The Global Open Knowledgebase (GOKb): open linked data supporting electronic resources management and scholarly communication

Based on a breakout session to be held at the 38th UKSG Annual Conference, Glasgow, March-April 2015

The Global Open Knowledgebase (GOKb), a partnership between Kuali OLE and Jisc, is an open data repository of information related to e-resources as they are acquired and managed by libraries. Because GOKb tracks change over time – titles, publishers, packages – and can be used to populate other tools with data, it is changing the way that libraries think about the knowledge base. Propagation of authoritative and enhanced data about e-resources has the potential to benefit all actors in the supply chain from publishers to libraries. GOKb can also serve as a platform to explore how open knowledge base data can contribute to the broader scholarly community infrastructure, particularly around open access (OA).

Background

As the products of scholarship that libraries acquire have moved from print to electronic, even as the article and journal still largely resemble their print ancestors, the business models that govern how they are sold, distributed and managed over time and their descriptive metadata have changed fundamentally. Current business models employ complex and highly customizable packaging, differential pricing and licensing, all of which are enabled by the fact that these are digital files with very low marginal distribution cost. The metadata ecosystem that moves data from publishers to libraries has spawned new roles for intermediaries (e.g. subscription agents and knowledge base providers) and standards initiatives (e.g. KBART). The focus of knowledge base providers and KBART is on managing and standardizing data to support discovery; however, neither addresses certain attributes of titles or packages related to current publisher business models that libraries need to support their management over time.

The Global Open Knowledgebase (GOKb) makes freely available KBART-formatted, standardized metadata about e-resources as well as enhanced metadata as described above to support management of e-resources. As an open data project that provides authoritative metadata about key publication entities, including change over time, GOKb has the potential to go beyond library management purposes and become a supplier of data to other applications in support of scholarly communication.

International collaboration

GOKb is a collaboration between the Kuali OLE partners[1] and Jisc, with support from the Andrew W Mellon Foundation, and is being developed by Knowledge Integration Ltd (K-Int) and Sero Consulting. The first public beta release was made available in December 2014[2]. The GOKb software is

KRISTIN ANTELMAN

University Librarian
California Institute of Technology

KRISTEN WILSON

Associate Head, Acquisitions and Discovery
North Carolina State University

'GOKb has the potential to go beyond library management'

made available under the Education Community License and GOKb data is freely available under a CC-0 license.[3]

GOKb was first conceived as the knowledge base for Kuali OLE. Because KnowledgeBase Plus (KB+) was already in development at that time, it was clear that the two projects had significant overlap and each could make better and more rapid progress through collaboration. There was also the broader recognition that shared solutions and approaches beyond national boundaries was the forward-looking approach, as both the data and the challenges libraries and institutions face in using this data are largely the same. As a result, the GOKb and KB+ projects are collaborating closely on future development. KB+ is planning a transition to use GOKb data and GOKb is leveraging user needs analyses conducted to guide development of KB+ to inform its Phase 2 development. GOKb is also providing data as originally intended to the Kuali OLE library management system (LMS) to support electronic resources management (ERM) for users of that software. Both KB+ and GOKb will consume GOKb data using the standard publicly available GOKb application programming interface (API).

'shared solutions and approaches beyond national boundaries'

The international value proposition has proven to resonate beyond GOKb's original partners in the US and UK with serious interest coming from Germany, France, Sweden and Japan in contributing data to GOKb and/or participating in the project.[4] International partners will be able to designate themselves as 'curators' of publishers or packages, which they take responsibility to manage over time. Because it is more challenging to maintain currency and comprehensiveness across all knowledge bases for non-English publishers and packages, it is considered that enabling deposit and maintenance by those stakeholders with the greatest incentive and knowledge of the data is the most promising approach to disseminating this data widely to all knowledge bases and, through them, to all libraries.

A GOKb core value proposition is to make well-structured data available and reusable. GOKb's support for those two tenets of open data is described in more detail below.

GOKb data

GOKb manages metadata about Packages, Titles, Platforms, Organizations and Licenses that is true at the global (national or regional) level. It does not manage data true only at the local level, such as entitlements or negotiated license terms. A key distinction between GOKb and a discovery-based knowledge base is that GOKb tracks changes over time: titles moving between publishers or in and out of packages; platform migrations; organizational changes and their relationships to titles, providers and platforms; identifiers associated with key entities that enable systems integration.

