Bringing GOKb to Life: Data, Integrations, and Development

Kristen B. Wilson

North Carolina State University Libraries, kristen_wilson@ncsu.edu

Follow this and additional works at: <https://docs.lib.purdue.edu/charleston>

Part of the [Cataloging and Metadata Commons](#), and the [Scholarly Publishing Commons](#)

An indexed, print copy of the Proceedings is also available for purchase at:

<http://www.thepress.purdue.edu/series/charleston>.

You may also be interested in the new series, Charleston Insights in Library, Archival, and Information Sciences. Find out more at: <http://www.thepress.purdue.edu/series/charleston-insights-library-archival-and-information-sciences>.

Kristen B. Wilson, "Bringing GOKb to Life: Data, Integrations, and Development" (2014). *Proceedings of the Charleston Library Conference*.

<http://dx.doi.org/10.5703/1288284315649>

This document has been made available through Purdue e-Pubs, a service of the Purdue University Libraries. Please contact epubs@purdue.edu for additional information.

Bringing GOKb to Life: Data, Integrations, and Development

Kristen B. Wilson, North Carolina State University Libraries

Abstract

The Global Open Knowledgebase (GOKb) project is developing a repository of freely available data that describes electronic journals and books as they are offered in the academic publishing supply chain. Since the first partners release in May 2014, the project has taken major steps toward realizing its goals. This article will include a general project overview and update, followed by discussion of data collection, integration, and development initiatives that are already underway among the project partners. Readers will also learn about next steps for GOKb and opportunities for broader community involvement.

Why GOKb?

Electronic resources knowledge bases are essential tools for libraries, providing management information and powering public-facing discovery services. Most libraries spend significant amounts of time managing knowledge base data, not only in their primary knowledge base but also in related tools like the integrated library system (ILS), electronic resources management system (ERMS), and subscription agent platform. My own experience at North Carolina State University Libraries illustrates both the extent of work required to manage a knowledge base and the duplication of effort that can accompany this process.

Like most large libraries, NC State uses a commercial knowledge base product. We often notice that the holdings information in our knowledgebase changes unexpectedly, introducing errors into our systems. Sometimes, large groups of titles disappear from a package. Or the coverage dates for a certain package are replaced with incorrect data. Whenever possible, we investigate these issues, figure out what the correct information should be, and send it back to our knowledgebase vendor. To their credit, our vendor is able to fix the majority of the problems we report. However, we often hear the same explanation for these errors: they come from the publisher, and sometimes the publisher sends the same errors again and again. In the past, staff in my unit have tried contacting publishers directly and asking them to fix faulty data, but haven't gotten very far with this approach. Dealing with data issues tends not to be a priority for most

publishers, and that's understandable given that they are the business of selling content—not metadata. Still, isn't it worth thinking about how we might address these data errors at their source? And consider this—when an individual library finds problems and reports them to a knowledge base vendor, it is doing a service for all the customers of that vendor. But what about libraries that use a different vendor? I have to imagine there are many librarians out there spending a lot of time correcting knowledge base errors and potentially duplicating one another's effort. Why not harness their efforts in a central location to benefit the entire community?

These questions get to the heart of what the Global Open Knowledgebase (GOKb) project hopes to accomplish. First, to harness the power of the library community, who in many cases are already doing significant knowledge base work, to create a repository of high quality metadata about electronic resources. And second, to engage publishers and vendors to help improve the quality of this data across the supply chain.

To these ends, GOKb (<http://gokb.org>) is a freely available, community-managed data repository that contains key publication information about electronic resources as represented within the supply chain from content publishers to suppliers to libraries. The community management aspect of GOKb means that librarians can directly influence the quality of the data and contribute their work back to a central repository. And the fact that GOKb's data is freely available under a Creative Commons 0 (<http://creativecommons.org/publicdomain/zero/1.0/>) can be used by anyone,

for any purpose, without attribution. These two features, taken together, uniquely position GOKb to become a central knowledge base that is managed by its users, independent of any one commercial product, and freely available to any organization that can benefit from its data.

