Smarter Libraries through Technology

Start-Ups Inject Innovation into the Industry

By Marshall Breeding

The library technology industry is dominated by a set of large organizations, most of which have been in business for multiple decades. These organizations provide an array of complex and sophisticated products that libraries need to support their work. They work with large numbers of libraries and create products that will have broad appeal. These large, well-established organizations have less tolerance for risk and less ability to focus on potential products based on emerging technologies or for smaller niches.

In the context of these large companies, new start-ups and smaller initiatives can insert new innovations into the mix. The efforts of these companies can produce new categories of products, address an under-served niche, or posit some new approach to library services. Organizational trajectories can also vary. Some may fizzle—such are the risks of new start-ups. But others will find success with their products and business models and may grow over time as independent companies or may be tapped for acquisition by an established player.

A variety of the companies and products can be seen as emerging start-ups rather than developing from established industry players. Each has made interesting contributions to the arena of library technologies.

SFX, the original link resolver, emerged as an entrepreneurial initiative in Ghent University in Belgium in the late 1990s. Being able to create reliable links to articles in a scalable manner was a critical problem in the emerging realm of electronic journals and this tool addressed this need. The quantity of articles available had quickly made hardcoding manual links unsustainable. The work of Herbert Van de Sompel led to the development of the SFX context-sensitive link resolver and the formation of the OpenURL standard. Ex Libris acquired the rights to SFX in 2000 and commercialized it to become one of its key offerings, which also launched a new competition in this newly-established genre.

BiblioCommons was founded as a start-up in 2007 by Beth Jefferson based on her research in the area of teen literacy, including the Perfink Project in collaboration with the Toronto Public Library. The company experienced a relatively slow start. Following an initial deployment in some pilot libraries, it worked to strengthen its technical platform. Since about 2009, the company has attracted some of the largest and most prestigious public libraries in the United States, Canada, and beyond to its discovery service, which embraces concepts from the social networking sphere. The company has remained independent and has grown to be one of the major providers of patron-facing interfaces for public libraries.

Biblionix was formed as a very small family business in 2006 to support and market a new web-based automation system for small and mid-sized public libraries. The development of its Apollo Integrated Library System (ILS) was originally created beginning in 2006 for the Westbank Community Library by family member Xan Charbonnet. In the decade since, the company has remained independent and has amassed a customer base of almost 650 libraries. Biblionix remains a small company,

but leverages a modern multi-tenant platform to efficiently support its ever growing customer base.

SIPX (Stanford Intellectual Property Exchange) began in 2012 as a start-up out of Stanford University led by Franny Lee as a technology initiative to help students reduce the expenses for course materials by leveraging materials already owned by the library and through more sophisticated tools for copyright management and licensing. Following its launch as a company in 2013, the company found a strong reception to its product, leading to its acquisition by ProQuest in 2015. The SIPX technology is now under the jurisdiction of Ex Libris, which has integrated SIPX with its Leganto product for managing course reading lists. They are also continuing to develop and support it as an independent product.

ByWater Solutions began in 2009 as a small consulting firm providing support services for the open source Koha ILS. The company has found a warm reception to its services and is now established as the leading Koha support company in the United States with almost 700 library customers spanning over 1,000 individual branches.

TIND Technologies was recently established in 2013 as a start-up out of the CERN high-energy physics research facility in Switzerland. TIND provides support services for the open source Invenio platform originally created for the CERN library, which has since been adopted by a relatively small number of libraries with similar configurations in support of research organizations. TIND has attracted some attention in the United States following its selection by Caltech libraries. The company remains small and in its early stages of business development.

Koios is a very recent start-up focusing on helping libraries market themselves through better visibility on the web. The company originally developed a browser plug-in that would layer in results from a person's local library as they performed searches on web platforms, such as Amazon or Google. Koios recently launched Libre, a service able to create lists of featured items as digital displays as a means to improve a library's visibility on the web. Chattanooga Public Library has partnered with Koios as one of the first to deploy Libre.

Yewno, also a start-up out of Stanford University, has ramped up very quickly to develop and market an innovative new search technology for libraries. This issue of *Smart Libraries Newsletter* provides an in-depth look at Yewno's product, technology, and business development.

