The top left data value, 76.7, is male life expectancy in 2004-6 for the County Borough of Newport, and in the Datacube example that translates to the following RDF triples, presented here using Turtle:

```
<http:…dataset/le1/newport/2004/M> a qb:Observation;
 qb:dataSet      <http:…/dataset/le1>;
 eg:refArea      os:7000000000025499;
 eg:refPeriod    <http://reference.data.gov.uk/id/…
  gregorian-interval/2004-01-01T00:00:00/P3Y>
 sdmx-dimension:sex sdmx-code:sex-M;
 eg:lifeExpectancy 76.7.
```

The third line in the above Turtle code uses an external reference to a URI defining Newport published by the Ordnance Survey, os:7000000000025499, which resolves to an RDF document identifying the administrative unit and its relationships with other units, expressed semantically using predicates such as "contains" and "touches," but derived algorithmically from polygon data: http://data.ordnancesurvey.co.uk/id/7000000000025499.

The "extent" predicate within that document is then used to link to a separate RDF document specifying an "abstract geometry" for Newport, which begins as follows: http://data.ordnancesurvey.co.uk/id/geometry/71446.

```
<rdf:RDF>
 <rdf:Description rdf:about="http://data.ordnancesurvey.co.uk/id/geometry/71446">
 <rdf:type rdf:resource="http://data.ordnancesurvey.co.uk/ontology/geometry/AbstractGeometry"/>
 <geometry:asGML rdf:parseType="Literal">
  <gml:Polygon srsName="os:BNG">
   <gml:exterior>
    <gml:LinearRing>
     <gml:posList srsDimension="2">
      325302.2 177813.4 325303.3 177813.8 ….
```

This document consists of a minimal wrapper of RDF around Geographical Mark-up Language (GML), a namespace defined by the Open Geospatial Consortium (OGC); and in fact the bulk of the document is simply a sequence of coordinate pairs defining the boundaries of Newport. For now, this uses the "abstract geometry" class defined as part of the Ordnance Survey ontologies. A comment within the RDF notes that this is "a superclass of all geometry types such as points, lines and polygons. This is currently a place holder class

and likely to change when some standard way of representing geometries in RDF is agreed." John Goodwin, who built this system for the Ordnance Survey, notes "I suspect the way boundaries have been encoded in the OS linked data will not change much when GeoSPARQL is a standard - hopefully it will mainly be a change of namespaces for the geometry ontology" (e-mail of 19th September 2011).

Embedding geospatial data within the semantic Web is a major topic of current research, with a substantial involvement from the Open Geospatial Consortium (OGC). They have defined GeoSPARQL as an extension enabling coordinate data to be included within semantic queries. This would mean, for example, that statistics covering a given location could be identified via linkage to external polygon definitions, although it seems likely that these would need to be held on the same triple store for acceptable performance. OGC published a candidate GeoSPARQL standard in 2011, and the opportunity to comment closed in August 2011. Defining this standard will not mean that available triple stores necessarily support spatial querying, although where the stores are based on databases which already support such querying, such as Oracle and Postgres/PostGIS, this should be relatively easy to achieve: http://www.opengeospatial.org/standards/requests/80.

Linkage between statistics and geography might occur at the dataset level, and simply indicate the area within which the data were gathered. However, the real power of geographical linkage depends on identifying the variables or dimensions within the dataset that identify which geographical area specific cases or cell counts relate to. Once this is done, linkage to external data defining polygons would enable the right software to (a) create actual maps from the data, using the polygon definitions to create base maps, and (b) subset data based on area.

Relevant entities in DDI

Geographic coverage of data is dealt with in all versions of the DDI. In DDI-Codebook this is expressed in multiple elements within the Study Description section. Information on the Geographic Coverage of the data includes the scope or geographic coverage of the data, and any additional levels of geographic coding provided in the variables. In DDI-Codebook this maps to the Dublin Core Coverage element.

