Smart Libraries Newsletter

News and Analysis in Library Technology Developments

ALA TechSource
alatechsource.org

50 East Huron Street, Chicago, Illinois 60611-2795, USA

Smarter Libraries through Technology

Addressing Disruption in the Discovery and Access to Scholarly Literature

By Marshall Breeding

Finding and gaining access to scholarly articles can be complex. Although libraries work hard to provide the best environment possible for the discovery and access to scientific literature, many gaps remain in terms of content available and in simplicity of use. The eventual transition of the realm of scholarly publishing to open access business models promises broader availability and simplified access. In the short term, however, it presents challenges to the resource management and access ecosystem based on library subscriptions.

Researchers can encounter several different scenarios as they attempt to gain access to a scholarly article, book chapter, or other information resource. Today, the majority of articles are restricted to paid institutional or individual subscribers. An increasing proportion are published as open access and can be viewed by anyone regardless of a paid subscription. Many subscription-based articles are also available in open access repositories. In some cases, the open access copy may be a pre-print, or it may be the final version. Agencies and foundations that fund scientific research increasingly stipulate mandates for any articles or reports based on that research be made available on open access repositories, even if they are also published in subscription-based publications.

Even if the scholarly publishing arena were to achieve a complete transformation to open access models, much of the body of existing literature would remain behind paywalls. Libraries should expect to support multiple models of access indefinitely.

This mixed mode of publishing complicates the task of the index-based discovery services that most libraries invest in. These products were originally designed around the prevailing subscription model. They populate their indexes with citation metadata and full text representing the broadest possible array of scholarly and professional content. Most libraries will configure their discovery service to reflect their active subscriptions. A library may opt to suppress resources not available within their subscriptions or enable requests for these items through some type of expeditated electronic document delivery service.

Gaps continue to remain in the coverage of index-based discovery services for subscribed resources. These products depend on maintaining central indexes with metadata and full text provided by providers of scholarly content. Although most content providers contribute content to each of the major discovery services, some gaps in participation remain.

Discovery services must also handle open access content. Since all users are eligible to access these resources regardless of institutional affiliation, they can be included in the results returned by the discovery service. The challenge lies in identifying open access content and providing reliable links. Some hybrid journals include both restricted and open access articles. Although NISO has issued a Recommended Practice on Access and License Indicators that includes identifying open access content, its implementation throughout the scholarly publishing and discovery service ecosystem remains imperfect and incomplete.

Another area for improvement in discovery services involves articles published in subscription-based journals that have open access versions available elsewhere. There are services available to identify open access articles, such as Unpaywall Data (formerly known as oiDOI). Link resolvers can be configured to tap into these services, though the implementation of this capability is currently somewhat experimental.

Libraries usually maintain a proxy server to enable researchers

IN THIS ISSUE

ISSN 1541-8820 May 2018 Volume XXXVIII Number 5

to gain access to restricted resources from outside the institutional network. Although initiatives are underway to eventually move to other authentication mechanisms, IP authentication continues to dominate as the most common way to provide access to restricted resources to off-campus users. These proxy services work best when the user accesses the resources through the search tools provided by the library and can thus embed the specially formed links needed to access IP-authenticated resources from outside the institutional network address space. Users will usually need to authenticate to the library or institutional network as they use these proxied links. These access mechanisms can be more difficult to invoke when using Google Scholar from outside the university network.

The fragility of the official processes established for researchers to gain access to the scientific literature has led to the popularity of Sci-Hub, a site which provides pirated copies of articles in violation of the copyrights of publishers. Despite the ethical and legal concerns, Sci-Hub attracts considerable use since it provides access to a large proportion of scientific literature without the complications of access restrictions of the official ecosystem. Sci-Hub currently contains about 65 million articles and represents a significant portion of the content in the scholarly publishing arena. Despite legal actions taken by the multiple publishers, the site continues to function since it operates in international domains outside the reach of legal enforcement. Sci-Hub has caused a major disruption in the scholarly publishing arena. The current subscription-based business model, which turns away those unable to pay, as well as the body of search tools, linking technologies, and access methods that can be difficult to navigate invites disruption. Publishers and other organizations are increasingly interested in simple, efficient, and legal ways to access these materials. In the music arena, legal services like Spotify eventually supplanted copyright-infringing sites such as Napster. The scholarly communications arena likewise stands in need for new business and technical models able to solve the issues that fuel interest in Sci-Hub.

