

The Academy Unbound

Linked Data as Revolution

Philip Evan Schreur

Linked data has the potential to revolutionize the academic world of information creation and exchange. Basic tenets of what libraries collect, how they collect, how they organize, and how they provide information will be questioned and rethought. Limited pools of bibliographic records for information resources will be enhanced by data captured at creation. By harvesting the entire output of the academy, an immensely rich web of data will be created that will liberate research and teaching from the limited, disconnected silos of information that they are dependent on today.

A revolution is at hand, one that is potentially as world-altering as the development of the web. And, as are most truly transformative revolutions, it is driven by a simple concept: in this case, linked data. Linked data has the potential to change most aspects of the universe of information creation and exchange. As a primary purveyor of information, the academy will be at the nexus of this revolution. The information infrastructure of this world will be dramatically altered as basic tenets of what it collects and how it collects, organizes, and provides information are questioned and rethought. Much has been said about linked data, its ties to the Semantic Web, and its application for libraries, but what is it exactly and how does it work?

This paper aims to answer these questions by defining linked data, discussing problems with libraries' focus on bibliographic records, and exploring the potential of linked data as a solution in a rapidly evolving global discovery environment. A new discovery approach developed by the Bibliothèque nationale de France is presented as a service that takes advantage of the potential of linked data.

What Is Linked Data?

Linked data has so much potential because it is imbedded in the fabric of the web. As more aspects of professional and private life move to the cloud, the way in which information is stored and linked on the web becomes crucial. The four tenets of linked data are simple: (1) use URIs (Uniform Resource Identifiers) to name resources on the web; (2) use HTTP URIs so someone can find the resources; (3) have the information provided by the link be useful, and provide

Philip Evan Schreur (pschreur@stanford.edu) is Head, Metadata Department in Technical Services, Stanford University Libraries, Stanford, California.

Submitted March 12, 2012; tentatively accepted, pending modest revision April 4, 2012; revision submitted May 7, 2012, and accepted for publication May 12, 2012.

that information using standards (RDF, SPARQL); and (4) provide links to other URIs so people can discover related information.¹

Linked data are commonly published using the Resource Description Framework (RDF).² Each expression in RDF has a subject, a predicate, and an object. This simple structure allows anyone to make simple assertions about anything, for instance, *The Raven* (subject) has author (predicate) Edgar Allan Poe (object). Ideally, both the subject and object would be represented by URIs (a string of characters used to identify a name or resource on the web) and the statement itself expressed using an XML-based syntax. By using RDF, applications can exchange information on the web without loss of meaning. Because RDF is a widely used model, information expressed with it can be used by many applications and applications can be developed to take advantage of this growing pool of data.

RDF is a model of entities and relationships. As such, it is well adapted to be the basis of an entity-relationship model in support of linked data.³ A strength of this model is that it allows anyone to make assertions about anything. Equally powerful is that any two entities may be linked and, through this process, an immensely rich web of data can be created. Although nothing requires that these statements are true (e.g., “*The Raven* has author Philip Schreuer” is an equally valid statement in RDF as “*The Raven* has author Edgar Allen Poe”), anyone may correct invalid statements. Through this iterative process of data use and correction, the web of data becomes richer and more reliable; this is crowdsourcing at a truly international level.

Breaking the Tyranny of Records

Since the days of the card catalog, libraries have focused on bibliographic records. These discreet bundles of information supply metadata about resources in collections. Their record structure is carefully controlled, and access points such as names, subjects, or series come from recognized thesauri and carefully curated authority files. Even with the transition to online catalogs made possible by the development of MARC by Henriette Avram in the 1960s, the focus has remained on records.⁴ The information they contain is fractured into various fields and subfields and stored in relational databases where they can be associated and maintained. Theories of bibliographic control arose over time.⁵ The possibility of effectively organizing and retrieving resources by using controlled description, analysis, and classification seemed achievable as libraries dealt with finite information in a closed system.

The integrated library system (ILS) was developed to take advantage of bibliographic records within all of the library's functional areas (acquisitions, cataloging,

circulation, etc.), and bibliographic utilities such as OCLC and SkyRiver help libraries exchange these records between their various systems. Collaborative efforts such as the Program for Cooperative Cataloging (PCC) and the American Library Association (ALA) help maintain standards for these records so libraries can be assured of quality data as they exchange their work.

This preoccupation with bibliographic records has drawbacks. First, many institutions began to favor their own particular version of a bibliographic record. Even though OCLC might espouse the use of the master record in its database, libraries are free to alter and enhance the copy of that record in their local databases. Corrections to perceived errors in other's cataloging, missing data elements, and local practices can be incorporated into an in-house version of the record designed to meet local users' needs. Large numbers of staff are dedicated to this work at enormous cost. As the number of records grows, so does the cost of attempting to maintain them.

A second drawback is that a relational database is, by definition, a closed system. For a patron to discover a resource in the online catalog, a bibliographic record for it must be present in the system. Bibliographic records for many of the resources a library owns may never appear in the catalog because of a shortage of staff to create the records needed. In considering the amount of resources on the web, the problem grows by orders of magnitude. Within a world of limited staffing and records in relational databases, consistent access to the web of data is impossible.

Linked data, however, is not focused on bibliographic records but individual statements of fact. No discreet records need to be maintained in a local ILS, and no master records need to be maintained in a world-wide relational database. Instead, massive collections of triples in triple stores (i.e., RDF databases) will suffice.⁶ By bypassing the *a priori* need for a record, linked data frees libraries from the cycle of record creation, maintenance, and deletion. Valuable staff time can be freed from these activities, and the confines of the relational database can be broken.

