Title: Extending the value of a CRIS with Research Data Management

Background

In recent years, research data management (RDM) has become increasingly important in the context of managing the research lifecycle. This has in part come from the birth and growth of Open Science and the need to make research transparent, reproducible, and accessible. Mounting concerns over research integrity and data quality have led to the promotion of initiatives such as the FAIR principles (Wilkinson et al., 2016) for RDM.

Governments and funding agencies have created policies and mandates around the communication of science and for research governance, leading to the need for institutions to collect and report on their research activities and outputs. This makes the management of research data a complex but fundamental process for institutions and it becomes of vital importance to have high quality research data. This includes information on their researchers and organization structure, projects, funding, publications, datasets, patents, ethical reviews, etc.

Current developments show that CRIS's play a key role in the management of an institution's research data, and by providing advanced and comprehensive tools to manage and report on this data they enable both efficient communication of research activities and outputs and collaboration between researchers. CRIS's have also evolved to include 'smarter' functionality for the discovery and linking of related research data, such as the datasets created and used in the research, software, ethical reviews and, more recently, Data Management Plans (DMPs). In addition, CRIS's are now integrating with research data archives, specialized RDM tools and institutional repositories guaranteeing the compliance and preservation of research data.

The Pure RDM solution

As a CRIS, Pure facilitates the management of an Institution's Research Information (RI) by providing extensive functionality for the management, reporting, and showcasing of the research data. This data is used for various purposes and processes-- from completing national assessments to fulfilling funder requirements, or to enable research collaborations.

To evaluate their research activity and build their research narrative, researchers and research administrators need to be able to trust that the data they enter and register in Pure is of the highest quality. The value of CRIS's in the management of research data is the emphasis that is put on data quality and integration and interoperability. In Pure we define data through the '5Cs of data quality'- the data must be Correct, Complete, Current, Compliant, and Connected. The 5Cs of data quality describe concepts such as accuracy and consistency of the data, completeness, relevance, and availability (and accessibility) (Azeroual & Schöpfel, 2019).

Also, and perhaps more importantly when speaking about RDM, data must be interoperable. This is addressed in Pure as in most CRIS's by the adoption of common standards for data exchange such as CERIF (*Main Features of CERIF | EuroCRIS*, n.d.) and Dublin Core (*DCMI: DCMI Schemas*, n.d.) for publications, or OpenAIRE[1] (Dvořák et al., 2018; *OpenAIRE Guidelines for CRIS Managers — OpenAIRE Guidelines Documentation*, n.d.) for datasets. Pure relies on these defined schemas when converting and storing records, files, datasets, and in particular the metadata fields associated to these objects. While

[1] Based on the DataCite metadata schema.

for textual publications (i.e. articles, proceedings, books, etc.) the data is mostly well-structured, the full research cycle also includes content such as research artifacts (e.g. datasets, software), funding content (e.g. awards, projects) or related processes such as ethical approvals and data management plans (DMP). Some of this research data can be unstructured, therefore requiring a conversion to a structured (and processable) format.
By supporting over 30 different content types, Pure can already accommodate a wide variety of research activities, processes, and outputs that can vary in nature, form, and source. The research data is enriched in Pure[2] and, through its relational database, connections to related content are found.
This makes Pure, and in general CRIS's, powerful and fundamental tools when trying to monitor and register the full research cycle comprising of different stages, processes, and outputs.
Although the value of CRIS's in facilitating the management of research data has been shown (Guillaumet et al., 2019; Jetten et al., 2019), RDM remains a complex task that must cover different aspects and stages of the research cycle. CRIS systems currently offer a large suite of features to lighten the burden of RI management but there is a need for different solutions and, to be able to ensure that researchers and research managers have the best tooling in place to manage each part of the journey, there is a need to integrate the different pieces. By putting in place new integrations with external systems, CRIS solutions are trying to ensure that all pieces of the RDM ecosystem are present and tied together.

