Disambiguation without de-duplication: Modeling authority and trust in the ORCID system.

Author: Geoffrey Bilder
Version: 4
Date: 16-MAR-2011

1. Introduction

ORCID's mission statement starts as follows:

"ORCID, Inc. aims to solve the author/contributor name ambiguity problem in scholarly communications by creating a central registry of unique identifiers for individual researchers and an open and transparent linking mechanism between ORCID and other current author ID schemes."[1]

Meanwhile, the first three items from the ORCID beta scope say that:

1) ORCID will focus on currently active researchers.
2) ORCID will be a hybrid system of self- and organization-asserted identity.
3) Data will come from individuals and organizations.[2]

There seems to be a paradox here. On the one hand ORCID is saying that our goal is to create unique identifiers for individual researchers, yet we are creating a database comprised of both self-asserted records and many (potentially) externally asserted records- a process that is highly likely to result in the deposit of duplicate records for some researchers. In other words, the mission statement implicitly describes a system where there is a one-to-one relationship between an identifier, a record and a researcher- whereas the scope statement describes a system with n-n-1 relationship between identifiers, records and a researcher.

This paper will make the case that most of the "areas of disagreement" documented in Thorisson's white-paper only arise due to the assumption expressed in ORCID's mission statement that the only way to create an effective researcher disambiguation system is to assign unique identifiers to individual researchers. As soon as we set aside the assumption that we require a one-to-one correspondence between and identifier, a record and a researcher, then it is possible to explore alternative architectural and technical approaches to solving the author ambiguity problem- approaches that actually exploit the repetitions and/or conflicts between multiple records in order to build a simple, flexible and resilient trust metric into the ORCID researcher disambiguation system. In order to build an effective researcher and contributor identification system, disambiguation is essential but de-duplication and uniqueness are not.

The rest of this paper will attempt to describe (at a very high level) an architectural approach that can exploit repetition and/or contradiction in deposited records. Although this proposed approach is very different from what ORCID developed for its "alpha", there are relatively simple ways in which the more traditional alpha system (or similar) can evolve to work with the architectural approach proposed below.

---

[1] http://orcid.org/mission-statement

[2] http://orcid.org/content/orcid-beta-scope

ORCID's decision to accept potentially duplicate records from multiple parties was a pragmatic one, based on two broadly shared convictions of ORCID participants. The first was that, in order for ORCID to succeed and become self-sustaining, it needed to build a critical mass of data quickly and that neither an organizationally-asserted system nor a self-asserted system would, by themselves, allow for the creation of this critical mass quickly enough. The second was that, in order for ORCID to create records for non-active researchers (that is, researchers who were unlikely or unable to self-claim records), the system would depend on a variety of third-parties to provide organizationally-asserted records.

The process of gathering requirements for and developing the ORCID alpha helped to highlight the tensions raised by the seemingly contradictory visions presented by ORCID's mission statement and its scope statements. The alpha of the ORCID system was based on Thomson Reuters's (TR) codebase for the ResearcherID (RID) system, which, in turn, is focused on self-asserted claims where it is relatively easy to enforce a 1-1-1 relationship between identifier, record and researcher. Even though the TR system allows universities to 'seed' records on behalf of their faculty, these records are only made public once they are claimed by the researcher and from then on they are controlled entirely by the researcher. There is no real provision in the RID system for creating organizationally managed records and, similarly the RID system has no support for the problems that arise when multiple parties deposit records describing the same researcher.

Still, the experience of building the alpha helped ORCID come to several important conclusions. First, that, in the short term, there was value in focusing ORCID's initial efforts on the seemingly more tractable problem of registering active researchers. This is echoed in the first line of ORCID's scope statement. Second, that ORCID needed to start a parallel track of requirements gathering, research and and prototyping in order to explore the technical issues relating to dealing with multiple records for the same researcher. To this end, we have convened a subgroup of the Technical Working Group (TWG) that is exploring algorithms for detecting records that appear to be about the same person and trying to understand what particular metadata is needed in order to get those algorithms to work optimally.

But this work on disambiguation algorithms and metadata requirements, while important, will not address some of the difficult and important policy issues that running a hybrid system will raise. Gudmundur Thorisson noted in his "Summary report on ORCID core system requirements and current status of development"[3], there are still a number of "significant and unresolved issues concerning control and authority over profile information sourced from third party profiles." In short, with a hybrid system designed to accommodate records from multiple sources, how do we decide which data to trust?

---

[3] https://sites.google.com/site/openrid/technical-working-group/orcid-requirements-summary-nov10

It is not enough to determine that X records from Y different sources are likely to be referring to the researcher named Josiah Carberry [4]. We also have to have a clear policy about what we do with the records after that determination. Do we merge them? Do we privilege one of them? If we merge them do we discard the original records? If we privilege a record, do we discard the non-privileged ones? If we discard records or data, how to we keep track of the provenance of the data in the new, merged record? If we merge records, who then "owns" the resulting record and who has the right to correct and edit it? What are the IP implications of such a merged record? How do we decide these things?

At first, the answer to the question of "what data do we trust? "might seem obvious- "we trust data from authorities and authorities own and control the data" But this simply begs the question- "on what basis and in which contexts do we consider a party to be "an authority?"

Brown University might be an authority when it comes to telling you what Josiah Carberry's affiliation and title are. But if Josiah has a joint affiliation with Wesleyan University, then they are also authorities on his affiliation and title. Meanwhile the Society of Psychoceramics will be an authority on his membership in the society, The Journal of Psychoceramics (via CrossRef) will be an authority on whether he authored the paper with the CrossRef DOI 10.5555/pceramics-feb-29-1970 and Josiah himself will be an authority on his preferred email address, the location of his blog, his research interests and his privacy preferences.

And this is just the start of the complications that can arise. We must also consider the possibility that some of the records might be wrong or partially wrong. That they might be wrong due to deliberate deception or honest mistake. And that even authorities might be wrong about critical data.

Finally, our disambiguation algorithms are never going to be 100% accurate- so it will also be a reasonable possibility that the several records identified for Josiah Carberry are not all, in fact, about the same researcher and that the apparent discrepancies between the records are not the fault of the party who submitted them, but are the result of the disambiguation algorithm itself.

It soon becomes clear that the "areas of disagreement" that Thorisson's report highlighted are not merely technical quibbles. If, as ORCID's mission statement implies, the ultimate goal of the ORCID system is to have one canonical identifier and one canonical record for each researcher registered in the system, then the questions about how these records are created, how they are managed and corrected and who ultimately "owns" them are of paramount importance and need to be settled well before any public ORCID system is built and launched.

---

[4] http://en.wikipedia.org/wiki/Josiah_S._Carberry

Alternatively, we can revisit some of the assumptions that we have thus-far made about what is required in order to “solve the author/contributor name ambiguity problem.”

2. Terminology

The terminology used in discussing identity systems is often used differently in different contexts. Before discussing the proposed architecture it will be helpful to briefly define some terms.

Party

The term “party” will be used to refer to a person or an organization (i.e. a legal/juridical person). Examples of parties include:

a) Josiah Carberry
b) Brown University
c) ORCID

Claim

A claim is a statement about a party or about another claim. We use the term claim as opposed to “assertion” because the word claim implies a certain level of uncertainty. Examples include:

a) Josiah Carberry’s title is “Professor”
b) Claim “a” was made by “Brown University”

Identity Record

An “identity record” (or, simply “record”) is a set of claims made by one party about itself or another party. For example, Brown University might create a set of claims stating the following:

a) Josiah Carberry’s title is “professor”
b) Josiah Carberry’s is a faulty member at “Brown University”
c) Josiah Carberry’s department is “psychoceramics”

Taken together, these claims can be thought of as a “record.”

