Linked Data Authority Records for Irish Place Names

Catherine Ryan[1,2], Rebecca Grant[1], Eoghan Ó Carragáin[2], Sandra Collins[1], Stefan Decker[3], and Nuno Lopes[1,3,4]

[1] Digital Repository of Ireland
Email: {c.ryan,r.grant,s.collins}@ria.ie
[2] National Library of Ireland
Email: eocarragain@nli.ie
[3] INSIGHT @ NUI Galway, Galway, Ireland
Email: stefan.decker@deri.org
[4] IBM Research, Smarter Cities Technology Centre, Dublin, Ireland
Email: nuno.lopes@ie.ibm.com

Abstract. Linked Data technologies are increasingly being implemented to enhance cataloguing workflows in libraries, archives and museums. We review current best practice in library cataloguing, how Linked Data is used to link collections and provide consistency in indexing, and briefly describe the relationship between Linked Data, library data models and descriptive standards. As an example we look at the Logainm.ie dataset, an online database holding the authoritative hierarchical list of Irish and English language place names in Ireland. This paper describes the process of creating the new Linked Logainm dataset, including the transformation of the data from XML to RDF and the generation of links to external geographic datasets like DBpedia and the Faceted Application of Subject Terminology. This dataset was then used to enhance the National Library of Ireland's metadata MARCXML metadata records for its Longfield Maps Collection. We also describe the potential benefits of Linked Data for libraries, focusing on the use of the Linked Logainm dataset and its future potential for Irish heritage institutions.

Keywords: Linked Data, place names, authority data, metadata, multilingual data, libraries

1 Introduction

One of the key roles of a librarian is organising knowledge and the description of resources to allow for their efficient retrieval, sharing and preservation. Over time, librarians have adapted well to changing technologies, increasing amounts of information and the changing needs of users. Also libraries attract diverse user communities, who demand more information sources and engage with information in different ways - through social networks, tagging, sharing or the development of their own applications. These new resources and communities

have led to the development of a number of standards to catalogue the information necessary for the retrieval, sharing and preservation of resources. Such standardised controlled vocabularies, authorities and classification systems are used in libraries, archives and museums to create consistent access points for their catalogue records, to organise materials and to improve search results.

Controlled vocabularies provide structured terminologies which represent the relationships between given concepts, and suggest preferred terms. Such vocabularies can be used by cataloguers to choose appropriate subject terms for concepts such as an event or location which has undergone multiple name changes over time, or an author who uses a pseudonym. Controlled vocabularies can also be known as authority records, an index of controlled terms compiled and maintained by an institution.

Authority records or authority files not only capture the preferred heading for a particular term, but can also include additional contextual information e.g., birth dates, death dates, titles, etc. Some examples of controlled vocabularies, authorities and classification systems include the Getty Vocabularies,[5] Library of Congress Authorities,[6] and Iconclass.[7]

One such database providing a controlled multilingual vocabulary of Irish place names with authoritative contextual information is Logainm.ie.[8] The Logainm dataset includes a list of Irish place names, their validated translations between the English and Irish languages, and administrative hierarchy information (stating in which other place names they are included). Linked Logainm has created a Linked Data version of Irish place name data held by Logainm.ie, publishing Irish place names data as Linked Data and providing information and support to encourage the uptake of the dataset in Irish libraries.

In this paper we describe current cataloguing practices in libraries in Section 3, give a brief overview of Linked Data principles and technologies in Section 2, and present how libraries are moving towards Linked Data and beginning to incorporate its principles and technologies (particularly in the area of authority data), in Section 4. Section 5 describes the process for transforming Logainm's dataset into the Resource Description Framework (RDF) [24] and the work of linking to existing library subject heading schema available as Linked Data, in this case DBpedia,[9] LinkedGeoData,[10] Geonames,[11] and the Faceted Application of Subject Terminology (FAST),[12] a dataset released by the Online Computer Library Centre (OCLC). Finally, Section 6 describes the use of the newly generated dataset for enhancing a collection of maps, in our example the National Library of Ireland's Longfield map collection, and Section 7 presents the conclusions and future work.

[5] http://www.getty.edu/research/tools/vocabularies/
[6] http://authorities.loc.gov/
[7] http://www.iconclass.nl/
[8] http://www.logainm.ie/
[9] http://dbpedia.org/
[10] http://linkedgeodata.org/
[11] http://www.geonames.org/
[12] http://experimental.worldcat.org/fast/

2 Linked Data Preliminaries

Linked Data [5] refers to data published on the Web following a set of principles designed to promote linking between entities on the Web. This allows these entities to be connected and enriched, and facilitates linking between related resources.

The Linked Data realm or Semantic Web operates in a similar way to the traditional Web, where text, image, video and other documents from all over the world are published and connected to each other, with clickable Uniform Resource Locators (URLs) pointing to their locations. By making these documents available and using hyperlinks to make them discoverable, they can be connected into a global information space made up of other interconnected documents, or the "Web of Documents".

When implementing Linked Data technologies it is the data, rather than documents, that can be published to the Web in formats that are semantically understood by computers. This data can be anything –a person, a place, a subject, a book– and can come from sources including the geographic, governmental, library, medical, scientific, statistical, user-generated, and cross-sector domains. By creating links between diverse datasets, Linked Data breaks down the barriers between them, allowing for greater discovery, sharing and reuse of the data. [32]

To enable this linking, each entity (for example a place name or personal name) must be given a unique identifier, generally in the form of a Uniform Resource Identifier (URI). Having determined these URI identifiers, Linked Data reuses other data models such as the Resource Description Framework (RDF) [24] to represent the data about each entity and specify the links, and their type, between two URIs. Data can then be linked together and connected across a global data space to create a "Web of Data".

Best practices have been developed to facilitate the publication and connection of structured data in the Web of Data. These are known as the Linked Data Principles [2]:

- Use URIs as names for things;
- Use HTTP URIs so that people can look up those names;
- When someone looks up a URI, provide useful information using the Linked Data standards (RDF, SPARQL);
- Include links to other URIs so that people can discover more things.

Semantic Web technologies such as URIs, data models, knowledge representation languages, vocabularies and ontologies are currently being used in sectors including healthcare, government, energy, IT and eTourism to improve search, data integration, content management, and discovery.[13] These technologies are providing similar benefits to cultural institutions such as Europeana [14] and the BBC [27]. The application of Linked Data technologies to diverse datasets has the potential to open up and greatly enhance metadata and digitised collections generated by libraries, archives and museums globally.

