Final Report & White Paper
IMLS Grant Number LG-46-11-0082-11
Sparks! Ignition Grant for Libraries and Museums

Amy J. Hatfield, Brian E. Dixon, Elaine N. Skopelja, Aaron Springer, Rose Jones

1. Administrative Information

- Ruth Lilly Medical Library, Indiana University School of Medicine, Indianapolis, IN
- ***Transforming Special Collection Digital Content Silos into Digital Knowledge using Semantic Web Technologies***
- Award Amount and Total Project Cost
    - Award $25,000
    - Total Project Cost $24,610
- Grant Time Period August 1, 2011 – July 31, 2012
- Project Director Elaine Skopelja MALS, AHIP
- Project Partner Center for Biomedical Informatics, Regenstrief Institute, Indianapolis, IN

2. Project Summary

Historic digital collections are no longer a novelty. Libraries, museums, and archives have been digitizing special collection materials for years. Such special collections are created using various standards and are housed in many different systems and repositories. Initially, the impact of such individualization upon users was difficulty in searching and aggregating content across disparate systems that were not interoperable. Technologies such as OpenURL and link resolvers were implemented to address the issue of linking content from disparate systems, and federated search engines were layered on top to allow users to locate and access content from the included systems. However, the linking technologies have limitations and are not as user friendly as hoped. Librarians have discovered that end users have difficulty navigating search result sets produced by aggregated systems, and the data used to identify accessible resources are not always kept up-to-date leading to dead ends (McCracken, Arthur, 2009; O'Neill, 2009; Sugita, Horikoshi, Suzuki, Kataoka, Hellman, Suzuki, 2007; Turner, 2004). Additionally, while the linking technologies provide access to content from many external resources, the retrieved content set does not identify relationships among the returned content. As a result, users are unable to synthesize related materials within the content retrieved.

On a smaller scale, even a single specialized digital collection's content can be housed in more than one content management system or database due to the special needs and handling of different content types within the collection. Some of a collection's content is optimally stored and retrieved through a repository, such as CONTENTdm or DSpace, while another type of the collection's content is better suited for 3-D display using specialized software. In order to access the complete set of the collection's available content, end users are forced to search multiple repositories or an aggregation of the content silos using the linking technologies, with the same limitations as previously discussed. By having to search multiple content management systems for a single collection, users are unable to synthesize related content from the various content silos. Just as in the case of aggregated content from

multiple collections and holdings there are missed opportunities for information discovery, content analysis, data pattern identification, and image analysis in a single collection.

The Ruth Lilly Medical Library's Digital Initiatives Group created an Indiana Public Health Digital Library collection. The Group has digitized about one-third of a historic *Indiana State Board of Health Monthly Bulletin* series, which was published for over 90 years beginning in 1899. The *Bulletin* was sent to all health officers and deputies in Indiana, as well as to individual subscribers and was used to disseminate public health directives, disease statistics, health legislation and general information to Indiana health officials. Each issue of the *Bulletin* included articles (items) ranging in size from 1/8 of a page to several pages. Also included were images and statistical tables including infectious disease outbreaks, morbidity (illnesses and injuries), and mortality (death rates and causes). The collection provides a historic portrayal of Indiana public health issues, medical history and vital statistics starting in the early 20th century. The collection includes digitized, full text (PDF) issues of the *Bulletin* tagged with Dublin Core metadata; historic photos, drawings and images (JPG) tagged with the Library of Congress' *Thesaurus for Graphic Materials (TGM)* (http://www.loc.gov/rr/print/tgm1/) metadata; and a vital statistics database with data culled from *Bulletin* tables containing vital statistics data and state disease reports. The *Bulletins*, and the other related content, are accessible through an Indiana University campus DSpace repository – IUPUIScholarWorks (https://scholarworks.iupui.edu/handle/1805/1640). However, the collection's content is currently decentralized. Each content component resides in its own digital space and each has to be searched individually.

To address the digital silo and aggregation limitation issues, the project team utilized Semantic Web technologies and methodologies to seamlessly integrate all components of the collection's decentralized content silos. We attempted to present the integrated "digital knowledge" with embedded relationships via linked data through an interactive interface for visual exploration. Instead of a dry list of documents, arranged in a table, the interface displays the returned *Bulletins*, articles and images in a scatter graph, visually representing how closely related each document is to the most relevant document. The visualization of the content set allows users to synthesize relationships among the content.

