Check for updates

OPINION ARTICLE

ORCID for funders: *Who's who* - and *what are they doing?* - ORCID IDs as identifiers for researchers and flexible article based classifications to understand the collective researcher portfolio [version 1; peer review: 1 not approved]

Christian Herzog, Giles Radford

ÜberResearch, Cologne, 50674, Germany

V1 First published: 20 May 2015, 4:122
<https://doi.org/10.12688/f1000research.6504.1>Latest published: 20 May 2015, 4:122
<https://doi.org/10.12688/f1000research.6504.1>

Abstract

For science funders, ORCID provides a persistent identifier that distinguishes one researcher from the others, and can facilitate workflows in grant submission, career tracking, and research impact(s). It makes life easier for the researcher – they can update their information in ORCID and make his/her past publications available to a funder as an ongoing service by just allowing this access as a one-time agreement. With these newly launched persistent tokens, researchers can grant a funder the right to update their grant record on ORCID once awarded – the metadata goes on an automatic roundtrip – effortless for the researcher, but the researcher stays in control, and can remove this right at any stage.

Having and sharing data is one aspect – but being able to understand true researcher activity is another – and even more challenging is to understand research activity in the aggregate. What are hundreds or thousands of researchers doing? Often a standard search will only answer or provide insights into a slice of the data. Research classification systems - like the **Fields of Research** (FOR) - provide sufficient aggregation, but these normally require manual tagging and curation of all the documents in a dataset. However, by using machine learning to automate tagging, it becomes possible to answer the 'what' question easily. This 'article-based classification' is realized using Natural Language Processing (NLP) technology.

With Dimensions, a portfolio analysis tool for research funders these capabilities are combined for research funders: allowing the researcher to provide controlled access to their ORCID profile and a solution environment for flexible article based classification, providing immediate access to analytical information on the researcher and institutional level – answering the questions 'who is who' and 'what are

Open Peer Review

Approval Status

1

version 1

20 May 2015

[view](#)

1. **Johanna R. McEntyre**, European Molecular
Biology Laboratory, European Bioinformatics
Institute (EMBL-EBI), Hinxton, UK

Any reports and responses or comments on the
article can be found at the end of the article.

they doing'?

Keywords

ORCID-CASRAI, Researcher Identification, Name ambiguity

This article is included in the **Research on Research, Policy & Culture** gateway.

ORCID

This article is included in the **Proceedings of the 2015 ORCID-Casrai Joint Conference** collection.

Corresponding author: Christian Herzog (christian@uberresearch.com)

Competing interests: Christian Herzog is the CEO and co-founder of ÜberResearch, a Digital Science Portfolio company, Giles Radford is employed by ÜberResearch.

Grant information: The author(s) declared that no grants were involved in supporting this work.

Copyright: © 2015 Herzog C and Radford G. This is an open access article distributed under the terms of the **Creative Commons Attribution License**, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. Data associated with the article are available under the terms of the **Creative Commons Zero "No rights reserved" data waiver** (CC0 1.0 Public domain dedication).

How to cite this article: Herzog C and Radford G. **ORCID for funders: Who's who - and what are they doing? - ORCID IDs as identifiers for researchers and flexible article based classifications to understand the collective researcher portfolio [version 1; peer review: 1 not approved]** F1000Research 2015, 4:122 <https://doi.org/10.12688/f1000research.6504.1>

First published: 20 May 2015, 4:122 <https://doi.org/10.12688/f1000research.6504.1>

Challenges on different levels - who, how and what?

Researcher identification has always been challenging for all research systems in a number of ways - one of them has been that researcher names are, obviously, not always unique. In addition they tend to express themselves in different permutations, especially over time. They may use a full first name, then a shortened first name, or just an initial only. And sometimes they might add a middle initial, sometimes not etc. This makes identifying the same person, over time, based on their name variations exceedingly difficult. Additionally, in an increasingly data-driven context of science funding and evaluation, the ability to attribute grants and publications to the individual researcher is in everyone's interest, and is now often required by employing institutions or funders. The answer of how to allow researchers to manage and share their research activities and outputs accurately is known as the ORCID ID system. Each researcher gets a unique identifier. It allows the individual researcher to manage their inputs and outputs in a single place and control and allow which organization/system can access it automatically (e.g. to provide the information in the context of a grant application without any effort other than providing the reading permit to the respective system - the rest then happens automatically). The question of how research activities and outputs can be related to a person is solved with the ORCID ID. It solves the 'who' question.

Driving adoption

However, obviously ORCID requires adoption in order for it to work as envisioned, which relates to the 'how' question. How can we make ORCID adoption universal? There are obviously a few gatekeepers in the research process with a critical role to play - primarily funders, publishers and research organizations. The benefits are obvious - having reliable information on researchers and

their related documents and activities is critical for funding decisions, saves costs and enables more specific support services.