Beyond the Curated Catalog

Libraries have spent decades, sometimes centuries, carefully selecting the resources to add to collections. This process of curation has focused collections on the needs of individual communities over time, giving individual collections their unique point of view. The records in local catalogs reflect the discovery needs of individual communities, a closely knit cycle of acquisition and access. Local catalogs, however, are flawed in several ways.

Large parts of collections never appear to any great depth in the library's catalog. Archival materials are a notable

example. These holdings are often accessible through paper finding aids, or, if electronic, described in the Encoded Archival Description format (EAD). Institutions also are notorious for creating stand-alone databases in a variety of transient formats for accessing discreet collections, but by doing so they isolate this content from broader discovery.

By strongly advantaging the purchase of large, generic e-book packages, e-book vendors inadvertently promote the breakdown of a collection's point of view. Institutions also might choose to load large quantities of copyright-free digital books (such as those available from the HathiTrust), making their collections even more generic. As individual collections become more generic, community ties to those collections reflected in the local catalog weaken.

Recently, libraries have explored means of loading nontraditional bibliographic records directly into enhanced discovery environments. The metadata for these resources, however, are often nonstandard, tailored to a specific collection, and do not integrate well within the broader context of the traditional catalog. By merging these collections, libraries can join silos of information, but inconsistent metadata choices keep them isolated at the data level.

As more and more substantial information appears on the web, many library patrons have shifted their discovery there. With Google indexing library catalog records, patrons have a good chance of finding library materials on the web but no chance of finding the breadth of data available on the web in library catalogs. At first glance, this movement of patrons to the web seems puzzling after libraries have spent so much time curating collections and carefully grooming catalog records, but the direction is very clear. As collections become less and less distinctive and data on the web becomes more and more pervasive, the process of discovery has turned outward. If libraries are to meet patrons' need for more comprehensive information discovery, they must move beyond curated catalogs and provide discovery environments based on the web and its architecture.

Beyond MARC

Libraries have always focused on the content of their local collections. With limited funds, libraries must select materials carefully, taking into account both current uses of the collection and anticipation of future needs. Once the content has been acquired, it must be made discoverable or the acquisition process is pointless. To do this, libraries have relied on bibliographic surrogates as discovery vehicles. With carefully controlled access points and consistent descriptive structure, these surrogates populate library catalogs and make library assets discoverable.

As impressive as these surrogates are, they pale before the resources themselves. A serial record gives no hint to

the plethora of articles a serial title contains. Multivolume monographs may contain any number of immensely important resources. Records for *Festschriften* and other collective formats may or may not include a table of contents note detailing the articles they contain. E-books offer a bit more flexibility. Publishers may assign ISBNs to chapters, paragraphs, charts, or other parts of a book if they intend to sell them separately, but these individual chapters will need descriptive metadata in their own right.

To help fill this gap, libraries look to vendors to supply enhanced content for resources they have acquired. Vendors such as Nielsen Bookdata Service (www.nielsenbookdata.co.uk) provide not only tables of contents, but book covers, author biographies, short and long descriptions, short and long reviews, and promotional information. These enhanced data can give patrons a better idea of what a resource actually contains and provide better tools for evaluating content quality. These enhanced data can feed more elaborate search tools that can give patrons new windows on discovery.

All this additional content can be problematic, however. It is overpowering, overwhelming, and challenges librarians' traditional concept of a discovery environment. Although library discovery environments have experienced tremendous improvements in searching, the presentation of content has been nearly static. The initial display screen typically replicates the initial card in the card catalog. Real estate is limited and only a selected amount of information can be displayed. Because much of the information is generated from MARC data, which was created to communicate information used to generate the cards for catalogs, the display of this information often carries the feel of that earlier display. The problem is that supplemental information, such as a three-page review, does not fit into this concept. These data enhancements often have no place in the MARC format, and, if they did, displaying this content would present challenges. A choice must be made as to what information is suppressed from the initial screen to make room for these enhancements, and the library runs the risk of patrons finding that enhanced data more irritating than useful.

In addition to these more traditional resources and enhanced descriptions, libraries have many untapped sources of content surrounding them. An academic environment is an incredibly rich one. Faculty and students are continually creating new data in their everyday activities. Class syllabi and reading lists capture curated data on particular subjects, datasets support research and publications, and all these data feed on themselves in a continuous cycle of growth, yet little is captured or shared across the academic enterprise.

One example can be seen in the research generated by students in a course in paleography taught at Stanford University. As a project, each class member chooses an unidentified, loose manuscript fragment in the Stanford collection. The student must identify the fragment of text it contains,

identify the style of writing, and give an approximate date of creation. Over the years, a tremendous amount of scholarship has been generated on these anonymous manuscript fragments, but all those data are trapped in the reports generated for the class. These supplemental data could be of tremendous help to scholars examining the medieval fragments Stanford holds and perhaps in matching them to fragments in other collections. But unless the paper files from the course are requested through the service desk in Special Collections, no one will ever see them.

This project is only a small example of the wealth of information that the academy creates in its day-to-day functioning. How should these data be made available? To create catalog records for each project in every class would be impossible and the records themselves would poorly represent the depth of information the reports contain. Minimal records, although simpler to create, would negatively affect the sophisticated faceting in new discovery environments. Discovery becomes the crucial factor and more bibliographic records cannot be the answer.

Role of Metadata

Metadata are critical in any discovery environment and metadata come in many guises. Publishers communicate bibliographic information through ONIX, a standard equally as complex as MARC but focused on communicating information within the publishing industry. Various schemas exist for encoding bibliographic data outside of the book trade such as DC or MODS. Metadata may be embedded within web documents themselves through a set of attribute-level extensions to XHTML following the RDF data model (RDFa).⁷ Publishers with large article databases may develop their own internal formats for organizing information. At present, however, most library discovery environments are based on MARC records for discreet resources: books, journals, maps, videos, etc.

