Selecting an Appropriate Web-Scale Discovery Service: A Study of the Big 4’s

Vinit Kumar
Babasaheb Bhimrao Ambedkar University, Lucknow - 226 025 India
E-mail: mailvinitkumar@gmail.com

ABSTRACT

Seeing the dynamic user demands the libraries are changing their approach to the user services. The web-scale discovery (WSD) service is the latest attempt in this direction. There are several players in the marketplace providing solutions for WSD with products having basic features and subtle features as well. As more players are entering the marketplace, it becomes challenging to select appropriate WSD system. It is also paramount for the library managers to be aware of the myriad range of features and underlying technology of WSD. This understanding will be for making informed purchase decisions. This paper attempts to explain in detail the components of a typical WSD system. Further, the paper evaluates the features of the Big 4’s in WSD. The paper concludes by discussing some of the parameters to consider while evaluating the WSD system.

Keywords: WSD; Discovery services; Primo central; WorldCat discovery services; Summon; WS-Discovery

1. INTRODUCTION

Libraries boast about the increasing number of subscriptions to journals, electronic content and databases whereas at the same time funding agencies always complain about the minimal full-text downloads of subscribed resources by the library patrons. The probable reasons could be the limited information literacy of users, irrelevant content, and disablments of the access system. The user’s needs are very dynamic and keep on changing rapidly and dramatically. If we see the OCLC study done in 2005 asking the respondents the question “Where do you typically begin your search for information on a particular topic?” 87 per cent of respondents answered, “Search Engines”. The above finding signifies that just increasing the number of subscriptions would not be of much help to the users, but there is a need for such system matching the expectations of the users for maximum utilisation of the resources could be made. This system should operate at the same level of sophistication as other popular Web destinations are operating. As warned by Coyle & Hillmann’, “If the libraries are to avoid further marginalisation, they need to make a fundamental change in their approach to user services”.

Web scale discovery (WSD) service is the latest attempt in this direction. WSD service serves as the “front door” for the researchers or “a single point of entry” for the library catalogue, digital repository, journals subscriptions, and subscribed online databases. In other words, a unified interface to all the internal and external resources of the library.

With the increased availability of content in the electronic format and its associated benefits, libraries started subscribing electronic content. Despite several advantages of electronic resources, they are too fragmented and prompt several issues concerning their management and delivery. Several attempts are ongoing for providing easy and intuitive access. One of the significant efforts is the implementation of a federated search of all the resources subscribed by the library. Federated search engines are a kind of meta-search engine wherein the query received from the user is sent simultaneously to multiple publisher websites and displays the received responses to the users as a single set. Although at the time of preliminary implementation of federated search in libraries, librarians regarded it as a “win-win-situation” but later with the passage of time several issues surfaced on the capabilities of federated searching. As in federate searching the query received from the user is sent to several resources simultaneously, to save the time the system puts a limit on the number of resources to be searched. The more the number of resources are searched the more will be the waiting time for the display of results. Another issue is the speed of searching; as the searching speed depends on the response of the slowest resource if any of the resources delays the response, the federated search also delays the display of results because it waits for the response from all the resources. There are other issues related to the display of results too, such as improper de-duplication, poor ranking and merging of obtained responses from multiple databases due to a variety of underlying architecture.

These issues led to increasing distrust about the service to the end users and resulted in the development of a new kind of systems, WSD systems wherein the metadata is pre-harvested from multiple resources (both local and external) and stored in a central index. Once the metadata is available in a centralised
location, better ranking methods, proper merging, efficient de-duplication and a customised display is possible.

1.1 WSD Components

In general, a typical WSD system involves following components like central index, a link resolver, relevance algorithms, and user interface.

1.1.1 Central Index

The central index is the base on which the success of a discovery system relies. Different service providers follow varied approaches for building a central index. The process starts with harvesting metadata from publishers and content providers. The harvesting of metadata involves the following methods:

- Scrapping publisher website,
- OAI-PMH compliant services.
- A single update file (in XML format) directly uploaded by the content provider through FTP to a specified location on discovery system.