3. Process

Methodology for Article Preparation and Metadata Assignment

The prototype database incorporated a three-year run (36 issues) dated 1917-1919 of the *Bulletins*. The technology coordinator for the project made the decision to assign article-level metadata within the *Bulletins* in order to achieve granularity within the test collection. The same was done for the images in the historic image collection. The goal in achieving this level of granularity was to demonstrate accurate relationship-linking among the *Bulletins,* articles and images.

This process proved to be a challenge. The project team tested three Adobe Acrobat Pro features in order to delineate and tag the articles:

1. First feature was the Bookmark feature. Descriptive metadata was assigned as the Bookmark text. This proved problematic in that the Bookmark text is not searchable.
2. The next feature was Named Destinations. Named Destinations text is searchable. This, too, proved problematic. When a user searched the *Bulletins*, the *Bulletin* would not open

to the article tagged with the Named Destination. Instead, while the correct *Bulletins* were returned, a user would have to search the full text of the entire *Bulletin* issue in order to find the correct article.

3. The final feature tested was Article Box. While the team could delineate the articles in the *Bulletin*, there was no way to add searchable metadata to each article.

Ultimately, the decision was made to manually extract the articles from the digitized images (TIF) of the *Bulletins*. Using Photoshop the articles were cropped and saved as PDF documents. Descriptive metadata was then assigned to each article document.

The images were already individual files with assigned metadata. No processing was applied to the image files.

Methodology for Resource Description Framework (RDF) and Triples

To semantically link *Bulletins* with articles and images, we employed the Resource Description Framework (RDF) http://www.w3.org/TR/rdf-primer/. RDF is a model and syntax for representing distributed data across the Semantic Web [Allemang and Hendler, 2011]. RDF has been used to link data and information in a variety of industries, including public health and library science [Holford et al., 2006; Kunapareddy et al., 2006; Liu, 2004]. The RDF syntax uses triples to represent specific metadata about a resource in a collection: 1) the subject, which identifies the resource in the collection; 2) the predicate, which denotes the relationship between the subject and the object; and 3) the object, which identifies the resource linked to the subject in a relationship defined by the predicate.

To implement RDF, the semantic web consultant first defined a schema for the prototype project. The developed RDF schema (SPARKS) extends an existing schema, a common practice in the Semantic Web. The SPARKS schema extends the one developed and maintained by the Dublin Core® Metadata Initiative (http://dublincore.org). *Bulletins*, articles, and images are described using many Dublin Core® attributes, including title, contributor, type, and source. Additional RDF properties were defined by SPARKS to describe the relationships between *Bulletins*, articles, and images. The semantic web consultant further defined unique classes that will enable future linkages to mortality data stored in the vital statistics database. However, those linkages were not implemented in this project due to time and resource constraints. The final schema appears in **[Appendix A]**.

In parallel with development of the schema, existing and newly assigned metadata was normalized from the various sources: *Bulletins*, articles and images. Standard attributes such as title and source were abstracted for each bulletin, article, and image. Subject headings were also abstracted, and the values were normalized using the Library of Congress Subject Headings (LCSH). Files containing the abstracted metadata were then processed using a Perl computing script to convert the metadata into RDF triples using the SPARKS schema. The article metadata schema can be viewed in **[Appendix B]**.The scripting algorithm created the links between *Bulletins*, articles, and images. The normalized metadata were then loaded into a relational database to enable execution of queries by the web application. Using the SPARKS search engine, researchers are presented with a visual and percentage-of-relevancy display of each document to the original search term or to the current focused document.

Methodology for Database and Application Development

Standard search engines provide a linear list of results, starting with the most relevant and descending to the least relevant. Although the prototype database includes such a listing, clicking on a document in the result list shows a scatter graph of documents based on their relevancy to the focused document. From a programming standpoint, creating RDF Triples is an interesting problem. First, storage for the Triples had to be addressed. The current standard used to store RDF triples is a Triples store platform such as 4store (http://4store.org/). 4store is a database storage and query engine that holds RDF data. However, due to restricted access to resources, instead of using the preferred 4store platform the programmer emulated the platform using standard MySQL database functionality.