In a linked data environment, these MARC records often are seen as a prime, preliminary source of information. All the effort that catalogers have put into controlled subject access, controlled names, classification, and consistent description has made MARC records extremely desirable. Because any library foray into linked data must begin with its collection of MARC records, a closer look at that format is worthwhile.

Figure 1 offers a typical MARC catalog record for a sound recording. The record is quite impressive. It gives a description of the medium, the contents, the years of performance, controlled subject headings, and analytical entries for all the individual musical works it contains, and it displays the information in an easily digestible structure for the eye. Anyone glancing at this record can see that it represents a


Author/Creator:	Kreisler, Fritz, 1875–1962.
Imprint:	Pembury, Kent : Pearl, [1975]
Format:	 Music – Recording 1 sound disc : 33 1/3 rpm, mono. ; 12 in.
Note:	Title from container spine.
Contents:	<ul style="list-style-type: none"> o Double concerto in D minor, BWV 1043 / Bach (with: Efrem Zimbalist, violin ; string quartet) o Four encores (with piano) : Chant sans paroles / Tchaikovsky (originally for piano). Air on the G string / Bach (originally for orchestra). Sarabande / Sulzer. L'abeille / Schubert o Violin concerto no. 4 in D major, K. 218 / Mozart (with: London Symphony Orchestra ; Sir Landon Ronald, conductor)
Participant:	Fritz Kreisler, violin, with various accompaniments.
Event:	Recorded 1904–1924.
Contributor:	Zimbalist, Efrem. Performer Ronald, Landon, Sir, 1873–1938. Conductor London Symphony Orchestra Performer
Included Work:	Bach, Johann Sebastian, 1685–1750. Concertos, violins (2), string orchestra, BWV 1043, D minor. Tchaikovsky, Peter Ilich, 1840–1893. Souvenir de Hapsal. Chant sans paroles; arr. Bach, Johann Sebastian, 1685–1750. Suites, orchestra, BWV 1068, D major. Air; arr. Sulzer, Joseph. Saraband, violin, piano, op. 8. Schubert, Franz, 1797–1828. Bagatelles, violin, piano, op. 13. Abeille. Mozart, Wolfgang Amadeus, 1756–1791. Concertos, violin, orchestra, K. 218, D major.
Related Work:	Mozart and Bach concertos.
Subjects:	Concertos (Violins (2) with string ensemble) Violin and piano music, Arranged. Violin and piano music. Concertos (Violin)

Figure 1. MARC Record for Sound Recording

recording of Fritz Kreisler performing a selection of violin music. The musical works are clearly articulated and responsibilities are clear from glancing at the record as a whole.

Much of the semantic meaning in this example can only be derived from the entire bibliographic record. The human eye can easily see that the main entry is Fritz Kreisler and that he is a violinist, that the piece by Joseph Sulzer is for violin and piano, and that the Mozart Violin concerto is accompanied by the London Symphony Orchestra conducted by Sir Landon Ronald. Also, if the patron liked the composition by Sulzer, he or she could follow the subject heading “Violin and piano music” to find similar works.

This dependence on a complete bibliographic record for semantic meaning is a holdover from the card catalog. The MARC format allowed these records to be transformed into electronic documents and shared internationally, but they are still bibliographic records, and to be understood, must be evaluated as a whole. Individual statements such as “Fritz Kreisler, violin, with various accompaniments” or “Recorded 1904–1924” are meaningless out of context. RDF, however, is a model that allows for the creation of a

series of independent statements. The publishing of MARC records as RDF triples has to overcome two great obstacles. The first is the concept of the bibliographic record and the second is the inability of the MARC communication format to convey semantic meaning clearly.

Realizing how much information the mind supplies is difficult. One sees from the author field that Fritz Kreisler is listed as a creator and, from the participant note, that he is a violinist. From the contributor field, one sees the recording includes Efrem Zimbalist and, from the contents note, that he also is a violinist. From the contents note, one sees that Kreisler performs a piece by Tchaikovsky (“Chant sans paroles”) that was originally for piano, from the analytical added entries that this piece is from Tchaikovsky’s work “Souvenir de Hapsal,” and from the subject headings that the correct Library of Congress Subject Heading (LCSH) term for this work is “Violin and piano music, Arranged.” Nothing in the bibliographic record itself, though, meaningfully links this information together. The human mind makes these logical associations

The MARC format was created to communicate clearly the information encoded in card catalogs, and, in this, has been very successful. Although perpetuating the concept of the bibliographic record, MARC very clearly articulates and differentiates all the elements in the record. By turning these bibliographic records into an electronic form, MARC has allowed for the development of the ILS, the online catalog, programs such as the PCC, and organizations such as OCLC. However, MARC is used exclusively by the library community. As libraries seek to encompass all the data generated both within their institutions and throughout the world, they must stretch far beyond MARC. In the semantic world of linked data, these MARC records themselves are inarticulate.

The need for a replacement for MARC has been discussed for many years. On October 1, 2011, the Library of Congress (LC) released the initial plan for its Bibliographic Framework Transition Initiative.⁸ The initiative will be international in scope and designed to “reap the benefits of newer technology while preserving a robust data exchange that has supported resource sharing and cataloging cost savings in recent decades.”⁹ One of the key requirements for the initiative is the “accommodation of textual data, linked data with URIs instead of text, and both. It is recognized that a variety of environments and systems will exist with different capabilities for communicating and receiving and using textual data and links.”¹⁰ Undoubtedly this effort will take years to complete, and considerations such as backward compatibility will be of prime concern. But this transition to an exchange format that has buy-in beyond the confines of the library community, is compatible with other communication formats, and has greater semantic understandability in a linked data environment will be the single most important element in the transition to linked data that the library world faces.

Why Linked Data?