GOKb Data

Data Processing and Workflow

Data is the heart of the GOKb project. By collecting and curating essential metadata about electronic resources, GOKb will provide a valuable service to libraries and vendors who wish to use that data in local projects or to improve the quality of an existing data set. GOKb's data is contributed by project partners and flows through several stages of processing to become part of the knowledge base.

The starting point for data management in GOKb is a tool called OpenRefine (http://openrefine.org). OpenRefine is open source software used for cleaning, transforming, and enhancing data. The GOKb development team has built a custom extension for OpenRefine that allows users to take

title list files that have been created by publishers or vendors and validate them against a predefined set of rules before loading them into GOKb. These files are currently downloaded from publisher web sites, but in the future they may be harvested in an automatic way or contributed directly by publisher partners themselves.

Once data has been loaded into OpenRefine, the GOKb extension identifies errors like blank fields, incorrectly labeled data, and ISSN conflicts. A series of error and warning messages appear in the OpenRefine interface, along with tools that help automatically address problems or single out values that are problematic (see Figure 1). Users can work through each error, and when they are all resolved, OpenRefine will present an option allowing the user to ingest the data into GOKb. The OpenRefine extension can be used for any file type, although the formatting will determine how much work is needed to complete the validation process. GOKb has been designed to work especially well with KBART-compliant files (http://www.niso.org/workrooms/kbart), and the project strongly supports industry adoption of this best practice.

OPEN

Refine

Taylor and Francis: Sociology and Related Disciplines Collection: 20141031

Permalink

Open...ExportHelp

Facet / FilterUndo / Redo0Errors 7/15

39 rows

Show as: rowsrecordsShow: 5102550 rows

« first « previous 1 - 10 next » last »

All	publication_title	print_identifier	online_identifier	date_first_issue_online	num_first_vol_online	num_first_is
1.	African Identities	1472-5843	1472-5851	2003-04	1	1
2.	Australian Feminist Studies	0816-4649	1465-3303	1997-04	12	25
3.	Ethnicity & Health	1355-7858	1465-3419	1997-03	2	1-2
4.	Innovation: The European Journal of Social Science Research	1351-1610	1469-8412	1997-03	10	1
5.	International Review of Sociology	0390-6701	1469-9273	1997-03	7	1
6.	Journal of Gender Studies	0958-9236	1465-3869	1997-03	6	1
7.	Journal of Intercultural Studies	0725-6868	1469-9540	1997-04	18	1
8.	Journal of Ethnic and Migration Studies	1369-183X	1469-9451	1997-01	23	1
9.	Journal of Youth	1367-6261	1469-9680	1998-02	1	1

title.identifier.issn

Import does not specify an title.identifier.issn column

title.identifier.eissn

Import does not specify an title.identifier.eissn column

DateFirstPackageIssue

Import does not specify an DateFirstPackageIssue column

DateLastPackageIssue

Import does not specify an DateLastPackageIssue column

KBARTEmbargo

Import does not specify an KBARTEmbargo column

CoverageDepth

Figure 1. The OpenRefine client featuring the GOKb extension.

Once data has been ingested from OpenRefine, it becomes part of the GOKb knowledge base and is accessible to users via a web application. Within this application, data can be edited further. Upon

ingest, GOKb will generate a series of review tasks to identify possible errors that could not be detected at the OpenRefine stage. Each review task is assigned to the user who completed the

initial ingest, although it can be reassigned to another user if desired. Users will resolve any review tasks generated by an OpenRefine ingest before approving the file for public use. There are also additional data quality activities that will take place in the GOKb web application, such as creating title history metadata or documenting information about open access rights. These activities, however, are large projects that will be addressed through a variety of methods as the knowledge base matures.

Data Model

GOKb's data model contains many of the same record types as traditional knowledge bases, along with a few unique features that improve on these existing concepts. Some of the key entities within the GOKb data model are the title, package, platform, title instance package platform (TIPP), and organization. Each of these records supports a

variety of metadata elements, as well as a series of relationships created by establishing linkages between the different record types.

