Interview with Geoffrey Bilder

Martin Fenner, Gobbledygook

February 17, 2009

Almost exactly two years ago, **CrossRef** invited a number of people to discuss unique identifiers for researchers (**CrossRef Author ID meeting**). One year ago Thomson Reuters launched **ResearcherID** (**Thomson Scientific launches ResearcherID to uniquely identify authors**). And two months ago Phil Bourne and Lynn Fink wrote about this topic in a **PLoS Computational Biology** paper (**I Am Not a Scientist, I Am a Number**).

So it comes as no surprise that we also talked about author identifiers at the recent **ScienceOnline09** meeting in North Carolina (both in the session **Impact Factors and researcher incentives** and over drinks afterwards). **Cameron Neylon** wrote down his thoughts after the meeting in a blog post (**A specialist OpenID service to provide unique researcher IDs?**), and this resulted in **a very interesting** discussion on **FriendFeed**. And at about the same time both **Jan Aerts** (**Who-o-o are you? Who who? Who who?**) and Christopher Leonard (**Some thoughts on unique author IDs**) independently wrote blog posts about the same topic. This week **Cameron Neylon** summarized the discussion in another blog post (**Contributor IDs – an attempt to aggregate and integrate**).

Science bloggers have put a lot of thought into the idea of a unique author identifier and I've collected more reading material about author ID at **Connotea** using the tag **authorid**. But I was also very curious to learn more about the work that has already been done. That's why I asked **Geoffrey Bilder** from **CrossRef** a few questions. In the end Geoffrey talked not only about author identifiers, but also about CrossRef, DOIs and many other aspects of scholarly publishing.

1. Can you describe what CrossRef is and does?

Let me start with what it does because this is a little less likely to make your eyes glaze-over.

CrossRef was originally founded by scholarly publishers to fight **link-rot**.

Web links have a half-life of about six years. That is, after six years a link is likely to break because the content that it pointed to has been moved. To a

lay-audience this might be a mere annoyance, but to scholarly and professional publishers broken links are anathema. The scholarly record is built on a foundation of links in the form of citations. If these online citation links break, the online scholarly citation record breaks.

But surely fighting link-rot should be simple, right? After all, the glory of the web is its decentralized architecture, one in which the domain name that you own can be used as a namespace for identifiers. **Tim Berners-Lee** has said that "cool URIs never die". Aren't "persistent URIs" merely a matter of being disciplined in the way that you mint URIs and in being conscientious about sensibly redirecting URIs when things change location on your web server?

Well, certainly the majority of broken links on the web are the result of careless web administrators not taking the time to structure and redirect their web site's URIs properly, but there are a significant percentage of links that will break despite the best efforts of webmasters. This is because in some cases the domain name in the link will change, and in these cases the whole "domain name as URI minting name-space" starts to crumble. When otherwise sensible technorati refer to "owning" a domain name, it makes me want to stick forks in my eyeballs. We do not "own" domain names. At best, we only lease them and there are manifold ways in which we could lose control of a domain name – through litigation, through forgetfulness, through poverty, through voluntary transfer, etc. Once you don't control a domain name anymore, then you can't control your domain-name-based persistent identifiers either.

Incidentally, another technorati meme that makes me want to self-harm with cutlery is the notion that "persistent identifiers don't matter in the age of the search engine. If a link breaks, we can just find the content again wherever it has moved." This, naturally, is the self-serving argument often used by Google and it's starry-eyed acolytes. Even just few minutes' reflection reveals the gaping hole in this approach.

The hole is this – how do you cite a specific copy of something if there are multiple **almost identical** copies of it located in different places? For instance, lets say there are 2 copies (X and Y) of an article which only differ in a few paragraphs, but those few paragraphs are crucial and likely to change the reader's interpretation of the work. How, in a world where persistent linking is maintained by search engines, do you create a citation link to article X instead of article Y? To create a search-based link that is more likely to resolve to article X, you would essentially have to encode the entire article in the search URI! Even this wouldn't guarantee you that the link directed a future reader to article X first; it is still possible that article Y might end up getting a higher ranking because more people have linked to it and it therefore has a higher page-rank (or equivalent). Believe me, even variations of the search engine scenario (using document hashes for citations, pingbacks, etc.) quickly unravel after a little reflection.