Identifier

A token that uniquely identifies a set of claims (identity record) or an individual claim. (e.g. 12345-67890)

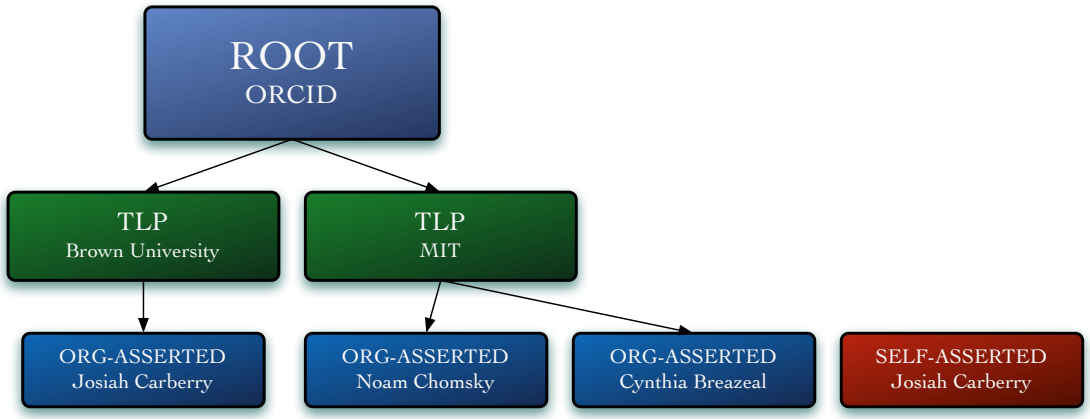
3. An Architecture That Exploits Redundancy of ORCID Records

Summary

The gist of the proposed architectural approach is to think of the ORCID system as being a collection of "claims" made by "parties." Some claims may be reasserted many times by many different parties, some claims by different parties might contradict each other and finally, some claims might be unique. Critically, this duplication, contradiction and uniqueness can tell us something about the credibility of a record or even of a particular claim. Particularly if one exploits a de-facto authority hierarchy that will exist for parties within the ORCID system.[5]

Hierarchy of Parties

Some parties who interact with the ORCID system will actually be **known** to ORCID. That is, the ORCID organization will deal directly with them as members, affiliates or partners. In contrast, there will also be a class of parties who will be relatively anonymous to ORCID. Individual researchers creating self-claim records may interact with the ORCID system, but they will not generally be dealing with the ORCID organization directly. In essence, we can think of this hierarchy as a tree where (borrowing from DNS terminology), ORCID is a root party who, in turn creates identity records for "top level parties" (TLP) representing its members, partners, affiliates, etc. These TLP's can, in turn create identity records representing other parties.

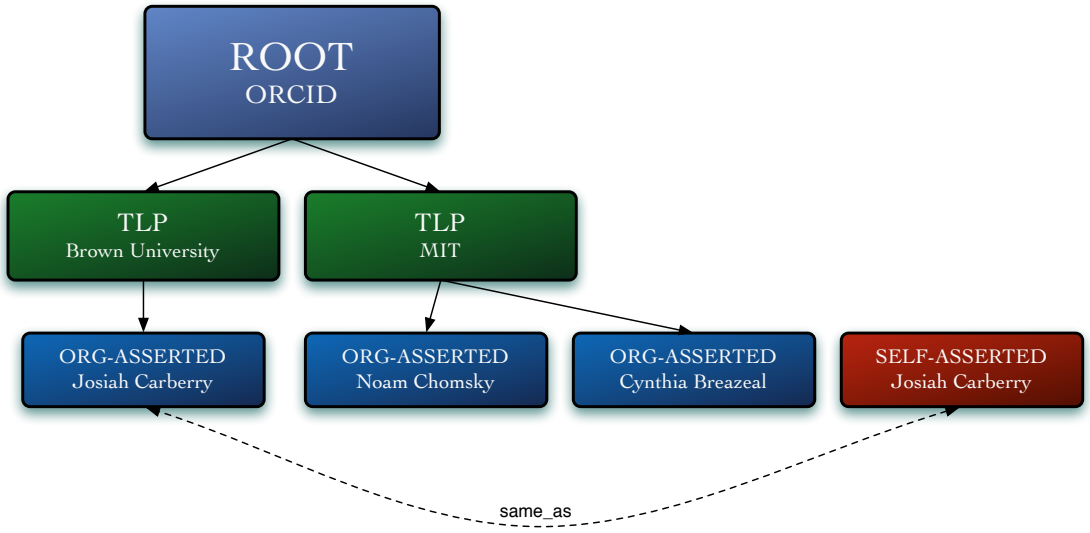


Note that, **initially**, self-claim identity records would sit entirely outside of this hierarchy.

Disambiguation

[5] The technically minded will, no doubt, immediately see the connection between what I am describing and triple/quad stores, graph databases, etc. And, yes, essentially what I am saying is that this architecture can be implemented using these technologies. However, I am also deliberately trying to avoid the use of semantic web jargon and overly technical or detailed examples. Forgive the handwavyness. The first more detailed draft of this was unreadable.

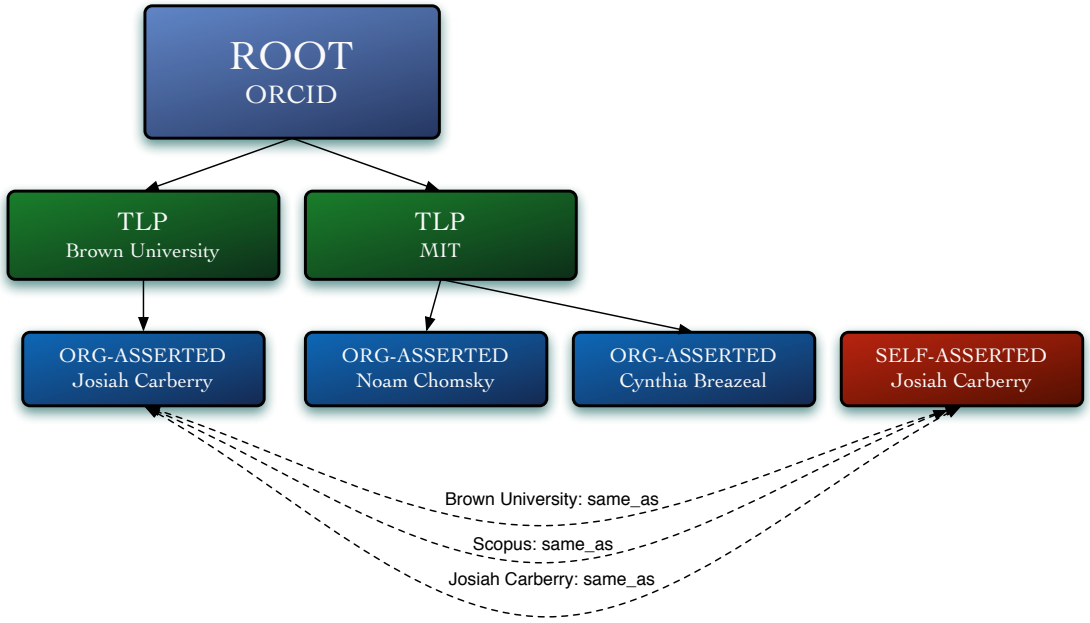
In this architecture, the process of detecting that two or more "identity records" are likely to be referring to the same "party" will still be important. As such the work that the TWG Profile Exchange Subgroup is doing on disambiguation algorithms and the metadata needed to make them effective will be vital. What is different about the architecture being proposed is that it would then eschew "de-duplication" of records that are disambiguated. Instead, disambiguation of records would simply consist of linking the records using a "same_as" claim[6]. The individual records would otherwise be left alone. No attempt would be made to merge them or to privilege one of them.



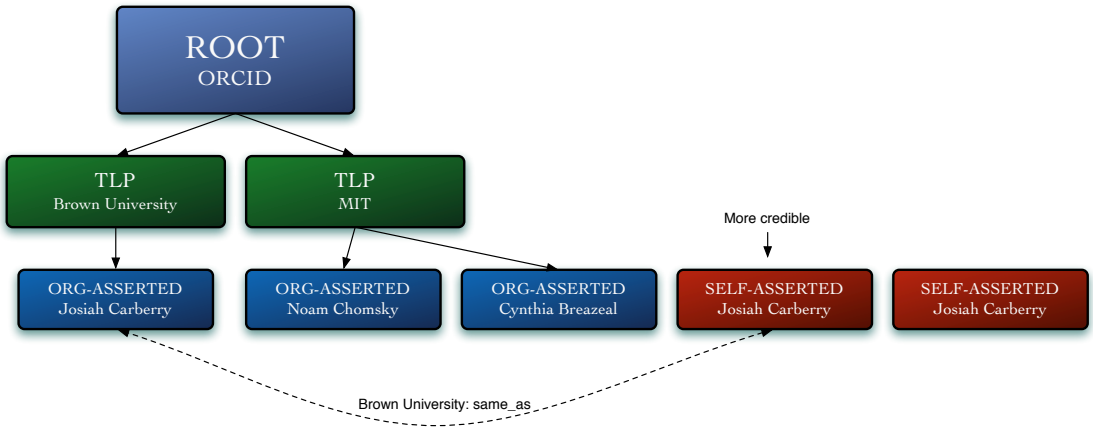
The process of disambiguation identity records through creating "same_as" claims has some interesting properties.

