Smarter Libraries through Technology

Linked Data Brings Challenges and Opportunities to Libraries

By Marshall Breeding

For a decade or more, libraries have been working toward increased adoption of semantic web technologies and linked data. As early implementors of online systems, libraries created their own set of protocols and metadata formats that were optimized for a technical world that pre-dated the web.

The MARC record format was devised in the early days of library automation. At that time computer storage was enormously expensive and network bandwidth was barely a trickle compared to what is available today. This record structure was extremely compact and designed to carry bibliographic information according to the cataloging practices carried forward from physical catalog cards.

The MARC standards proved to be extremely successful in enabling a global ecosystem of shared cataloging, resource sharing, and systems compatibility. All integrated library systems and library-oriented systems using bibliographic or authority records adopted the MARC formats. As a result, libraries can easily import catalog from many different external sources, receive records with shelf-ready materials, and generally exchange bibliographic records for a host of scenarios. An entire global ecosystem has amassed billions of MARC records,

distributed in central bibliographic services and in local integrated library systems in libraries. Protocols were also defined for automated search and exchange of MARC records, such as Z39.50. It is ubiquitously implemented in integrated library systems, cataloging services, and systems.

While this global ecosystem has proved invaluable to the library community, it also has major drawbacks. These formats find use only in libraries, while the broader information ecosystem is based on web standards, such as XML and JSON. As semantic web technologies became prominent, RDF and other forms of linked data have gained wide adoptions. Adhering to domain-specific standards, libraries became somewhat isolated from the broader web and the growing universe of information encoded as linked data.

In this context, there have been a variety of initiatives to pivot the library metadata ecosystem to the realm of linked data. These efforts have built on each other for the last decade, leading to an increasing mature linked data ecosystem for libraries. In some cases linked data replaces previous metadata formats and in others provides a bridge from MARC-based systems.

The Dublin Core Metadata Initiative was an early step out of an entirely MARC-based bibliographic universe. This effort, launched in 1995, devised a simplified metadata standard more appropriate than MARC for describing digital resources. Dublin Core was based on a finite number of metadata elements. Fifteen elements were defined in the original metadata set, with the option to create new ones to accommodate the needs of specialized communities adopting the standard. Dublin Core has gained wide use well beyond the library community.

Other projects have worked to provide an alternative able to replace MARC in the library bibliographic ecosystem with new structures based on linked data and semantic web technologies. These efforts emerged out of selected national libraries, major academic or research libraries, and OCLC.

COMET (Cambridge OPen METadata) was a joint project between OCLC Research and Cambridge University in the UK, with funding from Jisc to create and release a large set of

bibliographic records derived from the Cambridge University Library catalogs in multiple forms of linked data. This project tested the viability of the multiple linked data technologies to support library bibliographic content.

OCLC has been a consistent proponent of linked data for libraries. The organization has engaged in many projects with a diverse set of collaborators to develop, test, an implement linked data. OCLC has released multiple data products as linked data, including

- FAST (Faceted Access to Subject Terminology), a scheme based on the Library of Congress Subject Headings, relying on the individual components rather than complex coordinated terms. OCLC released the FAST headings as linked data in 2011.
- VIAF, the Virtual International Authority File
- WorldCat Works released as linked data in 2014. 197 million bibliographic work descriptions

The Library of Congress launched a major initiative to create a linked data alternative to MARC through its Bibliographic Framework Transformation Initiative. The aim of this project was to create a new framework for bibliographic information consistent with modern concepts of linked data that captures the content of MARC records, but without the constraints of its legacy formats and encodings. The Library of Congress contracted with Zepheira to collaboratively develop new framework mapped from MARC, now known as BIBFRAME.

Since that initial effort, BIBFRAME has continued to mature through ongoing revisions and is generally positioned as the successor to the MARC formats. It is seeing implementation by many different projects in the US and internationally, including experimental prototypes as well as production environments.

Despite the enduring persistence of MARC, BIBFRAME and other linked data frameworks will be a growing component of the global bibliographic infrastructure for libraries.

These are just some examples of the work that has taken place to move libraries into the realm of linked data. Despite a maturing set of tools, technologies, and viable frameworks, the operational library bibliographic ecosystem remains well entrenched in MARC. With the vast number of MARC-based systems deployed globally and the generally slow cycles of product development, it is difficult to imagine a wholesale shift in the foreseeable future. Recently created systems are naturally designed to accommodate BIBFRAME but must simultaneously support MARC. The massive installed base of more longstanding systems was essentially hard coded for MARC cannot not feasibly be reprogrammed to support BIBFRAME.