Various groups have sought to answer this question. From June 27 to July 1, 2011, Stanford University hosted several librarians and technologists to examine the use of linked data in the academic environment.¹¹ The hope was that participants at the linked data workshop could both confront the challenge of planning a multinational, multi-institutional discovery environment and lay the groundwork for its development. One of the most interesting products of the workshop was a series of value statements as to why a linked data approach was worth pursuing:

- Linked Open Data (LOD) puts information where people are looking for it—on the web.
- LOD can expand discoverability of our content.
- LOD opens opportunities for creative innovation in digital scholarship and participation.
- LOD allows for open continuous improvement of data.
- LOD creates a store of machine-actionable data on which improved services can be built.
- Library LOD might facilitate the breakdown of the tyranny of domain silos.
- LOD can provide direct access to data in ways that are not currently possible, and provides unanticipated benefits that will emerge later as the stores of LOD expand exponentially.¹²

As people shift to the web as their first point of discovery, library resources need to be represented there. Although catalog records may appear on the web, much of the meaning embedded in the MARC coding is lost. For the most part, the data in these records are treated by search engines as blocks of text. By using RDF, however, important information encoded by the MARC tags can be translated into triples that carry semantic meaning for machine processing. Each one of these elements can be recorded as a URI that can link these data points to matching data points within the web of data. The RDA (Resource Description and Access) Vocabularies have been published on the Open Metadata Registry, making its element sets and value vocabularies available for all to use.¹³ Standard thesauri and authority files are becoming available as source URIs for linked data through such sites as the Library of Congress Authorities and Vocabularies website (id.loc.gov). By intelligent conversion of library MARC records to machine resolvable RDF triples, the semantic meaning in the records can be realized in individual statements. By moving these statements to the web, the data becomes a vital, structural part of the Semantic Web.

This initial transition would fulfill many of the value statements of the Stanford Linked Data Workshop. By

converting bibliographic data to linked data, libraries move their data to the web (value statement 1) and so expand their discoverability (value statement 2). This linked, open data store then becomes available for anyone to experiment with developing new applications (value statement 3) and, because the bibliographic data are encoded as linked data, they can be understood semantically for machine processing (value statement 5). Because linked data are format agnostic, any schema (MARC, Dublin Core (DC), Metadata Object Description Schema (MODS)) or any file format (FileMaker Pro, Luna, etc.) can be converted to linked data. All the data silos that have evolved during the last fifty years can finally be broken down (value statement 6). Because the data are there on the web, they are available for anyone to correct and improve (value statement 4). Last, as libraries move beyond their metadata into the data itself and link at an elemental level, they exponentially expand the data to which scholars have direct access (value statement 7). But the academic environment is far richer than its MARC records and has much more to offer the web of data.

Discovery

Discovery is a key issue in any knowledge environment yet libraries' current relational databases have been pushed to the limit both by the growing volume of resources and the constraints of current systems. Traditional ILS environments have required that bibliographic records be ingested in the MARC format, and so much effort is spent mapping data elements from DC, MODS, EAD, article metadata, Excel spreadsheets, local schemes, etc., to MARC fields. Because many other data formats are not as complex as MARC, or complex in a different way, much semantic meaning can be lost in this transition. Because libraries' complex discovery environments depend on the complexity of the MARC format for many of their features, this new information fits poorly into the data store and new discovery features, such as faceting, become corrupted.

In a relational database, which is a closed system with all data contained within its indexes and related to specific fields within its environment, discovery relies on the fields defined in the system. Similar elements, such as author, are grouped by definition of the format in which the data were ingested. These fields are predefined and all data must fit into them. A user can explore only what is contained in a certain field or relationships defined by the records' construct, but cannot leave the environment to make deeper connections other than by the few links explicitly encoded in the records.

Practical ILS complications also are in play. The cost of an ILS system often depends on the size of the database that it must handle. This explosion of records can push libraries

into a costly size upgrade of their system. In an effort to embrace as much data as possible to compete better with the web, libraries are twisting their current environments to behave in ways they never were meant to act. Ultimately, the disconnection between their inherent structure and expectations will lead to a breakdown.

In contrast to a relational database, a linked data environment has no bounds. It is open and dynamic. Data are continually added as more and more RDF stores are exposed for harvesting and linking. In addition, a linked data environment is an interactive one. Users can make persistent links of their own and correct faulty links made by others. The system grows, enhances, and corrects itself with use. Work done by scholars in one part of the world becomes available to everyone as their data, and corrections to existing data, are recorded as RDF triples and are made available to linked data stores. This free and open exchange of data becomes the basis of a global discovery environment. Silos of information, frozen in separate environments, countries, and formats become available to all.

Bibliothèque nationale de France

An impressive demonstration of what is possible using RDF to fashion a discovery environment has been created by the Bibliothèque nationale de France (BnF) (data.bnf.fr). This platform pulls together information on major authors from various data sources worldwide. In addition, all the data found are published as RDF triples and made available to others under an open data license.¹⁴ More than six million triples are available for harvesting. For many authors, a tremendous amount of information is collated, including biographical information, editions and translations of their works, illustrators of their publications, works about the author, musical settings, biographical sketches, holdings from the BnF archives, and more.

The BnF clearly explains the data model.¹⁵ Not only are directions on how to understand and use the data presented, but a description of the data itself is made explicit. The BnF emphasizes the use of existing, registered schemas and vocabularies such as RDF, Simple Knowledge Organization System (SKOS), DC, Friend of a Friend (FOAF), and the various RDA vocabularies to foster interoperability. The result is an elegant presentation of a wealth of resources concerning a particular author and the availability of the data in RDF triples for others to make use of or build on under an Open Data license.

