Emerging practices for mapping and linking life sciences data using RDF - A case series

Web Semantics: Science, Services and Agents on the World Wide Web 14 (2012) 2–13

Contents lists available at SciVerse ScienceDirect

Web Semantics: Science, Services and Agents on the World Wide Web

journal homepage: www.elsevier.com/locate/websem

Ontology paper

Emerging practices for mapping and linking life sciences data using RDF — A case series

M. Scott Marshall [a,b,*], Richard Boyce [c], Helena F. Deus [d], Jun Zhao [e], Egon L. Willighagen [f], Matthias Samwald [g,h], Elgar Pichler [i], Janos Hajagos [j], Eric Prud'hommeaux [k], Susie Stephens [l]

[a] *Informatics Institute, University of Amsterdam, Amsterdam, The Netherlands*
[b] *Department of Medical Statistics and Bioinformatics, Leiden University Medical Center, Leiden, The Netherlands*
[c] *Department of Biomedical Informatics, University of Pittsburgh, Pittsburgh, PA, USA*
[d] *Digital Enterprise Research Institute, National University of Ireland at Galway, Ireland*
[e] *Department of Zoology, University of Oxford, Oxford, UK*
[f] *Division of Molecular Toxicology, Institute of Environmental Medicine, Karolinska Institutet, Stockholm, Sweden*
[g] *Section for Medical Expert and Knowledge-Based Systems, Center for Medical Statistics, Informatics, and Intelligent Systems, Medical University of Vienna, Vienna, Austria*
[h] *Vienna University of Technology, Vienna, Austria*
[i] *Department of Chemistry and Chemical Biology, Northeastern University, Boston, MA, USA*
[j] *Stony Brook University School of Medicine, Stony Brook, NY, USA*
[k] *World Wide Web Consortium, MIT, Cambridge, USA*
[l] *Janssen Research & Development, LLC, Radnor, PA, USA*

ARTICLE INFO

Article history:
Received 4 March 2011
Received in revised form
9 September 2011
Accepted 27 February 2012
Available online 3 April 2012

Keywords:
Linked Data
Semantic web
Health care
Life sciences
Data integration

ABSTRACT

Members of the W3C Health Care and Life Sciences Interest Group (HCLS IG) have published a variety of genomic and drug-related data sets as Resource Description Framework (RDF) triples. This experience has helped the interest group define a general data workflow for mapping health care and life science (HCLS) data to RDF and linking it with other Linked Data sources. This paper presents the workflow along with four case studies that demonstrate the workflow and addresses many of the challenges that may be faced when creating new Linked Data resources. The first case study describes the creation of linked RDF data from microarray data sets while the second discusses a linked RDF data set created from a knowledge base of drug therapies and drug targets. The third case study describes the creation of an RDF index of biomedical concepts present in unstructured clinical reports and how this index was linked to a drug side-effect knowledge base. The final case study describes the initial development of a linked data set from a knowledge base of small molecules.

This paper also provides a detailed set of recommended practices for creating and publishing Linked Data sources in the HCLS domain in such a way that they are discoverable and usable by people, software agents, and applications. These practices are based on the cumulative experience of the Linked Open Drug Data (LODD) task force of the HCLS IG. While no single set of recommendations can address all of the heterogeneous information needs that exist within the HCLS domains, practitioners wishing to create Linked Data should find the recommendations useful for identifying the tools, techniques, and practices employed by earlier developers. In addition to clarifying available methods for producing Linked Data, the recommendations for metadata should also make the discovery and consumption of Linked Data easier.

1. Introduction

Data integration is challenging because it requires sufficient domain expertise to understand the meaning of the data which is often undocumented or implicit in human-readable labels. *Linked Data* is an approach to data integration that employs ontologies, terminologies, Uniform Resource Identifiers (URIs) and the Resource Description Framework (RDF) to connect pieces of data, information and knowledge on the Semantic Web [1]. RDF makes it possible to use terms and other resources from remote locations together with one's own local data and terms. In effect, the ability to create assertions that mix local and remote namespaces makes it possible to publish and access knowledge

* Corresponding author. Tel.: +31 71 5269111.
E-mail addresses: mscottmarshall@gmail.com, marshall@science.uva.nl (M. Scott Marshall).

1570-8268/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
doi:10.1016/j.websem.2012.02.003

distributed over the Web using common vocabularies. Expressing information as Linked Data shifts some of the integration burden from data consumers to publishers, which has the advantage that data publishers tend to be more knowledgeable about the intended semantics.

The benefits of Linked Data are particularly relevant in the life sciences, where there is often no agreement on a unique representation of a given entity. As a result, many life science entities are referred to by multiple labels, and some labels refer to multiple entities. For example, a query for Homo sapiens gene label "Alg2" in Entrez Gene (http://www.ncbi.nlm.nih.gov/gene) returns multiple results. Among them is one gene located on chromosome 5 (Entrez ID: 85365) and the other on chromosome 9 (Entrez ID: 313231), each with multiple aliases. While a geneticist might refer to 'Alg2' without confusion among her lab colleagues, doing so in published data would present a barrier to future data integration as a correct interpretation would require the context in which these two genes are identified (e.g., the chromosome). If instead a Linked Data approach is taken to ensure that these two labels are semantically precise *a priori* (i.e., during data publication), then the burden of integration would be reduced.

There are several motivations to publishing Linked Data sets as indicated by the following potential use cases:

- Shareability: A *data provider or publisher* would like to make some existing data more openly accessible, through standard, programmatic interfaces such as SPARQL or resolvable URIs. A *scientist* wants to provide early access to data (pre-publication) to her research network.
- Integration: *A developer* desires to create and maintain a list of links between different RDF data sets so that she can easily query across these data sets.
- Semantic normalization: A *computer science researcher* is interested in indexing an existing RDF data set using a set of common ontologies, so that the data set can be queried using ontological terms.
- Discoverability: A *bench biologist* would like to be able to discover what is available in the Semantic Web related to a set of proteins, genes or chemical components, either as published results, raw data, or tissue libraries.
- Federation: A pharmaceutical company desires to retrieve data from sources distributed across its enterprise using SPARQL.

Participants of the World Wide Web Consortium, Health Care and Life Sciences Interest Group (HCLS IG) have been making health care and life sciences data available as Linked Data for several years. In 2009, members of the interest group published collectively more than 8.4 million RDF triples for a range of genomic and drug-related data sets and made them available as Linked Data [2]. More recently, members have begun enriching the LODD data sets with data spanning discovery research and drug development [3]. The interest group has found that publishing HCLS data sets as Linked Data is particularly challenging due to (1) highly heterogeneous and distributed data sets; (2) difficulty in assessing the quality of the data; (3) privacy concerns that force data publishers to de-identify portions of their data sets (e.g., from clinical research) [4]. Another challenge is to make it possible for the data consumer to discover, evaluate, and query the data. Would-be consumers of data from the Linked Open Data (LOD) cloud are confronted with these uncertainties and often resort to traditional data warehousing because of them.

This collective experience of publishing a wide range of HCLS data sets has led to the identification of a *general data workflow* for mapping HCLS data to RDF and linking it with other Linked Data sources (see Fig. 1). Briefly stated, the workflow includes the following steps for both structured and unstructured data sets:

1. Select the data sources or portions thereof to be published as RDF
2. Identify persistent URLs (PURLs) for concepts in existing ontologies and create a mapping from the structured data into an RDF view that can be used to transform data for SPARQL queries
3. Customize the mapping manually if necessary
4. Link concepts in the new RDF mapping to concepts in other RDF data sources relying as much as possible on URIs from ontologies
5. Publish the RDF data through a SPARQL endpoint or as Linked Data
6. Alternatively, if data is in a relational format, apply a Semantic Web toolkit such as SWObjects [5] that enables SPARQL queries over the relational schema
7. Create Semantic Web applications using the published data.

HCLS Linked Data developers may face many challenges when creating new Linked Data resources using the above general workflow. As such, the identification of practices for addressing such challenges is a necessary step to enable integration of health care and life sciences data sets. The questions listed in Table 1 summarize many of these challenges. The purpose of this paper is to provide practices that address these questions.

Before presenting the recommendations, we present real world case studies intended to demonstrate both the data flow shown in Fig. 1 and how some of the questions in Table 1 have been addressed by HCLS IG participants. The first case study describes the creation of linked RDF data from microarray data sets while the second discusses a linked RDF data set created from a knowledge base of drug therapies and drug targets [6]. The third case study describes the creation of an RDF index of biomedical concepts present in unstructured clinical reports and how this index was linked to a drug side-effect knowledge base. The final case study describes the initial development of a linked data set from a knowledge base of small molecules [7].

2. Mapping case studies[1]

2.1. Creating a representation of neurosciences microarray experiment results as RDF

Experimental biomedical results are often made available in semi-structured formats that can be used to assemble RDF representations. For example, microarray experiments typically include both the context in which the experiment was performed, including the anatomical location from which samples were collected, and the software used to extract and analyze the data as well as information regarding the list of genes that were deemed to be significant for the hypothesis under consideration.

As was reported at a recent conference [8], the results from three different microarray experiments, previously available only in spreadsheets, were represented in RDF format. The BioRDF task force of the interest group identified a set of concepts and relationships important to the description of a microarray experiment, including the list of differentially expressed genes (usually determined by statistical analysis and discussed in publications about the microarray experiment), as well as provenance items regarding institution, experimental context, significance analysis and data set description. Challenges were addressed in the following manner:

1. Three relevant neuroscience publications were identified from which microarray experimental data were to be extracted.

