

Discovery systems: Are they now the library?

John Akeroyd^{1,2}

¹Senior Consultant, CIBER Research, London, UK

²Honorary Research Fellow, University College London, London, UK

E-mail: John.akeroyd@gmail.com

Key points

- Library discovery systems have become successfully embedded in many academic and specialist libraries.
- Such systems require excellent metadata to ensure discoverability of content and increasingly publishers are charged with delivering that.
- Libraries' traditional role in organizing their collections for effective discovery is being usurped by these outsourced systems.

DISCOVERY

Discovery has become the *mot du jour* in the library, publishing, and scholarly communication worlds, and although it is used frequently and in multiple contexts, it remains ill defined. A strict dictionary definition would have it that discovery is the process of 'finding out or bringing to light that which was previously unknown; making known; an instance of this' (Trumble, 2002). In some senses, it might be argued that it is a mix of deliberate research – a targeted process designed to reveal something that is known to be there – and serendipity that is, coming across pertinent information whilst either researching for or otherwise handling information. Certainly, discovery is not just search *per se* but is the whole gamut of user actions and interactions, both digital and analogue, which might reveal something of value; it is the culmination of, or stages along, a journey. User actions could be prompted by a mix of system-driven functions, such as highlighting information sources deemed relevant to what is being worked upon or viewed on screen; recommender systems where users' needs are profiled by the system and content deemed relevant is highlighted; and communication through social media (or indeed any person to person channel) and through the formal processes of linking data, such as citations or Linked Open Data.

If discovery is poorly defined, then so too is discoverability. A lazy definition would simply be to say that it is the obverse of discovery – that is, ensuring that published information is so structured that discovery systems will ensure that it is exposed to the reader at the right opportunity. In practice, it means, at the very least, the provision of high-quality semantic metadata, whether that is implicit within the document itself or applied at some later point in the supply chain. If a resource is inadequately described, then inevitably, it will be hard to find.

LIBRARY DISCOVERY SYSTEMS

Global-scale discovery systems have become embedded throughout society and have largely centred on a few highly successful systems of which Google is by the far the most used. And these global search engines are not just deployed to answer every day questions but have become a starting point (and, sometimes, an end point) from much scholarly research. There is much published research to that effect; so, for example, Perruso (2016) reported that 70% of first-year undergraduates start with a general web search (although that declines as students go through their course), whilst Housewright (2013), in a large-scale survey in 2012, reported that over 30% of library users will start with a general search engine in preference to a library web site or other tools. Wolff, Rod, and Schonfeld (2016), in the Ithaka longitudinal series of studies, show web search engines to have been the primary route for search since 2010, and that continues today.

However, Google *et al.* have been challenged over the past 5 years (at least in the university domain) by the development and promotion of library-delivered, web-scale discovery systems, the prime examples of which are probably Primo and Summon (now merged), EBSCO EDS, and OCLC WorldCat Discovery. Such systems effectively consolidate all resource indexes (including, if required, the catalogue, institutional repositories, and any other data set deemed useful) into one search portal. Research undertaken by Spezi, Creaser, and Conyers (2015) and funded by UKSG showed that almost all UK universities have now implemented a library discovery engine, and the same would be true in the USA. Moreover, more recently, they have started to be deployed in public libraries (see, e.g. the UK Access to Research project, 2016), including major US city libraries, special libraries, and others. Such systems have been recorded as being very successful largely because they emulate the simplicity of Google in providing

a simple, single search box to an index of significant collections of information, including licensed and open access resources of a given library or consortium. Users, particularly students, have repeatedly expressed their satisfaction with these discovery systems through surveys and research, and there is evidence that they have become a significant starting point for search, challenging both the general search engines (and Google Scholar) and/or discrete bibliographic databases. Thus, Wolff reports a 'decrease in the use of specific electronic resources/databases', which 'may be reflective of library investments in discovery tools which are especially of use to faculty members in these fields due to the types of materials they use in their research (e.g. journals) and the interdisciplinary nature of social science research' (2016).

