New Uses for Knowledge Bases

The usefulness and ubiquity of knowledge base data in large research and academic libraries has led to much innovation in this space. While knowledge bases have evolved continuously since their introduction, the past five years have seen increased engagement with these tools across the field, leading to a number of exciting developments. These include new thinking about how knowledge bases are structured and the data they collect, increased used of APIs to integrate knowledge bases with new services, and trends toward greater automation and customization.

Enhanced Knowledge Base Data

The type of data found in a knowledge base has remained fairly consistent since the earliest implementations. Titles, packages, and holdings, along with their associated attributes, remain the core data elements. Recently, however, some knowledge base suppliers have begun to rethink the basics and explore enhanced data models for their knowledge bases. Two major efforts in this area include the development of a re-architected knowledge base by ProQuest and exploration of an enhanced, librarian-driven data model by the Global Open Knowledgebase (GOKb).

ProQuest's New Knowledge Base

In late 2015 ProQuest announced a new knowledge base designed to enrich its existing service, which has its roots in the original Serials Solutions knowledge base dating back to 2001.[1] Yvette Diven, product manager lead for management solutions at ProQuest, described work in four key areas for the company's knowledge base: scope, scale, systems, and services.

The scope of the ProQuest knowledge base will become more global and diverse through the inclusion of new electronic content types, including streaming audio and video titles. The new knowledge base also will feature a single data model that pulls together the traditional e-resources metadata, along with the contents of Ulrich's Periodicals Directory, authoritative information from MARC records, and article-level metadata from the Summon discovery index. The reengineered product will live in the cloud, making it more scalable. And an API will allow ProQuest to reuse this enriched metadata across all of its products and services and to share the data more widely with its customers.

Diven said that these changes will give ProQuest's customers a comprehensive view of their collections from within a single integrated product. The enriched knowledge base will also map the relationships between entities in a more sophisticated manner—making connections between, for example, an author and a title, an organization and the resources it publishes, and two journals published by the same entity. The knowledge base will also have the ability to track changes, helping users manage title and publisher changes and allowing them to see snapshot views of their collections over time.[2]

ProQuest has been gradually rolling out these changes across its existing product suite, and Diven describes the process as more of a continuum than a migration. The enhanced knowledge base was planned to be a major component of Intota, ProQuest's developing LSP. In a recent webinar, the company announced that the vision for Intota, including use of the new knowledge base, instead would be rolled into Alma, the LSP ProQuest recently took over with its acquisition of Ex Libris.[3]

GOKb's Enhanced Data Model

(Full disclosure: I am the principal investigator of the Global Open Knowledgebase (GOKb) project, and any uncited information regarding the project in this section comes from my personal experiences.)

GOKb is a community-managed, open-source project that aims to make e-resources metadata freely available to the library community. Like ProQuest, the GOKb has been innovating in the knowledge base space by addressing the data itself. The GOKb data model has been designed with the flexibility to model a complex environment and the transparency to work openly through a community contribution model.

The goal of creating a data model that can handle the current electronic resources landscape, as well as expand to accommodate changes in the market, led the GOKb development team to adopt the bill of materials (BOM) approach. Used widely in industry, the BOM model labels individual items as *components*, which can be bundled together into *combinations*. New component and combination types can always be created, and combinations can even be linked together to form larger combinations. In the current GOKb environment, three components—titles, packages, and platforms—are linked together to form a combination that represents a holding. But if, in the future, the knowledge base needs to accommodate, for example, article-level metadata, the model can easily accommodate this expansion. Articles can simply become a new type of component, bundled together into journals, which now become combinations.[4]

Like ProQuest, GOKb is working to track changes over time, including title changes and transfers between publishers. Using the BOM model, GOKb allows users to create linkages between two titles to represent a change. All of the titles linked together in this way can be pulled together to represent a comprehensive title family. Similarly, the BOM model also allows a linkage to be created between a title and the organization that publishes it. For any title, users can view all of the title-publisher linkages, along with associated dates, to see a comprehensive publication history.[5]

As a community-managed knowledge base, GOKb has also taken the unique step of building transparency into its data model. Since project partners from many different universities have a role in creating and maintaining data, it's important for users to be aware of who's doing what. To this end, GOKb has included fields in all of its components that allow users to see who has last updated that record. For core components like packages and titles, GOKb also includes several additional status fields. These include the name of an individual verifier, last verified date, and an approval
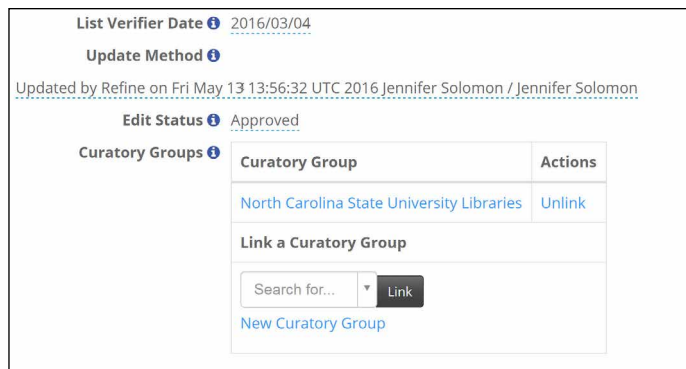