[13] `http://www.w3.org/2001/sw/sweo/public/UseCases/`

Enabling Linked Data to share structured data across the Web and across different organisations and individuals, concepts and relationships between resources must be explicit and defined in a format that is understood by computers. Without such an agreement on structures and vocabularies, words used in different datasets may be ambiguous and it may not be possible to establish correct relationships among datasets. Libraries have a long tradition of using vocabularies and ontologies to organise knowledge and Linked Data vocabularies offer cultural institutions the possibility of expanding research and discovery by linking together authority URIs, thereby sharing their data with and integrating data from other cultural institutions, publishers, business, social and other datasets. Ontologies created for the Semantic Web, e.g., Friend of a Friend (FOAF),[14] use URIs, are based on RDF and are expressed primarily in RDF Schema (RDFS) [8], Web Ontology Language (OWL) [19] and the Simple Knowledge Organisation System (SKOS) [21] which are outlined briefly below.

2.1 URIs

Uniform Resource Identifiers (URIs) are used to provide unique names for resources on the Web and sit at the heart of Linked Data and the Semantic Web. One such example is `http://dbpedia.org/resource/James_Joyce`, the URI for James Joyce in the DBpedia domain, or namespace. URIs are used to identify resources and express relationships between them. Such resources can be anything existing on the Web as well as "real world objects" such as people and places. To ensure stability in the Semantic Web it is important that URIs remain unchanged - that they are persistent. Using a URI instead of a string of text to identify an entity or resource removes any ambiguity between people or places that have the same name. This "things not strings" approach [30] to naming entities also removes the possibility of breaking the link between authorities due to a misspelling or misplaced or absent punctuation, e.g., currently many library management systems consider "James Joyce" and "James Joyce." (with a full stop) to be different entities.

2.2 Resource Description Framework

RDF [24] is a graph data model developed by the W3C for representing and exchanging information on the Web. RDF makes statements, called "triples", and they take the form *subject*, *predicate*, *object*, where subject is the entity or resource and object is another resource or value. The predicate is the relationship between them and is defined using predefined vocabularies. Several statements can be created about the same subject, objects can also act as subjects and link to other objects and, by combining these statements, a network of triples is created, a vast graph spanning out across datasets and the Web. RDF requires that URIs are used to name things and relationships and, by doing so, this data can be understood by computers, is persistent, is unambiguous and can be shared across the Web. RDF has also been serialised in XML and other formats.

[14] `http://www.foaf-project.org/`

2.3 Knowledge Representation Languages

While RDF is a basic data model which makes statements about resources, it does not define the relationships in these statements and the terms used to express them. To do this we use what are known as "knowledge representation languages". These are predefined vocabularies which have been developed from concepts arising out of RDF and they are used as building blocks for the creation of more complex and specific ontologies on the Web. These languages include RDFS and OWL. Developed for the Semantic Web, they can, depending on the capabilities of the software in question, infer relationships between things and resources both within and across datasets automatically, enhancing discovery in a way not previously possible.

RDF Schema RDFS [8] is the most basic knowledge representation language arising from RDF. While RDFS does not provide an extensive vocabulary of its own it does provide a basic core type system of classes and properties and indicates how they should (not necessarily must) be used together. These can then be used to describe the classes and properties of other resources and to express richer, more complex element sets or vocabularies on the Web. One such example of this is OWL.

Web Ontology Language OWL [19], itself an ontology expressed in RDFS, is more extensive and expressive than RDFS but essentially performs the same function, used primarily as a language for creating and expressing more complex ontologies on the Web. It is used to describe and define terms within a particular domain of interest or subject and to describe and define the relationships between them. It can be used to represent almost any concept or domain.

Many ontologies that have been developed use a combination of RDFS and OWL as their base languages. Examples of this include FOAF, the social ontology for describing people, their activities and the relationships between them which is now widespread on the Web.

Another example of an ontology built using RDFS and OWL is SKOS [21], another standard for the organisation of knowledge, specifically vocabularies, on the Web.

Simple Knowledge Organisation System SKOS [21] is a data model and vocabulary expressed in RDFS, OWL and Dublin Core Terms. It is designed to make controlled vocabularies such as thesauri, classification schemes, taxonomies and subject heading systems available on the Web with the ability to express their hierarchical, associative and other relationships.

2.4 The SPARQL Query Language

Information resources described using RDF, and expressed in RDFS, OWL, SKOS and others, are saved to database management systems (DBMS) known

as "triplestores". These database management systems are similar to relational databases but where relational databases store their information in tables, triplestores are specifically designed for the retrieval and storage of RDF triples. As relational databases use SQL to query tables in the database, so triplestores use SPARQL [18] to query and retrieve information stored in the RDF triple format. David Stuart offers an introduction to SPARQL for librarians in his book Facilitating Access to the Web of Data, in which he notes that SPARQL, while it may look complex, is not a programming language and can be learnt with a little practice [32].

3 Library Standards

This section describes current practices and standards in library cataloguing and recent changes that have laid the groundwork for Linked Data in libraries.

There are a number of standards used in libraries, archives and museums to create and encode authority records, including the MARC 21 Format for Authority Data;[15] the International Standard Archival Authority Record For Corporate Bodies, Persons and Families (ISAAR(CPF))[16] and the Metadata Authority Description Schema (MADS).[17]

MAchine-Readable Cataloguing (MARC)[18] is the data format commonly used in libraries to allow computers exchange, use and interpret bibliographic information. MARC records can be encoded using the MARCXML schema. MARC's current iteration, MARC21 is considered a barrier to opening up library data due to its complex nature and card-based, document-centric focus [11]. Addressing this, recent changes in the bibliographic landscape, and more particularly the development of Resource Description and Access (RDA)[19] and Functional Requirements for Bibliographic Records (FRBR),[20] have laid the groundwork for Linked Data in libraries.

3.1 Functional Requirements for Bibliographic Records

FRBR is an entity-relationship model that describes the bibliographic universe in relation to user needs and the variety of media now available. The ability to share library data and to link with other data on the Web requires that library data be "structured as individual statements that can interact in a meaningful way." As a model, FRBR describes the structure of bibliographic data, the different

[15] http://www.loc.gov/marc/authority/

[16] http://www.ica.org/10203/standards/isaar-cpf-international-standard-archival-authority-record-for-corporate-bodies-persons-and-families-2nd-edition.html

[17] http://www.loc.gov/standards/mads/

[18] http://www.loc.gov/marc/

[19] http://www.rda-jsc.org/rda.html

[20] http://www.ifla.org/publications/functional-requirements-for-bibliographic-records

elements and the relationships between them. The bibliographic relationships in FRBR have the potential to transform library catalogues from collections of records to a network of navigable relationships, linking data elements (FRBR entities) to each other within the catalogue and externally to datasets on the Web [10]. An in-depth description of FRBR can be found in [20] and the Library of Congress also provides a brief overview [33].

