

Full-text Global

Search Globally, Retrieve Locally

The One Great Scholarly Search Engine has become something of the Holy Grail of indexing..
(Larson and Arret, 2001 in Willinksy, 2006).

If you can't find the Holy Grail, create it!

A White Paper proposing Full-Text Global

The objective of this white paper is to propose a concept, business model and technology - **Full-Text Global** (FTG) - for further collaboration and development requiring expertise beyond that of the idea's author. No special consideration is sought, other than inclusion in the project if reaches the next phase – seed funding for development. It is presented here to develop discussion around the concept in principle and how it could be implemented, inviting comment from those with varying expertise. Specifically, the goal is to work towards developing:

- a single, comprehensive search application customizable for all varieties of access privileges to online scholarly content encountered worldwide at libraries
- ensure that this search application indexes with a logic of 'most direct route to highest quality form of all available material according to the user's preferences and privileges' – including all OA material and all material accessed through library privileges.
- Provide a search application indexing all titles and providing a filtered search to all full-text freely available materials (OA); available free for the 'subscriptionless researcher'; an individual anywhere in the world with no special privilege beyond access to the internet.
- Provide access to high-quality index and search to all people, taking advantage of increasing availability of gratis material through OA and other access programs that vary geographically.

The concept is geared towards peer-reviewed scholarly journal articles, however depending on demand it could be expanded to include all types of library material, with P/R research a filter option.

Rationale - Toward for a Commercial Open Source Software enterprise for a Federated Global Search of Scholarly Resources

- ❖ Available to all individuals worldwide with an internet connection
- ❖ Customizable for Libraries worldwide that grant access privileges to members
- ❖ Maximizing access and uptake of research in the Majority World

The world increasingly has access to full-text scholarly articles online, but as yet has no comprehensive and centralized search engine that can deliver full-text articles from the wide variety of routes to access that have emerged. This is a particularly acute and distressing problem for the developing world, where the benefits of open access and concession access programs will largely be lost if scholarly material goes unused due to the lack of a straightforward search engine that assists the user to search-navigate-retrieve articles. The rigour, quality and objectivity of any literature review aspect of any research project is

compromised if any article is selected in on the basis of convenience sample, a distortion that is introduced when the search engine used is either not comprehensive or not discriminating enough and buries relevant research too far down to find. Databases are limited by the contents in the database. Search engines can index metadata harvested globally without technical limitations, but suffer from other limitations as follows:

- a) They may not be sensitive to the user's access privileges by institution, country or region (particularly non-library search engines).
- b) They may not index open access full-text appropriately, particularly gold OA journals (particularly library search engines).
- c) They may not contain sufficient federated capabilities for filtering subjects, languages, peer-reviewed vs. primary source materials etc.

In essence, each provides a substantial degree of possibility of error and none are sensitive to access inequalities in a way that minimizes their impact. The improvements provided by Serials Solutions 360, in that regard, widen the information divide by serving to solve only the problems of the largest research libraries globally. There are two types of error and two types of access inequality that can be addressed by **Full-Text Global**, a comprehensive and customizable global federated search engine for scholarship. Subscription-rich patrons and subscription-poor researchers are prone to both errors.

Errors

- 1) Type I error – False Positive - Convenience Sample. Search is not sensitive enough, articles are included just because they are available. This is common with Google Scholar. Another possibility is that the refereed post-print is available to the user by a less convenient route, but they use the more readily available version that is not peer-reviewed, thus wasting the value of the peer-review process.
- 2) Type II error – False Negative - Article Not Found. The search leaves out articles that are relevant. This common to proprietary databases that are limited by the database composition, one may be unable to retrieve articles in searches such as Scholars Portal, PubMed etc. where articles appear under same keywords in Google Scholar. With Google Scholar, many hits point to pay-per-view (PPV) access. There is no instruction on how to retrieve the article by other means (library log-in, ILL, direct from author, document delivery service (INASP), etc.), so the article is passed up. Subscription journals and OA journals may not be retrievable within the same indexes and databases.

Access Inequalities

- a) Literature Access – Researchers in developing countries belong to institutions whose libraries typically carry few subscriptions. There are within-country divides in the North and South as well, where smaller libraries, rural and remote areas are at an access disadvantage. Open access creates a level-playing field for all who have internet access. Concession programs for developing countries such as HINARI make available thousands of journals to research libraries, and eIFL and INASP programs also create routes to access through bargaining. Much of this material requires access management, a technical hurdle for the South, both in set-up and in the context of low-bandwidth internet access with frequent interruptions that disrupt the user who has to log-in.