The decision about the method depends on the content provider’s underlying system architecture. A content provider applies either of the above methods or at times all of the above methods.

The collected metadata provided by content providers is raw data, and the central index cannot directly index raw data. The raw obtained data is normalised keeping in view the fields available in the index and de-duplication is done to remove the duplicate entries for the same item. Seeing the variety of content providers this process involves the application of complex computing algorithms. Metadata once harvested needs to be regularly updated with any new content added to the repository of content providers. The service provider based on the availability of new content usually sets the frequency of adding new records. For automatic updating of the central index, service providers write automated transfer routines, Load tables and indexing algorithms. The central index is maintained and usually hosted in a cloud environment by the service provider. The hosted environment helps them in the seamless update, maintenance, troubleshooting and feature addition.

1.1.2 Link Resolver

Metadata obtained from the content providers have a field storing the direct link to the full text of the document. The proper resolution of this link is paramount so that when the end user tries to access the full text, it becomes available. The full text involves text, figures, tables, and graphs. The service provider for this task usually deploys an open-URL compliant link resolver software. As the failure rate about the ability of link resolvers to generate a link to the full text is moderately high, the major players have developed specialised programs to increase the effectiveness. For example, EBSCO Discovery Service (EDS) uses smart links and custom links in addition to the open source link resolver.

1.1.3 Relevancy Algorithms

Before the display of results obtained from the central index, the ranking of results according to the relevance of search terms is made. Relevancy ranking involves defining the priorities among the fields of metadata. The service provider usually sets emphasis on criteria such as currency, availability of full-text, keyword appearance in the title, matches on the subject headings, author-supplied keywords, keywords within abstracts, match on the full text of the documents. Some service providers also allow the individual libraries to influence the algorithm as per their local needs. The popularity of the service highly depends on the quality of the relevancy algorithms.

1.1.4 User Interface

Most of the WSD service providers give high importance to the user interface (UI) as this part directly interacts with the end user. Includes following interface

- Single search box and advanced search options
- Faceted navigation to filter or drill down the results as per facets
- The inclusion of enriched content from services like Google Books, Syndetic Solutions
- User profiles
- Carts to easily mark items to save for later use
- Availability of search history
- Export of records to reference managers or an email address
- Sharing on social networking websites
- Spell checker/ “Did you mean?”
- RSS feeds for search alerts and email alerts
- Subject-specific profiles
- Research starters
- Mobile user interface.

The service providers also allow branding customisations to the interface such as logo, look and feel as per needs, Search Widgets to place on the library website, custom search boxes and Catalogue enhancements.

2. RELATED WORKS

The available literature on WSD services comprises of several studies underlining the importance of discovery service for libraries and patrons, their usability, case studies reporting the implementation at different libraries and studies reporting the comparison and evaluation of different available discovery tools.

Several systematic reviews are available in the literature reviewing the available literature on discovery systems. Moore & Greene presented a historical review of the literature on discovery systems with an aim to recapitulate the published available knowledge for the selection and evaluation of discovery tools. They also summarised the criteria proposed by different authors for the evaluation of discovery tools. Similarly, Frederiksen reviewed the available literature comparing the discovery services and published an annotated bibliography on discovery systems. Vaughan proposed features of a model discovery service by analysing the feedback from library staff, the Discovery Task Force report, and vendor interviews. The paper also reported the recommendation of discovery service to the University of Nevada, Las Vegas library. Similarly, Newcomer also reviewed the literature and
proposed recommendations for selecting a discovery tool for music-related content.