This is all a long-winded and ranty way of saying that the issue of persistent identifiers on the web is just a wee bit more complex than most people think.

So how does CrossRef address the persistent identifier issue?

From our perspective, the persistent identifier problem is much more a social problem than a technical one.

In fact, the technical part of our service is relatively straightforward. CrossRef provides a level of indirection (i.e. a pointer) between an identifier and a URL. When publishers put something online, they assign a CrossRef **Digital Object Identifier** (DOI) to it and submit a record for that item with CrossRef. The record includes the CrossRef DOI, basic bibliographic metadata for the content and a URL that points to the current location of the content. People citing the publisher's content are encouraged to use the CrossRef DOI for the citation instead of the publisher's URL. When a researcher clicks on a CrossRef DOI, the CrossRef service redirects the URL to whatever URL the publisher has currently registered for that CrossRef DOI. This means that the publisher can update their CrossRef DOI record to point to a completely new URI (including a new domain name) and any CrossRef DOI citations will continue to work. We provide a few other services based on this infrastructure too. So, for instance, we can resolve an **OpenURL** to a CrossRef DOI (by querying the publisher-submitted bibliographic metadata), resolve a free-text query to a CrossRef DOI and we can also return bibliographic metadata instead of redirecting if that is what the user wants.

So-far, so good, but this isn't anything that couldn't be accomplished using other redirection tools such as **PURLs**, **CNRI Handles**, **XRIs**, **NUmly Numbers**, etc.? The crucial question to ask of any such service is, "what guarantees that the publisher will actually update their URL pointers?" If the publisher doesn't update these pointers, then the links will break anyway. It isn't enough that a publisher decides to use PURLs, if they then don't update their PURLs- in perpetuity.

This is where it is important to explain the organizational structure and the social effect that this has on the service.

CrossRef was founded as a non-profit, membership organization for publishers. Note that we are entirely catholic in our definition of what a publisher is, so our membership includes commercial publishers, non-profit publishers, open access publishers, institutional repositories, NGOs and IGOs, Video publishers, Wiki-based publishers, etc. We are also open to publishers of all disciplines (humanities, social sciences, sciences, professional), geographies and content types (journals, books, database records, videos, etc.)

In practice, what unites our membership is a concern that their content should be considered worthy of trust by professional researchers. One way in which researchers assess the trustworthiness of content is by determining how it sits within the scholarly record. Does it provide evidence for its assertions in citations? Do other people cite it?

When a publisher joins CrossRef, they agree to a **set of enforceable terms**

and conditions that govern the way in which they use CrossRef's persistent citation infrastructure. Specifically, they agree to:

- Register DOIs within a week of something being published online
- Update the URLs associated with a DOI when said URLs change
- Link citations in their content via the DOI

In joining CrossRef they also agree that CrossRef can fine them or throw them out of the service if they do not meet the terms and conditions of the service. Note that the penalty of being thrown out can be quite severe as it effectively means that the publisher would become invisible in the online scholarly citation record. In short, the system has a built-in social feedback loop that strongly enforces good citizenship.

As I said, the technical infrastructure of CrossRef is pretty mundane, and it is the social aspect of the service that does the most to guarantee the persistency of CrossRef citation links.

2. What are your responsibilities within CrossRef?

Thinking of, gathering the requirements for, designing and (most importantly) launching new services.

Last year we launched a plagiarism detection service called **CrossCheck**. This year I am working on Contributor ID, another project tentatively named Cross-Mark and a bunch of smaller projects designed to encourage the use of DOIs in citations.

