Chapter 5

# Collaborative Knowledge Bases

The idea of an open, central, and collaboratively managed knowledge base is as old as the knowledge base itself. The first project of this type was the Jointly Administered Knowledge Environment (jake), which began at Yale University in 1999. The goal of the project was to track e-resources metadata and relationships in an open-source environment. Librarians with an interest in the project were encouraged to contribute by collecting journal title lists, correcting errors, and promoting the project with publishers and vendors. By banding together, jake participants could help reduce the duplication of effort that occurred when individual libraries each had to research and document the same information about e-journals.[1] While jake shut down for good in 2007 and never existed as more than a simple online reference of e-resources metadata, it helped set the stage for future efforts to develop open community knowledge bases.

Culling engaged in a significant discussion of the centralized knowledge base in his 2007 report to UKSG. He pointed out that vendors, like librarians, also engage in duplication of effort when it comes to managing e-resources metadata. Each knowledge base supplier must build and maintain its proprietary product in isolation—even though these products all strive to describe the exact same universe of resources. He proposed as an alternative a single central knowledge base that would use web services to provide its data freely to anyone who wished to use it. Culling concluded that while a centralized solution might be possible in the long-term future, it would require significant investment and management from an organization that had the resources to support it.[2]

Another eloquent plea for a centralized knowledge base came from Singer in a 2008 article. He disputed the notion that a single entity would need to manage such a knowledge base and instead pointed to successful projects like *Wikipedia* and the Internet Archive, which harness the power of many invested users to manage open, dynamic content. Singer acknowledged the difficulties of creating such a service, including modeling complex data and coordinating the involvement of large numbers of data managers. However, he believed the payoff in implementing this model would ultimately be worth the cost:

> The knowledgebase crisis is not going away, and as the digital universe expands, especially to new and different formats, it will only get more difficult to manage. By tapping into the power of the entire community—from the beginning of the publishing chain to the end-user—the knowledgebase becomes self-sustaining and finds new and interesting uses along the way.[3]

While Singer's vision has certainly not become reality yet, several projects that have emerged over the past five years demonstrate that the desire remains to collectively improve knowledge base data and ease its flow across the supply chain.

## Community-Managed Knowledge Bases

### The Global Open Knowledgebase

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

The project most closely aligned with the grand vision for knowledge base collaboration is the Global

Open Knowledgebase (GOKb.) Not unlike jake, the project aims to provide a fully open, community-managed knowledge base that describes electronic journals and books and their relationships. The three major ambitions for the GOKb project are improving data quality across the supply chain, reducing duplication of effort, and encouraging interoperability between systems. GOKb's focus on openness, collective effort, and enhanced data model (described in chapter 4) all contribute to its work in these areas.

The GOKb project began as a joint venture between Jisc and the Kuali OLE project. In addition to support provided by these institutional project partners, GOKb also employs one full-time staff member, the GOKb editor. The editor is responsible for setting the policies that define how the data is managed and for coordinating the community members who can contribute various forms of effort to GOKb. Contributions include collecting and loading KBART-formatted title lists into the knowledge base, addressing data errors and anomalies identified during the loading process, and engaging in other data enhancement activities, such as researching and documenting title history information.[4] As the lead school on the project, North Carolina State University has engaged heavily with GOKb, contributing staff time to pilot a data-loading initiative, and several other Kuali OLE partners have contributed to the data-loading process as well.[5] GOKb has also been successful in attracting librarians unaffiliated with its major partner projects to work with the knowledge base in more lightweight ways—particularly in areas such as researching title changes and documenting them in the knowledge base.

GOKb's data is freely available under a Creative Commons 0 (CC0) license, which means that it can be used by anyone, for any purpose, without attribution.[6] While GOKb was originally created to support the Knowledge Base Plus (KB+) and Kuali OLE services, the fact that the data is in the public domain means the project can have a much broader impact. Other open-source projects in need of knowledge base data are free to use GOKb, and—just as importantly—publishers and vendors can consume the data as well. As GOKb grows and the quality of its data improves, publishers at the top of the supply chain can also use GOKb's data to improve their own data, while knowledge base suppliers can integrate the data into their services. Vended knowledge bases could at some point even replace their proprietary knowledge bases with GOKb or mirror some or all of its content rather than maintaining the same information themselves.