'GOKb tracks changes over time'

GOKb has developed a data model based on the widely used 'Bill of Materials' approach that is both hospitable to the variety of data and relationships involved in the target domain and also fundamentally geared to expressing clear relationships between data objects and therefore to publishing linked data.[5]

GOKb metadata can also support data manipulation and quality control for other applications related to electronic collections, such as usage management tools. This has already been demonstrated by the use of KB+ data to push canonical data about titles and organizations to COUNTER users to normalize data received in usage reports, saving dozens of hours of staff time.

'GOKb supports management of multiple identifiers for titles and packages'

In order to support integrations with multiple external systems and data suppliers, as well as to support day-to-day management of e-resources, system implementations and migrations, GOKb supports management of multiple identifiers for titles and packages. Those identifiers may be global (ISSN) or proprietary (e.g. publisher or intermediary title IDs). The GOKb co-referencing service, described in more detail below, enables query, including machine-based query, of GOKb data given any set of identifiers in hand.

GOKb also mints identifiers for each entity, most of which are of use internally, but two of which have broader potential. The first is the title instance package platform (TIPP) ID, which represents what libraries license and purchase and for which they manage usage statistics. If this identifier appeared in purchase orders, invoices and local LMSs, commonly desired management tasks, such as calculating cost per use and linking titles to COUNTER statistics, could be more easily performed. The second identifier of interest more broadly is the authorized organization name identifier, which is discussed in more detail below in the context of the Organization Name Linked Data service.

GOKb interfaces

GOKb has three principal interfaces: an OpenRefine interface (through a custom GOKb extension) to manipulate and validate data for batch ingest into GOKb; a web version for viewing and editing individual or sets of data elements directly; and a broadly functional API to extract the data from GOKb (see Figure 1).

OpenRefine with GOKb extension

The principal interface to GOKb for data loading and updating data is OpenRefine, an open source tool that enables collaborative workspaces for data management. Through a custom GOKb extension, OpenRefine supports the import of data files from publishers and vendors and their normalization by GOKb data editors to the GOKb data specification, which is based on KBART (see Figure 2). Once validated, data is ingested into the GOKb application, thus guaranteeing that data loaded into GOKb is complete and well structured.

Web interface

The GOKb web interface[6] (see Figure 3) enables the data to be searched and viewed and for editors to manage review tasks that are generated as a result of the OpenRefine ingest process and edit data directly. Enhanced functionality to support co-operative management of the data is planned

'Enhanced functionality to support co-operative management of the data is planned'