3.2 Resource Description and Access (RDA)

The development of RDA was based on the FRBR framework and its entities, attributes and relationships terminology. Intended to succeed the Anglo American Cataloguing Rules (AACR2)[21] as a content standard for resources in the evolving bibliographic environment, RDA is a set of guidelines designed for all types of resources found in the digital world. As a set of guidelines RDA implicitly states which elements are required in a bibliographic description but does not set these out in a machine-readable format [12]. RDA was developed by the Dublin Core Metadata Initiative / RDA Task Group (now the Bibliographic Metadata Task Group)[22] who used FRBR entities (Work, Expression, Person etc.) to define classes and RDA data elements (title, form, name, date of birth, etc.) to define properties. RDA value vocabularies[23] were also developed to define content such as format and carrier type.

4 Linked Data in Libraries

The current information environment is changing rapidly, with new formats and data sources now present in most libraries. While some bibliographic datasets and authority lists are now available as Linked Data, a number of issues inherent to current library data mean that it is not generally connected to or "of the Web". With the development of new skills, librarians can bring their collections into the Web of Data, engaging in the structuring and organisation of data and helping users to access, use and republish data themselves [32].

In order to implement Linked Data technologies, libraries need to begin focusing on data elements rather than the record, as it is the data elements and values which can be defined in ontologies and vocabularies, encoded and shared across the Semantic Web [1].

Bibliographic Data In terms of bibliographic data, emerging cataloguing practices are moving closer to integrating library data with the World Wide Web. The FRBR element set has been published online and represented in RDF along with the rest of the FRBR family of standards and they are now in the process of being consolidated [28]. RDA value vocabularies were also developed and defined

[21] `http://www.aacr2.org/`
[22] `http://wiki.dublincore.org/index.php/Bibliographic_Metadata_Task_Group`
[23] `http://rdvocab.info/`

using RDF and include RDA Book Format (12mo, 8vo, etc.) and RDA Carrier Type (computer disc, film reel, etc.) among others. The RDA element set and value vocabularies have all been published in RDF, resulting in a definition of library cataloguing data that can be shared at the data level, opening up many possibilities of interaction with the Semantic Web.

Other initiatives such as Bibliographic Framework Initiative (BIBFRAME)[24] have been set up to progress the linking of bibliographic data rather than bibliographic records. Developed using FRBR concepts and using RDA as a primary content type, BIBFRAME aims both to be a model to which different content models can be mapped and, using Linked Data standards, to transform the MARC format from a flat structure into an RDF-based triples structure of entities and the relationships between them [25]. The BIBFRAME model is currently in draft and subject to change. The Online Computer Library Centre (OCLC) has also been working with Linked Data and in 2012 began adding Linked Data to WorldCat.org using Schema.org, a lightweight and cross-domain markup vocabulary set up by Google, Yahoo, Microsoft and Yandex that specifies normalised markup for webpages, in a way that reduces ambiguity, and makes the integration of the data into search engines more efficient.[25]

Authority Data As the bibliographic landscape is reconceptualised, work has been ongoing on the conversion of existing authority data and value vocabularies used in libraries to Semantic Web formats. By assigning URIs to authority terms, these terms become suitable for use in the Semantic Web and cover a range of concepts including subjects, properties, and real world objects such as people and places to which subjects and properties can relate.

The benefits of linking authority data include improved search, the potential to enhance discovery systems and integrate data from multiple sources, disambiguation and the addition of extra value through the contextualisation of resources. By encoding classification systems, controlled vocabularies and authority records in a machine-readable Linked Data format, these resources are brought into the networked environment and can continue to improve search and discovery systems on the Web. Using both preferred and alternate labels to include alternative spellings and names for people, places and concepts allows the precision of results to be improved. Systems can be built which reflect the hierarchical nature of vocabularies and the relationships between the entities within them. Linking to external datasets allows data to be pulled into online catalogues automatically, adding valuable contextual information to digital heritage collections. By facilitating the reuse of data, new and more creative applications can also be developed. [34, 9, 17, 16]

[24] `http://bibframe.org/`
[25] `http://schema.org/`

4.1 Relevant Datasets

We now describe some of the most important datasets used in the Library domain.

Library Datasets Using URIs to uniquely identify each term and RDFs, OWL and SKOS for structure, OCLC have published a Linked Data version of the Dewey Decimal Classification System,[26] and the Faceted Application of Subject Terminology (FAST), a Linked Data version of the Library of Congress Subject Headings. OCLC have also been involved in the creation of the Virtual International Authority File (VIAF),[27] where national authority files from across the world have been matched and linked together.

Other authority work includes the publication of the Dublin Core terms,[28] the Library of Congress Name Authority File, Library of Congress Genre/Form terms, and the Library of Congress Children's Subject Headings.[29]

Geographic Datasets Geographic data forms a substantial portion of the Linked Data landscape. Some of the most relevant data providers and related approaches are LinkedGeoData, GeoLinkedData, Geonames, and DBpedia.

LinkedGeoData [31] consists of a mapping from OpenStreetMap (OSM) data to RDF. The OSM data model contains three types of place names: nodes, ways, and relations. It includes links to DBpedia and GeoNames that are created based on geographic location, the name, and type of place name. GeoLinkedData [7] contains information specific to the Spanish national scenario and mostly relates information about coastal areas with other Spanish national statistics. The UK Ordnance Survey is the mapping agency for Great Britain and exposes some of its data as Linked Data [15]. While originally exposing only simple indexes of place names (gazetteers), they have now defined custom ontologies to describe the relationships between place names such as topological relations (e.g., *borders*, *spatially contains*).

GeoNames is a worldwide geographic database, freely available and also exposes its data as Linked Data. DBpedia [6] publishes information extracted from Wikipedia as Linked Data. Although not specifically targeted at geographic data, it includes geospatial entities along with point representations for their locations.