In DDI-Lifecycle, relations between geographies are defined by GeographyStructure (e.g., ParentGeography). The type of Geography is captured in GeographyDomain (structures the response domain for a geographic point to ensure collection of relevant information). GeographicLocation contains information on the specific geometry of areas and can be expressed as coordinates. This is defined in the dataset (examples include countries and sub-national levels). These geographic areas can be defined within the DDI instance or an external structure can be referenced. Geographic references can also exist for organizations, persons, title statements, etc.

Target data on the Web

- Linked Data identifiers assigned to geographic entities (online gazetteers)
- From these identifiers, more information such as bounding boxes and boundaries might be discoverable

Possible link predicates

- owl:sameAs from codes in the DDI-RDF to the external URI on the Web
- rdfs:seeAlso when linking to vernacular geographies, e.g., dbpedia -- found within, beside, etc.
- owl:unionOf

Example target datasets

Databases of physical features:

- LinkedGeoData.org is an RDF version of Open Street Map, and currently comprises about 2 billion triples, gathered mainly by volunteers using GPS.
- http://geo.linkeddata.es provides such data for Spain, and in more generalized form for Europe.
- National mapping agencies may have similar databases of features, such as the UK Ordnance Survey's MasterMap system, which certainly could be exposed as RDF, but there are large copyright issues.

Place gazetteers and vernacular geography:

- **DBpedia.org** derives from Wikipedia and contains very large numbers of entries tagged with point coordinates.
- **Geonames.org** brings together the US Board on Geographic Names and National Geospatial-Intelligence Agency gazetteers with certain other official sources, and then extends them via crowd-sourcing. This resource provides URIs and an RDF view of entries, but no current SPARQL end point. It holds only a single point coordinate even for the largest areas, such as "Europe".

URIs for reporting geographies linking to polygons:

- UK work was described above, and is available as RDF and via a SPARQL endpoint:
  - http://data.ordnancesurvey.co.uk
  - http://api.talis.com/stores/ordnance-survey/services/SPARQL
- At the time of writing, the Ordnance Survey may well be the only national mapping agency publishing reporting geographies as Linked Open Data. The Spanish GeoLinkedData.es initiative, noted above, is experimenting in this area and may have material available.
- Open Street Map is attempting to include boundaries, but coverage and reliability are unclear: you cannot map boundary lines with a GPS.

URIs for reporting geographies without coordinate data:

The following provide URIs for administrative units, but currently do not provide links to coordinate data:

- The UN Food and Agricultural Organisation (FAO) maintains the FAO Geopolitical Ontology as RDF, providing a URI for each nation state. As of September 2011, it included South Sudan,

which became independent in July 2011. It also included certain defunct states, such as the Soviet Union, with end dates:
    - http://www.fao.org/countryprofiles/geoinfo/geopolitical/resource
  - Eurostat makes available URIs for European countries and regions, and also a list of the nation-states of the world derived from the CIA World Fact Book. However, this site appears at the Free University of Berlin and it is unclear how it is being sustained, so for nation states the FAO listing is preferrable:
    - http://www4.wiwiss.fu-berlin.de/eurostat/directory/countries
    - http://www4.wiwiss.fu-berlin.de/eurostat/directory/regions
    - http://www4.wiwiss.fu-berlin.de/factbook/directory/countries
  - URIs for the sub-divisions of the United States, even the states, are problematic,. Data published by the US Census Bureau from the 2000 Census have been converted into RDF by Joshua Tauberer, creating about 1 billion RDF triples. This includes a wealth of geographical information, but the site provides only downloads and a SPARQL endpoint, so there are not resolvable URIs:
    - http://www.rdfabout.com/demo/census

Historical units

The US National Historical GIS provides digital boundaries for historic states and counties within the US. The Great Britain Historical GIS provides identifiers for a large number of historical units and unit types in Britain, Estonia, Ireland, and Sweden, plus the nation-states of Europe since 1815, although polygonal boundaries are held only for a subset. Neither system is currently accessible as RDF but the GB system is actively investigating this, and also exploring the potential for a global listing of historical units in collaboration with the Center for Geographic Analysis at Harvard and the Center for World History at the University of Pittsburgh.