As an example, figure 2 shows the beginning of the entry for Edgar Allen Poe as of February 2012. Certain elements in the display are generated automatically and can vary with time. By selecting "Liste des auteurs" and "P" in the alphabetical display, a user may browse to the entry for Poe.

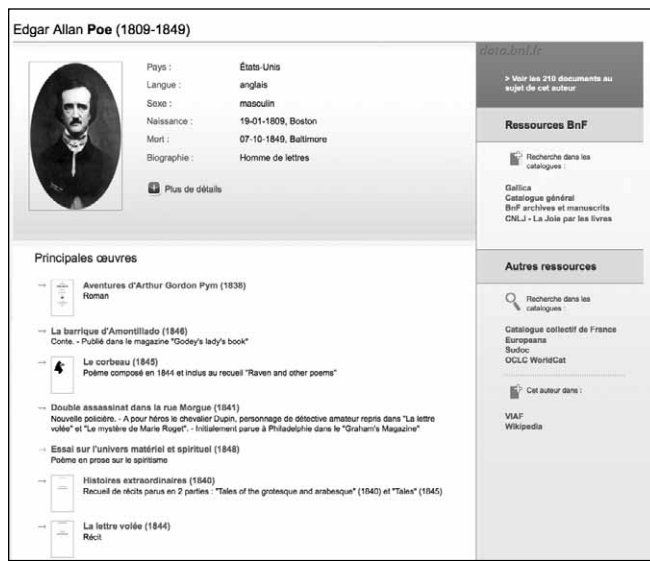


Figure 2. Beginning of Edgar Allen Poe Entry

The entry begins with brief biographical information and a portrait. Beneath the author's entry is a list of Poe's major works and then a list of all of his works broken down by field of activity (e.g., author, librettist, etc.). To the right are listed additional resources at the BnF itself (Gallica, Catalogue général, BnF archives et manuscrits, Centre national de la littérature pour la jeunesse), other resources on the web (Catalogue collectif de France, Europeana, Sudoc, OCLC), and other sources of biographical information (Virtual International Authority File (VIAF), Wikipedia).

Of special interest might be Poe's poem, "The Raven." By selecting this entry, the user navigates to a page devoted to this work; see figure 3. On this page is the original language, date of creation, subject headings, and original title. Beneath it, all the editions of the work as well as sound recordings of the work are listed. By following the links to the right, the user can pursue the holdings in Europeana where one discovers illustrations from "The Raven" by Édouard Manet and a videorecording of a lecture on the works of selected poets including Edgar Allen Poe. By following the additional links provided in Europeana, one can find other topics to explore, such as an image of the Poe family headstone courtesy of the North Ayrshire Council Museum Service and images of other famous illustrations of Poe's works from collections across Europe. By following the links to Gallica, one can find and access digitized editions of "The Raven," including the one translated by Mallarmé and illustrated by Manet in 1875, articles on Poe, works discussing his psychological disposition, correspondence of Mallarmé on Edgar Allan Poe, and so on.

Returning to Poe's page (figure 2), 637 entries follow his name. They include, as one would expect, editions and



Figure 3. Page Devoted to "The Raven" ("Le corbeau")

translation of his works. Also included in the list are musical settings of his works by such composers as Darius Milhaud (*Les Cloches*), a facsimile of the autographed manuscript of *Tales and Songs from the Bible of Hell* by Henri Pousseur, and works based on Poe's stories such as the unfinished opera *La Chute de la maison Usher* by Debussy and a horror movie *Le Vampire et le sang des vierges* based on *The Pit and the Pendulum*. A tremendous amount of information concerning Poe, his works, and how they have been used and adapted has been pulled together from various silos of information and brought together for the first time.

One of the most fascinating academic additions is the link to the BnF Archives et manuscrits. Through it, one can find archival materials related to Poe and his works. One such archive is that of Robert de Montesquiou, a famous personality in his own right at the end of the nineteenth century, with social relationships with Bernhardt and Cocteau, among others. His collection contains two portraits of Poe. Another connection appears in the archive of Antonin Artaud, a French playwright, poet, actor, and theater director. His collection includes a partial translation of Poe's *Annabel Lee*.

Two aspects of the future of discovery in a linked data environment are essential in this extraordinary site. The first is linked data's ability to pull together related information that has been kept in separate silos. In the Poe example, one sees traditional bibliographic information concerning many of Poe's works and use of his works, links to the museum community's objects concerning Poe (the Poe family headstone), additional holdings of works related to Poe from across Europe (Europeana), digital versions of his works, and archival holdings from the BnF related to Poe. The amount and depth of information available to the user is amazing. Equally important is that all this information is available to others to use and upon which they can build. The metadata model behind the data has been clearly and publicly articulated and standard, international vocabularies are used for data interoperability. The BnF made the data available freely under an open license agreement so anyone can easily ingest the data and build additional discovery tools on top of it. As more and more sites publish their data

as RDF triples and make them available, the global web of data becomes richer and richer.

Fundamental Change

Fundamental changes to the basic functionality of the library in an academic environment are inevitable as the shift toward linked data occurs. Core functions such as collection development, preservation, cataloging, reference, and patron services will evolve as the idea of an information ecosystem takes hold and libraries find their place within it.

The concept of the collection and what libraries should collect has been under discussion for several years. These responsibilities, then, become more complex as libraries try to capture and preserve not only the entire output of the academy but also those materials in which academy members are interested for their work. The act of curation will expand from the process of selection in a limited world of resources to include the collation of linked data nodes in particular subject domains. Models that have been based on bibliographic records are coming under elevated levels of stress lately. As libraries make the shift to a new paradigm, they must thoughtfully reconsider what it is they do and not carry over what they have done in the past simply because the model is familiar.