QUALITY METADATA

However, whether through a library discovery system or Google or another large-scale search system, all such systems depend on excellent quality metadata for success to achieve effective retrieval – this has been, and will continue to be, a major challenge for all parties in the supply chain; not all content providers invest in sufficient metadata. Good examples would be: library catalogue records that are only indexed at the title level (especially historic collections that may lack any subject metadata at all) and digitized images may lack any metadata at all and even some books, especially eBooks, tend to lack sufficient subject metadata to make them stand out in search systems. (The exception is probably abstract and indexing services that have always been designed with excellent discovery in mind.) There are also problems where multilingual collections or collections built from disparate sources are brought together with consequential differing levels of metadata.

But these concerns are addressable, and to some extent resolvable, with much research being undertaken as to how to semantically enhance resources either at the record level or at the search interface. EDS, for example, is now grafting onto the discovery platform-relevant taxonomies so as to compensate for poor indexing by expanding queries at the search stage. Other systems, such as Europeana, have auto-indexed their collections by ingesting vocabularies, such as DBpedia, into their metadata records and thus providing enhanced opportunities for word matching. But even these techniques have their limitations in that, even though they can overcome language barriers and provide synonyms, they cannot index at a deeper level of granularity than is already the case, nor can they easily deal with facets such as sentiment or humour. Perhaps more important is the increasing interest and application of more advanced search techniques – especially given that the keyword and key phrase searching that are implicit in most of these bibliographic systems are becoming dated. Thus, Ransom (2016) reports on new functions to be implemented in the Primo and Summon platforms, and artificial intelligence, text mining, and conceptual search are all likely to play part in the future. (One such pilot project is Yewno (2016), which is a cluster-based search engine that enables

discovery through concept visualizations rather than listings of citations derived through probabilistic calculations.) All these systems enable discovery that goes beyond just focused searching into serendipity. Thus, what we are seeing is a slow but profound shift in the way that libraries deliver content. We have moved from a position where libraries were traditional print collections augmented by digital content to hybrid libraries combining both print and digital (perhaps with different and often difficult interfaces) to a digital first premise with a single unified search system as a starting point. Users are then led into other resources and platforms, such as publisher's sites; they are at the start of a journey into the full text, with suggestions, references, and alerts as follow up.

Thus, from a publisher perspective, if a publisher wants to sell into libraries or to ensure the highest levels of discoverability or get their content adopted, then they must be providing high-quality metadata themselves or be assured that it is in place further along the supply chain, perhaps being applied by aggregators or other intermediaries. Publishers will have to ensure that they have access to the expertise to make their content work in these new environments. Indeed, it appears to be increasingly less of a library concern as digital content becomes more and more aggregated, and its indexing and delivery is less the direct concern of any given library and more that of a mix of publisher, aggregator, and system providers.

THE LIBRARY ROLE

These developments raise the question: has the library's traditional role as indexer of its collections now been usurped so that web-scale discovery systems are now, de facto, the library? It has been argued that the library catalogue has ceased to be the primary route to a library's resources (Kortekaas, 2014). Thus, the library will have much less, if any, direct responsibility for enabling discovery but will be more concerned with access and fulfilment – that is, ensuring that the maximum possible array of relevant full-text digital resources are available, suitably licensed, and easily and effectively delivered to the users. It implies that libraries must provide seamless authentication systems so as to enable ease of access and that there should be no significant barriers to access, even where license arrangements apply. Indeed, even other traditional library functions, such as inter-library lending or the delivery of print books, could be served through *click and collect systems* from within the discovery system so as make their availability as seamless as possible. Thus, the discovery and associated functions are effectively outsourced, and the discovery system itself has, in effect, become the library. And much, if not all, metadata creation and its deployment will no longer be a local library concern but produced at a national or international scale. And where there is a lack of suitable metadata, record augmentation will be achieved retrospectively or at the point of search by highly developed systems using commonly available vocabularies. Libraries, at least in respect of their traditional role as collection managers and organizers, will be left to focus on

contract management and ensuring effective access through a reliable infrastructure. This is not to say that libraries will diminish per se, but their function will shift to that of exploitation, support, and even origination rather than curation.

There are at least two caveats to this argument. First, although the discovery engine itself may well be beyond the control of libraries, which engines to use and in what context still remain decisions for the library and its users. It is unlikely that one discovery system will be capable of dealing with all usage scenarios – it is already the case that the bespoke discovery systems implicit in some bibliographic databases are better aligned to searching that data than the generalized web-scale systems. Thus, libraries will need to be able to advise on the best tools for particular requirements and to customize those tools to meet the needs of their local populations. They will need to manage the discovery landscape and be prepared to provide and support a mixture of routes into content so as to ensure both an easy journey for users and also one that is ultimately successful for users. The task is that of ensuring an effective customer journey and not just that of providing a simple start point.

