IS THE JUICE WORTH THE SQUEEZE?

Bruce Heterick, Vice-President, JSTOR
Andrew Wells, University Librarian, UNSW Australia

Abstract

As the implementations of web-scale discovery services (WSDS) like Primo, Summon, EDS, and WorldCat Discovery Services proliferate, and as libraries continue to make significant investments to purchase, implement, and maintain these services, it is increasingly important to understand if these investments are helping libraries achieve the impact they originally anticipated. What was the original rationale for these services? What were the impacts that were expected? How are those impacts being measured? What are the early results? Are the results supporting the investments? Are there other options?

At the same time, many content providers are making similar investments to make their content discoverable in these services and to support libraries in leveraging these investments on behalf of their faculty, students, and researchers. The promise, of course, is greater discoverability and increased usage for the content provider. What, exactly, are the investments that these content providers are making? Given that there is a culture of providing content for indexing for free to these services, how are these investments being recovered by the content provider? Are these investments resulting in the expected outcomes?

Introduction – Development of Discovery Services and Providers

Discovery services have been a significant area of focus and expense. In the early 2000s, initial players were library system vendors, such as Ex Libris with MetaLib/SFX, Innovative's Encore and Voyager's Encompass. Some libraries developed local discovery systems. Library system vendors have continued to develop these services, and launch new products such as Primo from Ex Libris. Serials Solutions launched Summon, subsequently acquired by ProQuest. Other types of players have entered the market in recent years, including EBSCO Publishing introducing the EBSCO Discovery Service (EDS). OCLC is another major provider with the WorldCat Discovery Service. Some libraries have also developed their own web-scale discovery services. In Australia, a notable example is the National Library of Australia's Trove.

The early versions of these services used federated searching (also known as metasearching) which sent a user's search query to multiple target databases. The search results were returned and displayed to the user: this display could be configured and grouped according to the service's functionalities. There were inherent performance restrictions in federated searching as search queries were sent to a limited number of target databases in real time. "The slow performance, limited number of content targets that could be included in a query, and the limited number of results returned were factors that impacted the success and satisfaction with metasearch."[1] The Z39.50 standard was an important component of these services.

In 2009, discovery service providers started to release index-based discovery services which featured a different model aimed at increasing their scope and performance. The model exploits a central index populated with metadata and/or full text for a wide range of content that a library supplies to its users. Content providers play an enhanced role in this model, as their metadata must be supplied to the central index. A user search is performed in the central index, not the native interface of the content provider. Link resolvers such as SFX navigate the user to content. This short description underestimates the technical complexities of these new services. In comparison to federated searching, the key point is the content provider is playing a new role through the supply of metadata to a central index at the cost of losing direct access to their own platform. Librarians see this as a benefit to content providers: “the purpose of these discovery services is not to re-publish material represented in the index, but to provide an additional channel for connecting library users with that content.”[2] For content providers, it was a new ‘squeeze’.

The take-up of discovery services represents significant investments by libraries and discovery service providers. This growth has affected content providers because they cannot ignore these developments. Library users access their content via web-scale discovery services. The increased functionality of these services has depended on content providers making metadata and application program interfaces (APIs) available. The coalition of companies providing the services and their customer libraries has been a very powerful and influential one. At the same time, content providers invest resources in their own web-sites and platforms.

This paper provides viewpoints from both a university library and a content provider. Why did libraries and system providers develop these complex discovery layers? Do they represent a return on investment and an improved user experience? From the content provider’s point of view, what is their return of investment through cooperating with discovery service providers?

The Library Rationale

The concept for these services emerged in the mid-1990s. This was a time of very rapid growth in the World Wide Web and online publishing. Journals began their transition from print to online. Librarians were confronted with an enormous amount of easily and freely accessible online content, instead of the bounded and highly controllable print and physical collections in their buildings. There were courageous attempts to help users find content – these were the days of subject gateways and portals. The Dublin Core metadata standard delivered a simple and flexible method of resource description for this complex and rapidly growing information environment. Librarians realised universal bibliographic control was a goal that belonged to a former time.

