Published in final edited form as:

J Libr Metadata. 2015 ; 15(3-4): 157–176. doi:10.1080/19386389.2015.1099967.

Transforming the Medical Subject Headings into Linked Data: Creating the Authorized Version of MeSH in RDF

Barbara Bushman [Assistant Head],

Cataloging and Metadata Management Section National Library of Medicine 8600 Rockville Pike,
Bldg. 38 Bethesda, MD 20894

David Anderson [Librarian], and

National Library of Medicine 8600 Rockville Pike, Bldg. 38A Bethesda, MD 20894
andersondm2@mail.nih.gov

Gang Fu, Ph.D. [Research Fellow]

National Center for Biotechnology Information National Library of Medicine 8600 Rockville Pike,
Bldg. 38A Bethesda, MD 20894 Fug2@ncbi.nlm.nih.gov

Abstract

In February 2014 the National Library of Medicine formed the Linked Data Infrastructure Working Group to investigate the potential for publishing linked data, determine best practices for publishing linked data, and prioritize linked data projects, beginning with transforming the Medical Subject Headings as a linked data pilot. This article will review the pilot project to convert the Medical Subject Headings from XML to RDF. It will discuss the collaborative process, the technical and organizational issues tackled, and the future of linked data at the library.

Keywords

National Library of Medicine's Medical Subject Headings; linked data; resource description framework

Introduction

The concept of the semantic web was first mentioned in a 2001 article (Berners-Lee, Hendler, & Lassila, The Semantic Web, 2001) and since that time the semantic web has grown to include a growing number of linked data services (Williams, 2011). The library community has embraced linked data and the semantic web as a means to better expose data about their collections and encourage the reuse of library data on the web. As a response to the increasing desire to participate in world of linked data, in 2011, the World Wide Web Consortium (W3C) released a set of recommendations on linked data for libraries (Baker, et al., 2011). As more and more national libraries, including the Library of Congress, the National Agricultural Library, the British Library and others, began releasing various datasets as linked data, it became evident that the National Library of Medicine (NLM®)

should also participate in this arena. When the Library of Congress introduced the Bibliographic Framework Initiative (BIBFRAME)¹ as a linked data framework to replace the MARC communication format for bibliographic data, NLM was invited to serve as an Early Experimenter (Library of Congress). While bibliographic data such the Library of Congress Subject Headings were available as linked data for use in BIBFRAME experimentation, NLM's Medical Subject Headings (MeSH®) were not yet publicly available for use in a linked data environment.

In 2013 NLM conducted an environmental scan of linked data at peer institutions and a survey of NLM's "datascape" in order to generate recommendations on how NLM could participate in the semantic web. The internal report detailed how other national libraries were publishing their vocabularies and resource description metadata as linked data. The report also illustrated how other organizations have mapped NLM datasets to the W3C web standard for linked data, Resource Description Framework (RDF), and published them as linked data. The internal report asked many questions including: How can NLM best disseminate authoritative terminologies across the web? How can NLM describe resources for a variety of audiences? How will NLM manage internal data in the future? Answers to these questions and others raised in the report could be found in experimenting with publishing NLM data as linked data.

Also in late 2013/early 2014, researchers at NLM's Lister Hill National Center for Biomedical Communication presented a paper to the 2014 American Medical Informatics Association (AMIA) Meeting analyzing the six different versions of MeSH in RDF currently available (Winnenburg & Bodenreider, 2014). The purpose of this paper was to examine the existing versions in order to identify the desired features for an authoritative version of MeSH RDF. Those features included completeness, usability, linkability, currency, availability, and transparency. As part of the research for this paper an internal prototype for generating MeSH RDF was created. Also around this time, the National Center for Biotechnology Information at NLM released a beta version of PubChem RDF,² and staff from the Technical Services Division were participating in the development of BIBFRAME. Each of these efforts were being conducted independently and without communication among the areas of the library regarding standards and best practices.