These companies and initiatives represent a sample of the relatively new organizations that have emerged within the library technology arena. I appreciate the impact and innovation they bring to libraries. In the context of a heavily consolidated industry, these new companies and initiatives expand the choices available to libraries. It will be interesting to follow the trajectories of these organizations to see how they impact the ongoing dynamics within the sphere of library technology.

Yewno Advances as a New Type of Discovery

The genre of index-based discovery services has become well established, with the majority of academic libraries adopting products such as Ex Libris Primo, Ex Libris Summon, or OCLC's WorldCat Discovery Service. These products operate through indexing vast quantities of full text, metadata, articles, book chapters, and other content items that represent the body of scholarly content of interest to libraries. They make use of indexing and search technologies, such as Apache SOLR, Elasticsearch, or proprietary equivalents, to provide a broad-based search environment. This environment is able to retrieve and rank search results based on the relevancy of full text matching, assigned subject terms, or other factors. EBSCO Discovery Service gives heavy weighting to subject terms assigned by domain experts to optimize relevancy. These discovery services make use of facets and other text-oriented tools to enable searchers to navigate through search results to identify items of interest.

A new product has recently launched called Discover from Yewno for Education, which takes quite a different approach to the discovery of scholarly and educational resources. Yewno Discover (a play on "you know") provides a new type of discovery environment, which presents a visual search experience that enables researchers and students to explore and select information resources based on concepts rather than keywords. It also allows users to discover information resources as they traverse the connections or relationships among those concepts.

Yewno has developed an intuitive visual interface that can be easily understood by all types of searchers. It relies on a very sophisticated set of technologies, which the company describes as combining computational linguistics, machine learning, and graph theory.

User Interface

Yewno Discover positions itself as offering a next-generation discovery environment based on a fundamentally different set of principles than the current slate of discovery products. Yewno Discover presents a visual interface and is based on concepts rather than terms or keywords. The interface includes a search box, which is used to enter an initial set of query terms, a visual graph in the middle pane, with a definitions box and a context bar on the right. Once the searcher enters a query, the visual graph representing the concepts contained within the documents matching the search appears, with the original terms in the center. The definitions box and context bar are populated based on an initial rendering of the search results and change dynamically based on the current context and selections.

The central concept node will be surrounded by related concepts with lines to show how they relate to others. The design is based on a highly interactive environment where the searcher can click on anything they see. The searcher can single click on a node to understand its specific network of connections, or can double click on any node to expand it and explore its underlying concepts. As the user explores and navigates through the concept graph, chains of related concepts are built, continually displaying the paths of connection. Clicking on the lines connecting nodes will present the concepts held in common.

Yewno Discover presents a visual interface and is based on concepts rather than terms or keywords.

At any point, the searcher can click on the context bar to view lists of documents relevant to the currently selected context. The list presents documents, titles, and excerpts. When an individual article is selected, the searcher is linked to its publisher's platform for viewing or downloading. Consistent with other types of discovery environments, metadata and excerpts of the documents are available within the interface, but the documents themselves are fulfilled by the original publisher. Access to documents is also contingent on library subscriptions, with other payment or request options offered for items not covered by institutional license arrangements.

It is important to keep in mind that the concepts presented through the interface may not be literally represented in the text of the documents. Yewno Discover presents concepts it has extracted from each document, not the keyword or terms that may literally appear. This approach can be especially helpful when working across disciplines that may use different terms to refer to the same concept.

Underlying Technology

Yewno has developed what it calls an inference engine that ingests documents and derives semantic data, which the search and retrieval capabilities present through its visual interface. As with conventional search products, documents are ingested into its platform. But instead of building textual indexes, as technologies such as Apache SOLR or Elasticsearch would, Yewno adopts algorithms from the realm of computational linguistics and machine learning to identify and extract concepts contained within its entire, interdisciplinary content set. Yewno projects those into a multi-dimensional semantic space and ultimately builds a semantic network of associated concepts. The platform is currently optimized for multidisciplinary scholarly content but can also be applied to other content domains or business sectors.