Figure 4.1
A GOKb package record contains information about the group responsible for its maintenance and the individual user who last edited it.

status (see Figure 4.1). Packages can also be assigned a curator—an institution that has claimed responsibility for managing that particular group of titles. While users outside the curator group can still edit the package, the system provides a warning message and encourages users to communicate with a curator before making major changes. Taking inspiration in part from the model used by *Wikipedia*, the goal of these fields is to encourage communication and trust between the users of GOKb.

Knowledge Base APIs

Application programming interfaces (APIs) are sets of tools for building and interacting with software applications.[6] In recent years, many types of library systems and services have begun offering APIs that allow users to build their own integrations with a vended product. Knowledge bases are no exception to this trend. OCLC currently offers an API for the WorldCat Knowledge Base, and both ProQuest and Innovative are planning to introduce them soon as part of their knowledge base enhancements. These APIs are beginning to give knowledge base customers the flexibility to create custom solutions using knowledge base data.

OCLC's WorldCat Knowledge Base API is available in a sandbox version to anyone, but in production only to libraries that use the knowledge base. The API can provide article, e-journal, or e-book citations; links to e-resources customized with a user's account identifiers; proxy information; and browse and search features similar to an A-to-Z list.[7]

Brian Cassidy, senior web developer at the University of New Brunswick (UNB), shared some details about his library's use of the WorldCat Knowledge Base API to create a custom discovery tool. The library's website features several search tabs for different types of e-resources, including databases, journals and

newspapers, online reference works, e-books, and videos. All of the results returned by these searches are drawn straight from OCLC's knowledge base via the API. Users can search for specific titles, browse lists of collections, and link out to their desired resource (see Figure 4.2).

Very little of this functionality came out of the box, Cassidy said, but rather was all designed in-house by staff at UNB. To use the API, UNB's system provides OCLC's API with a web service key that authenticates it as a valid user and authorizes the API to release customer information. The UNB website makes web requests to the API and receives JSON or XML code in return, which it can then use to create the custom search environment. UNB is also preparing to integrate its custom search with OCLC's WorldShare License Manager, which will facilitate the display of permitted uses along with the search results.[8]

Since the WorldCat Knowledge Base API is fairly new, it will take time before more libraries can experiment with the functionality and discover new ways of using it. Stephanie Doellinger, section manager for data services at OCLC, said that creating homegrown *A-to-Z* lists and search interfaces continues to be the most popular use of the service at this time. Jodie Stroh, OCLC's product manager for Collection Manager, suggested that the API could also potentially be used to expose a library's unique digitized collections. A library could create a custom collection in the knowledge base with links to archives, photographs, or videos. That metadata would then be available to other knowledge base users to expose through implementations of OCLC services.[9]

Integrations with Other Products

Knowledge bases have also proved a practical way to communicate information about an institution's collection to outside services. Most academic libraries already use knowledge bases to support core discovery and management tools—usually all centralized with a single large vendor. But the same holdings information stored in a knowledge base is often required by other services as well. Rather than duplicate the effort of describing the same collections information in two (or more!) places, librarians are working to find creative ways to reuse their knowledge base metadata to help support a broader array of products.

Steve Oberg, assistant professor of library science, described how Wheaton College, in conjunction with the CARLI consortium, has been using holdings information pulled from the SFX knowledge base to support its implementation of BrowZine, a browsable interface for scholarly journals. For the service to work correctly,

Figure 4.2
The University of New Brunswick Libraries' custom search interface is generated using the WorldCat Knowledge Base API and contains direct links into WorldCat.

customers need to communicate to BrowZine exactly which journals their library subscribes to. And while BrowZine provides a way to manually input local holdings, doing so would duplicate work that consortium members have already done in their shared implementation of SFX. With this in mind, Oberg and his colleagues began working with Ex Libris to create a solution that would allow them to use their knowledge base holdings to communicate with BrowZine.[10]

The resulting process involves procuring a weekly export of all active full-text holdings from SFX, which is output to a zip file and stored in an accessible directory on the SFX server. BrowZine then fetches that file and uses it to rebuild each library's holdings information. This system builds on existing SFX functionality that allows customers to set up export profiles based on locally defined criteria. Oberg said that CARLI was one of the first BrowZine users to implement the automated system and the first consortial user to do it. Since then, BrowZine has expanded this functionality to work with other knowledge bases and makes the process available as a standard part of its service.[11]

The experience of integrating BrowZine with the SFX knowledge base has prompted Oberg and the CARLI SFX committee to pursue a new research project that will explore ways to make use of the SFX data with other services, including possibilities like WorldCat Local, Google Scholar, and ILLiad.