  - http://www.nhgis.org
  - http://www.port.ac.uk/research/gbhgis

Issues with this use case

In mainstream GIS, the world is defined as a continuous space of coordinates, not as sets of discrete objects, so conventional GIS systems are not easily worked with using Linked Data constructs. Systems such as Google Maps and Bing Maps may be relevant to constructing map-based user interfaces, but do not currently expose discrete features with URIs. Similarly, one cannot link via RDF to an Open Geospatial Consortium-defined Web **Map** Server, but conversely one could in principle link to an OGC-defined Web **Feature** Server. Target datasets need to be catalogues of geographical items -- gazetteers, not maps, but ideally gazetteers which include definitions of the items as points, lines, or polygons.

There are, however, three broad types of items which can appear in geographical catalogues:

- **Physical features** that exist in the landscape such as houses, roads, rivers, mountains -- things you can touch, and measure with surveying equipment. These are the main focus of national mapping agencies and, in open data, of the Open Street Map project which collect data mainly by assembling data

gathered using GPS systems. However, most statistics do not relate to physical features, not even to large ones such as oceans and mountain ranges.

- **"Places"**. These are tags that people use to refer to locations that have particular meaning to them, and especially to locations where groups of people live: "populated places". Note that a settlement is a grouping of houses, and therefore of physical features, but the decision that a particular set of houses forms a settlement needing a name is a subjective one, and this is even truer of the delineation of parts of cities. Given that places are defined and named entirely through popular usage, crowd-sourcing is arguably the most appropriate way of gathering information about them; modern methods of aerial survey and satellite mapping gather no information about them at all. However, the typing applied to the major crowd-sourced gazetteers complicates semantic use. For example, "Dagstuhl" is defined in DBpedia as a "computer science research center," while Geonames defines it as a "populated place," meaning the village. Linking such items, directly or indirectly, via the owl:sameAs predicate could lead to the deduction that villages and research centers are the same thing. rdfs:seeAlso may be safer.
- **Administrative divisions and other statistical reporting units**. Administrative divisions are defined in law, by national and local governments. Their boundaries sometimes follow landscape features, but often follow arbitrary straight lines, for example, most of the boundary between the USA and Canada. Most statistical data are necessarily gathered for areas, not points or lines, and while surveys and projects could define their own systems of reporting units, in practice most base their reporting geography on administrative geographies. We therefore need to link statistics to formal definitions of reporting units published by the relevant governments, and including definitions of the actual boundaries, as in the UK example discussed above. As listed above, it is currently easier to find URIs for "places" than for administrative units, especially if associated coordinates are required, but "places" will rarely be appropriate objects to link statistics to. Most national mapping agencies already have the necessary digital data, and the availability of appropriate URIs is changing fast.

The other large issue is, of course, what to link from. Linking whole datasets to a geographical unit defines overall coverage, but does not indicate the amount of detail: is it a national total or village by village? Linking individual data items to geographical units enables new approaches to sub-setting as well as geo-spatial analysis, but will require much greater effort and technical integration to expose underlying data elements.

Use Case 4: Linking related studies

Researchers who are not already familiar with the most relevant statistical datasets for their work usually seek datasets on the particular topic (or related topic) of their research, among other possible attributes (such as geographic coverage and granularity, time period). They may also want to find studies that were based on or followed prior studies, not limited to longitudinal designs continually undertaken by the same organization. This would be particularly useful for social science literature reviews, and content analyses. The datasets may be available from the same provider (e.g., ICPSR, UKDA, GESIS), which has already subject-indexed their data holdings, but they may also be available only from different providers. An example from the area of political election results broken down by county in the USA are Dave Leip's Election Atlas and POLIDATA. By

asserting relationships between such studies in their respective DDI-based documentation, the researcher can be informed by one study of other potentially relevant studies.

Relevant entities in DDI

DDI entities such as Series, Topic, Data Collection Situation, Data Sources, Related Study, and Other Materials, all provide useful scenarios for linking to related resources. Defining relations between and among data series, data topics, sources of data, and related studies may provide researchers with a better understanding of the connectivity of studies, datasets, and research publications.