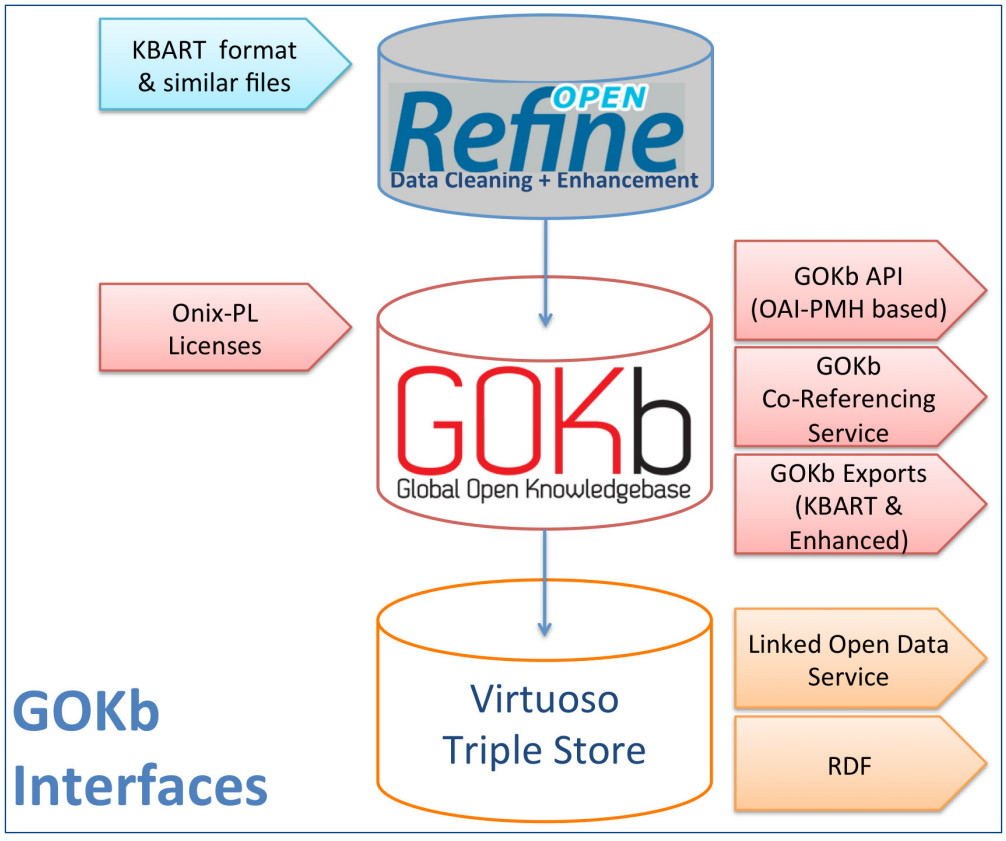

Figure 1. GOKb interfaces

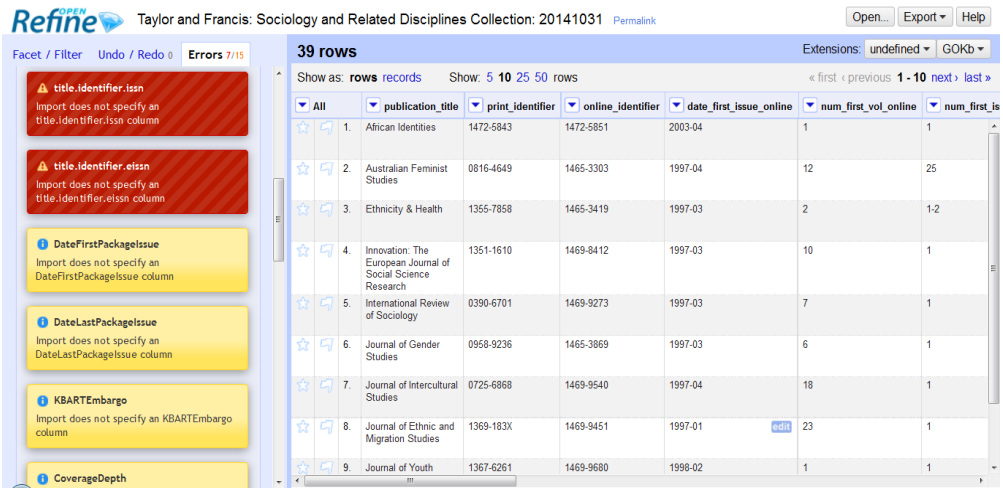

Figure 2. OpenRefine with GOKb extension

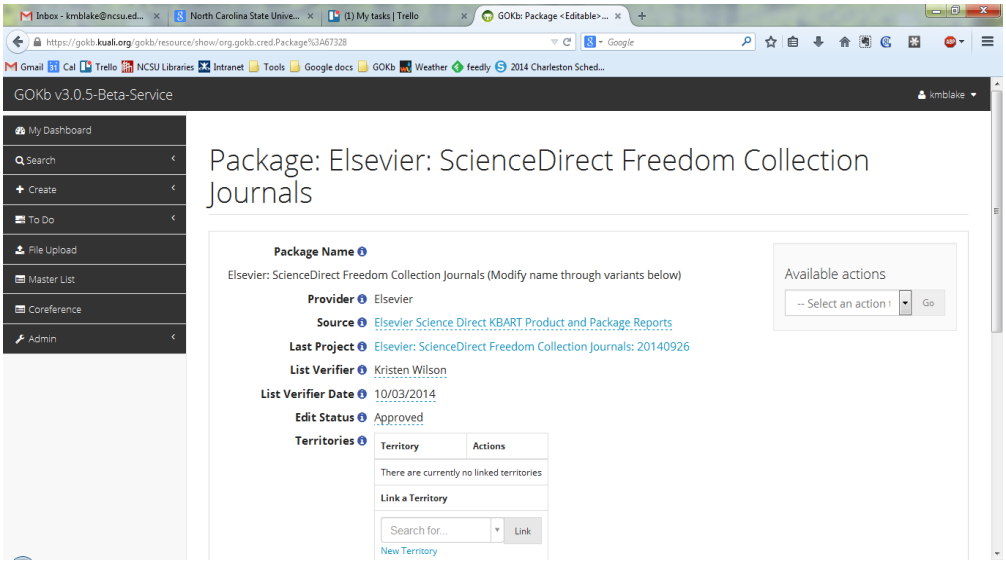

Figure 3. GOKb web interface with package selected

once requirements are better understood and as more editors, including international partners new to the project, begin working in the GOKb environment.

The web interface will also serve as a useful resource across a range of functions for staff in any library, from assessing title lists to troubleshooting access problems caused by changes in titles, organizations or platforms.

API

The GOKb API allows external services to extract data from GOKb. The API also enables credentialed users to contribute changes back to GOKb, thus creating a scalable, systematic contribution path for updated data (i.e. bypassing the OpenRefine ingest process). The API, which uses the OAI-PMH protocol, is being developed and tested with the first core applications using GOKb data, KB+ and Kuali OLE.

Co-referencing service

The GOKb co-referencing service supports linking multiple identifiers with key GOKb entities, allowing users to create crosswalks between identifiers. Users can search GOKb for any identifier and return a list of all the other identifiers associated in some way with that same component, including limited by a specific namespace. It should be noted that the service is not a registry; it makes no assertions of identifier correctness or semantic equivalence between an entity in GOKb (e.g. organization) and an analogous entity in

another system. The co-referencing service is available both through the GOKb web interface and through JSON or XML queries through the GOKb API.[7]

Benefits of open knowledge base data across the supply chain

Propagation of authoritative and enhanced data about e-resources has the potential to benefit all actors in the supply chain from publishers to libraries.

'the potential to benefit all actors in the supply chain'

An open repository has the potential to support a more efficient data flow across the system and bolster more widespread use of the evolving KBART recommended practice. GOKb can also function as a useful place to host KBART data so that publishers don't have to maintain it on their own websites or maintain multiple bilateral distribution channels to knowledge base providers and libraries (see Figure 4). The advantage of this model is that publishers could direct other knowledge base vendors to GOKb as the canonical source of their data, thereby leveraging community effort to push out high-quality data across the supply chain.

