Organisation identifiers: current provider survey

Geoffrey Bilder, Josh Brown and Tom Demeranville

This document is based on an analysis of the current landscape of Organisational Identifier provision, conducted during the second quarter of 2016 by Geoffrey Bilder (Crossref) and augmented by further research conducted by Josh Brown and Tom Demeranville (ORCID). This public discussion paper has been prepared as a complement to the Organisation Identifier requirements document published by Crossref, DataCite and ORCID in March 2016. It informs our analysis of the current landscape of identifer provision, and is intended to be read alongside companion documents concerning the governance and functional requirements for Organisation Identifiers.

Introduction

The scholarly communications sector has built and adopted a series of successful open identifier and metadata infrastructure systems. Resource identifiers (through Crossref and DataCite) and person identifiers (through ORCID) have become foundational infrastructure to the industry.

Scholarly communications is notable for its extensive development and use of open identifier and metadata infrastructures. Identifiers have become widely used and firmly established enabling infrastructures for the industry. Our adoption of identifiers has been driven by necessity. With the steady increase in research outputs, and the increasing number of active researchers from both academia and industry, research stakeholders find they need to be able to automate workflows to scale their systems efficiently.

Funders want to be able to track the outputs that arise from research they have funded. As a result, institutions find themselves having to regularly analyse and summarise the research their faculty produce. Faculty, in turn, are facing increasing accounting bureaucracy in order to meet all the reporting requirements that are cascading through the system. Finally, publishers are seeking to make the manuscript submission and evaluation process more efficient as well as to increase the discoverability and richness of their content.

DOIs and ORCID iDs have become deeply embedded in a host of traditional scholarly communications tools, including manuscript tracking systems, hosting services, bibliographic management tools and discovery tools. Additionally, scholarly identifier systems have played a key role enabling the development of entirely new application categories, including alternative metrics services, CRIS systems and services for content tracking and promotion.

Context

Resource identifiers and person identifiers are two key pillars of the infrastructure that is powering these systems. Yet there is third pillar that is needed to truly deliver potential efficiencies across the scholarly community, and that is organisational identifiers. Despite many players attempting to do so, there is still no equivalent to Crossref, DataCite or ORCID providing a robust, open and stakeholder-governed identifier system for organisations. There are several players attempting to do so, but none of them has yet to combine the required services and non-functional requirements to become a trusted infrastructure provider to the sector.

Wishing to close this gap, Crossref, DataCite and ORCID have been collaborating to:

- Explore the current landscape of organisational identifiers
- Collect the use-cases that would benefit our respective stakeholders in scholarly communications industry
- Seek to bring the community together to forge a consensus that can drive the delivery of an organisation identifier solution that meets an agreed subset of the use cases collected.

In pursuit of these goals we have presented our analyses and solicited feedback at several important stakeholder events. We have also met with individual publishers, funders, institutions, researchers and service providers to better understand what gaps exist in existing organisational identifier offerings. Following that we have met with several of the parties who have been working in the organisational identifier market to see what, if anything, they might be able to do to address these perceived gaps.

We organised workshops at the Coalition for Networked Information (CNI) Spring Meeting and the Force2016 conference. Both sessions were attended by a wide variety of stakeholders including librarians, funders, institutional administrators, standards organisations, publishers, system integrators, and organisational identifier providers. The goal of these meetings was to validate our understanding and characterisation of the problem and expand our set of use-cases. We encouraged all participants to collaboratively expand a draft “Minimum viable product requirements” document that we made public via Google Docs and subsequently published.

The consensus of the groups was that organisation identifiers could play a critical role in improving the workflows of numerous stakeholders. Similarly, there was consensus in support of the argument that there was still no organisation identifier service that was performing an equivalent function to those of Crossref, Datacite and ORCID for resource and person identifier infrastructure.

We then presented the issues and complexities that had been identified and asked for both feedback and more use cases. At the CNI meeting we were confined to asking for feedback on the document linked to above. At the Force meeting we broke out into interest groups and developed use cases and requirements. We also encouraged the workgroups to modify the shared requirements document directly.

