Australian Partnership for Sustainable Repositories
Benchmark Statistics Project (BEST)

Guiding Principles

Version 4 November 2007

Contents

Appendices

1. Executive summary

The aim of the BEST project is to facilitate the harvest, aggregation, analysis and presentation of repository content and usage data. In this initial phase the participating repositories come primarily from research institutions, which has affected the identification of user groups—repository administrators/managers, meta-researchers/policy makers, and researchers/authors—and the choice of priority questions for the aggregator service to address. The BEST Reference Group, comprised of representatives from participating repositories, has had a key role in guiding the direction of the project, identifying priorities, and defining terms and standards.

While many repositories present information on their holdings and usage, and some existing services combine some content information, there is currently no federated service to combine and compare usage statistics from Australian repositories. Likewise, there are no industry standards for the storage and presentation of repository statistics, and so the BEST project takes an approach that uses or is compatible with widely used systems and standards, is Open Access compliant, platform agnostic, and extensible to meet future requirements.

This service will, in the first instance, be of use to repository administrators and managers, meta-researchers and policy makers, and primary researchers and authors of repository content. The priority queries address these user groups.

Current variation in the format and definitions of repository content and usage data limit the potential uses of an aggregator service. By participating in the BEST service participating repositories agree to comply with general principles relating to the processing of usage data, and use the definitions and formats specified in this document and the accompanying Event Interchange Model in the reports that will be made available for harvest by the aggregator service. Inevitable discrepancies and variations from these rules will be documented by repositories and this information made available with the aggregated statistics.

The BEST reference group (or the relevant succeeding body) will review this document and make revisions as necessary.

2. Purpose

This document is intended as a general guide for repositories that are registered with the BEST service and those that might be considering registering in the future. It will act as an informal agreement between participating institutions—outlining what the service should aim to provide, and what needs to be done to make it possible.

This is not a static document; these principles will be reviewed regularly so that the service continues to address real needs and work with current practices of the partner institutions.

3. Scope

This document seeks to facilitate the harvest, aggregation and analysis of usage data from repositories registered with the APSR Benchmark Statistics (BEST) service. Most of the repositories participating in the pilot project have a research focus, and the following three user groups have been identified:

Repository administrators and managers

Meta-researchers and policy makers

Researchers and authors

It is anticipated that the users of the service will change as new potential applications emerge, and as the capacity to include different types of repositories grows. At present the pilot service is aimed at these three groups, and the content of the reports harvested from individual repositories and the presentation of the aggregated data will be designed with these groups in mind. The priority use cases/queries are presented at Appendix B, and those that have been identified as useful but outside the scope of the BEST service at this stage are at Appendix C.

Limitations

The participating repositories will provide the data in the usage reports in good faith, understanding that it will be freely available on the World Wide Web. None of the participating repositories, or the BEST project team, shall be responsible for any non-intended use of the data, or for any erroneous interpretations.

End users should be aware that there is much variation in the definition and processing of data relating to repository contents and usage, and that participating repositories and depositors may have an interest in manipulating statistics. Every effort is made to ensure data integrity and align relevant definitions and practices. However, no standards are currently in use for the harvesting or aggregation of repository usage data, and it is outside the scope of this project to endorse and enforce an industry standard.

The aggregated usage information is limited to the reporting of events (as defined in the Definition of Terms), and does not infer any reliable assessment of user intent or academic or other impact of the resource. It may be possible in the future to make comparisons between the usage information and other measures of research impact, such as citation indices.

4. Reports

Reports should be provided in the format specified in the BEST Standard Statistics Specification and be made available through the use of web services.

5. Processing

Any filtering of repository usage information will be performed by repositories before the data is presented for harvest.

It is recognised that repositories have different processing practices and standards. The general principles and minimum requirements for the filtering of usage data and repository statistics before they are presented for harvest follow. Terms in italics are used as defined for the BEST project in the Definition of Terms at Appendix A.

Visitors

Information about visitors to repositories may be useful to repository managers, policy makers and meta-researchers. The main purpose of examining visitor information is to assess the interest in and use of the material in the repositories by the target audience by counting access to the metadata and full versions of repository items. It may be of interest to also be able to view non-human and internal visits. To make the raw data useful:

- human visitors **must** be separated from crawlers

The AWSTATS list of crawlers **should** be used to identify known crawlers. Also, any IP address accessing the file '/robots.txt' will be classified as a crawler, as will any IP address that downloads more than X% of the repository holdings in a single day. Further filtering is optional.