As a result of these activities, and as an attempt to answer the questions raised in the internal report, in February 2014, NLM formed the Linked Data Infrastructure Working Group. The Working Group brought together representatives from existing NLM linked data projects as well as other stakeholders to coordinate efforts across NLM. The business case for the establishment of the group was to help NLM provide a more consistent user experience, coordinate data publishing activities across NLM, ensure scalability and consistency of linked data publication systems, and facilitate better linking between NLM datasets and datasets on the web. NLM publishes datasets on the web using a variety of formats and while the publishing of NLM data has required a great deal of cooperation among areas of the library, NLM has not yet undertaken any broad efforts to publish data using RDF. While

¹BIBFRAME, <http://www.loc.gov/bibframe/>

²PubChem RDF, <https://pubchem.ncbi.nlm.nih.gov/rdf/>

other organizations have mapped NLM data to RDF and published it on the web, the permanency and currency of that data is not guaranteed. NLM's participation in the web of linked data will ensure that users can link and reuse consistent, permanent, and authoritative NLM data. The Working Group was charged to investigate the potential for publishing NLM linked data, determine best practices, and prioritize linked data projects. The Working Group chose to transform MeSH as a pilot project and developed and built an infrastructure for transforming, storing and publishing NLM linked data.

MeSH RDF Pilot Project

Publishing MeSH in RDF was selected as the pilot project because it is widely known and used in the health and medical community. There was demonstrated interest in having MeSH available as linked data based on the existence of third party versions identified in the previous studies and there had been requests for MeSH as linked data from other libraries experimenting with BIBFRAME. Probably the most important internal reason to select MeSH as the pilot was that there was an existing prototype that could be modified as needed. The goals of the pilot included providing an authoritative version of MeSH RDF, ensuring its maintenance and preservation, developing an infrastructure for publishing NLM linked data, and increasing NLM's knowledge of MeSH use cases. MeSH RDF has been produced by third parties, but there is no guarantee that those third parties are updating, maintaining, and/or preserving it. In order to be able to update, maintain, and preserve the MeSH RDF, NLM would need to develop an infrastructure for publishing linked data that is sustainable and scalable. While it is well known that MeSH is utilized for cataloging, indexing, and mapping to other vocabularies, other uses are less known and understood. It is anticipated that publishing MeSH RDF will result in learning about additional use cases and enable greater functionality across the web.

NLM released the initial beta version of MeSH RDF on November 17, 2014 to coincide with the AMIA Meeting and the presentation of the aforementioned research paper. Since the release of the beta MeSH RDF was considered to be a soft-launch, there was no NLM official publicity associated with its release. The product was a true beta version and not finalized; its existence was announced via presentations and demonstrations at the AMIA meeting, as well as presentations at subsequent meetings, word of mouth, and social media.

Methods

MeSH is NLM's controlled vocabulary thesaurus used for cataloging acquired books, journals and audiovisuals as well as for indexing articles for the MEDLINE®/PubMed® database (National Library of Medicine, 2013). MeSH is a three-tier-structured terminology. The first tier consists of Descriptors, Qualifiers, and Supplementary Concept Records (SCRs). Descriptors, also referred to as main headings, are used to indicate the subject of an indexed or cataloged item. Qualifiers, also referred to as subheadings, are used in conjunction with a Descriptor to express a particular aspect of a subject (e.g. Liver—drug effects). SCRs are used to index chemicals, drugs, and other concepts such as rare diseases. Each SCR is mapped to one or more Descriptors (National Library of Medicine, 2014). The second tier consists of Concepts, which are groups of synonymous Terms. The third tier

consists of Terms, which are grouped into Concepts. Each Descriptor/Qualifier/SCR record groups together a set of Concepts. Each Concept groups together one or more Terms. The three-tier-structure of the MeSH terminology is depicted in this example from documentation describing concept structure of MeSH (National Library of Medicine, 2014):

Cardiomegaly [Descriptor]	
Cardiomegaly	[Concept, Preferred]
Cardiomegaly	[Term, Preferred]
Enlarged Heart	[Term]
Heart Enlargement	[Term]
Cardiac Hypertrophy	[Concept, Narrower]
Cardiac Hypertrophy	[Term, Preferred]
Heart Hypertrophy	[Term]

MeSH RDF Data Model

The MeSH RDF project intends to facilitate data sharing and linking using semantic web standards and technologies. RDF is a standard data model that decomposes knowledge information into small pieces with well-defined semantics (meaning). Each piece of information is represented as a RDF statement, also called a “triple” composed of a subject, a predicate and an object. In RDF, the data in the triple statement are represented, where possible, using Uniform Resource Identifiers (URIs), persistent web identifiers that can represent resources and relationships. A collection of triples constitutes a knowledge graph which can then be queried using SPARQL, an SQL-like query language that can produce both triples and tabular data.