First, because the process is essentially non-destructive (i.e., it doesn't discard any information) the system can support multiple parties making their own "same_as" claims. So, for instance, Scopus and Brown University could make organizationally-asserted "same_as" claims about Josiah Carberry and Josiah Carberry himself could "self-claim" those records- thus creating "same_as" claims as well. In order to support this, the system simply needs to record the provenance of the "same_as" claims. If multiple parties using different disambiguation techniques create multiple, duplicate "same_as" claims, this effectively tells you something about the likelihood that the relationship is accurate.

[6] Again, sem-web people please realize that I am using "same_as" very loosely here and am not trying to make the argument (yet) that owl:sameAs is the correct relationship to use for implementation.



Second, as soon as we have “same_as” claims created by a top level party (TLP), we effectively have a way of slotting self-claims into the authority hierarchy. So, for instance, if there are two self-claim records claiming to be about Josiah Carberry- One authentic record created by Josiah and one bogus record created by Josiah’s nemesis, Joel Feinberg- but only one of these records has a “same_as” relationship created by Brown University linking it to Brown’s record, then the record confirmed by Brown clearly has a more credible provenance.



Using “same_as” claims to link multiple identity records which refer to the same party also allows us to more easily deal with many of the specific concerns that ORCID participants have expressed about the proliferation of records and identifiers.

For one thing, as long as records are disambiguated using “same_as” claims, then it doesn’t really matter technically which of the several identifiers that refer to a particular party is used. Josiah Carberry can use the identifier on his self-claim record when submitting manuscripts to the Journal of

Psychoceramics and Brown University will still be able to find his list of publications by querying the system using the identifier assigned to the organizationally-claimed record that they deposited.

Similarly, if Scopus were somehow the first party to deposit an identity record describing Josiah Carberry- third parties could use the identifier assigned to the Scopus-deposited record and be assured that they will be able to easily map it to the record Josiah Carberry eventually creates himself and to the record that Brown University deposits.

It is important to note that the above makes a distinction between "depositing a record" and "disambiguating records." It is entirely possible for a third party to deposit records without them (or ORCID) trying to disambiguate them first. Similarly, it is possible that a third party might provide disambiguation services, without necessarily depositing records. In order to do the later, all the party needs to do is deposit "same_as" claims to the system.

These characteristics of the proposed approach open up some intriguing possibilities. First of all, it removes what might be a very high barrier to allowing third parties to submit records into the system. Second, it allows ORCID to work with a wide variety of "disambiguation" partners who might have excellent disambiguation technology, but who do not have data to contribute to the system. As discussed above, identical "same_as" claims from multiple parties only reinforces the likelihood that the subject records are, in fact, about the same party. As more "disambiguation partners" confirm "same_as" claims, ORCID users can become more confident of them.[7]

Trustworthiness of claims

Establishing "same_as" relationships between records not only gives us a macro-level way of assessing the trustworthiness of identity records, but it also gives us the ability to evaluate the credibility of individual claims within a record. This, in turn, helps us to address many of the concerns that have been raised by ORCID members concerning who ultimately "controls" a record.
As an example, let's consider the case of a researcher "claiming" a record that has been deposited on their behalf by their institution. Up until now we have modeled this process of "claiming" a record as follows[8]:

1) Brown University deposits a record describing Josiah Carberry
2) Brown University sends email to Josiah with a link to the record asking him to claim it.

---

[7] It is beyond the scope of this paper- but there might even be a potential ORCID revenue stream in creating a "disambiguation market" where third parties are given discounted access to realtime ORCID data and where the discount is tied to the number and accuracy of "same_as" claims that they deposit into the ORCID system using their disambiguation technology.

[8] http://wokinfo.com/media/pdf/univ_hk_cust_profile_palmer.pdf

3) Josiah follows the link and claims the record. Once he has done this, he "controls" the record. He can edit and change the value of any field (other than the identifier) for the record. Similarly, he can control the privacy settings of these values.
4) Optionally- he might delegate the ability to edit his record to another party- the library, a PA, etc. He can also revoke this delegation if he wishes.

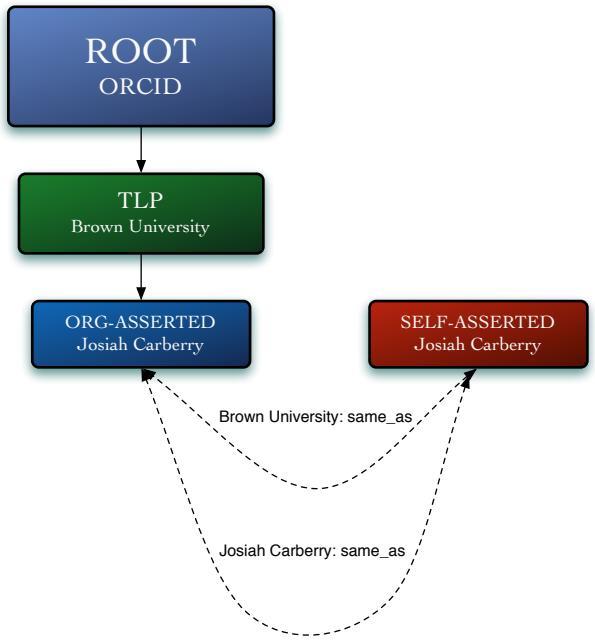
The problem raised by this scenario is that Josiah could forget to update his contact information when he moves offices, forget to update his bibliography when he has published new articles and change his title from "Assistant Professor" to "Professor". In short the accuracy of the record could deteriorate quickly.

The ability to delegate editing rights to a librarian or other authority might mitigate this problem to an extent, but it has some very important limitations. First, it relies on the delegated authority to notice these disparities and make corrections quickly. Second, between the time a disparity is introduced and the time it is corrected, third parties will have no way of knowing that particular claims in the record are potentially incorrect. Third, when a researcher removes the ability for the delegated authority to edit their record (for instance, when they move to a new institution), then information related to their past affiliations might be hard or impossible to correct. In short, the delegation mechanism might be useful in enabling busy researchers to allow others to maintain their records- but it is not a robust mechanism for ensuring the integrity of said record.

Again, these problems largely stem from the assumption that we need a one-to-one relationship between an identifier a record and a researcher. We can instead exploit the presence of multiple records for a researcher if we change the above scenario slightly so that it results in two records- one controlled by Brown University and the other controlled by Josiah Carberry. In this case, the process would look like this:

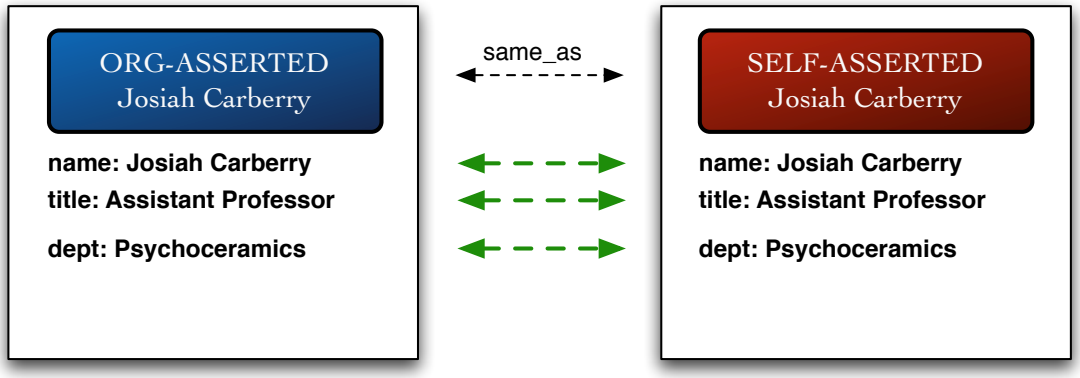
1) Brown University deposits a record describing Josiah Carberry
2) Brown University sends email to Josiah with a link to the record asking him to claim it.
3) Josiah follows the link and claims the record. Once he has done this, a **copy** of the Brown University record is made, is assigned a new identifier and **two "same_as" relationships are recorded linking the two records**- one on behalf of Brown University and one behalf of Josiah Carberry. Josiah "controls" the new record. He can edit and change the value of any field (other than the identifier and the above created "same_as" relationship) for the record. Similarly, he can control the privacy settings of these values. Note that Brown University continues to control the original record that they created.
4) Optionally- Josiah might delegate the ability to edit his record to another party- the library, a PA, etc. He can also revoke this delegation if he wishes.

