Structure of the ORCID Identifier

Geoffrey Bilder

Version History

V1	2010-04-16	
V2	2010-04-17	
V3	2010-04-21	
V4	2011-11-03	Slight edits to reflect phase 1 status. Update to recommend ISO 7064 MOD 11-2 checksums.
V5	2011-11-09	Reflect changes recommended by TAG from meeting of 2011-12-07
V6	2012-03-30	Reflect agreement to initially limit ORCID assignment to a block which will be guaranteed to not collide with ISNIs.
V7	2012-08-02	Make it clearer that the the conclusions and appendix of this document represent the definition of the ORCID identifier structure. The rest is just background and rationale for the final recommendation.

*[**NOTE**: a summary of the structure of the ORCID Identifier also can be found on the ORCID Developers Portal: http://dev.orcid.org/structure-orcid-identifier]*

Introduction

ORCID is in preparing to launch and ORCID system based on technology and code that was originally donated to the ORCID effort by Thomson Reuters. The donated code had several dependencies on proprietary back-end systems and was also missing several features that ORCID has identified as being essential to the launch of a successful system.

Current development work consists of replacing proprietary dependencies and implementing these missing features.

One of the proprietary dependencies that has been replaced is the structure of the actual ORCID identifier used for profiles. This paper discusses this original Researcher ID identifier structure and explains how and why this identifier structure has been changed in the ORCID system.

If you just want to know what the final, defined ORCID identifier structure is, we recommend that you skip to the Conclusions section of this paper and look at the examples in Appendix A. On the other hand, If you want to understand the rationale for the recommendations, read on.

Initial Status

The software donated by TR for development of ORCID was configured to use the identifier scheme used in the Researcher ID system.

The current Researcher ID identifier is of the format AAA-NNNN-YYYY where:

AAA = Alpha Character values from A-Z. The value is in uppercase and ranges A - ZZZ
NNNN = Numbers from 0-9. The value ranges are between 1001 - 9999
YYYY = The year in which the identifier is minted.

The maximum length of the identifier is 13 characters (ZZZ- 9999-2010) and minimum length is 11 characters. (A-1001-2010). [1]

Thus, a typical Researcher ID identifier currently looks like this:

`A-4031-2008`

Issues with the current Researcher ID identifier structure

Some ORCID participants have expressed concerns about using the current Researcher ID identifier as the ORCID ID structure. These concerns span policy, usability and technical issues.

First, there is a general policy concern that, in using the Researcher ID identifier as the native ORCID identifier, we will:

- Appear to be privileging Thomson Reuters identifiers over other third party identifiers that are registered and mapped via the ORCID switchboard. This would, in turn, be perceived as undermining the independence of the ORCID system and the purported equality of the ORCID partners.
- Make it more difficult for the Researcher ID and ORCID systems to evolve in different directions should that become necessary or desirable.
- Make it more difficult to determine what an ORCID identifier "resolves to". Does it resolve to an ORCID profile or to a Researcher ID profile?

[1] http://sites.google.com/site/openrid/technical-working-group/researcherid-documentation

In short by using the Researcher ID identifier for ORCID, we might misleadingly conflate the two systems.

Second, there are a number of technical concerns that have been raised:

a. Running out of identifiers. With the fixed structure of the current identifier system, it is possible that we could exhaust the number of identifiers that can be issued in a year. This is particularly true if we have many partners submitting multiple, duplicate records for current authors and if we have significant deposits of records for inactive authors.
b. Semantic overloading. The current structure of the Researcher ID identifier has several points where users could (mistakenly or otherwise) assume semantic meaning to parts of the identifier. It seems to be a rule that the more semantic information is contained in an identifier, the more brittle it becomes. This is potentially a serious problem for the ORCID identifier, which is supposed to be a persistent, long-term identifier for the scholarly record. We will discuss this in further detail below.

Semantics in persistent identifiers

The the most recognized problem with embedding semantic information in an identifier is that semantic information changes over time. Even the most apparently permanent properties of objects change over time. Personal names change, organization names change, even country names and boundaries change. Indeed, the longer the time scale, the more likely some semantic information is going to change.

And this problem can even apply to semantic information that is:

a) Inadvertently or indirectly incorporated into an identifier.
b) Mistakenly “projected” onto a semantic identifier.

For example, people get attached to seemingly abstract data such as phone number area codes or license plate sequences. This attachment, in turn, makes it much harder for governments and organizations to update or changes these numbering sequences.