An important aspect of RDM is the archiving (and long-term preservation) of data, which is typically stored in Institutional Repositories (IR) or, especially when it comes to large datasets, in specialist data repositories. While Pure already offers a range of integrated repository functionalities and thus the possibility to archive datasets internally, development is ongoing to expand archiving options by putting in place connections with external (institutional) repositories. An example of this is the recent integration between Pure and Digital Commons, an IR that includes features such as professional-grade open access publishing capabilities, specialized research data management tools and integrated faculty profiles. In addition to providing IR functionality, Digital Commons also includes Digital Commons Data (DCD). DCD can be used by researchers to create and share in-progress projects with collaborators, keeping track of the status and related RI.
However, not all research data is deposited in the IR. Typically, research data is distributed over a large number of data repositories (including domain specific repositories, institutional repositories and generalist repositories) meaning that, in order to keep track/register the complete research cycle a CRIS must enable connections (and subsequent normalization of data) to all these external sources, which in itself represents a further complexity when trying to provide complete coverage of an institution's research activities and outputs.
To respond to these complexities and make sure to cover the full range and variety of data sources, Pure recently implemented an integration with Data Monitor. Data Monitor is a specialized RDM tool that harvests research data from 2000+ generalist and domain-specific repositories, and normalizes the metadata following the OpenAIRE metadata schema. In this process, the research data is cleaned up by removing non-reliable and non-data sources, duplicates, datasets with dead links, and non-data records. The data is also enriched, when possible, by adding related publications, author, and institutional links

[2] Dedicated functionality such as *Available updates* and *SHERPA/RoMEO*.

captured from other sources such as DataCite, Scholix and Scopus (*Data Monitor - Track and Analyse Your Institutional Research Data*, n.d.) and using advanced machine learning algorithms.

The integration between Pure, Digital Commons and Data Monitor provides an end-to-end solution for research administrators. Research data is (automatically) imported into Pure through the integration with Data Monitor, where it is matched to related research organizations, persons, and content after being handled by the built-in tools in Pure for deduplication, link checking and reporting. This integration eases the discovery of new datasets produced at the institution that are not yet registered in Pure and presents these as candidates for import into Pure. This helps ensure that the institution has the highest possible recall of the datasets relating to their institution, and moreover, that the datasets are linked to other relevant entities such as the Person records of the dataset producers. At the same time, the integration with Digital Commons provides users with an IR where they can archive and preserve their research data. Digital Commons Data also allows users to create on-going projects that can be tracked in the DCD workspace and shared with collaborators. Additional integrations to external specialized (and open source) systems can then be implemented in Pure, satisfying different customer needs that relate to different parts of the research cycle.

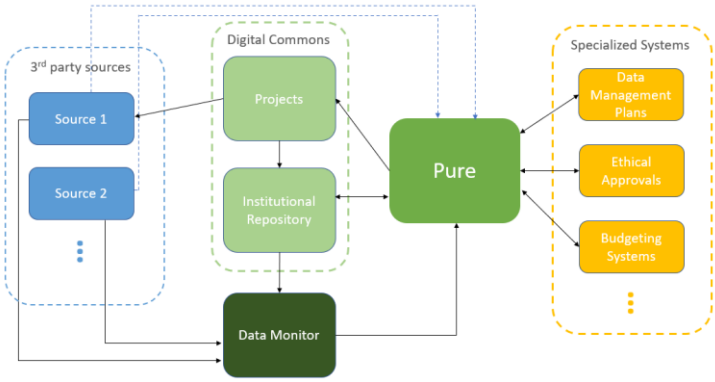

Fig. 1- The Pure RDM solution.

Case Studies

Between 2019 and 2021 we worked together with two institutions using Pure as their CRIS to develop ideal workflows for users to perform across different products, depending on the tasks that are relevant from an RDM perspective. In this paper we present these case studies and discuss the insights emerging from them, as well as the conclusions drawn from these studies and next steps.