The real estate of libraries’ web-sites became another tool to navigate users around this new world. Kortekass captures this practice from 2002 at the Utrecht University Library where a locally developed discovery tool named Omega had been launched: "Looking for printed material? Search the WebOPAC. Looking for electronic journals or e-books? Search Omega. Looking for specific disciplines or materials? Search the dedicated electronic databases.”[3] There was no way to present these resources in an integrated way, so giving clues to the user about where to go to find resources was the next best thing to do. However, search engines were giving people a simple, one-step approach to finding resources at a global scale. Searchers did not seem to mind retrieving thousands of results either. The success of services such as Google was noticed by librarians, who saw that despite lack of precision and relevance in the search results, users loved them.

University libraries started to realise that they were at the beginning of a major evolution from print to online, accompanied by emphasis on access to information anywhere and anytime replacing the longstanding ownership of physical resources model. This evolution has continued at a gradual pace, but the direction is clear: a majority of libraries are moving to an access model. For research libraries, journals are the dominant information resource. Increasingly, libraries will be prepared to license this content without perpetual access provisions “because they can provide access to far more content at a far cheaper price than perpetual access allows”.[4] The volume of available information, both paid and free, has grown rapidly. Libraries also want to make their special and legacy print collections visible too. Some libraries began to see that their value might lie elsewhere: because libraries no long had a monopoly on the provision of access to information, value could be added to the content they provide by improving discovery and delivery services to grow usage of library-provided resources. And if it could be a Google-like experience, even better.

For UNSW Library, an opportunity to tender for a new system in 2001 asked vendors to respond to these issues and concerns. In the UNSW library system tender written in 2001, specifications for an "Information Access System" were summarised as "the key tool for provision of seamless and integrated access to print and digital resources. The system will enable the Library to move from a traditional collecting to an electronic access paradigm. It will help Library staff manage this shift in service and provide information on usage of networked information resources. It will assist users to find information in networked information resources." This led to the acquisition of MetaLib and SFX from Ex Libris. In 2008, the Library became a Development Partner for Primo. In 2011, the OPAC was decommissioned so discovery of library content, whether owned or licensed, was via Primo only. UNSW Library was a true believer in web-scale discovery services.

UNSW Library Effort (The Squeeze)

The effort and resources required to implement, maintain and enhance Primo are significant. Being a development partner required extra commitment. Several groups in the Library are responsible for the application support and regression testing required for the 2-3 releases expected each year. Primo is locally hosted, so there are infrastructure and performance issues to be regularly monitored. It is one of UNSW Australia’s enterprise systems.

A high level Primo Enhancement Group meets regularly to make decisions about the configuration of Primo. There is constant work on facets, filters, relevance and ranking to obtain the best fit between a user’s search and the results. Web-scale discovery services cannot merely be ‘turned on’ or ‘used out of the box’. They require constant experimentation as new information resources are added and new functionality is delivered. University libraries pay high prices for scholarly content, so it is critical that the discovery service leads users to high quality content. UNSW Library estimates that the web-scale discovery service consumes 5% of its human resources and 10% of its information technology budgets.

How does UNSW Library justify this investment?

UNSW Library User Experience (The Juice)

A good place to start is to look at the number of searches each year. This has grown steadily as shown in Figure 1.

Figure 1: Annual Primo Search Statistics at UNSW (2010-2014)

Annual Primo Search Statistics		
Year	**Searches**	**% Increase**
2010 (from 5 Jun)	1,796,821	
2011	4,016,620	123.5%
2012	5,169,570	28.7%
2013	5,944,727	15.0%
2014 *	7,128,732	19.9%

The significant increase from 2010 to 2011 arises from the decommissioning of the ALEPH OPAC in 2011. This is a lot of activity, but it does not reveal much more. There are more questions for which we seek answers: Are people finding what they are looking for? Where do the searches start from? Are there more searches because it is harder to find something? Which campus groups are using it – are there differences between students and academics, or disciplinary differences? In 2014, a survey of faculty staff provided some information relevant to these questions.

Ithaka S&R local faculty surveys have provided another way to consider return on investment. Five of the Australian Group of 8 libraries completed local faculty surveys using an instrument designed by Ithaka S&R in New York. The surveys gather responses from academic staff only: undergraduate and postgraduate students are not surveyed. Even so, the results are informative.

