Multidisciplinary Research on Geographical Information in Europe and Beyond
Proceedings of the AGILE'2012 International Conference on Geographic Information Science, Avignon, April, 24-27, 2012
ISBN: 978-90-816960-0-5
Editors: Jérôme Gensel, Didier Josselin and Danny Vandenbroucke

An Infrastructure for Publishing Geospatial Metadata as Open Linked Metadata

James Reid
EDINA,University of Edinburgh
Causewayside House,
160 Causewayside, EH9 1PR
Edinburgh, UK
james.reid@ed.ac.uk

William Waites
Visiting Researcher, Laboratory for
Foundations of Computer Science
School of Informatics, University of
Edinburgh, UK
wwaites@tardis.ed.ac.uk

Ben Butchart
EDINA,University of Edinburgh
Causewayside House,
160 Causewayside, EH9 1PR
Edinburgh, UK
b.butchart@ed.ac.uk

Abstract

This paper outlines work that was carried out to investigate how geospatial metadata might be transformed and published as Open Linked Data. A linked-data corpus was produced by harvesting extant metadata catalogues for content and a range of transformation and publishing methodologies explored to develop a pipeline infrastructure for geospatial metadata linked data production. In the context of the project three strategies were considered – translation in the style of a metadata “crosswalk”, mapping to existing relational database schema and derivation direct from UML. These alternative strategies and their attendant pros and cons are reported on.

Keywords: metadata, linked data, vocabularies,harvesting

1 Introduction

EDINA[1]as a JISC[2] national data centre makes available, inter alia many geographical datasets and services for use by the UK academic community. To support discovery of these resources it operates metadata catalogues that one can search and browse to find specific items of interest. These metadata are typically made available both in HTML form for human consumption and in a machine readable form, often as ISO XML for compliance with international obligations.

Catalogue systems which publish geospatial metadata and which provide interfaces for their query and maintenance have been promoted by the Open Geospatial Consortium[3] (OGC) for some time. The Catalogue Services (CS) specification which provides just such a standard method of defining, querying and organising stores of geospatial metadata [1] is the de facto standard for discovering information about geospatial data and services in a standards compliant fashion. The CS specification defines an HTTP protocol binding named Catalogue Services for the Web (CSW) which underpins resource discovery across and within Spatial Data Infrastructures (SDIs) such as the UK Location Information Infrastructure. The UK academic sector has its own thematic SDI and already has its own geospatial discovery service through GoGeo[4].

Independent of this, there is a growing trend to make such metadata catalogues available in a uniform way using techniques from Linked Data[5] and the semantic web. Reasons for this include the ability to correlate metadata across different catalogue implementations that may use very different internal representations, to facilitate linking and annotation of the metadata by third parties and to provide a basic level of referential infrastructure on top of which more difficult questions of provenance and quality can be addressed.

2 Context

An example of the broader relevance of CSW and geospatial metadata for discovery purposes, is the recommendation issued in the context of INSPIRE by the Network Services Drafting Team [2] (2008) to SDIs in the European Union to derive the base functionality of discovery services from the ISO profile of CSW. However, CSW is, arguably, not ideally suited for the modern Web where search engines are the users’ default gateway to information[6]. The reasons why CSW might be regarded as sub-optimal from a purely 'discoverability' perspective are (adapted after [3]):

- Search engines are poorly optimised to index Deep Web databases. The term 'Deep Web' refers to the database content that is effectively hidden from search engines behind Web forms and back-office applications. Surfacing Deep Web content is a research problem that has concerned the search engine community since its description by [4]. From this perspective, the underlying content of SDI metadata repositories are opaque and hidden behind catalogue application interfaces; as such, SDI metadata forms part of the Deep Web. Hence, the 'findability' via search engines depends on the success of crawling processes that require the analysis of the Web interface, and then the automatic generation of appropriate queries.