There are four tables and a view. Three tables store the object, subject and predicate values from the Triple, as well as unique ID numbers for each. The fourth table shows individual Triples, stored as three unique ID numbers. The view recombines the values into text that is used by the script to find matching values.
As RDF Triples are gaining in popularity, it makes sense to design SPARKS to be able to move into the RDF Triple world. The code is designed to be adaptable to changing the database source.

4. Project Results

There were several issues encountered and lessons learned.

In preparing the articles and assigning the metadata, several issues hindered the processing in a timely fashion. The issues included:

1. The testing and research involved in delineating individual articles within the *Bulletins* took much longer than estimated and was labor-intensive.
2. The assignment of article-level metadata was additionally time- and labor-intensive, especially due to the archaic terminology used in the early 20th century and the wide range of public health topics, e.g. infectious disease outbreaks, food safety, school hygiene, patent medicines, etc. Multiple metadata schemes were utilized, including MESH (Medical Subject Headings) http://www.nlm.nih.gov/mesh/MBrowser.html, PHIN (Public Health Information Network) http://www.cdc.gov/phin/, Library of Congress Subject Headings (LCSH) http://authorities.loc.gov/, and an archaic medical terminology database—Rudy's List of Archaic Medical Terms (http://www.antiquusmorbus.com/). A spreadsheet was compiled that included SEE and SEE ALSO references (e.g. consumption = tuberculosis), which basically documents the relationship of one term or concept to another. A relational database with this unique information will be uploaded and publicly available in the near future **[Attachment A]**.
3. Public health morbidity and mortality data from the vital statistics database does not translate well into RDF triples.

As the grant period progressed throughout the year the database and application development was postponed as the articles and metadata were necessary for their

development. In order to expedite the preparation of articles and metadata assignment for the database and application, graduate student volunteers from the Indiana University School of Library & Information Science were recruited to aid in the process. Students were eager to work on a digitization project and it was not a problem to attract volunteers.

Another issue encountered was the inability to utilize University resources. The project programmer needed a web server, database server and Triples store software, 4store in order to develop the prototype database and application. Due to University IT security policies the programmer could not get the administrative permissions to install necessary software and apply server configurations. While part of the team worked on the processing of articles and metadata, the programmer was attempting to negotiate access to the needed resources. This further impeded the progress of the project.

Unfortunately our team was unable to successfully model the morbidity and mortality data from the Indiana Vital Statistics specialized collection (http://inpub.medicine.iu.edu/database/index.html). Existing RDF models, including the Ontological Web Language (OWL), do not enable translation and aggregation of health data easily. Our project team was unable to successfully translate the morbidity and mortality data, with its inherent temporal and co-variate relationships, into a RDF model consistent with the monthly *Bulletins*, articles, and images from the other specialized collections. A simplistic model that relates individual data points (e.g., number of deaths) to a disease concept (e.g., tuberculosis) is achievable. However, modeling statistics' temporal (year, month), regional (state, county), and relational (co-occurrence with other disease) characteristics was challenging. Furthermore, RDF queries that would efficiently aggregate and combine statistical data were challenging to define and implement. Additional time and resources would be required to fully model and translate the complex flat files in the vital statistics database to a robust RDF model. For now, large public health data sets may be best suited to traditional, Bacchus-Normal Form (BNF) relational databases.

Due to the delay caused by article/metadata processing and resources negotiation, the actual time left in the grant period was limited. Once prototype development began, it was a rushed venture. While we were on the cusp of achieving true relational associations among the *Bulletins,* articles and images, the grant period was coming to a close. However, an infantile prototype was developed and can be explored at http://database.sparkthought.net/ **[Attachment B].** It was very close to what we hoped to achieve.

Recommendations

Although the rapid prototype project is still in its infancy, the project team would recommend other institutions embrace semantic web technologies and explore their application in digital environments – however that may be defined.

- Institutions should start with small, easily defined collections. The experience is just as useful, but results will be both easier to see and modify and then revise as needed.