The Working Group produced a draft set of triple statements from NLM’s existing prototype by adapting an eXtensible Stylesheet Language Transformation (XSLT) to generate RDF data from the existing MeSH XML. The semantics (i.e., the types and relations) in the MeSH RDF model are defined in the ontology as classes and predicates. The Working Group discussed the advantages and disadvantages of utilizing existing ontologies such as SKOS to represent MeSH in RDF, but eventually decided to develop a specific MeSH RDF vocabulary to define the types and relationships expressed in the XML files. The vocabulary was manually prepared using Protégé, a widely-used ontology editor.

After examining the initial results, the Working Group found that some of the peculiarities of MeSH made for some difficult choices regarding the data model. The following questions arose: How should MeSH RDF classes be expressed? How should MeSH RDF handle Descriptor-Qualifier combinations? How would MeSH RDF express hierarchical relationships? How loyal should MeSH RDF be to the data model implicit in the MeSH XML? As the original data model was adjusted to answer the first three questions, the answer to the fourth question became clear. While the MeSH RDF model would be informed by the XML, it would become impossible for it not to diverge.

MeSH RDF Classes

MeSH RDF has several classes of things. Classes are used to define types of things in RDF. The Working Group chose “meshv:” as a prefix for classes and predicates defined in the MeSH RDF vocabulary and “mesh:” as a prefix for instances of classes. Prefixes are used in SPARQL queries and data representations as a shorthand for URIs. MeSH RDF classes and subclasses are arranged as follows:

```
owl:Thing
  meshv:Concept
  meshv:Descriptor
    meshv:CheckTag
    meshv:GeographicalDescriptor
    meshv:PublicationType
    meshv:TopicalDescriptor
  meshv:DescriptorQualifierPair
    meshv:AllowedDescriptorQualifierPair
    meshv:DisallowedDescriptorQualifierPair
  meshv:Qualifier
  meshv:SemanticType (deprecated)
  meshv:SupplementaryConceptRecord
    meshv:SCR_Chemical
    meshv:SCR_Disease
    meshv:SCR_Protocol
  meshv:Term
  meshv:TreeNumber
```

The MeSH RDF technical documentation includes lists of MeSH RDF classes and predicates accompanied by their definitions.^{3,4}

Descriptors—MeSH Descriptors (meshv:Descriptor) indicate either major topics/subjects, geographic locations, publication types, or “check tags.” These are represented by subclasses of meshv:Descriptor:

```
meshv:TopicalDescriptor
meshv:GeographicalDescriptor
meshv:PublicationType
meshv:CheckTag
```

A typical instance of a Descriptor, mesh:D015242 (Ofloxacin), is depicted in Figure 1.

³MeSH Classes, <http://hhs.github.io/meshrdf/classes.html>

⁴MeSH Predicates, <http://hhs.github.io/meshrdf/predicates.html>

A Descriptor may also have links to other Descriptors, or instances of other MeSH RDF classes, like Qualifiers, Concepts, Tree Numbers, and so on. For example, D015242 has the following relationships:

```
meshv:allowableQualifier      mesh:Q000744 ;
meshv:pharmacologicalAction   mesh:D000892 ;
meshv:concept                 mesh:M0333653 ;
meshv:treeNumber              mesh:D03.438.810.835.322.500 .
```

Qualifiers—MeSH Qualifiers (meshv:Qualifier) are used to give additional context to an associated Descriptor. A typical instance of Qualifier, mesh:Q000008 (administration & dosage), is depicted in Figure 2.

MeSH has rules governing which Qualifiers can be used with a given Descriptor as well as which Qualifiers cannot be used with a given Descriptor. In MeSH XML, each Descriptor is associated with a set of allowable Qualifiers. In MeSH RDF allowed combinations are instances of meshv:AllowedDescriptorQualifierPair (see Figure 3 for D015242Q000008), and disallowed combinations are instances of meshv:DisallowedDescriptorQualifierPair (see Figure 4 for D000005Q000293). Both types are subclasses of meshv:DescriptorQualifierPair.