[1] Throughout the case studies, the approach taken to address any of the 14 questions shown in Table 1 is noted with a bold Q and the number of the question.

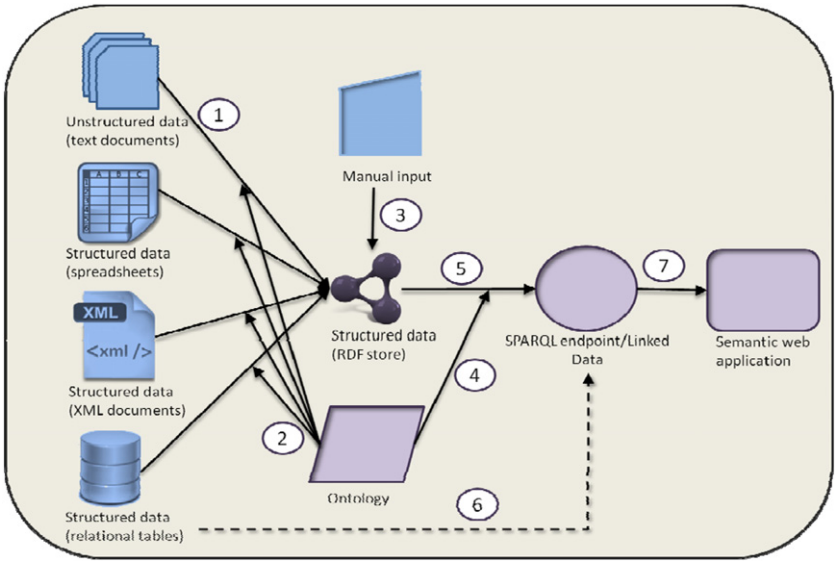

Fig. 1. Data flow depicting the ontology-driven mapping of structured and unstructured data into RDF format and the subsequent use of that data by a Semantic Web application via a SPARQL endpoint. See Section 1 for a description of the numbered elements.

Table 1
Fourteen questions that a Linked Data creator might have when creating an RDF Linked Data set following the workflow in Fig. 1.

Q1: What are the tools and approaches for mapping relational databases (RDBs) to RDF?	**Q8:** How should metadata and provenance be handled?
Q2: Are some relational schemas easier to map to RDF than others and is a direct equivalence mapping better than a transformation?	**Q9:** Under which license should I make the data available?
Q3: How should the RDF representation be mapped to global ontologies or reference terminologies?	**Q10:** Does data always need to be provided as Linked Data, or is a SPARQL endpoint enough?
Q4: How to interlink instances to other existing data sources?	**Q11:** How can I make it easier for people and programs to find and use my published RDF?
Q5: Does all data need to be mapped to the same ontologies in order to achieve interoperability?	**Q12:** What tools make use of the Linked Data once it is available?
Q6: How should the URIs and namespaces be determined?	**Q13:** How to convert non-relational information to RDF?
Q7: What should be done if there are gaps in the current ontology landscape?	**Q14:** Can I use automated reasoning with Linked Data?

From the supplementary material, spreadsheets containing the experimental results after statistical analysis were collected.

2. Contextual descriptors that were deemed relevant for representing the provenance of microarray experiments were extracted from the publications and the resulting list was manually mapped to URIs available through services at BioPortal [9] from the National Center for Biomedical Ontology (NCBO). These were identified manually because the NCBO SPARQL endpoint was not yet available. To ensure interoperability with other LODD data sets, the following domain ontologies were selected using NCBO's BioPortal: NIFSTD [10], MAGE-TAB [11], and DOID [12]. The complete list of these descriptors is available at http://biordfmicroarray.googlecode.com/hg/biordf.rdf# (**Q3**).
3. URIs for entities such as genes and diseases were reused from existing data sets such as Entrez Gene (made available as RDF through the Bio2RDF project), which provides information about the chromosomal position of the gene, its presence in multiple taxonomic groups and publications where the gene is mentioned, and the Human Disease Ontology, available from BioPortal. The criteria for this selection were based on the availability of metadata for such URIs, either when dereferenced or exposed as SPARQL endpoints (**Q4**).
4. Microarray experiment results depend substantially on experimental context, making it important to link the data not only to the institution where the experiment was performed, but also to the experimental conditions and, for example, the statistical package used to analyze the results as well as the version of the RDF documents produced. The vocabulary of interlinked data sets (VoID [13]) and the Provenance Vocabulary [14] were used for creating those assertions (**Q8**).
5. Because there is a gap in the ontology landscape regarding statistical methodologies and workflows (**Q7**), only some of the analytical and statistical details of how each gene list was produced could be located in the associated literature and linked to known ontologies. To address this challenge, the HCLS BioRDF task force engaged with stakeholders, such as with members of the Functional Genomics Data Society (FGED) and the ArrayExpress team at the European Bioinformatics Institute (EBI). The software typically used for microarray analysis (R Bioconductor packages [15]) was found to provide an easy method to report results linked to gene ontology annotations. Incorporating an RDF output option into widely-used tools such as Bioconductor would further benefit Linked Data efforts by incorporating the desired semantics into the output of the microarray analysis.

The representation of microarray experiment results in RDF enabled the assembly of queries that test hypotheses about drugs potentially useful for the treatment of Alzheimer's Disease, for example by looking up drugs in DrugBank related to genes that have been differentially expressed in a microarray experiment. Example queries that demonstrate the above data integration as well as query federation are available at: (http://purl.org/net/biordfmicroarray/demo).

2.2. DrugBank as RDF

DrugBank [16] is a repository for FDA-approved drugs that contains information about a drug's chemical and pharmacological

properties as well as sequence, structure, and pathway information on each of the drug's known biochemical targets. DrugBank's source data is accessible as either an XML data dump or Drug Cards, a structured flat file format used to describe information about drugs. DrugBank has been mapped to the RDF format as part of the LODD task force activities [2] RDF DrugBank has been interlinked with many other drug-related data sets that are part of the LODD task force including DailyMed [17], which provides complementary information about marketed drugs, and LinkedCT [18], which publishes clinical trial information about drugs.

In brief, the source data in Drug Cards format was downloaded and loaded into a relational database. This relational database was then published in RDF using the D2R server [19]. D2R enabled the task force to both publish a SPARQL endpoint on top of the relational database [20] and make the RDF data available as a physical data dump [21] (**Q10**).

The D2R server requires a mapping file so that it can convert SPARQL queries into SQL. Members of the LODD task force defined the D2R mapping file based on DrugBank's XML data schema. The concepts and properties used to describe the RDF representation of DrugBank were all defined under a local namespace (http://www4.wiwiss.fu-berlin.de/drugbank/) (**Q6**).

There are two areas for improvement of the initial release of RDF DrugBank. Firstly, no upper ontology was used to describe DrugBank's drug-related concepts (**Q3**). The recent proposal of the Translational Medicine Ontology (TMO) [3] published by the W3C HCLS TMO task force provides an ideal solution. Another drawback is that the LODD task force's linking of DrugBank to other drug-related Linked Data resources, was based solely on the string labels of the drugs (**Q4**). However, the RxNorm ontology [22] provides a mapping between drug names (both branded and generic) and product codes, which could provide important guidance on this relatively crude mapping. Using the TMO to describe DrugBank and making use of RxNorm to improve our data interlinking are part of the future release plan of RDF DrugBank.

2.3. Creating a Linked Data semantic index of concepts present in de-identified and unstructured clinical data

Unstructured "free text" data is a prevalent form of biomedical data that is sometimes quite useful for research. For example, the written protocol for a pre-market drug trial contains information that can indicate if the study qualifies for inclusion in a systematic review or meta-analysis of a drug's efficacy or safety. In the post-market research context, the notes taken by a clinician during a patient encounter might contain evidence of a drug's beneficial effects that could complement other sources of pharmacovigilance data such as prescription claims data. In this use case, we demonstrate how the Linked Data workflow can be used in conjunction with NLP tools to construct a *Linked Data semantic index* (LDSI) of unstructured clinical notes that may enable researchers to identify clinical notes that might be of use in their research.

The University of Pittsburgh's NLP repository (Pitt Repository) [23] is a collection of clinical data that may be shared with other qualified researchers after they go through an approval process and sign a data sharing agreement.[2] The repository contains more than 100,000 reports from over 20,000 patient encounters that occurred in 2007 at the University of Pittsburgh Medical Center. Several different types of reports exist in the system including radiology (RAD), cardiology, consultation, operative and emergency department reports (ER), in addition to progress notes, history and physical exams (HP), and discharge summaries (DS). The 'free text' for each note has been embedded in simple XML (Fig. 2) that includes other tags indicating the report type, subtype, date of creation, unique ID, and ICD-9 triage or discharge diagnoses assigned during a patient visit. Currently, researchers are able to request specific reports by limiting type and/or constraining the search to particular ICD-9 diagnoses but are unable to know in advance how many reports mention certain medications or procedures as free text. Also, because ICD-9 diagnoses are used for billing, they can be an inaccurate representation of the current or historical conditions experienced by a given patient.