Perhaps the greatest shift will happen within library's technical services departments. Much time and effort has gone into the creation and exchange of bibliographic records. These records are expensive to create and to maintain. A tremendous and costly redundancy results because they are stored at the network level and at the local level, often with significant local variations. Because of the cost of these records, they can be created for only a small percentage of the materials the library owns. As the question of what libraries collect shifts toward what they provide access to, this model of record creation and exchange becomes insupportable.

If libraries are to provide access to both the output of the academy and all the data of which it makes use, they will not be creating metadata surrogates for every resource. In addition to more traditional bibliographic records, they will need to harvest and store data as they are created by the patrons themselves. The only thing on which libraries can count is that these data will have no consistency.

Libraries will always need a database of defined library assets. Orders for items will need to be placed, items that are purchased will need to be tracked financially to the satisfaction of the university, physical items will need to circulate, and collections will continue to move both within the library's facilities themselves and to remote storage. For all of these activities, clear, accurate, and enduring records will need to be maintained. The area of discovery is where real change will take place.

Librarians have been very forgiving in the area of bibliographic description. Many catalogs contain bibliographic records created according to the cataloging rules in place at the time the item was cataloged. Records created according to any of these rules sets may come in many different variations from brief, to core, to full, to completely locally defined. Libraries may favor records created by the most authoritative institutions (e.g., the LC) or according to a consistent, demanding standard (e.g., PCC) but, will accept nearly any type of record either as a permanent representation of the object or with the hope that someone, somewhere will upgrade it at a later date. Brief records created by vendors and record loads from other countries further complicate the mix. What remains of unquestionable benefit, however, are the controlled access points associated with these records.

Controlled access points are divided into two main areas in bibliographic records: authoritative headings for names and titles, and subject terms selected from recognized thesauri such as LCSH. Each has a primary position in the world of linked data but each presents unique problems that will have different solutions.

Authority Control

Through the use of authority control, works by a specific individual may be pulled together under one heading, for instance, whether the author writes under various pseudonyms (e.g., Samuel Clemens/Mark Twain) or appears under various forms of names (e.g., Philip Schreur, PE Schreur, P. Evan Schreur). This basic principle of collation has been a given in library catalogs and will remain a basic tenet of linked data. How it will be achieved, however, is less clear.

A first option is through crowdsourcing. An individual may not know that Samuel Clemens and Mark Twain are the same person, but someone certainly does and may link these forms of name at any time. Not only individuals, but services such as Sameas.org or Freebase (www.freebase.com) may make these associations as well. Not all the associations made may be correct, but the associations are open to review and, over time, the links between statements becomes more and more consistent. According to this model, creating and publishing the raw data to the web as soon as possible is better than waiting for an authoritative version of it to appear, as libraries have done with bibliographic record creation in the past.

A second option would be to create an entry for the person in some recognized, central registry. By doing this, the effort needed to create this heading would be necessary only once, and from that point on this heading would become the link with which other data could be associated. This approach has downsides, however. First, many different registries for names exist: various national authority files, domain specific data centers such as Mimas (mimas.ac.uk) or registries such

as Open Research and Contributor ID (ORCID) (<http://about.orcid.org>) and growing international database made available through the VIAF (<http://viaf.org>) and International Standard Name Identifier (ISNI) (www.isni.org). A person's name, however, may be registered in any one or several of these systems. If librarians expect to harvest data created by authors themselves, authors cannot be expected to search each database to see if the name they wish to use already exists. Given the amount of data that the library will be ingesting, catalogers cannot be expected to establish this number of additional headings by traditional means.

Two shifts in practice would make this option more practical to implement. First, the rules for the form of entry in many of these databases need to change. In the past, these headings had to have a unique text string; it was the string itself which provided the differentiation. The ability to make these unique strings was difficult to teach and limited to a subset of catalogers. As linked data nodes, this is no longer necessary. The only requirement is that enough information is in the authority record itself so that a correct link can be made. Headings themselves might be identical. This change would open the door to almost anyone to register a name or for names to be registered automatically in the process of data creation.

Second, the various name registries must be linked to each other to provide an interlaced web of names. In this way, simple name registries could be established for particular domains or geographic locations (e.g., ORCID or the LC Name Authority File), and the links between these registries could be created to join names at a global scale. Various databases and registries (e.g., VIAF, ISNI) are attempting to do this now and the preliminary results are promising. Undoubtedly, no single approach will resolve this problem and a combination of both is the most likely to occur.

Subject Access

Subject access is by far the more difficult area to handle. Subject analysis has always been considered the most professional aspect of cataloging. A cataloger must not only assign appropriate terms from a defined subject thesaurus—e.g., LCSH, Art and Architecture Thesaurus (AAT)—but must apply them consistently across all languages and historical periods. Subject analysis poses two great challenges with linked data. First, a wide variety of thesauri are used internationally, e.g., LCSH, and Répertoire d'autorité-matière encyclopédique et alphabétique unifié (RAMEAU), and subject domain thesauri such as ATT. Second, the assignment of the subject terms is an intensely human process. The cataloger must not only be familiar with the subject area and language of the resource, but also be intimately aware of the thesaurus itself to assign the most appropriate terms. By applying these terms consistently, subject content is linked

across language families, historical periods, and formats.

To transition to a linked data environment, as with names, the various subject thesauri must be linked. In this way, parallel topics assigned by various thesauri can be linked in an automated way and subject content pulled together across subject and geographic domains. The LC has already accomplished preliminary work in this arena with the linking of LCSH and RAMEAU.¹⁶ The automatic assignment of subject headings is by far the more difficult goal to achieve. Preliminary attempts have been made by organizations such as HighWire Press in which subject assignments for journal literature are made through semantic analysis and automatic taxonomic indexing.¹⁷ As libraries move away from the exclusive creation of records to capturing additional data at the point of origin, skilled technical services staff will be using the same techniques that they have employed in the analysis of individual resources to develop and maintain systems capable of handling the automated assignment and reconciliation of controlled terms within linked data environments.