Despite the enduring persistence of MARC, BIBFRAME and other linked data frameworks will be a growing component of the global bibliographic infrastructure for libraries. Even though library systems may not soon change to BIBFRAME as their internal bibliographic record structure, linked data plays an important role as a bridge to the web. MARC records can be transformed into BIBFRAME and other linked data syntaxes to achieve benefits such as better representation in web search results or integration with other information services. Also, the internal use of MARC does not preclude the delivery of catalog listings or other resource pages with embedded linked data references, such as schema.org and BIBFRAME. In this way it is possible to both leverage the continuity of MARC as a mature and universally implemented standard and to gain new benefits from linked data.

In this issue of *Smart Libraries Newsletter*, we feature EBSCO Information Services' acquisition of Zepheira, recognized as the leading linked data services firm in the library industry. Zepheira has played an important role for the promotion and deployment of linked data in the library sphere, notably through its work supporting the development of BIBFRAME.

EBSCO Information Services Acquires Zepheira

Zepheira, a company that has played a prominent role in the promotion and implementation of linked data to libraries and related institutions, has been acquired by EBSCO Information Services.

Details of the Acquisition

Following its merger into EBSCO, Zepheira operates as an independent division and will continue to be led by company

co-founder and CEO Eric Miller. A total of eight individuals from Zepheira will become employees of EBSCO, including the core teams involved with product, architecture, and engineering. Within the EBSCO corporate structure, the Zepheira group will report to Gar Sydnor, Senior Vice President, Analytics and Hosting Services. Sydnor leads EBSCO's involvement with FOLIO, User Experience, and its portfolio of SaaS technologies, including EBSCO Discovery Service, and Stacks. Zepheira will work closely with EBSCO's NoveList team, led by Danielle Borasky.

With this acquisition, the Library.Link subscription service from Zepheira joins the EBSCO suite of SaaS products. It will be promoted through the EBSCO sales teams and through ongoing business partners such as SirsiDynix. The Zepheira team within EBSCO will continue to develop and support the service.

With the expertise and vision of Zepheira, we can anticipate that linked data technologies will also be incorporated into EBSCO's product suite, though any specific developments have not been announced. Although Zepheira is a small company, its influence has been substantial. Its acquisition into EBSCO represents an important event in the library industry.

Financial terms of the acquisition were not disclosed.

Common Value in Open Source Technologies

Zepheira developed a suite of open source tools through its linked projects. These include a suite of utilities that together make a toolkit for transforming MARC records into linked data and for creating and editing BIBFRAME entities (https://github.com/zepheira). These tools have been developed and released as open source software. Though Zepheira does not own exclusive rights, it is able to provide value-added services based on its expertise with the tools and the linked data domain. These tools likewise become part of the growing slate of open source projects allied with EBSCO, including the FOLIO library services platform.

Miller stated, "Zepheira found EBSCO to be a good match with its core values, in that both embrace open source software and have complementary approaches to technology and business culture. EBSCO brings the resources to advance many of the concepts and products to the global sphere. An important objective going forward will be making library lending on the web, outside the library's own catalog or website, the new

We can anticipate that Zepheira's linked data technologies will also be incorporated into EBSCO's product suite.

norm. As part of EBSCO, the means to accomplish this vision are within reach." (source: telephone interview March 11, 2020.)

Libhub Initiative

As part of its work to help libraries adopt linked data, Zepheira launched the Libhub initiative in September 2014. Libraries were invited to join this initiative to gain access to educational resources and technologies to explore the benefits of linked data for their organizations.

Library.Link as Core Linked Data Infrastructure

Library.Link is Zepheira's core product and the infrastructure on which it has developed, refined, and populated as it works with libraries to make their collections more available through the web. Its core concept is to channel users to the library's resources even when they search on the web and not the library's own catalog interface. The service amplifies the exposure of a library's collections by exporting its database of bibliographic records and holdings in MARC format, transforming the records into BIBFRAME and other linked data entities, and merging the results into the Library.Link network. The service also creates geolocation data associated with each library location. As Google and other web search engines harvest the Library.Link network, library content becomes better positioned in search results, directing users to the item in the library's own catalog. The geolocation data enables Google and other services to present library borrowing options for libraries in close proximity.

Library.Link is offered as a subscription service to libraries, both directly from Zepheira and through its business partners. Libraries pay an annual fee, which covers the initial transfer of data from the local ILS to Library.Link, configuration of local infrastructure, and incremental monthly transfers.