5 Creating Linked Data Authority Records from the Logainm Placenames Database of Ireland

Acknowledging current practice in library cataloguing and the increasing implementation of Linked Data technologies in libraries, the Linked Logainm project

[26] `http://www.dewey.info/`
[27] `http://viaf.org/`
[28] `http://dublincore.org/documents/dcmi-terms/`
[29] `http://id.loc.gov/`

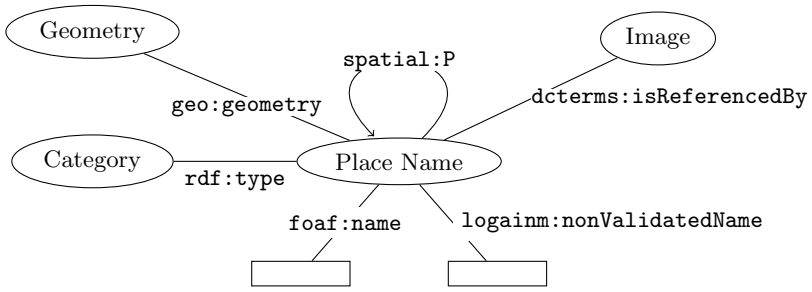

Fig. 1. Initial Schema for representing Logainm place names in RDF

has generated an authoritative linked dataset of Irish place names for use by the National Library of Ireland, and the broader heritage community both in Ireland and internationally. We will now describe the process undertaken to create the Linked Logainm dataset and its use by the NLI as an authority record.

Logainm.ie is an online database containing just over 100,000 Irish geographic names, including authoritative Irish language translations and historical variants. Furthermore approximately 50,000 place names include geographic coordinates. This dataset is generated by the Placenames Branch of the Department of Arts, Heritage and the Gaeltacht, and the database was created and is maintained in collaboration with Fiontar,[30] the Irish language department of Dublin City University. The Logainm dataset is intended as a resource for researchers including educators, students and genealogists. As a bilingual authoritative list of place names, it is also used as the basis for cataloguing and key wording collections from heritage institutions including museums, archives and libraries nationally and internationally. While Logainm's complete dataset has been made available by Fiontar on a request basis, the inclusion of Linked Data in their website allows immediate access to structured data which can be used by cataloguers, as well as computer scientists and application developers.

We next present a set of requirements for the dataset, based on the Linked Data principles and other requirements specific to Logainm:

Place name identifiers: Each place name is identified by a URI under the new sub-domain `http://data.logainm.ie/`. URI identifiers for place names follow the pattern `http://data.logainm.ie/place/{LogainmID}`, where `{LogainmID}` is the place name identifier from the Logainm dataset.

Names in Irish and English: Data representing the Irish and English names is retrievable from the respective URI. To represent the different languages, we use XML "language tags" associated with values of the same property (e.g., foaf:name). Another approach would be to use different properties for English and Irish names, as done by the UK Ordnance Survey: "hasOfficialName" and "hasOfficialWelshName" for Wales [15]. Furthermore, place names include the concept of a validated place name to enable the admin-

[30] `http://www.dcu.ie/fiontar/`

istrative process of translation, and also include alternate spellings for the place name in Irish.

Types for Place Names: Each place name has a Category, e.g., *Barony*, *Town*, *County*.[31] Each type is identified by a URI `http://data.logainm.ie/category/{categoryID}`, where {`categoryID`} is the category identifier from the Logainm dataset. Place names are contained in (possibly multiple) other place names: The Logainm dataset contains information regarding hierarchical inclusion of place names. This information represents the administrative structure of the place names.

Geographic Coordinates: Place names are generally represented by one coordinate according to the Irish Grid Reference format [26]. In some cases, place names can have multiple coordinates: e.g., for Rivers. For town lands and other types the coordinate approximates the geographic centre, while for rivers the two coordinates are used represent the source and mouth.

A depiction of the Linked Logainm schema is presented in Figure 1. The generated RDF follows the NeoGeo vocabulary [29], where each place typed as "Feature". The NeoGeo vocabulary also defines several spatial relations between entities, for our data we are relying on the `is_part_of` relation (`spatial:P`).

5.1 Creating Linked Logainm

The content of the Logainm database were provided as an XML dump that included all English and Irish place names along with their type, and if available, the geo-location of the place in Irish Grid Reference format. The approach of using an XML dump rather than other techniques, for example RDB2RDF [13], is due to Fiontar's plans to migrate the Logainm database from a relational database to an XML database.

Based on the team's previous expertise, the translation of the Logainm database dump (in XML) into RDF was performed using XSPARQL [4], by developing a query that transforms the input XML into the target RDF schema. As an alternative, other options to translate XML into RDF could be used, for instance custom XSLT queries or other languages similar in spirit to XSPARQL, e.g., XSLT+SPARQL [3]. Furthermore, if we had access to the relational database, this translation could be performed by any RDB2RDF tool.[32] A partial example of this query is presented in Figure 2. The presented example creates a subset of the RDF data from the input XML, that contains:

(i) the place name identifiers (generated in line 3 with the help of an auxiliary function);
(ii) the connection to the original page in Logainm's website (line 5); and
(iii) the (possibly multiple) types of the place name (lines 6–8).

[31] A description of the types in the Logainm dataset can be found at `http://www.logainm.ie/en/inf/help-categs/`.
[32] `http://www.w3.org/TR/rdb2rdf-implementations/`

```
for $place in places/place
let $id := fn:data($place/@id)
let $URI := local:createURI($id, "place")
construct { <{$URI}> a spatial:Feature ;
                    foaf:isPrimaryTopicOf <{fn:concat("http://
                        logainm.ie/",$id,".aspx")}> .
           { for $type in $place/type
             construct { <{$URI}> a <{ local:createURI(fn:data(
                 $type/@id), "category") }> }
           }.
           }
```

Fig. 2. XSPARQL transformation

The target RDF representation contains the information in the original Logainm dataset and also includes the geo-location of the place following the World Geodetic System (WGS) coordinates, the reference coordinate system used by the Global Positioning System (GPS). These coordinates were translated from the provided Irish Grid Reference format coordinates and used to aid the determination of links between the Logainm dataset and other Linked Data sources on the Web.

The Logainm RDF dataset has been deployed in the Logainm website and the SPARQL endpoint is available at `http://data.logainm.ie/sparql/`. The resulting RDF dataset contains approximately 1.3 million triples.