Within our authority hierarchy, we end up with the following:

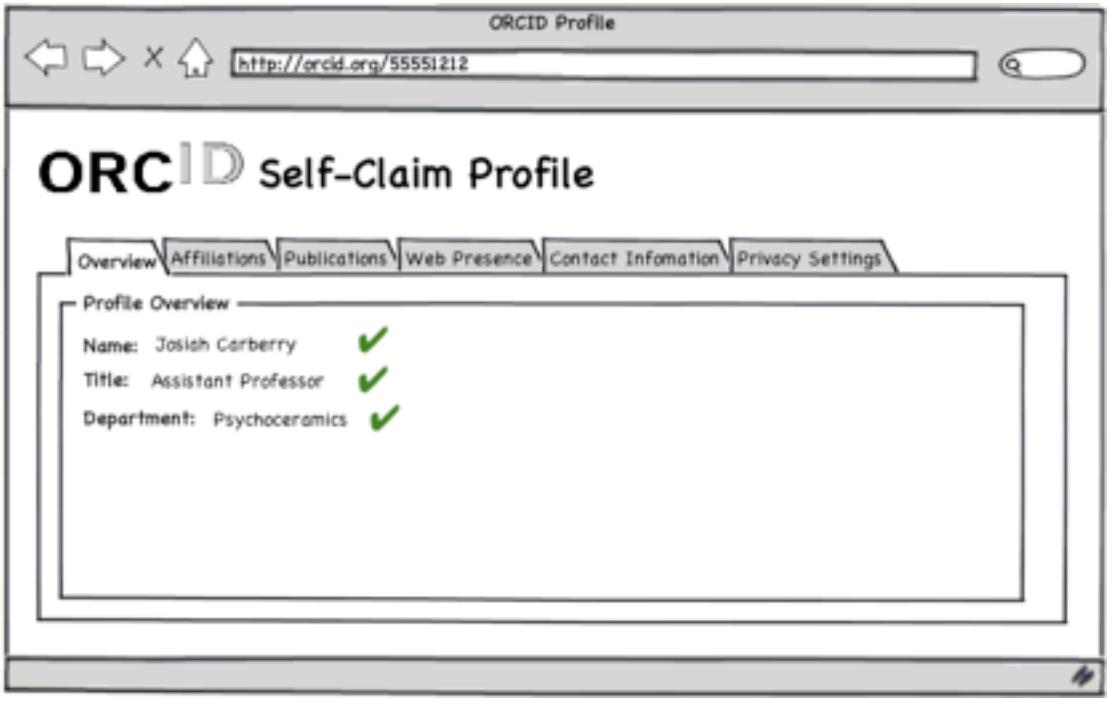


From the researcher's point of view, the process will look identical to the first, but, critically, behind the scenes, two records will have been created and "tied together" using "same_as" claims. This, in turn, makes the process of tracking and highlighting discrepancies in the records much easier.

Essentially, we start out with the following two, identical records. The Brown University record "verifies" all of the claims in Josiah's self claim record and Josiah verifies all the claims in Brown University's record.:

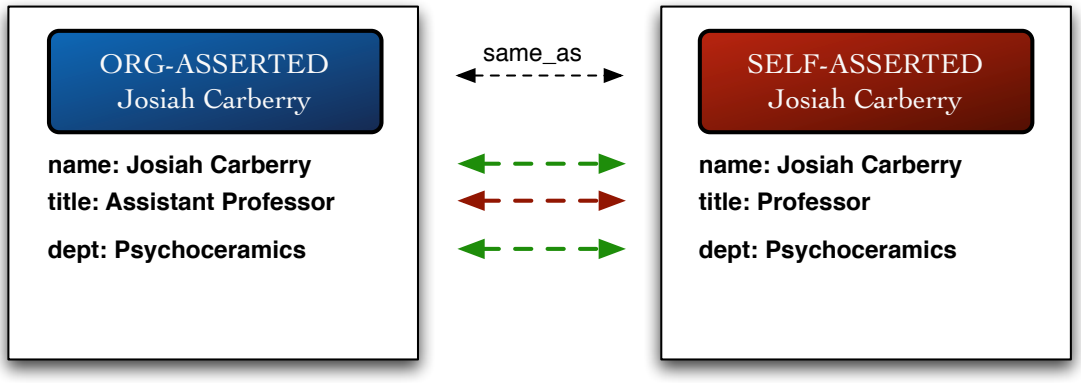


In the user interface, the fact that a claim has been verified by a Top Level Party (TLP) could be highlighted.

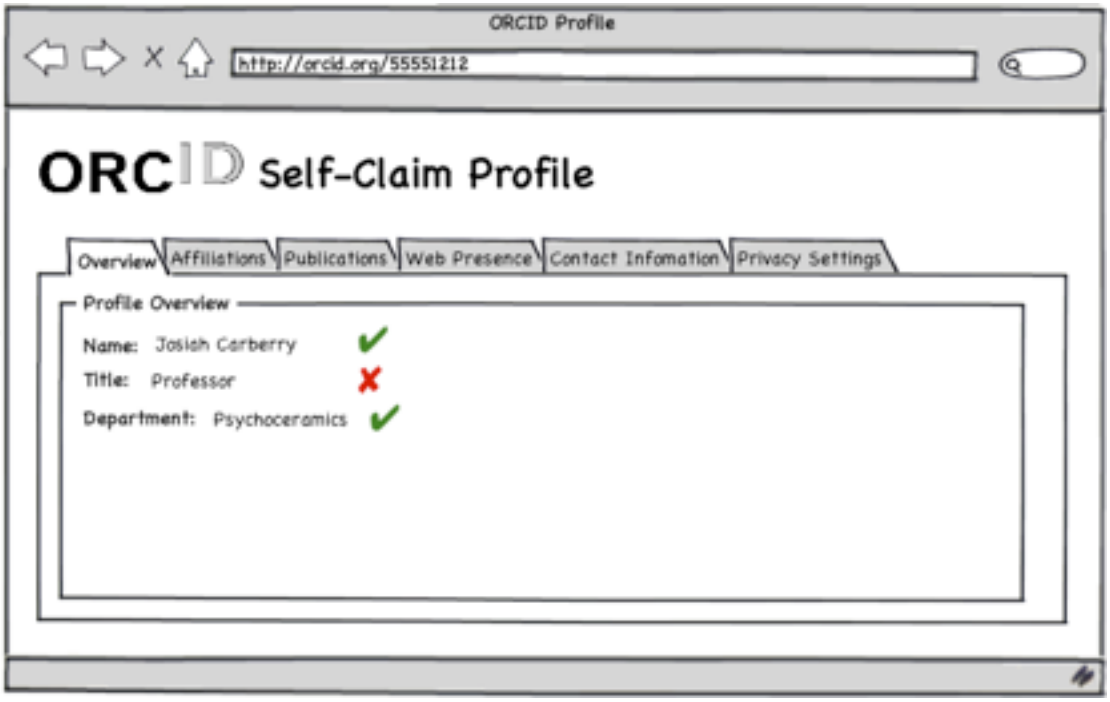


Now, it will be very easy for both Brown University and Josiah Carberry to monitor when discrepancies occur between the two records. The ORCID system, could, in fact, provide alerting tools to notify parties whenever discrepancies arise between records that have been linked with “same_as” claims. These tools could also provide the parties with a simple way to “reconcile” the discrepancies.