In the area of discovery, the survey asks three key questions about where they commence searches for particular types of information needs.

Here are the UNSW results.

1. For academics wanting to locate information on a research topic, 17% turn to the Library, compared to 55% using search engines and 27% using a specific web-based scholarly resource, either provided by a commercial or non-commercial content provider. The Library comes third.
2. For academics wanting to locate known secondary sources, the results are different. The Library narrowly comes first at 38%, followed by 32% on general purpose or scholarly search engines, then by 25% on a specific scholarly database.
3. Searching for new journal articles or monographs in their area of research interests, has the Library behind again on 17%, with 46% using a specific scholarly database, then 31% on search engines. The Library comes third.[5]

The summary results for the five G08 libraries have slightly different percentages, but the relative positions are the same.[6] The same relative position was found in the Ithaka 2012 survey of faculty in the United States. Academics turn to the Library to retrieve known items. For research purposes, whether locating information or finding newly published items in the area of interest, the Library is consistently third and well behind. While this summary does not give justice to the rich findings of the survey, such as disciplinary differences and trends over time, there is a consistent message. When it comes to choosing a service for research purposes, searchers are more likely to use a search engine or a specific online resource from a content provider. So most of our 'squeeze' at UNSW Library is to produce the 'juice' of known item searching. At the moment, UNSW Library management judges that the return on investment as satisfactory, but it raises a question about how much more investment is needed or would indeed be worthwhile to make Primo a satisfactory research tool for users. Even if we did put in more effort, would they really abandon Google, Pub Med or whatever their favourite place is?

At UNSW Library, we recognise that our users are using native interfaces of commercial and non-commercial content providers, as well as the major search engines. They have a legitimate need to do so. Maybe users don't want as much juice as we think they do ... so what does a content provider make of all this, and what do they see?

Content Providers and Libraries

It is a common observation that libraries and their suppliers are living through disruption. Every year, there is evidence of this disruption, such as the continuing demise of subscription agents. Web-scale discovery systems are an interesting product of this disruption and it is useful to see their development from the perspective of a content provider.

Libraries placed new demands on content providers as the online age dawned. Preservation and perpetual access provisions were negotiated in licences, as libraries and their users were fearful that access might be lost in the future. Libraries influenced content providers to participate in long-term preservation services such as Portico and CLOCKSS, or established national or international schemes. They also encouraged content providers to participate in web-scale discovery systems (WSDS) by supplying their article level metadata and/or full-text to these services for indexing purposes. Libraries believed this would improve the user experience by providing a one-stop for searching heterogeneous information resources, hopefully increasing usage of licensed information resources, thereby justifying the significant investment in them. While encouraging this, libraries had no visibility of whatever commercial arrangements between the content providers and discovery system had to be made. It is fair to assume that costs incurred here are recovered from the purchasers of these services, that is, libraries.

In parallel, content providers have continued to make significant investments in their own web-sites, platforms and discovery systems. Some of them have significant brand value to library constituents – e.g. JSTOR, Web of Science, ScienceDirect – and are important starting places for research in certain disciplines. While those platform investments are sometimes seen as a necessary evil by libraries, there is no doubt that they are welcomed and supported by the constituents that the library serves.

These developments have provided constant source of conflict between content providers and libraries. Libraries place more value (and effort) into their own web-scale discovery systems and sometimes actively discourage content providers making their own investments. The ICOLC statement in response to the Global Financial Crisis advised content providers to decrease investment in their own platforms, or at least not ask libraries to pay for it.[7]

Competition between web-scale discovery systems has also produced conflict, notably the very public EBSCO decision not to have its own content indexed in Primo.[8] The very fact that the WSDS business is competitive is a good thing for libraries. Competition typically translates into lower prices and a more robust set of services. However, because the market dynamics are somewhat skewed (e.g. libraries pay for WSDS services, but WSDS does not typically pay the publisher/content provider a royalty for the right to index that content that makes the service valuable) and convoluted (e.g. some WSDS providers are also content providers themselves), there are inevitable conflicts amongst the collaborators. The Open Discovery Initiative (ODI) from the U.S.-based National Information Standards Organization (NISO) has recognized these potential conflicts and has attempted to bring the various stakeholders together to adopt a code of practice[9], but these efforts are still nascent and their potential impact has yet to be seen.

