Blink Project:
Linked Open Data for Countway Library
Final report for Phase 1 (June-Nov. 2012)

Prepared by Sophia Cheng

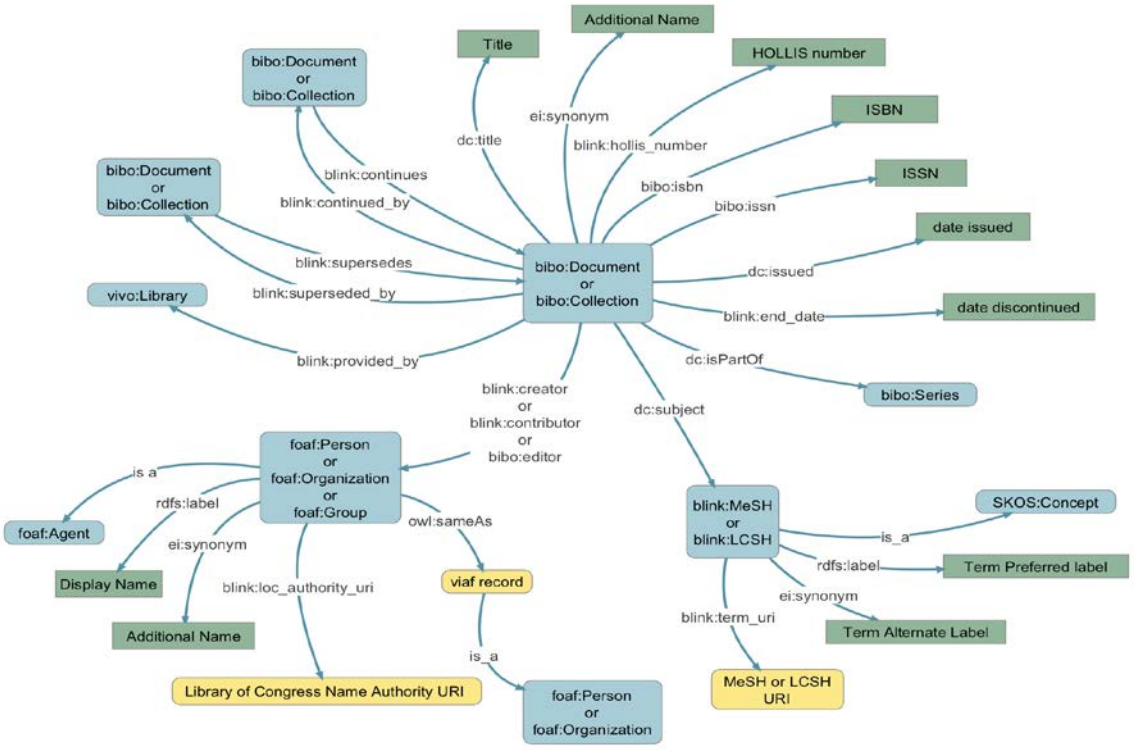
Summary
We propose to improve the usefulness and discoverability of Countway Library's digital collections by exposing bibliographic information as Linked Open Data, and then utilizing these new data points to enhance searches. We believe that this will make the data much more useful to our researchers, and it will also allow us to link our bibliographic data to other linked data sources, such as the Harvard Catalyst's Profiles, the Virtual International Authority File (VIAF), and eagle-i resources. This project will also address a deficiency that currently exists in the library's public catalogs: the entry terms/see-references for subject headings and the variant forms of personal and corporate names are not included in searches, resulting in inaccurate search results.

Approach

1) *Ontology selection*
   After a survey of available ontologies for representing bibliographic data, we found that there was no single ontology that encompassed everything we wanted to represent. We decided to select parts from several ontologies and created additional terms as a custom ontology when we could not find a term that was appropriate. We selected terms from the following existing ontologies:

   - bibo (http://bibotools.googlecode.com/svn/bibo-ontology/trunk/doc/index.html)
   - Dublin Core (http://dublincore.org/documents/)
   - SKOS (http://www.w3.org/2004/02/skos/)
   - VIVO (http://vivoweb.org/)
   - foaf (http://xmlns.com/foaf/spec/)

   The current blink ontology has the following structure:

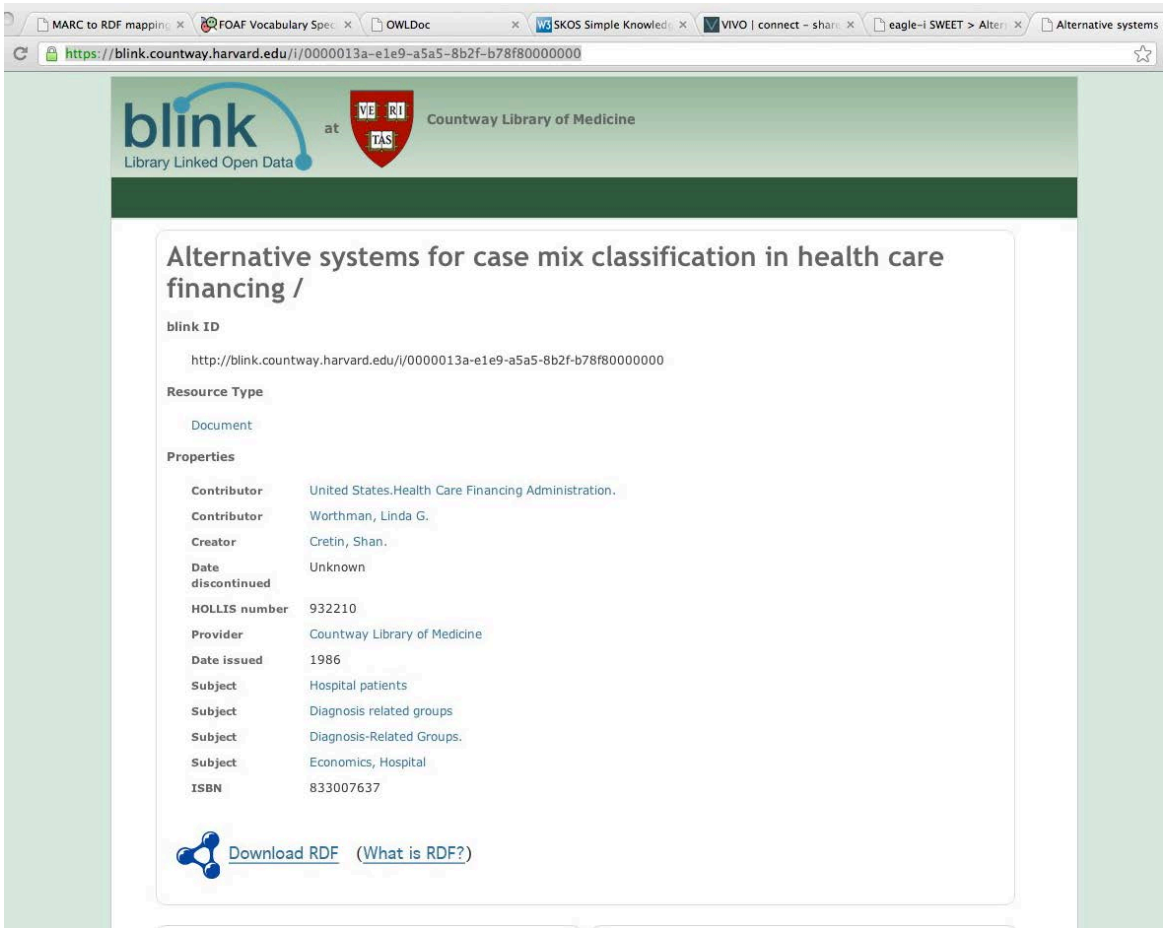


2) *Repository and software tool selection*

To expedite the development process, we chose to reuse and customize existing eagle-i components for this project. Since members of our team have developed the eagle-i software, we have intimate knowledge of how it functions and are able to quickly and efficiently customize the software for use in this project. Additionally, the software itself is robust, provides features that we need as well as nice "extra" features.

Our bare minimum software needs to support this project are a triple store repository for storing the data and some way to manipulate the repository for loading and editing the data. From the eagle-i project, we are able to reuse the repository web application. This application contains a set of APIs that abstract away the low level interactions with the underlying sesame triple store. We are also using the ETL scripts that make use of the repository APIs to extract, transform and load data into the repository. These scripts include generation of a spreadsheet template based on the ontology that is then filled with data to be transformed into RDF and loaded.