Descriptor-Qualifier Pairs—The Working Group chose to avoid blank nodes where possible. In RDF, blank nodes are unclassified instances, and while they are not prohibited in RDF, the Working Group preferred to express instances of explicit classes. Therefore, the Working Group created new full-fledged instances rather than use blank nodes to represent complicated data elements. To illustrate this practice, the ‘EntryCombination’ element in MeSH XML specifies how to use an appropriate Descriptor or Descriptor/Qualifier combination to replace a prohibited Descriptor/Qualifier combination. If Descriptors and Qualifiers are not pre-coordinated, this relationship would require a blank node to relate the pre-coordinated pair to the resource it describes. Instead, URIs were created based on the combinations of Descriptors and Qualifiers. Instances of the resulting class are called Descriptor-Qualifier Pairs. For example, the Descriptor-Qualifier Pair of D015242 and Q000008 is represented as mesh:D015242Q000008. This will make it possible to link MEDLINE citations and library catalog resources directly to Descriptor-Qualifier Pairs.

Supplementary Concept Records—MeSH Supplementary Concept Records, or SCRs (meshv:SupplementaryConceptRecord) account for the large volume of chemical names, rare diseases and treatment protocols that are found in the biomedical literature but are not included in MeSH headings. It has three subclasses: meshv:SCR_Chemical, meshv:SCR_Disease, and meshv:SCR_Protocol. A typical instance of an SCR_Chemical, mesh:C025735 (Aeron), is depicted in Figure 5.

Concepts—MeSH Descriptors, Qualifiers, and SCRs may group a set of MeSH Concepts (meshv:Concept). A MeSH Concept represents a unit of meaning. Each MeSH record (meshv:Descriptor, meshv:Qualifier, or meshv:SupplementaryConceptRecord) consists of

one or more meshv:Concept, and each meshv:Concept consists of one or more synonymous Terms (meshv:Terms). Descriptors and Concepts are related via meshv:concept, except in cases where the Concept shares a label with the Descriptor. In those cases, Descriptors and Concepts are related via meshv:preferredConcept. A typical instance of a Concept, M0000001 (Calcimycin) is depicted in Figure 6.

Terms—MeSH Terms (meshv:Term) comprise the basic unit of the MeSH terminology. Terms related to a particular MeSH Concept are strictly synonymous, and they often have lexical permutations, such as ‘Abnormality, Congenital’ as opposed to ‘Congenital Abnormalities’. In MeSH RDF, Terms have a property of meshv:prefLabel and meshv:altLabel to describe this lexical difference. A couple of typical instances of Terms are depicted in Figure 7.

Tree Numbers—MeSH Descriptors are organized into 16 categories (A for anatomic terms, B for organisms, C for diseases, D for drugs and chemicals, etc.) referred to as “trees”. Each Descriptor is categorized in at least one place in the trees, and a given Descriptor may appear in multiple trees as appropriate. Therefore, a Descriptor may have different parents or children in different trees (National Library of Medicine, 2014).

Whereas MeSH XML does not explicitly express the hierarchical relationships found in the MeSH trees, MeSH RDF has made parent relationships explicit by elevating Tree Numbers to first class citizens and relating them to each other via meshv:parentTreeNumber. MeSH RDF also expresses broader relationships between Descriptors via meshv:broaderDescriptor. Because MeSH’s multiple-tree structure may suggest strange hierarchical relationships between the ancestor of a Descriptor in one tree and the descendant in another tree (i.e. ‘eyebrows’ may be erroneously inferred as the descendent of ‘sense organs’ via the Descriptor, “eye”), relationships between Descriptors are expressly non-transitive (see Figure 8 and Figure 9). Note that Qualifiers and Concepts occupy special hierarchies independent from Descriptors and Tree Numbers, and they are related via non-transitive properties specific to those classes.

Converting the MeSH XML data to RDF enables logic-based reasoning to derive and validate new statements that are not explicitly asserted in the XML. MeSH RDF’s vocabulary designates relationships between Tree Numbers as transitive properties and these transitive relationships can only be found using the meshv:parentTreeNumber relationship between Tree Numbers. For example, we can traverse MeSH trees to retrieve all of the ancestors for Descriptor D005138 using the following SPARQL query:

```
SELECT ?treeNode ?ancestorTreeNode ?ancestor ?alabel
FROM <http://id.nlm.nih.gov/mesh2014>
WHERE {
  mesh:D005138 meshv:treeNumber ?treeNode .
  ?treeNode meshv:parentTreeNumber ?ancestorTreeNode .
  ?ancestor meshv:treeNumber ?ancestorTreeNode .
  ?ancestor rdfs:label ?alabel
```



```
}
ORDER BY ?treeNode ?ancestorTreeNode
```

By running a SPARQL query with logic-based inference, all of the direct and indirect ancestors (i.e., the parents of parents (grandparents)) of Descriptor D005138 can be retrieved, otherwise, only the direct ancestors can be retrieved. Alternatively, the property path (syntax - `meshv:parentTreeNumber+`), can also be used in a SPARQL query to navigate the transitive closure without inference. Broader or parent relationships which were merely implicit in the MeSH XML have now been made explicit in the MeSH RDF.