Innovative Interfaces began its involvement with the Libhub Initiative in March 2015. By June 2015 Innovative began offering a service for an initial output of a library's bibliographic records to Library.Link with monthly updates thereafter. Innovative no longer offers this product directly, though existing customers continue to receive support and renewals through Zepheira.

SirsiDynix entered into a partnership agreement with Zepheira in June 2015. By January 2016 SirsiDynix launched

its BLUEcloud Visibility product based on Library.Link. This product provides an administrative interface within BLUEcloud Central for initializing the harvesting of bibliographic records and selecting configuration options. Libraries can integrate additional information, such as library hours and geolocation data for branch locations. The partnership between SirsiDynix and Zepheira for BLUEcloud Visibility remains in effect.

EBSCO Information Services has been a longstanding partner with Zepheira and its linked data services. In the five years preceding this acquisition, EBSCO has collaborated with Zepheira, supporting the Libhub Initiative and providing a linked data integration service through NoveList.

EBSCO joined the Libhub Initiative as a sponsor in October 2015 and began working with Zepheira to create integrations between NoveList and Library.Link. In September 2016 EBSCO released NoveList Select for Linked Data service. It builds on the basic Linked Library Service by layering additional content to enhance the linked data associated with a library's collection. In the same way that a library can subscribe to NoveList Select to enrich the presentation of records in its online catalog, NoveList Select for Linked Data adds additional entities that aim to increase the opportunities for library collection items to be discovered through web search.

The NoveList Select for Linked Data service also enables libraries to present library content, such as book carousels, on web resources outside of the library's own website. Libraries can create targeted collections that can be placed on community pages for book clubs or any other organizations or events. The code snippets for implementing the functions are provided through a subscribing library's administrative console.

The Library.Link service has seen its largest adoption among public libraries. These libraries seek to engage users beyond the confines of their physical facilities, websites, and catalogs. Amplifying their visibility on the general web aligns well with their mission. A number of major academic and special libraries have also worked with Zepheira to integrate their collections via the Library.Link network.

Library Lending on Google Knowledge Panels

The vagaries in the appearance of library collection items in Google search results recently made a giant leap forward. Beginning in about Fall 2019, Google enhanced its Knowledge Panel for books to include library lending in addition to purchase options (see figure 1). This feature relies on its Knowledge Graph database that supplements the standard Google indexes. Library content in the Knowledge Graph comes from

Figure 1: Library locations in Google Knowledge Panel
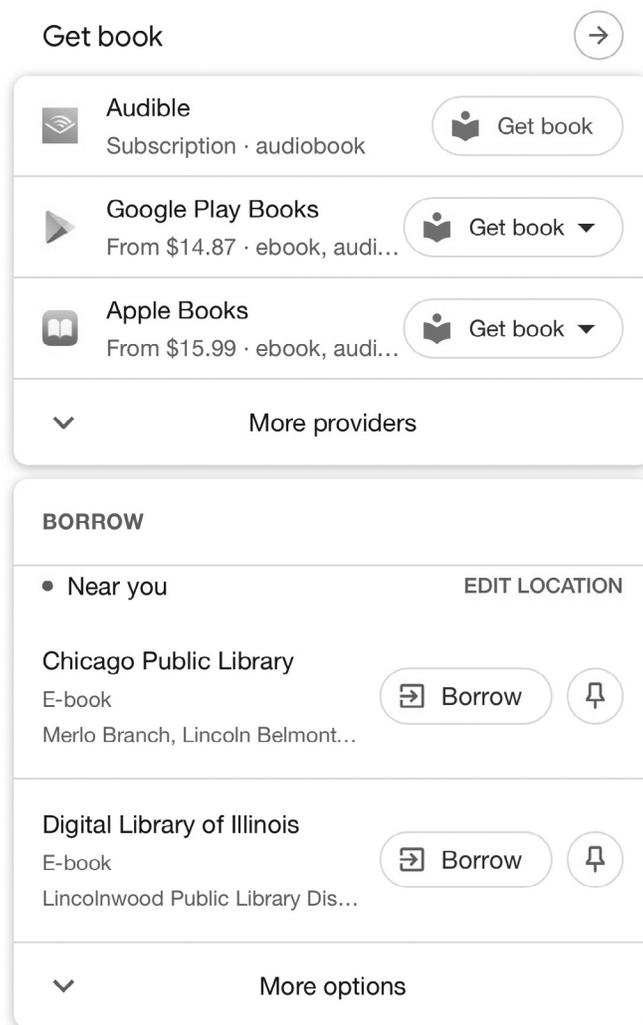
a variety of sources, including links and structured data within the library's website and linked data repositories such as Library.Link. The lending options are based on nearby libraries that offer a copy of the title. Selecting a library from the available options links to its online catalog where it can be requested.