- Applications are increasingly becoming Linked Data friendly. The Linked Data community, which has grown significantly over the last few years, promotes a Web of data based on the architectural principles of the Web [5]. Linked Data is less a technology than a set of best practices for publishing, sharing and connecting data and information using Uniform Resource Identifiers (URIs) that are resolved to Resource Description Framework (RDF) documents. RDF is a W3C recommendation for modelling and exchanging metadata [6]. As an aside, the original Geography Markup Language (GML) model was based upon RDF and vestiges of this ancestry is still evident today. In the UK, the Cabinet Office has published guidance

[1] edina.ac.uk
[2] www.jisc.ac.uk
[3] www.opengeospatial.org
[4] www.gogeo.ac.uk
[5] linkeddata.org/

[6] Several studies of research behaviour in the UK HFE sector support the view that 'googling' is a default reflex when seeking out resources. As noted by in 'Information behaviour of the researcher of the future', UCL, 2008, “they [resource providers] need to make their sites more highly visible in cyberspace by opening them up to search engines”

on the design of URIs[7] which forms the basis for the UK Location Programmes approach to publishing Linked Data[8] (see below).

- The evolution of metadata vocabularies towards RDF based models. Well known metadata vocabularies have evolved to models based on RDF with an emphasis on the linking of metadata descriptions. The abstract data models of Dublin Core Metadata Initiative (DCMI)[9] and the Open Archive Initiative (OAI)[10] have evolved alongside development of the RDF data model. This has resulted in abstract models based on the RDF data model [7] which emphasis the use (and reuse) of entities rather than using plain literals as the value of properties. This evolution ultimately enables the effective hyperlinking of metadata and traversal queries using query languages and protocols, such as SPARQL [8].

Whilst CSW are undoubtedly useful to enable discovery [9] and indeed are essential to the development of regional, national and global SDIs, they are nevertheless disjoint with the operational model of Deep Web crawlers. Popular search engines have developed several techniques to extract information from Deep Web databases without previous knowledge of their interfaces:-

> “The operational model for Web crawlers described in Raghavan (2001), based on (1) form analysis, (2) query generation and (3) response analysis is widely accepted. It models queries as functions with n named inputs X1..Xn. where the challenge is to discover the possible values of these named inputs that return most of the content of the database. This approach is suitable for CSW HTTP GET requests. However, the constraints are encoded in a single named input as a CQL string (see Nebert et al. 2007), or an XML Filter (Vretanos, 2004). This characteristic is incompatible with the query model of the Deep Web crawlers. Researchers working for search engines, such as Google (see Madhavan et al. 2008), discourage the alternative operational model that is based on the development of ad-oc connectors as non-sustainable in production environments.” [10]

Consequently, geospatial metadata is not as 'open' as it potentially could be because of the formalised constraints imposed by the (pre-)existing geospatial metadata query standards e.g. the GoGeo CSW metadata repository effectively resides behind an opaque interface[11] from the point of view of other (non-geospatial) communities. The CSW interface does not define a simple Web API to query and retrieve metadata. Some communities that potentially could use CSW are accustomed to simple APIs and common formats for purposes of 'mash-up'. For example, many geo-mashups and related data services [10] use Web APIs to access and share data built following the REST architectural style [11]. These APIs are characterized by the identification of resources by opaque URIs, semantic descriptions of resources, stateless and cacheable communication, and a uniform interface based on the verbs of the HTTP protocol which sits in opposition to the style adopted by CSW.

3 Approach

We addressed these perceived shortcomings by producing Linked Data from extant ISO19115/19139 records. In doing so we evaluated a number of alternative strategies for producing the RDF.
Metadata 'crosswalking'. There are several geographic metadata crosswalks to the Dublin Core vocabulary which may be viewed as the lowest common denominator metadata baseline. We adopted the use of a Dublin Core crosswalk to implement uniform mappings from geographic metadata schemas to the RDF data model. This approach consists of three steps: I) Apply a metadata crosswalk from the original metadata schema (ISO) to the Dublin Core vocabulary.; ii) add additional metadata such as provenance of the record, original information model or crosswalk identification; iii) Apply the profile for expressing as RDF the metadata terms.
An alternative approach was to publish direct from the relational data store underpinning the metadata resource , direct to RDF. A further extension to this approach was to produce the RDF direct from the Unified Modelling Language (UML) representations of the underlying schemas using a visual modeling approach.

4 Results

We worked with two types of metadata catalogues. 1) Catalogue Services for the Web,(CSW) services, important particularly because they are required for the implementation of the EU INSIPRE directive. We harvested, amongst others, the Scottish Location Information Discovery Service CSW for this purpose. 2) Customised catalogues that themselves aggregate data from various sources as exemplified by GoGeo and are typically implemented on top of a relational database schema.