Paradigm Shifts

*The word is about, there's something evolving,
whatever may come, the world keeps revolving
They say the next big thing is here,
that the revolution's near,
but to me it seems quite clear
that it's all just a little bit of history repeating¹⁸*

This shift to a linked data model could affect every aspect of the knowledge environment, creating an ecosystem in which to work and share information. One should note, however, that linking data are not a new interest in libraries. Through the use of controlled headings and MARC linking fields, libraries have always tried to link related information. By enabling their discovery environments to accept more than MARC records and utilize the power of linked data, libraries will be able to digest a much larger part of the information world. The RDF model is simply a technique for doing so. Libraries could redo what they do now using RDF, but to do so would simply be “all just a little bit of history repeating.”¹⁹ The importance of linked data as a technique is not in its ability to allow libraries to do the same things they do now in a new way, but rather its importance is the paradigm shift it allows libraries to make. For this shift to take place, four crucial changes are needed in the way data are created and managed:

First, this linked data ecosystem presupposes a free and open exchange of data, not a restricted exchange of records. Data becomes something that is shared and built on, not a commodity itself. The current model of academic

bibliographic record exchange is filled with limitations. Vendors charge for specialized access to enhanced content. If access to information is purchased, its use and reuse is restricted.

The paradigm shift to free and open data exchange is happening already. In September 2011, the Conference of European National Libraries (CENL) voted overwhelmingly to support the open licensing of their records.²⁰ Eventually all of the bibliographic records that these forty-six nations produce will be available for re-use for any purpose under a Creative Commons Universal Domain Dedication, or CC0. The first data available will be the data these nations supply to Europeana, but the rest will be made available as well. As mentioned earlier, the BnF is making their data available under open license. In February 2012, the Ontology Engineering Group (Spain) announced the launch of their linked data initiative built on the library data of the Spanish National Library.²¹ This initiative will publish both bibliographic and authority data in an RDF triple store under a CC0 license. At the time of this writing, they have published approximately 2.4 million bibliographic records and 4 million authority records generating 58 million RDF triples. The German National Library also has set up a Linked Data Service where it plans to distribute its entire stock of bibliographic and authority data free of charge for noncommercial use.²² In July 2011, the British Library announced the release of a new approach to publishing the British National Bibliography (2.8 million titles) according to a data model developed in consultation with Talis.²³ OCLC is making its experimental linked data available under an Open Data Commons Attribution License.²⁴

The second important change is the shift from the creating, maintaining, exchanging, and storing records to the linking of individual statements. RDF is a model used for the creation of individual statements. These individual statements can be linked to each other, creating a dynamic web of data far richer than anything seen previously. Because these are statements, not records, patrons are able to finally reach into documents to discover links at the chapter or even paragraph level. All the data that have remained hidden to discovery, on the web or in library catalogs, are exposed. Given the cost of cataloging, accessing these data any other way would be impossible. The transition will be a dramatic one. Businesses and environments designed around the exchange of records will eventually fade unless they revise their business and data models.

Third, libraries must move from the exclusive creation of records to the inclusion of data captured at the source. Currently, libraries have a talented cadre of catalogers creating records for individual bibliographic resources. As libraries come to embrace the entirety of information production, this model cannot stand. As these people shift their attention to broader issues of authority control and controlled

access, another solution is needed for creating the additional data libraries wish to use. The only solution will be to have the act of data creation itself generate the RDF triples that can be used to link to other information in the web of data. Researchers cannot be expected to create standardized information and controlled headings as library catalogers have done in the past. New, automated means of data capture and enhancement will need to be developed. The world of linked data are a self-improving one as people make use of the data. Discovery exclusively based on highly structured records created by catalogers will shift to one based on a more heterogeneous environment with quality assured through a combination of automated processes and iterative use.

Finally, libraries will need to focus on effectively managing statements in triple stores, not on adding more records to the catalog. In the current environment, libraries are continually searching for more resources. They put enormous effort into adding non-MARC records to their discovery environments. Finding aids in EAD, records from digital repositories in DC or MODS, even article level metadata by the millions are forced into these environments. Unfortunately, what makes these environments work so well is the complexity of the MARC format and the sophisticated work on controlled access points (names, subjects, etc.) that libraries have spent many decades creating and maintaining. Faceted search is one of the most notable of these recent developments. It has opened the discovery experience in unmatched ways. However, these facets must appear in the records in the database and the data must be entered in a clear, controlled way. As new, less controlled records are added to this environment, techniques such as faceting produce poorer and poorer results. This creates an endless cycle of record loading and optimization. The focus is on a limited set of records with the best, precoordinated headings possible.

With linked data, the problem is exactly the opposite. The amount of data will be overwhelming. Libraries will move from a closed, relational database to something nearly infinite. Applications will need to be developed that will pull together information into usable domains, i.e., subareas of knowledge that can be exploited at the discretion of the individual patron. Often called the "killer app," this development will be central to linked data's acceptance. Without it, the graph of data will become impossible to navigate.

Conclusion

Moving to a linked data environment, the model discussed throughout this paper, has the power to completely alter the way academia creates, maintains and explores data. The entire work pattern of the academy will be changed. The dependence on creating metadata surrogates for discreet resources and associating them within a relational database

for discovery will be a thing of the past. Industries built on creating, maintaining, and sharing of these records will need to radically reinvent themselves. Programs devoted to the creation and standardization of these records will need to vastly expand their scope.

The first steps in the paradigm shift are evident. The LC is spearheading an effort that will replace the MARC formats as the means of data communication and will be embracing the shift to a linked data model. National libraries are publishing their bibliographic records as RDF triples and are making them freely available under a CC0 license. Applications such as those developed by the BnF or Linked-Sailor (linksailor.com, a linked data browser) are already exploiting this freely available data.