- As mentioned previously, statistical datasets are a separate problem and are better suited to Bacchus-Normal Form (BNF) relational databases.
- Locating existing metadata schema to ensure that the project is not “reinventing the wheel”. However, blending different metadata schemes is more time consuming than it looks.
- Ensure that server resources are available and accessible, with the proper administrative permissions and support in place as this was a stall-issue for this project.
- Be sure to have application experts on the team – whatever the software being used for content manipulation may be – in order to avoid long delays in research and problem solving.

Next Steps

Further work will include:

1. Optimizing accurate relational associations among the content types: *Bulletins*, articles and images.
2. Optimizing the user interface and result set presentation.
3. Prototype evaluation will be implemented and documented.
4. Collaboration with other metadata and ontological researchers to model and describe morbidity and mortality data from the vital statistics database using semantic structures.

In the future, the team plans to seek funding to further develop and refine the prototype. The concept is achievable; now it needs to be refined, expanded and proven.

Resources

Indiana Public Health Historic Collections
https://scholarworks.iupui.edu/handle/1805/1640

Indiana Public Health Historic Collections Search Page Infantile Prototype
http://database.sparkthought.net/)

References

Allemang D, Hendler J. *Semantic Web for the Working Ontologist: Effective Modeling in RDFS and OWL*. Waltham, MA: Morgan Kauffman Publishers, 2011.

Holford ME, Rajeevan H, Zhao H, Kidd KK, Cheung KH. Semantic Web-based Integration of Cancer Pathways and Allele Frequency Data. *Cancer Information* 15 (8):19-30, 2009.

Kunapareddy N, Mirhaji P, Richards D, Casscells SW. Information Integration from Heterogeneous Data Sources: a Semantic Web Approach. A*MIA Annual Symposium Proceedings* 2006:992.

Liu, Yan Quan. Best practices, Standards and Techniques for Digitizing Library Materials: A Snapshot of Library Digitization Practices in the USA. *Online Information Review* 28.5 (2004):338-345.

McCracken P, Arthur MA. KBART: Best Practices in Knowledgebase Data Transfer. *Serials Librarian* 56(10-4):230-235, 2009.

O'Neill L. Scaffolding OpenURL results: A Call for Embedded Assistance*. Internet Reference Services Quarterly* 14(1/2):13-35, 2009.

Sugita S, Horikoshi K, Suzuki M. Linking Service to Open Access Repositories. *D-Lib Magazine* 13(3/4): 2007.

Turner S. Resource Integration in the Library: Linked Resolvers and Federated Searching. *Mississippi Libraries* 68(3):63-66, 2004.

APPENDIX A: SPARKS Schema

```
<?xml version="1.0" encoding="UTF-8"?>
<rdf:RDF xml:lang="en"
   xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
   xmlns:dcterms="http://purl.org/dc/terms/"
   xmlns:dc="http://purl.org/dc/elements/1.1/"
   xmlns:sparks="http://aaronspringer.com/sparks"
   xmlns:base="http://aaronspringer.com/sparks">
  <sparks:Bulletin rdf:about="1234567890.htm">
      <!--Bulletin info-->
      <dc:title></dc:title>
      <dc:creator></dc:creator>
      <dc:subject></dc:subject>
      <dc:description></dc:description>
      <dc:contributor></dc:contributor>
      <dc:date></dc:date>
      <dc:type></dc:type>
      <dc:format></dc:format>
      <dc:identifier></dc:identifier>
      <dc:source></dc:source>
      <dc:language></dc:language>
      <dc:relation></dc:relation>
      <dc:coverage></dc:coverage>
      <dc:rights></dc:rights>
   </sparks:Bulletin>
   <sparks:Article rdf:about="URL to Article">
      <dc:title></dc:title>
      <dc:creator></dc:creator>
      <dc:subject></dc:subject>
      <dc:description></dc:description>
      <dc:contributor></dc:contributor>
      <dc:date></dc:date>
      <dc:type></dc:type>
      <dc:format></dc:format>
      <dc:identifier></dc:identifier>
      <dc:source></dc:source>
      <dc:language></dc:language>
      <dc:relation></dc:relation>
```

```
        <dc:coverage></dc:coverage>
        <dc:rights></dc:rights>
    </sparks:Article>
    <sparks:Image rdf:about="URI">
        <dc:title></dc:title>
        <dc:creator></dc:creator>
        <dc:subject></dc:subject>
        <dc:description></dc:description>
        <dc:contributor></dc:contributor>
        <dc:date></dc:date>
        <dc:type></dc:type>
        <dc:format></dc:format>
        <dc:identifier></dc:identifier>
        <dc:source></dc:source>
        <dc:language></dc:language>
        <dc:relation></dc:relation>
        <dc:coverage></dc:coverage>
        <dc:rights></dc:rights>
    </sparks:Image>
    <sparks:mortalityReport>
        <sparks:month></sparks:month>
        <sparks:year></sparks:year>
        <sparks:region></sparks:region>
        <sparks:population></sparks:population>
        <sparks:disease></sparks:disease>
        <sparks:numDeaths></sparks:numDeaths>
    </sparks:mortalityReport>
</rdf:RDF>
```