Several attempts have been made by researchers in reporting the comparisons and evaluation results of different discovery tools. Foster and MacDonald\(^a\) compared Serial Solutions’ Summon and EBSCO Discovery Service based on their usability. They also studied the behavioural aspects of users while using the search features of individual systems and compared the overall user experience at the Illinois State University. The findings are of importance for the librarians making purchase decisions to understand the user experience and usability of Summon and EDS. Similarly, Djenno\(^9\), et al. compared the usability of Summon and WorldCat Local at the University of Illinois at Chicago. They evaluated the selected two discovery systems based on first the chances of continuing with the particular discovery tool and second to study the search behaviour of users. Their study found faculty preferring Summons than WorldCat Local. Similarly, Chickering and Yang\(^4\) compared fourteen major discovery tools by framing a benchmark comprising of sixteen criteria based on the highly expected features reported in the literature regarding the “next generation catalogues” with an aim to update the librarians about the features of the available tools. Rondi\(^12\) of University of Texas library published a comparative report on Summon, EDS and WorldCat Local using the criteria such as “content, user interface/searchability, pricing, and contract options”.

There are few studies comparing the online academic databases with discovery tools. Asher, A.D., Duke, L.M., & Wilson, S.\(^13\) studied the search efficacy of EDS, Summon, Google Scholar and subscription-based databases. The study reported that the users highly rely on the default search settings provided by the vendors and showed their inability to evaluate sources. Ketterman, E. and Inman, M.E.\(^14\) compared PubMed and Summon and found that the Summon discovery system produces quality results, but it cannot be a replacement of PubMed, a subject-specific database.

There are several case studies reported in the literature reporting the implementation and evaluation of discovery systems at different libraries. Bull, S., Craft, E. and Dodds, A.\(^15\) studied the usability of Primo-based discovery service at the University of Birmingham. Nicholas et. al.\(^16\) studied the usability of Primo in a mid-sized academic research library. Similarly, Cassidy, E.D., Jones, G., McMain, L., Shen, L. and Vieira, S.\(^17\) studied the usability of EDS among the students at the Newton Gresham Library of Sam Houston State University (SHSU) and reported that there is a demand for local customisations in the EDS interface as there is a good level of confusion among the students about the source types and icons, facets/limiters, relevancy ranking, integrated search of EDS. In a similar study, Thompson, J.L., Obreg, K.S. & Abate, L.E.\(^18\) reported their experiences of implementing EDS in an academic health science library. Pinkas M.M. et al.\(^19\) reported their path to selecting a discovery tool for the University of Maryland Health Sciences and Human Services Library.

3. OBJECTIVES
   The objective of this paper is to elucidate in detail the components of a typical WSD system. Further, the paper aims to evaluate typical features of discovery layers in the Big 4’s of WSD. The paper aims to discuss some of the parameters to consider while evaluating the WSD system for purchase decisions.

4. RESEARCH METHODOLOGY
   The features selected in this study have been developed from the available literature dealing in detail on the comparative evaluation of WSD services. The list is derived from Moore & Greene’s listed criteria\(^6\) based on their systemized review of the literature. The features are further tallied from Vaughan’s\(^7\) proposed features of model discovery service which were based on the feedback of library professionals. Although the features list is not a comprehensive list of all features, it helps to understand the availability of the majority of features in selected WSD services. Also, the scope of the comparison is limited to the features only and does not attempt to compare the usability of selected WSD services, based on user’s feedback.

5. SCOPE
   Web Scale Discovery systems have moved from academic research level to production level. Several vendors are seeing it as a global opportunity as a business product and are heavily competing among academic libraries through their products. The major products commonly referred to as the Big 4’s in WSD in the marketplace are WorldCat Discovery Services from OCLC, Summon from Serials Solutions, EBSCO Discovery Service from EBSCO and Primo Central from Ex Libris. Although there are several products available in the marketplace this study is limited to the above four discovery layers only seeing their popularity and time since inception.

5.1 WorldCat Discovery Services
   OCLC released WorldCat Local in the year 2007 after completing a pilot project at University of Washington. Initially, WorldCat Local was developed as a customised and localised copy of OCLC’s WorldCat catalogue. Now it is available in two versions, a preview version WorldCat Local “quick start” and premium version WorldCat Local. The “quick start” version is available to the subscribing libraries of WorldCat catalogue whereas non-subscribing libraries can purchase the WorldCat Local. With the passage of time, OCLC updated its discovery layer and named it as WorldCat Discovery Service. The upgrade involved a refurbished interface and inclusion of more content in its central index due to an increasing number of collaborations with the content providers.\(^12,20,21\) It is an amalgamation of two OCLC products, WorldCat Local and FirstSearch.

