



13th International Conference on Current Research Information Systems, CRIS2016, 9-11 June
2016, Scotland, UK

A CRIS in the Desert: The Implementation of Pure at KAUST A Case Study in Information Exchange

Daryl Grenz^{a*}, Thibaut Lery^b, Manus Ward^b, Eirini Mastoraki^b, Mohamed Baessa^a

^aKing Abdullah University of Science and Technology (KAUST), University Library, Thuwal, Saudi Arabia

^bKing Abdullah University of Science and Technology (KAUST), Office of Sponsored Research, Thuwal, Saudi Arabia

Abstract

The integration of research information systems with existing university processes has tended towards information exchange models in which the CRIS ingests information from existing systems and takes on functions that were previously distributed across several independent solutions. This paper draws upon the experience of the implementation of a CRIS at the King Abdullah University of Science and Technology (KAUST) to posit a model in which functions remain distributed so as to take advantage of the strengths of each system. The functions discussed include institutional reporting, publications tracking, preservation of research outputs, provision of public access, researcher identity and profiling, and metrics analysis. The systems reviewed include a CRIS (Pure), a locally developed publications tracking system, a hosted DSpace repository, a locally developed ORCID integration, and a metrics dashboard (PlumX). The interactions between these systems forms a network of services to our research community, with each node connected to several others, and we discuss how we arrived at the current arrangement, as well as its drawbacks and advantages. The still limited use of standard data exchange formats like CERIF XML is discussed as a constraint that increases the costs of adding to and maintaining the network of services. At the same time we look at how increased standardization should make this distributed approach sustainable, allowing institutions like ours to mix and match complementary systems to achieve an optimal set of research information services for our needs.

© 2017 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

Peer-review under responsibility of the Organizing Committee of CRIS2016

Keywords: interoperability; distributed model; Pure; publications tracking; institutional repositories; ORCID; PlumX

1. Introduction

As a recently founded research-intensive university, King Abdullah University of Science and Technology (KAUST) has a unique opportunity to build processes for the management of research information that are relatively unencumbered by legacy systems, data and workflows. At the same time, the geographic isolation of the institution, combined with its newness and resulting lack of shared institutional knowledge among stakeholders, presents real

* Corresponding author. E-mail address: daryl.grenz@kaust.edu.sa

challenges in achieving our goals. In this paper we outline our work within this environment to implement processes for managing research information that not only meet the immediate needs of stakeholders in the university, but also support maintenance of system-independent workflows and validated data that will be useful in the long-term.

1.1. Background

Since opening its doors in 2009, KAUST has grown to support approximately 800 masters and doctoral students, 400 postdoctoral fellows, 300 researchers and 150 faculty in an array of scientific and technological disciplines, with a special focus on food, water, energy and the environment. A network of core laboratories provides access to the best in research equipment and facilities, while an active economic development program supports the commercialization of research breakthroughs through startups and collaboration with corporate partners. The ability of the university to set its own research priorities and support them effectively is further enhanced by the baseline funding for each faculty member and the existence of an internal funding body that supports a number of competitive program and project funding schemes.

2. Implementation of Pure

2.1. Initial approach

The Pure CRIS provided by Elsevier is designed as a comprehensive solution that aggregates information about research outputs, incomes, projects, research staff, students, organizational units, external collaborations, grant applications and more, connects the entities in meaningful and complex ways, and then provides methods for visualizing and reporting on these relationships with flexibility and precision. As with any research information system project, an important initial challenge is identifying all of the relevant sources of information both internal and external to the institution, and then looking at what data they provide and how it can be exchanged. For our implementation of Pure we outlined the needed types of information and expected system relationships and then proceeded to approach them one by one, starting with information about people, then about research outputs, and finally about projects and awards.