The details of the various use-cases can be found in the document which we shared with the participants, and we will not reiterate the list here. It is enough to observe that there are a number of best practices associated with administering PID infrastructure systems that we rarely make explicit because they are so intrinsic to everything that we do. They include things like ensuring that organisation identifiers:

- Are globally unique
- Are stable
- Are discoverable
- Are resolvable
- Are not recycled
- Are documented
- Have appropriate metadata associated with them
- Are interoperable with other identifiers through relationship metadata
- Can can be merged/split when necessary
- Are expressed as HTTP(S) URIs
- Support content negotiation for machine representations
- Support discovery APIs
- Have transparent, non-profit governance
- Offer the ability for organizations to manage their own records

A companion paper to this document will discuss the practical ramifications of functional requirements in the context of the use cases set out earlier this year.

In addition, our discussions surfaced a core group of non-functional requirements which describe the qualities of a system as opposed to specific functional behaviour. For example, we know from experience that uptake of scholarly infrastructures is tied to the perceived reliability of the system and trustworthiness of the organisation running it. Reliability and trustworthiness are just two of a few broad categories of non-functional requirements that the community has expressed are critical for adoption. In general the non-functional requirements identified by the community are as much about the infrastructure itself as they the party responsible for managing and maintaining the infrastructure and the principles under which that party operates.

A second companion paper to this document sets out the expectations for governance of such an organisation in detail, and is designed to address many of these non-functional requirements.

Landscape overview: current players

The following is a brief overview of current players in the industry. Many of these organisations provide identifiers and value-add services to a defined section of the scholarly communications community. They are, for the most part, commercial entities and have achieved a degree of sustainability by focussing on the use cases of their customers/members. While, as noted above, they are delivering solid services, none of them yet meets the requirements identified in our community consultation. In this overview of the current provider landscape, we examine key organisation identifier providers and assess their offer in relation to these community requirements..

Open Funder Registry

The Open Funder Registry (neé FundRef - http://www.crossref.org/fundingdata/registry.html) was created to meet the need funders have to track published literature resulting from research that they fund. The registry was seeded with a donated funder taxonomy that was developed for internal use by Elsevier. Crossref assigned a DOI to each entry in the registry and released the registry under a CC-0 license.

The registry has grown from about 2K entries when it was first launched in 2013 to about 12K entries today. Crossref members can suggest new entries for the Funder Registry by including them in their normal Crossref deposits. Any entry in a Crossref deposit that does not have an associated Funder identifier is automatically passed on to the Elsevier team as a candidate for inclusion in the system. The Open Funder Registry currently has no mechanism for allowing organisations to edit and manage their own records directly, however organisations can request additions or modifications via the Open Funder Registry's product manager. Several US federal agencies, for example, have worked closely with Crossref to provide detailed funder metadata, including organisational hierarchies.

Governance of the registry is provided by a multi-stakeholder Crossref working group, which serves in an advisory capacity to Crossref staff and the Crossref board. The maintenance and updating of the registry is performed by the team at Elsevier that created the original taxonomy. Elsevier currently perform this function free-of-charge. Provision of the registry and updates via Crossref are funded as part of Crossref's core infrastructure.

Crossref has close to a million DOI records that now have at least one Open Funder Registry identifier associated with them. Crossref also provides a set of free, open APIs for accessing the Funder Registry and for searching Crossref metadata by funder identifiers.

Crossref also provides an SLA-backed version of the same API which provides guarantees on uptime and response time.

There are several perceived shortcomings of the Open Funder registry:

- The scope of the registry is limited to funders. This in turn means that it cannot provide an adequate basis for a more general organisational identifier without significantly increasing its coverage.
- Crossref provides no clear way for non-members to update or correct their own records.
- The Open Funder Registry is governed by a working group that provides an advisory role to the Crossref board. While the working group does represent multiple stakeholders, including funders and institutions, the Crossref board does not. In theory the Crossref board could override the wishes of the Open Funder Registry working group. This runs counter to one of the non-functional requirements, which is that the governance of a broader organisational identifier infrastructure should include wide stakeholder representation.
- The service currently depends on the goodwill of Elsevier and Crossref to cover its operating costs.

International Standard Name Identifier (ISNI)

ISNI is an ISO standard, governed by the agency ISNI IA (http://isni.org/). The database held 9.12 million records at the time of writing, which describe both individuals and organisations. Data are contributed to the system by 37 organisations. A large part of the database is derived from the Virtual International Authority File (VIAF) which is hosted by OCLC.