The emergence of browser extensions designed to facilitate access to scientific articles can be seen as at least one step in the direction of addressing these challenges. These extensions, once installed, can enable access to articles through a single click regardless of publication model. These tools simplify the work of the researcher and benefit publishers. Providing a better experience for finding official copies of articles, including open access versions and pre-prints, helps deter interest in rogue services.

This issue of *Smart Libraries Newsletter* examines this genre of web browser extensions as a strategy for providing simplified access to scientific literature. These extensions interoperate with existing components provided by libraries, including discovery interfaces, link resolvers, and proxy services. They also supports researchers working with Google Scholar and other non-library search services. The acquisition of Kopernio by Clarivate Analytics illustrates the growing role of these browser extension in the scholarly communications ecosystem.

Clarivate Analytics Acquires Kopernio

In recent years, several companies in the scholarly publishing industry have expanded their portfolios with new workflow and analytics products and services. In a move consistent with this trend, Clarivate Analytics has acquired Kopernio, a company that has created a web browser extension that facilitates access to scholarly articles. The acquisition of the product further extends the tools Clarivate Analytics offers in scholarly communications arena.

Kopernio: A Browser Extension for Viewing Articles

Kopernio bills itself as doing something similar for scientific literature as Spotify does for music, providing a simple and legal way to access and save scholarly articles. Services such as Google Scholar make it relatively easy to discover a scholarly article, but gaining access to the PDF of the article itself can be much more challenging. Researchers often encounter dead-ends, paywalls, or authentication issues. Kopernio was born out of the frustration researchers encounter in the complicated and unreliable processes they face in finding the information they need for their work.

Kopernio belongs to the genre of browser extensions that facilitate access to scholarly articles. Once installed in the user's web browser, it automates the steps involved in obtaining the PDF of articles. It is not a search or discovery service, but a tool that streamlines access to the PDF of an article with a single click once an article of interest has been identified. For articles not available through library subscriptions, it attempts to locate versions that may be available as open access.

The key function of Kopernio lies in helping a researcher find the best version of the full text of an article—almost always provided in PDF format, independently of the search tool or interface used to find it. It works behind the scenes, activating a visual bar in the browser, which presents a "View PDF" button that loads the full text of the article with a single click. Kopernio takes care of the details behind the scenes to find the best PDF version of the article, formulating the correct version of the link needed to view it.

Kopernio works with common scholarly search environments, such as Google Scholar and PubMed. Its work begins once the researcher lands on an article of interest. Without the browser extension, finding the full text PDF of an article from its citation can be a convoluted process, depending on whether it is available through the library's subscriptions or if the researcher is working from the campus network, or other complications.

Legal Access to Resources

Kopernio operates entirely within the realm of the official scholarly communications ecosystem and does not rely on pirate sites that may not respect the copyrights of publishers or scholars. It does not circumvent the access restrictions on articles to which the researcher's institution does not subscribe, but rather attempts to find open access versions that may have been deposited on disciplinary repository, such as ArXiv, an institutional repository, or on the personal blog of its author.

Researchers continually deal with obstacles in the scholarly publishing ecosystem. Scholarly articles may be available directly from publishers via library subscriptions, on disciplinary pre-print servers, institutional repositories, or posted on blogs of researchers. Much of the literature resides within proprietary resources that require institutional subscriptions for access. Those attempting to access these articles without an active subscription will encounter a paywall prompting for payment. Libraries aim to purchase subscriptions to the most important resources and provide basic mechanisms to enable access to those affiliated with the institution to gain access. Kopernio interoperates with link resolvers that libraries provide for the resource covered within their subscriptions. For those not available within institutional subscriptions, it determines if an open access version is available.