The Content Provider Experience

Research shows that WSDS can affect usage for the content provider – both positively and negatively.[10] The usage impacts can be tied to any number of variables: type of content (full-text vs. abstract), brand recognition amongst users, quality of metadata (subject metadata, primarily) provided to the WSDS for indexing, relevancy ranking algorithm of the WSDS, on-going maintenance and prioritization (updating administrative settings) of WSDS by library staff, etc. In some cases, these variables are within the control of the content provider (e.g. quality of metadata); sometimes they are in the control of the WSDS provider (e.g. relevancy algorithm); but often they are in the control of the library (e.g. on-going investment in staff/resources to stay current).

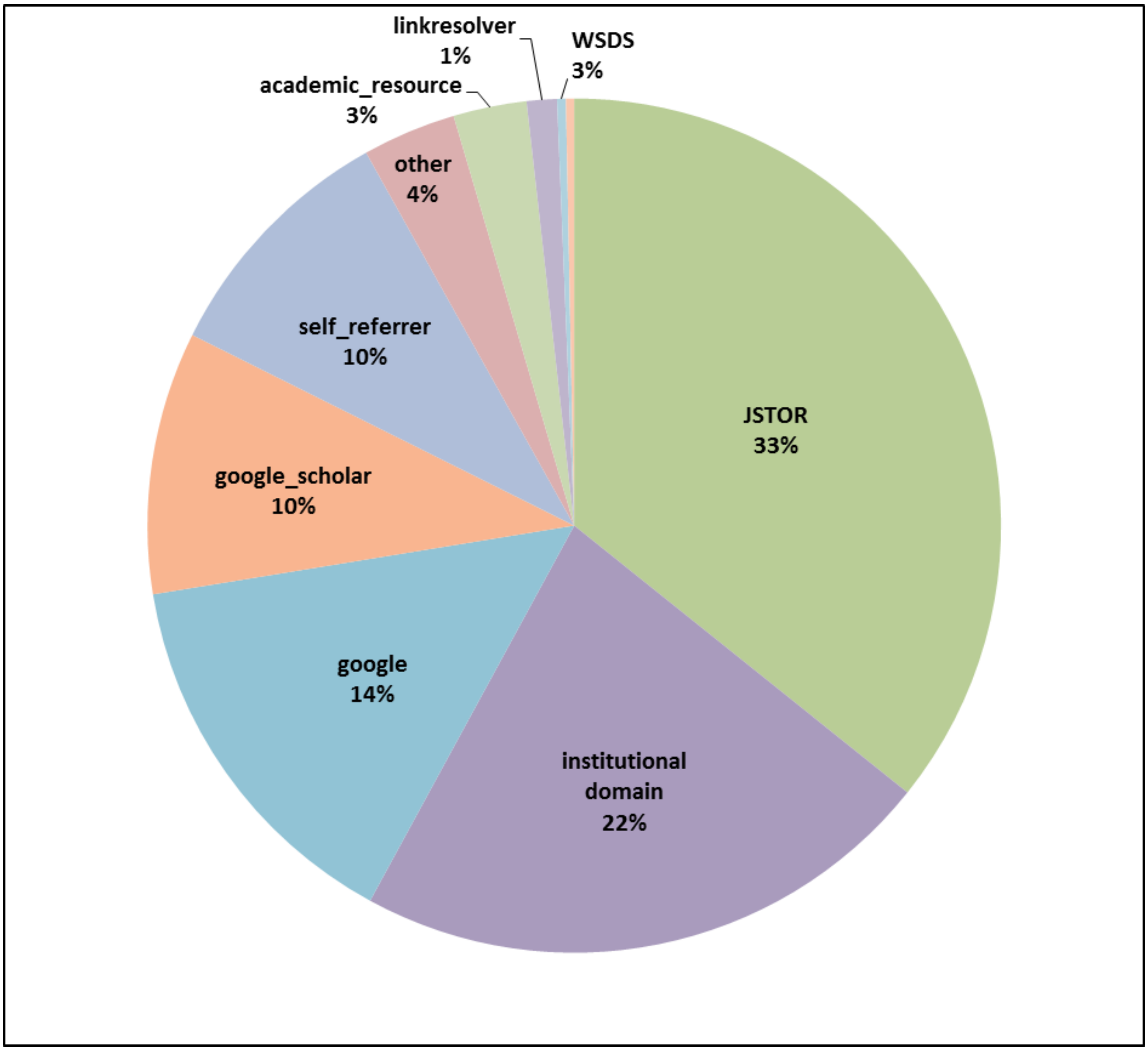
For most content providers that depend upon libraries as the intermediary to get their content to the end-user, usage is the "holy grail". A low cost-per-use (CPU) can keep a journal/aggregation in the budget, while a high CPU can put that journal/aggregation on the list of resources to be evaluated for reduction/elimination. So, understanding the impact of WSDS on usage is of considerable importance to most content providers.