It is possible to create an RDF-formatted LDSI into the Pitt Repository that can be posted publicly for interested researchers to explore using concepts mentioned in the free text that are present in other Linked Data sources. The identification of BioPortal concepts [9] present in specific reports can be done automatically using the freely available NCBO Annotator [24]. The NCBO Annotator takes as input unstructured text and outputs the string location within the text of concepts from any of more than 200 biomedical ontologies. The system can be run from a browser, by submitting unstructured text to a REST Web Service or by running a pre-configured virtual machine [25]. For this use case, the annotator was run over 1960 notes (498 RAD, 482 HP, 491 ER, and 489 DS). The code used to generate this example is available at [26]. The resulting semantic index contains 116,855 triples and is accessible as a SPARQL endpoint [27].

The following decisions were made during the creation of this data set:

1. XML output was requested from the NCBO Annotator so that the program's annotations could be more easily integrated into the RDF semantic index (see Fig. 3).
2. The RxNorm [22] drug ontology was selected so that the NCBO Annotator would identify drug concepts that could be linked with the SIDER node of the LODD (**Q3**). This decision was made after manually determining that most drug terms in the clinical notes would be identified by the Annotator using RxNorm and that many of the RxNorm concepts were exact string matches to SIDER terms.
3. The URIs for concepts present in the reports were provided by the NCBO Annotator and links to a definition for each concept was created using `rdfs:isDefinedBy` and the NCBO RDF term service [28] (**Q6**).
4. Resources in the resulting semantic index were linked to resources in the Linked Data version of SIDER [29] by configuring the SILK program [30] to create `owl:sameAs` mappings for exact string matches between `rdfs:label` attributes for the two resources (**Q4**).
5. Where possible, the guidelines of the Banff Manifesto [31] and recommendations of the W3C BioRDF task force [32] were followed. For example, the resulting index contains descriptive meta-data (e.g., Dublin Core Terms "title", "creator", "date", "publisher", "license"), no blank nodes, and only open source tools (Python [33] and RDFLib [34]) were used throughout the process to create the LDSI (**Q8, Q9**).

Fig. 4 shows a sample extracted from the final LDSI by querying for all clinical notes containing drugs associated with the side effect *hyperkalemia* in the SIDER knowledge base. While the new Linked Data set makes it easier for interested researchers to know if free text clinical reports available for secondary research mention specific biomedical concepts of interest, there are some limitations to the current version. The data set is explicitly linked to only one node in the LODD data cloud and the graph provides metadata on its source and license but not about the software used for its creation (**Q8**), nor is it registered in such a way as to facilitate discovery (**Q11**). Finally, the precision and recall of the index has been estimated (data not shown) but is not made explicit within the graph. We plan to address these issues in future work.

[2] An Institutional Review Board (IRB) approved the use of de-identified data for the purposes of this research (IRB# PRO11010367).

```
<report>
  <checksum>...</checksum>
  <subtype>CHEST</subtype>
  <type>RAD</type>
  <chief_complaint>COUGH</chief_complaint>
  <admit_diagnosis>786.2</admit_diagnosis>
  <discharge_diagnosis>382.9,465.9,530.81,</discharge_diagnosis>
  <year>200X</year>
  <download_time>2011-12-12</download_time>
  <updatetime/>
  <deid>v.6.XX.XX.X</deid>
  –
 – <report_text>
    EXAMINATION: TWO VIEWS OF THE CHEST ON **DATE[Jan XX 200X]. HISTORY: Cough and congestion.
    DISCUSSION: The cardiothymic silhouette is of abnormal size. There is definite consolidative infiltrate and effusion. There is
    also distention of the gastric air bubble. IMPRESSION: There is definite consolidative infiltrate.
  </report_text>
</report>
```

Fig. 2. A fictional report from the University of Pittsburgh NLP repository. The NCBO Annotator and other open source tools were used to create Linked Data semantic index (LDSI) of biomedical concepts present in the report text and link it to the SIDER.

```
<?xml version="1.0" encoding="UTF-8"?>
<success>
  ... meta-data on use of the NCBO Annotator ...
  <data>
    <annotatorResultBean>

      <statistics>
        ... statistics regarding annotation process ...
      </statistics>

      <parameters>
        ... paramaters to control the program and its output ...
      </parameters>

      <annotations>
        <annotationBean>
          <score>28</score>
          <concept>
            ... meta-data on an annotated concept ...
            <fullId>http://purl.bioontology.org/ontology/NCIM/C0013687</fullId>
            <preferredName>effusion</preferredName>
          </concept>
        </annotationBean>

      <ontologies>
        ... the ontologies from which concepts were identified
      </ontologies>
    </annotatorResultBean>
  </data>
</success>
```

Fig. 3. A sample of the results returned from the NCBO Annotator. The Annotator provided Persistent URLs (PURLs) to each biomedical concept it recognized in the report text. These were both used in the RDF semantic index and mapped to SIDER 'drug' and 'side effect' entities.

2.4. ChEMBL as RDF

ChEMBL is a Creative Commons-licensed database of bioactive drug-like small molecules [7,35]. Within the context of pharmaceutical bioinformatics, ChEMBL was converted into RDF and made available as a SPARQL endpoint and a Linked Data resource [36]. The data has since been used for general data analysis [37] as well as the development of predictive models for certain biochemical properties [38]. In this latter study, a platform was developed to enable the user to interactively select data from the ChEMBL database. A set of SPARQL queries is used to create lists or protein targets for selection by the user. Other SPARQL queries are used to retrieve drugs and drug-like molecules that target the selected proteins.

The mapping of the relational database content of ChEMBL to RDF is a project in progress, and illustrates the choices that need to be made in addressing several of the questions in Table 1.

Since the data can be downloaded as SQL statements [39], a conversion to structured data was unnecessary. To create RDF for the ChEMBL data, SQL was used to first populate a local mySQL database. Then, custom scripts, written in PHP [40], were used to create RDF statements, making use of existing PHP code that was available for generating HTML pages (**Q1**). The script was modified to generate simple *N*-Triples to make it easier to upload the large ChEMBL RDF graph by uploading it in parts. The script source code uses a customized mapping of tables and relations to classes and predicates. For the classes and predicates where no existing ontology was used, a namespace was created using a domain under control of the person conducting the conversion (http://rdf.farmbio.uu.se/chembl/onto/#) (**Q6**) Adoption of established ontologies is in progress (**Q5**).

```
<?xml version="1.0" encoding="UTF-8"?>
  <rdf:RDF
   xmlns:bl="http://purl.org/net/nlprepository/test#"
   xmlns:dc="http://purl.org/dc/terms/"
   xmlns:owl="http://www.w3.org/2002/07/owl#"
   xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
   xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
  >
  <rdf:Description rdf:about="http://purl.org/net/nlprepository/test#">
    <dc:publisher>Department of Biomedical Informatics, University of Pittsburgh</dc:publisher>
    <dc:creator>Richard D Boyce, PhD</dc:creator>
    <dc:license rdf:resource="http://www.opendatacommons.org/licenses/by/"/>
    <dc:title>Test Semantic Index for the University of Pittsburgh NLP Repository</dc:title>
    <dc:date>2011-06-30</dc:date>
  </rdf:Description>

  <rdf:Description rdf:about="http://purl.org/net/nlprepository/test#report_106">
    <dc:Description>DS</dc:Description>
    <bl:reportSubtype>ORTHO DISCHARGE</bl:reportSubtype>
    <dc:Identifier rdf:datatype="http://www.w3.org/2001/XMLSchema#integer">106</dc:Identifier>
    <dc:Date>2007</dc:Date>
    <bl:ncboAnnotion rdf:resource="http://purl.bioontology.org/ontology/RXNORM/225036"/>
    ... other drug entities ...
  </rdf:Description>

  <rdf:Description rdf:about="http://purl.bioontology.org/ontology/RXNORM/225036">
    <rdfs:label>Lovenox</rdfs:label>
    <rdfs:isDefinedBy rdf:resource="http://rest.bioontology.org/bioportal/rdf/44775/225036"/>
    <owl:sameAs rdf:resource="http://www4.wiwiss.fu-berlin.de/sider/resource/drugs/772"/>
  </rdf:Description>

  ... other reports and resource descriptions ...
</rdf:RDF>
```

Fig. 4. A sample of the LDSI extracted by querying for all clinical notes containing drugs associated with the side effect *hyperkalemia* in the SIDER knowledge base. The top portion of this example shows the RDF metadata for the entire LDSI. Then follows an RDF description for each report declaring its type, subtype, date of creation, and all drug entities found by the Annotator in the report text. The complete definition for each drug entity is available from the resource assigned to the `rdfs:isDefinedBy` attribute in the entity's description. An `owl:sameAs` tag is present in the description if the annotated entity was mapped to a SIDER drug.

In ChEMBL, compounds are often molecules with an exact chemical structure, and there are many cases where a drug compound is in fact a mixture of chemical structures [41]. Therefore, strongly typing compound instances would have led to inconsistencies. The use of an existing ontology, like CHEMINF [42] or the Drug Ontology [43], is in progress, but the database does not contain all necessary information to decide the correct type of the entries in the compounds table. At the time of writing, this ambiguity is being discussed with other members of the open chemical community (**Q7**) and existing ontologies may be extended to capture the complexity of the drug concept accurately (**Q2**).

