ORCID API Hack Day – 24th April 2018, London

In April developers and practitioners gathered at Jisc's London offices to explore the potential of the ORCID API and to look at how it could be used to deliver new value and functionality to universities and researchers.

This 'hackday' was an opportunity to discuss and explore ideas and experiment with a variety of APIs and data sources - a day where failure was as interesting as success and where those attending could be playful in their approach to the challenges on hand.

In advance of the day several use cases or areas for investigation were discussed leading to five documents:

- Delegating ORCID tokens workflows
- Sharing ORCID IDs from Institutional Repositories
- Gaining intelligence from ORCID and Sherpa
- Trends and Visualisations from ORCID Members reports
- Core and ORCID

These topics were meant to help inspire and focus work at the hackday, but not meant to limit what participants worked on - so if people arrived with their own ideas, or discussion on the day took the work in other directions, that was fine too.

On the day itself, after some brief introductions to the ORCID API and the use cases participants were left to talk and work together, and at the end of the day we all came back together to share what we had done (or tried to do) during the day.

Presentations

At the end of the day there were reports from participants on what they had looked at (either individually or as part of a group) during the day. There were 8 topics reported on altogether, which were:

Sharing ORCID IDs (presented by John Salter)

This group had been looking at a major barrier to the sharing of ORCID IDs between repositories and aggregation services (such as Core and IRUS-UK) - that there is no clear place to include an ORCID in the OAI-DC metadata schema that most OAI-PMH compliant repositories use to share metadata about publications.

This group came up with a proposal for a "dc-oa-plus" metadata schema, which would support the addition of IDs to any element in the OAI-DC schema.

Aggregators and harvesting services could 'prefer' the OAI-DC-plus schema where it is available from a repository, but continue to use OAI-DC if that is all the repository offers.

By trying to add support for IDs of all types, including ORCID IDs, with the smallest possible changes, it was hoped that support for the OAI-DC-Plus schema could be easily added to existing repository software such as EPrints and DSpace.

Defining API for related identifiers (presented by John Salter and Paul Needham)

One group at the hackday identified the need for a very lightweight and compact way of sharing information about relationships between identifiers. For example, if you have an ORCID ID in a system, you might want the ability to request a list of related DOIs.

The focus of this discussion was around the need for machine-to-machine communication that kept the payload (information shared) to a minimum by not including any human readable information (such as bibliographic metadata) but rather just sharing lists of IDs.

Once the IDs have been retrieved, the system can decide which ones it is interested in and where it needs to request further information about the ID (e.g. from ORCID, CrossRef, DataCite etc.)

The group looked at a number of existing schemas, but these were seen as over-engineered for the simple sharing of lists of IDs. The group wanted to look at existing schemes but felt that a very minimal service which could communicate the necessary data in a very compact way was required.

How good is getting ORCID IDs via RIOXX (presented by Matteo Cancellieri)

This was an attempt to measure and display the number of ORCID IDs supplied to the [Core aggregator](#). Firstly, the number of ORCID IDs gathered through the [RIOXX Metadata Application Profile](#) was charted, and it was found that the [White Rose Research Online](#) and The [Open University's Open Research Online](#) repositories were sending the most ORCID IDs through RIOXX. It was also noticeable that most, or possibly all, the repositories sharing ORCID IDs in this way were running the [eprints software](#).

In addition there was an attempt to search for ORCID IDs appearing in the full-text items across all the papers harvested by Core. This work wasn't completed during the hackday (because "doing a grep on 11 million documents takes a while!") but it might offer an alternative approach to obtaining ORCID IDs from institutional repositories who were not able to share their ORCID IDs via an appropriate metadata profile.

Nudges in the Open Access workflow (presented by Adam Field and Katie Shamash)

This idea was that APIs and data from multiple services (e.g. [ORCID](#), [Core](#), Sherpa, ...) could be used to help nudge items through an Open Access workflow by prompting the appropriate people to do things at an appropriate time via email. An example situation might be:

- Given an ORCID ID for a researcher, the ORCID API could be used to get information on publications and/or funders
    - Services such as CORE or oaDOI could be used to check that the item is available open access
    - Services such as Sherpa could be used to know what a funder requires and what the publisher/publication allows in terms of Open Access
- Given this information, an 'action' email could be created and sent to the researcher that nudges them to do something (such as upload a PDF to a repository) in order to move the item forward through the Open Access workflow - until the item is compliant with funder or institutional requirements

Event data (presented by John Salter and Joe Wass)

The idea of 'Event data' has come from [a collaboration between CrossRef and DataCite](#). In this context an 'event' is essentially a link between two things, and consists of an assertion such as "this tweet mentioned this DOI", together with other

data about the assertion - such as the provenance of the assertion, a timestamp, evidence for the assertion etc.