3. What did you do before starting to work for CrossRef?

In the early nineties Allen Renear and I co-founded Brown University's Scholarly Technology Group, where we were charged with providing advanced consulting and support to Brown's research community. In the mid-nineties I grew tired of the politics, resource constraints and institutional paralysis that seems to grip so many universities and I decided to do something as far away from the academic sphere as possible. In short, I worked at a management consultancy doing R&D for their IT group. In 2000 I was lured into managing the web development efforts for an Information Architecture firm called Dynamic Diagrams. In 2001 we were bought by **Ingenta** in the UK. I became Ingenta's CTO and I moved to Oxford in 2002. I left Ingenta in 2005, did a brief spell of consulting for publishers in 2006 and joined CrossRef in 2007.

4. What are your thoughts on how an author identifier should look like?

First of all, I think we need to stop talking about "author" identifiers. One of the first requirements we found when interviewing publishers, researchers and librarians is that we would ideally like to be able to identify any party who

contributes to the scholarly literature in any way. That is, we would also like to be able to identify reviewers, editors, correspondents, bloggers, commenters, etc. This is why we have taken to calling our project the "CrossRef Contributor ID" project. This isn't just playing with words either. For instance, as soon as you start thinking about things like "how do you accommodate reviewers" in this system you need to think of things like pseudo-anonymity. That is, you want somebody to be able to get credit for doing reviews in a way that doesn't necessarily reveal who reviewed what. In turn, the pay-off for designing a system whereby anonymous reviewers might be credited with reviews could be profound. It might ultimately result in researchers having much more incentive to review if reviewing were something that could be counted and rated in the same way that authorship is.

Second, I think that people conflate a lot of issues when they talk about "author identifiers" [sic]. Are they talking about the simple token used (e.g. a unique string or a number assigned to an individual like a social security number), are they talking about an authentication mechanism (e.g. **OpenID**, **Shibboleth**) or are they talking about the profile information associated with an identifier (e.g. publications, affiliation, contact info, etc.)? Obviously, these all overlap in some ways, but how they relate and what you choose to focus on depends largely on your use cases.

Third, speaking of use cases, our requirements gathering has identified two broad categories of use cases that, though related, have profoundly different implementation implications. One category of use cases identified revolves around "knowledge discovery" and the other category of cases revolves around "authentication."

The "knowledge discovery" use cases are probably the most obvious things that people would like to be able to do with a contributor ID such as:

- Determine what IDs authored/edited/reviewed document X
- What documents where authored/edited/reviewed by ID Y
- What IDs are related to ID Z and what is the nature of that relationship (e.g. co-authored, edited, reviewed)
- What (subject to privacy settings) is the profile information for ID Z (e.g. institutional affiliation, email address, etc.)
- All the author IDs and their respective publications where the institutional affiliation recorded by the author is X
- Etc.

At this point I feel obliged to point out that the bulk of our requirements gathering has been focused on trying to understand the needs of our member publishers. The reason I mention this here is that the bulk of the "authentication" use cases that we identified are all focused around making publisher back-office systems less cumbersome. So, for instance, publishers are interested in using a "contributor id" for:

- **single sign-on** (SSO) for manuscript tracking systems

- Disambiguating contact information for use by editorial offices, royalty payments systems, copyright clearances, etc.
- Automatic updating of email addresses for table of contents (TOC) alerts and other automated email communications
- Automated tools for detecting potential reviewers, including tools for detecting potential conflicts of interest
- Synchronization with publisher web site user profiles and granting researchers customized, privileged access to content based on profiles
- Understanding all of the manifold ways in which an individual "contributes" to a publisher or a field (e.g. As an editor, reviewer, letter writer, conference chair, etc.).
- Etc.

As I said, these are very publisher-focused use cases, but this is not to say that we are not interested in the use cases posed by librarians, researchers and funding agencies. We have actively been talking to people from each of these constituencies and we are trying to understand if there are ways in which we can help them. For instance, we have recently been speaking to a group of researchers who are interested in using some sort of authenticated contributor ID as a mechanism for controlling who gets trusted access to sensitive genome-wide aggregate genotype data.