The data were made available via a SPARQL endpoint at [44], but the main access point for querying the endpoint is a web page using an interface called SNORQL (**Q12**) [45]. This web page serves two purposes: first, the SPARQL endpoint is a web front-end that comes with the Virtuoso software (http://virtuoso.openlinksw.com/), and we chose not to program using the Virtuoso APIs in order to simplify software upgrades; secondly, the SNORQL endpoint provides a simplified user interface [46] with direct links to query all classes and properties. The SNORQL endpoint is customized, links to the data source, and lists the license under which the data is available (CC-SA Unported [35]) (**Q9**).

ChEMBL data were also made available as Linked Open Data [47] enabling individual resources to be dereferenced (**Q10**). This allows other RDF resources to link to entries in the ChEMBL-RDF data set. The hosted data is available as RDF in various serialization formats and also as HTML making it more discoverable (**Q11**).

3. Emerging practices for handling issues that a Linked Data publisher may encounter

The variety of goals and issues presented in the four case studies suggest that no single set of rules would be able to address all of the heterogeneous information needs that exist within the HCLS domain. However, discussion within the HCLS IG has led to the following set of recommendations that address each of the 14 questions listed in Table 1.

Q1. *What are the tools and approaches for mapping relational databases to RDF?*

Relational databases (RDBs) are the mainstay of data management and a prevalent source of data. A popular expression in the Semantic Web community is '*leave the data where it lives*' and create an RDF view—resulting in synchronized and always up-to-date access to the source data. Many Semantic Web practitioners, including some in the pharmaceutical industry and in the HCLS IG, prefer to operate from an RDB foundation. In fact, some prefer to create a schema and import unstructured data sources into an RDB before working on the RDF rendering of that data. For this reason, key technology will be found in the area of mapping RDB to RDF (RDB2RDF).

The survey by the RDB2RDF Incubator Group [48] provides a review of the state-of-the-art for mapping between RDB and RDF. Mapping the content of a relational database into RDF is an active area of research that has led to the development of new tools and approaches such as SWObjects [49]. Entities in life science DBs often have complex relationships. The semantics of these entities and their relationships can often be expressed using existing life science ontologies. This not only promotes the reuse of existing knowledge resources but also has the potential to enhance the interoperability between different RDF data sets. Hence, for mapping life science data to RDF, one of the most important aspects is the capability of representing the domain semantics that is not explicitly defined in the relational algebra of RDBs.

The tools for RDB2RDF must be able to support customized mapping definitions. Of nine tools reviewed by Sahoo et al. [48], three of them (Virtuoso [50], D2R [19] and Triplify [50]) support manual definition of the mappings, enabling users to use domain-semantics in the mapping configuration. Practical experiences have shown that the automatic mappings generated by tools like D2R provide a welcome starting point for customization [51]. Apart from customizable mapping definitions, the tools should also support the use of existing bio-ontologies in the mapping process, using services such as those hosted at NCBO's BioPortal [9,52].

An important feature of RDB2RDF tools is the ability to create a 'virtual view', or a semantic view of the contents of a (non-RDF) source database. For example, in addition to creating an RDF version of database contents using its mappings, D2R can provide a SPARQL endpoint and a *Linked Data interface* directly on top of the source non-RDF database, creating a virtual 'view' of databases. Such a 'semantic view' will guarantee up-to-date access to the source data, which is particularly important when the data is frequently updated with new information.

SWObjects creates semantic views by rewriting queries, from SPARQL to SQL, as well as from one RDF vocabulary to another. One of the advantages of using the relatively new SWObjects, is that the mappings used to create RDF views of RDB's are expressed using the SPARQL `CONSTRUCT` standard rather than a specialized, non-standardized language. In SWObjects, SPARQL `CONSTRUCT` statements form rules that can be chained together. This also makes it possible to pass constraints through a chain of `CONSTRUCT` statements that effectively rewrite a query to be more efficient and reduce the computational cost of query execution. A challenge of using SWObjects is that all queries and desired types of information must be anticipated in order to create the RDF views. Configuring such RDF views therefore requires knowledge of SPARQL, SQL, and of how the anticipated queries would translate into SQL. Ideally, this activity would be supported by an "SWObjects map editor tool".

The user experience in terms of performance depends on many architectural and design factors, including the optimization of the back-end RDB, size and structure of the data, and the specific queries required by the application. For example, in a federation, there are potentially invisible costs such as support for entailment regimes (materialized or on-the-fly), etc. However, query rewriting alone is a linear exercise that should not add any unexpected algorithmic overhead.

Developer costs in terms of maintenance, on the other hand, are more straightforward. Generating and storing RDF will require synchronization whenever either: the source data model, or the target RDF model, or the mapping logic between them changes. Therefore, query rewriting will generally lower developer costs that would otherwise be devoted to synchronization of code, data, and GUI. The main cost is the time of the developer to become familiar with configuring a query rewriting tool such as SWObjects.

Q2. *Are some relational schemas easier to map to RDF than others, and is a direct mapping better than a transformation?*

An RDB schema can vary in complexity from a single table to hundreds of tables with supporting reference tables. Many HCLS databases, such the UCSC Genome Browser [53], are complex and organized specifically for research and discovery. Tools that map RDB into RDF, like D2R or Virtuoso, provide an automated process to generate a mapping file [48], which converts every table into a class. For tables with a large number of columns this strategy can translate into a significant reduction of the initial time investment required for converting the contents of a database. In practice, a completely automated process is a convenient first step in publishing Linked Data, however it does not enable the fine-grained control that is needed to create RDF representations that align well with existing ontologies or related data sets. Also, it is often unnecessary to convert every table into a class and can create scaling problems. Domain-specific

enhancements during the mapping process create a much more accurate representation of the meaning and interrelatedness of statements within and across databases. Furthermore, they have the potential to drastically reduce the size and complexity of the resulting RDF data set [54].

RDB schemas can vary in their level of normalization as quantified by *normalized forms* [55]. The normalization process seeks to reduce the occurrence of functional dependencies within tables by breaking apart tables with many columns into component interlinked tables. The component tables often do not directly reflect the underlying "real world" objects which would be desired for an RDF representation. In practice, many databases are not normalized because the overhead of working with the schema is not worth the extra consistency and space savings that may result. For Online Analytical Processing (OLAP) applications in particular, highly normalized RDB schema designs can increase the complexity of SQL to such an extent that analysis becomes impractical.

In dimensional modeling [56], a logical design technique for data warehouses and OLAP, data are grouped into coherent categories that are easier to understand. This makes the mapping from dimensional representations to RDF, RDF Schema, or Web Ontology Language (OWL) classes more straightforward, and enables the default automated mapping process to yield better results. Furthermore, hierarchies in the dimension tables may help to indicate RDF classes and their relationships providing a richer data set.

How much data restructuring is needed depends on the underlying RDB schema structure and data contained, as well as the intended application of the interlinked data. These issues have also been faced by designers of traditional data warehouses and their data extract, transfer, and load (ETL) processes. In this context, linked data publishers can learn from recommended practices for building and populating data warehouses [57]. We recommend that the data publisher become as versed as possible in the underlying RDB schema, data content, and "business rules" that generated the stored data so as to best understand how the data should be structured in a Linked Data representation.

Q3. *How should the RDF representation be mapped to global ontologies or reference terminologies?*

NCBO's BioPortal is a convenient tool which can be used to identify public ontologies that best map to the entities in biomedical and clinical data sets. BioPortal contains URIs for concepts from more than 200 biomedical ontologies and reference terminologies including SNOMED CT and NCI Thesaurus. Selected terms from these ontologies can be collected in a *common resource ontology* (CRO) to avoid repeating class definitions for every data source [58]. For classes available in public ontologies, it is recommended that the CRO be built as a comprehensive representation of a domain by importing a standard set of orthogonal ontologies using the guidelines described in MIREOT [59].

Using a CRO offers some advantages:

- Scientists may have strong preferences for particular ontologies. When there is no general agreement about which ontology to use, one can use the definition of a proxy class in the CRO. The proxy can be linked to a number of public ontologies using URI aliases. Another advantage of the proxy approach is that in depth knowledge of the ontologies to reference by proxy is not necessary for the selection of terms for use in the RDF.
- Building a SPARQL query requires knowledge about the ontology or ontologies used during the mapping process. This information can be retrieved from the CRO.
- Using semantic wikis such as Semantic MediaWiki [60], the CRO can be maintained or extended by the scientists themselves.
- Semantic MediaWiki can store its data in RDF if the "triple store connector" plug-in is installed, enabling its use as a metadata repository for data source discovery. While at the time of this writing SPARQL is not integrated, Semantic MediaWiki extensions like RDFIO can be used to make the RDF data available as SPARQL [61].
- Custom definitions can be included within the CRO; this is an important step as it may happen that some class definitions required for mapping the RDB schema to RDF will not be available as public ontologies.

A challenge in making data available as Linked Data is the high level of expertise necessary to understand ontologies and dialects of description logic employed by OWL. For example, understanding and using the full potential of OWL 2 can be challenging, even with ontology tools like Protégé [62]. Fortunately, creating RDF (or an RDF view) that makes use of global ontologies does not always require comprehensive knowledge about OWL and ontologies. In the life sciences, tools such as the NCBO Annotator [24] can be used to automatically identify relevant concepts in existing ontologies.