Much remains to be proved, however. Services must be developed that will allow RDF triples to be generated as an automated by-product of the work of the academy. Notions of data enhancement and correction in an academic setting by crowdsourcing have to be demonstrated. A killer application based on linked data principles that can replace current discovery methods has yet to be developed.

Models such as the Stanford Linked Data Technology Plan attempt to resolve many of these questions.²⁵ As implementation techniques are planned, developed, and adopted by an initial set of institutions, the transition to this new model will gather momentum. The next few years will be critical. True beginnings do not happen often and revolutions can be swift and unexpected. Libraries must be leaders in this revolution. Information creation and exchange is the *raison d'être* of the academy. The time has come for a pivotal change in the entire information ecosystem and libraries cannot afford to let history simply repeat itself.

References

1. Tim Berners-Lee, Linked Data, July 27, 2006, www.w3.org/DesignIssues/LinkedData.html (accessed Feb. 29, 2012).
2. W3C, RDF: Resource Description Framework (RDF), Feb. 10, 2004, www.w3.org/RDF (accessed Feb. 29, 2012).
3. Tim Berners-Lee, Relational Databases and the Semantic Web, September 1998, www.w3.org/DesignIssues/RDB-RDF.html (accessed Apr. 26, 2012).
4. Henriette Avram, MARC: Its History and Implication (Washington, D.C.: Library of Congress, 1975), www.eric.ed.gov/ERICWebPortal/contentdelivery/servlet/ERICServlet?accno=ED127954 (accessed Feb. 29, 2012).
5. Doralyn J. Hickey, "Theory of Bibliographic Control in Libraries," *Library Quarterly* 47, no. 3 (1977): 253–73.
6. W3C, Category: Triple Store, www.w3.org/2001/sw/wiki/Category:Triple_Store (accessed Feb. 29, 2012).
7. W3C, RDFa in XHTML: Syntax and Processing: A Collection of Attributes and Processing Rules for Extending XHTML to Support RDF: W3C Recommendation, Oct. 14, 2008, www.w3.org/TR/rdfa-syntax (accessed Feb. 29, 2012).
8. Library of Congress, Bibliographic Framework Transition Initiative, www.loc.gov/marc/transition (accessed Feb. 29, 2012).
9. Ibid.
10. Library of Congress, News and Announcements, A Bibliographic Framework for the Digital Age, Oct. 31, 2012, www.loc.gov/marc/transition/news/framework-103111.html (accessed Mar. 22, 2012).
11. Council on Library and Information Resources, Linked Data for Libraries, Museums, and Archives: Survey and Workshop Report, www.clir.org/pubs/abstract/reports/pub152 (accessed Feb. 29, 2012).
12. Report of the Stanford Linked Data Workshop, 27 June–July 1 2011 (Washington, D.C.: Council on Library and Information Resources, 2011), 20, www.clir.org/pubs/reports/pub152/reports/pub152/LinkedDataWorkshop.pdf (accessed Mar. 25, 2012).
13. Open Metadata Registry: Supporting Metadata Interoperability, The RDA (Resource Description and Access) Vocabularies, www.rdvocab.info (accessed Feb. 29, 2012).
14. Data.gouv.fr, License Ourverte/Open Licence, www.data.gouv.fr/Licence-Ouverte-Open-Licence (accessed Feb. 29, 2012).
15. Bibliothèque nationale de France, Semantic Web and Data Model, Resources Description Framework (RDF) and the Semantic Web in the data.bnf.fr Project, data.bnf.fr/semantic-web-en (accessed Mar. 22, 2012).
16. Library of Congress, Authorities and Vocabularies, id.loc.gov (accessed Apr. 2, 2012).
17. TEMIS and HighWire Press, "HighWire Press Partners with TEMIS to Semantically Enrich Publishers' Content," Oct. 25, 2011, highwire.standord.edu/PR/TemisHighWirePartnership.pdf (accessed Mar. 26, 2012).
18. Propellerheads featuring Shirley Bassey, "History Repeating," Decksanddrumsandrockandroll, recorded 1997, Wall of Sound, compact disc. The lyrics found at "Propellerheads f/ Miss Shirley Bassey History Repeating Lyrics," ST Lyrics, www.stlyrics.com/lyrics/theressomethingaboutmary/historyrepeating.htm (accessed Feb. 29, 2012).
19. Ibid.
20. CENL, "Europe's National Librarians Support Open Data Licensing," Sept. 28, 2011, <http://app.e2ma.net/app2/campaigns/archived/1403149/f14691f55d5483aff43360a9b4a7d35> (accessed Apr. 9, 2012).
21. Biblioteca Nacional de España, Data Links at the BNE, www.bne.es/en/Catalogos/DatosEnlazados (accessed Feb. 29, 2012).
22. Open Knowledge Foundation, Open Bibliography and Open Bibliographic Data, "German National Library Goes LOD & Publishes National Bibliography," Jan. 26, 2012, openbiblio.net/2012/01/26/german-national-library-goes-lod-publishes-national-bibliography (accessed Feb. 29, 2012).
23. Talis Consulting, "Significant Bibliographic Linked Data Release from the British Library" (July 14, 2011), consulting.talis.com/2011/07/significant-bibliographic-linked-data-release-from-the-british-library (accessed Feb. 29, 2012).
24. OCLC, Linked Data at OCLC, www.oclc.org/data.html (accessed July 9, 2012).
25. Council on Library and Information Resources, Stanford Linked Data Workshop Technology Plan: 30 December 2011, www.clir.org/pubs/reports/pub152/LDWTechDraft_ver1.0final_111230.pdf (accessed May 7, 2012).