So, if (as in our previous example), Josiah edited his title and changed it to “Professor”, this could be detected within the system automatically and email could be sent to the Brown University library indicating that there was a discrepancy between the two records. The library could decide to either “accept” the change, at which point the change would be propagated to the library’s record as well, or they could decide to “reject” it, at which point the discrepancy would remain and be easily detectable. Importantly, anybody querying Josiah’s record between the time that the discrepancy was introduced and the time the discrepancy was resolved, would have a clear indication that the title claim was in dispute.



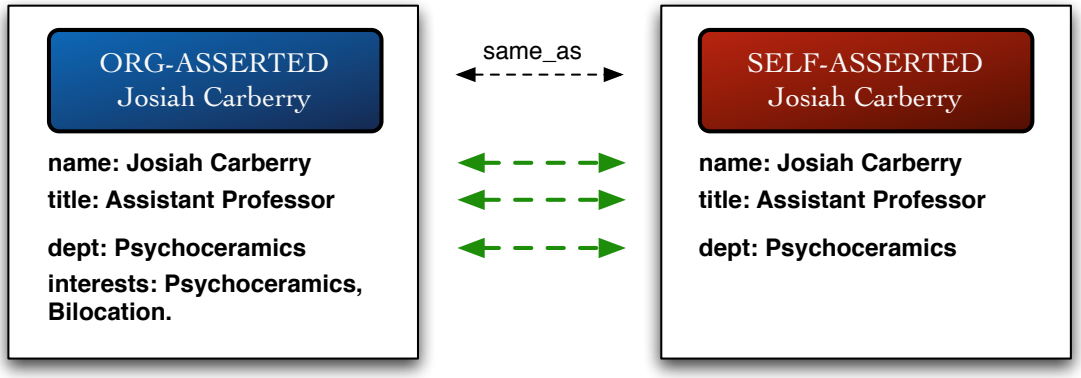
Again, in the user interface, this kind of discrepancy could be indicated fairly easily.



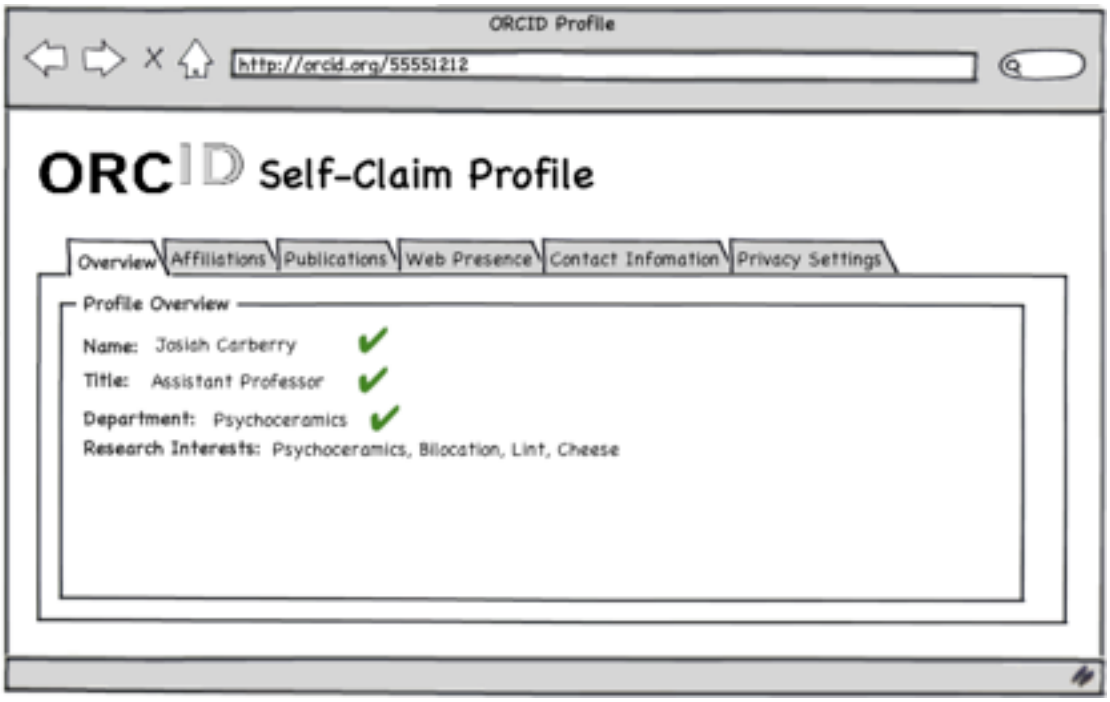
Of course, the same mechanism could work the other way around. If the library made a change to their record which conflicted with what was recorded in Josiah's record, Josiah could receive email asking him to "accept" or "reject" the change. Again, if there was disagreement, this would be indicated in the records and in the user interface until the two parties worked out a mutually acceptable conclusion.

Linking records in this way, not only allows for the detection of conflicting information, but it also allows the system to alert parties when additional information has been added to a record.

So, for example, Josiah Carberry might add a self-claim to his record listing his research interests. In this case, the Brown University library could be alerted to this addition and choose either to add it to their record or ignore it. Seeing as the research interests of Josiah Carberry are probably not an item over which Brown University feels it has an authoritative say, they could choose to ignore it and this too would be easily reflected in the record. In this case, the claim being made by Josiah is not being confirmed or disputed by a TLP, at which point it is just recorded without any confirmations or warnings.

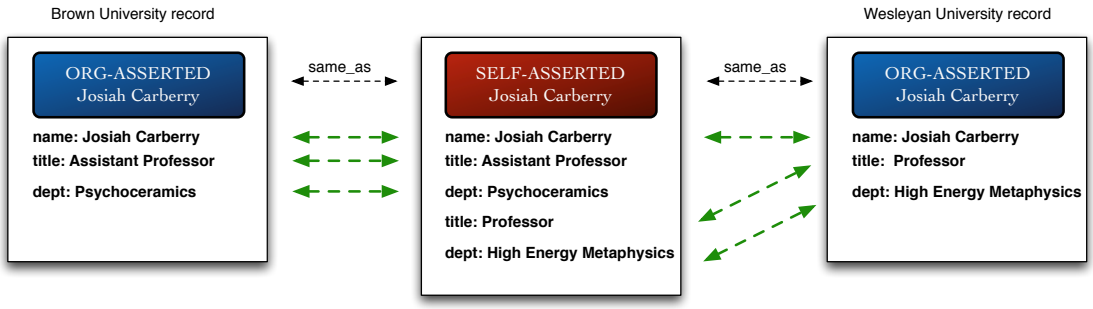


As well as in the user interface:



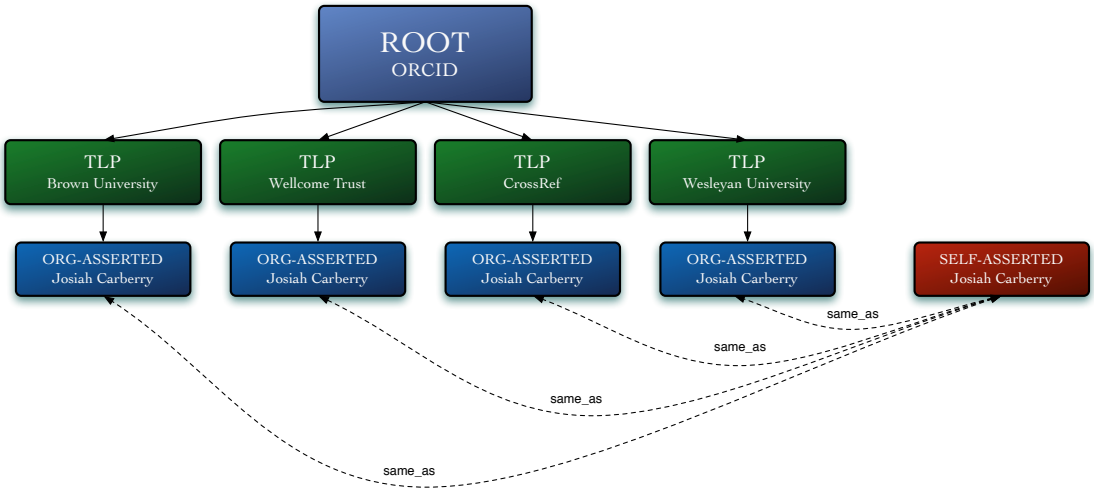
The same basic mechanisms can work across multiple TLPs all of whom maintain records pointing to the same researcher. So, for example, if Josiah Carberry left Brown University and joined Wesleyan University, then Wesleyan would also create a record and ask Josiah to “claim” it.