5.2 Linking the datasets

From the relevant sources of Geographic Linked Data, we elected to generate links to DBpedia, LinkedGeoData, and Geonames. We used the Silk Link Discovery Framework [22] to generate the links between the Logainm RDF dataset and the other target datasets. The Silk framework compares entities from the different datasets according to a pre-defined set of rules and assigns a normalised value (in the interval $[0, 1]$) to the similarity between entities. The entities with highest similarity value are considered to be equivalent.

Next we present the set of comparison rules that we devised for establishing the links between the datasets. The final rule was based on the following similarity values:

- place name (n);
- type of the place name (t);
- name of the parent place (p); and
- geographic coordinates (g) (if available).

For defining the different similarity values we used the functions and aggregation operations provided by the Silk framework. For example for comparing the place

Table 1. Mapping of types between the different datasets

Type	DBpedia	LinkedGeoData	GeoNames
townland	Populated Place	Locality	LCTY, PPLF
population centre	Populated Place	Town, Village, Suburb, Locality	PPLS, PPL, PPLL
town	Town	Town, Village	PPL, PPLS
mountain or mountain range	Mountain, Mountain Range	Mountain Pass, Peak	MT, MTS, PASS, PKU, PKSU
village	Village	Village, Hamlet	PPL
island or archipelago	Island	Island	ISL, ISLET
river	River	River	STM
monument	Monument	Monument	MNMT
city	City	City	PPL
valley	Valley	NaturalValley	VAL

names we used the provided string comparison function based on the jaro distance metric or for comparing the geographic coordinates we used the provided spatial comparison function wgs84. Both p and g allow us to distinguish between place names that have a similar name but are located in different parts of Ireland, e.g., Newcastle in Dublin, Newcastle in Cork, and Newcastle in Galway.

To calculate the aggregated comparison value, we used a weighted comparison, according to the following formula, applied over the previously determined correspondence of types t:

$$\frac{1}{2}n + \frac{3}{8}p + \frac{1}{8}g \ . \tag{1}$$

This formula places a higher weight on the string comparison values, both of the place name and the place name it is included in, rather than the geographic location. This meant that we were not overly penalised by any errors in the geographic coordinates of the place names and could still detect links for the place names that do not have any geographic information. This formula can still be iterated, for example a more fine-tuned approach, as presented in [31], can be investigated. This formula was applied over the Linked Logainm RDF dataset and the results were evaluated by domain experts in Fiontar. A summary of the produced results and their evaluation is presented in Section 5.3. Also noteworthy is that for determining the links to GeoNames, we omit p from the formula, since hierarchical information in GeoNames is not freely available. We also generated the GeoNames RDF from the provided data dumps (in tabular separated format) since RDF dumps are another premium feature.

We also take into account the type of the place name (t), as such this weighted comparison is only performed between entities of types that have been considered similar. Similar to the approach followed in [31], the small number of types allows for a manual process which, in our case, was performed by domain experts in Fointar. The results of this matching (for the most relevant types) is presented in

Table 1. The most problematic types to match were "townland" and "population centre", as these can be mapped to different types in the target datasets. We have taken the approach of matching against a more general type (as in the case of DBpedia) or matching against several types (as in LinkedGeoData and GeoNames).

Linking to the FAST dataset. Given our interest in enhancing library metadata, we investigated the possibility of linking to existing library subject heading schema available as Linked Data. The Online Computer Library Centre (OCLC) has released the Faceted Application of Subject Terminology (FAST) dataset as Linked Data. FAST is a subject heading vocabulary which is derived from the Library of Congress Subject Headings (LCSH), the most widely used subject vocabulary in the library domain. Although the full LCSH dataset has also been published as Linked Data, the simplified FAST syntax and, in particular, the presence of GeoNames references in the FAST data made it an easier target for matching with Linked Logainm. However, since FAST in turn contains links to the Library of Congress Linked Data Service,[33] Linked Logainm will also be related to LCSH. We looked at reusing the previously described rules and similar process to determine the links, however the FAST data is not structured in such a way that would make this possible. As an alternative, we have decided to leverage the links to GeoNames present in FAST and, combined with our own links to GeoNames, establish an initial set of links from Logainm to FAST. Using this approach we matched approximately one third of the Irish geographic entities present in FAST (500 out of 1,400). In order to obtain a complete matching from Logainm to FAST, the datasets were manually linked along with the link evaluation process (described in the next section). From this manual linking we determined approximately 1,000 links to FAST, and for the remaining entities no adequate match was found.

5.3 Link Evaluation

From the similarity value that the Silk framework assigns to each link we consider only those above 0.95, i.e., the result of Equation (1) is above 0.95.[34] The evaluation of the rules presented in the previous section generated a set of approximately 16,000 links to the different datasets. A breakdown of the number of links by types and to the different datasets is presented in Table 2. It is noteworthy that the vast majority of links was established for "townlands", which was also the type that was matched to an higher-level type in the target ontologies, and such an approach may introduce errors in the generated links since the matching is being performed against a larger number of entities (all the matched type and its subtypes).

[33] `http://id.loc.gov/`

[34] This specific value was decided by the project partners as an acceptable tradeoff between the number of generated links and the (possible) link errors introduced by the automatic process.

Table 2. Number of links between the different datasets

Logainm		# Links			
Type	# Entities	DBpedia	LinkedGeoData	GeoNames	Total
townland	61,104	747	4,970	7,024	12,741
population centre	2,226	505	1,151	970	2,626
town	849	560	688	605	1,853
mountain or mountain range	372	63	115	111	289
village	142	79	90	10	179
island or archipelago	1,087	20	26	120	166
river	930	12	4	82	98
monument	245	22	36	39	97
city	8	8	7	5	20
valley	111	1	6	9	16

In order to determine the precision of the links generated between Logainm and the other datasets we manually checked a subset of these links (further details on the evaluation for each dataset are presented below). The task was to examine the information provided by each pair of URIs (by accessing the URI with a web browser) and deciding if the suggested matching was correct or incorrect. Since Logainm is a manually curated database, our main focus was to ensure the correctness of the generated links, thus maintaining the dataset's reputation of trusted quality data. As such, we are aiming at a higher precision of the generated links rather than covering all the place names (higher recall). From this manual checking of the links we estimate that the precision of the generated set of links is 97%.