These two separate catalogue implementations, lend themselves conveniently to the two alternative strategies for RDF production. In the case of CSW the crosswalk approach is the obvious solution whilst for database based schemas, schema mapping and UML derivation approaches seemed more appropriate. For the CSW “metadata crosswalk” approach, we applied XSLT transforms. For the purposes of establishing a Linked Data production flow-line, the metadata are harvested and stored in an intermediate database (triplestore) and it is against this intermediate database that queries were performed and representations (RDF and others) were published.

5 A Note on URI Governance

Implicit in any approach that mints URIs are certain assumptions. Firstly, that a URI can be clearly and unambiguously defined and assigned to identify particular resources. Secondly that URIs are stable and persist. Axiomatically, the essence of a persistent URI is it's immutability. The aforementioned Cabinet Office paper, “Designing URI Sets for the UK Public Sector”, endeavoured to address this issue but it's adoption and adaptation to location (geospatial) information in “Designing URI Sets for Location”[12] highlighted an issue of political sensitivity –

[7] “Designing URI Sets for the UK Public Sector” downloadable at: http://www.cabinetoffice.gov.uk/sites/default/files/resources/designing-URI-sets-uk-public-sector.pdf

[8] See e.g. http://location.defra.gov.uk/wp-content/uploads/2011/03/INSPIRE-UK-Location-and-Linked-Data.pdf

[9] http://dublincore.org/documents/dcmi-terms/

[10] http://www.openarchives.org/

[11] GoGeo also supports a Z39.50 target interface but this is not heavily used.

[12] http://location.defra.gov.uk/wp-content/uploads/2010/04/Designing_URI_Sets_for_Location-

specifically the fact that in its original presentation the guidance neglected to allow for sub-national URI schemes. At time of writing, the guidance is being recast to allow for Scottish, Welsh and Northern Irish URI naming schemes to be adopted if required by the devolved administrations[13].

6 A Note on Target Representation[14]

As far as possible we chose well known and widely used vocabularies (a set of agreed terms). The core vocabulary is, Dublin Core Terms[15], hereafter simply referred to as DC. Unfortunately, DC does not contain a notion of a collection except indirectly as an underspecified range for dc:isPartOf and does not contain any particular way to refer to a representation of an abstract dataset such as a downloadable resource. It also contains no particular mechanism either to talk about people and organisations who might be authors or maintainers of datasets or to represent geographical or temporal extents. For all else, bar the question of geographical extent there are, fortunately, solutions that are more or less well established.

For describing data catalogues there is a specific vocabulary designed to augment DC called the Data Catalog Vocabulary or DCat[16]. Of particular interest are the concepts of dcat:Catalog, dcat:CatalogRecord and dcat:Distribution. Also used is dcat:Dataset which is simply an alias for dc:Dataset. Further predicates are used for expressing other metadata such as keywords and spatio-temporal granularity. For referring to people, either natural or corporate, common practice is to use the Friend-of-a-Friend or FOAF vocabulary[17]. Where the precise nature of the entity in question is unclear foaf:Agent[18] was used. Where it is clear that a natural person is being referred to, the more specific foaf:Person was used and if it was an organisation we used foaf:Organisation. For more involved descriptions of people and organisations and their relationship, the Organisation Ontology may be used.

We make a distinction between the *description of a catalogue record* and the *description of the dataset itself.* This is in fact quite important. Frequently, the metadata record may be changed or updated in isolation, even when no change to the dataset itself has been made. The dataset may be maintained by one person or organisation and this person may have no influence whatsoever over the metadata catalogue. Separating out the concepts is important in order to be able to express this difference. This separation is well known and is already expressed through existing geospatial metadata fields. It is also worth noting that where a dataset has no natural dereferenceable URI, it is given a non-resolvable one in the urn:uuid namespace (where it has an identifier that is a conformant UUID). This non-resolvable URI is essentially a constant that is used to refer to the dataset in third-party documents. As in the current project we were concerned with publishing catalogue records and not dataset descriptions as such, it was more expedient (elaborated on in the section on the use of named graphs in the metadata crosswalk section below), to adopt this approach - although it is not a necessary feature of the representation.