Q4. *How to interlink instances to other existing data sources?*

A challenge that becomes increasingly relevant as linked data grows is the comprehensive review and identification of the data sources that may contain instances which can be linked to a converted RDF data set. This is a necessary step for making the data "5-star" as per Berners-Lee's definition (i.e., available on the Web in a machine-readable, non-proprietary format and linked to other RDF data sources) [63]. Creating interlinks between data often requires extensive research work on deciding and choosing a target linked data set. Once a target data source is identified for linking, connections established between the two data sets will enrich both. Ways to achieve this include constructing links using a scripting language that performs string processing, using a SPARQL `CONSTRUCT` query if both data sets are loaded into a single triple store, or using more specialized frameworks such as Silk [64]. Domain knowledge can provide more appropriate guidance on how links between data sets should be established and, in our practical experience [51], has been found to be effective for correcting automatically-created interlinks such as gene synonyms and genome locations. Structure descriptors, such as SMILES strings and InChI identifiers may be used to establish links between data sets containing drugs and small molecules.

Q5. *Does all data need to be mapped to the same ontologies in order to achieve interoperability?*

The more entities and predicates there are in common between two data sets, the easier those two data sets can be integrated ('joined' in a query), without loss of information or gain in noise. This is demonstrated in the example in Q10. The use of *any* ontology when mapping data is already an improvement over *tags* or unstructured text which often force integration to rely simply on lexical or 'string' similarity. Indeed, if the same ontologies or URIs are used in different data sets, the level of effort required is much less than if the ontologies are different. In case different ontologies have been used in each data set, it is sometimes possible to use alignment information between the two ontologies in order to translate to a single mapping.

Q6. *How should the URIs and namespaces be determined?*

Before transforming data into RDF (Fig. 1, step 5) or creating an RDF view of the source data (Fig. 1, step 6) one must decide on the URIs and namespaces to be used. Berners-Lee has clearly outlined good practices for Linked Data [65] and more information can be found in the Linked Data Tutorial [66]. Here are some of the general guidelines:

- Reusing existing URIs improves the connectivity of your data set with other data sets.

- Creating links that resolve to URIs published by others is highly recommended and necessary if the data source will be published as Linked Data.
- New URIs should be coined only if no existing URIs can be found. Use BioPortal for matching entities and their URIs (including ontologies from Open Biomedical Ontology Foundry [67]). Use http://identifiers.org/ to find URIs for information artifacts, such as a database records and gene accession numbers.
- If you create new URIs, be sure to have control over the namespace.
- Use PURLs (Persistent URLs) where possible. PURLs are meant to address changes of ownership and finance by redirecting the persistent URL to point to the current hosts and domains.

A number of projects can potentially supply URIs for the biomedical domain. The intention of Shared Names project [69] is to supply a common set of names or URIs for entities described in bioinformatics data records. Shared Names is based on a federation of PURL servers that create a vendor neutral and stable namespace for common URIs. Bio2RDF [68] is already widely used and serves URIs for many of the most common biomedical identifiers. An identifier system called MIRIAM from the European Bioinformatics Institute (EBI) has recently announced its adoption of URIs. A related consortium http://identifiers.org/ is developing URIs for the HCLS domains in cooperation with prominent members of the Semantic Web community, including Bio2RDF. We recommend using URIs for information records, such as, for example, a gene accession number.

We do not attempt here to describe all the technicalities of creating proper data URIs. Further information can be found in existing best practice documents [66,70].

Q7. *What should be done if there are gaps in the current ontology landscape?*

The life sciences domain is very dynamic and evolving. When a particular phenomenon cannot be described in enough detail with existing ontologies, it is good practice to contact the authors of the most relevant ontology to alert them of the gap. When coverage of existing ontologies does not supply the necessary detail, the creation of a specialized ontology might be unavoidable. This has been done, for example, with the Experimental Factor Ontology (EFO) [71] when the Ontology of Biomedical Investigation (OBI) [72] could not yet supply the needed terms. Of course, when such an ontology is used, it should also be available in the public domain for those who would access the published RDF.

When using an ontology, a common concern is to identify which entities should be classes and which should be instances. In general, most data should be described as an instance of an ontological class where possible. Classes might be used as a value in the metadata for a graph, for example, to indicate that a particular class of data is being provided in the data set.

Q8. *How should metadata and provenance be handled?*

Before making data accessible either via a Linked Data interface or a SPARQL endpoint, we must consider a number of augmentations to the data in order to make it more discoverable and trustworthy. Provenance is an important type of metadata because it describes the origins and processes that led to the data set. The nature of experimental data is determined by the experimental design and conditions at the time of the experiment. Such information can be important for the analysis and interpretation of the data. In a biomedical data set, laboratory protocols, instrumentation and handling conditions can all be relevant to particular types of analysis. This information can be thought of as equivalent to the *Materials and Methods* section of a biomedical article, which is meant to make it possible to reproduce the results discussed in the article. The best way to express this type of information in ontological terms is an area of ongoing investigation [8].

In the context of RDF publishing, another type of metadata describes the RDF itself, including provenance information describing the processes used to produce the serialization, the versions of any ontologies used to describe the data, and an unambiguous reference to the original data. The purpose of this metadata is to supply essential information about the RDF to make it discoverable on the Web. A list of essential information about the RDF would include: label of the data set, creator(s), original publishing date, date of last issue, frequency of updates, data source, version, license, vocabularies used, and the software used in the transformation. Ideally, it would be possible to reproduce the RDF using the original data and the processes and parameters described in the provenance for the RDF. For this reason, it is preferable to use and refer to the open source software that was used in the process. Other good practices would be to refer to the SPARQL endpoint that serves the graph and provide a representative query. Many of these practices are already possible with the VoID vocabulary [13] and the Provenance Vocabulary [14]. Also, the PROV Ontology (PROV) is in the process of being standardized by the W3C Provenance-Work Group with strong input from the HCLS community [73].

The ability to name a graph with a URI enables the storage of metadata statements about the graph in RDF. We recommend that such data be stored in each graph and that such information be aggregated for all graphs at the SPARQL endpoint. A promising approach from the W3C standards track is SPARQL 1.1 Service Description [74]. Most triple stores currently supply some form of named graph (*quad store*) functionality and, although it has not yet been standardized by the W3C, this seems to be on track for standardization [75,76].

Q9. *Under which license should I make the data available?*

We think that it is very important to annotate relevant information on the copyright and redistribution/modification rights of any Linked Data set. Such information should be presented in the location where users are expected to interact with the data, such as on individual web pages for resource, or on the SPARQL endpoint page. Without such statements it is unclear how data can be reused or modified. We think it preferable that a true Open Data license is used, such as proposed by the Panton Principles [77] and the Open Knowledge Foundation [78], the CC0 waiver [79], or the PDDL [80]. Data that is in the public domain (that has no copyright owner), should be labeled with the creator of that data set, as well as a clear notice about the public domain nature. For example, the Creative Common's Public Domain Mark 1.0 can be used [81].

In many cases the type of license is determined by the original data source. If this information is not clear, the original source should be asked to provide such information. Regardless of which license is chosen, we suggest including an RDF triple with the predicate <http://www.w3.org/TR/void/#license>, the URI of the published RDF graph as the subject and, ideally, the URI to a license ontology as the object. The HSLS IG is currently investigating whether an appropriate license ontology can be made available in order to complement VoID's effort.

Q10. *Does data always need to be provided as Linked Data, or is a SPARQL endpoint enough?*

Although a Linked Data interface (such as that provided by D2R) is preferable because it makes resource URIs dereferenceable, not everyone has the means to create one. The advantage of resolvable URIs is that when SPARQL endpoints do not use the same ontologies or link to each other, the RDF supplied by a resolved URI can sometimes make it possible to dynamically discover relevant remote data. Fortunately, a Linked Data interface can always be added to the URIs, which are accessible through a SPARQL endpoint, at a later time and can still serve to interconnect data sets, as long as ontological identifiers and common URIs are used.