Data contributed to the system are processed by a Data Quality Team that is split between the British Library and the Bibliothèque Nationale de France. ISNI has 9 registration agencies mainly focused on registering ISNIs for library materials. They also handle additions and corrections for their respective constituencies. Ringgold (see below) is an ISNI registration agency for organisation names.

ISNI Data is licensed under a custom "ISNI International Agency Information License" described as open, which has an attribution requirement similar to CC-BY. ISNI reserves the right to switch to a "more restrictive license" in future if they detect "misuse of the 'Information', in particular by dissemination of corrupted copies of it". The license does not define what "misuse" or "corruption" mean.

ISNI provides basic read-only public and member APIs based on the Search/Retrieve via URL (SRU) standard. The public API provides a subset of publicly available data. The member API provides access to all non-confidential metadata elements as well as different indices on which to search. Updates are handled by offline manual processes involving emailed spreadsheets and/or XML files.

ISNI's revenue model includes an initial fee of €250 with an €999 annual fee. Membership and annual fees are not scaled according to the revenue of the member. ISNI also charges for registering ISNIs with the priced scaled to the number of ISNIs registered. Prices range from US$25 for a single ISNI through to block pricing equivalent to €0.10 per identifier for up to single batch uploads of up to 50K records, and €0.05 per identifier for up to 3 million.

There are several perceived shortcomings of the ISNI system.

- The license applied to ISNI data is a non-standard "open" license and includes requirements (e.g. attribution) that are not considered best practice for data. The licence is not fixed and could become more restrictive at any time.
- The ISNI database is not focused. For example, it includes identifiers for people and organisations. The system is not limited to research and scholarly communications and instead covers all creative works and their contributors.
- The ISNI system was not designed with organisation identifiers in mind.
- The ISNI organisation is not transparent. It is unclear how big it is, whether it has dedicated support staff or the extent of its technical resources.
- Information about the ISNI sustainability model is unavailable, so cannot be assessed.
- Non-members cannot access all public data in a machine-readable format.

Ringgold

Ringgold (http://www.ringgold.com/) focuses on providing organisation identifiers for institutions in the scholarly supply chain. Specifically, the identifier has been designed to help disambiguate institutional subscribers to scholarly content. There are over 400K Ringgold Institutional identifiers which includes aggregators, consortia, licensees, subscribers, subscription agents, and institutions.

Ringgold is also an ISNI registration agency and has many of its records have been mapped to ISNIs. Ringgold's data is not available under an open license except for that metadata that they submit to ISNI which is held under the above-mentioned ISNI license. ISNI has a limited public API that restricts users to 10 queries a day. Ringgold's revenue is based on providing services to publishers including database cleanup, auditing and consulting on business intelligence.

Ringgold is a provider of organization identifiers in the ORCID Registry, in an agreement in which ORCID exposes Ringgold organization names, identifiers, and basic metadata for linking with ORCID records. Per the license, these data are made available to the community for free to reuse.

There are several perceived shortcomings of the Ringgold system outside of those already mentioned in relation to ISNI:

- Ringgold metadata is proprietary.

- Ringgold is an entirely private organisation and does not have a transparent, stakeholder-driven governance structure.
- There is limited access to the data via API or user-driven search interface
- There is restricted capability for organizations to manage their own metadata.

Publisher Solutions International

Publisher Solutions International (PSI - http://www.publishersolutionsint.com/) has built a database of ~70K institutional identifiers and related metadata which it uses to help publishers fight subscription fraud, collect business intelligence and verify IP addresses. The identifiers are the basis of their IP Registry - http://theipregistry.org/ - a newly launched service to enable institutions and publishers to exchange verified IP address ranges via a centralized registry.

PSI data includes hierarchies, city and country, and related organizations. The data is derived from publisher systems with manual cleanup done by PSI. It naturally focuses on organisations that consume the products of research. PSI data is proprietary and not currently available via an API.

There are several perceived shortcomings:

- PSI data is proprietary and not currently available via any APIs.
- PSI is focused on addressing subscription auditing and IP address verification.
- PSI is a small, commercial organisation and there is no transparent, stakeholder driven governance structure.

GRID

Digital Science has built a database of ~64K institutional identifiers and related metadata. The GRID (http://grid.ac/) system was initially designed to serve Digital Sciences in-house projects but they made the decision to release a subset of all the metadata under a CC-BY license. GRID has a data curation team and adds entries through entity extraction and matching against the full text of research papers and grants. When an extracted entity does not match an entry in the GRID system, it is sent to the curation teams for consideration.