The rigour employed in filtering out non-human visitors (robots, spiders, crawlers) will vary between repositories. Some studies have shown that rigorous, manual filtering for crawlers can reduce the number of valid visitors by almost 50 per cent when compared to more standard filtering practices.[1] Filtering rigour depends in part on resource availability, and so it will be necessary to acknowledge that variation between repositories will affect the data.

- double counting **must** be minimised

Visitation data can be over-represented by multiple requests from the same user, which may be a result of accidental double-clicks, use of browser 'back' 'forward' and 'refresh' functions, system crashes, and repeat accesses of material to obtain additional information. To minimise double counting, repeat visits from the same visitor will be excluded where they occur within X time period.

- internal visits (for demonstrations or administration for example) **should** be separated from external visits seeking information from the repository

Some legitimate human visitors are not seeking to obtain information from the repository. These would include accesses that are used to demonstrate the repository, depositors accessing their own material, and visits by repository managers or administrators. In many cases, separating visits from the host domain will be a suitable way of segregating these visitors.

Information access and retrieval

All three user groups are likely to be interested in the aggregated measures of popularity/impact/use, which could be broken down by author, item, subject, or time. It is important to emphasise that the data is not perfect and is not a surrogate for research impact, and to note that repositories and depositors may have an interest in inflating their usage statistics. The approach used in the BEST project is to report on access, views and downloads of information. Any interpretation about the intent or actual use of the information involved is up to the end users. The BEST definitions of *Item*, *Access*, *Retrieve, View* and *Full text* **must** be followed for repository reports.

6. Data Re-Use

[insert something about any agreements that may be desirable from participating repositories/for end users. Philippa Stevens is looking into whether OAK material can be reworked for this purpose]

7. Governance and Compliance

The BEST project is guided by a Reference Group, comprising representatives from the participating repositories. The Reference Group provides input into the approach to and delivery of the project outputs, and ensures that the capabilities and aspirations of their repositories are represented. Membership may change during the project, and at the conclusion of the project period the Reference Group will advise on the next phase of the project.

---

[1] http://www.bepress.com/download_counts.html and http://logec.repec.org/about.htm#stats

The BEST project team, based in the Australian Partnership for Sustainable Repositories, is responsible for coordinating the BEST project, managing project funds, maintaining communication with the BEST Reference Group, and ensuring that the project milestones are met. At the conclusion of the project period the project team will cease to have formal responsibility for the BEST service.

In order to be classed as BEST compliant and have statistics included in the BEST aggregator service, repositories are required to comply with the minimum processing standards outlined in this document, and provide reports that in the format specified by the Standard Statistics Specification. All repositories **must** provide an explicit statement of compliance, which will:

- Outline filtering methodology including the separation of robots and internal visitors, the exclusion of double clicks, and the treatment of internal administration/demonstration activity.
- State any variations in definitions or usage of the terms defined in the BEST Definition of Terms or the other selected controlled vocabularies (for resource type and file format, for example)

This statement of compliance will be made available with the BEST aggregated statistics. The BEST Reference Group will assess these statements to ensure that participating repositories comply with BEST standards.

8. Review of the Guiding Principles

These Guiding Principles will be reviewed by the BEST Reference Group as required, and will be considered at least annually at the first meeting in the calendar year.

Appendix B – List of priority queries for the BEST service

1. Origin of visitors (country, city)
2. Total unique visitors
3. Total items viewed (metadata level) by time, subject, author
4. Total items retrieved (full text) by time, subject, author
5. Top 'n' viewed items (metadata); cumulative and over time
6. Top 'n' retrieved items (full text); cumulative and over time
7. Total items by subject
8. Total items by resource type
9. New additions by time, subject
10. Top 'n' items, authors, documents; cumulative and over time
11. Total items by file format
12. Accessibility

Appendix C – Low priority queries for future consideration

1. Frequency of user logins
2. Total user logins
3. Total new logins
4. Total user logins by user type
5. Total external logins
6. Total number items
7. Total items by collection
8. Most active daily users
9. Total items by depositor
10. Total requests for home page
11. Total local logins