Even when the library has subscribed to the resource, affiliated users must be authenticated to gain access. Researchers can record their institutional credentials securely within Kopernio. These credentials are then available to login to the library's proxy service or the institutional authentication service to gain access to restricted resources available through library subscriptions. Kopernio avoids the need to install a VPN (virtual private network) client or other complex mechanisms to access subscribed resources from off-campus.

Kopernio facilitates access to PDF copies of articles regardless of the search environment employed. A researcher can begin with general search tools, such as Google or Google Scholar, to find the landing page for an article of interest. It automatically detects and activates links for articles covered by library subscriptions and offers suggestions for open access versions for articles not covered. Once the browser loads the page with the citation, it works behind the scenes to find the PDF of the article:

- If the researcher is working from their own institutional network and their library has obtained a subscription, then no additional processing is needed for Kopernio to present the "Access PDF" button. Kopernio can also locate pre-prints or other versions of the article if the researcher is interested.
- If the researcher is working from outside the institutional network, Kopernio can use saved authentication credentials to transparently login to the library's link resolver to present proxied versions of the link to the desired article.
- If the researcher's library does not have an active subscription with the publisher of the article, Kopernio will perform a search against the oaDOI database and other relevant resources to attempt to locate a copy of some version of the article on an open access repository.

Kopernio cannot provide a link to the PDF in all cases. No library subscribes to every scholarly resource, and open access versions are not always available. Its performance, however, should improve over time in tandem with the growth in open access publishing.

In addition to its core features for viewing PDF versions of articles, Kopernio includes additional features for storing and sharing them. It provides a storage area for the user, called a locker, to place any articles accessed for future reading. The article locker resides in cloud storage provided by Kopernio, so that it can be accessed across a user's multiple devices.

Kopernio Company Background

Kopernio traces its development to "Canary Haz" developed by Dr. Peter Vincent and Benjamin Kaube, both affiliated with the Imperial College London, to simplify finding scientific papers. Canary Haz Limited was established in July 2016. Kopernio was founded in July 2017 by Jan Reichelt and Benjamin Kaube as the successor company to Canary Haz. Reichelt

served as its Chief Executive Officer. In February 2018, Kopernio received funding from Innovate UK, an agency of the UK government to accelerate innovation. At this time, the product was renamed from Canary Haz to Kopernio. Its acquisition by Clarivate Analytics in April 2018 represents a very rapid trajectory from the development of a prototype product, the launch of a startup, and initial rounds of investment to the culmination of its strategic acquisition by a major commercial enterprise.

Clarivate Analytics Background

Clarivate Analytics is a major corporation involved in scholarly publishing, offering competing products and services to Elsevier, Digital Science, and others. The company traces its corporate history to the Intellectual Property and Science division of Thompson Reuters. In 2016, this part of the organization was separated and acquired by investment firms Onex and Baring Private Equity Asia for $3.55 billion. The company is based in Philadelphia, has over 4,000 employees, and is active in over 100 countries. Its products in its Scientific and Academic Research group include Web of Science, Endnote, InCites, and Publons.

Details of the Acquisition

Clarivate Analytics will become the primary investor in Kopernio and will accelerate its development and integrate it with its other products and services. The founders of Kopernio will take new roles within Clarivate Analytics.

Jan Reichelt joined Clarivate as the Managing Director for Web of Science and will continue to oversee the strategic development of Kopernio. He had previously co-founded Mendeley along with Victor Henning in 2008. Mendeley was acquired by Elsevier in April 2013.

Ben Kaube joined Clarivate as the Managing Director for Kopernio. Kaube, along with Freddie Witherden had previously founded and served as the Chief Technical Officer of Newsflo, a service to monitor the coverage of scientific research in scientific and popular news media. Newsflo was acquired by Elsevier in January 2015.

