Using the IR as a Research Data Registry

Daryl Grenz, University Library, KAUST, Thuwal, Saudi Arabia, daryl.grenz@kaust.edu.sa; Eirini Mastoraki, University Library, KAUST, Thuwal, Saudi Arabia, eirini.mastoraki@kaust.edu.sa; Han Wang, University Library, KAUST, Thuwal, Saudi Arabia, han.wang@kaust.edu.sa; Mohamed Ba-essa, University Library, KAUST, Thuwal, Saudi Arabia, mohamed.baessa@kaust.edu.sa

Session Type

- Presentation

Abstract

As data and software become increasingly common research outputs, universities have an opportunity to expand their existing efforts to record affiliated publications so that they also capture information about research data releases. At KAUST we have taken several steps to put our repository on a path towards becoming a reliable registry for information about the existence and location of research data released by affiliated researchers. These included developing a process to retrospectively retrieve and register information about datasets with machine-readable relationships to publications already in the repository, and updates to our active publications tracking procedures so that data availability statements are retrieved at the time of harvesting and checked for references to research data. The presentation will conclude by discussing how these efforts help put the repository in a position to provide expanded services in support of improved research data management, including access to and preservation of research data not explicitly linked to a formal publication.

Conference Themes Addressed

- Content - research data, digital preservation, persistent urls, archiving
- Infrastructure/Integrations - integrations between systems, changing technical environments
- Reuse, standards, and reproducibility - for example: software, data, content types
- New open technologies and standards

Keywords

Institutional research data registry, dataset tracking, data-publication linking, data availability, data citation

Audience

repository managers, developers, librarians.

Background

At OR2016 we presented about efforts to make our institutional repository a comprehensive and current registry for information about research published by affiliated researchers. As the variety of avenues for public release of research continues to expand beyond traditional publication, universities have a similar interest in staying aware of where and how researchers are disseminating research data and code. Furthermore, this represents an opportunity to start developing services that interact with researchers at earlier stages in the research life cycle, not only at the point of final publication. In order to justify the resources needed to sustain and grow repositories at the institutional level they need to support both administrative reporting and active research, this work proposes using the repository as a dataset registry in order to provide a basis for new or improved services in both areas.
First, we developed a process to identify datasets deposited in external repositories with relationships to already deposited publications by querying the Data Literature Interlinking Service and Datacite with publication ids such as DOIs and PubMed IDs. We then developed and applied criteria for reviewing the retrieved records to clarify the type of relationship that they had to the publication and determine the suitability of their inclusion in our repository, before generating metadata records (including all of the elements necessary for data citation along with additional fields available in the retrieved Datacite224 records) for the datasets. We then loaded the records to the repository, as well as updates to the existing publication records so that their relationship to the dataset record would be displayed.

Second, we included steps in our publications tracking process to check for references to data being made available separately in support of the publication. We particularly looked for data and code references in the form of DOIs, accession numbers, and Github repository links, while also recording information about any other URLs that were mentioned for the code or data discussed in the paper. Two major issue areas that came up repeatedly and require additional attention were the treatment of supplemental data released via the journal publisher site without being assigned unique identifiers distinct from the publication DOI, and of data and code released via institutional lab and research group websites.

The presentation will discuss these particular projects within the context of related activities including discussions with faculty about the types of data produced in their research and their current practices in managing it, and continued improvements to our repository integration with our university’s current research information system, Pure, that included export of dataset records and their relationships to publications from the repository to Pure. The faculty consultations in particular have helped us plan for follow on projects by giving us an early gauge of their interest in registering DOIs for the data they release and in providing for archiving of their data through institutional or external services.

Conclusion

Just as an institutional repository can function as a source for authoritative information about publications by affiliated researchers, it can also provide a valuable service as a registry for research data that supports those publications, or that is otherwise released by affiliated researchers. For institutions looking to take proactive steps toward better management of research data, implementing a process to monitor existing research data dissemination practices may be a good first step.

Repository System

Is your submission related to a specific repository technology? Please select one technology by deleting those not applicable or add another technology not listed below.

- DSpace

References