Availability of MeSH RDF

As one of the goals of the pilot project was to develop a broader knowledge of the uses of MeSH, MeSH RDF was made available both for bulk download, and through a SPARQL endpoint, as well as a SPARQL query editor for querying and exploring the data.⁵ From NLM's FTP site anyone can download the entire set of RDF files, the original XML files used to generate the RDF, the XSLT transformations, and MeSH RDF vocabulary.⁶ The members of the Working Group believed strongly that linked data should be linked open data and therefore access should not be restricted due to a license or memorandum of understanding. The case for open access was presented to NLM's administration and permission was granted to offer all of these files on the FTP site without any restrictions. These files are in addition to the full data files of MeSH available for download in XML, ASCII, and MARC formats.

MeSH RDF on the Web

In addition to modelling MeSH RDF data, the Working Group established a web presence for MeSH RDF that includes a read-only SPARQL endpoint, a SPARQL query editor, a browseable interface, RESTful interface for URIs, documentation, and GitHub repositories. The interface is powered by a stack that includes OpenLink's open source Virtuoso server and an open source front-end Java application deployed in Tomcat, which borrows heavily from the European Bioinformatics Institute's Lodestar platform.⁷ The Java application connects to the Virtuoso RDF store through JDBC connection pooling with reasonable concurrency and scalability handling.

Uniform Resource Identifier (URI) Patterns

All MeSH RDF data is either represented as a URI reference or encoded as a literal string. NLM minted two base URIs for MeSH RDF:

- id.nlm.nih.gov/mesh/vocab# (meshv: prefix) for classes and predicates.
- id.nlm.nih.gov/mesh/ (mesh: prefix) for instances of classes.

⁵SPARQL Query Editor: <http://id.nlm.nih.gov/mesh/query/>

⁶Download MeSH RDF, <ftp://ftp.nlm.nih.gov/online/mesh/>

⁷LodeStar on GitHub, <https://github.com/EBISPOT/lodestar>

Class instance URIs are constructed by appending id.nlm.nih.gov/mesh/ to identifiers. For example:

- Descriptor D015242: <http://id.nlm.nih.gov/mesh/D015242>
- Qualifier Q000008: <http://id.nlm.nih.gov/mesh/Q000008>
- Supplementary Concept Record C025735: <http://id.nlm.nih.gov/mesh/C025735>
- Concept M0000001: <http://id.nlm.nih.gov/mesh/M0000001>
- Term T000002: <http://id.nlm.nih.gov/mesh/T000002>
- Tree number A09.371.613: <http://id.nlm.nih.gov/mesh/A09.371.613>

Classes have a base URI of <http://id.nlm.nih.gov/mesh/vocab#>. For example:

- <http://id.nlm.nih.gov/mesh/vocab#Descriptor>
- <http://id.nlm.nih.gov/mesh/vocab#Qualifier>
- <http://id.nlm.nih.gov/mesh/vocab#SupplementaryConceptRecord>

Predicates also have a base URI of <http://id.nlm.nih.gov/mesh/vocab#>. For example:

- <http://id.nlm.nih.gov/mesh/vocab#allowableQualifier>
- <http://id.nlm.nih.gov/mesh/vocab#annotation>
- <http://id.nlm.nih.gov/mesh/vocab#registryNumber>

SPARQL Endpoint

MeSH RDF's SPARQL endpoint⁸ makes it possible to query MeSH RDF directly using SPARQL query language. This SQL-like RDF query language serves as an important component of semantic web standards and technologies, which can query both local and remote RDF databases through data access protocols. NLM has provided a few SPARQL queries that can be selected and run immediately using the SPARQL query editor and additional SPARQL queries live in the MeSH RDF documentation.⁹ Each SPARQL query has its own shareable URL. SPARQL queries can output HTML, XML, JSON, CSV, TSV, RDF/XML, RDF/N3, JSON-LD or Turtle depending on the nature of the query. The number of results is currently constrained to 1,000 rows of data or 10,000 triples. The SPARQL query editor produces tables with links to a browseable interface.