Some areas have even more need than others. Organizations in Asian countries often have even greater challenges in this area of name ambiguity, due to a lot of synonymous names, and confusion of 'first' and 'last' name which often leads to data transposition. These countries are therefore able to profit from ORCID - allowing the researcher to distinguish their artifacts from others. The emergence of community and industry solutions like ORCID; CrossMark and FundRef provides a very cost efficient, standard based and effective implementation¹.

Funding organizations are key players, in that they are generally at the start of the research process - and in a unique position to drive the adoption of systems like ORCID. They can apply rules making it mandatory to have an ORCID ID. Creating an ORCID ID isn't difficult and we have already seen the ongoing benefits, like the ability to extract previous publication lists (required for the application process) from ORCID rather than the applicant manually submitting them, but there are downstream benefits too; for instance when other parties, such as publishers require the same type of information in the context of publishing the results of the funded research.

Requesting, or even insisting, that the researcher must have an ORCID ID is one thing - but how do the relevant publications, activities and grants get into the ORCID record? It still requires effort from the researcher to tag their publications and grants to their profile, and they have to remember to do this regularly, which is a burden and easily forgotten - so it needs to be made as simple as possible to reach two goals: completeness and quality of the data associated with an ORCID ID (Figure 1).

Figure 1. Workflow for assigning grants to an ORCID record using the ÜberWizard.

Making it easy for the researcher - the ÜberWizard for ORCID

One answer to the challenge of how to make things easier for the researcher is the ÜberWizard for ORCID. Developed by ÜberResearch when ORCID introduced funded grants as a data type, it allowed the researcher to add their grants from many different funders to their ORCID record in one simple step. The advantages are clear: the researcher can assign all their grants from participating funders in one step (and one wizard rather than searching for the right wizard per funder) and the data in the ORCID record is therefore correct and complete since it has been pulled from a consolidated grant database compiled by ÜberResearch and authenticated by the researcher.

Every funder can integrate their own grant portfolio into this database to save on the costs of developing their own routines of how to expose their funded grants for integration into ORCID records – but the more important aspect is to simplify the process for the researcher. ÜberResearch provides this service free of charge to support the integration ORCID identifiers in the funding workflow and to support the researcher. In addition funding organizations can make use of the global grant database for portfolio comparison and analysis purposes (The global award database can be analysed with Dimensions for Funders <http://www.uberresearch.com/dimensions-for-funders/>, which is available at no cost for small funders, see <http://www.uberresearch.com/ubershare/>).

Roundtripping of metadata - the next level of integration

However, this interaction model with the ÜberWizard still asks a lot from the researcher – they have to use the ÜberWizard to bring the

data into their ORCID record. But with new functionality launched by ORCID in 2015 (<http://orcid.org/blog/2014/11/21/new-functionality-friday-auto-update-your-orcid-record>), this can be made much easier – based on trusted relations and automation (Figure 2). The grant application is normally the first step in a research cycle, and with the functionality of a ‘long-lived token’ the funding organization can request permission from the applicant to read and update their ORCID record: This permission starts to send the metadata on a round trip. ORCID starts to become a (hidden) infrastructure working for the researcher; an awarded grant or related information can be pushed automatically into the researcher’s ORCID record once the grant appears in the global grant database fueling the ÜberWizard for ORCID (Figure 3).

How funders can push for adoption- the FCT drives national adoption of ORCID IDs in Portugal

The Portuguese Fundação para a Ciência e a Tecnologia (FCT) made it mandatory for funded researchers to have an ORCID ID. This obviously results in a high adoption rate and becomes a *de facto* national roll out – making it far easier for the researcher to share their inputs and outputs going forward, and enabling the national Portuguese funder oversight on all Portuguese researchers’ activity. In a recent research assessment exercise, 15,000 researchers registered their ORCID ID and about 10% also added funded projects to their record, using the ÜberWizard for ORCID. This is expected to increase during 2015 when scholarship grants are included and the connection to the national CV system is realized (Personal communication with João Moreira, FCT).

Figure 2. Giving permission to read and write into the ORCID record.

Figure 3. Workflow of assigning grants to an ORCID record using persistent tokens and the ÜberWizard.

Understanding the research activities – ‘the what?’

With the ORCID ID in place the relation between input/output and researcher is in place – the ‘who’ question is solved. But this is still on the metadata level and does not create insights into ‘what’ the researcher or an entire population of researchers is doing. What would be required is an identifier system to tag the content of the documents – preferably automatically. It is clear that the use case is not to understand every article or publication, but to be able to cluster large numbers of documents in high-level categories in order to understand distribution across research topics, disciplines and trends. This is currently done in some areas with journal classifications where subject categories are assigned on the journal level. This works as expected – quite well for highly specialized journals, but not at all well for multidisciplinary journals, and it is not possible to have the same classification for non-journal documents².

