STM Best Practice Recommendations for Federated Search of Copyright-Protected Works

INTRODUCTION

STM publishers collectively and individually invest a great deal of effort and resources in making their sites searchable, in refining and improving site navigation, and in supporting content discovery and integration tools

As more and more search services are supplied by specialised providers and intermediaries, and as users themselves engage in customized searches of copyright-protected works, STM publishers wish to offer a set of best practice guidelines to enable providers and users of federated search tools to get the maximum results from this activity, legally and in keeping with customary and contractual provisions. The principles underlying the best practices for federated searching should also be respected by the broader community of vendors and service providers that assist their customers to programmatically or automatically integrate with publishers' websites.

WHAT IS A FEDERATED SEARCH?

Federated search is defined by Wikipedia as:

> An information retrieval technology that allows the simultaneous search of multiple searchable resources. A user makes a single query request which is distributed to the search engines participating in the federation. The federated search then aggregates the results that are received from the search engines for presentation to the user.

In other words, Federated Search software/interfaces enable users to simultaneously search multiple library catalogues (OPAC's), Web Sites (Google, Bing, etc.), subscription and citation databases, and other information sources in one integrated process. Federated search tools, as opposed to spiders or crawlers or other types of content scrapers and harvesters, retrieve unique search results by enabling the collection of results of a single search across multiple data sources, which are then presented to the user in one integrated list of search results. Federated Search software is hosted by either a specific vendor or the subscribing institution itself. The benefits of Federated Search to the end user include speed and breadth of search as well as a single search solution in contrast to a variety of interfaces.

PUBLISHERS' APPROACH TO FEDERATED SEARCHING

Publishers understand that their customers would wish to employ federated search technologies across multiple publisher platforms. In fact, when asked to allow federated searching, many publishers consent, usually through an agreed licence with agreed interfaces (Application Programming Interfaces – or APIs). Through agreed APIs, the publishers can measure usage, and the parties can ensure that the federated searches do not compromise data integrity or system performance. The licence mechanism also allows publishers to be compensated for the use where appropriate.

STM publishers have begun participating in industry efforts to review and validate the Internet Protocol (IP) address ranges used to authenticate customers. These techniques are also expected to be deployed to asses compliance with APIs and licences for federated search and also for permitting text and data mining activities which may rely as a preparatory step also on federated searching.

As a result of deploying these techniques, examples of unauthorized IP ranges have been located which, upon review, are controlled by federated search providers. Separately, STM members report having experienced system slowdowns caused by unauthorized data and text mining through unauthorized IP addresses controlled by federated search providers.

Publishers have to be careful to guard against federated search requests from suspect entities or persons who might seek to engage in unauthorised access and intellectual property infringement or related abuse of networks and systems. To assist publishers to distinguish legitimate federated searching from system abuses, STM suggests the following.

RECOMMENDED BEST PRACTICES

As a matter of good business practice, a publisher should be informed when a federated search provider's IP address range is included in licence between that publisher and the customer. When this does not happen, the publisher cannot configure their systems to balance user demand for content and this can result in system slowdowns for all customers.

Notifying their business partner may also be required of a customer under its contracts with publishers. In any event, to avoid investigation and potential liability for copyright infringement, the federated search provider itself should insist on prior notification of the IP range the search provider intends to rely on.

1. Federated search providers should not circumvent access controls without the knowledge and consent of the publisher. For content freely available from publisher sites, federated search providers should observe publisher requirements and rules with respect to robots, crawlers, spiders, and similar technologies. In most cases, this will nowadays involve the entering into an agreement on APIs, ie the grant of a licence for an interface - not for content. Such agreements do not, as a rule provide means to or authorise the authentication for access to publisher's content.

2. Federated search providers should not maintain or authorize the maintenance of databases containing copyrighted material from publishers without the express written consent of the publishers or their agents.

3. Any agreement with publishers for federated search should include agreed APIs, or some other mutually agreed upon method, for access to the content by the federated search provider. Fees, if any, should be stipulated or if applicable waived therein, depending on established trade practice between the parties.

4. Federated search providers should maintain appropriate security measures to protect any content (including metadata) lawfully downloaded from publisher websites. This includes ensuring the security of proxy servers which may be used for such purposes.

5. To the extent that customers or users are able to access content through the federated search provider, the search provider should employ appropriate protections to ensure that only authorized users from subscribing customers (institutions or commercial customers) can access the content. This includes protecting content against password sharing and bulk downloading.

6. Federated search providers and publishers should consider agreeing on carrying out security audits from recognised providers.

7. Federated search providers should immediately notify publishers of any violations of any of the activities or situations listed above.

8. Prior to accessing content through APIs, federated search providers should request their customers to confirm in writing that the customer has fulfilled its responsibility that the credentials for access have been met

9. Where federated search providers continue to rely on IP ranges for access, whether supplied by their customers (licensees of publishers) or the publisher's agents, or obtained independently from the publisher, the IP ranges should periodically be reviewed in order to avoid that IP ranges may be added incorrectly, or remain incorrectly listed after an IP range change, or otherwise without the knowledge and consent of the publisher concerned.

*****

Copyright and Legal Affairs Committee, STM

January 2013