Browsing MeSH RDF

Each URI dereferences to an HTML page with links to related data. For example, pointing a browser to <http://id.nlm.nih.gov/mesh/D000900> will produce a page that expresses all of the relationships from and to the Descriptor, "Anti-Bacterial Agents." From each URI page, it is possible to browse to any related URIs.

⁸SPARQL Endpoint Base URL, <http://id.nlm.nih.gov/mesh/sparql>

⁹MeSH RDF Sample Queries: <http://hhs.github.io/meshrdf/sample-queries.html>

RESTful Interface

Class instance URIs can be dereferenced through HTTP 303 redirection to a document in one of many formats. The allowed RDF formats listed in Table 1 can be specified either in the HTTP accept header or in the URL extension. The following URIs will return HTML, RDF/XML, Turtle, JSON-LD and N-Triples data, respectively:

- <http://id.nlm.nih.gov/mesh/D015242.html>
- <http://id.nlm.nih.gov/mesh/D015242.rdf>
- <http://id.nlm.nih.gov/mesh/D015242.ttl>
- <http://id.nlm.nih.gov/mesh/D015242.json>
- <http://id.nlm.nih.gov/mesh/D015242.n3>

Documentation

The MeSH RDF documentation developed for the pilot project includes detailed data model information, diagrams, sample data, sample SPARQL queries, cheat sheets and helpful links.¹⁰ The Working Group built the documentation website using GitHub Pages and Jekyll. Note that MeSH RDF remains in beta as of the writing of this paper, and classes and predicates may evolve.

Outcomes and Lessons Learned

Collaborative Efforts

The members of The Working Group are quite versed in XSLT and RDF, but there were cases where the group could not reach consensus as to the best approach. The Working Group decided that it would be beneficial to procure RDF/linked data/semantic web consulting services to assist with determining the best approach in these cases. Consultants from Zepheira were engaged to assist with this effort. To help develop possible use cases for MeSH RDF, the Working Group identified a set of institutions/organizations that agreed to assist NLM with testing the beta version. The beta partners were asked to document how they plan to utilize the beta MeSH RDF and to share their needs, use cases, and experiences either through email, a GitHub site which was established for this purpose, and/or conference calls. The GitHub site is publicly available and comments have been received from other interested parties not previously identified in the group of beta partners.¹¹

User Feedback

NLM is tracking downloads of MeSH RDF data and usage of the MeSH RDF interface. NLM utilized WebTrends software to conduct analytics on the MeSH RDF webpages and discovered a number of organizations utilizing the site that were not included in the known set of beta partners. These additional users included pharmaceutical companies, academic institutions, foreign biomedical institutions and others. WebTrends also provides data on which specific pages are being accessed most often and which SPARQL queries are being

¹⁰MeSH RDF Documentation, <http://hhs.github.io/meshrdf/>

¹¹MeSH RDF GitHub Issues: <https://github.com/HHS/meshrdf/issues>

run. More in-depth analysis would still need to be done to determine the impact and value of this data. WebTrends also showed that a significant number of users were accessing the site via Twitter. As a result, the Working Group decided to develop a set of weekly tweets to promote the beta MeSH RDF and solicit additional feedback.

Potential use cases of MeSH RDF identified by the beta partners included: investigating compatibility of other SPARQL endpoints, integrating MeSH RDF with other RDF datasets, replacing literal strings of MeSH with URIs, replacing a third-party version of MeSH RDF with the NLM version, mapping to other ontologies, and enriching other existing linked data services. The beta partners requested that NLM consider non-English representations of MeSH linked data to enable mapping and linking to other multilingual vocabularies. Some users suggested that NLM map hierarchical relationships to SKOS, a standard vocabulary for thesauri, however, MeSH's peculiar 3-level structure makes it difficult to map its hierarchical relationships to SKOS (Winnenburg & Bodenreider, 2014). While the Working Group preferred to express MeSH RDF predicates using NLM's own namespaces, this does not preclude future mappings to shared vocabularies. Use of NLM's namespaces also does not preclude external users from mapping MeSH RDF classes and predicates to other vocabularies as use cases require.

The beta partners also expressed a need for “tree top” Descriptors – a short list of Descriptors that sit at the top of MeSH trees and act as entry points into the trees. Tree top Descriptors would improve the navigability of browseable MeSH trees on the web. After review, it was determined that this enhancement was beyond the scope of the Working Group and the request has been forwarded to the appropriate section at NLM which will consider the addition of tree top Descriptors to the content of MeSH.