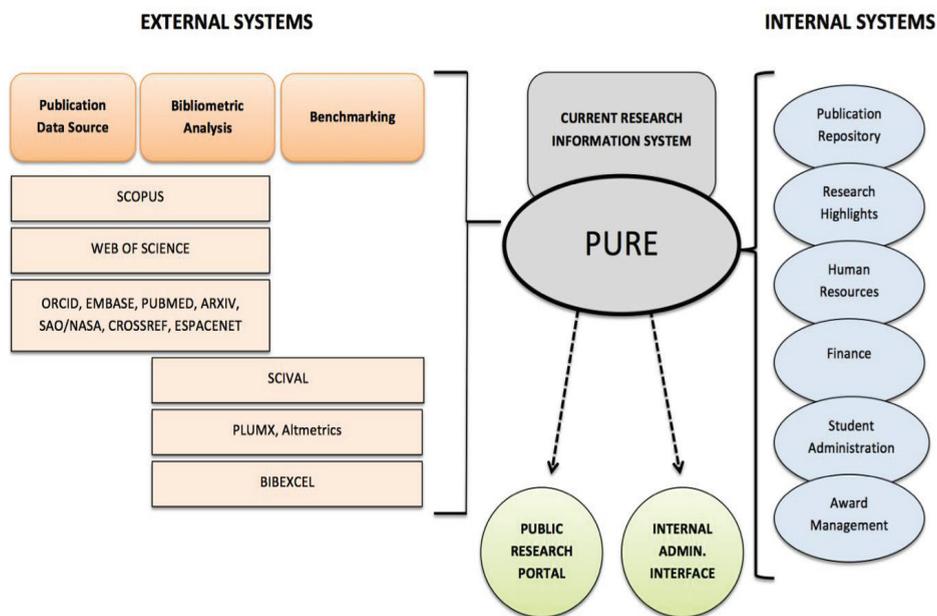


Fig. 1 Overview of systems related to the implementation of Pure at KAUST

2.2. Challenges encountered

As we moved through this process we also tried to assess how the particulars of the information and how it was structured would impact its utility for specific analysis and reporting use cases. For example, with our university's strong focus on global collaboration, we need to be able to evaluate the relative effectiveness of different types of collaboration in producing high-impact research, but reliably identifying collaborating institutions by type and geographic location based on the available information can be challenging.

In some cases the way in which information was transformed as it was ingested into Pure proved to be a complication. For example, as a new institution that is attracting researchers with established track records in research at past institutions, one potential point of reference is their relative productivity at KAUST in comparison to their earlier careers. We initially struggled to get pre-KAUST publication histories imported into Pure in a way that allowed us to do this type of analysis effectively. Tools such as Pure (Elsevier) and Converis (Thomson Reuters) use 'Organization' and 'Person' identifiers as the base data on which all other data is layered. This is perfectly logical for older more established Research Institutes where there is relatively little movement of researchers. However, this has become a major issue at KAUST as all of our Faculty arrived after 2009 with many having long productive research histories prior to their relocation. What we identified early on in the implementation phase of Pure was that an individual's research history and in particular their prior affiliations were being compromised on import and replaced with their current affiliation in KAUST. In many established research institutes this can be equated as "noise", however, in KAUST this noise currently constitutes 60-70% of the total research outputs associated with our faculty. This has raised serious data quality issues that have significantly delayed the implementation of Pure at KAUST as an appropriate solution is identified.

The peculiarities in how information was provided by internal systems also proved to require a great deal of intermediary process development in order to structure and control the information to the point where it was suitable for ingest into Pure in a usable form. The creation of these intermediate processes poses a problem for the sustainability of information transfer from internal systems, requiring ongoing monitoring and correction. Even in situations where a standardized information exchange protocol was expected to be useful, such as in the connection between Pure and KAUST's commercially hosted DSpace repository, a custom solution proved to be necessary. In this context, it is clear that the set-up process for the deployment of Pure had to span over several years to standardize the collection of the relevant data, as shown in Figure 2.