Benefits Beyond Web Search

According to Miller, the vision of Library.Link and its other services goes beyond exposure of library catalogs in web search results. According to Miller, the key intent has always been on creating actionable data using web standards that can be applied to many different types of use.

An example of such use can be seen in Zepheira's collaboration with the Internet Archive to facilitate its service for Controlled Digital Lending (CDL) (https://controlleddigitallending.org). This service is based on the premise that a digitized book can be lent under the same conditions as a print book. As the Internet Archive digitizes physical books, it can allow public borrowing of that ebook by one borrower at a time, as long as the original book is withdrawn from circulation. If a library subscribing to Library.Link wants to contribute to the Internet Archive's CDL program, it can have their holdings automatically transmitted and processed to determine copies that could be transferred from local to digital lending. This process would match to titles already digitized, so no additional scanning would be necessary. Each copy of a title that library transfers increments the count of simultaneous borrowers allowed.

An Acquisition of Talent

The acquisition of Zepheira brings a team of leading experts and proponents of linked data into EBSCO, Miller in particular brings important expertise about the domain of linked data to the company. This infusion of linked data expertise will help shape EBSCO's ongoing development of metadata and indexing products and its expanding set of technology products and services.

Eric Miller has a distinguished career in metadata and linked data initiatives. From October 1988 through 2000, Miller served as a Senior Research Scientist for OCLC and was one of the co-founders of the Dublin Core Metadata Initiative launched in 1994. Dublin Core has been adopted broadly as the metadata format for the description of digital collection materials, as well as a wide range of other internet-based resources. Following his departure from OCLC, Miller worked as a Research Scientist at Massachusetts Institute of Technology in the Computer Science and Artificial Intelligence Laboratory and led the MIT SMILE project to develop open source tools for semantic web technologies. (source: LinkedIn). During this period Miller also served as the Activity Lead for the Semantic Web Initiative of the World Wide Web Consortium (https://www.w3.org/People/EM).

Miller co-founded Zepheira in February 2007, along with Kathy MacDougall and Uche Ogbuji, as a business to deliver products and services to libraries and related organizations based on linked data concepts and technologies. Since its founding, Miller has served as Zepheira's Chief Executive Officer.

Zepheira Milestones

- February 2007. Zepheira founded by Eric Miller, Kathy MacDougall (left Zepheira in July 2018), and Uche Ogbuji
- July 2007. Zepheira engaged by **OCLC** for assistance in the redesign of its Persistent URL service. (PURL). The software created for this service was created as open source. This service, created by OCLC in 1995, was transferred to the Internet Archive in September 2016.
- May 2012. The **Library of Congress** initiates contract with Zepheira for assistance in mapping MARC bibliographic formats into linked data through their Linked Data Initiative. As part of its work with the Library of Congress, Zepheira developed a set of open source tools for BIBFRAME transformation and related tasks.
- May 2014. The **National Library Board of Singapore** engages Zepheira for services related to the transformation of their collection metadata into linked data.
- September 2014. **National Library of Medicine** engages Zepheira to develop its linked data architecture.
- September 2014. Zepheira launches Libhub Initiative to help libraries learn about and implement linked data.
- 2015. Zepheira initiates a series of partnerships with ILS vendors. Subscription service for libraries enables export of catalog records for transformation into linked data and exposure through Library.Link network.
- February 2020. Zepheira acquired by EBSCO Information Services

Smart Libraries Q&A

What changes can we implement to ensure our catalog and discovery services are protecting our patrons' privacy?

How can we help ensure our patrons' confidentiality in regard to reader privacy, reading statistics, and digital information access?

Safeguarding the privacy of patrons as they use library resources is one of the basic values of the library profession. In the context of library catalog and discovery services protecting privacy involves several different areas of concern.

A typical session that includes a patron using a library search tool must be treated with the same degree of care as the data related to the borrowing of a physical item from the library. The need to protect circulation records is well accepted and the related technical and operational processes are followed by almost all libraries that lend materials to the public. When an individual searches for content on a library catalog or discovery interface, it involves personal data of equal or higher sensitivity than circulation. The network address, browser cookies, or other identifiers can easily be resolved to the personal identity of the searcher. The text entered into the query box, identifying content of interest, is transmitted across the internet from the web browser to the servers processing the search. The search results returned to the user and items selected, links to resources, or even specific items read or downloaded become part of a package representing a very sensitive transaction between an individual and the library.