The title record in GOKb describes an electronic resource at the work level. It includes basic administrative metadata, including some fields that help users see the status of the title and whether or not it has been approved by an editor. This record also includes extensive title history metadata. A set of published from and to dates show the years that a title was actively publishing, and a series of title history events can be used to link titles together. Each event describes the situation before and after a change occurs. Figure 2 illustrates a title family that contains a simple title changes and a title split. The title record also contains identifiers for the title, a list of current and past publishers, and a list of all the packages to which the title belongs.

Title: Acta Materialia

Title ⓘ Acta Materialia (Modify title through variants below)

Status ⓘ Current

Status Reason ⓘ Empty

Edit Status ⓘ Approved

Current Publisher ⓘ Elsevier (Organization 4487)

Imprint ⓘ Pergamon

Published From ⓘ 01/01/1996

Published To ⓘ Empty

Title History ⓘ

Available actions

-- Select an action -- Go

Date	Before	After	Actions
01 January 1996	<ul style="list-style-type: none"> Acta Metallurgica et Materialia (01 January 1990 To 31 December 1995) 	<ul style="list-style-type: none"> Acta Materialia (01 January 1996 To) 	Delete
01 January 2000	<ul style="list-style-type: none"> Nanostructured Materials (01 January 1992 To 01 January 1999) 	<ul style="list-style-type: none"> Scripta Materialia (01 January 1996 To) Acta Materialia (01 January 1996 To) 	Delete

Figure 2. A GOKb title record with two title history events.

Like most knowledge bases, GOKb contains records to represent packages and platforms. The package record brings together a set of titles that are commonly sold as a group by their publisher or a third party. The platform record represents a web site where electronic resources content is

hosted. GOKb links together the combinations of title, package, and platform to form another record called the title instance package platform, or TIPP. The TIPP record represents what might be traditionally thought of as a holding—the unique, salable entity that libraries purchase and manage.

By creating the TIPP record, GOKb can not only store metadata about this entity, such as URL and coverage dates, but also assign it an identifier and make statements about its relationships to other records.

GOKb also creates linkages between the records mentioned above and organization records. Organizations represent the key players involved in making e-resources available to libraries: publishers, content providers, platform providers, vendors, and licensors. Organizations within GOKb can assigned one or more of these roles, and associated with resources with respect to which they play those roles. Organizations also mark the start of GOKb's foray into linked data. Thanks to contributions from on an organization name linked data project done at NCSU (<http://www.lib.ncsu.edu/ld/onld/>), we have been able to prepopulate GOKb's organization records with a rich set of metadata, including alternate names and URLs for other linked data sets like the Library of Congress Name Authority File (<http://id.loc.gov/authorities/names.html>) and the Virtual International Authority File (<http://viaf.org/>).

Integrations

While the GOKb web application is a useful tool in itself, GOKb's data is also available via an API, which will allow users to create integrations with local systems. GOKb's current partners, Kuali OLE (<http://www.kuali.org/ole>) and Knowledge Base + (<http://www.kbplus.ac.uk/kbplus>), have already been working on creating integrations with GOKb. The API and coreferencing service allow any user to do the same at a variety of scales—from a full-scale integration with a downstream system to a simple queries using a tool like OpenRefine.

The primary function of GOKb's API is to allow external systems to extract data from GOKb. The API can be used to retrieve data such as the list of titles in a package, along with their associated metadata, or the title history information for a particular title. GOKb uses the Open Archives Initiative Protocol for Metadata Harvesting (<http://www.openarchives.org/pmh>), so users familiar with that standard should find GOKb's APIs easy to adopt. Additionally, authorized users will be able to send updates back to GOKb via API.

This piece of the API will become critical for partners who find themselves interacting with GOKb's data primarily through a downstream system.

Integration between GOKb and KB+ is currently under development. KB+ is a service developed by Jisc Collections (<https://www.jisc-collections.ac.uk>) to help libraries in the United Kingdom manage their electronic resource purchases. KB+ provides information about publications, subscriptions, and licenses. Data managers for the KB+ project have been building their own knowledge base of information about UK-specific journal packages within their local environment. The ultimate goal for KB+ is to shift from managing these packages locally and instead manage the global data in GOKb, while continuing to allow libraries to customize the data to match their individual entitlements in KB+. This model gets back to the earlier point about reducing duplication of effort. By doing part of their work in GOKb, the KB+ data managers can maximize the impact of their efforts by contributing them to a central community system that feeds many other projects in addition to their own.

