Project Report of the

Benchmark Statistics Service (BEST)

Contents

Attachments

1. Background

As part of the Australian Partnership for Sustainable Repositories' aim to improve the management of scholarly digital assets, the Benchmark Statistics Service project (BEST)[i] was designed to enhance the type and quality of statistical information about repository holdings and usage. The problem to be solved relates to the strategic need for better, standardised, statistical information to inform a wide range of policy and funding decisions within the scholarly communications cycle. In order to address this need support was secured for APSR to manage a short pilot project, which was initially envisioned to include the production of a pilot harvesting and aggregation service. Due to personnel and time restraints the project scope was revised in October 2007 to identify an approach and initiate the design of a pilot service, providing the framework for further development.

2. Approach

The BEST project was managed out of the APSR office, and directed by a reference group representing the participating institutions (see Attachment F). The reference group met formally once during the project, corresponded by telephone and email, and were able to access project documents on the BEST Wiki.

A list of related work was compiled and any additions sought from the Reference Group. Scott Yeadon reviewed all related projects, summarised their status and identified potential for collaboration (see attached Review of Related Projects).

One of the main challenges encountered in designing an automated service to combine the statistics of numerous repositories is the variation in the manner in which content and usage information is structured, stored and presented. While some standards—such as OAI-PMH support and Dublin Core standards for metadata—are likely to be followed by most institutions, there are no widely used standards governing the type and format of repository content and usage information. This is due to conceptual issues (what constitutes a measurable 'item' in a repository to which metadata should refer; how are complex collections identified and linked, and their usage measured) as well as difficulties with identifying and enforcing standards (where standards exist they are usually specific to individual repositories, and may not be strictly enforced to encourage the use of the repository by contributors).

It was recognised that defining universal standards and enforcing compliance was well beyond the scope of a short project. Instead, the BEST project considered existing standards and works in progress, and defined a minimum set of terms and conditions to guide the BEST aggregator service.

3. Outcomes

Progress achieved toward the milestones, as identified in the project specification document, is summarised in the table below.

Milestone	Progress	Follow up
1. Specification document	Complete (Oct 2007)	
2. Evaluation report	Complete (Oct 2007)	
3. Reference group report	Complete (Nov 2007)	
4. Code of Practice document	Drafted; renamed 'Guiding Principles' (Dec 2007)	Add section on licencing/user agreements; circulate for comment and endorsement
5. Project technical specifications	Partially complete (Dec 2007)	Circulate Event Interchange Model (EIM) for comment, design aggregator service and front-end
6. Conference presentation	Outline complete (Dec 2007)	Select main points relevant for event; refine powerpoint
7. Partner audit reports (gap analysis)	Incomplete	Identify work needed for each repository to expose metadata and event information as described in the EIM
8. Project documentation	Complete (this document)	
9. Project handover	Cancelled (service not ready to be transferred to APSR National Services)	

Instead of the project being transferred to APSR National Services (milestone 9) this document, along with the accompanying technical and project documentation and record of correspondence with the reference group, constitute a record of the current status of the project.

The major focus of this project was on defining an approach for the development of a pilot aggregation service. Importantly, this involved reaching agreement on priority use cases and identifying and defining the data elements and formats that would be required to address the main questions that could be asked of the aggregated statistics set. The main outcomes of this process, which could be used as a starting point for future work, are summarised in the Definition of Terms, the Priority Scenarios, and the Guiding Principles (see attachments). These three documents were developed from discussions at the reference group meeting in October, and were further refined by email.

Preliminary feedback suggests that most data elements are able to be drawn directly (or with minimal mapping) from web logs or OAI providers. The Event Interchange Model was developed fairly late in the project and, due to seasonal circumstances (Christmas) it was not possible to circulate this to the reference group for comment, or to perform a more in depth gap analysis on what work would need to be carried out at the repository level to make the harvest of standard reports possible.

Deciding on authority lists for subject and item type metadata elements generated quite a bit of discussion. Ultimately it was agreed that one authority list was needed to make subject information useful, and that the RFCD[ii] codes were the best option.

Many repositories not currently using these will need to start using them anyway, to meet RQF reporting requirements.