In this case, because Josiah already had a self-claimed record, the assertions in the Wesleyan record would simply be copied into his self-claim record along with the old assertions from his previous employer, Brown University.

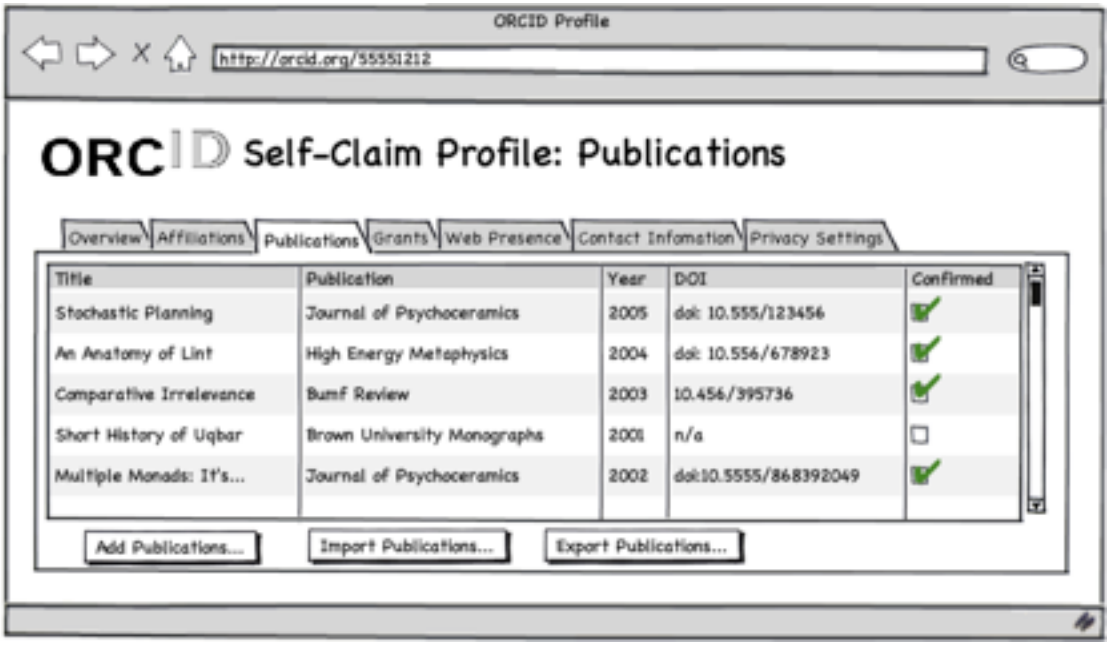


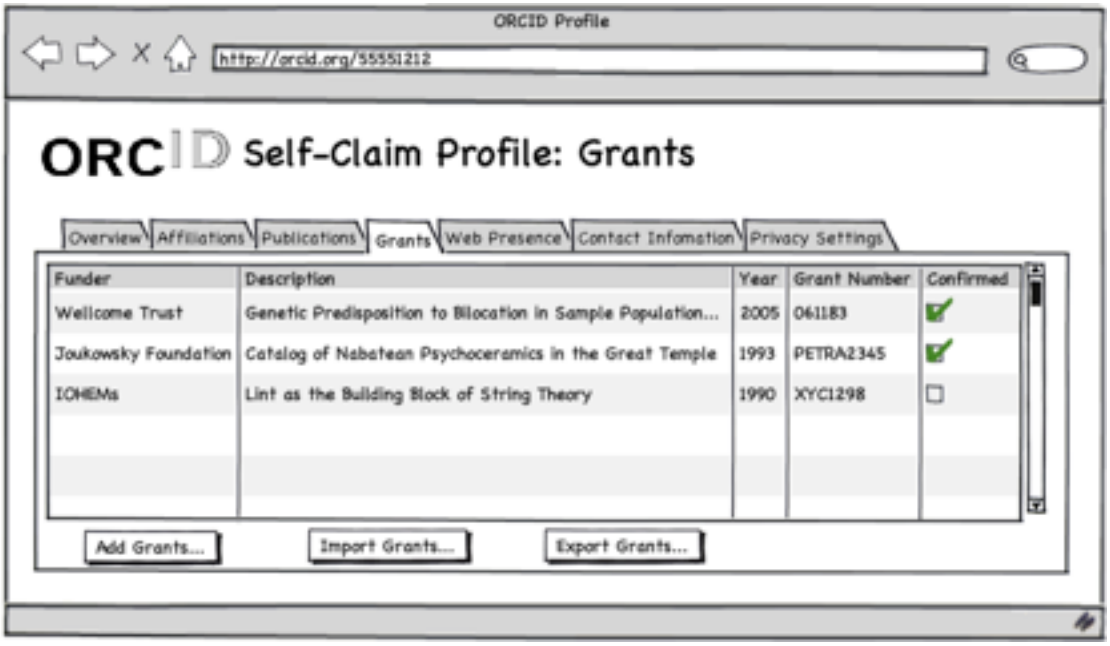
Then, if Josiah modified anything that conflicted with his Wesleyan associated claims, Wesleyan would be notified. And if he modified anything with his Brown-associated claims, Brown would be notified. And, of course, he would continue to be notified if Brown or Wesleyan changed anything.

It is important to note, that although the above examples involve a relatively small number of parties and are focused very much on the relationship between one particular class of of “Top Level Party” (i.e. A University), the same basic mechanisms can work across multiple TLPs of various classes, all of whom maintain records (again- just a set of claims) pointing to the same researchers. So, for instance, the Wellcome Trust would be a TLP which would deposit records including claims that particular parties had received particular grants. CrossRef would be a TLP who, on behalf of publishers, would deposit records including claims that particular parties had been authors of particular DOIs. All of these would, of course, be tied together (directly and indirectly) using “same_as” claims.



And the user interface conventions could be very similar to that used in the case of the University.





To summarize, if we model the ORCID system as a database of claims made by one party about itself or about another party, then we can actually exploit the uniqueness, duplication and contradiction of claims in order to create a rudimentary but effective built-in trust metric. A third party using the ORCID API or user-interface could assess the credibility of individual records in the database using a number of methods.

First, at the record level, one could easily determine the provenance of a record. Is it a self-claim record or has it been deposited by a Top Level Party? If it has been deposited by a TLP, which TLP deposited it? One might, for instance, conclude that a record deposited by a well-known university library is likely to be more credible than a record deposited by a relatively unknown,

startup social networking site or a self-claim record. Secondly, one could see how many “same_as” claims exist for the record in question. One might conclude that a record that is linked to many others by “same_as” claims is likely to be more credible than one that has no such links. Additionally, A record that is linked to by a TLP gains some transitive credibility. A record that is linked to from multiple TLPs even more so.

Even individual claims could be assessed using the same mechanisms. If Josiah Carberry’s name, affiliation and title claims are duplicated by multiple TLPs, then those particular claims are likely to be credible. If, on the other hand, his self-claim to have won the *Ig Noble Prize* in 1991 is not repeated anyplace else, then it should probably be treated with some skepticism. This is not to say that all unverified claims should be discounted if they are not repeated by a TLP. For instance Josiah’s self-claims about his research interests, his preferred contact information, the location of his blog, etc. are all claims that could probably be taken at face value without confirmation by another party.

4. Metadata Implications

Clearly, implicit in a system where the credibility of records and claims is tied to the repetition and/or contradiction in claims, is that such claims need to be made using a standard vocabulary. The system will not work if Brown University” deposits a record for Josiah Carberry with a “title” claim and Wesleyan deposits one with a “job_title” claim. This is not different from a system based on “de-duplication” of records. However, what does probably change is our conception what what constitutes “minimal metadata” for a deposit.