Competing and Related Products

Kopernio belongs to a growing genre of products employing similar techniques and technologies. Several browser extensions have been developed to address various aspects of the challenges in accessing scholarly articles. Each addresses the issues differently with different features and capabilities.

Unpaywall

Unpaywall is a free browser extension created by Impactstory, a non-profit organization specializing in developing tools for open access scholarly publications. It offers the same basic capability as Kopernio in enabling researchers to bypass paywalls for restricted articles not available through their institutions subscriptions.

Impactstory developed a large-scale citation database designed to improve access to open access scientific literature. Originally known as oaDOI, but now branded as Unpaywall Data, this resource currently indexes over 100 million scientific papers and includes links to the full text for those available as open access. The resource currently links to about 18 million open access articles. This data facilitates discovery of open access resources to users of Unpaywall and other tools. Impactstory provides access to the Unpaywall Data via an API and copies of the database to other non-profits and through commercial arrangements. Link resolvers can integrate with the Unpaywall Data API to provide links to open access copies for articles where full text is not available through the library subscription.

The Unpaywall browser extension operates as a front-end to the Unpaywall Data, detecting citations and overlaying an unlocked padlock icon for those that can be accessed for free when open access copies are found.

In a related event, Clarivate Analytics entered a partnership in June 2017 with Impactstory. Clarivate contributed data for 18 million open access articles from its Web of Science citation database to oaDOI. This partnership provided a significant expansion of open access resources represented within the resource, benefiting those that use Unpaywall and other tools based on oaDOI.

Impactstory was established as a non-profit in December 2012 by Jason Priem and Heather Piwowar. The organization has received grants from the National Science Foundation and the Albert P. Sloan Foundation in support of their research and development of their services and products.

Lean Library

Lean Library offers a browser extension branded as Library Access that facilitates access to scholarly resources. It emerged from a prototype originally developed at Utrecht University Library, which gained distinction through its strategy not to offer its own discovery service but to encourage the use of Google Scholar and other disciplinary resources.

The Library Access browser extension from Lean Library facilitates access to scholarly content provided by the library. The extension sits in the background until the user opens a resource available with the library's subscriptions, and then it

presents the versions of the link from the institution's proxy server. Library Access works to make it easier to access library-provided resources, removing any complications related to authentication from on-campus and off-campus locations. Like Unpaywall and Kopernio, it also attempts to find open access copies when subscribed versions are not available.

Lean Library focuses on the perspective of the library to assist its users in gaining easier access to subscribed resources and to provide analytics related to usage that can be used to inform decisions on developing its collection of electronic resources. Lean Library appeals to libraries through its policy of not sharing or selling usage data. The data it collects is anonymized and not shared beyond the library.

Lean Library was founded by Johan Tilstra in October 2016 with assistance from UtrechtInc, an incubator that provides services and financing to new startups. Prior to launching, Lean Library Tilstra was affiliated with Utrecht University.

Perspective on the Acquisition

Clarivate Analytics competes with Elsevier as a provider of workflow and analytic products and services for scholarly communications. The acquisition of Kopernio, which was created by individuals that founded companies that Elsevier acquired, illustrates the competitive nature of the current arena of tools related to scholarly communications workflows. Within this sector, Elsevier, Digital Science, and Clarivate Analytics are each working to build product portfolios that will give them more traction and engagement with researchers in the production and consumption of scientific data and literature.

Clarivate, as a company with interests in developing products and services on analytics surrounding scholarly communications, will likely benefit from data added into its ecosystem from Kopernio. Usage data created from the perspective of a researcher's web browser represents an intimate view of the patterns of work of researchers and students, including search behavior and content resources queried and accessed. Although it is not known what data Clarivate Analytics may collect and use as it integrates Kopernio into its ecosystem of products and services, it seems natural for a company deeply engaged with research analytics to see the potential for value. Clarivate Analytics has shown involvement in this sector previously through its partnership with Impactstory. The acquisition of Kopernio as a service that incorporates Impactstory data into its internal workflow can also be seen as leveraging that investment.

