Title: Developing a model approach for community led CRIS aggregations

Type: Business/technical paper

Authors: James Toon[1], Alexander Kujath[1], Eli Khazzam[2]

1 Elsevier, Amsterdam, Netherlands

2 New Jersey Economic Development Authority, New Jersey, United States

Keywords

Research Information Management, Community repositories, CRIS Aggregation, Deduplication, Harvesting, Web Services, Pure

Summary

This abstract is intended to provide information on how we have developed an easily deployed and scalable CRIS aggregation service intended to provide value added services for research communities, based on research information sourced via CRIS systems and/or repositories used by participating institutions. The goal was to build a common, inter-institutional information source based on the Pure data model. The proposed presentation will describe the technical approach taken and outline some core challenges encountered. The paper and presentation will also invite discussion around future opportunities made possible because of the new aggregation service.

Background

Repository aggregations work best where they have a singular purpose in mind (such as for a specific subject community or special interest area) and where users contributing to the aggregation feel there is value to be gained from actively participation – i.e. that it is possible to demonstrate high visibility of relevant content of high quality, provide useful services to those persons who need to find and use it, and provide active feedback to contributors that their content is being found (and ideally used) by the right people.

Demand has increased in recent years for such aggregation services that are keen to take advantage of the breadth of research information being captured by organisations through use of CRIS systems. Examples of these use cases include organisations wishing to foster collaboration between universities and industry, or where federated institutions wish to provide a singular view on research activity across a region. Other use cases may also include funder-based aggregation of research outcomes to demonstrate return on their investments. As part of our presentation, we will provide a specific example of how we have met such a demand through use of an implementation case study, the New Jersey Economic Development Authority[1].

We have developed an approach, based on Pure, that allows community managers to implement a CRIS aggregation service from across multiple systems. The solution provides a view of data available for showcasing on public portals and reporting or benchmarking on behalf of the community. The solution

[1] New Jersey Economic Development Authority http://www.njeda.com/

can be quickly deployed and easily maintained, allowing owners to focus efforts on data curation and acquisition to improve the value of the services they are providing.

Although the developed solution is as automated as possible to reduce administrative burden, the aggregation cannot be entirely passive and needs to be actively managed by a community owner, who has an incentive to ensure that the data in the repository is well formed and curated to best service the mission of the repository. Examples of active curation include adding community specific subject classifications not present (or even relevant) in the source repositories, building relations between content types and supporting scholarly communications.

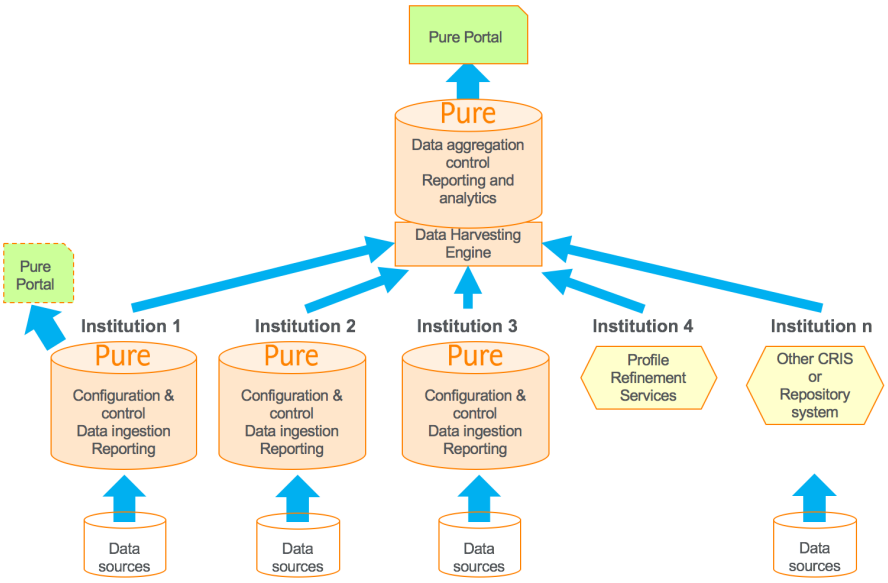
Solution

Our technical implementation is based on the principle that community owners wish to deliver enhanced value to their users whilst limiting the administrative burden on the participating research institutions. Further focus is made on creating a solution that can be scaled upwards or downwards to meet with the needs of the community I.e. as participating members join/leave.