In a claim-based system, even a simple deposit by CrossRef of a claim that says nothing more than “the party with the identifier 55551212“ was the author of DOI:1028/746585” is of value. So, minimally, individual claims by TLPs about parties that already exist in the database should be allowed.

The slightly more complicated issue is that of a TLP depositing a “new record.” That is a set of claims about a party that has yet to be linked via a “same_as” relationship to any other party in the system. So, for example, let’s say Scopus had a record for a researcher named “J. Carberry” and they deposit the record without linking it with a “same_as” relationship to any existing record in the ORCID database. In essence, they are creating what appears to be a “new party”. Now, they might do this for one of two reasons:

a) They are legitimately the first to deposit a record for Josiah Carberry.
b) Their record cannot be matched definitively to an existing record for Josiah Carberry. This could be because:

    1. The algorithm is imperfect (see below).
    2. The metadata is flawed.

3. The metadata is incomplete.

It has to be pointed out, that from a practical standpoint it will be impossible for ORCID to distinguish between cases a and b with 100% precision. This is both true of a system based on “de-duplication” and for the proposed system that eschews de-duplication. The critical difference with the proposed claims-based system is that a third-party querying the ORCID system and obtaining the Scopus-deposited record would be able to easily see that the record had no “same_as” relationships. They could then either decide that the record was not credible enough to use, or they could use it secure in the knowledge that if, in the future, the Scopus record was eventually matched with other records relating to Josiah Carberry they could use the identifier assigned to the Scopus-deposited record and discover all of the “same_as” relationships and the records they point to. This latter situation might happen if, for example, disambiguation technology improved or if subsequently another record was submitted that contained a piece of metadata which allowed existing technology to link otherwise disambiguate-able records.

In other words, there would seem to be a big downside to rejecting records from the ORCID system because they had too little metadata. The worst that will happen is that the claims in those records will never be repeated and that, therefor, they will clearly lack credibility. On the other hand, if the claims in those records *are* eventually repeated and/or contradicted, then they might help to enhance and or expand on existing records for the same party.

5. What gets to be called “an ORCID?”

This leaves us with another issue that has frequently been raised in relation to the possibility of the ORCID system containing multiple records pertaining to the same party. That is- what does one call the “identifier” associated with each of those records? Again, does one of the identifiers get “privileged” and called “the ORCID?” If so, what do the identifiers for the other records get called? How does one decide which record gets privileged? Again, these are questions of “control”, “authority” and “trust.”

One could argue that only identifiers from records from “authorities” get called “ORCIDs”. In other words, only records from TLPs. But this presents a number of problems. For instance, Josiah Carberry might work for Brown University and Wesleyan University - both of whom are TLPs. Does he then have two ORCIDs? Similarly, are all TLPs considered equal- or are some privileged? If so, how? And finally, it would seem pretty perverse to the researcher if *they* had a self-claim record and yet whenever they were asked to record their ORCID, they were expected to record the identifier from a record (or records) that they don’t manage.

More to the point, from a practical level- does this distinction matter in a system where Josiah’s identifier can automatically be mapped to the Brown and Wesleyan identifiers and where any conflicts in claims can easily be

spotted. Surely, this, more than anything underscores the authoritative roles of the TLPs?

And finally, if you call only TLP identifiers “ORCIDs” what do you call of the others?

So then, does it make sense to only call the identifiers associated with “self-claim” profiles an “ORCID?” Again, from a practical, technical level, this makes virtually no difference. If Josiah Carberry used his self-claim identifier when registering with Nature’s manuscript tracking system and then accidentally used the Brown University record’s identifier for registering with the Society of Psychoceramics, we would still be able to link the identifiers back to Josiah. And besides, if we do call only self-claim profiles “ORCIDs” then this again leaves us with the question of what to call the other ones.

The question of “what gets called an ORCID” is really another manifestation of the assumption that we somehow have to de-duplicate or collapse records into something canonical and, as such, all of the same complications and issues are raised.

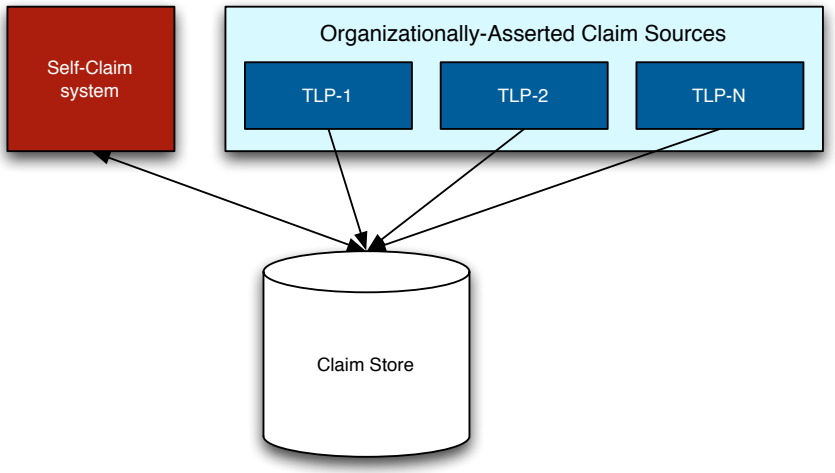
The alternative is to call all the identifiers ORCIDs and to focus our efforts on making sure that people understand ORCIDs can be used for disambiguation, but that it is entirely possible that multiple ORCIDs will point to one party. The critical point is that this relationship between the ORCIDs will be exposed and that resolving any one ORCID will effectively be the same as resolving any of the ORCIDs that are associated with it via a “same_as” claim.

Our instinct is to want one “canonical” identifier, but in the described architecture, having such an identifier makes no technical difference at all. Any decision to designate a certain set of these identifiers as canonical would have to be made on political or usability grounds and balanced against the difficulties we would encounter in justifying that decision.

6. How Does This Fit With the Approach Used in the Alpha?

At first glance, it would appear that the architectural approach proposed here is orthogonal to the architecture developed for the TR RID system and upon which the ORCID Alpha was developed. Indeed, it would appear that the architecture being proposed would be incompatible with any of the technology stacks that the TWG has identified as being potential replacements for the RID code (e.g. RePEc, VIVO, BibApp) should ORCID fail to come to licensing terms with TR. What does this mean for the ORCID beta?

Fortunately, this is not the case. The self-claim component of the overall ORCID platform can be built as a discrete system and can be run productively on its own until it is eventually integrated into a claim store built on the claim model described. In a sense, the self-claim platform can be viewed as just one of several sources of data for the eventual claim store.



But this does have some implication for the beta scope statement. Although the scope statement starts by saying that the beta will focus on active researchers, the scope goes on to state that data will come from individuals and organizations. To a limited extent this is fine. There were originally two major drivers behind ORCID's need for organizationally asserted records:

1) To expedite the signup procedure of researchers by allowing them to "claim" already existing records.
2) To allow for the disambiguation of inactive researchers.

And to this, the claims-based approach being proposed here adds a third:

3) To provide ORCID system with a claim-verification mechanism.

The first of these drivers can be easily accommodated within the architecture model of the alpha in one of two ways:

a) By having the researcher's institution deposit a record on their behalf, which they can then "claim".
b) By having the researcher enter the ID of another identity system from which they want to copy profile data.

Neither of these requires the existence of the proposed claims database or any of its verification functions.

The second of these drivers (disambiguating inactive researchers) has already been de-scoped from the beta. As the first scope statement says, the ORCID beta will focus on active researchers. The ability to handle inactive researchers will be a feature of a subsequent iteration of the platform.

The third of these drivers (verification of claims) has also been de-scoped from the beta. Although many of the wireframes for the alpha included examples of claim verification (publishers confirming publications, institutions confirming affiliation, etc.)- there seemed to be broad consensus that a useful

beta could be launched without these features- as long as they could be added easily in the next iteration of the system.

So it would seem that an ORCID beta based on the general architecture (if not the code base) of the ORCID alpha will meet the requirements set out in the beta scope statement and can be easily expanded to deal with our additional use cases when it is eventually coupled with a claim store.