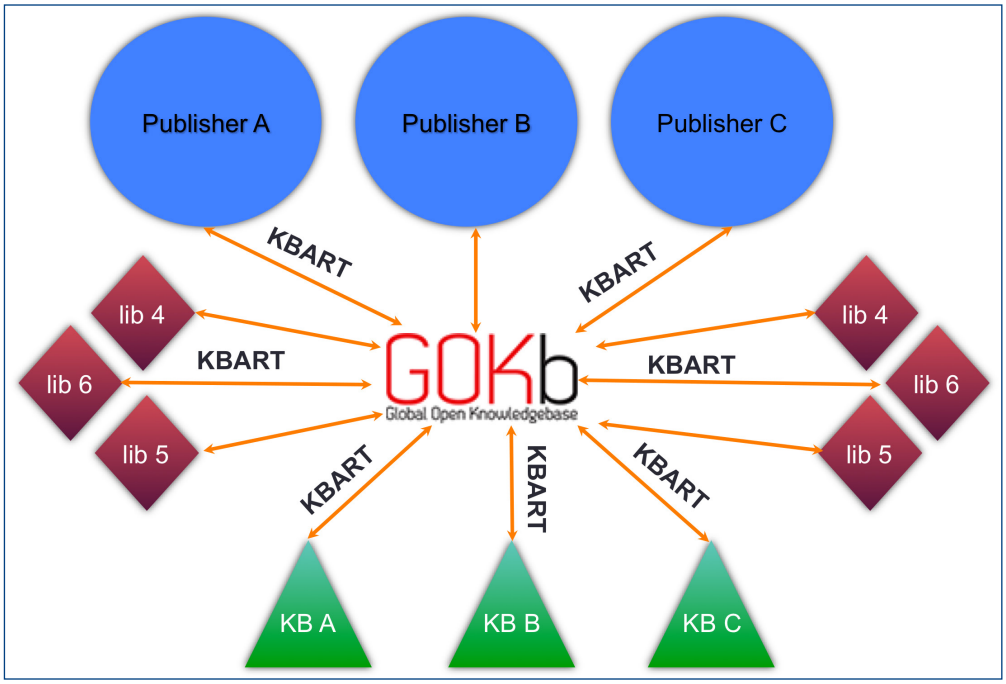

Figure 4. GOKb support for data dissemination

Linked open data

The linked open data schema (see Figure 5) is a visual representation of the GOKb linked data model, still in development, that will be used to represent GOKb data as resource description framework (RDF) triples.[8] The schema is an almost direct mapping to the GOKb relational data using properties from existing linked data vocabularies such as Dublin Core, SKOS and BIBFRAME, as well as properties that will be created for a new GOKb linked data vocabulary. Using this data model, the GOKb linked data store will be maintained and updated by a background task that transforms GOKb's relational data into named RDF graphs that can be expressed in various serializations of RDF, such as RDF/XML, N3, N-Triples, and JSON-LD.

While this model simplifies some of the complex semantics in the relational data for a direct expression better suited for the RDF triple model and SPARQL queries, the benefit of this approach is that GOKb can make actionable and repurposable statements about key entities of interest to libraries and other shared services, such as the title as supplied in a given package on a given platform (the TIPP). The project will also co-ordinate with the KBART working group to release KBART2-formatted data as RDF.

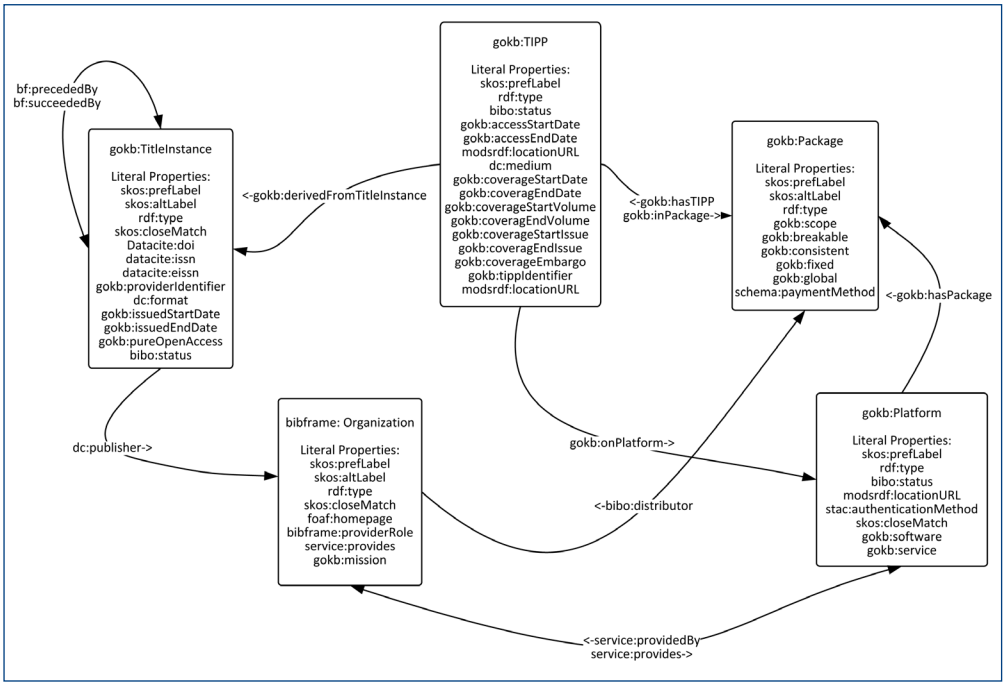

Figure 5. GOKb linked open data schema

Organization Name Linked Data (ONLD)

Metadata describing organization names is highly inconsistent, being based on the source and purpose for which data is supplied. There are several strong use cases for having practical control over these names within library systems. Authorized names and identifiers for organizations support what should be simple management queries such as identification of all titles from a given publisher or supplier. Additionally, with organizational mergers and acquisitions, formerly independent publishers often become imprints owned by another publisher, with the acquiring organization being the entity with whom the library would typically maintain a license agreement, but the acquired organization is still important for title identification and management purposes.