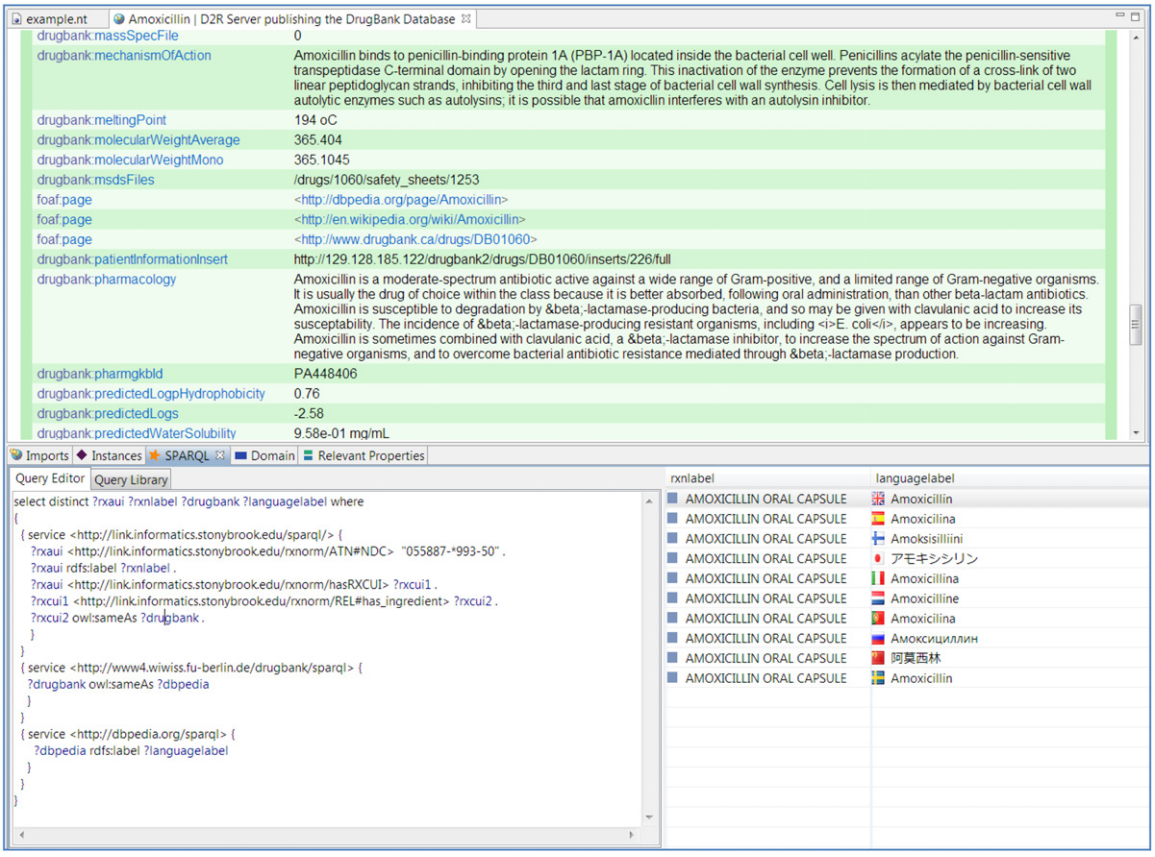

Fig. 5. A federated SPARQL connecting an Amoxicillin Capsule identified by the NDC code 055887-*993-50 to different language representations of the active ingredient. The screenshot from TopBraid Composer Free Edition (Versions 3.5.0) which utilizes Jena's ARQ show the results of running the federated SPARQL query. Note this query can be executed anywhere in the world where the three SPARQL endpoints can be accessed without installing any data locally on the client. *Top:* DrugBank provides more detailed chemical data on the active ingredient shown in the dereferencing of the URI. Different language labels are obtained from `rdfs:label` in DbPedia. *Bottom left:* source of the federated query. *Bottom right:* result set of executing the query. The RDF standard allows literals to be encoded in Unicode.

Fig. 5 illustrates this point using the example of a distributed (federated) SPARQL query for the string name, in multiple languages, of a drug product that is represented by an NDC code. The query is first applied to an SPARQL endpoint hosting an RDF version of the RxNorm "Rich Release Format". This resource includes, in addition to RxNorm, the names and attributes from additional drug ontologies. Because RDF RxNorm drug ingredients are linked to DrugBank through the CAS registry number of its active ingredients, and DrugBank is linked to DBPedia, the query can request all language renderings of the drug product present in DBPedia (see [82] for further details). While it is not expected that the average user would write a distributed query such as that shown in Fig. 5, software agents acting on a user's behalf could do so.

Q11. *How can I make it easier for people and programs to find and use my published RDF?*

An important part of improving the utility of the Web is by documenting the reliability and performance of information services. In the area of biomedical information services, BioCatalogue [83] describes and tracks the features and performance of thousands of bioinformatics services. The CKAN Data Hub registry makes it possible to "find, share and reuse open content and data, especially in ways that are machine automatable" [84]. The Data Hub registry has its own SPARQL endpoint for machine discovery. A SPARQL Endpoints Status page has been recently initiated by an independent source [85] that makes use of the Data Hub to provide an overview of reliability for the SPARQL endpoints in the Data Hub. Complementing this effort with descriptions of concepts, properties, and links to third party data sources may help users more easily query a new SPARQL endpoint.

Q12. *What tools make use of the Linked Data once it is available?*

There are a number of Linked Data browsers that enable browsing of Linked Data such as Disco [86], Tabulator [87], and Openlink Browser [87]. An overview of these types of browsers is available at [88]. There are also RDF crawlers such as Sindice [89], SWSE [90] and Swoogle [91]. While generic Linked Data browsers are useful for getting an overview of the raw data available on the web, they may not be practical for all end-users because the user interfaces are generally not very user-friendly and "on-the-fly" aggregation of distributed data is often slow. Fortunately, custom applications can be built that utilize SPARQL endpoints in a Service-Oriented Architecture (SOA) based on distributed vocabulary resources and Linked Data. In our experience, applications built using this approach are generally faster than applications that rely on URI/PURL resolution. Moreover, this approach generally makes it possible to create both web applications, such as those based on AJAX libraries [51], and stand-alone applications based on the same architecture and relying on the same API (SPARQL).

Q13. *How to convert non-relational information to RDF?*

Even though the ideal situation is to create an RDF view or directly map information from an RDB to Linked Data, there may be a situation in which information in other formats (CSV, XML, etc.) should be transformed into RDF. The xCurator project offers an end-to-end framework to transform a semi-structured (XML) source into high-quality Linked Data [92]. Also for XML, Gloze [93], which is part of Jena, uses the information available in the XML schema to convert bi-directionally between XML and RDF. CSV4RDFLOD [94] can be used to transform CSV (Comma Separated Values) flat files into RDF. Google Refine [95] is a general

data "Cleansing" tool that works with a plethora of formats: CSV, Excel, XML, etc. The Google Refine RDF extension [96] can be used to export a Google Refine project as RDF.

Q14. *Can I use automated reasoning with Linked Data?*

Automated reasoning is the process by which the axioms implicit in an ontology are made explicit by a program called reasoner. The reasoner infers the axioms that are implied by the assertions made in an ontology. For example, if *A* is a subclass of *B* and *B* is a subclass of *C*, the reasoner will infer that *A* is a subclass of *C*, since "sub class of" is a transitive relationship. Automated reasoning can be used to infer the class hierarchy of an ontology, check its consistency, perform queries against the ontology, or determine to which class an entity with certain features belongs.

Even though OWL automated reasoning is not yet efficient for large knowledge bases, algorithms are improving continuously. For example, OWL 2 now presents three profiles (QL, RL and EL) that optimize different kinds of reasoning. Automated Reasoning can be used in a data set to materialize (or "insert") inferred triples exploiting the axioms of the ontology [97]. Reasoning can also be used to check the compliance of the data against the ontology, especially with tools like Pellet ICV [98]. Finally, some triple stores offer the possibility of exploiting OWL semantics. For example, in Virtuoso, transitivity can be used in SPARQL queries using the TRANSITIVE keyword. Moreover, OWLIM offers the possibility of exploiting fragments of OWL semantics by approximating such semantics with rule sets [99].

In order to support automated reasoning, it is important that linked data creators consider carefully which entities they model as classes and instances. The discussion of what is best modeled as a class vs an instance is outside the scope of this paper but is covered in resources such as [100]. Typical practice by linked data developers has tended to describe most data as an instance of some ontological class and to use classes as values in the metadata for a graph; for example to indicate that a particular class of data is being provided in the data set.

4. Recommendations

We have proposed a set of practices that authors publishing HCLS data sets as Linked Data may find useful. Here, we highlight some of the most important points:

Create RDF views that anyone can use

- Use a mapping language to create an RDF view of the data when possible, rather data conversion and migration.
- When possible, use vocabularies that are openly available from an authoritative server like that provided by OBO and the NCBO for HCLS data.
- When faced with uncertainty about the proper term from an authoritative domain ontology, use a CRO that can redirect references to proper terms until they are replaced.
- Use rdfs:label and rdfs:comment generously to provide information to user interfaces.

Publish RDF so that it can be discovered

- Publish open access data whenever possible, as well as any associated software.
- Publish a URL to the software and mappings that you used to create the RDF.
- Register your data in the CKAN Data Hub. If it is a biomedical SPARQL endpoint, register it in BioCatalogue.
- Assign a graph URI to the RDF graph and add provenance and metadata about the graph URI to the graph itself. This practice makes it possible for visitors and crawlers to find out what is in the graph using SPARQL.

5. Conclusions

We have supplied four case studies of creating and publishing RDF for life sciences data sets and proposed recommended practices. Although our suggestions to the questions that may arise during Linked Data creation (Table 1) are oriented towards the HCLS domain, there is no reason why such practices could not be applied in other domains. For example, efforts are underway for a general source of ontologies and terminologies called the Open Ontology Repository [101] that would be much like BioPortal, but useful for scientists and researchers outside of the HCLS domain.

Finally, the principles and practices identified in this report have emerged from community practice and agreement rather than from a top-down approach. These principles and practices are being incorporated into a W3C Health Care and Life Sciences (HCLS) Interest Group Note [102]. However, they necessarily reflect the state of the art in the field and some of our practices may shift as new tools and resources are made available to the community.

Acknowledgments

We thank the reviewers and several colleagues for their comments on the manuscript, especially Claus Stie Kallesøe and Lee Harland. We acknowledge Mikel Egaña Aranguren for contributing ideas that were integrated into this paper while working on a W3C IG Note with a similar theme. We thank the participants of the Linked Open Drug Data task force and the W3C Health Care and Life Science (HCLS) Interest Group. Support for HCLS activities was provided by the World Wide Web Consortium (W3C). RB was funded by grant K12HS019461 from the Agency for Healthcare Research and Quality (AHRQ). The content is solely the responsibility of the authors and does not represent the official views of AHRQ.

