BEST
Review of Related Projects

Author: Scott Yeadon
Date: 26 September 2007

1. Purpose

The purpose of this document is to provide a brief outline of the international projects reviewed as part of the BEST initiation process and make a recommendation to guide subsequent BEST technical development.

2. Introduction

In establishing the BEST project a number of existing projects were reviewed to determine whether concepts, code, standards and schemas could be applied. It is worth noting that the key goals of APSR projects include the encouragement of international collaborations, building on the output of other projects, developing lightweight platform- and repository-agnostic solutions, and promoting the use of common and emerging standards over customised solutions. In addition, the use of OAI-PMH for discovery and exchange of information in APSR software is encouraged based on the popularity and stability of the standard. Several leading projects in the area of repository usage statistics aggregation and analysis were identified and reviewed in the context of the above goals.

3. Approach

Each of the project websites and documentation available from the website were reviewed and project participants contacted where required to provide further information or advice. A brief description of each is provided below, for more detailed information refer to the individual project websites, URLs are cited at the end of this document.

4. Projects Considered

Interoperable Repository Statistics (IRS)[1]

According to the project website, the IRS was to “investigate the requirements for UK and international stakeholders...and build distribution and collection software for repositories as well as generic analysis and reporting tools. The project will design an application programming interface (API) for gathering download data, and implement this for data providers who are using common IR software” . In a follow-up discussion with Dr Leslie Carr, OAI-PMH harvesting was considered as a possible collection mechanism, however the project found harvesting event information at an item level was not workable. The IRS ultimately delivered a test implementation based around ePrints but able to be used in other contexts. It requires the use of a specific database schema and the population of it by any adopting repository. An access point to Citebase (an existing citation impact service) has been developed to provide some measure of publication impact. While the project has defined a very useful set of deliverables, these have not as yet been made available on the project web site.

SUSHI[2]

SUSHI comprises a NISO Working Group examining the problem of machine-to-machine statistics harvesting and the availability of statistics to customers (repositories, vendors, etc) in a standard container. SUSHI makes use of Web Service technology (SOAP, no REST equivalents) to enable repositories to request reports from COUNTER and other sources as long as compliance with the schema (essentially COUNTER-based) is retained. The SOAP requests encapsulate a fixed set of reports which can be requested by repositories. This approach also lends itself to a harvest model, where a set of known reports (essentially contracts fulfilled by repositories) could be exposed to aggregators periodically or ad hoc. An initial version of the software is available for use (v0.1)

COUNTER[3]
COUNTER "is an international initiative serving librarians, publishers and intermediaries by setting standards that facilitate the recording and reporting of online usage statistics in a consistent, credible and compatible way." COUNTER publishes Codes of Practice providing the conditions/rules parties must subscribe to in order to be COUNTER-compliant. Currently COUNTER covers online journals, databases, online books and reference works.

MESUR[4]
MESUR is a LANL initiative funded by the Andrew W. Mellon Foundation "to investigate metrics derived from the network-based usage of scholarly information". A related paper, possibly a precursor to MESUR, provides a comprehensive description of an architecture for a statistics aggregator service[5]. The MESUR site contains additional publications which may assist in informing the design process.

Developing Archival Metrics in College and University Archives and Special Collections[6]
An international collaboration attempting to obtain agreement on the need for core metrics from a range or archival environments and subsequently developing a standardised set of metrics in the archival and special collections domain. The focus appears to be on a standard way of self-assessing an archive in relation to its value to its customers (researchers and global community).

5. Recommendation

In making a recommendation a number of factors (outlined in the introduction) were taken into consideration. MESUR was deemed to provide a comprehensive model (via research papers) and suggests a promising longer-term solution should another iteration of BEST development proceed in 2008. Given the time constraints on the project (approx 3 months from start to completion) the familiarity with OAI-PMH and the expectation that openURL is a direction many of our related projects will ultimately take, the ideas from MESUR have been selected as the start point for the BEST development. It is worth examining the IRS implementation to determine whether any data models, code, schema, etc can be re-used. The COUNTER concept of a Code of Practice is also recommended. In the pilot phase this should be limited to identifying those policies of event reporting needing clear agreement between partners (e.g. approach to robots/crawlers, double-click, information provided, etc).

6. Technical Approach (Preliminary)

Note the following is based on initial ideas taking into consideration time constraints and APSR goals, it is not based on a detailed design exercise which will take place subsequent to this review. Some or all of this approach may change.

It is recommended that the concepts as well as the actual XML schema outlined in [5] be adopted (or at least considered during the design process) in conjunction with OAI-PMH. For the pilot, deployment and use of link resolution servers will not be in scope. Instead a set of agreed URLs (essentially REST-like service points) will be used by participating repositories that will act as a resolver service for URLs linked to from the aggregator. It may be possible to implement this functionality entirely in OAI-PMH where a repository has a Data Provider service, using requests such as "ListRecord" to drill down to item-level metadata.

The OAI-PMH approach outlined in [1] will be used. That is, repository event information will be harvested rather than focussing on the exchange of reports or harvesting events as item-level metadata. This will require the establishment of some simple event vocabularies to ensure equivalence across repositories. In addition to events, as in SUSHI certain fixed reports may have to be generated (e.g. object format summary, object-by-subject summaries, etc) where information is unable to be ascertained from events or where the overhead of subsequent follow-up requests for metadata information by the aggregator is unlikely to scale.

[1] http://irs.eprints.org/
[2] http://www.niso.org/committees/SUSHI/SUSHI_comm.html
[3] http://projectcounter.org/
[4] http://www.mesur.org/Home.html
[5] J. Bollen and H. Van de Sompel. An architecture for the aggregation and analysis of scholarly usage data. http://www.citeulike.org/user/scholze/article/669192
[6] http://www.si.umich.edu/ArchivalMetrics/