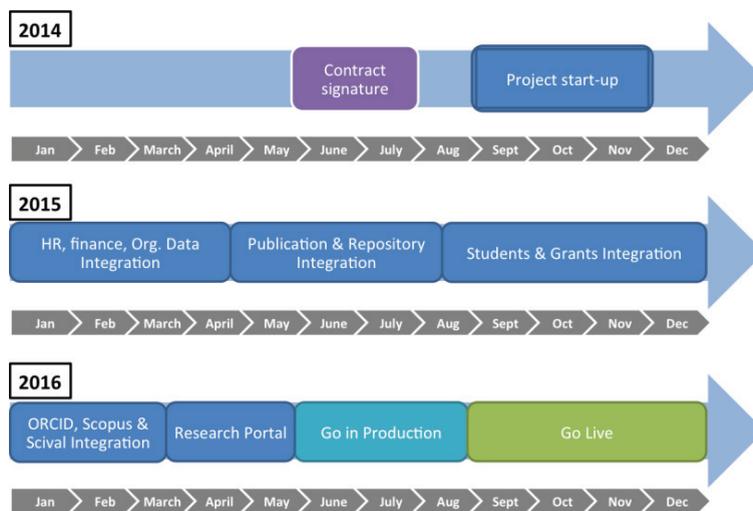


Fig. 2 Timeline for the deployment of Pure at KAUST

Getting broad-based institutional buy-in for the project also proved challenging due to the diversity of global backgrounds held by stakeholders. CRIS systems are becoming more common in European and American research institutes where there is a need to monitor productivity and impact given that budgetary provisions are linked to these metrics. KAUST is in a unique position with its faculty, students and staff coming from over 90 different countries, many of whom are unfamiliar with and do not inherently see the value of an integrated tool for research information management and reporting given our existing institutional access to online tools from Elsevier (Scopus and Scival) and Thomson Reuters (Web of Science) that can fulfill similar needs.

3. Relationships between systems

3.1. Current arrangement

Over the course of the project we have moved towards accepting a more fluid arrangement of internal systems than originally intended, as shown in Figure 3. While the specifics of the relationships between systems are still evolving, we have clearly started to realize benefits from the process of beginning to implement a coordinated institutional approach to research information. For example, the work put into organizing information about the people at KAUST from human resources and student administration systems in preparation for use in Pure paid early dividends when the improved information was fed into the Open Access policy compliance tracking and ORCID creation initiatives. Building upon that success we were then able to use ORCID identifiers and improved publications information in our institutional repository as the basis for adding the PlumX metrics portal to our network of systems.[1] PlumX in turn has provided an opportunity for us to look at adding more connections to other elements of the university’s online presence. We have started to do this by having stories on the KAUST Discovery research highlights website treated as blog mentions for related publications in PlumX. Working with systems such as PlumX to provide a richer understanding of how our research outputs are shared and reused has also contributed to our decision to emphasize greater modularity in our systems approach, so that we place less emphasis on moving towards having a single system that will optimally meet diverse needs, and more emphasis on having the flexibility to choose the best system for a given function. Meanwhile we have come to appreciate the value of having a store of validated research information and a series of defined workflows that institutional stakeholders can cooperate on maintaining and using without over-reliance on a single technological solution.