The first publication of GOKb linked open data builds on the Organization Name Linked Data (ONLD) service published by North Carolina State University (NCSU) Libraries.[9] The data is derived from a tool originally developed at NCSU to manage variant forms of name for journal and e-resource publishers, providers and vendors in a local ERM. The data is represented as RDF triples using properties from the SKOS, RDF Schema, FOAF and OWL vocabularies.

GOKb in support of open access

GOKb can currently support knowledge and use of open access (OA) resources in several ways. These include loading OA resources, both journals and e-books (with the potential for linkable chapter-level entities), as well as recording OA attributes of these resources. Additionally, GOKb data can help support increasingly important collection analysis use cases related to the intersection between OA and subscribed collections.

In discussing how GOKb can support OA, it is important to note that GOKb manages data at the title level; it does not manage data at the article level. While GOKb's Bill of Materials data model could be readily adapted to support articles (or e-book chapters) – and this work will be piloted in Phase 2 of GOKb development in 2015 – it is an open question as to whether GOKb is the best and most appropriate home for a freely available repository of article metadata. Still, there are several ways in which GOKb as currently scoped can support institutional management of the transition to OA.

'GOKb manages data at the title level; it does not manage data at the article level'

Analysing the intersection between open access and subscribed collections

It goes without saying that libraries still make substantial investments in traditional published collections. At the same time, they are faced with the need to realign internal resources toward support for functions closer to *process* than *product* in the research lifecycle.[10,11] It is imperative therefore that library staff are able to conduct standard management tasks against electronic collections with maximum efficiency and analyse the relative value of resource investments. It is still the case that a typical library's infrastructure, in terms of data, software tools and integrations, is not adequate to properly analyse electronic collections. One obvious example is the complexity of calculating cost per use, which requires mapping the title (at the instance level, which is what is purchased) with the cost for that title (which may not be properly itemized or referenceable on the invoice) and the usage statistics (which may not identify titles as they are identified in internal systems).

'a typical library's infrastructure … is not adequate to properly analyse electronic collections'

An emerging use case for collection analysis is to support activities related to the transition to OA. A fiscally relevant question libraries should be able to answer is to what degree the subscribed collection overlaps with OA content. To take a simple example, a library seeks to compare two toll-access journals that are roughly equivalent in importance and cost. The content of Journal A over the past five years is 5% OA (through author self-archiving, i.e. green OA) and the content of Journal B over the past five years is 95% OA (through 100% delayed OA after 12 months and 15% OA in the most recent year through a combination of hybrid OA and author self-archiving). Clearly, the relative value of Journal A is much greater to the library's community than Journal B. Libraries will be at a disadvantage in negotiating subscriptions and packages if the subscribed/OA overlap cannot be taken into account relatively efficiently. This is particularly critical now while there is significant concern about 'double-dipping' with hybrid OA journals and there is no consensus among publishers or libraries about how the extent of OA content in hybrid journals affects subscription costs. One current model, which discounts the subscription cost only to the institution of the author who pays to make their article OA, appears to be a form of double-dipping; it should be of particular concern across the library community that it not become the de facto model.

'there is significant concern about 'double-dipping' with hybrid OA journals'

In analysing the impact of OA on subscribed resources and their relative value, one must look separately at the different modes of OA currently in evidence: 'gold' journals (fully OA), hybrid OA journals (toll journals whose authors may purchase 'gold' OA), delayed OA (journals that become fully OA after an embargo period), and 'green' OA (author self-archiving). In terms of collections analysis, gold OA journals stand relatively to the side; they can be made part of a library's discovery environment but are not analysed as part of its licensed and subscribed collection. Hybrid, delayed and green OA present a far greater challenge, especially as the distribution between these types changes with years from publication and increased adoption of all OA options over time.[12] By definition, hybrid, delayed and green OA intersect with the licensed and subscribed collection (see Figure 6). The challenge of course is that while gold, hybrid and delayed OA can describe journals, ultimately OA is defined at the article level. An additional complication is that libraries manage aggregates: bundles of articles (journals) and bundles of journals (packages).