Below is a breakdown of the link evaluation per dataset and some of the problems in the matching:

DBpedia: For DBpedia we manually checked all the generated links, and determined a precision of 98%. Some common issues that were encountered in the generated matching were:

(i) Since Logainm contains more fine-grained information, it can contain different entries for "towns", "population centre", and "townland" with the same name. However DBpedia contains only an entry for "town" or "population centre". For example, Adrigole is a "population centre" and a "townland" in Logainm (with two distinct identifiers, `http://data.logainm.ie/place/1412693` and `http://data.logainm.ie/place/8649`), while in DBpedia Adrigole is only a "village" and both Logainm entities are matched to the same DBpedia entity: `http://dbpedia.org/resource/Adrigole`.

(ii) Another issue, although less common, is the discrepancy between types in Logainm and DBpedia. For example, Kentstown is a "townland" in Logainm (`http://data.logainm.ie/place/38671`), while in DBpedia classifies it as a "village" (`http://dbpedia.org/resource/Kentstown`).

One possible solution would be to define relations between the different definitions using SKOS (`skos:broader`, `skos:narrower`, etc) and include these mappings in our linking rules

LinkedGeodata: For LinkedGeoData we have checked a random set of 500 links from all the generated links (7.5% of all generated links). Within this subset the precision was of 96%. Also for LinkedGeoData a common source of errors were "townlands", often matching other types in LinkedGeoData. Also the geographic coordinates in between Logainm and LinkedGeoData are often above the defined offset in our rules (1km), especially for "townlands". This suggests that increasing the value of the offset in our rules may provide further links to LinkedGeoData but may also increase the number of incorrect links.

GeoNames: The links to GeoNames provided very accurate results, from the subset of 500 links (6% of the total generated links) we checked, the precision was 99.6%.

FAST: The links to the FAST dataset were manually generated, based on an initial set of approximately 500 links with GeoNames. In this initial set of links, 1.2% were found incorrect, possibly due to errors in the links established between Logainm and GeoNames. Overall it was not possible to determine links for approximately 12% of the entities in the FAST dataset to Logainm entities, frequently because no hierarchy nor geographic coordinates are provided and is thus impossible to distinguish between place names with the same name across Ireland. Also the FAST database frequently conflates "town", "townland", "parish", and other types. We followed the approach of matching to "town" in Logainm.

A workflow for editing incorrect links was put in place so that any incorrect links that may be discovered can be fixed.

6 Applying Linked Logainm to Library Metadata

Our example collection, the Longfield Maps,[35] are a collection of 1570 map surveys carried out in Ireland between 1770 and 1840. Derived from the maps themselves, the existing metadata records include subject headings for counties, baronies, and occasionally parishes. The emphasis on baronies and parishes in this metadata, as well as the presence of minor geographic features in the surveys, make this collection particularly suitable for linking to the geographic entities found uniquely in the Logainm dataset. The place names stored in the metadata about the maps are in English, preventing any searches for place names represented in the maps to be specified in Irish. We describe ways in which this problem may be overcome by relying on Linked Data from the Linked Logainm project. Our initial approach is to enhance the existing metadata by recording the corresponding identifier in the Linked Data version of Logainm.

[35] `http://catalogue.nli.ie/Collection/vtls000282687`

```
<marc:datafield tag="522" ind1=" " ind2=" ">
  <marc:subfield code="a">
    Barony of Coshma, County Limerick, Province of Munster,
        Ireland.
  </marc:subfield>
</marc:datafield>
<marc:datafield tag="651" ind2="7" ind1="">
  <marc:subfield code="2">logainm.ie</marc:subfield>
  <marc:subfield code="a">Coshma</marc:subfield>
  <marc:subfield code="0">
    http://data.logainm.ie/place/145
  </marc:subfield>
</marc:datafield>
```

Fig. 3. Enhanced MARCXML catalogue record

The structured data published on the Web enables developers to reuse Logainm's data to build applications, taking advantage of query languages like the SPARQL Protocol and RDF Query Language (SPARQL) [18] that allow the user to go beyond string matching for searching for place names. For example by using SPARQL one can retrieve only entities of a specified type, with specific values for any property, or simply count the number of entities in a dataset.

6.1 Enhancing the Longfield Maps collection

As previously stated, a key use case which motivated the Linked Logainm project was the potential re-use of Logainm data by cultural heritage organisations and information professionals such as archivists and librarians. Some of the potential benefits identified at the start of the Linked Logainm project (with examples from the Longfield Maps) were:

- The potential to link to other digital objects or information from other Open Datasets. For example, by relating objects from other projects (e.g., the Royal Irish Academy's Historic Towns Atlas Project)[36] to Logainm entities they could be presented alongside the Longfield maps. Furthermore, by linking Logainm entities to international Open Datasets like GeoNames and DBpedia, contextual information about those locations could also be imported into the library catalogue.
- The potential to enhance discovery by drawing on Irish-language and historical forms of place names found in Logainm. For example, a user searching for "Ceara" or "Cera" could be directed to maps for the barony of Carra.
- The potential to enhance discovery by drawing the hierarchical information in the Logainm dataset. For example, a user searching the Longfield maps for a townland name not found in the existing records could be directed to maps for the related barony.

[36] http://www.ria.ie/research/ihta.aspx

– The potential to use Linked Logainm, along with the linking techniques described above, as a source against which to clean and normalise legacy metadata during conversion to a standard schema.

Having established the Linked Data URIs for the Logainm dataset, place names in the National Library's MARCXML metadata records of the Longfield Maps collection were compared and linked to place names in the Logainm dataset. The MARCXML records contained place names as subdivisions of Topical Subject Headings (i.e., the MARC *650* field); however, we decided to use information from the Geographic Coverage Note field (i.e., MARC *522* field). Although, this field normally contains free-text, uncontrolled values, in the case of the Longfield Map records the information had been entered with sufficient consistency to allow predictable parsing. Most importantly, the information in this field not only included a place name (n), but also hierarchical information such that the name of the parent place (p) could also be taken into account as per the linking methods described above. We manually checked approximately 300 of the 1570 linked records (19%) to determine that the URIs were correct and no errors were found.

In order to show the potential to link to other sources, we have created a demonstrator website that combines information about Irish places from various sources such as DBpedia (via the established links) and the Longfield Maps but also other content from Europeana, the Placenames Branch's digitised archival records, and the Irish Historic Towns Atlas' Dublin volumes. This demonstrator is available at `http://apps.dri.ie/locationLODer`.