Item type was less clear. Initially the reference group discussed using the authority list and definitions being developed by MACAR[iii]. However, on reviewing the draft list it was thought that more granularity, and being able to separate non-scholarly from scholarly works, would be necessary if the service is to be used for research assessment. Linda Butler has suggested that most repositories probably use fairly standard categories, based on those defined by DEST in the mid-1990. The list of types used by the ANU is available online[iv] and this could be used as a basis for a BEST controlled vocabulary if other repositories use a similar breakdown that could be mapped to these types. It is likely that item types in the 'other' category in this system would need to be expanded to fill the needs of non-academic repositories. In particular types for datasets, still and moving images, manuscripts and sound would need to be added.

4. Lessons Learned

The meeting of the reference group in October was a very useful way of getting rapid input and feedback on the project design and approach from most of the participating repositories. Achieving the same results purely through remote communication would have been a much longer and perhaps impossible process.

Almost all partner institutions were actively involved in refining the scope and approach of the project, and identifying priority queries for the aggregator service to cover. Correspondence suggested that the ANU Supercomputing Facility felt that the service would only cover repositories holding academic research outputs, and that ANUSF contributions would not be useful. Increased involvement of ANUSF would be desirable as a way of including a different type of repository, and to begin to test the extension of the best service to other repositories.

The lack of a single authority list for authors limits the usefulness of queries using this field. At an aggregated level, the kinds of questions that are of interest require that an individual person has a unique identifier that can be used by any repository, so that all past and present statistics regarding that person and their work can be included. Significant variation currently exists within institutions in the representation of individual authors' names. This issue is compounded by cross-institutional differences. Due to its nature an author authority list would need to be maintained by a national body, and should be compatible for use with international lists. One possible solution could be to use the preferred name from the National Library of Australia's Australian Name Authority File, which is being enhanced and made publicly available as part of the *People Australia* service[v].

5. Future developments

The BEST project has described how a pilot usage statistics aggregation service could work with a limited number of repositories, facilitated engagement between the partner organisations, and produced the preliminary technical design documentation. This section lists some of the immediate activities required to build and implement the service, as well as some of the issues that were identified as being worthy of further consideration but were beyond the scope of this project.

It would be ideal if future work, under the Australian National Data Service or a similar program, could start early in 2008 so that momentum is maintained and the views and issues identified during this project still apply. If there is a considerable gap then additional work may be required to re-establish the activity of participating organisations and review the Guiding Principles and technical documentation.

- Identify new governing structure and organisation:
  How will future development of this service be resourced, and where will the central contact point be? How will the service be maintained in the long run?
- Review list of priority queries, particularly addressing issues with queries involving ranked lists.
- Finalise choice of authority lists for resource type and file format
- Design technical framework for aggregator service, and front end for users.

- Review Guiding Principles.
- Complete gap analysis and resulting repository-level activities:
  The preliminary analysis will need to be checked against the reviewed technical specification, and any remaining gaps in the capacity of participating repositories to provide reports as specified in the Standard Statistics Specification will need to be addressed.
- Code harvesting and aggregator service
  This will require a programmer to build the harvest, aggregation, and presentation sections of the BEST service (letters A through D on the functional diagram).
- User documentation
  The information, instructions and disclaimers to accompany the pilot service. Much of the information should be able to be drawn from existing project documents, the technical documentation and the Guiding Principles.
- Test and refine pilot service
  The pilot service will need to be tested using reports from the participating repositories, and any problems addressed.
- Publicise and launch pilot service
- Expand service capabilities and national distribution
  This would include integration of the service with other related applications, including ORCA[vi] and AONS[vii]. Additional queries that were identified by the BEST Reference Group as being of potential interest but beyond the scope of this project could be addressed here (see Guiding Principles). The service should be extended to as many national (and international?) repositories as possible.
- Gather feedback from service users

Out of scope developments:

Tracking user behaviour, threads, object proximity information

Determine discipline norms for access events (e.g. average downloads/month for discipline X) so that comparison of usage statistics is more meaningful

---

[i] http://www.apsr.edu.au/best/index.htm
[ii] http://www.abs.gov.au/ausstats/abs@.nsf/66f306f503e529a5ca25697e0017661f/955FFA4EB1B23847CA25697E0018FB14?opendocument
[iii] Metadata Advisory Committee for Australian Repositories (http://www.arrow.edu.au/macar).
[iv] http://www.anu.edu.au/ro/publications/categories_guide.doc
[v] http://www.nla.gov.au/initiatives/peopleaustralia/
[vi] http://www.apsr.edu.au/orca/index.htm
[vii] http://www.apsr.edu.au/aons/index.htm