Content Addressed

Yewno depends on content ingested into its inference engine. The company has developed arrangements with a variety of publishers that contribute content for ingestion that can then be discovered via the Discover interface. This content is provided on the basis that the platform will facilitate discovery and link users back to publisher sites for fulfillment. Yewno is able to provide publishers with data describing the documents access via its interface and the concepts that led to document views.

Some of the major content providers currently partnering with Yewno include: Taylor & Francis, SpringerNature, JSTOR, Cambridge University Press, Oxford University Press, PNAS, Alexander Street, MIT Press, Stanford University Press, Association for Computing Machinery, Stanford University Press, Island Press, the National Academy of Sciences, Wiley, PubMed, Gale, SPIE, Annual Reviews, the Institute of Physics, the Association for Microbiology, and BioOne as well as repositories of open access content such as bioRxiv, arXiv.org, MIT's DSpace repository, and Harvard University's DASH repository. Yewno plans to continually expand its publisher partners.

Yewno recently announced that it has analyzed and ingested over 100 million texts into its platform. Although specific numbers are not disclosed, the index-based discovery services, such as Primo, Summon, EBSCO Discovery Service, are estimated to address between 1.5 and 2 billion content items. It will be interesting to track whether Yewno is able to expand its coverage to match those of the established discovery products.

Early Adopters

Although Yewno has been in development for about two years, it remains at a relatively early stage in technology development, content acquisition, and marketing. The company made its public debut of Discover at the June 2016 ALA Annual Conference in Orlando, FL.

An extensive beta trial ran from April to December 2016 across beta test sites including a variety of large universities, smaller colleges, and a national library. Libraries currently working with Yewno include Stanford University, the Bavarian State Library, New York University, The Harvard Library, Stonehill College, University of Oxford, University of California, Berkeley, and MIT Libraries. Yewno is actively marketing Discover to other educational institutions in the United States, Canada, Germany, and the United Kingdom.

Yewno recently announced that it has analyzed and ingested over 100 million texts into its platform.

Yewno Company Background

Yewno was founded in 2014 by Ruggero Gramatica, an entrepreneur and business executive with a PhD in Applied Mathematics and a background in technology across the telecommunications and biomedical sectors. The company emerged out of the vision of Gramatica and Michael Keller, the Vice Provost, University Librarian, and Director of Academic Information Resources at Stanford University, to create a concept and inference engine able to address information discovery across disciplines and in multiple domains.

Gramatica had previously been the CEO of Thermametrics (formerly mondoBIOTECH), a small biotechnology firm involved in the research problem of identifying peptides that could be used in the treatment of rare diseases through the review of scientific literature. The company's original approach was based on work performed by human experts who would work though scientific papers to identify potential associations. Gramatica led the development of a new algorithmic framework able to ingest tens of millions of citations and abstracts from PUBMED and other biomedical databases in order to algorithmically identify associations across biological entities and construct hypotheses of various mechanisms or actions. This framework platform was able to identify dramatically higher numbers of potential treatments for rare diseases than was possible through the manual work of a team of scientists. Gramatica maintained ownership to the intellectual property, which was perpetually licensed to Thermametrics.

Michael Keller, a member of the Board of Directors of Thermametrics, invited Gramatica to Silicon Valley to lead a pilot of a new Inference Engine based on Gramatica's original mathematical framework. The framework was applied this time to develop a similar platform for the broader realm of scholarly literature that Search&Learn used for the biochemical domain (as related in a presentation to the February 2017 Charleston Conference).

The first phase of the company saw the initial development of a new technology infrastructure and interface prototype, which ultimately became the Yewno discovery platform. During this time, they were also making agreements with publishers to supply content for ingestion. Following this initial two-year period of self-funding by its founders, Yewno was able to attract funding from investors to expand its operations. In November 2015, Yewno secured a $10 million

investment from Pacific Capital Group headed by Silvio Scaglia. In November 2016, the company secured a second round of investment with an additional $5 million from Desmond Shum of GO VR LLC and an additional $1.5 million from Pacific Capital, closing the Series A fund raising with a total of $16.5 million.

