Towards a Comprehensive and Up-To-Date Institutional Repository: Development of a Publications Tracking Process

Mohamed Baessa, King Abdullah University of Science and Technology (KAUST), mohamed.baessa@kaust.edu.sa
Daryl Grenz, King Abdullah University of Science and Technology (KAUST), daryl.grenz@kaust.edu.sa
Han Wang, King Abdullah University of Science and Technology (KAUST), han.wang@kaust.edu.sa

Session Type
✓ Presentation

Abstract
As our repository matured we found that a concerted effort was needed to establish it as a resource that institutional stakeholders knew they could turn to for reliable information about the research outputs produced by researchers affiliated with or funded by our institution. The initial impetus for this project was the adoption of an institutional open access policy, which necessitated a process to track the publications of institutional authors, deposit items with applicable open access licenses, and notify authors when an accepted manuscript was needed. To support this process we developed a local publications tracking system independent of our hosted DSpace repository that regularly queries publisher and indexer APIs for new publications, checks for relevant permissions policies, identifies institutional authors and helps us send manuscript request emails. We also harvested records for past items and modified our repository to accommodate metadata-only records. The success of these tracking and harvesting services in making our repository comprehensive and up-to-date has allowed us to rely on our repository as the key source of publications information for additional integrations that update ORCID records with publication information, populate a PlumX metrics dashboard and, most recently, support the implementation of a current research information system (CRIS).

Conference Themes
✓ Supporting Open Scholarship, Open Data, and Open Science
✓ Integrating with the Wider Web and External Systems

Keywords
publications tracking, external integrations, OA policy compliance, institutional repositories, current research information systems
Audience
Repository and CRIS managers and developers interested in publications tracking.

Background
Establishing an institutional repository as a comprehensive and up-to-date resource for information about a university’s research outputs supports the repository’s ability to provide open access to research and integrate with add-on services that are of value to researchers and other university stakeholders.

Presentation content
The landscape of potential sources for information about the publication of research outputs is vast, ranging from publishers and indexing services to search engines and metadata aggregators. The quality and currency of the metadata available is also quite varied.

While it is possible to rely on major indexing services (such as Scopus or Web of Science) for information about a large proportion of a university’s research output, they are not fully comprehensive. Furthermore, information about new publications becomes available in these services at unpredictable intervals, often several months after online publication of an article or conference paper. This presents an obstacle to the deposit of accepted manuscripts in compliance with open access policies, as researchers are more responsive to manuscript requests close to the date of publication, and also impedes the provision of accurate reports and of services that help researchers automatically maintain online profiles.

More immediate information about new publications can be obtained from individual publisher APIs (each with their own distinct query methods and response formats), the Crossref API, Google Scholar (with wide scope, but lack of a formal API and standard response metadata), Pubmed/PMC, and from emerging repository-oriented services like SHARE. At the same time, major repository systems do not include processes for harvesting publications information from indexing or publisher services, often only supporting OAI-PMH based harvesting from other repositories or from services like the JISC Publications Router. This means that a repository needs to rely on either a CRIS or a standalone harvesting system to interact with most major external metadata sources.

In our case, the decision to be a more comprehensive resource for publication information came only gradually as we began implementation of an institutional open access policy, starting with a process to track and harvest publications that had been assigned Creative Commons licenses and also to request accepted manuscripts from authors when needed, and realized that there was no simple method of achieving either of these goals without first having a reliable list of all publications. As we experimented with the use of feeds, notifications and manual exports from
several sources we slowly moved towards using APIs when available so as to partially automate update and deduplication processes.

On the repository side, this process also prompted minor modifications to support the addition of metadata-only records in a way that helps users to distinguish them from those with deposited files. Overall we were largely able to rely on existing features of DSpace (such as CSV import of metadata-only records) in our process. However, as we began to launch additional services based on our more complete record of institutional publication, such as ORCID integration and a PlumX metrics dashboard, we found that more questions arose about what it meant to offer a complete picture of our university’s research.

Two additional pieces of our tracking system were born out these inquiries, one to track patent publication events and the other to identify publications with non-affiliated authors which acknowledge KAUST grant funding support. This latter project is somewhat unique to our institution as the existence of a funding body within the university means that a significant proportion of all publications associated with KAUST are products of work by external authors with KAUST funding. While Web of Science, Scopus and Crossref/Fundref all provide some support for tracking based on funding acknowledgement, no single source was able to fully meet our needs in this area, again leading us to attempt to combine sources in a quest for comprehensiveness. This quest will clearly continue as we also consider whether to track public releases of software and research data.

We initially considered our publications tracking project to be a stop-gap measure as our institution went through a lengthy process of selecting and moving towards implementation of a CRIS system with automated harvesting tools, only putting in the minimum amount of time and effort needed to have a functioning process. Over time however we realized that our publications tracking system had evolved with several novel features, such as integral use of ORCID IDs and the ability to identify funding acknowledgments, that we would like to retain and strengthen, leading us to begin restructuring the system as an application that we can share publicly in the hopes that it will prove useful to other institutions working to make their repositories more comprehensive and up-to-date.

**Conclusion**

Before a repository can serve as a meaningful store of publications metadata for use in tracking compliance with open access policies or integrating with other research reporting services, it should first establish a process to track research release events in external systems. Such publications harvesting is not typically a feature in institutional repository software, but is common in current research information systems. In lieu of a full-feature CRIS, a repository can benefit from having a standalone publications tracking system.