7 Geographical Metadata

Geographical metadata, by which we typically mean an expression of the coverage (spatial) area of a particular dataset- most commonly this will be a (minimal) bounding box but it may be, in principle, a shape of arbitrary complexity. Whilst there is well established practice for representing point data in RDF, there does not appear to be any such consensus when it comes to representing even a shape as simple as rectangle[19]. What is reasonably certain is that the dc:spatial predicate should be used to link from the dataset to its spatial extent. However the range of this predicate is simply a dc:Location and is not further specified other than such an entity must describe “a spatial region or named place”.

One approach, taken by NeoGeo[20] is to attempt to create a completely granular “native” RDF description of the geometry. This is not incorrect but was regarded as inappropriate for several reasons. Firstly, there is no support in any current software for presenting such data to the user as a map rendering. Secondly, correspondence with the authors of NeoGeo suggests that they are primarily interested in publishing such geometries as first class entities in themselves whereas we assume that the geometries in our application are unlikely to be of primary interest outside of the context of the datasets which they annotate. Lastly, the complexity involved in creating indexes for data expressed in this way can be considerable. The Ordnance Survey on the other hand has opted to encode *any* geographical component on data which they publish simply by encoding fragments of Geography Markup Language (GML) as literals [12]. This works well in that it is easy to index and many more tools exist that understand GML natively – at least relative to the “native” RDF approach. A third approach is that presented in the GeoSPARQL W3C member submission[21] which, though the specification document conflates vocabulary with query language extensions, allows for both approaches.

Ver0.5.pdf

[13] Scottish Government are adopting a http://{resource}.data.scotland.gov.uk pattern where {resource} maps to the United Nations Classification of Functions of Government (COFOG) headings, using a cofog01, cofog02 approach rather than the descriptive label e.g. "general public services", "defence", etc. To improve the overall quality of this approach work is being undertaken with the UN and Rensellaer (http://rpi.edu/) to produce a SKOS RDF version of the COFOG.

[14] Throughout, in the examples given below, Turtle notation is used (http://www.w3.org/TeamSubmission/turtle/). Except where confusion and ambiguity might arise, prefix declarations are generally omitted and well known or standard prefixes are used.

[15] We have studiously avoided the use of the legacy Dublin Core Elements namespace. For this reason where the prefix dc: appears it is taken to refer to the Dublin Core Terms namespace.

[16] http://dvcs.w3.org/hg/gld/raw-file/default/dcat/index.html

[17] http://xmlns.com/foaf/spec/

[18] a subclass of dc:Agent

[19] Where by “rectangle” is meant the shape described by geodesics in standard coordinate system like WGS84 implied by the coordinates of two diagonally opposite points aka MBR or minimum bounding rectangle

[20] http://geovocab.org/doc/neogeo.html

[21] See http://geosparql.org/

We chose to construct literals using the Well Known Text (WKT)[22] and annotate them provisionally with terms from the GeoSPARQL namespace in the hope that the useful parts of this vocabulary will be incorporated at a later date into a more widely used standard i.e. as the standards are still undergoing comment and revision cycles, it is too early to say which approach represents the 'best' for future proofing purposes and consequently we have applied a 'best guess' logic. Considering the example in Figure 1, though we have used the WKT as the lowest common denominator representation of geospatial data, there is room alongside for other equivalent representations in e.g. GML as the Ordnance Survey does - or indeed, in expressly materialised granular RDF as with NeoGeo. Yet it also retains the properties of being easy to display to users using common tools and easy to index when importing into an RDF database.

```
procedure Harvest(source, start)
    for xml in source modified since start do
            rdf ← Transform(xml)
            Store(rdf)
    end for
end procedure
```

Figure 1. Harvesting algorithm

Retrieved documents are transformed using XSLT and are then stored in an RDF database[23]. For this project a specialised CSW client was written and this implements the query and actual fetching of ISO19139 documents from a server such as Geonetwork[24] (which is used in national and thematic based catalogues in the UK). The storage step is also important in that it must extract from the intermediary RDF data a suitable graph name, a URI that will be used to both identify a catalogue record and to group the statements that belong in this record. Catalogue records are typically "several levels deep" thus it is not sufficient to just, say, consider all statements with a given subject in order to obtain a complete catalogue record. In order to save on the expense of complex queries, the data relating to a particular record is therefore grouped together into a named graph during harvesting. This also simplifies updating records that have changed since it is merely necessary to replace the entire graph rather than do complex queries to determine exactly which statements need to be retracted before any new ones are added. The bulk of the logic, however, is done by the transformation step. In this case it uses an XSLT stylesheet[25] to transform ISO19139 XML data into RDF/XML which may then be stored. The structure of this stylesheet is straightforward, however it is perhaps appropriate to give some account of some of the specific mappings that we made. Salient mappings that we adopted are reproduced in Table 1.