Next Steps

The initial beta version of MeSH RDF was static and based on the 2014 version of MeSH XML. In the subsequent June 2015 release, the data was updated to the 2015 version of MeSH XML. With this release NLM began to incorporate daily updates to the MeSH Supplementary Concept Records. As a result of feedback received from the beta partners, language tags were added for all strings with literal values. In addition, the MeSH descriptor Central Nervous System Diseases was added in seventeen languages as a proof-of-concept to demonstrate multilingual capability. Full translations of MeSH cannot be added at this time due to licensing agreements with the foreign translators. The June 2015 beta release was accompanied by a full public news announcement, tweets, and updated documentation. The landing page is directly linked from the NLM home page as well as other relevant areas of the NLM website.

It is anticipated that MeSH RDF will be transferred from beta to production in the fall of 2015 to update it to the 2016 version of MeSH and synch the MeSH RDF production cycle with the MeSH XML production cycle. Prior to moving to production, issues surrounding best methods for archiving the RDF data, technical documentation, and the vocabulary will need to be resolved.

In addition to converting it from beta to a production version, NLM will experiment with incorporating MeSH RDF into other linked data activities. The team responsible for creating and maintaining PubChem RDF will be investigating how they can facilitate data sharing and integration between outside parties, MeSH and PubMed. NLM's ongoing experimentation with BIBFRAME will include representing subject headings in bibliographic records with the URIs available from MeSH RDF. Possibilities exist for other NLM services and applications utilizing MeSH to investigate and experiment with how they might benefit from the publication of MeSH as linked data.

Through the process of transforming the XML elements into RDF many questions were raised about the purpose of some of the elements and whether or not they were needed in RDF. As stated earlier, for the pilot project, a decision was made to keep the RDF as in-synch with the XML as possible, and subsequently perform a review of the RDF elements. Any peculiarities in the MeSH XML discovered during this review, such as elements deemed to be internal use only, were forwarded to those at NLM responsible for developing and maintaining the MeSH vocabulary for a possible review and revision of the MeSH XML. Some elements have been marked for deprecation as a result of this process.

Conclusion

In conclusion, the NLM Linked Data Infrastructure Working Group was convened to bring those at NLM interested and involved in publishing linked data together to establish consistent policies and procedures. One of the biggest benefits of the MeSH pilot project was bringing together representatives of many areas of NLM to work on a project supported at an institutional level. This collaborative effort led to a great deal of knowledge sharing across the library with regards to linked data, data modeling, application development, version control, and workflow management. The discussions that led to MeSH RDF began to align thinking about data publishing around a common framework. Despite a growing consensus around linked data standards and best practices, there were a number of cases where the path forward was less than clear. Data modeling choices were not always obvious and required a great deal of debate. That debate continues, and the reverberations of that debate may have an impact on the future of MeSH as well as other NLM data, interfaces, and application program interfaces (APIs).

Releasing MeSH in RDF is a major accomplishment and will benefit many in the healthcare and library communities, however, the publication of MeSH RDF is not truly linked data. The promise of linked data will only be realized once a critical mass of NLM resources is available in RDF and external services link to NLM RDF data. Producing linked data will provide an opportunity for NLM to re-examine the workflows around the development and maintenance of user interfaces and APIs. Publishing MeSH as linked data paves the way for NLM to consider producing more data in RDF, including MEDLINE/PubMed, bibliographic data, other terminologies and more. The foundation provided by MeSH RDF and its underlying infrastructure provides the beginnings of a linked data ecosystem from which data can be published more expressively. Ultimately, users of NLM data should be able to retrieve any subset of NLM data at almost any level of granularity and RDF linked data provides a framework for achieving that goal.

Acknowledgements

The authors would like express their thanks to the members of the NLM Infrastructure Working Group and the consultants from Zepheira for all of their efforts towards producing MeSH RDF and their contributions to this article. This work was supported by the Intramural Research Program of the National Institutes of Health, National Library of Medicine.