"It's not just an SFX thing." Oberg said. "It's something people need to think about a lot more. It's how to leverage all of your investment in your knowledge base, to reduce duplicate work and make sure your access is consistent. You want to make sure that whatever path users choose, they're able to get to your resources. To me the knowledge base is the key part there."[12]

Delivery of Library-Specific Holdings

Another trend that highlights the importance of efficiency and accuracy of knowledge base data is the

move toward delivery of library-specific holdings directly to knowledge bases. Up until now the supply chain has primarily focused on the delivery of global information from publisher to vendor, with the library supplying the localization component. But movement in this space suggests this equation may soon change. Since publishers must keep track of their customers' holdings to manage access and billing, it makes sense that these publishers could also communicate those holdings directly to knowledge base suppliers on behalf of their customers.

The knowledge base providers I spoke with agreed that library-specific holdings would continue to be a key area of expansion in the knowledge base space. Stephanie Doellinger said that OCLC's customers love their existing vendor feeds with Elsevier, Ebrary, and EBL, and that the addition of new feeds is one of their top requests.[13] Oliver Pesch from EBSCO echoed the importance of custom feeds, but stressed that additional functionality would be necessary to make them work for management as well as discovery.[14]

Pesch has been working with Elsevier and others throughout the supply chain to submit a proposal to NISO for work that would enhance the KBART best practice with new functionality to help support delivery of customized holdings, in addition to general efficiency improvements. The proposal notes that libraries care about management metadata and specifies the need to develop a best practice for delivering feeds that include both entitlement and packaging information. It also includes work in the area of automating delivery of data from publisher to knowledge base supplier (for both global and custom feeds) using a web service. The proposal has been endorsed by EBSCO, GOKb, Elsevier, and Project COUNTER.[15]

Decision Support

On the management front, knowledge bases are beginning to be thought of as tools to aid in what has come to be known as *decision support*—the process of gathering information to help with selection and ongoing maintenance of e-resources. Gold Rush, an electronic resources management system and knowledge base offered by the Colorado Alliance of Research Libraries, has made a niche for itself in the decision management space. Gold Rush is a smaller nonprofit service, and many of its customers also subscribe to the larger vended discovery services, said George Machovec, the Alliance's executive director. Rather than try to compete in this arena, Gold Rush has focused on developing a suite of services that allow its customers to delve more deeply into the analytics space.[16]

Gold Rush Decision Support allows libraries to compare packages and to determine the unique and common titles between the two. The service supports
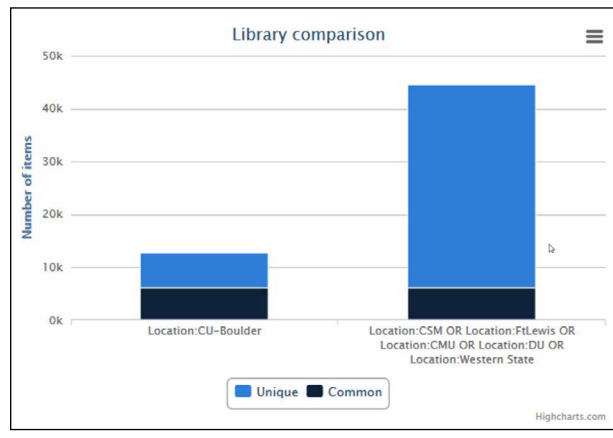

Figure 4.3
The Gold Rush Library Content Comparison System expands the concept of a traditional knowledge base overlap analysis tool to include MARC records and to compare holdings across multiple libraries.

full-text packages, as well as indexing and abstracting services. Machovec said that these features help libraries make better decisions about what to purchase, but can also support maintenance activities in unique ways. For example, the Decision Support Tool can be used to compare a library's holdings in a publisher collection to the holdings found in the Portico or CLOCKSS packages to help investigate compliance with archiving best practices.

A unique feature of Gold Rush is the Library Content Comparison System, which allows libraries to upload their MARC records to a knowledge base–like space (see Figure 4.3). This service is particularly useful if multiple libraries within a consortium subscribe to the service. Participants can then compare their MARC holdings against their peers using a matching algorithm based on various pieces of data, including the title, publisher, and fixed fields. It's a clever repurposing of the knowledge base concept to support different types of collection needs.