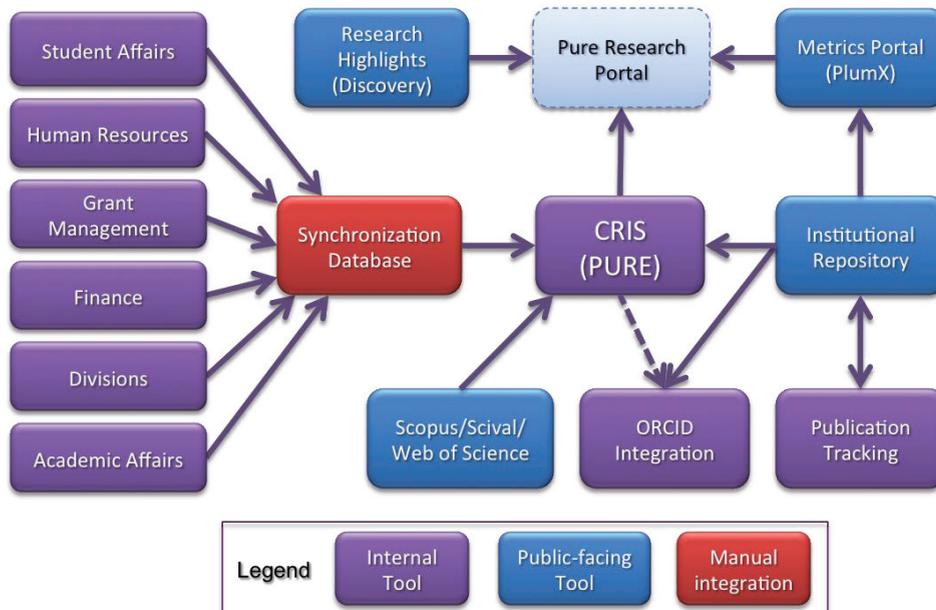


Fig.3 Current relationships within the KAUST CRIS network

3.2. Proposed model

Our experiences up to this point with gathering and organizing the research information at KAUST has prompted us to take a more system-neutral approach to our long-term thinking about how we will meet our institution's various needs regarding research tracking and reporting. In this view, what we are primarily interested in is having information validated and then stored in a way that can be reliably reused by any number of potential systems with features best suited for a specific purpose, as presented in Figure 4. While still in its early stages, the trend towards commitment to CERIF as the standard for conceptualizing research entities and their relationships and, more recently, towards the use of the CERIF API to share information between systems, holds the most potential for making such a model feasible. The initiatives by ORCID to experiment with exposing their data via a CERIF API [2], and by OpenAIRE to harvest data made available by CRIS systems as CERIF XML [3], are encouraging signs that the broader community can coalesce around these methods of information exchange. The next step will be for vendors and open source software developers to incorporate functionality for exposing and ingesting CERIF formatted information into their systems.

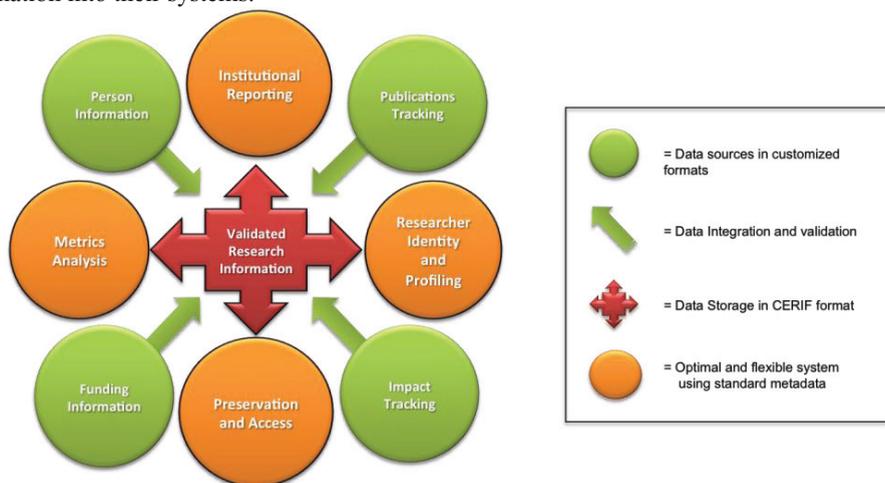


Fig. 4 Generic functional model

4. Remaining challenges

4.1. Moving towards comprehensive research information management as an institution

In those areas of the KAUST research information environment that have been brought into the research information management process through the implementation of Pure, many processes for bringing information together and validating it lack automation and consistency. This places barriers in the way of being able to have a body of data that is both of high quality and also truly up-to-date. Furthermore, as we look at the university as a whole we can see that up to this point in time only a few key units have fully engaged with the process of trying to combine their disparate internal methods of managing information relevant to the research enterprise at KAUST into a cohesive whole.