The visionary changes to which GOKb aspires are still a ways off. GOKb currently includes about 400 packages, compared to the tens of thousands found in most commercial systems. The scale of data that a comprehensive knowledge base needs to cover has proven difficult to achieve with only a single staff member and a couple dozen volunteers. The development team for the project is currently working on a new data loader that will allow multiple files to be loaded at once, opening up the possibility of consuming larger data sources in an automated way. New partners will also be necessary to achieve scale. Library partners are needed to help monitor data quality and collect enhanced data like title and publisher changes. And a large-scale partner—possibly even another knowledge base—will likely be required to collect the amount of data needed to be truly comprehensive.

Still, GOKb exists as an excellent proof of concept of the open, collaborative knowledge base. My experiences working with this project have convinced me that the library community values work in this area and that many individual librarians would be willing to contribute to an easy-to-use, well-managed knowledge base effort. I believe, also, that buy-in from other stakeholders, including publishers, knowledge base vendors, and standards organizations, is essential to meeting this goal. The vision for a centralized knowledge base remains valid, but it cannot be fully realized without the engagement of key players across the supply chain.

### WorldCat Knowledge Base

OCLC has also begun to explore a community management approach with its WorldCat Knowledge Base. While this product is not open source or intended to be a cross-product solution, OCLC has gone further than any of the other vended knowledge base products in inviting librarians to be part of the management process.

The WorldCat Knowledge Base operates using a cooperative approach that allows customers to view changes made to the knowledge base and vote on whether to approve or deny them. The voting window for each change is open for five days. If a change gets ten votes in either direction during this window, it will be implemented or rejected accordingly. If fewer than ten votes are received, the change will be automatically accepted when the voting window closes. Jackie Fahmy from OCLC said that users tend to cast more negative votes for errors and problematic changes, and simply let the voting window expire for the changes that don't affect them.[7] Votes also tend to come from a small number of very active libraries that want a lot of control over their data. To increase participation, OCLC has considered implementing a notification service, which would allow users to receive alerts when changes occur in specifically chosen packages.

OCLC also offers its users the ability to create custom packages that can be shared globally with all of its knowledge base customers. In addition to supplying typical knowledge base data, users can also


**Figure 5.1**
The BACON knowledge base helps French publishers improve their metadata and assigns a quality label to those who meet certain standards.

link up the holdings in custom collections with the appropriate MARC records. Fahmy said that creating custom collections for packages where the publisher doesn't provide KBART files or MARC records is a popular use case. Participants in consortial deals have also taken advantage of the cooperative management functionality. Fahmy described how one North Carolina library created custom collections to represent some of the content it receives from NCLIVE, a statewide consortium. The packages were then made available to other OCLC libraries in North Carolina that had access to the same content. In this way, individual librarians, many of whom are doing knowledge base work anyway, can have an impact beyond just their own institutions.

"Being a cooperative ourselves here at OCLC, we thought it was a good idea to allow our knowledge base to be a cooperative, too," Fahmy said. "We're reliant on data from providers and knowing that not everything is perfect, we wanted to give users the ability to make data changes for everybody. By doing this, we're giving the cooperative and the librarians the ability to own this data and make it what they need it to be."[8]

## National Knowledge Bases

National knowledge base projects have also taken up the banner of open, collaborative data management. These projects are most often run through a government agency, national library, or large university, and they attempt to create central knowledge bases describing resources specific to a certain country. National knowledge bases have tended to emerge in countries where there is already a high level of national collaboration, including the United Kingdom, Germany, France, and Japan, among others.