Target data on the Web

- ICPSR (most data descriptions available in DDI-XML)
- Council of European Social Science Data Archives (CESSDA) (most data descriptions available in DDI-XML and the UK Data Archive may have some data already described in RDF) and other data archives
- DOIs for datasets (i.e., International Polar Year data)

Possible link predicates

- owl:hasValue
- owl:intersectionOf
- owl:unionOf
- rdfs:subClassOf

Example target datasets

- ICPSR repository http://www.icpsr.umich.edu/icpsrweb/ICPSR/
- CrossRef (metadata for 46 million DOIs -- http://crossref.org/)

Issues with this use case

A challenge is the lack of a widely accepted common controlled vocabulary for describing studies topically, including the challenge of domain specific ontologies. While most data archives use a standard vocabulary, often their institutional context presents barriers to crossing discipline-specific concepts and notions. When we describe the world of data, for instance, it is not assumed that we are describing all the data in the world, and that it can be described within a common understanding and context.

Use Case 5: Linking to licenses

When data are covered by a specially-negotiated license, that license clearly needs to form part of the metadata. However, much of the data created by government agencies and academic projects is covered by one of a limited number of "open" licenses, generally allowing at least non-commercial use without payment. The exact license used will have a large impact on, for example, how different datasets may be integrated,

and on semi-commercial uses. As a result, researchers may well want to locate datasets which are covered by the same or compatible licenses.

The best known organization defining open licenses is Creative Commons (http://creativecommons.org). They maintain a family of licenses at fixed addresses which can be used as URIs. For example, this links to the full legal text of their Attribution-Noncommercial-Sharealike license: http://creativecommons.org/licenses/by-nc-sa/3.0/legalcode.

Creative Commons have also done work on expressing the licenses themselves in RDF, which would greatly assist in doing automated searches for data which were covered by not identical but compatible licenses: http://creativecommons.org/ns.

- The Free Software Foundation has worked with Creative Commons to restate their licenses as RDF: http://www.fsf.org/blogs/licensing/2009-06-rdf.
- An example of an open license developed to cover government data is the UK's Open Government License for public sector information. This can currently be reached at the following address: http://www.nationalarchives.gov.uk/doc/open-government-licence/open-government-licence.htm.

Relevant entities in DDI

- Study unit
- Variables (e.g., if there are embargoes on specific variables like income)

Target data on the Web

- License URIs
- URIs identifying additional conditions such as community norms

Possible link predicates

- dc:rights
- cc:license
- Waiver Vocabulary, http://vocab.org/waiver/terms/

Example target datasets

- Creative Commons licenses
- cc0 (CC Zero) – Creative Commons license for public domain
- Open Data Commons Public Domain Dedication and License (PDDL)
- PDDL community norms (for legally non-binding behavioral norms such as "attribute me if you use this)

Issues with this use case

One issue is the applicability of licenses created in one jurisdiction when the data are used in another, and the consequent implications for datasets assembled from supposedly "open" data from different countries. One striking difference between the US and Europe is over the so-called "moral right" of authors, and presumably dataset creators, to be acknowledged and for the integrity of their work to be respected. A lack of

consistency in a political and legal framework for data sharing and copyright presents additional challenges to a Linked Data Web.

Additionally, the issue still remains about how to identify data in the public domain, which is a legal status rather than a license. A researcher can place their work in the public domain, for example, by using the Open Data Commons Public Domain Dedication and License (PDDL): http://opendatacommons.org/licenses/pddl. However, this is not really relevant to a data librarian recording that data created by, say, the US government is in the public domain. One solution would be to link to dBpedia: http://dbpedia.org/page/Public_domain. What is clear is that the predicate linking the dataset to the license should be that provided by Dublin Core, i.e., dc:rights.

In the longer term, RDF may enable specially written licenses to be defined in a way that means they can be machine processed to establish which datasets meet specific requirements.