Many important areas of information remain outside of the current network of related systems and will require a concerted effort to bring together through relationships that are mutually beneficial to the involved stakeholders. Examples of this include relevant parts of the university's online presence such as faculty and researcher profiles and research group publication lists, as well as information about the use of core parts of the research infrastructure such as laboratory facilities and equipment. Developing a shared institutional approach to research data management

planning and working with researchers to follow through with it in ways consistent with their preferred practices and workflows will also be a major undertaking that the university is still only in the earliest stages of dealing with. We expect however that having a solid basis of existing information in Pure, as shown in Figure 5, will prove a valuable mechanism in engaging more internal partners in the project as they see how their particular pieces of the puzzle can be shown to be connected in meaningful ways to the broader picture of research at KAUST.

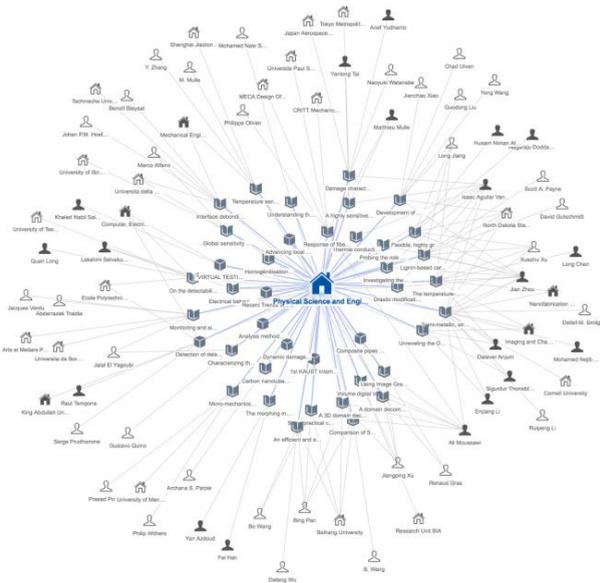


Fig. 5 Example of relationships between a department, research outputs, grants, internal and external collaborators and external institutions as visualized in Pure.

4.2. Global research information environment

The wider environment of external research-related information sources is also a key part of the picture. While there are many positive developments, such as the increased use of ORCID at different levels throughout the research information ecosystem and the incorporation of standards like CERIF as an integral part of major CRIS systems, there remain real gaps as well. One area where a real international effort is needed is in the use of a common set of organizational identifiers to identify affiliations and funders across different systems.[4]

5. Conclusion

As a young institution, we see the process of implementing a CRIS as an opportunity to establish a complete and accurate history of the research activities of the institution, so that going forward meaningful conclusions can be drawn about the effects of changes in strategy, or the outcomes of major investments. This means that while we knew at the start that gathering high quality information would be a key part of the project's success, we did not fully appreciate the level of difficulty to be encountered in bringing reliable and complete information into the system, whether from internal or external sources. Seemingly quick wins, such as bulk import of publication histories through Pure's Profile Refinement Service, turned out not to be viable options for us due to the data inconsistencies and errors introduced through such approaches. The emphasis on having information of a high enough quality to truly inform decision-making brings with it a tension that we will continue to work through, namely between casting a net for data that is broad enough for us to be comprehensive, and deep enough for us to feel confident in our analysis of the results.

Acknowledgements

This work was supported by King Abdullah University of Science and Technology (KAUST).

References

1. Baessa M, Lery T, Grenz D and Vijayakumar JK. Connecting the pieces: Using ORCID to improve research impact and repositories *F1000Research*, 2015, 4:195, <http://dx.doi.org/10.12688/f1000research.6502.1>
2. Demeranville, T. ORCID and CERIF. euroCRIS Strategic Membership Meeting, Barcelona, 2015, <http://hdl.handle.net/11366/422>
3. Houssos N, Joerg B, and Dvořák, J. OpenAIRE Guidelines for CRIS Managers 1.0, Zenodo, 2015, <http://dx.doi.org/10.5281/zenodo.17065>
4. Smith-Yoshimura, K et al. Addressing the Challenges with Organizational Identifiers and ISNI. Dublin, Ohio: OCLC Research, 2016, <http://www.oclc.org/content/dam/research/publications/2016/oclcresearch-organizational-identifiers-and-isni-2016.pdf>