How can GOKb help with these collection management challenges? The GOKb TIPP entity properly describes what libraries need to map cost and usage statistics data. This well-structured and referenceable data (through the TIPP ID) can be used to simplify the usage analysis. The resultant significant time saving using data in KB+ has been demonstrated by the use of the KB+ data to help normalize data taken from Jisc Journal Usage Statistics Portal (JUSP). More generally, GOKb can provide additional information around journal titles, change events and organizations that an LMS or ERMS could utilize to better structure its data for analysis over time.

'well-structured and referenceable data (through the TIPP ID) can be used to simplify the usage analysis'

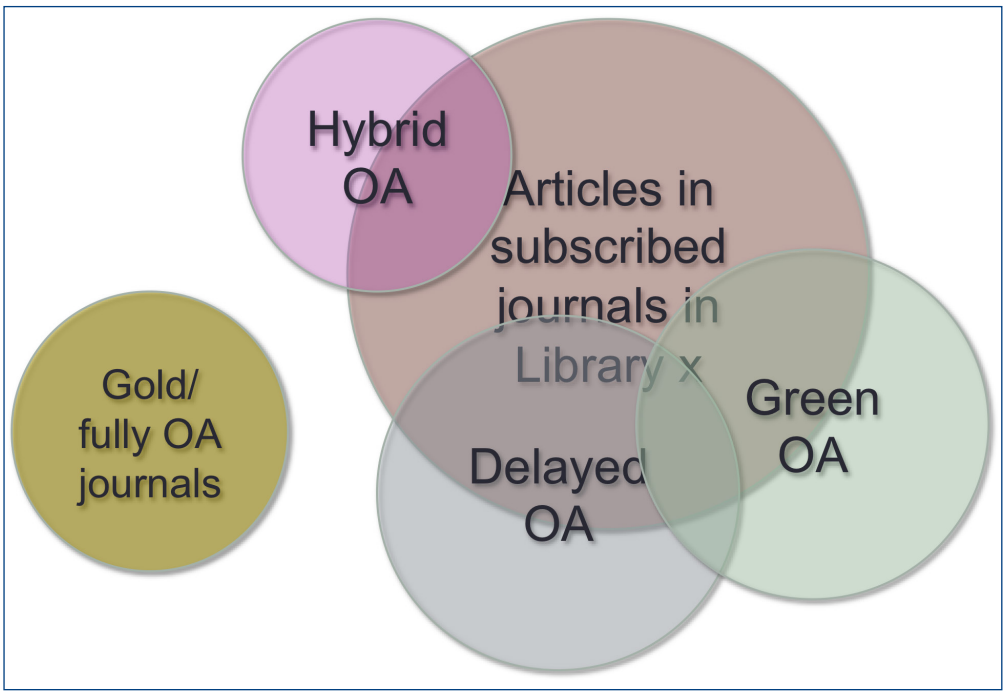

Figure 6. Intersection between open access and subscribed collection

For green OA, as both GOKb and SHARE[13] become more developed, the GOKb API and co-referencing service could provide useful data to support comparison of the local collection with green OA available across institutional repositories worldwide. GOKb indicates whether journals are gold/fully OA, hybrid OA or delayed OA. One could use the co-referencing service to map between the future SHARE registry and a given subscribed collection by taking journal-level information from citations in SHARE and comparing those with an institution's subscriptions. The result would give an estimate of green/gold overlap with the subscribed collection. But given the haphazard nature of what gets deposited into institutional repositories, only a comprehensive article registry, with consistently applied OA indicators at the article level,[14] would enable the library to conduct a thorough overlap assessment.

Potential futures

Higher education currently faces two significant challenges around scholarly communication: compliance with national and funder requirements to demonstrate the impact of research funding in terms of scholarly output and to transition that output to OA. Institutional and shared system-wide infrastructure will both have to be developed to support this increased level of institutional engagement with the research lifecycle and its metrics. Apart from the library's role in managing article processing charges (APCs), if libraries were able to answer the analytical questions related to OA and subscribed collections described above, there would be the opportunity to demonstrate that library materials inflation is not simply a library or institutional challenge but one that should be approached in conjunction with demonstrating research impact and transitioning to OA.

