EPiC Series in Computing

Volume 78, 2021, Pages 108–114

Proceedings of the European University Information Systems Conference 2021

Organization Information gone Wild: ROR, Entity IDs and The Organization Ontology

Marius Politze[1]*

RWTH Aachen Universtiy, Aachen, Germany
politze@itc.rwth-aachen.de

Abstract

While building services for individuals from academia, uniquely identifying a person is a challenge that was widely addressed in several contexts like eduGAIN. Sometimes, alongside the "who?", information systems also need reliable information about the "from where?". During the past years several alternative standards came up to tackle that problem from different directions. In this paper we would like to introduce some of them: Research Organization Registry (ROR), eduGAIN Entities and The Organization Ontology and give an opinionated overview of how they can work together.

1 Motivation

For several years we have developed IT systems that support business processes. In many cases these systems would support a set of processes in a single facility like a university or other research institution. Only rare rarely allowing access across organizational boundaries and if only based on existing cooperative projects. With the dawn of new initiatives spanning across organizations like EOSC on international or NFDI and NHR on, german, national level more and more facilities are becoming a service provider on a national or international level. The imminent question is how to provide authentication of external users. Within the academic context, Géant established eduGAIN, a multi federation approach that provides the necessary interfaces, infrastructure and implementations for decentral authentication.

Apart from user authentication projects may also require obtaining affiliation information from the users. For federated information systems we were naturally looking for a persistent and globally unique identifier that can be used to refer to an organization. In the project Coscine [6], we utilized DFN-AAI (the German sub-federation of eduGAIN) for Authentication and ROR for identification of research organizations and faced exactly the problem of synchronization of both catalogues.

*https://orcid.org/0000-0003-3175-0659

S. Bolis, J.-F. Desnos, L. Merakos and R. Vogl (eds.), EUNIS 2021 (EPiC Series in Computing, vol. 78), pp. 108–114

2 A Wild Organization ID appeared!

Generally speaking an organization identifier should conform to the same requirements that is posed to other identifiers encountered on the web. The most prominently known in the context of academia probably is the *Digital Object Identifier (DOI)* [4] that is based on the *Handle* System [3]. The Handle system provides ways to construct, issue and maintain globally unique and persistent identifiers for (mostly but not exclusively digital) resources. In addition, it provides a (machine operable) mechanism to resolve the identifiers to their digital representation, mostly a (human readable) web page or a structured document that gives more information or access to the resources. For this purpose the handle system offers a set of HTTP endpoints that act as a resolver. While DOI and handle share the same resolution mechanism, DOI furthermore defines endpoints to deliver more elaborate, machine readable, meta data on the represented object, most importantly other, related identifiers and more subject specific information like *authors* or *title* if the DOI represents a text document like a scientific publication. The choice of the identifier system and resolver highly depends on the actual use case. A meta resolver *N2T*[1] lists more than 900 different identifier resolution systems.

When looking for an organization identifier it is as such only natural that it should fulfill the same requirements. While it would be technically feasible to create a Handle for each organization that is encountered during the lifetime of a system, these IDs would be hardly suitable for exchanging information between systems and would require constant mapping to other schemes. Hence it was our goal to find a more well-defined and curated organization identifier.

2.1 eduGAIN *Entities Database*

While there is no claim in that direction, the eduGAIN *Entities Database*[2] fulfills a lot of criteria for a catalog of academic organizations: machine-readable, curated in a distributed manner, provide a unique, relatively persistent ID and additionally a reference to their logon interface definitions. For most IdPs there are additional information like organizations' names or websites. Mapping users' origin to Entity IDs is obviously embedded into the logon flow. With more than 4200 entries worldwide the catalog is also quite extensive. What else would we ask for? By design the Entities Database only contains currently active IdPs: Organizations that do not have an IdP are not listed, also eduGAIN is somewhat limited to public academic organizations. For example the "German National Library of Science and Technology" (TIB) is not available even though they certainly are a valuable part of the academic landscape, and so are many other international private or commercial research bodies.