Topics for future consideration

The use cases presented in this paper were selected as yielding the most potential benefit to the community of social science data users; however, the contributors also contemplated other potential linkages between DDI instances and external resources that might be developed into use cases in future work. These include, in no particular order:

1. **Organizations that are expected to receive data in the future**: In the chain of custody of a dataset, it may already be known which organization(s) or unit(s) will, or should, be receiving the data in the future. For example, following a funding agency-required data management plan, a research team may already have arranged to have their university's library or data archive house their research data for long-term preservation after the completion of their project. If DDI-based metadata could link to information about the future-stages-of-lifecycle custodians of the described data, it would aid in providing researchers, administrators and funding agencies assurance about the future availability and preservation of the documented dataset.

2. **Related materials**: It may be useful to be able to discover additional information about the funding for a study, or other studies funded by grants from same agency. Coverage of a study in news/media and other non-scholarly information dissemination channels could also be enabled.

3. **External events that may have influenced collected data**: Linked Data structures might be able to help data users understand factors that may have influenced the study outcome reflected in the data and reported in newspapers, etc. -- for example, a major strike that occurred during the administration of a survey on attitudes towards employers; or a plane crash that occurred and was reported on while interviews about traffic safety were conducted in a population.

4. **Outputs from statistical analyses**: Linking from datasets to analyses that were run against them, but not published (graphs, tables, etc.) can be useful. Typically, these would likely not be an external resource from the organization that archives and documents the dataset, but in a distributed

repository environment (e.g., "institutional" vs. "data" repository at a large university), this could be possible.

5. **Access control processes/levels employed**: In a virtual research environment where access levels to data within an organization are defined and described, the DDI instance could conceivably be used to declare which of these levels should be given access to the dataset(s) it describes.

6. **Certification of repository that houses data**: Some organizations undergo a formal certification of "trustworthiness" for their digital repository, or use the certification criteria as a checklist to help assure its technical and organizational stability. Examples of such certification efforts are the original *Digital Repository Certification* project by the Research Libraries Group and the National Archives and Records Administration; the subsequent *Trustworthy Repositories Audit & Certification (TRAC) Criteria and Checklist*; the *Data Seal of Approval* by DANS; and the *Certificate for Document and Publication Services* by DINI. While such a certification is an attribute of the repository in which data may be stored, an argument could be made for having a DDI instance link to the information about the certification criteria and processes at the time data are deposited in a certified repository, as an indication that at the time the data deposit was made, it was under the assumption that the repository met such criteria.

7. **Information about data quality**: Linkage to external descriptions of processes and guidelines followed for assuring the quality of the data could allow DDI-based metadata to assure potential future users of a dataset to gauge the data's quality. This could also serve a statistical literacy purpose.

8. **Linking to information on statistical methods**: Documenting the full life cycle of a dataset would clearly mean recording what analytical techniques have been applied to it, and the ultimate statistical download system would arguably be able to suggest "related datasets," partly based on which other datasets had had the same analytical techniques applied to them. More critically, some derived datasets can only be fully described by identifying the specific techniques used to derive them from raw data. Note that there are three slightly different use cases identified above, and this affects what predicates are used: (a) this technique was applied to this dataset to create some unspecified output; (b) this technique was applied to this dataset to create this specific publication; and (c) the technique was applied to dataset A to create dataset B. For now, the best source of URIs for statistical techniques is probably dbpedia, as Wikipedia includes a wide set of articles on specific techniques which appear to be reasonably scholarly; for example: http://dbpedia.org/page/Kolmogorov-Smirnov_test.

APPENDIX A

Appendix A provides detailed information relating to the DDI elements that can serve as the end-points to RDF expressions that describe linkages between data and resources.

Use Case 1: Linking related publications

DDI Codebook 2.x	DDI Lifecycle 3.x
Study: titlStmt, serName, subject, keyword, topcClas, abstract, sources (data) relMat: titleStmt, producer, etc., relStdy ---titleStmt, producer, etc., relPubl--titlStmt, producer, etc., OtherRefs, var: labl.	OtherMaterials (element "Citation" for bibliographic information), type is defined by free-text, e.g., "publication" or "report" or "results". Definition of a reference from an OtherMaterial as external resource to a DDI Item (it could also be a publication) by using the Relationship/RelatedToReference element. The definition of relationtype by RelationType.