However, given that the content is available in the document itself, why not taken the approach of deriving the ‘tags’ or classifications from the document itself? Which classification systems could be used?

Research classification systems and the automatic assignment of categories

Research classification systems are used for structuring and simplifying portfolios for use cases like trend analysis, reporting and strategic decision making. The systems can span the entire science portfolio, like the Fields of Research (FOR) codes as part of the [Australian and New Zealand Standard Research Classification](#) (ANZSRC) system, the [OECD Frascati classification of science and technology](#)

(FOS) or discipline specific like the [Research, Condition and Disease Categorization](#) (RCDC) system used by the National Institutes of Health.

Some of these systems are in use in some countries with some funders, meaning a small subset of grants can be interrogated by a small subset of classification systems. Most have been assigned manually to documents, but some have been assigned using semantic routines (e.g. the RCDC system, albeit only on NIH grants). Based on the use case of being able to get insights on the portfolio level in large document databases without the unmanageable burden of the manual effort in reading and classifying all the documents, ÜberResearch started to work with funding organizations, as development partners, to develop the routines and tools to be able to assign various classification systems to document databases – for example the FOR coding system from Australia/New Zealand. Using machine learning approaches and a large dataset of manually coded documents as a training set, we were able to derive a model which can now be applied on a document level to any document – achieving a consistent ‘tagging’ without the bias normally introduced by different human coders or professional groups. In addition to the FOR codes, Dimensions has automated the RCDC classification system used by the NIH; the health categories of the [Health Research Classification System](#) (HRCS) and a first implementation of the [Common Scientific Outline](#) (CSO) coding. The approach used to derive the model using machine learning routines will be discussed in a separate paper once the evaluation of several of the classification systems has been concluded.

Implementation of a 'babel fish' of research classifications

As a result of these efforts, it is possible to assign different research classification systems automatically to all documents – grants, publications, and others - at marginal costs, which allows funders or research organizations to use different ones for different purposes. The research classification approach is implemented in ÜberResearch's Dimensions – together with the corresponding analytical functionalities. Dimensions has been developed to serve as an 'applied babel fish' system (see 3) for research classifications.

Examples of the classification results associated with use cases

The examples below (see Figure 4 and Figure 5) have been taken from ÜberResearch's analytical tool Dimensions, analyzing a global grant database covering more than 1.4 million grants with a total funding volume of more than \$760 billion US. The examples show how article-based classifications can be surfaced for end users in an application to provide strategic insights on the global funding landscape. These approaches can also be realized in other tools – they should be seen as illustrations and examples.

Figure 4. Insights in a population. The screenshot shows all grants of the European Research Council (ERC) and the most funded organizations –together with the FOR codes receiving the highest funding amounts. It allows a user to understand quickly the focus of a group of research institutions, especially if looked at over time. Screenshot taken from ÜberResearch's application Dimensions for Funders.

Researcher
Stephen O'Rahilly
University of Cambridge Gb

Last 10 years

Figure 5. Person activity overview. The aggregated profile for a researcher (in this case only based on grants) shows on the right hand a profile across different classification systems – indicating clearly the different use case: FOR codes are relatively general, the RCDC systems provides as expected more details and the HRCS system adds a broad classification in areas – for what it has been built. Screenshot taken from ÜberResearch's application Dimensions for Funders.

Conclusions

With the open researcher and contributor ID, coupled with the corresponding infrastructure, a power approach has been established, which is constantly refined to make it easier for the researcher. This ‘hidden’ piece of infrastructure is ‘doing the right things automatically’ – while at the same time keeping the researcher in the driving seat in terms of who sees and gets access to his or her data.

This will help finally solve the challenge of name ambiguity and lack of links between inputs and outputs of researchers, whilst also removing much of the burden from the individual researcher. But it requires adoption and that is only possible with incentives or some ‘light’ pressure – by, for example, funding organisations making it mandatory (even if that feels, initially, like additional effort) – but it will pay off for the researcher downstream: for example when he

or she submits their next manuscript, applies for their next grant or moves between organizations. Publishers are in a similar gatekeeper role for strengthening the ORCID approach – creating the scenario that researchers require an ORCID ID. Researchers, too, can immediately see the benefit: having ones grants and publications underrepresented can be both frustrating and, potentially, damaging. A full and complete record will help in all that they want to do.

The ‘who’ question is increasingly solved by ORCID – and with the right incentives, and more and more gatekeepers adopting it – it will solve the challenge of knowing the relations between works and the individual. But what about the second approach of establishing flexible identifiers for the ‘substance’ or content of the research activities which are generated from the works directly using computational routines?