GRID identifiers are machine and human readable, expressed using schema.org/Organisation RDFa markup and include links to other identifier schemes. These identifiers include Open Funder ID, ISNI, OrgRef, Wikidata, UCAS ID, government databases and more. The GRID database is released monthly as a downloadable file. There is currently no mechanism for an organisation to make its own changes to records other than requesting modifications via email.

GRID is sustained as part of Digital Sciences day to day operations, as it is required by their other offerings. In addition, the GRID sustainability model includes paid-for services around data cleanup and matching. GRID also has a set of paid-for APIs for external clients. GRID doesn’t publish any public price lists for its service.

GRID has some limited public user interfaces for searching database and for extracting affiliation data from full text.

There are several perceived shortcomings of the GRID system:

- The CC-BY 4.0 license is not considered best practice for data that is made openly available for others to incorporate into their systems and merge with data from other sources making attribution difficult.
- GRID is managed by an entirely private organisation and there is no transparent, stakeholder driven governance structure.
- There is no clear method available for organisations to update their own records, although the GRID team are considering ways of enabling this.

LEI

The Global Legal Entity Identifier Foundation (GLEIF - https://www.gleif.org/en/) is a not-for-profit organization that was created to create and manage the Legal Entity Identifier (LEI). This identifier infrastructure was created in response to the global financial crisis of 2008 as a means of keeping track of hitherto opaque corporate and financial relationships. The LEI database contains ~ 450K identifier records which it has released for download under a CC-0 license.

LEIs are created by a series of LEI Issuers who generally operate on a national basis. The GLEIF is supported by the Issuers who, in turn, charge end-users to register LEIs. The basic charge for a registration request is US$ 200. Maintenance or update requests are charged at sliding rate between US$ 8 and US$ 100. All deposits and maintenance requests are reviewed by a validation team before they are accepted into the system. GLEIF runs a number of public APIs and discovery tools.

There are several perceived shortcomings of the LEI system:

- The focus of the data is currently on legal entities in the financial sector. At present the data model does not support listing of alternate entity names that would resonate in scholarly communications use cases.
- Depositing and maintaining LEIs at scale could be relatively costly due to the per-registration charge.
- The management and governance of the organisation, while transparent and international, is not made up of stakeholders from, or by and large connected to, the scholarly communications community, although there has been interest in expanding their board to include such representation, indicated by the addition of Sloan Foundation on their Board.,
- Registration is done by country-based authorities, and at present does not include non-profit entities.

OrgRef

OrgRef (http://www.orgref.org/web/index.htm) compiles information from open sources - mainly Wikipedia, but also ISNI and VIAF - about universities, funders and other organizations involved in scholarly communications. At the time of writing there are 31k entries. The aim of the project “is not to be completely comprehensive, but to share information about the most significant organizations which are involved with academic content”.

The data is available to download under the same terms as Wikipedia - Creative Commons ShareAlike. However, as the dataset contains information derived from ISNI and VIAF, it is also restricted by the licences they use. The dataset also links to GRID. There are no APIs. Updates and suggestions are handled through email. It is unclear how often new versions are released, at the time of writing the latest is less than a month old.

DataSalon, a small commercial company, runs OrgRef and their business model is to offer services for disambiguation and matching to publisher data.

There are some perceived shortcomings of OrgRef:

- The data set is much smaller than other services have.
- There is no transparent governance.
- DataSalon is a purely commercial organization.
- There are no APIs available to access or update data.
- It does not handle alternative names for organisations.

Conclusion

This assessment of current organisation identifier provision, whilst not exhaustive, is based on a mix of publicly available information and direct conversations with providers. There may be other providers for which insufficient information was available for us to compare them to the scholarly communications community’s requirements. Others may come forward. In any case, it is clear that as things currently stand, work remains to be done in order to address the needs of our community.

The companion documents to this analysis will set out in detail the proposed form that this work can take. The perceived shortcomings identified in this document fit within two categories, issues of governance and openness, and functional needs (such as the ability for organisations to update their own records). However, there is a third requirement which must be addressed: community engagement. As we move towards the next steps in eliminating the shortcomings in current provision, we will widen and formalise participation in our discussions, to ensure that a representative section of the international community is actively involved in shaping the process of seeking a community-led, comprehensive organisation identifier solution for scholarly communications.