By using the eagle-i repository, we also get a SPARQL workbench to build and execute SPARQL queries on the data as well as nice dissemination pages for the data:

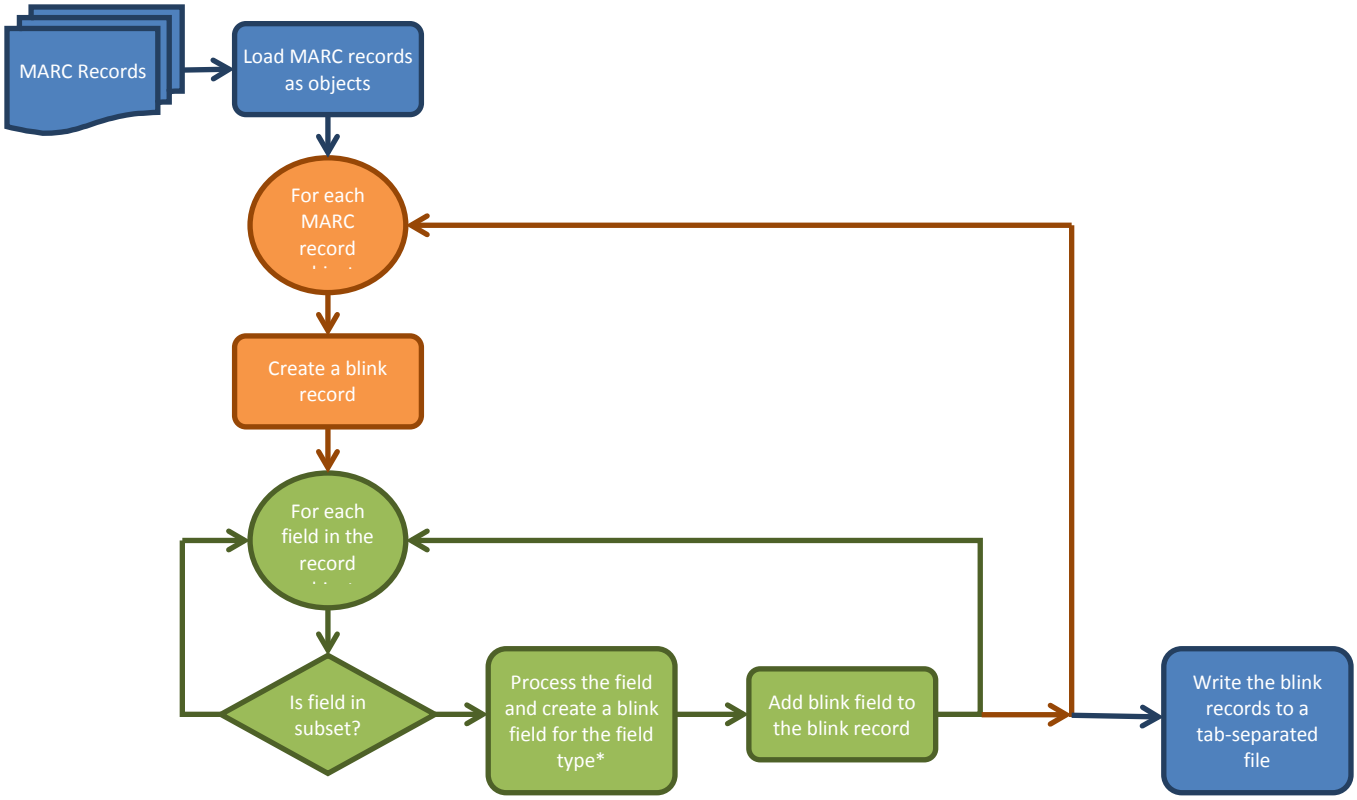


There are two additional web applications, model and SWEET, which required very light customizations to make them usable with this project. The model application is a web application that provides a user interface for browsing the ontology, https://blink.countway.harvard.edu/model/. The SWEET application is also a web application that provides a user interface for curating and entering data, https://blink.countway.harvard.edu/sweet/. We are also using the SWEET application to browse the data.

3) *Test records*

A set of 67 representative records from the Countway Digital Library was selected as our test set to transform and load. These records were provided in MARC format. Our first step in this process was to select a subset of all available MARC fields and subfields to use for mapping to RDF for this proof of concept (see appendix I). The next step was to map this subset to the ontology (see appendix II).

With the mappings we defined, we developed a rough pipeline to convert the MARC records into a tab-separated file conforming to the template generated by the ETL scripts.