It should be noted however, that this table is greatly simplified for in many cases it makes sense to produce more than one RDF statement, particularly when one considers type assertions and necessary bits of indirection, for what appears as perhaps a single element in the ISO19139 (e.g. temporal extent where the expression of a time interval in RDF can be quite verbose), and vice-versa where it takes many elements (at least eight) in the ISO19139 to express a geographical bounding box but because we opted for the simpler representation, the RDF requires only three. We experimented not only with CSW services run and maintained by ourselves but also those run by others, finding some anomalies in some elements e.g. to distinguish between various types of dates, such as publication or modification date, it appears that the controlled vocabulary and localisation support in many catalogues is incomplete. This required a number of special cases in the transform, e.g. "publication" vs. "publicatie" as found in Dutch catalogues for example as appears in the gmd:dateType/gmd:CIDateTypeCode/codeListValue elements. This complicated the transformation requiring manual intervention where automated harvesting and RDF production proved problematic. Nevertheless, this approach worked surprisingly well and provided us with a generic pipeline for harvesting geospatial metadata from any CSW and producing Linked Data outputs, albeit by making some default choices on the vocabularies used and representation simplification.

8 Database and UML to RDF Approaches

Database Mapping Approach

For our second approach we used the popular D2R software[26]. D2R is a toolset for publishing relational databases on the Semantic Web. It enables RDF and HTML browsers to navigate the content of the database, and allows applications to query the database using a formal query language - SPARQL *(SPARQL Protocol and RDF Query Language)*[27]. In

Table 1: Crosswalk mapping of ISO1939 .

ISO19139	RDF
gmd:MD_Metadata	dcat:CatalogRecord
gmd:fileIdentifier	dc:identifier
	also used in construction of URI
gmd:contact/gmd: CI_ResponsibleParty	foaf:Agent
gmd:identificationInfo/gmd: MD_DataIdentification	dcat:Dataset
gmd:identificationInfo/srv: SV_ServiceIdentification	dcat:Distribution
gmd:distributionInfo/gmd: MD_Distribution	dcat:Distribution
gmd:CI_Citation/gmd:title	dc:title
gmd:CI_Citation/gmd:abstract	dc:abstract
gmd:MD_Keywords	dcat:keyword
gmd:MD_LegalConstraints	dc:accessRights
	dc:rights
gmd:MD_Constraints	dc:rights
gmd:MD_Resolution	dcat:spatialGranularity
	dcat:temporalGranularity
gmd:EX_GeographicBoundingBox	dc:spatial
	geo:Geometry
gmd:EX_TemporalExtent	dc:temporal

[22] See http://www.geoapi.org/3.0/javadoc/org/opengis/referencing/doc-files/WKT.html

[23] In our case using 4Store (http://4store.org/)

[24] http://geonetwork-opensource.org/

[25] The current version of this stylesheet is available at https://bitbucket.org/ww/gold/src/tip/static/csw/md_metadata.xsl

[26] http://sourceforge.net/projects/d2rq-map/

[27] http://www.w3.org/TR/rdf-sparql-query/

time:Interval

this approach, queries are expressed in the SPARQL query language (whether expressly composed or implied by the act of fetching or dereferencing a resource identifier) and are transformed to more traditional SQL queries that are then executed against the underlying database, the results being returned according to the standard practice for SPARQL query results. Mappings between the database schema and the RDF can become quite onerous to hand-craft and we developed an approach to making this less tedious by extending the capabilities of a visual UML tool[28].