5.2 Summon WS-Discovery
   After six months of development time, the Serials Solutions released its WSD solution, Summon in July 2009. Primarily targeting the academic libraries Summon is available as hosted software-as-a-service model to minimise the maintenance cost for the subscribing libraries. The product includes the Summon service and the Summon index. The pricing of the product is not publicly disclosed and depends on several factors such as a size of the organisation, the number of users, and level of
required customisations. The Summon index indexes content sourced from several publishers and journals. Including content providers like ProQuest, LexisNexis, IEE, and Emerald, etc. The unique aspect of Summon is its coverage of Newspaper articles apart from the other content types such as thesis and dissertations, proceedings, music scores, and books. The Serials Solutions regularly releases a list of participating publishers on its website.

5.3 EBSCO Discovery Service
Starting the development of EDS since 2008, EBSCO released its Web Scale Discovery solution in early 2010. EDS emanate out of the EBSCOHost, an already established Abstracting and Indexing service of EBSCO. EDS provides a unified interface for the resources subscribed by the library as well as the resources locally available in the library. The EDS offers a complete solution as per the requirements of a WSD service. It provides Central Index, User interface and other tools like EBSCO’s LinkSource link resolver (Smart Link and Custom link) and Widgets. The price of the product relies on the factors like the size of the local resources, the variety of local resources and the level of customisation required by a library. EBSCO offers this product as a hosted cloud-based platform and restricts local installation which saves maintenance cost from the part of libraries and allows the content provider to implement the updates quickly. EBSCO’s central index includes content from EBSCOhost and other publishers additionally it also covers items from open access resources such as open access repositories and open access journals. The content providers include major publishing houses such as LexisNexis, Cambridge University Press, IEE, Ingenta, Emerald, Elsevier, Wiley, Taylor and Francis, etc. Recently EBSCO has entered into a collaboration with Web of Science to provide citation counts.

5.4 Primo Central
Primo Central is an extension of Primo the next generation discovery layer of Ex Libris. The third version of Primo was officially released in 2010 as Primo Central Ex Libris’s WSD solution. Primo Central consists of the Primo discovery layer and central index. The local library or Ex Libris can host the discovery layer while the central index is hosted by Ex Libris only. The price of the product depends on hosting chosen by the instance and other factors such as the user base and the number of resources to be included in the central index. The Primo Central Index harvests millions of electronic scholarly resources from aggregators, primary and secondary publishers, and open-source repositories. Ex Libris collaborates with the content providers through a Primo Central Publisher Program available on its website. The publishers already in an agreement are IGI Global, LexisNexis, Wiley-Blackwell, Thomson Reuters, etc. apart from these several open access repositories like archive.org, ePrints, DOAJ, etc. are harvested.

6. Big 4’s in WSD: Comparative Analysis of the Features.
The analysis of the selected big 4’s of WSD is based on the content analysis of their websites and evaluation of each discovery layer based on the available trade literature and methodology mentioned in methodology section. Table 1 presents an overview of the features available in the Big 4’s of WSD.

It could be said from the above analysis of the features of Big4’s that all the above-selected discovery layers attempt to add new features apart from the typical features of WSD. However, to rank these web services a detailed analysis of features using weighted checklists would be required.

7. SELECTING DISCOVERY LAYERS
Since the discovery layer is emerging as a vital service in any library and particularly in academic libraries and the more and more discovery service providers are entering the marketplace the selection of the appropriate layer for a library becomes challenging. It becomes more challenging due to the presence of too many parameters to consider while evaluating the discovery layers. On the basis of the features in Table 1, this section discusses some of the parameters that one should consider while selecting an appropriate discovery layer.

7.1 Content Comprehensiveness
The coverage of content by a discovery service is an important factor while considering the discovery service for adoption. One should note that not everything subscribed by the library will be covered under WSD tools because some publishers have not yet started to release their metadata for the general public or are ready to sign a contract with such third-party discovery systems. The broader the content coverage, the better the discovery system will be for a library. Due to this, almost all the major players keep on increasing their content coverage.