Given the sensitivity of these data, there are steps that should be taken to ensure patron privacy. The first and most fundamental action is to use encryption. If the transaction is not encrypted, it can be captured by unknown third parties using readily available network eavesdropping software or equipment. Configuring the service to use the HTTPS protocol provides end-to-end encryption that cannot be penetrated. All ecommerce sites depend on this protocol to protect

I would recommend that libraries regularly perform a technical review of their systems to assess the treatment of personally identifiable data.

credit card numbers and other sensitive financial data. Configuring library websites or the servers running catalogs and other search services today should be considered an essential requirement. Chrome and other web browsers currently display prominent warnings for any site that continues to run the unencrypted HTTP protocol. I have been tracking the use of HTTPS vs HTTP on library websites for the last several years. Recent scans of all of the public library websites in the US reveal that about 15 percent still use HTTP; about 6 percent of academic libraries use HTTP.

Other measures can be taken to protect data possibly stored on the search service and falling within the bounds of patron privacy. Almost all search services create logs or other types of records for each transaction. These logs support important functions such as statistical reporting and analytics. To ensure privacy, it is essential to anonymize these records. This can be accomplished by truncating IP addresses to identify only users' network or domain and not a specific device. It is important to ensure that all copies of the transaction be anonymized, including raw web server logs in addition to transactions captured in databases within the application.

I would recommend that libraries regularly perform a technical review of their systems to assess the treatment of personally identifiable data within their internal systems. This review would include inspection of log files, backups, and application logs to verify that they are handling these data elements as expected. This privacy review could be incorporated into periodic reviews that the organization would perform regarding its backup and disaster recovery processes.

Many search applications enable logged-in users to view their history of searches or items checked out or downloaded. This feature inherently stores data that associates content use with a specific user. Such data could be accessed through a security breach or a legal demand. At a minimum, patron profile data and associated transaction histories should be encrypted as a barrier to access by an unauthorized intruder.

Given the vulnerability possible with storing content items in a patron's profile, libraries can limit this feature to individuals that give specific agreement. Many library services will have

opt-in or opt-out options controlling the storage of this type of data. An opt-out by default policy would ensure that data is not saved without patron permission. This issue requires alignment between a library's privacy policy and the technical configuration of all the applications within its service portfolio.

Many, if not most, of the search services offered by the library will be implemented on the technical infrastructure provided by an external vendor. The major indexed discovery services such as EBSCO Discovery Service, Ex Libris Primo and Summon, and WorldCat Discovery services are almost always deployed this way. Socially oriented services such as those from BiblioCommons tap into an even greater set of personalize data, likewise hosted on vendor-provided infrastructure. In these cases the library must work closely with the vendor to ensure that the technical operation of the service matches their expectations in regard to the treatment of personally identifiable data, opt-in or opt-out retention of search history; and that the vendor's privacy policies and the technical behavior of the system matches the library's own privacy policies.

Limiting the collection of personal data can be counter to the interest of the library in delivering personalized services and in performing detailed analytics on the usage of its services. Expectations regarding these capabilities are set by the commercial environment that puts massive effort into extracting all possible personalized data from both online and in-person activities. Libraries, consistent with our distinct interest in protecting private data, cannot necessarily replicate the full extent of personalization and targeted marketing seen in the commercial arena. It is possible, however, to build effective services based on anonymized data, category and demographic data, as well as opt-in personalized data. This difference in values means that any marketing and analytics services used by libraries needs to be built around a different set of assumptions than those developed for the commercial arena. That requirement does not necessarily mean avoiding commercial customer relationship management or marketing engines but populating them and using them in ways that respect the library's privacy policies and practices.

Questions or suggestions for topics in future issues?

YOUR COMMENT COUNTS!

Contact Patrick Hogan at **phogan@ala.org**

ALA TechSource
alatechsource.org

April 2020
Smarter Libraries through Technology

Smart Libraries Newsletter

Marshall Breeding's expert coverage of the library automation industry.

TO SUBSCRIBE

To reserve your subscription, contact the Customer Service Center at **800-545-2433, press 5 for assistance**, or visit **alatechsource.org**.

ALA TechSource purchases fund advocacy, awareness, and accreditation programs for library professionals worldwide.