Similarly, in the publishing industry, CrossRef has had publishers imbue the relatively arbitrary four digit DOI prefix with “branding” significance and, as a consequence, has had publishers request that CrossRef change the DOIs for existing publications that have been transferred from one publisher to another.

The problem with the current structure of the Researcher ID identifier is that semantic meaning can be mistakenly projected onto the identifier in at least three different ways.

First, the three-letter sequence at the start of the identifier could conceivably increment to spell out words or acronyms. What happens when a researcher legitimately decides that they want to change their ORCID ID because starts with KKK or ASS or WTF? One could, conceivably write code to prevent the assignment of IDs with *known* acronyms and words- but there is no way to write a filter that flags potential future acronyms/words that might cause offense or concern.

Second, current Researcher IDs are generated in sequence. This means that, in the case of a bulk-load from an institution, it might appear that the institution has a designated

“prefix” or “name space”. For instance, users might mistakenly conclude that all Brown University faculty ORCIDs fall within a certain range.

Finally, as Joel Plotkin has already pointed out, by including the year of minting in the identifier, we would be indirectly revealing the ‘age’ of the researcher associated with the profile. This, in turn, could indirectly enable age discrimination and thus run afoul of various privacy jurisdictions.

In general, the only way to avoid these kinds of inadvertent or indirect assignations of semantic meaning to identifiers is to make them as opaque as possible.

How opaque is opaque?

The easiest way to generate mostly opaque identifiers is to make them entirely numeric; however, even with this approach, there are two problems that we might need to try to address:

a. The aforementioned problems with people potentially interpreting sequences as name spaces. In other words, the identifiers would have to be assigned out-of-sequence.
b. People assigning semantic significance to the numbers themselves. For instance, users in Asia might be reluctant to accept identifiers with too many “fours” in them. Westerners may eschew numbers with thirteen or “nine-eleven,” etc. Normally, this might seem like a pretty eclectic and edge-case concern, but given that these identifiers are going to be assigned to people (not things) and, given that not all of these people are going to be scientists, it probably merits some thought as to how we can work around these concerns. One option would be to give a researcher who is “claiming” a profile to “pick” an identifier from a series of generated alternatives.

Opaque Identifiers and usability

Some common objections to opaque identifiers and, particularly, to numeric identifiers are that they are hard to recognize, remember, transcribe, etc. The following discusses various ways in which these concerns could potentially be addressed.

Segmentation

ORCID could follow the lead of the ISNI identifier and simply recommend that the identifier be segmented into four-digit groups. For instance:

1422 4586 3573 0476

This is common practice and would make transcription, etc. slightly easier.

The problem with using spaces as a delimiter is that they will make it more difficult to use the ORCID identifier in web applications because the spaces will have to be encoded. A more web-friendly approach would be to delimit the number with dashes. For example:

1422-4586-3573-0476

Length

Although we should probably start with an identifier length that guarantees us a range that will last a few generations, there is probably little reason to adopt a fixed length for the identifier. The major argument for fixed-length identifiers is that it makes it easier to format them, but this can as easily be done using padding. For example:

0001-4586-3573-0476

Checksums

A common approach to making identifiers more transcribable is to include a check-digit at the end of the identifier. The checksum allows one to quickly detect a variety of transcription errors including transposition, single digit errors, etc. The ISSN, ISBN, ISNI, etc. all use check digits.

The ISNI uses the ISO7064 mod 11,2 algorithm for generating check digits. This seems to combine both efficiency with the ability to check most common transcription errors.

Affordance

The issue of "recognizing" an identifier is important. Ideally, people seeing an ORCID ID should be able to quickly recognize it as such. A number of identifiers (DOI, ISNI, PMIDs) make themselves "recognizable" with a prefix. For example, the DOI uses a pseudo-scheme syntax as follows.

doi:5555/12345566

One problem with this approach is that, with web applications, it is not immediately clear that one can make said identifier actionable by prepending it with http://dx.doi.org.

Indeed, CrossRef has recently revised its display guidelines to recommend that CrossRef DOIs be displayed as HTTP URIs.[2]

It seems that we could address the issue of affordance **and** make the ORCID identifier more in keeping with web architecture and linked data principles[3] if we simply said that the ORCID should be prepended with http://orcid.org/. Combined with the above recommendations on segmentation, this would mean and ORCID identifier would look like this:

http://orcid.org/1422458635730476

or like this:

http://orcid.org/1422-4586-3573-0476

[2] http://bit.ly/rkN1lz
[3] http://en.wikipedia.org/wiki/Linked_Data

Shortening identifiers

URL shortening services work by encoding a normal integer into a non-decimal base (base 62, often). This means that the integer can be expressed with a combination of letters and numbers. For example, using a simple base 62 encoding, the following identifier (including a checksum):

1422458635730476

Could be expressed as:

6VvBHUnrI

Again, if combined with the above recommendations on segmentation, affordance and checksums, a "shortened" ORCID identifier could be expressed like this:

http://6VvBHUnrI

or like this:

http://6Vv-BHU-nrI

Note that one problem with the URL shortening approach is that you again introduce the possibility of inadvertently introducing inappropriate words/acronyms along with undesirable numbers. For instance, who knows if anybody might take exception to the above identifier containing the initials for "Banaras Hindu University"?