The interesting thing to note about these "authentication" use cases is that they have far more stringent requirements than the "knowledge discovery" use cases. In other words if you are only trying to address the knowledge discovery problem, it might be fine to use automated techniques to disambiguate authors and assign IDs to them. State-of-the-art mechanisms for automatic disambiguation of authors from a defined corpus can be 96-97% accurate, which sounds pretty good. At least until you realize that CrossRef has ~200K new article DOIs deposited each month, each of which on average has about 3 authors. This could potentially leave you with upwards of 20K in mis/un-identified authors. This error rate might be an acceptable tradeoff for knowledge discovery type applications, but it certainly isn't suitable for authentication type applications.

Speaking of authentication, I think the fourth thing to note is that, though I think **OpenID** will probably play an important role in any service we provide, by itself it makes a pretty bad identity token and would provide little utility on its own. This all gets back to some of the issues that I raised above when discussing persistent identifiers: URI-based identifiers are fragile because they depend on the domain name. What happens if your OpenID is tied to a domain that you don't control (e.g. a company, an institution a country)? How can you guarantee that, should you leave that company/institution/country that they will do the right thing and let you maintain or redirect that identifying credential?

The traditional geeky response to this scenario is "don't get yourself into that situation. Only tie your **OpenID** to a domain that you own." (Insert forks in eyeballs). Again, you do not "own" a domain name. You lease it. What

happens if you lose control of it due to litigation, forgetfulness, poverty, divorce, death? Death? Yes, what happens when somebody dies? When I die does the not-yet-born Georgia Bilder get to buy "my" domain "gbilder.com" and make it the basis of her identity? Mmm... Gets kind of complicated doesn't it?

Of course, lots of the same issues can be raised with CrossRef, right? What guarantees that CrossRef won't become evil and co-opt all of our identities? This, of course is the big fear underlining the knee-jerk reaction against "centralized systems" in favor of "distributed systems". The problem with this, as I mentioned in the **FriendFeed thread** is that my personal and unfashionable observation is that "distributed" begets "centralized." For every distributed service created, we've then had to create a centralized service to make it useable again (ICANN, Google, Pirate Bay, CrossRef, DOAJ, ticTocs, WorldCat, etc.). This gets us back to square one and makes me think the real issue is- how do you make the centralized system that eventually emerges accountable? This is, of course, a social issue more than a technical issue and involves making sure that whatever entity emerges has clearly defined data portability policies and a "living will" that attempts to guarantee that the service can be run in perpetuity- even if by another organization. For the record, I don't think adopting the slogan "don't be evil" is enough ;).

Anyway- I could go on talking about what the contributor ID "should look like" for a very, very long time, but I think that the above probably addresses some of the major points that are raised when the topic is discussed.

5. What are the benefits (and maybe disadvantages) if CrossRef manages the author identifier?

I think the biggest potential disadvantage that CrossRef has is that it is a consensus-based organization that is governed by sometimes fierce competitors. This aspect of the organization can sometimes slow things down. On the other hand, this can also be a huge strength for us. Once a consensus is agreed, we can move very quickly and push uptake across the industry.

Research increasingly transcends institutional, geographic and discipline boundaries, so I think another advantage that we have is that we are well positioned to provide a service that is similarly unconstrained.

Finally, I think that we have a very interesting advantage by virtue of the fact that our infrastructure is already integrated upstream in the publication process. There is a useful property of the system that we are designing in that, as researchers used the CrossRef identifier in their interactions with publishers and this data is fed back into our system via DOI deposits, you could start to develop a trust-metric based on the types of claims attached to an author's profile. For instance, an author profile that consisted of nothing but self-claims (e.g. I claim I wrote paper X) might not be very worthy of trust whereas an author profile that consisted of publications that had been verified by the publisher (by virtue of those publications having been processed along with the CrossRef contributor

ID) would have far more credibility. You can start to see an interesting hierarchy of publication claims emerging such as:

- Proxy claims (Leigh claims Geoffrey wrote article X)
- Self Claims (Geoffrey claims Geoffrey wrote article X)
- Verified claims (Geoffrey claims Geoffrey wrote article X **and** the "Journal of Psychoceramics" confirms this claim)
- Verified Proxy Claims (Geoffrey (who has already been verified as an author of article X) claims that Kirsty was also an author of article X)