We were able to group the fields into several types based on the subfields of interest:

- Date
- International Serial Number
- Linking
- Name
- Subject
- Title

For example, because of the non-standard way that Personal Name fields, we have decided to treat this field as a string by concatenating the relevant subfields. Our plan is to have this information available as a guide for disambiguating names when linking to external sources. Another example is fields related to subject. We have decided for simplicity to focus on only the MESH and LCSH terms. For this field type, we attempt to resolve the text from the MARC record with the actual term URI, i.e. the MESH term "Economics, Hospital" will have a link out to http://purl.bioontology.org/ontology/MSH/D004469.

4) *Identify and link to external data sources*

We have identified the following as useful external data sources to link to:

- NCBI's MeSH
- LC Subject Headings
- LC Name Authority File
- VIAF

At this time, we are able to automatically look up and find the corresponding term URI for MeSH and LC sources.

5) *Develop a user interface*

In addition to the user interfaces that are part of the eagle-I component, a very simple user interface was developed to demonstrate the utility of searching records that are enhanced by linked open data. This interface allows users to search by subject and see a side-by-side comparison of results using the 'conventional' search and with the search enhanced with linked open data. It is important to note that this interface does not actually do synonym expansion or actually do the search via HOLLIS. Instead, it mimics the searches using SPARQL query. For the 'conventional' search, the SPARQL query does a keyword search over the data. For the 'improved' search, the SPARQL query does a keyword search over the data, including fields that would ideally be populated by the linked nature of data.

'Conventional' SPARQL query

```
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX bibo: <http://purl.org/ontology/bibo/>
PREFIX blink: <http://countway.harvard.edu/ont/blink/>
PREFIX dcterms: <http://purl.org/dc/terms/>

SELECT DISTINCT ?document ?document_label ?document_author ?secondary_record
?secondary_record_label ?linking_property
WHERE {
?document a bibo:TYPE .
?document rdfs:label ?document_label .
OPTIONAL { ?document blink:creator ?document_author_uri . ?document_author_uri rdfs:label
?document_author } .{
?document ?linking_property ?secondary_record  FILTER( ( ?linking_property !=
blink:provided_by ) ).
?secondary_record rdfs:label ?secondary_record_label .
FILTER ( isLiteral(?secondary_record_label) && REGEX( ?secondary_record_label,
"QUERY_TERM", "i") )
}
}
```

'Improved' SPARQL query

```
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX bibo: <http://purl.org/ontology/bibo/>
PREFIX blink: <http://countway.harvard.edu/ont/blink/>
PREFIX dcterms: <http://purl.org/dc/terms/>

SELECT DISTINCT ?document ?document_label ?document_author ?secondary_record
?secondary_record_label ?linking_property ?secondary_match
WHERE {
?document a bibo:TYPE .
?document rdfs:label ?document_label .
OPTIONAL {
?document blink:creator ?document_author_uri . ?document_author_uri rdfs:label
?document_author} .
{
?document ?linking_property ?secondary_record  FILTER( ( ?linking_property !=
blink:provided_by ) && ( ?linking_property = dcterms:subject ) ) .
?secondary_record rdfs:label ?secondary_record_label .
?secondary_record ?any2 ?secondary_match FILTER( isLiteral(?secondary_match) && REGEX(
?secondary_match, "QUERY_TERM", "i") )
}
}
```

6) *Query re-writing tool*
   We have a prototype of the query re-writing tool that will allow searches to also consider entry terms and variant forms of names. For example, searches for the term "cancer," which in the MeSH thesaurus is an entry (or non-preferred) term for the MeSH heading "Neoplasms" will, in the Blink tool, map to "neoplasms" and the tool will add the preferred term to the query. Our prototype tool has been set up to use the Unified Medical Language System (UMLS) and MeSH.

Challenges

There were three main challenges we faced during this project. The first of these challenges was the selection of the ontology. We attempted to reuse existing ontologies as much as possible. However, we found that the existing ontologies developed for bibliographic purposes were developed with monographs in mind. In our use case, we have many serials with properties that could not be correctly mapped to any of the existing terms. In the end, we chose to create custom ontology terms such as "Supersedes" for MARC field 785 (Preceding entry). Another problem we faced was that some of the ontologies we chose had conflicting annotations and required manual debugging to resolve. Finally, the size of the MeSH ontology posed a problem. We initially planned to load MeSH into our application and thereby giving us access to MeSH defined synonyms. As it turns out, we were unable to do this because of limitations in hardware. Instead, for the demonstration, we manually added the synonyms to the respective records in our repository.