Related News: Reestablishment of ISI

In February 2018, Clarivate Analytics revived the Institute for Scientific Information (ISI) as a research and innovation center within the company, addressing a variety of issues, such as developing new metrics for scientific literature. The original ISI was founded in 1960 and developed bibliographic indexing products, including the Science Citation Index and the Social Sciences Citation Index, which eventually became the core components of the company's Web of Science product. ISI was acquired by Thompson Scientific and Healthcare in 1992, taking the name Thompson ISI. Through subsequent business transactions, the heritage of ISI lives within the portfolio of Clarivate Analytics.

The new ISI division of Clarivate Analytics will be led by Samantha Burridge, Director of Strategy and Transformation.

Smart Libraries Q&A

Each issue, Marshall Breeding responds to questions submitted by readers. Have a question that you want answered? Email it to Samantha Imburgia, Associate Editor for ALA TechSource, at simburgia@ala.org.

What tools and strategies do you recommend we implement to help make our materials more discoverable online, such as through Google searches?

In addition to providing catalogs and discovery tools, it is also important for libraries to do all that they can to make their materials easily found though general search tools, such as Google, Google Scholar, Bing, Microsoft Academic, and other services. While libraries work hard to implement the best discovery tools that they can offer from their own websites, most users begin their research elsewhere. Libraries can employ a variety of techniques to improve the discoverability of their collection materials and services.

Ideally, information about libraries would be found just as easily on the web as other types of organizations. At the broad level, libraries themselves are well represented in the main search engines. In most cases, a query for the name of a

library or for "a library near me" will return successful results. But the collections of libraries are less well represented than the inventories of commercial establishments. Searching for the title of a book or an article through a general search engine will not usually turn up copies in libraries as much as e-commerce options. The performance of library collections on search engines can be improved by taking proactive measures, though the changes will be at best gradual and uneven. I am not aware of any technique or technology able to provide reliable access to a library's collection through Google and competing search engines.

Library collections are especially difficult to expose to search engines. The basic technologies libraries use to manage and provide access to their resources by default tend to be isolated from the broader information universe. Most library catalogs and discovery services initially were developed without extensive implementation of linked data and without a strong emphasis on search engine optimization. The MARC formats underlying these systems were designed for efficient storage and must be transformed to RDF, XML, or other linked data structures for better interoperability with the broader information ecosystem and for optimized performance in search engines.

A variety of search engine optimization techniques can be implemented to improve the performance of library collections in Google and other search engines.

One of the general principles of positive coverage by search engines lies in having interesting content delivered with a clear and uncluttered presentation. Each item of content needs to be presented visually to human readers and have clean and well-structured coding behind the scenes. Optimal discoverability involves developing web-based resource pages, optimized both for human readers and for computer-based harvesting and indexing services. It is essential to go beyond the presentation of resources for users to ensure that each catalog or resource page exposes the item it describes in ways that search engines can understand and ingest.

A basic level of search engine optimization involves adding metadata to each content page. It has long been standard practice to include basic metadata tags in the header of a web page such as <title> and <description> to properly label the page in a browser and to define the snippet of text that will be shown in search engine results. Additional metadata tags can provide additional citation elements for bibliographic resources. Although Dublin Core was once commonly used for citation tags, those defined by Highwire Press are now preferred. Including these tags on resource pages for bibliographic items also facilitates their use with citation managers, such as Zotero, Endnote, and Mendeley.

Each resource in the collection should be presented through a unique and persistent URL. The canonical URL for a resource should also be simple and clean (for example: https://mylibrary.org/catalog/item/48282/gone_with_the_wind). Many library catalogs present much more complex URLs for each resource, which may include session keys or other elements that may make it difficult to discern the persistent and canonical URL to reference the resource.

A good search engine optimization strategy will also include generating a sitemap, following the sitemaps.org protocol. Sitemaps inform indexing services about all the unique pages on a website and can help them index a site more efficiently. Google and other search engines do not rely on sitemaps exclusively and will also crawl through all the internal and external links within a website.