A second caveat to this argument is that libraries also have an emerging role in managing and disseminating their host organization's intellectual capital, including both archival content and new outputs such as research papers, research data etc. This requires the library to act in effect as publisher for this material. According to Lorcan Dempsey, this trend 'places an emphasis on effective disclosure, thinking about search engine optimization, syndication of metadata to network hubs, or to other specialist network level resources... Libraries have to become much more interested in the discoverability of their resources, sometimes within the context of the collective library collection' (2016).

SUMMARY

The foundation of all libraries is that of organizing knowledge to ensure its effective exploitation in service to an organization mission. That changed only marginally with the emergence of digital content, where libraries, through online catalogues and databases, ensure routes into the digital content they have acquired. Library web-scale discovery systems have altered the landscape so that discovery and content metadata is now largely contracted and beyond the control of individual libraries. With increasing aggregation and integration with library management systems, such systems could be regarded as an end point. However, it is likely that there will continue to be a mix of discovery systems, some niche and some in competition, and libraries will need to ensure that the right tools are in place to maximize the user journey.

Meanwhile, publishers and content providers, whether commercial or otherwise, will need to ensure their content is maximized for all 'generic' discovery systems whilst conceivably providing their own bespoke discovery engines attuned to their own content. It is about ensuring quality semantic metadata, which, to ensure inter-operability, may well need to be either standardized or easily mapped onto more generic taxonomies so as to be effective in all domains and all languages. From a research perspective, it might well imply a resurrection, or – more correctly – a new use for library classification standards, such as Dewey or Universal Decimal Classification, both of which have been trialled as aids to retrieval in digital libraries – see, for example, Lin *et al.* (2015). As the roles of libraries and content providers in ensuring discoverability of scholarly content continue to evolve, so too will the tricks of the trade.

REFERENCES

- Access to Research. (2016). Retrieved from <http://www.accesstoresearch.org.uk/>
- Dempsey, L. (2016). Library collections in the life of the user: Two directions. *LIBER Quarterly*, 26(4). doi:10.18352/lq.10170
- Housewright, R., Schonfeld, R. C., & Wulfsun, K. (2013). *RLUK UK Survey of Academics 2012*. Retrieved from http://www.rluk.ac.uk/wp-content/uploads/2014/02/UK_Survey_of_Academics_2012_FINAL.pdf
- Kortekaas, S. (2014, June). *Thinking the unthinkable—Doing away with the library catalogue*. Webinar presented at the UKSG Webinar. Retrieved from <http://www.uksg.org/librariycatalogue>
- Lin, X., Khoo, M., Ahn, J.-W., Binding, C., Tudhope, D., Jones, H., & Massam, D. (2015). A DDC visual interface for metadata exploration. *Proceedings of the International Conference on Dublin Core and Metadata Applications*, 158–169.
- Perruso, C. (2016). Undergraduates' use of Google vs. Library Resources: A four-year cohort study. *College & Research Libraries*, 77(5), 614–630. doi:10.5860/crl.77.5.614
- Ransom, J. (2016, September 22). Ex libris initiatives blog [Web log post]. Retrieved from <http://initiatives.exlibrisgroup.com/2016/09/library-exploration-serendipitous-discovery.html>
- Spezi, V., Creaser, C., & Conyers, A. (2015). The impact of Resource Discovery Services (RDS) on usage of electronic content in UK academic libraries: Selected results from a UKSG-funded project. *Serials Review*, 41(2), 85–99. doi:10.1080/00987913.2015.1035991
- Trumble, W. R. (Ed.). (2002). *Shorter Oxford dictionary* (5th ed.). Oxford, England: Oxford University Press.
- Wolff, C. W., Rod, A. B., & Schonfeld, R. C. (2016, April 4). *Ithaka S+R US Faculty Survey*. Retrieved from <http://digitalcommons.unl.edu/cgi/viewcontent.cgi?article=1016&context=scholcom>
- Yewno. (2016). Retrieved from <https://yewno.com/>