From Open Repositories to CRIS. A Case Study

Joachim Schöpfel[1]

Otmane Azeroual[2]

Stéphane Chaudiron[1]

Bernard Jacquemin[1]

Eric Kergosien[1]

Hélène Prost[3]

Florence Thiault[4]

[1] Univ. Lille, ULR 4073 - GERiiCO - Groupe d'Études et de Recherche Interdisciplinaire en Information et Communication, F-59000 Lille, France

[2] German Centre for Higher Education Research and Science Studies (DZHW), Berlin, Germany

[3] CNRS, ULR 4073 - GERiiCO - Groupe d'Études et de Recherche Interdisciplinaire en Information et Communication, F-59000 Lille, France

[4] Univ. Rennes 2, UR 4246—PREFICS—Pôle de Recherche Francophonies, Interculturel, Communication, Sociolinguistique, Université Rennes 2, F-35043 Rennes, France

Keywords

Current research information systems (CRIS)

Institutional repositories (IR)

Open access

Grean road

Self-archiving

Research evaluation

Repository usage

French national HAL repository

Open Science

Data quality

Extended abstract

The development of Current Research Information Systems (CRIS) and Institutional Repositories (IR) initially involved distinct systems with different objectives, functionalities, standards, and user groups. They both contribute to the "fourth paradigm" (Hey et al., 2009) of data-intensive scientific discovery, representing a shift in scientific practices enabled by information and communication technology. Despite a historical separation and discussions on data ingestion, exchange, and interoperability, there has been a convergence and even merging of CRIS and IR (de Castro et al., 2014).

While academic librarians traditionally focus on managing open access and IR, their involvement in research evaluation and CRIS tends to be limited. Presently, IR often serves the purpose of monitoring and assessing institutional research performance, while CRIS, beyond metadata processing, has expanded to store, preserve, and disseminate research papers. A survey involving 84 institutions across 20 European countries indicates that 65% of them have established connections between their CRIS and IR (Ribeiro et al., 2016). This interaction and convergence between the two systems have notable impacts, manifesting in distinct ways.

Firstly, there is a trend where CRIS platforms now incorporate functionalities associated with IR, allowing the deposition of documents, traditionally a service of open repositories. Notably, Pure, initially a standard CRIS developed by Atira and now a research information management solution by Elsevier, has evolved into a campus-wide content management system, explicitly designated as a repository.

Secondly, open repositories are adopting functionalities traditionally linked with research information management. For instance, HAL, the French national repository, initially created for self-archiving research papers, has transformed into a platform with institutional portals and collections. To meet institutional demands for scientific monitoring, HAL has integrated features for scientometrics and research output assessment, adhering to national and international standards and identifiers.

Our paper will describe and discuss the evolution of open repositories, considering the recommendation of the 2002 Budapest Open Access Initiative for self-archiving as the "green road" to open access. Despite two decades, only a portion of researchers deposit their publications in open repositories, and some repository content originates from nonfaculty contributions. The paper aims to provide empirical evidence on this situation and assess its impact on the future of the "green road."

Methodology

Based on our assessment of the global situation of convergence between CRIS and IR (Schöpfel & Azeroual, 2021), we analysed the contributions to the French national HAL repository from more than 1000 laboratories affiliated with the ten most important French research universities, with a focus on 2020, representing 14,023 contributor accounts and 164,070 deposits (quantitative part) (Schöpfel et al., 2023).

Additionally, we carried out a survey with 400 laboratories and conducted interviews with senior scientists and information professionals of 50 laboratories, to learn more about the usage of HAL (qualitative part).

Results

The empirical findings of our quantitative analysis illustrate the extensive transformation of the French national HAL infrastructure from a self-archiving open repository, akin to arXiv, to an open platform incorporating publications and metadata (records) from diverse origins. In our analysis encompassing over 1000 laboratories from the top ten French research universities, only half of the 2020 deposits are self-archived, with the remaining half comprising mediated contributions, primarily

from non-faculty sources. This mediated contribution necessitates and reflects institutional support and assistance, serving three main purposes: firstly, fostering open access and direct scientific communication by generating content within the repository; secondly, ensuring the long-term preservation of resources, as all HAL deposits are backed up in a publicly accessible dark archive hosted by CINES in Montpellier; and thirdly, developing an infrastructure for monitoring and evaluating the scientific production of individual researchers, laboratories, and universities. Through the backing of institutions and contributions from laboratories and universities, HAL has evolved into a showcase for their scientific output. Additionally, it contributes data to the French Open Science Monitor. Importantly, this mediated contribution is not a temporary strategy aimed at creating critical mass during the initial period after the repository's launch. Our findings highlight mediated contribution as a substantial and integral aspect of the ongoing and permanent functionality of the repository. HAL finds itself positioned in the middle of the trajectory from being an open repository (following the green road, as advocated by the Budapest Initiative) toward a distinctive form of open research information management system.