To achieve this, we decided to use Pure as the core of what is known as the Pure Community Module. Using Pure allows us to utilize the robust underlaying data model to capture and maintain information needed for the community use case. We focused on the design of a dedicated harvesting service that allows us to collect information needed from multiple data sources and that, once aggregated, can be made available for further curation and then used for showcasing of experts and their outputs, but also to allow for centralised reporting and benchmarking.

Harvesting of data can be via multiple methods, including the Pure web service (where participating institutions are existing Pure users), via the Elsevier Profile Refinement Service or imported via the Pure XML schema in the case where participating institutions are not Pure users. The harvesting service is set up to query each individual participating research institution and constantly retrieve updates. These updates are aggregated in the Community Pure instance and made available for the individual use case of the community.

Figure 1. Community Module architecture



Key Challenges

When information is aggregated from different sources, a common issue is how to deal with duplicate content. To resolve this, we use an automatic deduplication mechanism that takes advantage of identifiers to process content. Duplicates are typically encountered when publications are co-authored between the research institutions of the community, so there is an individual record in two or more of system that are harvested for the same publications.

Where possible, we use the Scopus ID as the primary identifier in the deduplication process. this allows us to uniquely identify records and conduct a merge operation to reflect this in the Community instance accordingly.

For records without a Scopus ID available, we operate with a merge strategy using other unique identifiers where available, like DOIs. In case a possible duplicate cannot be identified without doubt, we have decided to use a conservative merging strategy to avoid false positive merging operations. Here a manual curation can take place, if needed or wanted. An operator of the Community Module can resolve these cases manually using the Pure built in deduplication processes.

Future Directions

Now the core processes for data harvesting and deduplication are in place and the Community Module is being rolled out for customers, we are starting to work on our roadmap for future development of the Community Module. Examples of opportunities are as follows:

Additional content families – The Community Module currently focuses on organisation, person and publication content families, however the Pure data model can accommodate many other types including project/grant data, press and media, datasets, equipment, knowledge exchange or impact.

Concept matching via fingerprinting technology – Pure makes extensive use of the Elsevier Fingerprint engine to derive subject concepts based on data contained in publication abstracts, award data, equipment descriptions, research interest statements etc. This ability to create structure from unstructured data allows users to identify potential collaborators and experts. Whilst this is already available as a feature in the Community Module and Pure Portal, we aim to extend this further by adding enhanced matching to other content families.

Third Party Integration - Additional third-party integrations can be used to further build aggregated content in the Community Module. For instance, press/media content could easily be added through continuous integration with the Newsflo service, or metrics data could be integrated via multiple sources (bibliometric and/or alternative metrics). With integrations like these, the reach and significance of the community's research activity can be investigated and used to optimise marketing and outreach strategies.

Re-distribution of enhanced content to participating CRIS instances – Where content has been developed and/or enhanced at the level of the community aggregation, there is the opportunity to feed these enhancements back into the participating institutions – so for example, publications identified in one participating member could be used to help populate a collaborating member institution who does not have the record in their CRIS.

Additional community use cases – Functionality could be developed as part of the Community Module to meet the needs of specific use cases, such as may be required by a research funder, government agencies or industrial sector.

Conclusion

Through the development of the Pure Community Module, we have shown how data from multiple sources, like CRIS systems, repositories and other data sources can be readily aggregated to provide value for community/sector groups and for institutions when seeking to showcase expertise and develop new collaborative opportunities.

Our goal was to create an easily deployed system that will enable communities to establish platforms used to curate, promote and showcase content based on a given need. We believe that the solution achieves this, whilst dealing with a variety of common issues surrounding aggregations, including harvesting of multiple data sources, problems of deduplication and name disambiguation. Our paper and presentation will provide a technical overview of how this was achieved, provide a case study of the Community Module in practice, and raise discussion around the opportunities and next steps for developing the module further.