- b) Access to indexing and search – Along with the cost of subscribing to journals that is prohibitive, the cost of subscribing to indexing services provides a major hurdle for research libraries in the poorest countries in the world. Despite the growth of OA and the concession programs, uptake is hampered by the problems of indexing – search, navigation and retrieval.

The movement for open access emerged out of the state of the publishing market in an online environment in the context of the ‘serials crisis’ (Willinsky, 2006). What has emerged is a system that creates wider access to research while introducing distortions into the process of searching, and these occur whether one is at a subscription-rich institution or a subscription-poor one, or if one is not an institutional member at all. This situation has created the following routes to search-navigate-retrieval.

- 1) Your institution’s subscriptions – search by various means and subscriber databases and indexes. Comprehensive search with Google Scholar is possible, but not federated, not sensitive to access privileges.
- 2) Green OA – articles deposited in institutional OA repositories (over 3000) or archived on author’s sites. Can be retrieved over Google Scholar (green arrow or ‘all versions’). May not be refereed version.
- 3) Gold OA – articles published in gold OA journals (over 4000), OR, in mixed access journals (either post-embargo or by author fee open access). Retrieveable in various databases, but indexing is not necessarily consistent between gold, mixed and non-OA resources.
- 4) UN Programs – HINARI/AGORA/OARE – articles available to low-income countries research libraries, available through access management at participating institutions. Article versions may also be in OA repositories.
- 5) eIFL – articles available to research libraries in developing countries by eIFL library consortia bargaining.
- 6) INASP – articles available through INASP PERii program. By document delivery. African Journals Online as well – document delivery.
- 7) Publisher’s Concessions – A number of large to small publishers provide free to heavily discounted access to resources, some simply by recognition of IP.
- 8) Physical library resources and ILL – Interlibrary loan. Depending on the library’s network of agreements for ILL.
- 9) Author e-print or re-print by request. If articles can’t be accessed by any of the other routes, they may be obtained directly from the author by request.
- 10) Public domain digitized.
- 11) ‘Grey market’ – articles may be obtained in ways that circumvent digital rights management of the publisher, but on the other hand may constitute fair use on the part of the end user. For instance, it was reported that physicians in Thailand resorted to purchasing unauthorized copies of medical articles in order to gain the knowledge they needed to save lives, while they were in no position to pay the purchase price for the article or subscription and had no institutional subsidy like most doctors here would.

In terms of the Northern libraries who purchase a number of indexing services to manage their resources, John Willinsky (2006) has stated that “Despite this considerable array of indexes and portals, earnest scholars and students still have to wend their way through

overlaps, gaps, and partiality in the coverage of the research literature that the indexes provide...”(p. 173). When we add OA repositories, OA journals, UN programs, publisher concessions by IP address etc., one wonders if the available wealth of resources is used effectively, and how often specific resources are passed up because of search-navigation issues. One can predict that the more barriers to retrieval, the more complex the search, and the more specified prior knowledge required of the user, the greater chance there will be an overall reduction in the quality, objectivity and cost (in time) of literature search and consequently the research endeavour itself.

Full-Text Global Business Model

“Having one place to search that would include relevant resources would make research less fragmented” (Larson and Arret, 2001 in Willinksy, 2006).

Full-Text Global would be a commercial open source software (such as Sugar CMS) that would serve to address indexing issues in a mixed OA/subscription environment with particular attention to providing a truly global service with regards to the opportunities and needs of



developing country institutions. It would also provide a unique service accessible to anyone with an internet connection, with its basic application for federated global article search and global full-text search. In that sense, it should serve people equally regardless of their location in the world, in the sense that it maximizes everyone’s access privileges and reduces the impact of access privilege inequality. While a pure OA environment is ideal and may be inevitable, the need to manage knowledge resources in a mixed environment provides the possibility of making this product commercially viable.

Full-Text Global will be a company that believes in a vision for Open Access, but provides a service for maximizing access in a world complicated by access barriers. It addresses the primary access inequality - access to literature - indirectly by addressing the second – access to indexing and search. This is because the privilege possibilities as they exist today and as they will increase in the future risk under-utilization because of the complications of search and retrieval. In addressing the indexing issue, FTG also serves to solve problems faced by the large research libraries of the North and provide a service that treats all articles on the basis of the preferences of the user and the route to retrieval, thus harnessing the full power of purchased and OA resources. It should serve to provide the most direct route possible to free access to highest quality version of any article in the world. It will do this by managing the routes to access and deductively and logically providing the link to the article, or information on the most direct alternative route listed above (except grey market!).

The non-profit business model would be as follows:

- 1) Commercial open source software (Creative Commons-type license) for a federated and layered search - similar to Serials Solutions 360, but with attention to ‘searching

all resources available to the patron by any route' as opposed to 'searching the library's acquired resources'.