One of the primary goals of national knowledge bases has been to improve the accuracy of data that commercial suppliers often struggle to provide. National knowledge bases tend to fall into two categories with regard to this goal. Some aim to describe electronic resource content purchased by libraries in their country, regardless of its origin. The KB+ project in the United Kingdom, for example, describes subscription deals negotiated by British consortia, along with some master title lists for popular publishers. KB+ data managers spend huge amounts of time verifying title lists and improving metadata. Earney noted that KB+ data managers spent more than 70 hours creating a single title list for at least one major publisher package.[9]

In Germany, two knowledge base–like projects also attempt to capture definitive lists of holdings on behalf of member libraries. The Zeitschriftendatenbank (ZDB), or German Journal Database, is a bibliographic database that contains MARC records representing the print and online journal collections held by more than 4,400 German and Austrian libraries.[10] The Elektronische Zeitschriftenbibliothek (EZB), or Electronic Journals Library, provides information about German-held online serials, with more of an electronic resources management perspective.[11] In both cases, small, dedicated staffs collect and vet information with the goal of providing highly accurate metadata.

Other national knowledge bases focus more on describing publications that originate from their homelands. In France, the BAse de COnnaissance Nationale (BACON) project has a mission to create high-quality knowledge base data describing French publications.[12] The idea for BACON came up when the ABES agency, which maintains a French union catalog, conducted a survey that revealed that most libraries were happy with their vended tools, but found that data about French publications was often missing or incorrect. BACON aims to close this gap by collecting title lists from French publishers, analyzing and correcting errors, and formatting the lists according to the KBART code of practice. The vetted lists are then made freely available through the BACON site and shared with the original publishers (see Figure 5.1). The E-Resources Database-Japan (ERDB-JP) is a similar effort to supplement the supply chain with knowledge base data describing electronic journals and books written, edited, or published in Japanese.[13]

The data created by national knowledge bases is designed to directly benefit constituent libraries. In some cases, the data serves mainly as a reference. It can be searched and browsed on the web, exported for

local use, or accessed via an API. Benjamin Bober, the manager of the BACON project, described one potential use case for the French knowledge base data. ABES has been working on developing a tool to analyze e-resource usage using EZProxy logs. The goal of the project is to pull each URL visited from the logs and determine which resource it corresponds to using knowledge base data. This data can then be compared with COUNTER usage data to confirm accuracy or used in place of COUNTER data when it is not available. Without accurate, KBART-compliant files for French publications, such a process would be all but impossible.[14]

Other national knowledge bases go a step further by building services on top of their data. The ZDB and EZB, both of which have been around since 1997, support core library services. The ZDB provides tools to facilitate interlibrary loan and document delivery, and the EZB offers a linking service and XML feeds that can be used to support integration of the data with other systems. KB+ powers a full-featured electronic resources management system (ERMS) that supports subscription management, licensing, and integration with the Journal Usage Statistics Portal (JUSP). ERDB-JP supports a link resolver and discovery service.

The value of national knowledge bases also extends beyond library-focused services. KB+, BACON, and ERDB-JP all have explicit goals of improving the supply chain by making their data freely available for reuse under a CC0 license. In many cases, the knowledge base data created by these organizations fills in vital gaps in the supply chain. From the perspective of Tomoki Ueno and Tomoko Kagawa, who help manage the ERDB-JP project, Japanese resources are often underrepresented in products that are aimed at predominantly American and European audiences. By creating an open knowledge base of Japanese materials, they hope to provide a reliable source for this missing data.[15] The consortial deals managed by KB+ can also be difficult for commercial knowledge bases to represent because the details are not included in public channels and official data feeds. Currently, OCLC, ProQuest/Ex Libris, and EBSCO are all using freely available KB+ to enhance their own knowledge bases.

"The collaboration with KB+ has been fantastic," Yvette Diven, product manager lead for management solutions at ProQuest, said. "The entire team there is trying to build up something that's repeatable, that's sharable for other groups. And they're very open to working with commercial knowledge bases. They've actually laid the groundwork so that others who are using the KB+ model can follow. They've broken that ground."[16]

Efforts to mend the supply chain extend all the way back to the source, as many national knowledge base creators also make it their goal to work directly with publishers. Such work is an explicit part of BACON's mission. The knowledge base assigns a special quality label to identify data that has been certified by ABES and adopted by the publisher to improve the data provided on its own platform. So far, only two publishers have earned the quality label, but Bober said that BACON is working closely with nine or ten other publishers and hopes to award more endorsements soon. "It's a win-win because we centralize the metadata, but also encourage publishers to use it on their own platforms," he said. "So far, I think publishers are quite happy with the work we have done."[17]

GOKb has been in contact with all of these national knowledge base representatives to discuss the logistics of incorporating their data into the global knowledge base. The open nature of the data means the only barriers to this type of collaboration are the technical and resource challenges of loading and updating large amounts of data. Enhancements to GOKb's data-loading process should make redistribution of this data a more realistic prospect before 2016 is out.