Again science funders have a critical role to play here – since they are driving most of the use cases: portfolio classification and reporting on how research funds have been distributed and the analysis of the input, output and impact are at the core of their mission. To know what has been funded in any given research topic can drive effective strategic decisions.

Such use cases - generating or assigning classifications based on the content of the documents -enable comparisons and interactions between funders and research organizations. To know how much

has been funded in a given research topic area requires a conversation using the same classification language, applied in the same consistent way. The approach is still in its infancy, since it takes time to replace established routines (primarily manual tagging of documents), but the increase in attention to the field and approach hint at a near future, where the classification routines are part of a hidden and effective infrastructure, like the ORCID system. That would solve the ‘what’ question. And if both ‘who’ and ‘what’ can be solved using (mostly) automatic routines then the research funding landscape overview can help drive funding policy decisions based on reliable data, which has to be a good thing for science in general.

Author contributions

Both authors have equally contributed to the article. Both authors have seen and agreed to the final content of the manuscript.

Competing interests

Christian Herzog is the CEO and co-founder of ÜberResearch, a Digital Science Portfolio company, Giles Radford is employed by ÜberResearch.

Grant information

The author(s) declared that no grants were involved in supporting this work.

References

1. Huh S: **Application of new information technologies to scholarly journals: ORCID, CrossMark, and FundRef.** *J Korean Med Assoc.* 2014; **57**(5): 455–462.
[Publisher Full Text](#)
2. Glänzel W, Schubert A, Czerwon HJ, *et al.*: **An item-by-item subject classification of papers published in multidisciplinary and general journals using reference analysis.** *Scientometrics.* 1999; **44**(3): 427–439.
[Publisher Full Text](#)
3. Terry RF, Allen L, Gardner CA, *et al.*: **Mapping global health research investments, time for new thinking—a Babel Fish for research data.** *Health Res Policy Syst.* 2012; **10**: 28.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)

Open Peer Review

Current Peer Review Status:

Version 1

Reviewer Report 16 June 2015

<https://doi.org/10.5256/f1000research.6980.r9075>

© 2015 McEntyre J. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Johanna R. McEntyre

European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Hinxton, UK

This article is about two tools implemented within UberResearch, a small company operating as part of the Digital Science portfolio. The first is a tool that allows researchers to add funding awards to their ORCID record (the UberWizard for ORCID); the second is a tool for funders ("Dimensions") that performs analytics on funding decisions using research classification systems. A machine-learning method has been used to tag documents (grant awards) with various categories of the research classification systems. The two come together as grants claimed by an individual researcher can be tagged and therefore categorized on an individual basis as well as for funders.

I wish the authors could have been clearer on this from the outset. I think this context would help orient readers better. Much of the abstract focuses on the details of funders being able to automatically update a researcher's ORCID with funding information and an abstract discussion on automatic tagging of research classification systems. The writing and narrative style could be improved - some sentences extremely long e.g. p3. "Creating an ORCID D (sic) isn't difficult funded research."

However the major comment I have on the article is that it describes outcomes without describing how they were achieved, and furthermore there is no way for readers to evaluate Dimensions as a product. More specifically:

1. Did UberResearch need to be a member of ORCID to provide the grant linking service?
2. How has the "global grant database" been generated - from where? What is its scope?
3. What protocols and standards are used to make the transactions between ORCID and the UberWizard? Are there any other tools available?
4. What methods of machine learning were used, how were they validated? The given info is: "The approach used to derive the model using machine learning routines will be discussed

in a separate paper once the evaluation of several of the classification systems has been concluded." I would be happier if this were published already.

5. As a reviewer or a reader, I can't access Dimensions so I have no means to evaluate it.

6. The article seems to be aimed primarily at funders, as is the Dimensions product - what is in it for the researcher, who must be the primary readership of F1000Research?

I have personally used the UberWizard for ORCID and it worked fine; I also saw a brief preview of Dimensions a few months ago, and was impressed. I see this article is an "Opinion" article - but nevertheless it is tough to evaluate when the methods and outcomes are not publicly available. I would agree that ORCIDs are important for effective research assessment, but it would be a more powerful position if there were some means to evaluate Dimensions directly. Would a product review be a more appropriate route?

Competing Interests: No competing interests were disclosed.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to state that I do not consider it to be of an acceptable scientific standard, for reasons outlined above.

The benefits of publishing with F1000Research:

- Your article is published within days, with no editorial bias
- You can publish traditional articles, null/negative results, case reports, data notes and more
- The peer review process is transparent and collaborative
- Your article is indexed in PubMed after passing peer review
- Dedicated customer support at every stage

For pre-submission enquiries, contact research@f1000.com

F1000Research