Appendix B: Article Metadata

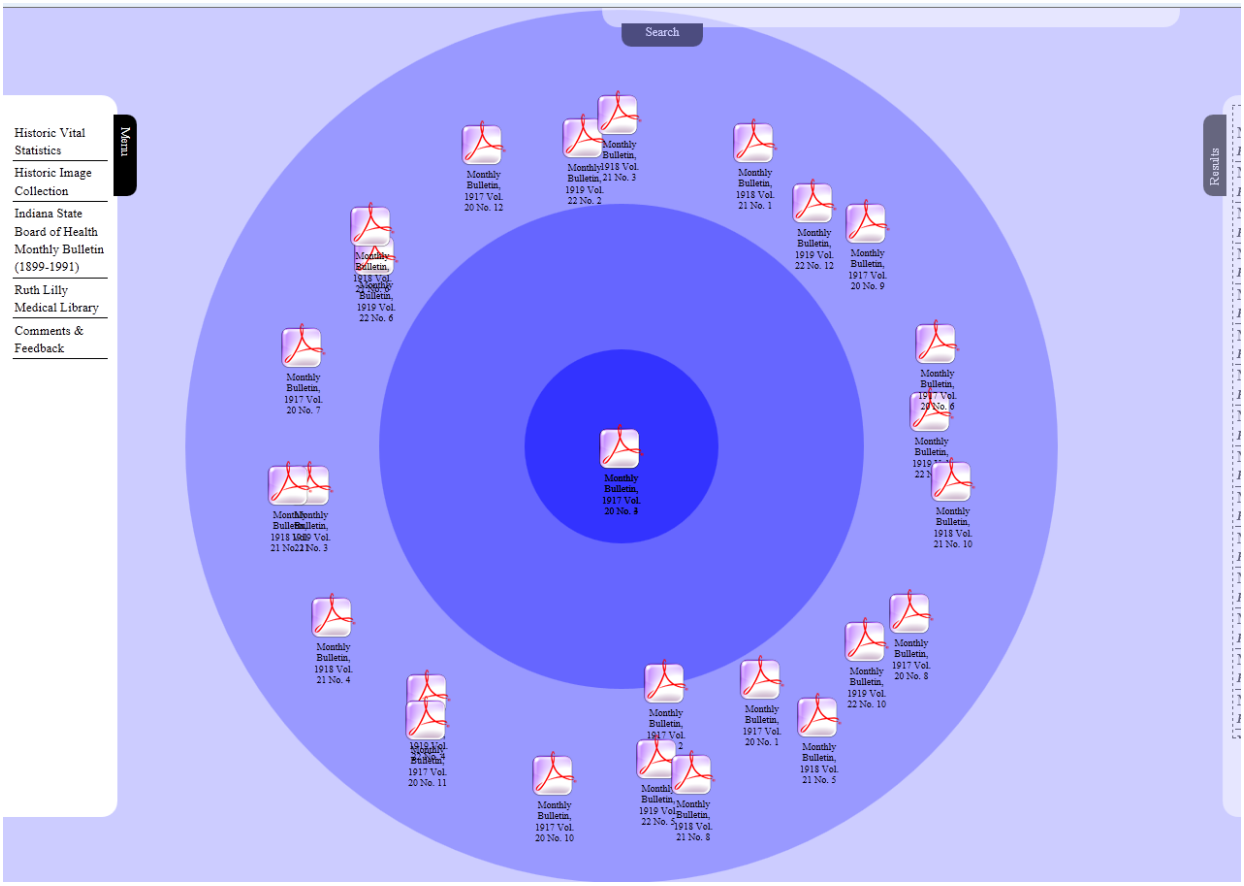

```
<?xml version="1.0" encoding="UTF-8"?>
<rdf:RDF xml:lang="en"
    xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
    xmlns:dcterms="http://purl.org/dc/terms/"
    xmlns:dc="http://purl.org/dc/elements/1.1/"
    xmlns:sparks="http://aaronspringer.com/sparks"
    xmlns:base="http://aaronspringer.com/sparks">
    <sparks:Image rdf:about="http://hdl.handle.net/1805/2584">
        <!--Article info-->
        <dc:title>Why Haven't You Told Your Child?</dc:title>
        <dc:subject>Sex</dc:subject>
            <!--List additional subjects here-->
        <dc:contributor>Bulletin of the Texas State Board of Health</dc:contributor>
        <dcterms:issued>1913</dcterms:issued>
        <dc:type>Text</dc:type>
        <dc:format>pdf</dc:format>
        <!--Below is URL to the Bulletin. Is there a better way to link these together using RDF?-->
        <dcterms:isPartOf>https://scholarworks.iupui.edu/bitstream/handle/1805/2584/im-iumed-iph-1913-
v16n6.pdf</dcterms:isPartOf>
        <dcterms:bibliographicCitation>Indiana State Board of Health. "Why Haven't You Told Your Child?"
Monthly Bulletin. 1913; 16(6):217-218.</dcterms:bibliographicCitation>
        <dc:source>IUPUIScholarWorks Repository</dc:source>
        <dc:language>English</dc:language>
    </sparks:Image>
</rdf:RDF>
```