A relatively high percentage of content providers support the indexing of their content in WSDS by providing their content (metadata and/or full-text) *without charge* to these services. This is done with the expectation that WSDS will drive more traffic to the content provider and improve the usage of that content from the institution. In some cases, usage does not increase. In some cases, for some content providers, usage can drop quite significantly. Is there a direct cause-effect between the implementation of these WSDS at some institutions and the increase/decrease in usage for some content providers? If there is a cause-effect, what variables are impacting those usage changes?

At JSTOR, we have spent considerable time and resources to understand these impacts. These investments are on-going; to the tune of over 500,000 USD annually, and if one were to look at the aggregate investments of the content providers to participate in these services, build infrastructure to support these services, and add staff to support these services, the total investment of the scholarly communications ecosystem would probably be surprisingly high. How are these investments being funded? Price increases? If libraries are paying annual maintenance costs for these WSDS, as well as price increases for the content to support the content provider's support of WSDS, how is the library measuring those increased costs versus the return it is getting on those new investments? How is the content provider measuring the value of those ongoing investments to meet the library expectations with regard to web-scale discovery?

Like many other content providers, JSTOR is actively tracking and evaluating usage patterns. Figure 2 illustrates the origin of content accesses (article views + PDF downloads) on the JSTOR platform in calendar year 2014:

Figure 2: Origin of Content Accesses on the JSTOR platform (2014)

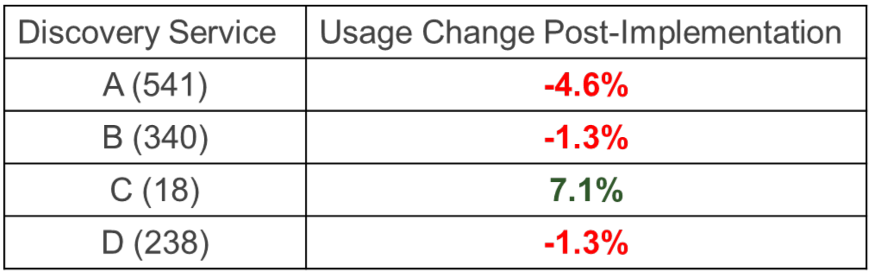


As the chart illustrates, a large percentage of content accesses on the JSTOR platform actually begin at www.jstor.org (~33%). There are also a good percentage of content accesses coming from Google and Google Scholar (~24%), and resources that have an institutional domain (~22%). Very little of the usage on the JSTOR platform is attributable to WSDS. Part of this low attribution is related to the way that usage from a WSDS is brought to JSTOR (e.g. via a link resolver), which obfuscates the actual origin of that content access. A small part of WSDS-based usage on the JSTOR platform is also embedded in the 'institutional domain' category (for those institutions that may have given their WSDS an .edu origin and do direct linking to JSTOR). However, at this point, Figure 2 illustrates clearly that only a small percentage of usage on the JSTOR platform – most likely less than 10% - is attributable to WSDS.

As a not-for-profit organization with a keen eye toward managing our cost base as effectively as possible, the data would suggest that the resources that do exist for discovery efforts at JSTOR would be most wisely spent on first improving the JSTOR interface to make it 'stickier' and drive more content accesses; next, the focus should be on the consistency and quality of indexing of the JSTOR platform in Google Web Search, as well as improving the authentication process for participating institutions, in an effort to better convert the tens of millions of Google users that are turned away from JSTOR access each year; efforts on improving the discoverability of the content on the JSTOR platform in Google Scholar would be next on the list. Efforts in the area of WSDS would not necessarily be a high priority for JSTOR by simply looking at the origin of content accesses on the JSTOR platform.