The integration between GOKb and KB+ is still under development, but enough work has been completed to convey clearly how the functionality will work. Within KB+, a user can access a list of all of the packages available through GOKb. This information is brought into KB+ using the GOKb API. The user can choose to create a new local package in KB+ based on a GOKb package, or to merge GOKb's package information with an existing package in the system. Creating or merging a package will pull in information from GOKb for all the titles in the package, including identifiers, coverage dates, URLs, and other details. Once a package has been established with a link to GOKb, users will receive updates whenever new information is available about that package or its titles, and they will have the option to accept or reject each change.

GOKb's other partner, Kuali OLE, is planning to integrate its product with GOKb using a very similar model to KB+. Kuali OLE is a full library management system, which includes traditional functionality like acquisitions, cataloging, and

circulation, in addition to e-resources management. In the Kuali OLE model, GOKb data will be used to create records like bibs, holdings, organizations, and platforms. Users will receive updates through a central change management dashboard, where they can choose which updates to apply or reject.

In addition to its API, GOKb offers a coreferencing service that will allow users to create crosswalks between identifiers. Users will be able to search GOKb for any identifier and return a list of all the other identifiers associated with that some component. Users can also search for within in specific namespaces, such as ISSNs or a publisher's proprietary identifier. This coreferencing service is available through the GOKb web interface and can also be accessed using JSON or XML queries sent by an external system.

Users can take advantage of the coreferencing service by using a lightweight tool like OpenRefine to retrieve additional identifiers for a data set. It's common that a user may have data that contains only a single identifier, for example the ISSN. Within OpenRefine, the user can add another column to her project based on a URL—in this case a URL that queries the GOKb coreferencing service. This new column will display a block of JSON code provided by GOKb. Within this code, the user will see any additional identifiers associated with an entity, such as an e-ISSNs or a publisher's proprietary identifier (see Figure 3). Tools like OpenRefine or Excel can be used to parse that information from the JSON code and make it into more regular or eye-readable data. As GOKb begins to collect more data and store more identifiers, this service should become extremely useful, and may aid in processes like system implementations or migrations.