7.2 Metadata Harvesting Efficiency
The efficiency to harvest the metadata as soon as it is available from the content providers is also an important factor as delay in metadata harvesting will lead to non-availability of “known-items” in the discovery system and would eventually generate distrust about the ability of the service.

7.3 Request Processing Speed
One of the reasons most libraries prefer discovery services over federated search engines is the latter’s slow request processing speed. The internal architecture of federated search engines requires sending user query to several databases and then post-process the received results before display. However, all the distributed databases are not equally efficient to post the responses on time causing to increase response time. Similarly, in the case of discovery systems, the request processing speed matters a lot. The request processing speed is measured as the time spent by the system after the user query is received and the results are displayed to the users. During the trial period, one should always do a small statistical evaluative study to check the request processing speed of the service.

7.4 Relevancy Ranking
Among the long list of retrieved results from the discovery service, it is important to order the results in such order that the
### Table 1. Comparison of features of Worldcat Local, Summon, EDS, Primo Central

<table>
<thead>
<tr>
<th>Feature</th>
<th>WorldCat Local</th>
<th>Summon</th>
<th>EDS</th>
<th>Primo Central</th>
</tr>
</thead>
<tbody>
<tr>
<td>Company</td>
<td>OCLC</td>
<td>ProQuest/Serials Solutions</td>
<td>EBSCO</td>
<td>Ex Libris</td>
</tr>
<tr>
<td>Year</td>
<td>2007</td>
<td>2009</td>
<td>2010</td>
<td>2010</td>
</tr>
<tr>
<td>Content coverage</td>
<td>WorldCat and other content providers</td>
<td>Several Content providers</td>
<td>EBSCOhost and other content providers</td>
<td>Several content providers</td>
</tr>
<tr>
<td>Relevancy ranking</td>
<td>Proprietary. Match on Keywords, (no full text), Proximity searches in title subject and author. Currency, Weights to library available records</td>
<td>Proprietary. Match on keywords (both metadata and full text), Proximity and frequency based relevance ranking, Weights based on the type of content, citation counts, currency</td>
<td>Proprietary (but disclosed the criteria). An algorithm based on Weights to Subject headings, author-supplied keywords, title, abstract, full-text. Currency. No weights to library available records.</td>
<td>Proprietary. An algorithm based on term frequency, field weighting, number of times a record has been accessed, and currency. Individual libraries can choose boosting metrics. Also, it includes content enrichment service data if subscribed by the library.</td>
</tr>
<tr>
<td>Linking resolver</td>
<td>Own link resolver</td>
<td>Uses library’s link resolver</td>
<td>Uses Smart links and Custom links and library’s link resolver</td>
<td>SFX Link Resolver</td>
</tr>
<tr>
<td>Result filters</td>
<td>Include author, content, format, language, topic, audience, and year</td>
<td>Content type, Subject terms, Full text, local availability, publication date, language</td>
<td>Dynamic facets are available depending on the content type. Filters like source types, Publisher, publication, Geography, content provider (local or external)</td>
<td>Title, author, and publication information, database, resource type, geographic, language, classification, genre, and title of the journal.</td>
</tr>
<tr>
<td>Export/RSS feeds</td>
<td>HTML, rich text, CSV, or the citation tag format RIS, Reference managers</td>
<td>Email, RSS feed, Reference Manager</td>
<td>Email, Reference managers, XML, MARC21, user-specific alerts as RSS feeds</td>
<td>Email, Delicious, Connotea, and reference management software, RSS feeds</td>
</tr>
<tr>
<td>Availability of widgets</td>
<td>WorldCat Local search box widget</td>
<td>Summon Search Box</td>
<td>EDS search box, widgets for openDOAR, JCR, Wikipedia, Scimago Ranking</td>
<td>Search box, Social networking buttons.</td>
</tr>
<tr>
<td>Web 2.0</td>
<td>Supports social sharing</td>
<td>-</td>
<td>Supports social sharing</td>
<td>Supports social sharing</td>
</tr>
<tr>
<td>Catalogue enhancements</td>
<td>Google Books Integration</td>
<td>-</td>
<td>Google book preview, Table of Contents from LOC, Goodreads user reviews</td>
<td>Supports Content enrichment services such as Syndetic Solutions, Amazon and Content Café.</td>
</tr>
<tr>
<td>Usage statistics</td>
<td>Adobe SiteCatalyst</td>
<td>Summon Analytics</td>
<td>EBSCO admin</td>
<td>Eclipse BIRT (Business Intelligence and Reporting Tools)</td>
</tr>
<tr>
<td>Mobile UI</td>
<td>Available with dedicated App for Android</td>
<td>Available</td>
<td>Available</td>
<td>Available, SMS integration</td>
</tr>
<tr>
<td>Customer support</td>
<td>Available</td>
<td>Available</td>
<td>Available (country level)</td>
<td>Available</td>
</tr>
</tbody>
</table>