Failure to implement standard web practices can also impair discoverability. Google and other search engines increasingly penalize sites that are not friendly to mobile devices. Those that have not implemented https may be considered less trustworthy and not ranked highly in search results.

Resources can also be delivered in ways that provide structured data throughout the body of the page. Coding can be embedded surrounding content elements to provide structure without impacting visual presentation. Schema.org defines a robust set of tags to ensure that each item of content on the page can be correctly interpreted by search engines. (See https://schema.org for additional information and examples.)

Enhancing web-based resource pages to incorporate structured data elements, such as schema.org, requires the ability to access the programming or templates of the server delivering them. In the library context, this level of control is more likely to be possible for locally developed resources. In many cases, the library may not have extensive control over the presentation of each item as it is displayed.

These techniques can be implemented to improve the discoverability of library resources. Library programmers or web services librarians may be able to customize the templates or adjust the programming to provide metadata and embed tagging for structured data. This level of access may be possible for discovery interfaces based on open source software or even for a proprietary system, where the vendor enables extensive customization. Some libraries may be limited in implementing them for their core collections because they do not have sufficient control over the way that their content management systems, catalogs, or discovery tools format and deliver resource pages.

A number of commercial services for enhanced discoverability have emerged to assist libraries not able to implement their own technical strategies.

Zepheira specializes in providing services to libraries and other organizations, primarily oriented around linked data. For example, the company worked with the Library of Congress to develop BIBFRAME to represent bibliographic data in linked data as a possible successor to MARC. Zepheira has also created a service to help libraries improve the discoverability of their collection resources. The LibraryLink Network they have developed enables library materials to be represented in search results on the open web. Subscribing libraries provide copies of their MARC records that are converted into linked data using BIBFRAME and published within the library's domain, so that they can be easily indexed by search engines. For more information, see https://zepheira.com/.

Zepheira's LibraryLink services is the basis of products distributed by other library technology vendors, including:

- SirsiDynix, branded as BLUEcloud Visibility (http://www.sirsidynix.com/products/bluecloud-visibility)
- Innovative, branded as Innovative Linked Data (https://www.iii.com/products/metadata-management/#linked-data)

Demco Software offers its DiscoverLocal service to assist libraries in making their collections and events discoverable through search engine results. This service follows a similar model as that from Zepheira, where a library provides its MARC records, which are then enhanced with geolocation data and converted into linked data, and then exposed to the search engines. The service includes reporting and analytics to enable the library to measure engagement with their services. See http://www.demcosoftware.com/products/discoverlocal/ for more information.

Koios, a relatively new startup, develops technology and services to help libraries market their services through improved presence in web search results. Their current offering Libre Ads is based on a set of services to acquire Google Ads for the library available through Google Ad Grants for nonprofits. More information on Koios is available at https://www.koios.co/.

Each of these strategies and services can improve the representation and placement of a library's materials and services in search results from Google and other search engines. Don't expect immediate and dramatic change. It will take some time for metadata to be harvested and indexed by the search engines, and even longer to gain top relevancy rankings. But even if the difference is gradual, better exposure in Google and other search engines should result in increased use and impact of a library's collection and services.

ALA TechSource
alatechsource.org

Smart Libraries Newsletter
American Library Association
50 East Huron Street
Chicago, IL 60611-2795 USA
Address Service Requested

NON PROFIT
US POSTAGE
PAID
PERMIT 4
HANOVER, PA

May 2018
Smarter Libraries through Technology

Smart Libraries Newsletter

Marshall Breeding's expert coverage of the library automation industry.

Editor
Marshall Breeding
marshall.breeding@librarytechnology.org
Twitter: @mbreeding

Managing Editor
Samantha Imburgia
312-280-3244
simburgia@ala.org

Digital Access for Subscribers
journals.ala.org/sln

Production and design by the American Library Association Production Services Unit.

Smart Libraries Newsletter is published monthly by ALA TechSource, a publishing imprint of the American Library Association.

alatechsource.org