

# The bX project: Federating and mining usage logs from linking servers

Johan Bollen (LANL), Oren Beit-Arie (Ex Libris) and Herbert Van de Sompel (LANL).

**SUMMARY:** The bX project aims at unleashing the power of usage information that is recorded on an ongoing basis by OpenURL-compliant linking servers. The project is a collaboration between the Digital Library Research & Prototyping Team of the Los Alamos National Laboratory and Ex Libris.

Interesting things can be done with usage information that has been recorded for a specific information source from a digital library. The starting point of the bX project is that even more interesting things can be done with usage information that has been recorded by a linking server, because such a server can record activities across multiple OpenURL-enabled information sources of a specific digital library environment. As such, logs from a linking server are highly representative of the activities and preferences of a user population that requests services from that specific linking server. The bX project further recognizes that even more interesting things can be done with usage logs that are federated across multiple linking servers. Indeed, as a federation of linking servers grows in size, the federated usage log database becomes increasingly representative of the activities of the global scholarly user base.

**RATIONALE:** The evaluation of science is still largely based on citation and authorship data which originates in a paper-based process of communication, i.e. journals publishing a limited selection of approved articles. This model, exemplified by the Institute for Scientific Information's Impact Factors and Journal Citation Reports, is rapidly being made obsolete by advances in the evaluation of web resources which favor large-scale usage over expert opinion and structural metrics over voting procedures. Usage data combined with structural Google-type metrics of quality forms a powerful combo for the evaluation of scholarly communication items in a future where the predominant model of scholarly communication will be user-driven, decentralized and focused on a wide variety of item types, e.g. data sets, software and educational resources in addition to journal articles.

Fig. 1. shows a taxonomy which classifies evaluation techniques on the basis of whether the data they use is authored, e.g. citation and hyperlinks, versus user-generated, e.g. ratings and purchase patterns, and whether the metrics used are frequentist, e.g. citation counts, versus structural, e.g. Google's PageRank. The taxonomy shows how the present evaluation of scholarly communication items remains confined to the use of frequentist metrics applied to citation data, i.e. a publication's popularity among a group of author experts as indicated as its citation frequency, whereas most advances relevant to emerging publishing and research models have been made in the other 3 quadrants.

**APPROACH:** The bX project has developed and implemented an architecture which allows the large-scale aggregation of logs generated by openURL-enabled linking servers and their subsequent analysis for the evaluation of scholarly communication items. Fig. 2 and 3 show more details.

The architecture functions according to the following three phases:

- 1) Linking server logs are serialized as XML-ized OpenURL ContextObjects and exposed by an OAI-PMH repository. The repository retains full control of *what* is exposed and *how*, e.g. anonymization of user IDs. A trusted third-party (or federation thereof) can harvest and aggregate logs from a range of repositories thereby creating a usage data set representative of a particular community which in principle could extend to the entire scholarly community.
- 2) Next the aggregated logs are subjected to datamining techniques which derive item networks from access sequences recorded in the logs under the assumption that similar items are accessed by similar users. These networks can be used to construct recommender services useful to both local institutions as well as third-party aggregators.
- 3) As a last step, structural metrics of quality can be derived from the generated item networks leading to more complete, fine-grained and reliable evaluation of scholarly communication. Log data furthermore is free from publication delays and can be used to track immediately contemporary trends in science.

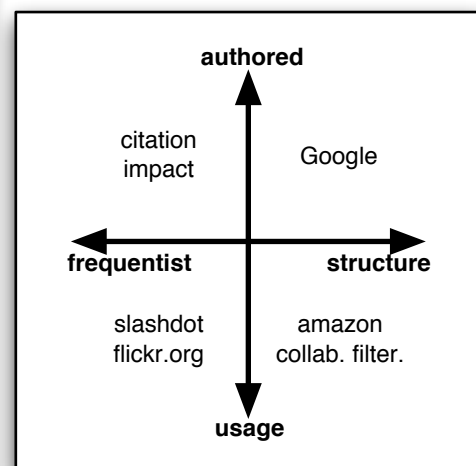


Fig. 1: Taxonomy of resource evaluation