- 2) Federated Search Service for Institutional Members - Customization service to integrate access privileges of institution, so that members can retrieve full-text available to them by privilege or by OA, from the comprehensive global search while logged in (without having to search Google Scholar and then return). This puts the customized privilege-sensitive search engine and results page on your library website, giving your patrons the best possible search capacity. The federated search would have three levels, two of which are common to the open version and one which is an option for libraries.
 - a. A single search bar, all-in-one keyword search. Simple, and effective (depending on purpose on strategy and keyword skill of user).
 - b. An advanced search filtering for advanced keyword, subject, author, title, discipline, years etc.
 - c. A specified search allowing the user to focus in on databases, and apply specific knowledge of them to their search.

Finally, the search results page, customized to the library patron's privileges will provide information on the best way known to retrieve any article (except grey market!).

- 3) Support Services to these libraries
- 4) Service for Access Management and indexing/search (2+3 above) to consolidate privileges through HINARI/AGORA/OARE, eIFL, INASP, OA + publisher concession access in the **South**. Payment scale model based an ability of institution to contribute - with contribution from UN, eIFL, INASP organizations. Sponsorship and shared programs for North-South-South partners and networks (to outfit entire network of partners with net contribution from them.)
- 5) Federated Open Access Search. For general internet users who have no additional access privileges. Will accept donations and generate a community of developers. This is essentially the basic software and one that can be skinned to any website. One can turn on and off a filter to go from comprehensive abstracts/article listings to only results that provide a full-text. Full-text search means that anyone in the world who can get on the internet will have a global federated search that indexes every journal article available to them (only hits that lead to full-text gratis articles are included, in essence creating the widest library possible for any general user at any given time. But there are several reasons for filtering back in the comprehensive index of titles.
 - a. Your library does not yet have **Full-Text Global**, and you want to use the open version to find the article and then return to log-in to retrieve it (as people often do with Google Scholar).
 - b. You want to ensure that your literature search is not biased by availability and convenience, to know what exists and/or to try alternative routes to access.
- 6) Provide both paid and open search services in as many languages as possible.
- 7) Provide a scholarly and principled *guide* to literature-based research, focused on a logic for routes of retrieval for online materials but also utilizing your closest physical library, ILL, author re-print, document delivery etc. etc. in as many languages as possible.
- 8) Provide a data-based map and accounting of the world's scholarly resources.
- 9) Phase II. Work on web 3.0 Semantic search capacities, *and* research and development into higher sophistication in computer-automated translation services,

towards providing a service of automated computer translation plus human translator peer-review to accelerate high-quality translation of the world's scholarly resources into as many languages as possible.

The advantage for endowed libraries in the North in choosing **Full-Text Global's** business model to provide the comprehensive search for online resources available to your patrons:

- a) the commercial open source commercial not-for-profit model builds sustainability into your purchases:
 - a. no proprietary rent-seeking and no legal hassles with regards to patents
 - b. freedom to build internal capacity to maintain and modify the software
 - c. any programmer globally can work with standardized open source code, and communities develop around open source platforms - you are not reliant on any one company for trouble-shooting, maintenance and improvements.
 - d. you pay for services, not for software and intellectual property.
 - e. payments cover costs for operating Full-Text Global, and 'profits' are invested in the social mission.

- b) **Full-text Global** capitalizes on open access resources by integrating them fully into your indexing services used by your patrons. Even though your library did not have to acquire OA materials, this does not change their value to your patrons. However, material that is green and gold OA and non O may not be searchable simultaneously and seamlessly with information in the result page providing the most direct route to the highest quality version available to the user and all access options. Your patrons' can log-in one time, and be assured of a single, comprehensive and federated search for all materials available to them, and though privileges differ between individuals, the service that maximizes the use of those privileges for anyone in the world would be **Full-Text Global**.

- c) **Full-text Global** has a Social Mission that improves the global environment for librarianship, with long-term benefits to your research library.
 - a. Improved access to literature and to search and retrieval through indexing worldwide directly translates into increased production of knowledge resources by all countries in all regions, a capacity currently well below potential in the vast majority of the world's societies. We may be described globally as operating at marginal capacity today due to the information divide.
 - b. Produced in a research and archival context, these resources will become available to your library's patrons, likely at no cost through OA or at an affordable price. This means more unique, diverse, authentic global sources for your patrons.
 - c. By purchasing **Full-Text Global** services, you will be contributing to indexing access and Open Access' uptake for everyone, regardless of subscription, through supporting the open version of **Full-Text Global**.

Reference:

Willinsky, John. 2006. *The Access Principle*. MIT, Cambridge Press.