References

[1] LinkedData, Linked Data—Connect Distributed Data across the Web, 2011. [Online] Available: http://linkeddata.org/ [Accessed: 31-Aug-2011].
[2] A. Jentzsch, J. Zhao, O. Hassanzadeh, K.H. Cheung, M. Samwald, B. Andersson, Linking open drug data, in: Triplification Challenge of the International Conference on Semantic Systems, pp. 3–6, 2009.
[3] J.S. Luciano, et al., The translational medicine ontology: driving personalized medicine by bridging the gap from bedside to bench, J. Biomed. Semant. (2011) 1–4 (Bio-Ontologies 2010 Special Issue).
[4] H.F. Deus, M.C. Correa, R. Stanislaus, M. Miragaia, W. Maass, J.S. de Lencastre Hermínia Almeida, S3QL: a distributed domain specific language for controlled semantic integration of life science data, BMC Bioinformatics 12 (1) (2011) 285.
[5] E. Prud'hommeaux, H. Deus, M.S. Marshall, Tutorial: query federation with SWObjects, in: Semantic Web Applications and Tools for Life Sciences 2010, 2010.
[6] C. Knox, et al., DrugBank 3.0: a comprehensive resource for 'Omics' research on drugs, Nucleic Acids Res. 39 (2010) D1035–D1041. no. Database.
[7] W.A. Warr, ChEMBL. An interview with John Overington, team leader, chemogenomics at the European Bioinformatics Institute outstation of the European Molecular Biology Laboratory (EMBL-EBI), J. Comput.-Aided Mol. Des. 23 (4) (2009) 195–198.
[8] H.F. Deus, et al. Provenance of microarray experiments for a better understanding of experiment results, in: ISWC 2010 SWPM, 2010.
[9] NCBO, NCBO BioPortal, 2012. [Online] Available: http://bioportal.bioontology.org/ [Accessed: 2012].
[10] NIFSTD, NIFSTD—Terms — NCBO BioPortal, 2011. [Online] Available: http://bioportal.bioontology.org/ontologies/45824?p=terms [Accessed: 11-Aug-2011].
[11] MAGE-TAB, MAGE-TAB model v1.1 prototype implementation, 2011. [Online] Available: http://wwwdev.ebi.ac.uk/microarray-srv/magetab/molgenis.do [Accessed: 11-Aug-2011].
[12] DOID, DOID, 2011. [Online] Available: http://www.berkeleybop.org/ontologies/owl/DOID [Accessed: 2012].
[13] K. Alexander, R. Cyganiak, M. Hausenblas, J. Zhao, Describing Linked Datasets with the VoID Vocabulary, 2011. W3C Interest Group Note 03 March 2011. http://www.w3.org/TR/void/. [Accessed: 10-04-2012].
[14] O. Hartig, J. Zhao, Publishing and Consuming Provenance Metadata on the Web of Linked Data, vol. 6378, Springer Berlin Heidelberg, Berlin, Heidelberg, 2010, pp. 78–90.

