One Repo Project Description

Sebastian Hammer [quinn@indexdata.com] - July 20, 2015

The One Repository project (1Repo) seeks to establish a single metadata set covering all open access scholarly articles. As a basis, this includes the metadata about the contents of relevant repositories and publications, but, depending on interest and funding, it may include additional sources that fit within the scope of open access scholarly communication.

The project has four specific goals which, in combination, offer tremendous value across a broad range of activities related to scholarly communication and higher education:

> 1) It will be comprehensive, covering as many sources as possible, including hard-to-extract sources such as smaller publications and non-standard repositories. Many other projects presently pass over these more challenging sources, prevalent outside of the STM sector, because of the technical challenges of including their contents.
>
> 2) It will be up to date, providing timely updates across all included sources, and adding new sources as they become known.
>
> 3) It will provide a uniform and high quality of metadata, in so far as the source data is available. This goal will be achieved by coordinating data models with other projects, and by working with each data source to seek a uniform quality and granularity of metadata. RIOXX is considered, both as a rich set of metadata elements and as one (among several) formats which will be made available for users of the 1Repo.
>
> 4) It will be completely free and open for all uses, including download in suitable data formats through OAI-PMH and other mechanisms, online searching through a user interface, through widgets which can be embedded in websites, and through APIs. The project will seek synergy with open linked data activities to maximize the value and leverage of the data as well as new modes of collaboration.

It is our belief that no other present efforts include these four attributes in practice, yet in combination, they have the potential to make the 1Repo a key infrastructure component and enabling service for an enormous range of other activities.

Index Data has the technology and the background to establish the 1Repo. Once established, the inherent value of the 1Repo will suggest a variety of sustainability models, but we currently seek funding to help us more quickly establish a critical mass of coverage.

The challenges to be addressed include both methodologies for curating the set of sources, which extend beyond those that are easily accessible through open standards, and the techniques for accessing these sources to retrieve metadata. Index Data's toolset includes a captive browser engine which is guided by our connector maintenance team to crawl through structured websites to extract metadata fields from HTML-formatted text, as well as normalization pipelines for cleaning up and repairing incorrect or inconsistent metadata (if necessary, the operators of repositories will be contacted directly to retrieve metadata in alternate forms). A team of data source curators and connector authors, to be extended to meet the needs of this project, operates the technology; indeed, the team and its approach to metadata, is arguably as important as the technology itself.

It is our premise that many metadata-based projects in the OA space limit themselves to the most accessible sources, or are hampered by the variable quality of the metadata returned. Our hypothesis is that if a public, high-quality metadata set can be established, applying a level of rigor comparable to that of commercial A&I services, this will have numerous applications and be disruptive in the space.