In terms of enhancing the National Library's metadata and catalogue, our initial approach is to record the corresponding identifier in the Linked Data version of Logainm in the bibliographic record. An excerpt of an enhanced record is presented in Figure 3. The URIs was stored in the MARCXML record by adding a new Geographic Name Subject Heading (a MARC *651* field). Standard MARC linking practices were followed as closely as possible: for example the provenance of the heading was encoded using the *subfield "2"* with a second indicator value of 7, while the URI itself was encoded in the Authority Record Control Number subfield (*subfield "0"*). This approach is adequate for maintaining the relationship between the MARCXML record and Linked Logainm; however, it is hoped that as new library encoding standards such as the BIBFRAME initiative emerge, more standard techniques for relating bibliographic data to Linked Data resources will be agreed.

6.2 Potential Uses of Linked Logainm

The potential uses of relating bibliographic data to Linked Data include the ability to enhance discovery and search. By encoding library vocabularies in a machine-readable format, we bring these advantages into a networked environment and can continue to improve search and discovery systems on the Web. We can build systems that reflect the hierarchical nature of vocabularies and the relationships between the entities in them. As a linked bilingual authoritative

database of Irish place names, Logainm offers all the above benefits and uses to libraries. Each URI in the database contains the authoritative form of the place name as well as any spelling variants and it does this in both Irish and English. The URI for that place name would be stored in the system and would reference both English and Irish versions of that name. A search performed in either language or containing a misspelling would then bring up results relating to the correct resource leading to more precise search results without loss of recall (the number of search results returned).

As it contains information on the counties, baronies, townlands, parishes, and other features of Ireland as well as the relationships between them, Logainm can also act as a knowledge base for hierarchical relationships between these place names and other entities in the database. This could be used in search and discovery systems for navigation and visualisation of content. Additionally, as Logainm links to external geographic datasets, information could be automatically pulled in to a discovery interface, including images and maps, historical data, census data for the geographic area, related content in other cultural institutions and more. This would allow for the display of disparate sources which reference Logainm alongside each other, improving contextualisation and understanding of these resources.

Linked Logainm can also be used as a source against which to clean and create consistency in metadata. Messy metadata is one of the challenges in linking collections. Institutions may hold duplicate records, records with variations in the spelling of personal- and place-names, and typos. These metadata errors can hamper the effective discovery of content, analysis of data and provision of quality services to end-users. The use of the Linked Logainm dataset for reconciliation allows the standardisation of Irish geographic place names in both English and Irish forms in catalogue records. By linking to the Logainm database, the content management system will have access to all the hierarchical and related features information stored in Logainm, offering the potential to pull in this additional information. Linked Logainm also offered authoritative URIs for Irish place names which are linked to other LOD datasets. This allows access to data from external datasets such as Wikipedia and Geonames, which can also be pulled in to provide additional contextual information. Tools that allow users to process and clean messy data are becoming available, for instance Open Refine, and it is possible to reuse Linked Data from Open Refine [23] and hence also the new Linked Logainm dataset.

7 Conclusions and Future Work

In this paper we presented an introduction to Linked Data, authority data and libraries, and our approach to enhance library records, specifically the National Library of Ireland's Longfield Map Collection, with extra authority information about the places that are contained in these maps. By using Irish specific Geographic Linked Data, based on Logainm's data, we presented possible options to also extend the library's catalogue to enable searching for place names in Irish.

We also detailed the process of transforming the Logainm dataset into RDF and how to establish the links to other external datasets, namely DBpedia, LinkedGeoData, GeoNames, and the OCLC FAST subject heading schema, along with an initial evaluation of the determined links. Although some issues presented in this paper are specific to the datasets and the Irish language, for example the alternate spellings of place names in Irish or the matching of categories between Logainm and the external datasets, the presented methodology, also similar to [31], can be used in other countries with place name translations in multiple languages. Possible workarounds for these issues could be to rely on annotating the data using SKOS (e.g., using `skos:narrower` or `skos:altLabel`) and consider such annotations in the linking rules.

Future Work. Further work can be done in the Silk rules to attempt to obtain a larger number of links. However it should be taken into consideration that the precision of these links should remain high. Another type of entities whose links can be improved are streets, currently a large number of entities in Logainm refers to street names. Even though streets are present in some of the datasets we are linking to — DBpedia includes information about the most important streets in Dublin and other cities; LinkedGeoData contains streets exported from OpenStreetMap — our current linking rules do not provide adequate links for streets. Further work is planned to enhance discovery of the Longfield Maps with the National Library's online catalogue. Rather than add to or modify authority data stored at the Library management system level at this stage, the planned approach is to index Irish and variant forms of place names found in Linked Logainm into VuFind, the Library's discovery interface. This system has existing functionality to provide search suggestions based on cross-references found in traditional library authority records (i.e., MARC *4XX* authority fields); this functionality will also work for the Linked Logainm forms once correctly indexed. Furthermore, the National Library is currently evaluating the use of the Linked Logainm dataset to help with vocabulary standardisation as part of a conversion of a legacy metadata set which contains approximately 18,000 distinct, uncontrolled Irish place names.

Acknowledgements. The work presented in this paper has been funded in part by Science Foundation Ireland under Grants No. SFI/08/CE/I1380 (Líon-2) and SFI/12/RC/2289. DRI gratefully acknowledges its core funding from HEA PRTLI cycle 5 (NAVR).

References

1. Thomas Baker, Emmanuelle Bermès, Karen Coyle, Gordon Dunsire, Antoine Isaac, Peter Murray, Michael Panzer, Jodi Schneider, Ross Singer, Ed Summers, William Waites, Jeff Young, and ML Zeng. Library linked data incubator group final report. Incubator group report, World Wide Web Consortium, October 2011. URL `http://www.w3.org/2005/Incubator/lld/XGR-lld-20111025/`.