[15] Bioconductor, Bioconductor—Home, 2011. [Online] Available: http://www.bioconductor.org/ [Accessed: 11-Aug-2011].
[16] D.S. Wishart, et al., DrugBank: a comprehensive resource for in silico drug discovery and exploration, Nucleic Acids Res. 34 (2006) D668–D672. no. Database issue.
[17] DailyMed, DailyMed: about DailyMed, 2012. [Online] Available: http://dailymed.nlm.nih.gov/ [Accessed: 2012].
[18] LinkedCT, About LinkedCT, 2011. [Online] Available: http://linkedct.org/about/ [Accessed: 11-Aug-2011].
[19] C. Bizer, D2R Map—Database to RDF mapping language, 28-Sep-2010.
[20] Drugbank, Drugbank SPARQL Endpoint, 2011. [Online] Available: http://www4.wiwiss.fu-berlin.de/drugbank/sparql [Accessed: 11-Aug-2011].
[21] DrugBank, DrugBank RDF dump, 2011. [Online] Available: http://www4.wiwiss.fu-berlin.de/drugbank/drugbank_dump.nt.bz2 [Accessed: 11-Aug-2011].
[22] S. Liu, W. Ma, R. Moore, V. Ganesan, S. Nelson, RxNorm: prescription for electronic drug information exchange, IT Prof. 7 (2005) 17–23.
[23] UPitt, University of Pittsburgh NLP repository, 2011. [Online] Available: http://www.dbmi.pitt.edu/nlpfront [Accessed: 11-Aug-2011].
[24] C. Jonquet, M.A. Musen, N. Shah, A system for ontology-based annotation of biomedical data, Med. Inf. (2008) 144–152.
[25] NCBO, NCBO virtual appliance—NCBO Wiki, 2011. [Online] Available: http://www.bioontology.org/wiki/index.php/Category:NCBO_Virtual_Appliance [Accessed: 11-Aug-2011].
[26] R. Boyce, Python Script to Convert a *U* of Pitt clinical note to linked-data RDF, 2011. [Online] Available: http://www.pitt.edu/~rdb20/data/convert-annotated-report-to-rdf.py [Accessed: 2012].
[27] R. Boyce, SPARQL endpoint for the *U* of Pitt clinical notes linked semantic index, 02/2011, 2012. [Online] Available: http://dbmi-icode-01.dbmi.pitt.edu:8080/sparql [Accessed: 2012].
[28] BioPortal, BioPortal REST services—NCBO Wiki, 2012. [Online] Available: http://www.bioontology.org/wiki/index.php/NCBO_REST_services#RDF_Term_Service [Accessed: 2012].
[29] SIDER, SIDER LODD, 2011. [Online] Available: http://www4.wiwiss.fu-berlin.de/sider/.
[30] R. Isele, A. Jentzsch, C. Bizer, J. Volz, Silk—a link discovery framework for the Web of data, 2011. [Online] Available: http://www4.wiwiss.fu-berlin.de/bizer/silk/ [Accessed: 11-Aug-2011].
[31] Banff, SourceForge.net: Banff manifesto—bio2rdf, 2011. [Online] Available: http://sourceforge.net/apps/mediawiki/bio2rdf/index.php?title=Banff_Manifesto [Accessed: 11-Aug-2011].
[32] HCLSIG, HCLSIG Bio RDF subgroup/MinimalInformationAbout AGraph—W3C Wiki, 2011. [Online] Available: http://www.w3.org/wiki/HCLSIG_BioRDF_Subgroup/MinimalInformationAboutAGraph [Accessed: 11-Aug-2011].
[33] Python, python programming language—official website, 2012. [Online] Available: http://python.org/ [Accessed: 2012].
[34] RDFLib, RDFLib, 2012. [Online] Available: http://www.rdflib.net/ [Accessed: 2012].
[35] CC-SA, CC-SA unported, 2012. [Online] Available: http://creativecommons.org/licenses/by-sa/3.0/ [Accessed: 11-Aug-2011].
[36] Z. Beauvais, Featured dataset: ChEMBL-RDF, with Egon Willighagen, Kasabi Blog, 2011.
[37] M. Samwald, M. Dumontier, J. Zhao, J.S. Luciano, M.S. Marshall, K. Cheung, Integrating findings of traditional medicine with modern pharmaceutical research: the potential role of linked open data, Chin. Med. 5 (1) (2010) 43.
[38] E. Willighagen, et al., Linking the resource description framework to cheminformatics and proteochemometrics, J. Biomed. Semant. 2 (1) (2011) S6.
[39] ChEMBL, ChEMBL FTP directory, 2012. [Online] Available: ftp://ftp.ebi.ac.uk/pub/databases/chembl/ChEMBLdb/releases/chembl_09/README [Accessed: 11-Aug-2011].
[40] E.L. Willighagen, chembl.rdf, 2012. [Online] Available: https://github.com/egonw/chembl.rdf [Accessed: 2012].
[41] E.L. Willighagen, chem-bla-ics, 2012. [Online] Available: http://chem-bla-ics.blogspot.com/ [Accessed: 2012].
[42] J. Hastings, L. Chepelev, E. Willighagen, N. Adams, C. Steinbeck, M. Dumontier, The chemical information ontology: provenance and disambiguation for chemical data on the biological semantic web, PLoS ONE 6 (1) (2011).
[43] H. Stuckenschmidt, et al., Exploring large document repositories with RDF technology: the dope project, IEEE Intell. Syst. 19 (3) (2004) 34–40.
[44] E.L. Willighagen, ChEMBL SPARQL endpoint, 2012. [Online] Available: http://rdf.farmbio.uu.se/chembl/sparql [Accessed: 2012].
[45] E.L. Willighagen, ChEMBL Snorql endpoint, 2012. [Online] Available: http://rdf.farmbio.uu.se/chembl/snorql/ [Accessed: 2012].
[46] SNORQL, SNORQL—GitHub, 2012. [Online] Available: https://github.com/kurtjx/SNORQL [Accessed: 2012].
[47] E.L. Willighagen, ChEMBL-RDF on Kasabi, 2012. [Online] Available: http://beta.kasabi.com/dataset/chembl-rdf [Accessed: 2012].
[48] S.S. Sahoo, et al. A survey of current approaches for mapping of relational databases to RDF, w3org, 2009.
[49] SWObjects, SWObjects, 2011. [Online] Available: http://www.w3.org/2010/Talks/0218-SWObjects-egp/#(1) [Accessed: 10-Aug-2011].
[50] S. Auer, S. Dietzold, J. Lehmann, S. Hellmann, D. Aumueller, Triplify, in: Proceedings of the 18th international conference on World wide web - WWW '09, 2009, p. 621.
[51] J. Zhao, A. Miles, G. Klyne, D. Shotton, OpenFlyData: the way to go for biological data integration, Data Integr. Life Sci. (2009) 47–54.
[52] N.F. Noy, et al., BioPortal: ontologies and integrated data resources at the click of a mouse, Nucleic Acids Res. 37 (2009) W170–W173. no. Web Server issue.
[53] D. Karolchik, The UCSC genome browser database, Nucleic Acids Res. 31 (1) (2003) 51–54.
[54] K. Byrne, Having Triplets—holding cultural data as RDF, in: Proceedings of the ECDL 2008 Workshop on Information Access to Cultural Heritage, Aarhus, Denmark, September 18, 2008, 2008.
[55] C.J. Date, SQL and Relational Theory: How to Write Accurate SQL Code, O'Reilly Media, Inc., 2009, p. 404.
[56] R. Kimball, M. Ross, The Data Warehouse Toolkit: The Complete Guide to Dimensional Modeling, Wiley, 2002, p. 436.
[57] R. Kimball, J. Caserta, The Data Warehouse ETL Toolkit: Practical Techniques for Extracting, Cleaning Conforming, and Delivering Data, Wiley, 2004, p. 528.
[58] R. Verbeeck, T. Schultz, L. Alquier, S. Stephens, Relational to RDF mapping using D2R for translational research in neuroscience, Bio-Ontologies Meeting, ISMB 2010 [Online] Available: https://docs.google.com/viewer?a=v&pid=sites&srcid=ZGVmYXVsdGRvbWFpbnxiaW9vbnRvbG9naWVzc2lnMjAxMHxneDoxYmE1ZWQ5ZjExNDc5NzYy [Accessed: 12-Apr-2012].
[59] M. Courtot, et al. MIREOT: the minimum information to reference an external ontology term, Aug-2009.
[60] M. Krötzsch, D. Vrandečić, M. Völkel, Semantic MediaWiki, in: I. Cruz, et al. (Eds.), The Semantic Web—ISWC 2006, vol. 4273, Springer, Berlin, Heidelberg, 2006, pp. 935–942.
[61] S. Lampa, Extension:RDFIO—MediaWiki, 2010. [Online] Available: http://www.mediawiki.org/wiki/Extension:RDFIO [Accessed: 01-Mar-2011].
[62] H. Knublauch, R.W. Fergerson, N.F. Noy, M.A. Musen, The Protégé OWL plugin: an open development environment for semantic web applications, ISWC, 2004, 3298, pp. 229–243.
[63] T. Berners-Lee, Is your data 5*?, 2007. [Online] Available: http://www.w3.org/DesignIssues/LinkedData.html [Accessed: 2012].
[64] C. Becker, C. Bizer, M. Erdmann, M. Greaves, Extending SMW + with a linked data integration framework, in: International Semantic Web Conference, 2011, pp. 2–5.
[65] T. Berners-Lee, Linked data—design issues, 2006. [Online] Available: http://www.w3.org/DesignIssues/LinkedData.html.
[66] C. Bizer, R. Cyganiak, T. Heath, How to publish Linked Data on the Web, 2007.
[67] OBO, The open biological and biomedical ontologies, 2012. [Online] Available: http://www.obofoundry.org/ [Accessed: 2012].
[68] F. Belleau, M.-A. Nolin, N. Tourigny, P. Rigault, J. Morissette, Bio2RDF: towards a mashup to build bioinformatics knowledge systems, J. Biomed. Informatics 41 (5) (2008) 706–716.
[69] SharedNames, Shared Names, 2011. [Online] Available: http://sharedname.org/page/Main_Page [Accessed: 2012].
[70] L. Sauermann, R. Cyganiak, Cool URIs for the Semantic Web, 2011. [Online] Available: http://www.dfki.uni-kl.de/~sauermann/2007/01/semweburisdraft/uricrisis.pdf [Accessed: 2012].
[71] EFO, Experimental factor ontology, 2011. [Online] Available: http://www.ebi.ac.uk/efo/ [Accessed: 2012].
[72] OBI, OBI Ontology, 2012. [Online] Available: http://obi-ontology.org/page/Main_Page [Accessed: 2012].
[73] P. OWL, The PROV Ontology: Mode and Formal Semantics. W3C Working Draft 13 December 2011. http://www.w3.org/TR/prov-o/ [Accessed: 10-04-2012].
[74] G.T. Williams, SPARQL 1.1 Service Description, 2011.
[75] SPARQL, SPARQL query language implementation report, 2011. [Online] Available: http://www.w3.org/2001/sw/DataAccess/impl-report-ql [Accessed: 11-Aug-2011].
[76] RDF, RDF working group charter, 2011. [Online] Available: http://www.w3.org/2010/09/rdf-wg-charter.html [Accessed: 11-Aug-2011].
[77] P. Murray-Rust, C. Neylon, R. Pollock, J. Wilbanks, Panton Principles, 2010. [Online] Available: http://pantonprinciples.org/ [Accessed: 11-Aug-2011].
[78] OKF, Open Knowledge Foundation — promoting open knowledge in a Digital Age, 2011. [Online] Available: http://okfn.org/ [Accessed: 11-Aug-2011].
[79] CCO, CC0 Waiver, 2011. [Online] Available: http://wiki.creativecommons.org/CC0 [Accessed: 2012].
[80] ODC, ODC Public Domain Dedication and License (PDDL), 2011. [Online] Available: http://opendatacommons.org/licenses/pddl/ [Accessed: 2012].
[81] PDM, Public Domain Mark 1.0, 2011. [Online] Available: http://creativecommons.org/publicdomain/mark/1.0/ [Accessed: 11-Aug-2011].
[82] C. Bizer, et al., DBpedia—a crystallization point for the web of data, Web Semantics: Science, Services and Agents on the World Wide Web 7 (3) (2009) 154–165.
[83] J. Bhagat, et al., BioCatalogue: a universal catalogue of web services for the life sciences, Nucleic Acids Res. 38 (2) (2010) W689–W694.
[84] CKAN, CKAN—the Data Hub, 2012. [Online] Available: http://thedatahub.org/ [Accessed: 2012].
[85] P.-Y. Vandenbussche, CKAN—Public SPARQL endpoints availability, 2012. [Online] Available: http://labs.mondeca.com/sparqlEndpointsStatus/ [Accessed: 2012].
[86] C. Bizer, T. Gauß, Disco—Hyperdata Browser, 2007. [Online] Available: http://www4.wiwiss.fu-berlin.de/bizer/ng4j/disco/ [Accessed: 11-Aug-2011].
[87] Tabulator, Tabulator: generic data browser, 2011. [Online] Available: http://www.w3.org/2005/ajar/tab [Accessed: 2012].

[88] LOD, LOD Browser Switch, 2012. [Online] Available: http://browse.semanticweb.org/ [Accessed: 2012].
[89] Sindice, Sindice – The Semantic Web index, 2012. [Online] Available: http://sindice.com/ [Accessed: 2012].
[90] SW, Semantic Web Search Engine, 2012. [Online] Available: http://www.swse.org/ [Accessed: 2012].
[91] Swoogle, Swoogle Semantic Web Search Engine, 2012. [Online] Available: http://swoogle.umbc.edu/ [Accessed: 2012].
[92] S.H. Yeganeh, O. Hassanzadeh, R.J. Miller, Linking semistructured data on the web, Interface, no. WebDB, 2011.
[93] Gloze: XML to RDF and back again, 2006.
[94] Timrdf, timrdf/csv2rdf4lod-automation—GitHub, 2011. [Online] Available: https://github.com/timrdf/csv2rdf4lod-automation [Accessed: 14-Nov-2011].
[95] F. Maali, R. Cyganiak, Google-refine, 2011. [Online] Available: http://code.google.com/p/google-refine/ [Accessed: 14-Nov-2011].
[96] Google, rdf extension for google refine, 2011. [Online] Available: http://lab.linkeddata.deri.ie/2010/grefine-rdf-extension/ [Accessed: 14-Nov-2011].
[97] S. Jupp, J. Klein, J. Schanstra, R. Stevens, Developing a kidney and urinary pathway knowledge base, J. Biomed. Semant. 2 (2) (2011) S7.
[98] Clark & Parsia, Pellet integrity constraint validator, 2011. [Online] Available: http://clarkparsia.com/pellet/icv/ [Accessed: 14-Nov-2011].
[99] D. Stoilov, Primer introduction to OWLIM – OWLIM42 – Ontotext Wiki, 2011. [Online] Available: http://owlim.ontotext.com/display/OWLIMv42/Primer+Introduction+to+OWLIM [Accessed: 02-Dec-2011].
[100] D. Allemang, J. a Hendler, Semantic Web for the Working Ontologist: Effective Modeling in RDFS and OWL [Paperback], first ed., Morgan Kaufmann, 2008, p. 352.
[101] OORP, Open ontology repository poster, 2011. [Online] Available: http://kcap09.stanford.edu/share/posterDemos/164/index.html [Accessed: 2012].
[102] HCLS RDF Guide, 2012. [Online] Available: http://www.w3.org/2001/sw/hcls/notes/hcls-rdf-guide/ [Accessed: 12-Apr-2012].