Principles and pragmatics of "as open as possible": persistent identifiers as the interface between research information commons and closed systems

Clifford Tatum[1] and Josh Brown[2]
05 March 2018

A Current Research Information System (CRIS) depends on the availability of open, reliable, trustworthy research information to function. The concept of FAIRness[3] (i.e. data that is Findable, Accessible, Interoperable and Reusable) underpins and facilitates this trustworthiness. A CRIS ingests information from multiple sources, usually in the form of metadata. Being able to locate information about research is an obvious precondition of actually using it. Accessibility and interoperability are essential to the process of obtaining and ingesting data in a useful form. Finally, a licence that defines permissible reuse is essential.

Once in the CRIS, research information from public sources is often merged with private data (for example, data drawn from Human Resources systems within researchers' home institutions) to enable robust analytics with a richer context. At an institutional level, this is sometimes enough. The institution now has information that provides strategic insight into their research portfolio and can make informed decisions about how they support that portfolio and build upon their strengths.

However, this mixing of public and private information complicates further re-use of the public information. Once in the CRIS, open data linked to private data (e.g. about persons) renders closed the previously open data which has been enriched through linkage to private information. By design, CRISs operate across the boundary between public and private information about research and mixes personal, organisational and financial information. However, because the data gets 'mixed', a lot of what goes into a CRIS goes dark. By this, we mean that it is no longer open and no longer findable, let alone accessible. In this context, a CRIS risks becoming an information 'black hole'. In a world that is increasingly reliant on open science to address its shared challenges, this is problematic to say the least. Given the reliance on the provision of open data for CRISs to function, it does not seem unreasonable to ask such systems, or their institutional users, to ensure their collection of open research information remains open. In the spirit of open science policies that prescribe data that is "as open as possible[4]" the question we raise is this:

How much is 'open enough' when it comes to sharing research information?

[1] clifford.tatum@surfmarket.nl https://orcid.org/0000-0002-2212-3197

[2] j.brown@orcid.org https://orcid.org/0000-0002-8689-4935

[3] FAIR data principles: https://www.force11.org/group/fairgroup/fairprinciples

[4] Here we draw on the oft-cited principle of open data, "as open as possible, as closed as necessary." See for example, EC 2016 H2020 Programme Guidelines on FAIR Data Management in Horizon 2020

In this paper we propose that diligent registration and use of persistent identifiers for research objects provides a pragmatic solution for *how much is enough*. We contend that robust analytics and strategic advantage need to be balanced against the health, independence and sustainability of the research ecosystem. A healthy ecosystem is one in which the principles of open science, supported by the practical precepts of FAIR research information, operate in a dialogue with the pragmatic need for privacy and protection in a competitive environment.

On the one side, we have a public research information space in which the principles of interaction are defined in terms of openness. Open science brings together concepts of access, citability, reuse, aggregation and sharing, together with a fundamentally transparent and accountable approach to research decision-making. FAIRness is one way of defining what it means to be open in practical terms. Each of the four pillars of FAIRness is a necessary, but not sufficient, condition for openness.

At a national level, the setting of priorities, the understanding of our strengths and the ability to deliver support in a way that is relevant, constructive and tied to social priorities and common needs is best served by the aggregation and analysis, reporting and evaluation using the widest possible range of research information, and the deepest possible understanding of how research is working[5]. In our view, the whole system works better for the widest range of stakeholders when the information map is filled in. For us to grasp the scale, and contours of the research landscape, we depend on an aggregation of information from multiple, often disparate sources all acting in concert.

There can also, of course, be moral reasons for keeping research information closed. We should note here that we are not speaking of research data, i.e. data collected for the purpose of conducting research. Nevertheless, the ethical considerations of open research data are in some ways also relevant to this debate. Such considerations relate to the privacy and security of information about individuals who contribute to research, as well as to the reputation and credibility of the organisations tasked with safeguarding that information. Pragmatically speaking though, we observe substantial tensions between the imperative to openness and the demands of competition between research organizations for hiring talented researchers and acquiring research funding or competing commercial interests among service providers.