6. How does your author identifier relate to other identifiers, e.g. ResearcherID, Scopus Author IDor OpenID?

OpenID is a different kettle of fish, and I discussed it already above. As for the others (I'd add **Author Resolver**, **RePEC**, **SciLink**, **MathPeople**, **Nature Network**, etc.), we've actually been talking to some of these parties in order to understand how they might relate to a CrossRef Contributor ID. One obvious difference is in the use-cases being addressed. All of the above are focused on "knowledge discovery" use-cases. None of them pretends to provide any sort of authentication services. It is also interesting to note that in a lot of the above cases, the parties see their author identification functionality as a means to an end. For instance, their primary application is "creating better metrics" or "running a social network" or "expert identification" for recruiting purposes. In these cases they don't necessarily see a CrossRef system as being competitive and, in fact, they think that such a service might even improve their primary application.

7. Can you talk about the current status and next planned steps of the ContributorID project?

We just ended lengthy period of investigation and requirements gathering. In the process we went down a few blind alleys. Now we are working on a prototype that we will test with a few publishers. It is hard to say how long this will take as we are just in the process of planning this phase.

8. Satisfying many different interests is one of the biggest challenges in creating an author identifier. What are the lessons learned from implementing the digital object identifier (DOI)?

I'll give you one tactical lesson and one strategic lesson.

The tactical lesson is foremost in my mind because I have recently been trying to build tools to encourage researchers to use DOIs in their citations. The problem arrises when a researcher occasionally encounters a DOI that is 80 characters long. There is just no way that a researcher is going to insert **that** in a citation. The tactical lesson here is that it is sometimes better to make an identifier opaque and short. This is also a tremendously unfashionable position to take, but I think that one of Clay Shirky's observations about hierarchical

categorization systems also applies to identifiers. If you make the identifier human-interpretable and add semantics, then people will be extremely tempted to start hard-coding ontologies into their identifiers. This makes said identifiers both long and inherently brittle. The ontologies will inevitably evolve, and then people will want to change the identifiers- at which point they will either break or you have a giant identifier mapping subsystem to create.

We see a manifestation of this syndrome already with the DOI. Each DOI has a four-digit "prefix" which is effectively a namespace for the assigning publisher. Note that I said the "assigning" publisher- this is not necessarily the publisher who currently "owns" the DOI with that prefix. What this often means is that, when publisher A acquires publisher B, publisher A will ask CrossRef if we can create new DOIs for all of publisher B's backfiles so that they all have the same prefix! The answer to their request is "no", but you wouldn't believe how stroppy publishers can get about this. They somehow imbue this ridiculous four-digit prefix with branding significance. This, of course, is absolutely mental, but it is a predictable form of mental. The French went mad when they had to replace their region-encoded license plates with opaque EU ones. People in the US go mad when they are given new area codes. In short, when people associate semantic significance in identifiers, you will face problems.

The strategic lesson is basically a recapitulation of the "technical vs "social" theme I've been banging on about. I think that, at first, even our membership thought of the CrossRef DOI as being a technical solution to a problem, not a social one. It has become much clearer to us over the years that CrossRef DOIs are only as persistent as CrossRef staff. That is, we sometimes have to bang on lots of heads and threaten members with fines and worse in order to make sure that they are meeting their terms & conditions. The good news is that CrossRef has become essential infrastructure for a wide variety of publishers who are often at each other's throats in any other circumstances. In many ways these "different interests" are our strength. Everybody wants it to work better, nobody wants to see it die and nobody wants it to be co-opted. We are working hard to put the social structures into place that will guarantee its longevity. Part of this is making sure that we are fiscally sound (which we are) and part of this is making sure that, even if we do disappear, other stakeholders can run the system if need be.

9. What can researchers interested in author identifiers do to help?

- Feed CrossRef more use cases.
- Let CrossRef know what you think will/won't work.
- Make sure you let your publishers know if you think this is a good idea. Naturally, I expect you will also let them know if you think it is a bad idea ;-)

I can be reached at **gbilder at crossref dot org**.