References

- Baker, T.; Bermes, E.; Coyle, K.; Dunsire, G.; Isaac, A.; Murray, P., et al. [Retrieved August 14, 2015] Library Linked Data Incubator Group Final Report. Oct 25. 2011 from W3C: <http://www.w3.org/2005/Incubator/ld/XGR-ld-20111025/>
- Beckett, D.; Berners-Lee, T. [Retrieved May 1, 2015] Turtle - Terse RDF Triple Language. Mar 28. 2011 from W3C: <http://www.w3.org/TeamSubmission/turtle/>
- Berners-Lee, T.; Connolly, D. [Retrieved May 1, 2015] Noation3 (N3): A readable RDF Syntax. Mar 28. 2011 from W3C: <http://www.w3.org/TeamSubmission/n3/>
- Berners-Lee T, Hendler J, Lassila O. The Semantic Web. Scientific American. 2001:34–43.
- Jupp S, Malone J, Bolleman J, Brandizi M, Davies M, Garcia L, et al. The EBI RDF platform: linked open data for the life sciences. Bioinformatics. 2014:1338–1339. [PubMed: 24413672]
- Kellogg, G.; Lanthaler, G. [Retrieved May 1, 2015] RDF 1.1 Test Cases. Feb 25. 2014 from W3C: <http://www.w3.org/TR/2014/NOTE-rdf11-testcases-20140225/>
- Library of Congress. [Retrieved April 28, 2015] Bibliographic Framework Initiative. (n.d.)from <http://www.loc.gov/bibframe/>
- National Library of Medicine. [Retrieved April 28, 2015] Fact Sheet Medical Subject Headings (MeSH). Dec 9. 2013 from <http://www.nlm.nih.gov/pubs/factsheets/mesh.html>
- National Library of Medicine. [Retrieved April 28, 2015] Concept Structure in MeSH. Oct 27. 2014 from http://www.nlm.nih.gov/mesh/concept_structure.html
- National Library of Medicine. [Retrieved April 28, 2015] Introduction to MeSH in XML format. Aug 19. 2014 from <http://www.nlm.nih.gov/mesh/xmlmesh.html>
- National Library of Medicine. [Retrieved April 28, 2015] MeSH Record Types. Aug 6. 2014 from http://www.nlm.nih.gov/mesh/intro_record_types.html
- National Library of Medicine. [Retrieved April 28, 2015] MeSH Tree Structures. Aug 6. 2014 from http://www.nlm.nih.gov/mesh/intro_trees.html
- Sauermann, L.; Cyganiak, R. [Retrieved April 28, 2015] Cool URIs for the Semantic Web. Dec 3. 2008 from W3C: <http://www.w3.org/TR/cooluris/#solutions>
- Sporny, M.; Longley, D.; Kellogg, GL.; Lindstrom, N. [Retrieved May 1, 2015] JSON-LD 1.0. Jan 16. 2014 from W3C: <http://www.w3.org/TR/json-ld/>
- Williams, A. [Retrieved August 14, 2015] The Growth of Linked Data. Jan 18. 2011 from readwrite: <http://readwrite.com/2011/01/18/the-concept-of-linked-data>
- Winnenburg, R.; Bodenreider, O. Desiderata for an authoritative Representation of MeSH in RDF. AMIA Annual Symposium Proceedings; Bethesda: American Medical Informatics Association; 2014. p. 1215-1227.

Figure 1.
RDF graph diagram showing a typical Topical Descriptor D015242 and its literals

Figure 2.
RDF graph diagram showing a typical Qualifier Q000008 and its literals

Figure 3.
RDF graph diagram showing a typical allowed Descriptor-Qualifier Pair D015242Q000008

Figure 4.
RDF graph diagram showing a typical disallowed DQ pair D000005Q000293

Figure 5.
RDF graph diagram showing a typical SCR C025735

Figure 6.
RDF graph diagram showing a typical Concept M0000001

Figure 7.
RDF graph diagram showing a couple of typical Term instances

Figure 8.

Diagram illustrating one Descriptor organized as the intersection of two overlapping MeSH trees, where 'Eye' belongs to both the 'Body Regions' tree (A01) and the 'Sense Organs' tree (A09)

Figure 9.

Diagram demonstrating how meshv:Descriptor are related to meshv:TreeNumber, and what semantic relations are between them

Figure 10.
Legend

Table 1

Allowed MeSH RDF Formats

RDF Format	HTTP Accept Header	URI extension
HTML	text/html	html htm
RDF/XML	application/rdf+xml	rdf xml
TURTLE	application/x-turtle	turtle ttl
JSON-LD	application/rdf+json	json json-ld
N3	application/rdf+n3	n3
N-TRIPLES	text/plain	N.A.