2. Tim Berners-Lee. Linked data design issues, 2006. URL http://www.w3.org/DesignIssues/LinkedData.
3. Diego Berrueta, Jose E. Labra, and Ivan Herman. XSLT+SPARQL : Scripting the Semantic Web with SPARQL embedded into XSLT stylesheets. In Chris Bizer, Sören Auer, Gunnar Aastrand Grimmes, and Tom Heath, editors, *4th Workshop on Scripting for the Semantic Web*, Tenerife, June 2008.
4. Stefan Bischof, Stefan Decker, Thomas Krennwallner, Nuno Lopes, and Axel Polleres. Mapping between RDF and XML with XSPARQL. *Journal on Data Semantics*, 1:147–185, 2012.
5. Christian Bizer, Tom Heath, and Tim Berners-Lee. Linked Data - The Story So Far. *Int. J. Semantic Web Inf. Syst.*, 5(3):1–22, 2009.
6. Christian Bizer, Jens Lehmann, Georgi Kobilarov, Sören Auer, Christian Becker, Richard Cyganiak, and Sebastian Hellmann. DBpedia - A crystallization point for the Web of Data. *J. Web Sem.*, 7(3):154–165, 2009.
7. Luis Manuel Vilches Blázquez, Boris Villazón-Terrazas, Victor Saquicela, Alexander de León, Óscar Corcho, and Asunción Gómez-Pérez. GeoLinked data and INSPIRE through an application case. In Divyakant Agrawal, Pusheng Zhang, Amr El Abbadi, and Mohamed F. Mokbel, editors, *GIS*, pages 446–449. ACM, 2010.
8. Dan Brickley and R.V. Guha. RDF vocabulary description language 1.0: RDF schema. Technical report, World Wide Web Consortium, February 2004. URL http://www.w3.org/TR/rdf-schema/.
9. Gillian Byrne and Lisa Goddard. The strongest link: libraries and linked data. *D-Lib Magazine*, 16(11/12), November 2010.
10. Karen Coyle. FRBR: the domain model. *Library Technology Reports*, 46(2):20–25, 2010. ISSN 0024-2586 (Print).
11. Karen Coyle. Library data in a modern context. *Library Technology Reports*, 46 (1):5–13, 2010. ISSN 0024-2586 (Print).
12. Karen Coyle. RDA in RDF. *Library Technology Reports*, 46(2):26–36, 2010.
13. Souripriya Das, Seema Sundara, and Richard Cyganiak. R2RML: RDB to RDF Mapping Language. W3C Recommendation, W3C, September 2012. URL http://www.w3.org/TR/2012/REC-r2rml-20120927/.
14. Europeana. Europeana data model primer. Technical report, Europeana, 2013. URL http://pro.europeana.eu/documents/900548/770bdb58-c60e-4beb-a687-874639312ba5.
15. John Goodwin, Catherine Dolbear, and Glen Hart. Geographical linked data: The administrative geography of great britain on the semantic web. *Transactions in GIS*, 12:19–30, 2008.
16. Alexander Haffner, Jeff Young, and Joachim Neubert. Cluster authority data, September 2011. URL http://www.w3.org/2005/Incubator/lld/wiki/Cluster_Authority_data.
17. Corey A. Harper and Barbara Tillett. Library of congress controlled vocabularies and their application to the semantic web. *Cataloging & Classification Quarterly*, 43(3/4), 2007.
18. Steve Harris and Andy Seaborne. SPARQL 1.1 Query Language. W3C recommendation, W3C, March 2013. URL http://www.w3.org/TR/sparql11-query/.
19. Pascal Hitzler, Markus Krötzsch, Bijan Parsia, Peter F. Patel-Schneider, and Sebastian Rudolph. OWL 2 web ontology language primer (second edition). Technical report, World Wide Web Consortium, December 2012. URL http://www.w3.org/TR/2012/REC-owl2-primer-20121211/.

20. IFLA Study Group on the Functional Requirements for Bibliographic Records. *Functional requirements for bibliographic records: final report*, volume 19 of *IFLA Series on Bibliographic Control*. K.G. Saur Verlag, Munich, 1998. ISBN 3-598-11382-X. URL `http://www.ifla.org/files/assets/cataloguing/frbr/frbr_2008.pdf`.
21. Antoine Isaac and Ed Summers. SKOS simple knowledge organization system primer. Technical report, World Wide Web Consortium, August 2009. URL `http://www.w3.org/TR/skos-primer/`.
22. Robert Isele, Anja Jentzsch, and Christian Bizer. Silk server - adding missing links while consuming linked data. In Olaf Hartig, Andreas Harth, and Juan Sequeda, editors, *COLD*, volume 665 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2010.
23. Fadi Maali, Richard Cyganiak, and Vassilios Peristeras. Re-using cool URIs: entity reconciliation against LOD hubs. In *Proceedings of the Linked Data on the Web Workshop 2011*, volume 813 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2011.
24. Frank Manola and Eric Miller. RDF Primer. W3C Recommendation, W3C, February 2004. URL `http://www.w3.org/TR/2004/REC-rdf-primer-20040210/`.
25. Eric Miller, Uche Ogbuji, Victoria Mueller, and MacDougall. *Bibliographic Framework as a web of data: Linked Data model and supporting services*. Library of Congress, 2012. URL `http://www.loc.gov/bibframe/pdf/marcld-report-11-21-2012.pdf`.
26. Ordnance Survey of Ireland. The Irish Grid – A Description of the Co-ordinate Reference System used in Ireland. Technical report, Ordnance Survey of Ireland, 1996. URL `http://osi.ie/OSI/media/OSI/Content/Publications/The-Irish-Grid-A-Description-of-the-Coordinate-Reference-System-Used-in-Ireland_1.pdf`.
27. Yves Raimond, Tom Scott, Patrick Sinclair, Libby Miller, Stephen Betts, and Frances McNamara. Case study: use of semantic web technologies on the BBC web sites, January 2012. URL `http://www.w3.org/2001/sw/sweo/public/UseCases/BBC/`.
28. Pat Riva. FRBR review group initiatives and the world of linked data. *JLIS.it*, 4 (1):105–117, 2013. ISSN 2038-1026.
29. Juan Martín Salas and Andreas Harth. NeoGeo Vocabulary Specification - Madrid Edition. Public draft, , February 2012. URL `http://geovocab.org/doc/neogeo/`.
30. Amit Singhal. Introducing the knowledge graph: things, not strings, May 2012. URL `http://googleblog.blogspot.ie/2012/05/introducing-knowledge-graph-things-not.html`.
31. Claus Stadler, Jens Lehmann, Konrad Höffner, and Sören Auer. Linkedgeodata: A core for a web of spatial open data. *Semantic Web*, 3(4):333–354, 2012.
32. David Stuart. *Facilitating access to the web of data*. Facet, London, 2011. ISBN 978-1-85604-745-6.
33. Barbara Tillett. What is FRBR?: a conceptual model for the bibliographic universe. *Technicalities*, 25(5), 2003. ISSN 0272-0884.
34. Daniel Vila-Suero. Library linked data incubator group: use cases. Incubator group report, World Wide Web Consortium, July 2011. URL `http://www.w3.org/2005/Incubator/lld/wiki/UseCases`.