University of Groningen

The project carried out with the University of Groningen began at the end of 2019 and centred around developing the optimal workflows and tools to optimize the discovery, import, linking, and registration of research data related to existing publications or other research activities.
In 2015, a university-wide data policy was launched specifying that research data should be registered and made publicly available, whenever possible, and followed in 2016 by the creation of a research data office, now called the Digital Competence Center. Between 2016 and 2019 over 600 datasets were added manually to Pure; an extremely time-consuming process due to the need to manually copy and paste, check, and curate the metadata of each dataset[3].

[3] around 20-30 minutes per dataset, for a total of about 300 hours.

The integration between Pure and Data Monitor enabled the automatic import of dataset metadata and linking to specific people, organizational units, publications, and other research entities. Over the 12 months in which the project ran, the number of datasets captured grew from just over 600 to 3,000. In addition, the enrichments added to the datasets in Data Monitor helped classify and correctly assign datasets (to persons, organizations, etc.), thus facilitating discovery of datasets and reducing the effort required to correctly register these datasets.
Following this project, Pure and Data Monitor worked on the introduction of a common standard for dataset metadata, compliant with the OpenAIRE guidelines.

University of Canberra

The collaboration with the University of Canberra started in mid-2020 and focused on the process of creating a Data Management Plan (DMP). The University of Canberra's data policy requires researchers to write an RDMP and, if required by the nature of the project, also obtain ethics approval. In the long-term UC is planning to make RDMPs mandatory for Higher Degree Research (HDR) students as a requirement of access to corporate data storage. In the longer term, this could also become mandatory for all research staff starting any research project. This project involved three different systems: ReDBox for the creation of the Research Data Management Plan, Pure as the Grant Management System and to update researcher profiles, and Infonetica for the Ethics Review process.
Two different workflows were developed for this case study. The first (and most relevant, covering 95% of the cases) focused on post-award workflows, for which both research data management plans (RDMP) and ethical approvals are required. The second workflow centered around non-funded projects, that are typically not recorded in Pure, e.g. a PhD scholarship which is paid directly by the institution.
For the first use case, grant applications are created and managed in Pure (in the Award Management module), before being sent out to the funder. If the application is awarded, an award record is created from the related application in Pure. The approval of the award record in Pure triggers the creation of the Research Data Management Plan (RDMP) in ReDBox and of an ethics approval in Infonetica.

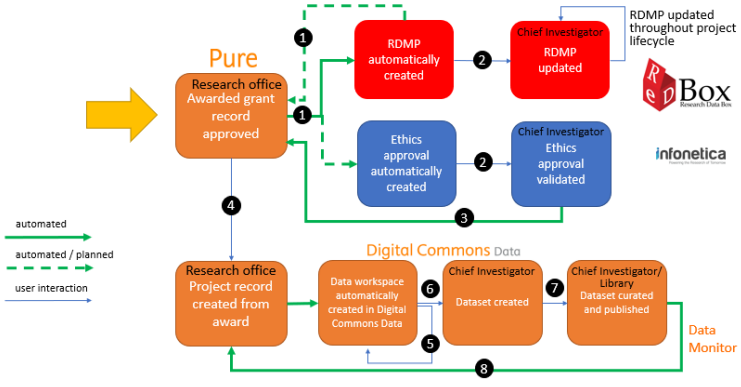

Fig.2- Architectural view of the integrated workflow.

The progress of both the RDMP and ethics approval are tracked in Pure and, once the RDMP and ethics approval processes are complete, the awardees are notified and a project is created, subsequently triggering the creation of a workspace in Digital Commons Data (DCD). All collaborators are invited to the project and researchers can link storage to the workspace, based on what is stated in the RDMP.

University of Canberra's data policy requires researchers to write an RDMP and, if required by the nature of the project, also obtain ethics approval. The current workflow for the second use case starts in ReDBox. Here the researcher enters the project information and answers RDM-related questions. The IT provisioning and ethical approval workflows proceed in parallel. If an ethics approval is required, an ethics application is created by the researcher in Infonetica. Once the ethics review has been completed, the researcher must return to ReDBox and update the RDMP, adding the information that the ethics approval has been cleared. Currently, there is no direct integration between the two systems.