Still, the question of whether to shorten or not does not need to be binary. We can follow the lead of the IDF and support both normal and shortened versions of the ORCID identifier (ORCLETs?) simultaneously as is done with shortDOI[4].

Ultimately, it is unclear if there is sufficient benefit to providing alternate, shortened forms of the ORCID. First of all, they are not all that much shorter than the long forms. Secondly, short versions might introduce unwanted character combinations which might cause offence and drive researchers to want to change their identifier.

Should ORCID use the DOI?

First it should be noted that, whether ORCID uses the DOI or not, we would still need to address most of the above issues. The IDF is fairly open-ended in allowing registration agencies to establish policies about how RA-specific (RA = Registration Agency) DOIs are generated and what they look like. An RA like CrossRef devolves DOI minting responsibility to its members and they, in turn, have adapted a wide variety of approaches to generating them (some of which have caused CrossRef big headaches). Other RA's may choose to generate "opaque" DOIs on behalf of their members using any of the techniques above. For instance, Bowker's "Actionable ISBNs" are simply DOIs comprised

[4] http://shortdoi.org/

of ISBNs.

Having said that, it is not clear that the DOI is a natural fit for ORCID. Although the word “switchboard” is used to describe both CrossRef and ORCID, there are some fundamental differences between way in which the two work.

For one thing, CrossRef has no content “of its own”. In this respect CrossRef is a “pure” switchboard. In merely serves as a level of link indirection that helps people find CrossRef members’ content even if it has moved. Even when CrossRef does “serve content” (as in its metadata services), it only does so as a proxy for its members.

ORCID certainly has “switchboard-like” functionality- that is, it will include the ability to map one identifier to another. But, crucially, ORCID will also have data of its own. In other words, ORCID is an actual destination. “Resolving an ORCID ID” means that you will be taken to an ORCID profile. From the ORCID profile, it might be possible to retrieve and resolve other identifiers (SCOPUS, MIT, etc.)- but an ORCID ID points at ORCID content. A CrossRef DOI does not point at CrossRef content.

Now, clearly, CrossRef is not the only DOI RA, and other RAs might choose to implement DOIs differently, but it remains the fact that, the fundamental ability of the DOI to map the identifier to a URL that might change is not a major use-case for ORCID. ORCID is establishing itself as a long-term persistent curator of researcher profiles. A publisher’s DOIs might eventually change and map to another publisher due to an acquisition, etc.- but and ORCID ID should always point to an ORCID profile on an ORCID-controlled domain.

Should we use the ISNI?

The ORCID board has summarized the relationship between ORCID and ISNI in the ORCID FAQ.

Having said this, ISNI identifier clearly addresses some of the identifier structure issues discussed above. That is, the ISNI identifier itself is opaque and has a check-sum. ORCID would probably do well to emulate ISNI in this regard.

Still, in order to leave open the possibility of closer ORCID/ISNI collaboration, both organizations have agreed that, at least initially, ORCID will issue identifier from a reserved block of 20 million identifiers which will not collide with any assigned ISNIs.

Conclusions

Given the above observations, we would conclude that the ORCID identifier:

- should be separate from the current Researcher ID identifier
- should be a number
- should be issued out-of-sequence
- should have a check sum
- should not be shortened or support a shortened alternative
- should be expressed as a HTTP URI
- shouldn't be a DOI

- shouldn't be an ISNI
- should initially be assigned from a reserved block agreed to by ORCID and ISNI in order to accommodate the possibility of closer future collaboration. The agreed block is 15000000 through 35000000.

Appendix A: Some sample, fake ORCIDs.

The following randomly generated example ORCIDs with checksums generated in accordance with ISO/IEC 7064:2003, MOD 11-2. The examples below are from the agreed block of reserved numbers designed to avoid collision with ISNIs.

ORCID
http://orcid.org/0000-0002-3843-3472
http://orcid.org/0000-0001-7051-1197
http://orcid.org/0000-0002-8205-121X