At the same time, JSTOR recognizes that libraries have made significant investments in WSDS, and the organization would like to help libraries leverage these investments as best they can. If collaboration with the various WSDS in this vein were improving the cost per use (CPU) of the JSTOR platform for participating institutions, this would be a win-win situation and there would be little discussion as to the value of the approach. However, the early analysis conducted on the impacts of usage on the JSTOR platform for institutions that have implemented WSDS has indicated mixed results (see Figure 3):[11]

Figure 2: Usage change at JSTOR institutions following WSDS implementation

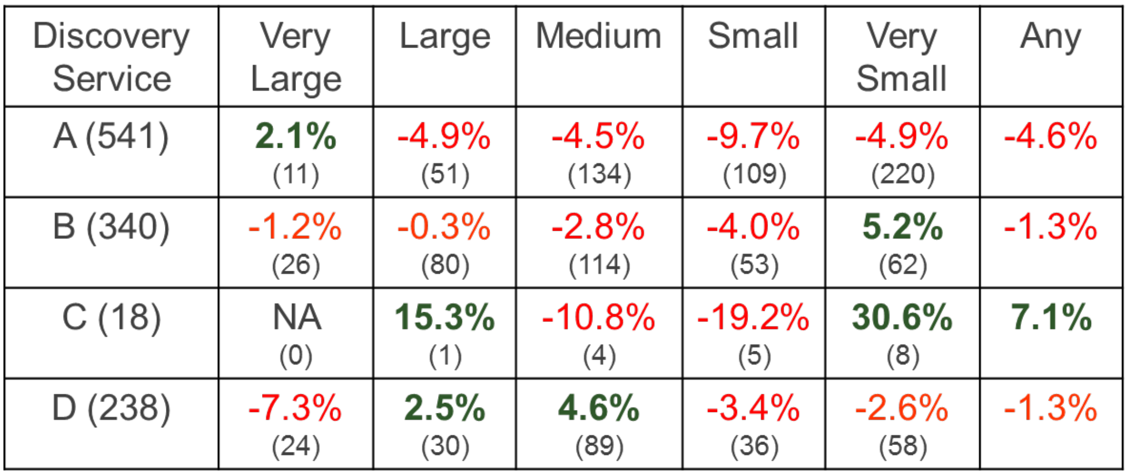


Discovery Service	Usage Change Post-Implementation
A (541)	**-4.6%**
B (340)	**-1.3%**
C (18)	**7.1%**
D (238)	**-1.3%**

For the purposes of this analysis, JSTOR worked with each of the major WSDS providers (EBSCO-EDS, Ex Libris-Primo, OCLC-WorldCat Discovery Service, ProQuest-Summon) to take the world-wide customer lists of each, along with the implementation date of the WSDS, and analyse the JSTOR usage at those institutions for the twelve (12) months prior to implementation and the twelve (12) months post-implementation. We found that the overall JSTOR usage for all worldwide participating higher education institutions had decreased -0.7% during the time period studied (August 2009 – September 2013). With the exception of Discovery Service 'C' – which is statistically challenged because of the low sample size – institutions that had implemented WSDS saw their usage on the JSTOR platform decrease more precipitously than the worldwide higher education average (-0.7%).

The data is even more interesting – and complicated to interpret – when institutions are broken out by JSTOR Class (Very Large, Large, Medium, Small, and Very Small):

Figure 4: Usage change at JSTOR institutions following WSDS implementation by JSTOR Class



Discovery Service	Very Large	Large	Medium	Small	Very Small	Any
A (541)	**2.1%** (11)	-4.9% (51)	-4.5% (134)	-9.7% (109)	-4.9% (220)	-4.6%
B (340)	-1.2% (26)	-0.3% (80)	-2.8% (114)	-4.0% (53)	**5.2%** (62)	-1.3%
C (18)	NA (0)	**15.3%** (1)	-10.8% (4)	-19.2% (5)	**30.6%** (8)	**7.1%**
D (238)	-7.3% (24)	**2.5%** (30)	**4.6%** (89)	-3.4% (36)	-2.6% (58)	-1.3%