Transforming the catalogue to Linked Data on the fly by translating requests into SQL queries against the relational database in which it is stored and then transforming the results has some advantages. Though it is slower than querying native RDF storage directly, chiefly because SPARQL queries may entail expensive self-joins when translated into SQL, there is no need to provision a large RDF database server to host what is essentially a cache of the original data. There is also no need to coordinate updates or periodically harvest the source. It may also be, with a well-normalised relational schema and judiciously chosen set of indexes, that the relational database has very much smaller resource requirements than an equivalent RDF store – this is particularly relevant when there is a significant volume of data involved. Realistically though, geospatial metadata repositories are seldom of sufficient size to make this a *prima facie* concern. Data currency is more of a consideration and in instances where the underlying geospatial metadata store is being regularly changed, it may be more appropriate to use a D2R type approach then try to re-synchronise and lock-step harvested resources. Data custodians need to balance the trade-off between periodic harvesting and change frequency in order to ensure overall data currency. This is of course a generic issue and not limited solely to the type of data publication work explored here.

The D2R server is a single Java program that takes an input file describing the coordinates of the database to be translated and the translation rules. The process for configuring it is relatively simple: 1. Generate a crude configuration file using the generate mapping program; 2. Refine and customise the mapping (the hard part!) ; 3. Run the d2r-server .

The result is an HTTP service that supports simple HTML and RDF representations of the resources in the database and a SPARQL endpoint through which arbitrary queries may be made. For production use a reverse proxy or embedding into a Java servlet system may be considered desirable.

We encountered two main difficulties with this approach. Firstly, whereas working from a widely published international standard as the source data format (say ISO), we can be reasonably confident that it will not change in the short term and therefore any mapping from that to Linked Data also will not need to change, the same is not true for information in a relational database where internal business rules and ad hoc schemas often prevail. The main use of relational databases is as internal data storage and as such there are frequently no external constraints on the schema changing. If the schema changes for whatever reason, the mapping for the D2R server must, of necessity, be revisited. Whilst this did not happen during the lifetime of this project, it can be reasonably expected that it will happen from time to time in the future as the services are extended or modified. The other difficulty is related to the first. A further property of stable, agreed international standards as data sources, is that they tend to be relatively well documented. On the other hand, it is much less common to have the same level of documentation for a relational database schema that is intended primarily for internal use. Consequently, the process of refining the mapping configuration for the D2R server is not trivial if the schema is at all complicated or subject to frequent changes. The relative merits of which approach to use – crosswalking or derivation direct from a database, are ultimately bound to issues of available infrastructure, skill sets, data currency etc. Table 2 provides a summary of the two approaches.

Table 2: Side-by-side comparison of Crosswalk and D2R based approaches .

Approach used	Crosswalk	D2R
Data currency	Cached	Live
Computing resources	RAM-bound for RDF DB	CPU and network bound
Moving parts / Administrative complexity	Harvesting / update machinery RDF storage HTTP/REST service	Just D2R Software
Configuration complexity	Simple	Complex
Skillsets for customisation	RDF (Turtle, RDF/XML) XSLT (for customising the transformation) HTML+JS+CSS (for customising human- readable output) Go language (for customising the behaviour of this particular implementation)	RDF (Turtle, RDF/XML) Java (customisation of behaviour, deployment, appearance) HTML+JS+CSS (customisation of behaviour)
Look and Feel		Somewhat more complex
SPARQL queries		Slower

9 Production from UML – a UML Profile for D2R scripting

The task of republishing data deposited in relational database as Linked Data principally involves two main technical challenges. Firstly, converting the tables, columns and relationships in the relational model to appropriate RDF constructs (that is, RDF classes and properties). Secondly, mapping columns and relationships to appropriate existing Linked Data vocabularies. This is the approach described above and is actually a fairly common problem which has lead to the evolution of tools such as D2R. These tools have emerged to assist dataset providers in mapping relational models to RDF and tend to treat the relational database as a virtual RDF graph enabling access through APIs such as Sesame[29] and Jena[30]. While the D2R platform greatly facilitated semantic access to data in our geospatial relational database, the creation of the necessary D2R mapping scripts is

[28] We used Enterprise Architect.

[29] http://www.w3.org/2001/sw/wiki/Sesame

tedious and error prone, as no visual authoring tools are currently available.

We extended our UML tool to support simple D2R constructs, thereby automatically generating D2r scripts from within a visual editing environment. Our approach was to define an XML profile with D2R constructs that could augment existing UML stereotypes used in standard database modelling (<table>, <column> etc). The data modeller first loads the relational model into the UML tool. Then, the data modeller imports the 'D2R profile' (our GOLD extensions), to expand their toolbox. These imported D2R tools provide access to constructs such as “ClassMap” and “PropertyBridge” that can then be dragged onto the tables and columns in the UML model to define the mappings. Once the modeller has completed all the mappings they can export the UML model to XMI[31]. The output XMI contains all the information that the data modeller specified in the UML modelling tool, including the stereotypes and TaggedValues from the D2R profile. An XSLT stylesheet can the be applied to the XMI file to generate the D2R mapping in RDF format. In the case of Enterprise Architect it was possible to specify the XSLT stylesheet as part of the XMI export so the export and xsl transform can be combined into one step.