7. Final Comments

Gudmundur Thorisson's review of the ORCID requirements noted both many broad areas of agreement as well as some areas of disagreement. These areas of disagreement tended to center around control and authority in the in the management of ORCID records. How does a record get created? Who manages the record? Who corrects a record? Who "owns" a record? Providing answers to these questions is vital to the integrity of the ORCID project. Our principles say that researchers will be able to "create, edit, and maintain an ORCID ID and profile" as well as "control the defined privacy settings of their own ORCID profile data." But what happens if researchers make false claims or hide information that might be embarrassing? Our first principle states that we will enable the "creation of a permanent, clear and unambiguous record of scholarly communication by enabling reliable attribution of authors and contributors. The scholarly record will be none of these things if it is also bowdlerised.

So how do we resolve this tension between providing researchers control of their profiles and preserving the integrity of the scholarly record?

In order to "square this circle," we have to stop conflating the terms "disambiguation," "profile matching," and "de-duplication." Disambiguation and profile matching" can tell us that several records are likely to be about the same party. "De-duplication" implies that we then need to somehow collapse those records into one. As soon as we do this, we introduce all of the issues of control, authority and ownership listed above. If instead we develop a system that takes advantage of the inevitable repetitions and contradictions that will exist between records we can instead build a platform that clarifies provenance, builds trust, and clearly delineates ownership.

Fortunately, this can be done in discrete steps. We don't need to develop the complete claim system in order to launch something useful. We can start by building an exclusively self-claim system and migrate it to include the "checks and balances" of a a hybrid system soon thereafter.

Appendix A: Examples using a toy DSL

Initially, I started by trying to write this paper using a toy domain-specific language with which to illustrate the concepts. I abandoned this attempt after realizing (i.e. being *told* in no uncertain terms) that this approach would not work for non-technical readers. Still, seeing how the DSl worked might be of marginal interest to those who enjoy this sort of thing.

I implemented the DSL using ruby and mongodb- simply because I didn't think I had time to faff around with a real quad-store or graph db. It really is a toy. Still it actually works- which is pretty cool.

The text below narrates an interactive session with the DSL using IRB. Comments are in italics. Commands start with a "$" and (partial) output from the commands is in blue.

==========

We start by creating our root party, 'ORCID.' This is the top of the authority hierarchy. It has no 'source' or provenance as it is the party upon which all other parties are created. Note that the "label" attribute is merely there to make the output human readable.

```
$ orcid = Party.create :label=> 'ORCID'

====================
Party: ORCID
Authority level: root
```

ORCID creates a new party on behalf of Brown University. This is a so-called "top level party", These are parties that have a direct relationship with the ORCID organization and for which ORCID is, essentially "vouching."

```
$ brown = Party.create :label=> 'Brown Library', :source => orcid

====================
Party: Brown Library
Authority level: Top Level Party (TLP)
```

ORCID creates another TLP, this time for CrossRef.

```
$ crossref = Party.create :label=> 'CrossRef', :source => orcid

====================
Party: Brown Library
Authority level: Top Level Party (TLP)
```

Brown University creates a record for Josiah Carberry.

```
$ b_jcarberry = Party.create :label=> 'Brown Library\'s record for Josiah
Carberry', :source => brown

====================
Party: Brown Library's record for Josiah Carberry
Authority level: Created by Top Level Party (TLP)
```

Brown University makes a few claims about Josiah Carberry.

```
$ brown.claims_that b_jcarberry, :institution, brown
$ brown.claims_that b_jcarberry, :identity_type, 'person'
$ brown.claims_that b_jcarberry, :givenname, 'Josiah'
$ brown.claims_that b_jcarberry, :surname, 'Carberry'
$ brown.claims_that b_jcarberry, :department, 'Psychoceramics'
$ brown.claims_that b_jcarberry, :title, 'Full Professor'
```

*Brown university alerts Josiah Carberry via email that a record has been seeded on his behalf and that he should "claim" or activate it. Josiah follows the link in the mail, which creates a *new* record for him.*

```
$ jcarberry = Party.create :label=> 'Josia\'s self-claim record'
```

```
====================
Party: Josia's self-claim record
Authority level: self-claim
```

Josiah is asked to examine the record submitted by Brown University and to accept it. He does, at which point his newly created record is seeded from the Brown University record.

```
$ jcarberry.seed_from_authority(b_jcarberry)
```

```
Claims:
       [13] Josia's self-claim record / same_as / Brown Library's record for
Josiah Carberry  (confirmations=1) (authorities=0)
```

Josiah then goes on to add some claims of his own.

```
$ jcarberry.claims_that jcarberry, :honor, 'Nobel Prize for Humor'
$ jcarberry.claims_that jcarberry, :authored, 'doi:10.5555/1065'
$ jcarberry.claims_that jcarberry, :authored, 'doi:10.5555/1111'
$ jcarberry.claims_that jcarberry, :authored, 'doi:10.5555/1234'
```

CrossRef also makes a claim about Josiah Carberry. This claim confirms one of the claims made by Josiah.

```
$ crossref.claims_that jcarberry, :authored, 'doi:10.5555/1234'
```

At this point it, if you look at the Brown University record, it will show you that it is the "same_as" the self-claim record by Josiah Carberry.

```
$ puts b_jcarberry
```

```
====================
Party: Brown Library's record for Josiah Carberry
Authority level: Created by Top Level Party (TLP)

Self Claims:
       [15] Brown Library's record for Josiah Carberry / same_as / Josia's
self-claim record  (confirmations=1) (authorities=1)
...
```

If we then look at Josiah's self-claim record, we can see each of his claims, as well as the number of times the claim has been confirmed by another party (confirmations) and the number of times the claim has been confirmed by a TLP (authorities). Note that Josiah's claim of the 'Nobel Prize for Humor' has not been confirmed and that only one of his DOI authorship claims has been confirmed.

```
$ puts jcarberry

Self Claims:
       [7] Josia's self-claim record / institution / Brown Library
(confirmations=1) (authorities=1)
       [8] Josia's self-claim record / identity_type / person
(confirmations=1) (authorities=1)
       [9] Josia's self-claim record / givenname / Josiah  (confirmations=1)
(authorities=1)
       [10] Josia's self-claim record / surnamename / Carberry
(confirmations=1) (authorities=1)
       [11] Josia's self-claim record / department / Psychoceramics
(confirmations=1) (authorities=1)
       [12] Josia's self-claim record / title / Full Professor
(confirmations=1) (authorities=1)
       [13] Josia's self-claim record / same_as / Brown Library's record for
Josiah Carberry  (confirmations=1) (authorities=0)
       [17] Josia's self-claim record / honor / Nobel Prize for Humor
(confirmations=0) (authorities=0)
       [18] Josia's self-claim record / authored / doi:10.5555/1065
(confirmations=0) (authorities=0)
       [19] Josia's self-claim record / authored / doi:10.5555/1111
(confirmations=0) (authorities=0)
       [20] Josia's self-claim record / authored / doi:10.5555/1234
(confirmations=1) (authorities=1)
```

So what happens if Josiah goes in and changes his "title" to "President"?

```
$ jcarberry.claims_that jcarberry, :title, 'Presendent'
```

Then we will easily see the conflict between his self-claim and the Brown University record for Josiah in claim number 12 (challenges)

```
$ puts jcarberry

Self Claims:
       [7] Josia's self-claim record / institution / Brown Library
(confirmations=1) (authorities=1)
       [8] Josia's self-claim record / identity_type / person
(confirmations=1) (authorities=1)
       [9] Josia's self-claim record / givenname / Josiah  (confirmations=1)
(authorities=1)
       [10] Josia's self-claim record / surnamename / Carberry
(confirmations=1) (authorities=1)
       [11] Josia's self-claim record / department / Psychoceramics
(confirmations=1) (authorities=1)
       [12] Josia's self-claim record / title / President  (confirmations=0)
(challenges=1)
       [13] Josia's self-claim record / same_as / Brown Library's record for
Josiah Carberry  (confirmations=1) (authorities=0)
       [17] Josia's self-claim record / honor / Nobel Prize for Humor
(confirmations=0) (authorities=0)
       [18] Josia's self-claim record / authored / doi:10.5555/1065
(confirmations=0) (authorities=0)
       [19] Josia's self-claim record / authored / doi:10.5555/1111
(confirmations=0) (authorities=0)
       [20] Josia's self-claim record / authored / doi:10.5555/1234
(confirmations=1) (authorities=1)
```