Currently the CrossRef Event data contains Events from the following sources:

Data source	Event type
Cambia Lens	Citations in Patents
Crossref Metadata	Links to DataCite registered content
DataCite Metadata	Links to Crossref registered content
Hypothes.is	Annotations in Hypothes.is
Newsfeed	Discussed in blogs and media
Reddit	Discussed on Reddit
Reddit Links	Discussed on sites linked to in subreddits
Stack Exchange Network	Discussed on StackExchange sites
Twitter	Mentions in tweets
Wikipedia	References on Wikipedia pages
Wordpress.com	Discussed on Wordpress.com sites

During the hackday there was some discussion about the possibility that Universities might serve as a source for Event data - for example the assertion that a researcher with a specific ORCID ID had authored a paper with a specific DOI could be treated as an Event, and might be the first link between that ORCID ID and that DOI.

Bristol Analytica (presented by Matt McGowan)

This was a fun and alternative way of looking at the ORCID data set, in this case using the option to access the [ORCID Public Data file](#), store it on [Amazon S3](#), and analyse it using [Apache Spark](#) running on an [Amazon EMR](#). Playing on the recent [Cambridge Analytica scandal](#), this piece of work set out to answer the question: "Could you influence an election result using data from ORCID?".

The controls and default settings used by ORCID to ensure researchers have full control over the data in their ORCID profile proved very effective, and so the answer turned out to "No!". However the ability to use Spark on a large EMR Cluster to interrogate a large data file in a short time was impressive, and the code written was able to find the most prolific authors in 2017 (in terms of publications in ORCID).

ORCID and reports for institutions (presented by Monica Duke, Owen Stephens and Oliver Lyttleton)

There were two strands of work that looked at the information or reports that institutions wanted to be able to access in relation to ORCID.

Firstly a specific piece of work looked at the regular "[member report](#)" institutions which are members of ORCID are able to get. This work looked at how these reports could be retrieved automatically and then visualised to help show trends over time and to have a way of exploring the data in an interative and visual way. This was achieved by using a single sample report, enhanced with some dummy data (as those working on the project didn't have direct access to the real reports). The data was accessed through a CSV file on a Google Drive (which is how the reports are made available to institutions), and then visualised using a Javascript API, with a focus on showing trends over time.

The second strand of work was a more general discussion about the sort of information that institutions want to see from ORCID. A common issue seems to be institutions often struggle to find out which (or how many) of their researchers are already using ORCID - this is a question that has been raised at previous ORCID events in the UK, including [the Jisc ORCID support workshop in June 2017](#). The ORCID member report may provide at least a partial solution to this perennial problem, although it cannot overcome the fact that if researchers do not share information on ORCID that links them to an institution, it is not possible to know which institutions, if any, they are affiliated with.

It was noted that the Swagger interface to the ORCID API (https://api.orcid.org/v2.0/) makes the API slightly more accessible, it still isn't an appropriate interface for users without technical skills and does not include any functionality such as the ability to download results.

Another suggestion in terms of information that might be interesting to be able to access is how many views there were of ORCID profiles on the website. While it might not be possible or appropriate to do this at the level of individuals, having information that showed that having an ORCID profile enhanced the visibility of your research online might be a powerful argument to present to researchers to persuade them to sign up to ORCID.

Delegating token workflows (presented by Neil Jeffries)

This was a problem that had been discussed in some detail before the hackday and described by Neil Jeffries in a series of blog posts:

- [Delegating ORCID tokens -background](#)
- [Revoking ORCID tokens](#)
- [Delegating tokens - the details](#)

The fundamental issue is that researchers may be asked to grant access to their ORCID profile to third party suppliers as part of their membership on an institution. This creates a complicated situation for all concerned, and raises concerns, especially in the light of [GDPR](#), that researchers may not be in a position to give informed consent to the use of their data across all the relevant systems and organisations, and that institutions may not be in a position to manage access to researcher data effectively.

This issue is a complex one, and not one that could be resolved at a hackday event. However, the day gave a chance to look at upcoming functionality from ORCID that might help either avoid or resolve the problem of delegating and revoking permissions in relation to third-parties.

The conclusion at the end of the day was that there was some potential in the approaches being explored and some confidence that both ORCID and UK ORCID consortium members were engaging with these challenges to find solutions.

Themes from the day

At least three themes emerged strongly on the day. One was the need to find ways of more effectively sharing and using IDs (of different types, but including ORCID IDs)

across systems and services. Another was the desire to improve the information available to institutions about the use of ORCID by their researchers and to make better use of that information. Finally there was an interest in how other services (and perhaps especially other Jisc services) might be used in conjuction with data from ORCID to offer better intelligence and improved workflows to both institutions and individual researchers.