For Linked Data production direct from a relational data store, particularly ones where table structure is more than very simple, this visual editing support for mapping production is both more intuitive from the data modellers perspective and significantly less error prone than conventional 'hand-crafting' approaches. It is also more adaptable and robust to underlying schema changes and more flexible where multiple target output schemas are required. Our D2R profile extensions are available on request.

10 Conclusion

We have taken a Linked Data approach and published geospatial metadata as RDF and explored alternate options for RDF generation – from crosswalking through well known vocabularies such as Dublin Core, to RDF generation direct from a relational database. In either case, there is an expectation that the outputs will be consistent yet the vagaries of geospatial data (quality aspects) mean that establishing which approach is more flexible or robust in any particular application instance, is necessarily subject to trial and error. We have established a basic workflow and infrastructure which support CSW harvesting (from any CSW) and automated Linked Data publication of that geospatial metadata by adopting well known and frequently used vocabularies e.g. FOAF, DCat. An open question remains as to whether or not Linked Data is the panacea to resource discovery and reuse that its proponents assert. A significant issue still to be overcome is to establish core, common vocabularies – particularly in respect to alternate and competing approaches to the representation of geometry information.

References

[1] Nebert, D., Whiteside, A. and Vretanos, P. A. (2007). Open GIS Catalogue Services Specification. OpenGIS Publicy Available Standard, Open GIS Consortium Inc.

[2] INSPIRE Network Services Drafting Team (2008)

[3] Lopez-Pellicer F. J., Florczyk A. J., Nogueras-Iso J, Muro-Medrano P. R. and Zarazaga-Soria F. J. (2010) Exposing CSW catalogues as Linked Data, Lecture Notes in Geoinformation and Cartography (LNG&C). Geospatial Thinking. 2010, vol. , p. 183-200. ISSN 1863-2246

[4] Bergman, Michael K. (2001) White Paper: The Deep Web: Surfacing Hidden Value, The Journal of Electronic Publishing (JEP), Volume 7, Issue 1, August, 2001. Available from: http://hdl.handle.net/2027/spo.3336451.0007.104

[5] Bizer, C.; Heath, T.; Idehen, K. and Berners-Lee, T. (2008) Linked data on the web (LDOW2008) WWW '08: Proceeding of the 17th international conference on World Wide Web, ACM, 1265-1266

[6] Miller, E. and Manola, F. (2004) RDF Primer. W3C, Available from http://www.w3.org/TR/2004/REC-rdf-primer-20040210/

[7] Nilsson, M.; Powell, A.; Johnston, P. and Naeve, A. (2008) Expressing Dublin Core metadata using the Resource Description Framework (RDF) [online] Dublin Core Metadata Initiative, DCM

[8] Seaborne, A. & Prud'hommeaux, E. (2008) SPARQL Query Language for RDF W3C. W3CRecommendation. Available from:http://www.w3.org/TR/2008/REC-rdf-sparql-query-20080115/

[9] Nogueras-Iso, J., Zarazaga-Soria, F.J., Muro-Medrano, P.R. (2005) Geographic Information Metadata for Spatial Data Infrastructures: Resources, Interoperability and Information Retrieval. Springer-Verlag New York, Inc., Secaucus, NJ, USA

[10] Turner, A. (2006) Introduction to neogeography O'Reilly Media, Inc.

[11] Fielding, R. T. (2000) REST: Architectural Styles and the Design of Network based Software Architectures University of California, Irvine

[12] Goodwin, J.; Dolbear, C. and Hart, G. (2009) Geographical Linked Data: The Administrative Geography of Great Britain on the Semantic Web Transactions in GIS, doi: 10.1111/j.1467-9671.2008.01133.x

[30] http://jena.sourceforge.net/

[31] The XML Metadata Interchange (XMI) is an Object Management Group (OMG) standard for exchanging metadata information