The next set of challenges we faced was developing the pipeline to convert MARC records into RDF. We had hoped to process a majority of the information in the records automatically and with little human intervention. We were not expecting the variation and lack of standard format for many of the fields. For example, it was not possible to develop code to parse the name field because of the many variations and different locations of the data. The pipeline also required several iterations of tweaking how the fields were processed because of the variability in the data.

Finally, we were unable to implement the full functionally of linked open data with the selected external data sources because they did not provide a SPARQL endpoint.

Next Steps

The next steps for this project include:

- Provide a SPARQL endpoint for external sources that do not have one by periodically loading their information into our repositories.
- Attempt to load the MeSH ontology into our application
- Refine the pipeline
- Refine the query re-writing tool, generalizing the concept mapper for more ontologies, pulling in LCSH, and connecting to the Blink site.

Administrative

- *Code Deposit*
- *Budget spent*
- *Outreach*

Appendix I
Selected Subset of MARC fields and subfields

Field Code	Field Name	Subfield	Notes
LDR	Leader	6, 7	
001	Control Label		
008	Fixed-Length Data elements - General Information	7-14	Dates for record
020	ISBN	$a	
022	ISSN	$a	
100	Main Entry (Personal Name)	$a $b $c $d $q	
110	Main Entry (Corporate name)	$a $b $c $d $q	
111	Main Entry (Meeting name)	$a $c $d $q	
245	Title Statement	$a $b	
246	Varying form of title	$a	
362	Dates of publication	first and second indicator, $a	
600	Subject added entry (Personal name)	second indicator, $0 $2 $a $b $c $d $q $t	MESH & LOC subjects only
610	Subject added entry (Corporate name)	second indicator, $0 $2	MESH & LOC subjects only
630	Subject added entry (Uniform title)	second indicator, $0 $2 $a $n $p	MESH & LOC subjects only
650	Subject added entry (Topical term)	second indicator, $0 $2	MESH & LOC subjects only
651	Subject added entry (Geographic name)	second indicator, $0 $2	MESH & LOC subjects only
655	Index term (Genre/form)	second indicator $0 $2	
700	Added Entry (Personal name)	$a $b $c $d $q $t	

710	Added Entry (Corporate name)	$a $b $c $d $n $t	
711	Added Entry (Meeting name)	$a $c $d $n $t	
730	Added Entry (Uniform title)	second indicator, $a $n $p	
740	Added Entry (Uncontrolled related/analytical title)	second indicator, $a $n $p	
780	Preceding entry	second indicator, $a $t $x $z $w	Select relationships only.
785	Succeeding entry	second indicator $a $t $x $z $w	Select relationships only.
830	Series Added Entry (Uniform title)	$a $n $p $v	
856	Electronic location and access	$3 $z $u	

Appendix II
Mapping of MARC to Blink Ontology

MARC field codes	Maps to	Notes
245	Title (http://purl.org/dc/terms/title)	Used to identify the resource
246, 730, 740	Additional Name (http://eagle-i.org/ont/app/1.0/synonym)	
100, 110, 111	Creator (http://countway.harvard.edu/ont/blink/creator)	
600, 610, 630, 650, 651, 655	Subject (http://purl.org/dc/terms/subject)	Separated into concepts of MESH and LOC
700, 710, 711	Contributor (http://countway.harvard.edu/ont/blink/contributor)	
830	Part of series (http://purl.org/dc/terms/isPartOf)	
001	Hollis Number (http://countway.harvard.edu/ont/blink/hollis_number)	
020	ISBN (http://purl.org/ontology/bibo/isbn)	
022	ISSN (http://purl.org/ontology/bibo/issn)	
LDR $7	Document type	
780	Supercedes (http://countway.harvard.edu/ont/blink/supersedes) Supercedes in part (http://countway.harvard.edu/ont/blink/supersedes_in_part)	
785	Superseded by (http://countway.harvard.edu/ont/blink/superseded_by) Superseded in part by (http://countway.harvard.edu/ont/blink/superseded_in_part_by)	
008 7-10	Date issued	

	(http://purl.org/dc/terms/issued)	
008 11-14	Date discontinued (http://countway.harvard.edu/ont/blink/end_date)	