## Conclusion

Enormous political and structural challenges stand in the way of fully implementing the open knowledge base vision. Purveyors of commercial products view the quality of their knowledge bases as a sales differentiator and would be rightly cautious in abandoning their proprietary systems for a communal approach. Individual libraries and librarians are often stretched thin and may believe that a vendor with paid staff could simply do the work better. Any change to the current situation will likely be a long and gradual one.

Still, in the interplay between open, commercial, and national knowledge bases and their users, it's possible to see how the vision for community knowledge base management might eventually play out. Participants in each type of knowledge base can contribute work that is natural and meaningful to their circumstances to the larger community. Commercial and global services would likely collect the data with the broadest application, national and regional groups would have an incentive to supplement it with specialized collections, and users across the board would contribute individual enhancements and corrections about the titles that are most important to them. Together these groups are already performing much of the work that would be needed to support more centralized knowledge base management across the industry.

The benefits of a truly central knowledge base to the field would be enormous. Nearly every trend I've written about in this report would benefit from

greater openness of knowledge base data. Shared identifiers (or even a shared data pool) could contribute to greater interoperability across tools and platforms, offering libraries choice and flexibility. Data enhancement efforts could increase, as open knowledge bases could be easily paired with other free data sets, especially those emerging with the rise of linked data. Broad availability of the data through APIs would increase creativity throughout the field, allowing libraries and individuals to create their own knowledge base–powered projects. And maximizing the number of users of a single data pool could have a big impact as different types of users contribute data changes and enhancements that can be applied across the entire supply chain. While achieving this vision will be no easy feat, the potential for great strides exists, and the first steps have already been taken.

## Notes

1. Daniel Chudnov, Cynthia Crooker, and Kimberly Parker, "jake: Overview and Status Report," *Serials Review* 26, no. 4 (2000): 12–17, http://dx.doi.org/10.1080/00987913.2000.10764619.
2. 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.
3. Ross Singer, "The Knowledgebase Kibbutz," *Journal of Electronic Resources Librarianship* 20, no. 2 (2008): 85, http://dx.doi.org/10.1080/19411260802272776.
4. 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://dx.doi.org/10.5703/1288284315649.
5. Eric M. Hanson, Xiaoyan Song, and Kristen Wilson, "Managing Serials Data as a Community: Partnering with the Global Open Knowledgebase (GOKb)," *Serials Review* 41, no. 3 (2015): 146–52, http://dx.doi.org/10.1080/00987913.2015.1064853.
6. "CC0 1.0 Universal Public Domain Dedication," Creative Commons, accessed April 29, 2016, https://creativecommons.org/publicdomain/zero/1.0.
7. Jackie Fahmy (knowledge base product analyst at OCLC) in discussion with the author, October 2015.
8. Ibid.
9. Liam Earney, "Maximizing the Knowledge Base: Knowledge Base+ and the Global Open Knowledgebase," *Insights* 26, no. 3 (2013): 244–49, http://dx.doi.org/10.1629/2048-7754.71.
10. Hans Lieder (head of the department for bibliographic services at the Berlin State Library) in discussion with the author, November 2015.
11. Evelinde Hutzler (librarian at Universitätsbibliothek Regensburg), e-mail message to author, November 19, 2015.
12. Benjamin Bober (librarian at ABES) in discussion with the author, December 2015.
13. Tomoki Ueno (librarian at the University of Electro-Communications) and Tomoko Kagawa (librarian at Ochanomizu University), e-mail message to author, November 16, 2015.
14. Bober discussion.
15. Ueno and Kagawa e-mail.
16. Yvette Diven (product manager lead for management solutions at ProQuest) in discussion with the author, October 2015.
17. Bober discussion.