The company currently has around 30 employees and is based in Redwood City, CA with a presence in London. Key personnel include co-founder and Chief Strategy and Business Development Officer Ruth Pickering, Chief Data Scientist Haris Dindo, Chief Operating Officer of Education Franny Lee, and Director of Product Development Ray Shan.

Franny Lee had previously led SIPX, another start-up out of Stanford University that created a platform for helping universities reduce costs in course materials, including advanced copyright management functionality. Lee joined ProQuest as General Manager and VP of SIPX in April 2015 when SIPX was acquired by ProQuest and worked there until November 2016. SIPX is now part of the product portfolio of Ex Libris, a ProQuest Company.

The company has also developed a product for publishers and content curators, called Yewno Unearth, which provides a topical hierarchy of their current catalog. Yewno Unearth offers more detailed information to inform their future content acquisitions operations, help sales and marketing efforts, and provide more granularity to strengthen search capabilities within their own platforms.

Leveraging the versatility of the algorithmic framework, which in the past was developed and successfully applied by Gramatica in the analysis of economic and financial cycles, Yewno is developing implementations of its technology optimized for other business sectors beyond the current Yewno for Education product, including finance, legal, and life sciences.

For more information, see **http://yewno.com**.

Major System Selections

This section provides a sample of some of the major contracts announced for library automation systems in the last month.

- **Universidad Peruana de Ciencias Aplicadas** has become the first university in Peru to select Ex Libris **Alma**. Other recent selections of Alma include the **University of Central Lancashire** in the UK, and the **Arizona University Libraries Consortium** (Arizona State University, Northern Arizona University, and the University of Arizona). The **State Library of Victoria** in Australia selected Ex Libris **Rosetta** for digital asset management and preservation.
- The **Maitland Public Library** in Florida has selected **Library•Solution** from the Library Corporation.

- **Biblionix** reported 28 new libraries selecting its **Apollo** ILS between August 2016 and February 2017.
- **The Glasgow Colleges Library Group** (including the City of Glasgow College, Glasgow Kelvin College and Glasgow Clyde College) and **Bradford College** have selected **Koha** with support from **PTFS Europe**.
- Libraries selecting **Koha** with support from **ByWater Solutions** include the Bowling Library at **Judson College** in Alabama and the Technical Intelligence Library at the **Hershey Company**.
- **Münster University of Applied Sciences** in Germany selected OCLC **WorldShare Management Services**.

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.

"Within the library field, I often hear comparisons of ILS or discovery search with Amazon and Google capabilities, with the deficiencies of the ILS/discovery being on the losing side. It is natural this occurs since search is so ubiquitous within the online world, but my suspicion for this library search weakness is that the library search market is so small compared to the giant internet companies most associated with search. Is this a correct assumption? Are the number of software developers employed by the companies most often involved in library software a good

comparison and a reason for this? Or is there something more fundamental at the root of it all, e.g. MARC21, RDA, not enough money spent? I would say I hear the comparison with library search vs 'big search' so often among my peers that the deficiencies of the library software are just a given and is not an assumption that requires being challenged, but why is this happening?"

I think that there is room—even the necessity—for multiple approaches to the discovery of library resources. There is not, and likely never will be, a single tool or platform that can prove comprehensive and seamless access to the content and services libraries provide to their communities.

Libraries center on a specific portion of the overall content available on the web. We select, license, purchase, and otherwise vet the specific resources that are deemed to be reliable, objective, and meaningful to the members of the community or institution the library serves. Optimizing a patron's access to that component of the vast content of the general web in the context of well-funded commercial interests presents a major challenge to libraries.

Library discovery needs to take place on multiple levels, each suited to the context, interests, and sophistication of the patron. It is important to make library-selected materials easily accessible for those who start with Google, Amazon, or other commercial search environments, to offer broad-based library-specific discovery services, as well as more complex search tools design for specific disciplines.

In the realm of global web search, it is essential for libraries to maximize their presence.

