Chapter 2

Building a Knowledge Base

The specifics of the process that knowledge base suppliers use to collect, normalize, and maintain their data may be of interest to a broader audience for two reasons. First, librarians interact with these tools intimately and invest a lot of time and effort aiding vendors in keeping knowledge bases up to date. A better understanding of how this work is done provides context for these activities. Second, as knowledge bases begin to reach beyond vended products into national, consortial, and institutional arenas, librarians may soon find themselves more deeply embedded in the process of knowledge base work.

The Supply Chain

Knowledge base metadata originates with the provider of an e-resource—usually the original publisher or a third-party content aggregator. These organizations create data files describing the products they sell. Each file generally represents a salable package made up of a specific set of titles. The lists may include e-journals, e-books, or a mix of both. For each title, the file provides the information needed to identify and access its content, such as unique identifiers, coverage dates, and URLs. This data, which is most often made available as a simple Excel or tab-delimited file, may be made publicly available on a provider's website or provided to knowledge base suppliers through FTP sites, e-mail, or other methods.

Knowledge base suppliers collect title list metadata from content providers and load it into their products, performing quality checks and normalization processes along the way to improve accuracy. These suppliers then distribute the data to libraries through their knowledge base software. Libraries, meanwhile, collect data about their local purchases from vendors, publishers, and subscription agents. They use their own records to identify the packages and titles they have purchased in their local implementation of a knowledge base. Once the appropriate resources are activated, the accompanying metadata is pushed out for use across the library's systems and services. Librarians often attempt to close the loop in this process by reporting changes and corrections to the data back to their knowledge base vendor or the content provider itself.[1]

Building a Knowledge Base

From the perspective of the group building a knowledge base, the first major step is collecting data from content providers. For commercial knowledge bases attempting a comprehensive list of scholarly publications, this process can be an enormous undertaking. Jackie Fahmy, knowledge base product analyst at OCLC, reported in an interview with the author that the WorldCat Knowledge Base contains data from more than 5,000 providers.[2] Oliver Pesch, chief product strategist at EBSCO, cites the numbers for EBSCO's Global Knowledge Base at more than 1,400 providers totaling more than 10,000 unique collections.[3] Smaller knowledge bases are often unable to achieve such a large scale and must limit their scope accordingly. Niche products like CUFTS, run by Simon Fraser University, and GoldRush, based at the Colorado Alliance of Research Libraries, scope their coverage to those collections specifically needed by their customers. National projects may address only publications native to their countries or packages purchased nationally at a consortial level. The Global Open

Knowledgebase (GOKb), an open-source, community-managed project, has begun work by focusing on priority packages and master lists, with a hope of increasing scale through evolving technology and partnerships.

While some content providers make suitable title lists freely available on their websites, others require special arrangements for knowledge base suppliers to access the data. Data can be delivered through several different mechanisms including websites, FTP sites, APIs, and occasionally even e-mail. FTP sites and APIs are ideal, as they allow providers to build harvesters that automatically pull in files, rather than forcing a human to visit the site and download the file manually.

Once a data pipeline is established, the next stage is normalizing and cleaning the data before ingesting it into the production version of the knowledge base. These validation processes aim at achieving consistency across all the resources represented in the knowledge base. Examples of the types of work done at this stage include checking for required fields, properly formatting data like dates and identifiers, and analyzing changes to the files from one update to the next. Additionally, knowledge base suppliers must compare the incoming data against the existing knowledge base. When discrepancies arise, they decide which version of the information is correct and choose to either replace the existing data or disregard the incoming change.

Much of the normalization work is automated, but a human element is still a key part of the process. OCLC, for instance, sets a 5 percent threshold for data changes in a content provider's file. If the threshold is exceeded, the file is flagged for manual review. [4] EBSCO monitors nearly all types of changes to the incoming data, as well as flagging new and dropped titles for review. Once all the review events are addressed, the file can be fully ingested for use in EBSCO's products.[5] The GOKb project divides data exceptions into groups of pre- and post-ingest tasks that are reviewed by its community contributors. Pre-ingest tasks focus on normalizing data before it becomes part of the knowledge base. The post-ingest tasks address discrepancies between old and new data, as well as anomalies that don't necessarily prevent ingest, but could cause problems for users down the road.