Use Case 2: Linking people and organizations

DDI Codebook 2.x	DDI Lifecycle 3.x
AuthEnty (Author of data collection [2.1.2.1] dc terms), othId, producer, fundAg, distrbtr, depositr.	Organization and Individual. Individual may be assigned some properties (keywords, position, email, or a researcher ID within a specified system). The elements Individual and Organization are connected by a n:m relationship. Other relations between Individuals in the context of the organization may be described by the Relation element.

Use Case 3: Linking geography

DDI Codebook 2.x	DDI Lifecycle 3.x
nation, geogCover, geogUnit, geoBndBox, boundPoly, polygon, point, universe.	Relations between Geographies are defined by GeographyStructure (e.g., ParentGeography). The type of Geography is captured in GeographyDomain (structures the response domain for a geographic point to ensure collection of relevant information). GeographicLocation contains information on the specific geographic areas defined in the dataset such as cities, countries, or states. The areas can be defined within the DDI instance or an external structure can be referenced.

Use Case 4: Linking related studies

DDI Codebook 2.x	DDI Lifecycle 3.x
timeMeth, origArch, serStmt, timePrd, collDate (cycle ex.), sources, collSize, complete, fileQnty, relStdy --- titleStmt, otherMat	Finding related studies by TopicalCoverage. DataCollection/Methodology/TimeMethod describes the time method or time dimension of the data collection.

Use Case 5: Linking to licenses

DDI Codebook 2.x	DDI Lifecycle 3.x
dataAccs, setAvail, accsPlac, origArch, avlStatus, collSize, complete, fileQnty, useStmt, confDec, specPerm, deposReq, conditions, disclaimer	Access restrictions can be definded at multiple levels (e.g., Archive, Collection, or Item level). Definition of DefaultAccess in Archive/ArchiveSpecific for the archive in general. The restrictions noted at this level apply to all holdings of the archive unless overridden for specified collections or items in the archive.

APPENDIX B

ACKNOWLEDGMENTS

The paper is one of several papers that are the outcome of a workshop held at Schloss Dagstuhl - Leibniz Center for Informatics in Wadern, Germany, September 11-16, 2011. The photo on the cover page shows the sculpture of a lion in the garden of Schloss Dagstuhl.

Workshop Title:

Semantic Statistics for Social, Behavioural, and Economic Sciences: Leveraging the DDI Model for the Web

Organizers:

Richard Cyganiak, National University of Ireland
Arofan Gregory (Open Data Foundation, Tucson, Arizona, USA)
Wendy Thomas (Minnesota Population Center [MPC])
Joachim Wackerow (GESIS, Leibniz Institute for the Social Sciences, Germany)

Link: http://www.dagstuhl.de/en/program/calendar/evhp/?semnr=11372

The authors of the paper would like to acknowledge others who participated in this workshop.

- Archana Bidargaddi, Norwegian Social Science Data Services (NSD)
- Thomas Bosch, GESIS - Leibniz Institute for the Social Sciences
- Franck Cotton, National Institute of Statistics and Economic Studies (INSEE)
- Richard Cyganiak, Digital Enterprise Research Institute
- Daniel Gillman, U.S. Bureau of Labor Statistics
- Arofan Gregory, Open Data Foundation (ODaF)
- Marcel Hebing, German Socio-Economic Panel Study (SOEP), DIW - Berlin - German Institute for Economic Research
- Jannik Jensen, Danish Data Archive (DDA)
- Stefan Kramer, Open Data Foundation (ODaF)
- Amber Leahey, University of Toronto
- Olof Olsson, Swedish National Data Service (SND)
- Abdul Rahim, Metadata Technology Inc., North America
- John Shepherdson, United Kingdom Data Archive (UKDA)
- Humphrey Southall, University of Portsmouth
- Wendy Thomas, Minnesota Population Center (MPC)
- Johanna Vompras, University of Bielefeld
- Joachim Wackerow, GESIS - Leibniz Institute for the Social Sciences
- Benjamin Zapilko, GESIS - Leibniz Institute for the Social Sciences

APPENDIX C