One of the takeaways from this work was that workflows are multiple and strongly dependent on the process that needs to be completed. Even within the same process, the workflow can vary depending on the type of content that is being managed. The University of Canberra currently has plans to track institutional centrally funded projects, where the funder is the office of the VP Research, in Pure. In this case, the optimal workflow will begin by creating the project directly in Pure and then triggering the creation of DMP and ethics approvals, as needed, with a workflow that is very similar to the one described before, with the only difference being the event that triggers the integration

Conclusions

As the focus on RDM increases and more and more countries and funding agencies require institutions to track and register all the research outputs and data describing their research activities, the need to effectively (and seamlessly) manage RI becomes more urgent. Although the value of CRIS's for the management, reporting and showcasing of (textual) research outputs is unobjectionable, the evolving research landscape and push towards Open Science leads to the further need of managing and linking different types of content.
In this paper we highlight some of the main RDM requirements, including high data quality and integrations with different types of external data sources and (open or institutional) repositories. We show how structured integrations with specialized tools, systems, API's and repositories not only alleviate the burden for research administrators, but also allow to 'adjust' the wide range of functionality provided by CRIS's to different types of research data, processes, and workflows. The two case studies discussed represent two of the use cases commonly encountered by researchers and research administrators and from which we provide insights, lessons learned, and future developments. We conclude that, although today's CRIS's provide advanced functionalities for the management of research data, there is still much work to be done in capturing, linking, and registering the different parts of the research cycle. Our approach of working together with the community to find and develop an optimized solution for a comprehensive management of RI has proven to be the right one, and we will continue to further develop and improve Pure based on the feedback and suggestions of our customers. Our goal is to reach a point in which connections between Pure and external systems can easily be set up, to be able to offer the research community the flexibility that they require in a continuously changing/evolving landscape.

Acknowledgements

The functionality and use cases presented in this paper are the result of a close collaboration with Elsevier's Data Monitor and Digital Commons teams. We would also like to thank the University of Groningen and the University of Canberra for their effort and valuable feedback during the course of these projects.

References

Azeroual, O., & Schöpfel, J. (2019). Quality Issues of CRIS Data: An Exploratory Investigation with Universities from Twelve Countries. *Publications 2019, Vol. 7, Page 14*, *7*(1), 14. https://doi.org/10.3390/PUBLICATIONS7010014

Data Monitor - Track and analyse your institutional research data. (n.d.). Retrieved March 9, 2022, from https://www.elsevier.com/solutions/data-monitor

DCMI: DCMI Schemas. (n.d.). Retrieved March 10, 2022, from https://www.dublincore.org/schemas/

Dvořák, J., Bollini, A., Rémy, L., & Schirrwagen, J. (2018). *OpenAIRE Guidelines for CRIS Managers 1.1*. https://doi.org/10.5281/ZENODO.2316420

Guillaumet, A., García, F., & Cuadrón, O. (2019). Analyzing a CRIS: From data to insight in university research. *Procedia Computer Science*, *146*, 230–240. https://doi.org/10.1016/J.PROCS.2019.01.097

Jetten, M., Simons, E., & Rijnders, J. (2019). The role of CRIS's in the research life cycle. A case study on implementing a FAIR RDM policy at Radboud University, the Netherlands. *Procedia Computer Science*, *146*, 156–165. https://doi.org/10.1016/J.PROCS.2019.01.090

Main features of CERIF | euroCRIS. (n.d.). Retrieved March 10, 2022, from https://eurocris.org/services/main-features-cerif

OpenAIRE Guidelines for CRIS Managers — OpenAIRE Guidelines documentation. (n.d.). Retrieved March 9, 2022, from https://guidelines.openaire.eu/en/latest/cris/index.html

Wilkinson, M. D., Dumontier, M., Aalbersberg, Ij. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J. W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., … Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data 2016 3:1*, *3*(1), 1–9. https://doi.org/10.1038/sdata.2016.18