GOKb has the potential to evolve to provide support for additional pieces of the research infrastructure, specifically related to the transition to OA. The project will respond to the findings arising from the Jisc Monitor work exploring shared services for open access.[15] This could include additional OA title types and enhancing GOKb to support articles and e-book chapter components.

'support for additional pieces of the research infrastructure, specifically related to the transition to OA'

GOKb can provide foundational data and also model new approaches through its extensible data model and identifier co-referencing service. GOKb is also well positioned to incorporate additional data elements and those involved in GOKb are enthusiastic about the opportunities to collaborate with other projects engaged in building an enhanced scholarly communication infrastructure.

Competing interests:The authors have declared no competing interests.

References and notes

1. Kuali OLE partners (as of January 2015) are: Indiana University, Duke University, University of Florida, Lehigh University, North Carolina State University, University of Chicago, University of London Library Systems Association (Bloomsbury), University of Maryland, University of Pennsylvania and Villanova University.
2. The first public beta release of GOKb: http://gokb.kuali.org/gokb (accessed 20 January 2015).
3. GOKb software is available on a GitHub project site. GitHub: https://github.com/k-int/gokb-phase1 (accessed 1 January 2015).
4. The project's partnership model is documented on the website: http://gokb.org/partners (accessed 6 February 2015).
5. For a more in-depth discussion of the data model and elements, see Wilson, K, Building the Global Open Knowledgebase (GOKb), *Serials Review*, 2013, 39(4), 261-265; DOI: http://dx.doi.org/10.1016/j.serrev.2013.10.002 (accessed 1 January 2015).
6. The GOKb web interface, ref. 2. http://gokb.kuali.org/gokb
7. For more detail, see Wilson, K, Bringing GOKb to Live: Data, Integrations, and Development. *Proceedings of the 34th Charleston Library Conference*, forthcoming.
8. This preliminary model will likely undergo significant revisions before it is implemented and a full vocabulary specification is released. The endpoint and sample SPARQL will be documented on gokb.org.
9. Organization Name Linked Data: http://www.lib.ncsu.edu/ld/onld/ (accessed 1 January 2015).
10. Dempsey, L, Malpas, C and Lavoie, B, Collection Directions: The Evolution of Library Collections and Collecting, *portal: Libraries and the Academy*, 2014, 14(3), 393-423.
11. Lavoie, B, Childress, E, Erway, R, Faniel, I, Malpas, C, Schaffner, and van der Werf, T, *The Evolving Scholarly Record*, 2014, OCLC Research: http://oclc.org/content/dam/research/publications/library/2014/oclcresearch-evolving-scholarly-record-2014.pdf (accessed 1 January 2015).
12. Björk, B-C, Roos, A and Lauri, M (2009). 'Scientific journal publishing: yearly volume and open access availability' *Information Research*, 14(1) paper 391: http://InformationR.net/ir/14-1/paper391.html (accessed 23 January 2015).
13. SHared Access Research Ecosystem (SHARE): http://www.arl.org/focus-areas/shared-access-research-ecosystem-share#.VJpBPcAA5w (accessed 1 January 2015).
14. NISO Access and License Indicators Working Group: http://www.niso.org/workrooms/ali/ (accessed 1 January 2015).
15. Monitor findings are due to be published May 2015. GOKb's co-referencing service could serve as a model for 'GUIDE' (Gathering Useful IDs Early), an area of work identified by Jisc Monitor and what in essence is a co-referencing service.

Corresponding author: Kristin Antelman
University Librarian
California Institute of Technology (Caltech), 1200 E California Blvd MC 1-32, Pasadena, CA 91125 USA
E-mail: kristin.antelman@caltech.edu

ORCiD: http://orcid.org/0000-0001-7604-6951

To cite this article:

Antelman, K and Wilson, K, The Global Open Knowledgebase (GOKb): open linked data supporting electronic resources management and scholarly communication, *Insights*, 2015, 28(1), 42-50; DOI: http://dx.doi.org/10.1629/uksg.217

Published by UKSG and Ubiquity Press on 5 March 2015