16 rows			Extensions: undefined
Show as: rows records		Show: 5 10 25 50 rows	« first < previous 1 - 10 next > last »
Title	ISSN	Co-referencing service	
3 Biotech	2190-572X	<pre>{ "requestedNS": "issn", "requestedID": "2190-572X", "gokbIdentifier": "org.gokb.cred.Identifier:41707", "count": 1, "records": [{ "type": "org.gokb.cred.TitleInstance", "id": "41709", "name": "3 Biotech", "gokbIdentifier": "org.gokb.cred.TitleInstance:41709", "sameAs": [{ "namespace": "gokb", "identifier": "org.gokb.cred.Identifier:41707", "namespace": "issn", "identifier": "2190-572X", "namespace": "eissn", "identifier": "2190-5738", "namespace": "springer", "identifier": "13205" }] }] }</pre>	
4OR	1619-4500	<pre>{ "requestedNS": "issn", "requestedID": "1619-4500", "gokbIdentifier": "org.gokb.cred.Identifier:41716", "count": 1, "records": [{ "type": "org.gokb.cred.TitleInstance", "id": "41718", "name": "4OR", "gokbIdentifier": "org.gokb.cred.TitleInstance:41718", "sameAs": [{ "namespace": "gokb", "identifier": "org.gokb.cred.Identifier:41716", "namespace": "issn", "identifier": "1619-4500", "namespace": "eissn", "identifier": "1614-2411", "namespace": "springer", "identifier": "10288" }] }] }</pre>	
ADHD Attention Deficit and Hyperactivity Disorders	1866-6116	<pre>{ "requestedNS": "issn", "requestedID": "1866-6116", "gokbIdentifier": "org.gokb.cred.Identifier:41725", "count": 1, "records": [{ "type": "org.gokb.cred.TitleInstance", "id": "41727", "name": "ADHD Attention Deficit and Hyperactivity Disorders", "gokbIdentifier": "org.gokb.cred.TitleInstance:41727", "sameAs": [{ "namespace": "gokb", "identifier": "org.gokb.cred.Identifier:41725", "namespace": "issn", "identifier": "1866-6116", "namespace": "eissn", "identifier": "1866-6647", "namespace": "springer", "identifier": "12402" }] }] }</pre>	
ADHESION ADHESIVES&SEALANTS	2192-2624	<pre>{ "requestedNS": "issn", "requestedID": "2192-2624", "gokbIdentifier": "org.gokb.cred.Identifier:41730", "count": 1, "records": [{ "type": "org.gokb.cred.TitleInstance", "id": "41732", "name": "ADHESION ADHESIVES&SEALANTS", "gokbIdentifier": "org.gokb.cred.TitleInstance:41732", "sameAs": [{ "namespace": "gokb", "identifier": "org.gokb.cred.Identifier:41730", "namespace": "issn", "identifier": "2192-2624", "namespace": "eissn", "identifier": "2195-6545", "namespace": "springer", "identifier": "35784" }] }] }</pre>	
AGE	0161-9152	<pre>{ "requestedNS": "issn", "requestedID": "0161-9152", "gokbIdentifier": "org.gokb.cred.Identifier:41735", "count": 1, "records": [{ "type": "org.gokb.cred.TitleInstance", "id": "41737", "name": "AGE", "gokbIdentifier": "org.gokb.cred.TitleInstance:41737", "sameAs": [{ "namespace": "gokb", "identifier": "org.gokb.cred.Identifier:41735", "namespace": "issn", "identifier": "0161-9152", "namespace": "eissn", "identifier": "1574-4647", "namespace": "springer", "identifier": "11357" }] }] }</pre>	
AI & SOCIETY	0951-5666	<pre>{ "requestedNS": "issn", "requestedID": "0951-5666", "gokbIdentifier": "org.gokb.cred.Identifier:41740", "count": 1, "records": [{ "type": "org.gokb.cred.TitleInstance", "id": "41742", "name": "AI & SOCIETY", "gokbIdentifier": "org.gokb.cred.TitleInstance:41742", "sameAs": [{ "namespace": "gokb", "identifier": "org.gokb.cred.Identifier:41740", "namespace": "issn", "identifier": "0951-5666", "namespace": "eissn", "identifier": "1574-4647", "namespace": "springer", "identifier": "11357" }] }] }</pre>	

Figure 3. Using OpenRefine to query GOKb's co-referencing service.

GOKb Development

GOKb has received a second round of funding from the Andrew W. Mellon Foundation to support development of the project from October 2014 through December 2015. During this period, the development team will focus on

enhancing existing functionality and exploring new directions.

There are several areas of ongoing work that will receive enhancement during Phase 2 of the project. The first area of development will be to continue to streamline the data loading process through OpenRefine. While the OpenRefine client

goes a long way toward guiding users through this process, there are still improvements to be made. Phase 2 will see the OpenRefine software become simpler to download and install, with more quick resolution options and tools that allow users to create reusable rules for specific packages and providers. We also want to look at the possibility of automating the data ingest process even further, so that data managers can focus more on data quality work instead of ingest.

Second, GOKb needs more features to support community management. Planned enhancements to the web application include editor dashboards that will allow for better monitoring of the general health of the system. These dashboards will highlight new contributions that need to be vetted and identify areas of missing or low quality data so they can be targeted for improvement. Editors will also need a snapshot view of open review tasks, so they can address or reassign outstanding work. We also plan to implement rules and options that will prompt review of high-value items and allow users to escalate problems to an editor.

Finally, GOKb will continue to develop ways to improve its overall data quality. One of the biggest areas of ongoing work needed to maintain any knowledgebase is documenting title histories and making sure that holdings are consistent with the years a title actually existed. During Phase 2, we plan to bring in more title history data from sources like KB+ and union catalogs, and to create review queues where users can focus on working through errors or conflicts related to this data.

