DATA-DRIVEN IMPROVEMENTS TO IR ACCESS AND VISIBILITY

Statement of National Need:

Reported activity data for institutional repositories (IR) are inaccurate and cannot be compared across repositories or time. Our research shows that existing analytics reports either significantly over-count or drastically undercount IR activities, such as file downloads. The Montana State University (MSU) Library seeks a $50,000 Planning Grant in the National Digital Platform project category to support a collaborative project with DuraSpace and MSU computer scientists to improve accuracy, comparability, and reporting capacity of IR analytics to better demonstrate value to stakeholders.

Log file analysis tools grossly over-report activity due to excessive robot traffic that is difficult to identify and exclude. Conversely, page tagging analytics services, e.g. Google Analytics, under-report activity by more than 50%, on average. Citable content downloads by humans is arguably the most important IR metric, and 93%-100% goes unreported when only Google Analytics is used.

Our research team developed the Repository Analytics and Metrics Portal (RAMP)[1], a free service that has attracted 35 subscribers in its first year of operation and has proven successful in providing more accurate analytics. RAMP can serve as a diagnostic and reporting tool for any repository platform, while also compiling a large dataset that offers exciting analysis and research opportunities.

RAMP is similar in principle to the IRUS-UK[2] project funded by JISC, but instead of analyzing log files it is based on Web 3.0 technology and leverages Google services to reduce costs and improve accuracy. Preliminary analysis of RAMP data has identified that downloads per IR item can vary by more than 300% across repositories. Based on this analysis, it appears that much of the IR under-utilization is due to basic Search Engine Optimization (SEO) deficiencies.

Libraries and their funders have invested an enormous amount of capital in IR since the early 21st century, but IR suffer from several deficiencies and many now call their value into question; however, our research suggests that IR *do* offer significant value when these deficiencies, such as SEO, are addressed.

In alignment with the National Digital Platform's aim of "expand[ing] the digital capacity and capability of libraries and archives," we seek a planning grant to develop a project work plan scoped around determining high-impact solutions to addressing IR deficiencies and improving RAMP's capacity to measure and assess the value and impact of IR.

We envision this work plan leading to the development of RAMP reporting tools that will give repository managers the following insights and capabilities:

1. The behavior of users as they seek scholarly information;
2. Accessibility of open access academic publications and whether there is duplication or gaps between what users are searching and what they're finding;
3. Visibility of IR content in search engines and whether it varies across organization or platform;
4. Whether repository metadata are aligned with user search terms; and
5. The ability to cost-effectively generate item metadata that improves knowledge discovery and reporting comparability across all IR.

[1] http://ramp.montana.edu – access with "demo"
[2] http://irus.mimas.ac.uk/

Proposed Work Plan:

We propose that IR activity reporting requires both technical remediation to address SEO deficiencies and metadata remediation that improves IR content visibility once the repositories are successfully harvested and indexed. The work plan proposes these topics and actions:

1. Management and reporting:
    a. Develop plan for RAMP to transition to DuraSpace management;
    b. Collect and prioritize IR manager requirements for standardized IR reporting;
    c. Develop project plan for incorporating SHARE metadata into RAMP analytics; and
    d. Develop project plan for an API that exposes RAMP data to researchers.
2. SEO auditing, monitoring and remediation plan:
    a. Design and pilot an IR SEO Audit on the top 3 and bottom IR performers; and
    b. Assess the importance of each IR SEO area for user search visibility and access.
3. Develop a research plan to apply advances in computer science that improve the quality, consistency, and usability of IR content metadata across IR:
    a. Conduct a pilot study predicting LCSH terms, or ORCID IDs, for a given item using supervised learning models;
    b. Conduct a pilot study that combines SHARE metadata with RAMP analytics data to evaluate the importance of each metadata field in user search visibility and access; and
    c. Develop a feasibility study and research plan for applying machine learning and graph theory algorithms to reduce the effort invested in creating consistently optimized and comparable IR content metadata. This could also help metadata consistency in SHARE.

Outcomes from this work plan: a sustainability plan for RAMP; improved IR reporting model; SEO audit plan for IR; a data-driven recommendation on the optimal amount and type of metadata per item; and a better idea of whether machine learning can improve the quality of IR content metadata.

Projected National Impact:

Our proposed work plan's IR reporting model, SEO audit, and recommendation on the optimal amount and type of metadata per item could help any IR whose managers seek to demonstrate their impact.

Budget Summary:

The investigators anticipate that $50,000 will be required to accomplish our proposed goals and outcomes. During the one-year grant period, funds will be allocated towards research personnel salary and benefits support, cloud computing services, intern salary and benefits, DuraSpace's contributions, and institutional indirect costs. An itemized budget for these costs is below:

Category	Item	Cost (USD)
Salary Support:	Patrick OBrien, Semantic Web Research Director	$20,563
Benefits:	Estimated at 37% of Salary Support	$7,608
Equipment:	Cloud Computing	$1,000
Student Support:	Internship for M.S. Candidate in Computer Science	$4,040
Subcontracts:	DuraSpace	$5,000
Indirect Costs:	F&A at 34.5% of Total Direct Costs	$11,789
Total Costs:		**$50,000**