# A Comparative Analysis of Institutional Repository Software

Siddharth Kumar Singh (singh84@purdue.edu), Department of Computer Science, Purdue University
Michael Witt (mwitt@purdue.edu), Purdue University Libraries
Dorothea Salo (dsalo@library.wisc.edu), University of Wisconsin-Madison Libraries

## Introduction

This proposal outlines the design of a comparative analysis of the four institutional repository software packages that were represented at the 4th International Conference on Open Repositories held in 2009 in Atlanta, Georgia: EPrints, DSpace, Fedora and Zentity [1]. The study includes 23 qualitative and quantitative measures taken from default installations of the four repositories on a benchmark machine with a predefined base collection. The repositories are also being assessed on the execution of four common workflows: consume, submit, accept, and batch. A panel of external reviewers provided feedback on the design of the study and its evaluative criteria, and input is currently being solicited from the developer and user communities of each repository in order to refine the criteria, measures, data collection methods, and analyses. The aim is to produce a holistic evaluation that will describe the state of the art in repository software packages in a comparative manner, similar in approach to Consumer Reports [2]. The output of this study will be highly useful for repository developers, repository managers, and especially those who are selecting a repository for the first time. As members of these respective communities and the organizations who support them are increasingly collaborating (e.g, DuraSpace), this study will help identify the relative strengths and weaknesses of each repository to inform the "best-of-breed" in future solutions that may be developed. The study's methods will be presented in a transparent manner with documentation to support their reproducibility by a third party.

## Related Work

Surveys of institutional repository deployment by Joan Lippincott and Cliff Lynch [3] and Gerard van Westrienen [4] were conducted as early as 2005 in the United States and 12 other countries, which were followed up in 2006 by Charles W. Bailey, Jr., for the Association of Research Libraries [5]. These sought to characterize the current state of institutional repository deployment and operation at the time. With support from the Joint Information Systems Committee (JISC) and other agencies, the United Kingdom has taken a leadership role in fostering the growth and development of institutional repositories. Resources such as the Repository Support Project [6] and the Institutional Repository Infrastructure wiki [7] provide a supporting context for this study and helped to formulate it. The scalability and performance of repositories has been explored using a community-based approach for Fedora [8] and in controlled experiments for DSpace by Misra et. al [9] and Lewis [10]. Lastly, analysis performed by the Sheridan Libraries at Johns Hopkins University to connect user requirements to repository functionality [11] informed our selection of the four, basic workflows (consume, submit, accept, and batch) to analyze and supported our decision to maintain a high-level, holistic focus in this study.

## Benchmark

The study was performed on a Dell Optiplex 755 personal computer with an Intel core 2 duo 2.66 GHz processor, 4 GB of memory, 145 GB of SATA primary hard drive, and an on-board single gigabit network adapter. The hardware specification came from the investigators' extrapolation of what machine is likely to be considered "current and typical" based on the equipment survey in ARL SPEC Kit 292 [11] and

hardware that was readily available to the investigators. Resource consumption footprints along with system requirements from the repository documentation were also considered.

EPrints, DSpace, Fedora were installed on the (Red Hat) Fedora 12 operating system and Zentity on Windows Server 2008. The latest stable release versions of the software were used, which at the time of the study were EPrints 3.1.3, DSpace 1.5.2, Fedora 3.3, and Zentity 1.0. Every step of each installation was recorded, and only the core distribution of the repositories were installed with their default settings.

In order to determine a base collections of objects and metadata that are typical of institutional repositories, the Directory of Open Access Repositories (OpenDOAR) was used as a population frame. OpenDOAR contains information about over 1,500 academic institutional repositories [12]. Its API was queried to return a list of OAI-PMH base URLs for these repositories, and 100 metadata records were harvested randomly from an English-language subset (920) of them. The content described by these records was used to characterize and populate a base collection that was then employed for testing and measurement of criteria that require a collection.

## Evaluative Criteria

These criteria were determined by the investigators with input from an external panel of reviewers, and they will be further refined by feedback from the respective repository developer and user communities. Each criterion includes a label, unit of measurement, method of data collection, analysis, and qualification. Measures are not comprehensive and only provide a good-faith indication of their criteria. When available, quantitative data are favored over qualitative data. In most cases, it is not possible or logical to compare repositories as apples-to-apples; for example, Fedora does not provide a user interface. For this reason, all criteria are presented in a context that qualifies the analysis. An abbreviated list of criteria currently being considered:

| | |
|---|---|
| Adoption | Number of running installations? How many downloads? |
| Maturity | Number and frequency of software releases? Duration of existence? |
| Support | What kind of support channels are available? |
| Installation | How easy is it to install the repository? |
| System Requirements | What hardware resources are needed to run the repository? |
| Globalization | Does the repository support multiple languages? |
| Platform Support | Can the repository run on different operating system platforms? |
| Scalability | How many objects does the repository support? |
| Authentication | How many authentication mechanisms are supported (e.g., LDAP)? |
| Access Control | How is authorization supported to selectively limit access? |
| Metadata Standards | What metadata are natively supported? |
| Plugins and Scripts | What 3rd party plugins are available to extend the functionality? |
| Object Format Support | What different object formats are supported? |
| Database Support | What databases can be used with a repository? |
| Storage Abstraction | What means for storing data are offered? |
| Sustainability | How is the repository project sustained into the future? |
| Interoperability | What standards are supported to enable the repository to integrate with other systems? |

| | |
|---|---|
| Developer Ecosystem | What tools and support exist for developers? Is the software easily extensible and programmable? |
| Search Engine Optimization | How well is repository content exposed to Internet search engines? |
| Upgrade | How easy and reliable are repository upgrades to perform? |
| Search | What is the accuracy and response time for the execution of a search? What query language is supported? |
| Performance | What is the bottleneck of the system? What is the response time for the application when an object is inserted/updated/retrieved/deleted while the system is under a typical user-load? |
| Migration | How easy is it to migrate content to another repository? |