Decision support has also become key for e-books, which can be purchased on a dizzying array of platforms, subject to complex technical limitations, and sold in unwieldy bundles of thousands of titles. A 2014 report from the Jisc e-book Co-Design Project aimed to understand the pain points surrounding e-book management and propose actions to address these issues.[17] The top pain points identified in the report included finding out what e-books are available, providing continuing and archival access to purchases, and managing e-book usage statistics.

Jisc's involvement with Knowledge Base Plus (KB+) and GOKb led to two recommendations for development in these systems to advance decision support. The first is advanced availability tracking, which aims to normalize e-book metadata and

identifiers and to track movement of e-books in and out of packages. The second, focused on decision support, proposes a data model that would allow libraries to contribute practical information about e-book management to the global knowledge base. Examples of the type of data to be tracked include license terms, formats offered, digital rights management (DRM) restrictions, platform characteristics, and device compatibility. Some universities in the United Kingdom are already tracking this information using spreadsheets and other local tools. If these attributes were stored in a global platform, they could reach a wider audience of users who could both benefit from and contribute to the decision support data. Since the publication of the Co-Design report, GOKb (with support from Jisc), has begun building a prototype for the e-book availability tracking and decision support functionality. A production release is scheduled for the fall of 2016.

Conclusion

The examples in this chapter demonstrate that the knowledge base has broken through the boundaries of core technical services work to become a key data repository that intersects with workflows across the library. New types of information are being added to knowledge bases that will allow users to manage streaming media formats, track changes over time, and automatically receive customized updates to their holdings. Export processes and APIs allow libraries to use their knowledge base data in more contexts than ever, supporting streamlined management of multiple tools and allowing for the creation of custom interfaces. And more sophisticated data means that knowledge bases can be used for new decision support purposes like availability tracking and collection analysis.

Even with all of this progress, more possibilities remain to be explored. The knowledge bases of the future may allow libraries to implement unmediated borrowing and purchasing at a greater scale through integrations with document delivery services. New data output formats, especially linked data, may support more fluid communication with external systems. And increased need for data to be open and reusable across multiple systems and services may improve interoperability and even lead to the adoption of more central, nonproprietary knowledge base solutions. Chapter 5 of this report will explore the beginnings of this last concept in greater detail.

Notes

1. Yvette Diven, "What's This I Hear about a New Knowledgebase? Part 1, Why a New KB?" *ProQuest Blog,* December 14, 2015, www.proquest.com/blog/pqblog/2015/Whats-this-I-Hear-about-a-New-Knowledgebase-Part-1-Why-a-New-KB-.html.
2. Yvette Diven (product manager lead for management solutions at ProQuest) in discussion with the author, October 2015.
3. "Ex Libris and ProQuest Product Strategy and Roadmap," webinar, ProQuest, January 13, 2016.
4. Kristen Wilson, "Building the Global Open Knowledgebase (GOKb)," *Serials Review* 39, no. 4 (2013): 261–65, http://dx.doi.org/10.1080/00987913.2013.10766408.
5. Kristen Wilson, "Bringing GOKb to Life: Data, Integrations, and Development," in *The Importance of Being Earnest: Charleston Conference Proceedings, 2014,* ed. Beth R. Bernhardt, Leah H. Hinds, and Katina P. Strauch (West Lafayette, IN: Purdue University Press, 2015), 607–13, http://doi.org/10.5703/1288284315649.
6. *Wikipedia*, s.v., "Application programming interface," accessed February 20, 2016, https://en.wikipedia.org/wiki/Application_programming_interface.
7. "WorldCat Knowledge Base API," OCLC Developer Network, https://www.oclc.org/developer/develop/web-services/worldcat-knowledge-base-api.en.html.
8. Brian Cassidy (senior web developer at the University of New Brunswick) in discussion with author, November 2015.
9. Stephanie Doellinger (section manager for data services at OCLC) and Jodie Stroh (product manager for Collection Manager at OCLC) in discussion with author, October 2015.
10. Steve Oberg (assistant professor of library science at Wheaton College) in discussion with author, November 2015.
11. "Working with Your Library Holdings," knowledge base, Third Iron, http://support.thirdiron.com/knowledgebase/topics/79846-holdings-reports-information.
12. Steve Oberg (assistant professor of library science at Wheaton College) in discussion with author, November 2015.
13. Doellinger discussion.
14. Oliver Pesch (chief strategist at EBSCO) in discussion with the author, November 2015.
15. Oliver Pesch, e-mail message to author, November 19, 2015.
16. George Machovec (executive director of the Colorado Alliance of Research Libraries) in discussion with author, November 2015.
17. Ian Chowcat, David Kay, and Owen Stephens working with Amy Devenney and Graham Stone, *eBooks Co-Design Report* (London: Jisc Collections, March 31, 2014), www.jisc-collections.ac.uk/Global/Projects/KB+/KB+%20Documents/140423-ebooks-co-design-report-final.pdf.