The qualitative analysis (interviews) enabled us to conduct a comprehensive evaluation of the tools and systems adopted by laboratories in their HAL practices. Some tools are situated upstream of HAL, focusing on monitoring scientific production, managing references, and/or supplying content to HAL. On the other hand, there are tools positioned downstream, utilized for updating internal databases, feeding website pages, analysing production, and generating assessments. The range of systems implemented by laboratories varies widely. Certain labs restrict themselves to managing references of their scientific production, employing tools for management, import, and export, where HAL essentially serves as a bibliographic database, sometimes replacing an internal database. In contrast, other laboratories go beyond this, with systems like LabMetry, which partially encompass the functions of a research information system. This includes monitoring scientific projects and providing information essential for assessments, activity reports, and evaluation campaigns (De Castro, 2018). However, it appears that only a few laboratories are utilizing functional and appropriate software.

The qualitative analysis uncovered another facet of this transformation. When questioned about potential alternatives for disseminating their publications, respondents compiled a list of sites distinguished not merely by their quantity but by their diversity, especially in terms of functionality. According to their perspective, HAL could be likened to, or even substituted by, open archives, preprint servers, databases, search engines, aggregation sites, social networks, and even scientometric evaluation tools or personal pages.

Despite their variations, what these systems share is that, in the eyes of researchers, they fulfil functions similar to those of HAL, at least to some extent. They can be utilized – within certain limitations – equally for managing references, disseminating publications, conducting assessments, and so forth. In this sense, they represent partial functional equivalents (Merton, 1949), devices with practices and services that are comparable, functionally speaking, to those offered by HAL. Importantly, this concept of functional equivalence is based on partial cognitive perspectives of scientific personnel, not stemming from a comprehensive comparative study of the services, functionalities, and practices of various devices, sites, servers, and platforms. This partial cognitive functional equivalence endows HAL with a distinctive position within the landscape of open science infrastructures, different from other open repositories.

Conclusions

This transformation is not exclusive to HAL and extends beyond the borders of France. Moreover, our objective is not to pass judgment on whether this shift is beneficial for open science or not. Instead, we aim to highlight a specific yet critical challenge: the impact of this transformation on the

significance of data and metadata quality. As the platform takes on monitoring and assessment functions, the quality of the data becomes a crucial criterion for the effectiveness and acceptance of the system's functionalities and services. This necessitates a comprehensive and ongoing evaluation of data quality and the implementation of specific measures to control and enhance data quality throughout the entire process, including upstream of data import and creation. This involves the FAIRization of data, encompassing a qualified and standardized utilization of the contributor field and rigorous control over input from other platforms.

Issues such as misspellings, homonyms, incorrect or missing identifiers, and erroneous attributions of scientific works are already significant concerns affecting the discoverability of resources in open repositories. However, as repositories and research information management systems converge, these challenges will become increasingly critical due to the potential adverse effects of poor data quality on institutions, projects, and individuals. Will the shift away from self-archiving lead to a decline in metadata quality? We don't believe so; given the gravity of the issues involved, we anticipate the opposite - a continuous enhancement of metadata quality through improved controls, standardization, and enhanced curation functionalities, including extensive use of persistent identifiers.

References

De Castro, P. (2018). Mapping the European CRIS infrastructure and its potential applications. *Antwerp ECOOM Workshop "Working with National Bibliographic Databases for Research Output"*, 10-11 September 2018. https://dspacecris.eurocris.org/handle/11366/705

De Castro, P., Shearer, K., & Summann, F. (2014). The Gradual Merging of Repository and CRIS Solutions to Meet Institutional Research Information Management Requirements. *CRIS 2014*, 13–15 May 2014, Rome, Italy. https://doi.org/doi:10.1016/j.procs.2014.06.007

Hey, T., Tansley, S., & Tolle, K. (Eds.). (2009). *The fourth paradigm. Data-intensive scientific discovery*. Redmond WA, Microsoft.

Merton, R. K. (1949). *Social theory and social structure*. New York: Free Press.

Ribeiro, L., de Castro, P., & Mennielli, M. (2016). *EUNIS – euroCRIS joint survey on CRIS and IR*. Final report. Paris, EUNIS. Retrieved from http://www.eunis.org/wp-content/uploads/2016/03/cris-report-ED.pdf

Schöpfel, J., & Azeroual, O. (2021). Current research information systems and institutional repositories: From data ingestion to convergence and merger. In D. Baker & L. Ellis (Eds.), *Future Directions in Digital Information* (pp. 19–37). Oxford: Elsevier Chandos. https://doi.org/10.1016/B978-0-12-822144-0.00002-1

Schöpfel, J., Chaudiron, S., Jacquemin, B., Kergosien, E., Prost, H., & Thiault, F. (2023). The Transformation of the Green Road to Open Access. *Publications*, *11*(2), 29. https://doi.org/10.3390/publications11020029