result that is most likely to be of interest to the user is placed on top. For achieving this, computer algorithms are applied to the result set. These algorithms differ as per the category of users for whom the system is intended for. For discovery services, the relevancy ranking is a significant factor, because of the poor ranking of results due to the poor ranking of results the user will be unable to find the records of their interests despite the availability of a record in the system (comprehensive coverage). Most of the discovery services have proprietary
relevance ranking mechanisms and use custom developed algorithms. Some vendors have disclosed the parameters that their algorithms take care while relevancy ranking. EDS claims the following parameters; Match on subject headings from controlled vocabularies, match on article titles, match on author keywords, match on keywords within the abstracts, match on keywords with full text. The relevancy ranking of discovery services can be evaluated by statistically studying the results obtained when searched for a ‘known-item’.

7.5 Integration with ILS and Other Systems

Discovery systems should have compatibility to integrate seamlessly with the Integrated Library Software (ILS) of the library. The integration should be without manual intervention like uploading files manually or from the database to vendor specified locations. The discovery service should provide tools to display the availability of items in a user-friendly manner. Apart from this, the discovery should also have the ability to broaden the search to other libraries to which the library has signed an ILL agreement.

A thorough verification of the discovery service capability to integrate with other library online services, such as institutional repositories, newspaper clipping service, library guides and learning management systems before finalising a purchase.

The above list is merely an indication of few facets to be considered while evaluating discovery layer a similar list should be developed by considering the local requirements of the library, as the WSD layer service efficiency highly depends on local needs. Thorough research on the available customer success stories and evaluation studies would also be helpful in easing the selection process. Similarly, requesting the vendors for a fully customised trial of the WSD service (including the catalogue and other library resources) and conducting local evaluative study will be the best way to conclude.

8. CONCLUSIONS

Resource discovery tools help the users to search the local library resources, institutional repositories, open access content and the subscribed external resources at the same time within a single interface.

In this paper, the technical aspects of a typical WSD service are discussed by giving the overview of the components of WSD service like central index, link resolver, relevancy algorithms and interface. The paper further outlined the features and a brief history of four major players of WSD, the Big 4s. Similarly, the primary criteria for selection of WSD layer for a library are discussed in detail including the content comprehensiveness, metadata harvesting efficiency, request processing speed, relevancy ranking, and integration with ILS and other systems.

WSD developments have seen a vast improvement in the coverage of central index to several user-friendly improvements in the user interface of these systems. As the WSD systems evolve over time and widely adopted by libraries, the limitations of the systems will be overtaken. On the similar lines, the vendors will introduce several innovations as the needs of the information seekers evolve.

REFERENCES


**CONTRIBUTOR**

**Dr Vinod Kumar** received his PhD in library and information science. Currently working as an Assistant Professor at the Department of Library and Information Science, Babasaheb Bhimrao Ambedkar University (A Central University), Lucknow, India. He is having more than seven years of experience in teaching and research of library and information science. He has successfully guided students leading to MPhil (LIS). His research interests are: Library online services, open data, research methods and digital libraries.