What is old is new again: Linked Data as bibliographic description, and the curation of digital collections harvested from the 'Net

The "how's" of librarianship are changing, but not the "what's".

Linked Data

Linked Data is a method of describing objects, and these objects can be the objects in a library. In this way, Linked Data is a type of bibliographic description.

Linked Data is a manifestation of the Semantic Web. It is an interconnection of virtual sentences known as triples. often possible to make explicit assuming information and new knowledge.

There are an increasing number of applications enabling libraries to transform and convert their bibliographic data into Linked Data. One such application is called the ALIADA. ¹

When & if the intellectual content of libraries, archives, and museums is manifested as Linked Data, then new relationships between resources will be uncovered and discovered. Consequently,

```
/ -- http://example.org/name ------> "Rome"
| -- http://example.org/description ---> "A long time ago..."
http://example.org/rome - | -- http://example.org/founding-date -> "1000 BC"
| -- http://example.org/sub-part ----> http://example.org/italy
\ -- http://example.org/type -----> http://example.org/city
```

Simple Arced Graph

Triples are rudimentary data structures, and as the name implies, they are made of three parts: 1) subjects, 2) predicates, and 3) objects. Subjects always take the form of a URI (think "URL"), and they point to things real or imaginary. Objects can take the form of a URI or a literal (think "word", "phrase" or "number"). Predicates also take the form of a URI, and they establish relationships between subjects and objects. Sets of predicates are called ontologies or vocabularies and they present the languages of Linked Data.

Through the curation of sets of triples, and through the re-use of URIs, it is

one of the purposes of cultural heritage institutions will be realized. Thus, Linked Data is a newer, more timely method of describing collections; what is old is new again.

Curation of digital objects

The curation of collections, especially in libraries, does not have to be limited to physical objects. Increasingly new opportunities regarding the curation of digital objects represent a growth area. With the advent of the Internet there exists an abundance of full-text digital objects just waiting to be harvested, collected, and cached. It is not good

¹ ALIADA will automatically publish a Linked Data cloud of datasets hosted by different library or collection management software. See https://github.com/ALIADA/ aliada-tool

enough to link and point to such objects because links break and institutions (websites) dissolve.

Curating digital objects is not easy, and it requires the application of traditional library principles of preservation in order to be fulfilled. It also requires systematic organization and evaluation in order to be useful.

Done properly, there are many advantages to the curation of such digital collections: long-term access, analysis & evaluation, use & re-use, and relationship building. Examples include: the creation of institutional repositories, the creation of bibliographic indexes made up of similar open access journals, and the complete works of an author of interest.

In the recent past I have created "browsers" used to do "distant reading" against curated collections of materials from the HathiTrust, the EEBO-TCP, and ISTOR. Given a curated list of identifiers each of the browsers locally caches the full text of digital object object, creates a "catalog" of the collection, does full text indexing against the whole collection, and generates a set of reports based on the principles of text mining. The result is a set of both HTML files and simple tabdelimited text files enabling the reader to get an overview of the collection, query the collection, and provide the means for closer reading.

How can these tools be used? A reader could first identify the complete works of a specific author from the HathiTrust, say, Ralph Waldo Emerson. They could then identify all of the journal articles in JSTOR written about Ralph Waldo Emerson. Finally the reader could use the HathiTrust and JSTOR browsers

to curate the full text of all the identified content to verify previously established knowledge or discover new knowledge. On a broader level, a reader could articulate a research question such as "What are some of the characteristics of early American literature, and how might some of its authors be compared & contrasted?" or "What are some of the definitions of a 'great' man, and how have these definitions changed over time?"

The traditional principles of librarianship (collection, organization, preservation, and dissemination) are alive and well in this digital age. Such are the "whats" of librarianship. It is the "hows" of the librarianship that need to evolve in order the profession to remain relevant. What is old is new again.

Eric Lease Morgan Artist- and Librarian-At-Large

October 12, 2015



Word cloud illustrating the "great ideas" written by Ralph Waldo Emerson