An example of our preliminary data collection for one criterion, adoption:

| Repository | Number of Installations | Number of Downloads |
|---|---|---|
| DSpace | 700 [a], 512 [d], 511 [e] | 146,984 [f] |
| Fedora | 172 [b], 26 [d], 73 [e] | 72,239 [f] |
| EPrints | 269 [c], 307 [d], 263 [e] | Not available |
| Zentity | Not available | Not available |

Sources:

a) http://www.dspace.org
b) http://www.fedora.org
c) http://www.eprints.org
d) http://roar.eprints.org
e) http://www.opendoar.org
f) http://www.sourceforge.net

## Workflow Analysis

The repositories will also be assessed on the execution of four common workflows:

| | |
|---|---|
| Consume | Quality navigation, browse, and search? Ease of finding and downloading content files? Usage data available? Email and RSS notifications of new deposits? Commenting, embedding, and other Web 2.0 functions? |
| Submit | Ease of signup? Usability of deposit-form ordering and layout? Appropriateness of displayed fields to content types? Auto-completion of fields? Flexibility to add, remove, or change fields on forms? Ease of error correction during and after deposit? |
| Accept | What quality-control steps are available, and to whom? What can be changed after deposit, and by whom? How good are notifications? Can items become stuck in a workflow? If a deposit is rejected, can the depositor edit and resubmit, or must the depositor start over from nothing? |

| | |
|---|---|
| Batch | What server privileges are necessary to perform a batch import? How complex is the batch-import format? How complex is the batch-import command invocation? |

## Current Status

Based on the definitions of the criteria that incorporated feedback from the external reviewers, preliminary data collection and analysis have been performed for 15 of the 23 criteria. These data are considered to be subject to change until feedback from the developer and user communities can be gathered and incorporated. Each criterion and its unit of measurement, method for data collection, and analysis has been posted as a thread in a public blog that has been advertised to the respective communities as a Request For Comments [13]. Commenting has been enabled on the blog, and comments will be gathered until April 2010, when the RFC period concludes. At this time, the criteria may be modified or new criteria added based on community feedback. The remaining data collection and analysis will be conducted in May and June; preliminary findings will be posted to the blog as the study progresses. Summary findings from the study will be ready to be reported at the conference in Madrid.

## Acknowledgements

Principals from each of the repository organizations reviewed the study design and the initial set of criteria and contributed feedback: Thorny Staples (Fedora), Tim Donahue (DSpace), Patrick McSweeney (EPrints), and Lee Dirks and Alex Wade (Zentity). Stuart Lewis (University of Auckland), Mark Newton (Purdue University), and Bill Helling (IUPUI) provided additional comments. Luo Si (Purdue University) provided expert advice on the search criterion.

## Notes

[1] The 4th International Conference on Open Repositories website, https://or09.library.gatech.edu.
[2] Consumer Reports magazine website, http://www.consumerreports.org.
[3] Lynch, Clifford A. & Lippincott, Joan K. (2005). Institutional Repository Deployment in the United States as of Early 2005. *D-Lib Magazine, 11*(9).
[4] van Westrienen, Gerard & Lynch, Clifford A. (2005). Academic Institutional Repositories: Deployment Status in 13 Nations as of Mid 2005. *D-Lib Magazine, 11*(9).
[5] Bailey, Charles W. wt al. (2006). SPEC Kit 292: Institutional Repositories. Association of Research Libraries.
[6] Repositories Support Project website, http://www.rsp.ac.uk.
[7] International Repositories Infrastructure wiki, http://repinf.pbworks.com.
[8] Fedora Performance and Scalability wiki, http://fedora.fiz-karlsruhe.de/docs.
[9] Misra, Dharitri, Seamans, James, & Thoma, George R. (2008). Testing the Scalability of a DSpace-Based Archive. *Proceedings of IS&T Archiving 2008.* Bern, Switzerland.
[10] Robot-generated Open Access Data website, http://www.jisc.ac.uk/whatwedo/programmes/reppres/tools/road.aspx.
[11] Choudhury, Sayeed. (2006). A Technology Analysis of Repositories and Services. Report to the Mellon Foundation, http://msc.mellon.org/research-reports/A%20Technology%20Analysis%20of%20Repositories%20and%20Services.pdf/view.
[12] The Directory of Open Access Repositories website, http://opendoar.org.
[13] Request For Comments: A Comparative Analysis of Institutional Repository Software, http://blogs.lib.purdue.edu/rep.