Because electronic resources products evolve so rapidly, the data collection and ingest process must be repeated on a regular and frequent basis. The Knowledge Bases and Related Tools (KBART) code of practice recommends that providers issue monthly updates, and many of the larger providers seem to be adhering to this schedule, or even exceeding it. [6] Smaller providers may update their files less often, but they may also have fewer changes to their metadata to warrant a higher frequency. As a result, knowledge base suppliers must be aware of the general schedule used by each provider so that they can continuously harvest, process, and load files to keep their products up-to-date.

To this end, commercial knowledge base providers employ sizable staffs whose job it is to maintain the knowledge base. Breeding reported in 2012 that the four major knowledge base suppliers (Ex Libris, Serials Solutions, EBSCO, and OCLC) employed between eight and twenty-nine full-time employees involved in knowledge base maintenance.[7] A smaller supplier, like CUFTS, has about five staff members who regularly work on the knowledge base, though none of them are full time. GoldRush relies on library school students working part-time to handle its file processing. GOKb employs one full-time staff member and relies on volunteer effort from project partners to help review errors and participate in data enhancement activities.

When It Goes Wrong

The validation stage of knowledge base creation is essential because the data being consumed is prone to errors—due to its complexity and its status as a secondary artifact of the publishing process. "The data we get isn't always clean, pristine data," said Yvette Diven, product manager lead for management solutions at ProQuest. "This metadata can be a byproduct of something that a provider does. If they're focused on publishing e-journals or e-books, this metadata can be a byproduct rather than the main product."[8]

The types of errors commonly found in knowledge bases are well documented. In an early analysis, Chen noted that content providers often failed to update their metadata frequently enough to capture titles added and dropped from their collections. She also provided several examples of data errors at the title level, such as incorrect coverage dates and URLs.[9] Cullen described similar issues broken down into a useful list that includes missing titles, titles listed in error, wrong identifiers for titles (ISSN, ISBN), incorrect coverage information, and incorrect embargo information.[10] Another error type frequently seen in knowledge bases involves the correct representation of serial titles over their life span, including title changes and transfers between publishers. While the introduction of the KBART code of practice has helped promote more frequent updates, metadata problems continue to be an issue for knowledge base suppliers and users, as the entire supply chain struggles to keep up with the volume of changes.

The consequence of bad knowledge base data can be felt across the internal and external operations of the libraries relying on it. The sharpest pain point is for end users of link resolvers and discovery tools, who may be incorrectly told their library has

no access to the article they're searching for—or, perhaps worse, directed to a resource they believe should be available, only to be faced with a pay wall or error message. Librarians also feel the frustration caused by knowledge base errors, which can make it difficult for them to manage their collections, reconcile title lists, analyze usage, and troubleshoot end user problems.

Because these errors have the strongest impact on librarians and library patrons, customers also play a role in helping to maintain the quality of knowledge bases. Every knowledge base supplier I spoke with provides a way for customers to report errors discovered through real-world use of the data. The suppliers then review these error reports, confirm proposed changes with the content provider, and edit the knowledge base if appropriate. While user participation in knowledge base maintenance certainly benefits users themselves and their knowledge base suppliers, Cullen rightly pointed out that the current model can also lead to inefficiencies. Librarians from different institutions will identify and report the same errors to various suppliers. And at times, suppliers are less likely to prioritize user error reports, leading to delays in these changes being applied.[11]

What's needed is additional effort to close the gaps in the supply chain by fixing problems at their source and building environments for greater collaboration. Most knowledge base providers already address the first aspect of this need by communicating known errors back to content providers whenever possible. Chapter 5 of this report will examine national and community efforts to improve the supply chain at a more grassroots level.

Knowledge Bases and Related Tools (KBART)

The biggest challenges surrounding knowledge base maintenance include the sheer volume of data that must be processed, the need to provide timely information, and the task of modeling complex and ever-changing collections of resources. The KBART code of practice was created to address these challenges by defining effective participation in the supply chain. The foundation for KBART was originally proposed by Culling in his 2007 report to UKSG, which identified the need to establish transparent guidelines for how best to format, deliver, and consume knowledge base data.[12] The original KBART working group was formed as a joint venture between UKSG and NISO in 2007. In 2010, the original recommended practice was released, follow by a Phase II revision by NISO in 2014.[13]

While the initial exploration for KBART covered broad ranging topics—including OpenURL syntax and compliance, the role of subscription agents, and the handling of e-journal title changes—the code of practice that emerged has so far focused mainly on the supply of title list files from content providers to knowledge base suppliers. KBART defines the method, frequency, and format of data exchange, along with a set of twenty-five fields to be included in each file. A KBART-compliant title list is a simple tab-delimited file, it can be delivered via a dedicated web page or FTP site, and the fields are all quite straightforward and eye readable.