As Figure 4 highlights, changes in usage after implementation of WSDS can vary widely by institution type (research vs. undergraduate, private vs. public, large FTE vs. small FTE, and especially country vs. country)

At JSTOR, this data was used not to make the case that WSDS are bad or unimportant, or to establish some cause/effect between the implementation of WSDS and usage on the JSTOR platform, but rather to illustrate the fact that the philosophical arguments as to the advantages of WSDS to *all* content providers can, in some cases, be disputed by the data. And, therefore, it is not a foregone conclusion that *all* content providers would find it in their best interests to participate in WSDS without truly assessing the ongoing costs that would need to be incurred to make that participation worthwhile. This data – along with nearly a year of effort to understand better how each of these WSDS worked – made it clear that JSTOR's ongoing investment into these collaborations would need to increase significantly if JSTOR – and its participating libraries - were going to see the advantages of these collaborations in increased usage of the content on the JSTOR platform.

There is an implicit assumption – in nearly all discussions of the value and impacts of WSDS implementations – that the library, content provider, and WSDS provider have goals that are aligned (or, at least, should be). It is not clear at all that this assumption is valid. For instance, there are few libraries that can detail success criteria for implementing a WSDS; and, fewer still that have identified any best practices to measure what "success" looks like. Without those, how can a WSDS or content provider really align with a library to maximize the desired impact of these collaborations? In some cases, the WSDS providers are also content providers themselves and compete with the very resources they are indexing to make discoverable in their WSDS. At the very least, this leads to suspicion from some content providers with regard to motives, relevancy ranking algorithms, and favouritism. These are but a small sample of the possible frictions that exist in this (un)holy trinity.

In the end, most content providers are simply interested in getting the content on their platform in front of as many students, faculty, and researchers as possible. If WSDS can assist in that effort – and in most cases it can – then that is enough. However, the JSTOR experience has shown that if the content provider truly wants to maximize its partnership with the WSDS, and by extension, provide a more robust experience for its library customers, there is a level of investment and ongoing financial commitment required by the content provider that is not insignificant. If the juice is worth the squeeze for libraries, and they are truly interested in making that discovery experience as robust as it can be, libraries are going to need to be prepared for content providers to expect to recover those investments, as well as increase their own level of investment within the library to actively manage this new environment.

References

[1] National Information Standards Organization (2014) Open Discovery Initiative: Promoting Transparency in Discovery: 10

[2] National Information Standards Organization (2014) 11

[3] Kortekaas S (2012) Thinking the Unthinkable: A Library Without a Catalogue (http://libereurope.eu/news/thinking-the-unthinkable-a-library-without-a-catalogue-reconsidering-the-future-of-discovery-tools-for-utrecht-university-library/. Accessed 21 November 2014

[4] Levine-Clark, M (2014) Access to Everything: Building the Future Academic Library Collection. Portal: Libraries and the Academy 14(3): 426

[5] Ithaka S&R (2014) UNSW Survey of Academics (unpublished)

[6] Ithaka S&R (2014) Group of Eight Survey of Academics 2013-2014 (unpublished)

[7] International Coalition of Library Consortia (2009) Statement on the Global Economic Crisis and Its Impact on Consortial Licenses (http://www.library.yale.edu.au/consortia/icolc-econcrisis-0109.htm). Accessed 21 November 2014

[8] Grant, C (2013). Do-They-or-Don't They: Ex-Libris and Ebsco Information Services (http://thoughts.care-affiliates.com/2013/06/do-they-or-dont-they-ex-libris-ebsco_13.html). Accessed 21 January 2015

[9] National Information Standards Organization (2014)

[10] Levine-Clark, M., McDonald, J., and Price, J.S.(2014) The Effect of Discovery Systems on Online Journal Usage: A Longitudinal Study, Insights 27(3):249

[11] Heterick, B (2013). Revisiting Plato's Cave: 2013 Charleston Conference Plenary Presentation (videorecording), available at http://you.tube.com/watch?v=ZkGSQIF0BI