In the realm of global web search, it is essential for libraries to maximize their presence. We can't deny the reality that most researchers start searching for resources through search engines Google, Google Scholar, Bing, or Baidu. Libraries therefore must have strategies to optimize the discovery and access to their resources in that context. The closed nature of library catalogs as well as obscure and dated metadata standards place barriers in the way of more organic inclusion of our content and services via the general web environment. Structures such as MARC21 were developed for the exchange of data between library systems, but do not fit well within the broader web architecture. Given this disconnect, libraries have to implement mechanisms that bridge the gap between library-specific metadata and technologies with that of the web. In the short term, libraries can implement search optimization techniques similar to those used in the commercial sphere. Examples include enhancing the way that resources are exposed on the web to embed structured metadata using schema.org, Open Graph, or other structures. Bibliographic

containers such as MARC21 and its associated AACR2 or RDA cataloging rules provided essential standardization for libraries to operate efficiently and cooperatively. But it is important to note that these standards are entirely unsupported and not understood outside our own domain. They provide very rich metadata, but in an obscure way from the viewpoint of World Wide Web Consortium (W3C) web standards. Creating cross-walks and mechanisms for expressing data in ways friendly to search engine indexing bots can greatly improve the performance of library resources in broader web search tools. In the longer term, libraries are shifting to more web-friendly models in or management of bibliographic resources. The BIBFRAME proposals to use linked data as the carrier for bibliographic information falls into this category. Moving from library-specific metadata structures to those consistent with the broader web has great potential to strengthen the relevancy of libraries on the web, but it will be an incredibly difficult and expensive transition.

I also see the genre of library-specific discovery services as important and ever improving. These products are optimized to access the content curated by libraries for their communities. They address the print, electronic, and digital resources that the library has selected, hopefully ensuring that the items discovered will be of high quality and easily accessed. While these discovery services have their limitations, they have advanced considerably since their introduction. These tools have largely superseded the traditional online catalog in academic libraries that address mostly the print resources managed within the ILS. Library patrons often did not understand the scope of online catalogs, especially the absence of the articles available through the library's subscriptions to electronic resources.

It is also important for libraries to offer tools and interfaces optimized for specific disciplines or types of content. The online catalog of the ILS, for example, can be seen in this category for those researchers that specifically want an advanced search tool for print resources. Specialists will continue to use the native interfaces to the key resources in their discipline rather than start with the broader discovery service.

Another key issue has to do with the objectivity of search and the degree to which a service respects the privacy of the searcher. The top-level Internet services follow a diametrically different business model and set of values than libraries embrace. They are designed to generate revenue through sales and advertising and to collect and exploit personal information to the fullest extent possible. Libraries, in contract aim to

provide objective content to their communities without cost and to safeguard their privacy. These ideals of objectivity and privacy may impose some restraint on the degree of personalization in the user experience offered by libraries compared to those like Google and Amazon. Hopefully libraries will continue to develop ways to enhance their services without compromising these essential values.

As the question highlights, those developing search services within the library domain will never have the resources at the level of the top-level Internet giants. But the consolidation of the library technology industry has created a set of companies with the development capacity that has never existed before. These organizations have the capacity to create more sophisticated and powerful technologies than were possible when the environment was more fragmented.

Finally, I don't see library-provided discovery services and the Internet giants as engaged in a win-or-lose competition. Those organizations have little interest in library-specific services. Given the divergent value propositions, I don't think that libraries can or should expect to rely on them entirely in the way that we deliver content and services to our communities. But to the extent that our users make use of those services, it is essential for libraries to be well represented. As libraries shift to linked data as the core model for managing and providing access to collections, we will be more natively part of the web ecosystem. That scenario will come with its own positive and negative implications, but should result in opportunities for libraries to create services more in line with the expectations set by the top level commercial entities.

Questions or suggestions for topics in future issues?

YOUR COMMENT COUNTS!

Contact Samantha Imburgia at **simburgia@ala.org**

ALA TechSource
alatechsource.org

April 2017
Smarter Libraries through Technology

Smart Libraries Newsletter

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

TO SUBSCRIBE

To reserve your subscription, contact the Customer Service Center at **800-545-2433, press 5 for assistance**, or visit **alatechsource.org**.

The 2017 subscription price is $85 in the United States and $95 internationally.

ALA Techsource purchases fund advocacy, awareness, and accreditation programs for library professionals worldwide.