The KBART standing committee continues work on the initiative, focusing on education and outreach. The committee conducts training workshops for implementers of KBART, provides endorsement for organizations that have demonstrated successful adoption, and maintains a registry of KBART-compliant file sources and contacts.[14] The KBART website currently lists forty-six endorsed organizations, and many unendorsed content providers use the code of practice informally.

The success of KBART has led to some discussion of additional uses and improvements for the code of practice. In a 2014 article, EBSCO's Oliver Pesch identified several new use cases for KBART, including the exchange of KBART data between vendors to allow customers to mix and match products; identification of lendable items for document delivery; and delivery of custom KBART files describing an individual library's holdings.[15] New uses for KBART and knowledge base data in general will be discussed more fully in chapter 4.

Conclusion

The knowledge base supply chain is really a complex web of players who create, consume, enhance, and make use of title list metadata. The process of collecting this information and transforming it into an accurate, consistent knowledge base is a monumental undertaking that can be accomplished at scale by only the largest vendors. At the same time, smaller players, including national, consortial, and open-source knowledge bases, focus on niche areas appropriate to their user bases—and in the process become experts on certain types of content. Individual libraries retain a key role in the supply chain by correcting and improving data issues discovered through real-world use. Together, these groups have managed to put together a system for the creation and maintenance of knowledge bases that has been quite successful—especially when judged against some of the early doubts about the products' feasibility.

Still, areas of inefficiency persist. Each of the large knowledge base providers essentially duplicates the efforts of the others. They all collect the same data and must handle the same errors and inconsistencies.

Closing the loop with content providers also remains a challenge. While libraries and knowledge base suppliers make some efforts to improve data at its source, publishers often lack the resources to acknowledge these changes or implement them in a meaningful way. And while KBART and other standards have made a big impact on the efficiency of data delivery, other areas of the supply chain—such as use of ISSNs and handling of title changes and transfers—could still benefit from additional codification. The following chapters of this report illustrate the extent to which these challenges are recognized across the supply chain and describe many new initiatives that aim to meet them.

Notes

1. James Culling, *Link Resolvers and the Serials Supply Chain,* final project report for UKSG (Oxford, UK: Scholarly Information Strategies, May 21, 2007), www.uksg.org/projects/linkfinal.
2. Jackie Fahmy (knowledge base product analyst at OCLC) in discussion with the author, October 2015.
3. Oliver Pesch (chief strategist at EBSCO) in discussion with the author, November 2015.
4. Fahmy discussion.
5. Pesch discussion.
6. KBART Phase II Working Group, *Knowledge Bases and Related Tools (KBART) Recommended Practice*, NISO Recommended Practice RP-9-2014 (Baltimore, MD: National Information Standards Organization, 2014), www.niso.org/apps/group_public/download.php/12720/rp-9-2014_KBART.pdf.
7. Marshall Breeding, "E-Resource Knowledge Bases and Link Resolvers: An Assessment of the Current Products and Emerging Trends," *Insights* 25, no. 2 (July 2012): 173–82.
8. Yvette Diven (product manager lead for management solutions at ProQuest) in discussion with the author, October 2015.
9. Xiaotian Chen, "Assessment of Full-Text Sources Used by Serials Management Systems, OpenURL Link Resolvers, and Imported E-Journal MARC Records," *Online Information Review* 28, no. 6 (2004): 428–34, http://dx.doi.org/10.1108/14684520410570553.
10. Culling, *Link Resolvers and the Serials Supply Chain*, 28.
11. Ibid., 31.
12. Culling, *Link Resolvers and the Serials Supply Chain*.
13. KBART Phase II Working Group, *Knowledge Bases and Related Tools (KBART) Recommended Practice*, NISO Recommended Practice RP-9-2014 (Baltimore, MD: National Information Standards Organization, 2014), www.niso.org/apps/group_public/download.php/12720/rp-9-2014_KBART.pdf.
14. KBART Registry, NISO, accessed April 29, 2016, https://sites.google.com/site/kbartregistry.
15. Oliver Pesch, "The KBART's Potential beyond OpenURL Linking," *Serials Librarian* 67, no. 3 (2014): 231–39, http://dx.doi.org/10.1080/0361526X.2014.960643.