ATTACHMENT A

Historic Public Health Metadata Sample
Includes Relational Terms and Definitions (if needed)

Sanatoriums	LCSH	**SEE ALSO Open-Air Sanatoriums**
Sand	LCSH	**SEE ALSO Water Purification Sand Filtration**
Sanitary Engineering	MESH	**Design, construction & maintenance of environmental facilities**
Sanitary Landfills	PHIN	**SEE ALSO Garbage, Waste Disposal, Solid**
Sanitation	PHIN	**Formulation and application of measures designed to protect public health or Disposal of sewage.**
Sarcoma	MESH	
Sarsaparilla	LCSH	**Flavoring used in root beer**
Satire	LCSH	**SEE ALSO Political Satire**
Sauerkraut	LCSH	
Sausage	Local	**USE ALSO Meat Products; SEE ALSO Frankfurters**
Scabies	PHIN	
Scarification	Local	**Archaic=vaccination method for smallpox making a small incision to insert the vaccine; Modern=deliberately scar the skin for beauty/other purposes**
Scarlatina	Local	**Archaic; USE ALSO Scarlet Fever**
Scarlet Fever	PHIN	
Scarlet Rash	Local	**Archaic; USE ALSO Scarlet Fever**
Scars	TGM	**SEE ALSO Scarification**
Schick Test	Local	**Diphtheria test; toxin injected into arm to test for a reaction, if none, then diphtheria antibodies are present**
Schiff's Reagent	MESH	
Schizophrenia	PHIN	
School Age Children	PHIN	**Ages 5-12**
School Attendance	LCSH	
School Breakfast Programs	LCSH	
School Buildings	LCSH	
School Buses	PHIN	
School Closings	LCSH	

LCSH = Library of Congress Subject Headings
Local = Archaic or colloquial terms researched/created by Primary Investigator
MESH = Medical Subject Headings (National Library of Medicine)
PHIN = Public Health Information Network (CDC)
TGM = Thesaurus for Graphic Materials (Library of Congress)

Attachment B. Indiana Public Health Historic Collections Search Page Infantile Prototype

Demonstrating search results for "Tuberculosis"