During Phase 2, GOKb will also be exploring new areas of development. One of the project's charges is to begin integrating ebook metadata into the knowledge base. We're hoping to collaborate with librarians, vendors, and publishers who are already the experts in this area to expand our data model to accommodate ebooks. We are also planning to expose data about all of GOKb's major entities as linked data, which will allow us to make connections with existing vocabularies and create new terms to describe concepts not currently covered. Linked data is one of the first priorities for development

and something we hope to accomplish fairly early on in Phase 2. Finally, GOKb hopes to collect more data about open access publications and to begin some investigations into the ways that this data might support other open access initiatives. This work will likely begin with an environmental scan and user research to discover what data is considered most valuable.

Community Building

Perhaps the biggest challenge for GOKb is that of collecting a significant amount of knowledge base data in a way that is scalable and effective. In addition to new technical development that will make the process easier, we also need to identify new partners who can contribute resources and expertise. Many of partnership use cases I will describe here are only in the beginning stages of development, but they represent powerful ways that GOKb can work with the community to build a robust knowledge base while also providing benefits to its partners.

Library partners will be key collaborators. Our current libraries partners include the Kuali OLE institutions, who are in the early stages of working on data collection and data quality activities. In the most basic model, these partners are getting title list files from publisher web sites, loading them into GOKb using OpenRefine, and working on cleaning up data problems in the web application. We'd like to find more library partners who would be interested in taking part in some or all of these activities—either because they'd like to use GOKb for a specific local project or because they see broader value the data we're providing to the community.

Another avenue for library partners that we'd like to explore is management of consorital packages. Often consortia struggle to keep track of the titles that are part of their custom deals and to communicate that information across members and to other knowledge bases. GOKb can serve as a central environment to host these custom title lists. It provides a publicly accessible location where all members of a consortium can go to download lists for internal use or distribution. Plus, any work that consortiums put into cleaning up the titles, platforms, and organizations related

to their packages will benefit the broader community.

Publisher partners are also important to GOKb, as they're the source of much of the data that makes up the e-resources metadata supply chain. There are a number of ways that GOKb can benefit publishers and the broader community at the same time. If a publisher is interested providing KBART-compliant files to knowledge base vendors, but doesn't have the time or expertise to implement the best practice themselves, GOKb data managers can help with the process by reformatting, vetting, and enhancing publisher metadata. GOKb can also function as a useful place to for publishers to host their KBART files so that they don't have to maintain them on their own web sites. The beauty of this model is that, once their data is properly formatted and stored in the GOKb, publishers can distribute this data to other knowledge base vendors, thereby pushing out high-quality data across the supply chain.

Finally, GOKb is also interested in partnering with library vendors, particularly other knowledgebase providers. GOKb is not a competitor to these services, but rather a collaborator that can help others working in the same space supplement their data. As GOKb grows, knowledgebase vendors are welcome to repurpose GOKb's data for their own products, and we especially hope they will include GOKb identifiers as a part of their data sets—and maybe even contribute their identifiers to GOKb in return. Sharing identifiers will help improve GOKb's co-referencing service, and allow users to crosswalk their data in situations like system implementations and

migrations. We also may be able to help other knowledge bases improve their overall data quality by pushing out files that have been vetted by the community, or even helping to clean up specific data sets in exchange for making them publicly available through GOKb.

Next Steps for GOKb

GOKb will make its web application, APIs, and co-referencing service available to the community as part of a public preview period scheduled to begin in early 2015. The preview will showcase the latest version of the GOKb software and contain a small amount of seed data illustrating the different record types. Additional data will continue to be added to the system throughout Phase 2 as more partners join the project. While we recognize that there is still a great deal of work to be done before GOKb has enough data to be truly useful to the community, we believe it's important to open up the service to a wider audience early on so we can incorporate feedback from potential users and contributors. During the public preview, users will be invited to use GOKb to search and browse metadata, export package information, and experiment with the system's API and coreferencing service. We are very interested in receiving comments and suggestions from users during this time.

Information about the GOKb public preview will be made available at <http://gokb.org> in early 2015. For more information about the project, please contact Kristen Wilson at kristen_wilson@ncsu.edu.