We aim to set out a case for a minimum threshold at which a research object's existence is openly knowable, while retaining the possibility of making access to its content conditional. Using persistent identifiers in a CRIS can provide a granular level of control for applying licensing logic or ensuring separability from the mix with private information. In this way, research objects that are persistently identified can be ingested by a CRIS while remaining findable.

We argue that the minimum required openness is tied to the principles of transparency that underpin open science, or findability in terms of FAIRness. The minimum set of open

[5] As an example, the promise of the 'Science of Science' depends on this definition of openness; see: Santo Fortunato, Carl T. Bergstrom, Katy Börner, James A. Evans, Dirk Helbing, Staša Milojević, Alexander M. Petersen, Filippo Radicchi, Roberta Sinatra, Brian Uzzi, Alessandro Vespignani, Ludo Waltman, Dashun Wang, Albert-László Barabási. Science of Science. Science 02 Mar 2018: Vol. 359, Issue 6379, eaao0185 DOI: 10.1126/science.aao0185

information to enable others in the community to interact with a research object is *knowledge of its existence*. The community should know where research information is, even if it's not accessible or reusable. The simple fact that the data is housed in a CRIS would ideally guarantee its technical interoperability. As a practical mechanism to achieve, if not full interoperability then at least data portability[6], we propose that CRISs should expose existing persistent identifiers (such as DOIs or ORCID iDs) for their data points wherever they are available. This acknowledges that when the data was acquired by the institution, it was open and it should be allowed to remain open.

More ambitiously, what is original in a CRIS is the curated aggregation of data that it enables. The CRIS should therefore provide open, persistent identifiers for the profiles or other collections of data that they house. As a proof of concept for such a 'compound object' identifier, the Research Activity ID (RAiD)[7] developed by the Australian National Data Service, is "persistent and connects researchers, institutions, outputs and tools together to give oversight across the whole research activity and make reporting and data provenance clear and easy." By analogy, a similar approach could be taken to provide open, persistent identifiers for the meaningful content of a CRIS whilst maintaining control over access to that content.

An open, persistent identifier is FAIR by definition, but the object it points to can be made findable, irrespective of accessibility or openness. Once the existence and whereabouts of research information is signposted by a persistent identifier, it can be made actionable by the consistent provision of a core set of metadata, including at least:

- Source data - the location and nature of the system/profile that hold the info
- Rights information - access restrictions defined
- Negotiation - ability or request or determine access electronically or manually

To provide a real-world example of how this approach can work, consider the case of peer review in journal publishing. Peer review is at present most often recognised via an acknowledgement on the journal web page or, independently, on a researcher's CV. However, that limits both the FAIRness of the peer-reviewer's contribution and transparency of the process[8]. By posting a review acknowledgement to a reviewer's ORCID record, the journal can share the fact that the activity has taken place, and link it to persistent identifiers for the publication, the journal, and potentially the publisher as well.

This is already occurring today, and it illustrates the flexibility of the balance we describe here. Some peer reviews are completely open, such as those in F1000 Research. In this case, the review citation contains additional persistent identifiers in the form of DOIs for the article and for the content of the review. In other cases, where warranted by peer review procedures, the journal simply acknowledges ambiguously that a researcher has performed peer reviews for them within a given time period, thus preserving the integrity of the double-blind review process.

[6] Tatum, Clifford (2016): Towards governance of PID portability for research evaluation. figshare. Presentation. https://doi.org/10.6084/m9.figshare.4212732.v1

[7] https://www.raid.org.au/

[8] Lamont, Michèle. 2009. *How Professors Think*. Harvard University Press.

The peer review example demonstrates ways in which the existence of information about research activities and events can be made knowable, even if the relationship to specific data points is obscured. In this way, persistent identifiers, which are actionable by design, serve as an interface between public and private aspects of research information. Persistent identification of research objects thus provides a minimum practical threshold to deliver visibility, traceability and citability of knowledge objects and processes. There are concrete steps that every sector in the research community can take to help to ensure that the minimum threshold set out here is reliably, consistently met. CRISs, as the institutional research information hub par excellence, play a vital role in ensuring that research performing organisations are playing their part.