2.2 Research Organization Registry

This is where the Research Organization Registry (ROR)[3] comes into play as it claims to be "a community-led registry of open, sustainable, usable, and unique identifiers for every research organization in the world". ROR offers a machine-readable and CC-0 licensed data dumps and an API for more than 100.000 research organizations including commercial and academia related nonprofit organizations like DFN or Géant themselves. ROR describes their goals as follows:

- Unique and persistent IDs for organizations in the research community

[1] https://n2t.net

[2] https://technical.edugain.org/entities

[3] https://ror.org

- IDs resolve to information about the entity: human- and machine-readable
- Open API/content negotiation
- Administrative facility to correct, manage, and crosswalk data, including assertion model and syncing with other PID providers
- Public data dump
- Common and uniform metadata set

Originally ROR sourced from the GRID[4] database and that is still its primary source of information. Both identifiers additionally link to the respective entities in Wikidata[5] allowing extended automated information discovery using linked data. Hence ROR offers a quite extensive catalog and additionally allows cross linking with other identifiers. However, at the time of writing is missing the cross link to the Entity IDs from the eduGAIN Entities Database.

3 It's dangerous to go alone!

For our current project at hand, we decided to use ROR for top level research organizations while still using eduGAIN during the authentication process. Currently, this requires mapping between ROR IDs and Entity IDs to get the membership relation of a user to an organization. Both approaches also fall short modeling the internal structures like departments. Having, Entity IDs and ROR IDs, in our project we need a way of aligning the organizational information from both.

3.1 Organizations on the Web: The Organization Ontology

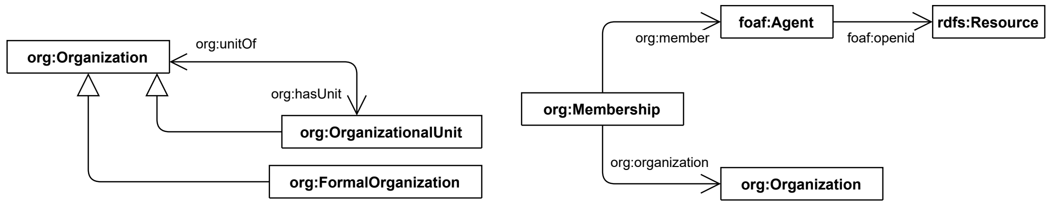
The data model of Coscine is based on RDF [8] therefore using W3C standard *Organization Ontology (ORG)* was a natural choice. ORG is designed "to enable publication of information on organizations and organizational structures" [7].

In the use case at hand we made use of two structuring element from ORG: Organization structures and memberships within organizations. For organizational structures ORG offers to model `org:OrganizationalUnits` as a subclass of `org:Organization` and provides a respective relationship `org:unitOf`. This effectively allows modeling structures like organizational hierarchies e.g. faculties and chairs as part of our university as shown in Figure 1a. In the latter case the class `org:Membership` is used to provide information about the employees working at a certain (sub-)organization. `org:Membership` makes use of the *Friend of a Friend (FOAF)* [1] ontology for referencing people within the organizations. In turn using the property `foaf:openid` allows our system to store the respective ID that a login mechanism like eduGAIN delivers for our users. Using this connection we can thus locate the user within their respective organizational structure as shown in Figure 1b.

At the root, every organization from ROR is modeled as a `org:FormalOrganization` using the ROR ID as the primary ID and assign the `org:identifier` for Entity IDs. Using ORG provided the additional benefit that it also allows modeling internal structures using organizational units and memberships. This essentially allows the data model of Coscine to provide a W3C standard compliant interface for internal structures while retaining compatibility with the existing catalogs from ROR and eduGAIN.

[4] https://grid.ac/

[5] https://wikidata.org



(a) Organization structures form trees or directed graphs of `org:OrganizationalUnit`

(b) Handling organization memberships uses `foaf:Agent`

Figure 1: Classes in ORG for modeling organization structures and memberships

3.2 ORG Practical Example

Putting all this together eventually leads to a small knowledge graph about the organization that can be easily represented using RDF as shown in the following example: At first we define the top-level organization based on the ROR ID.

```
<https://ror.org/04xfq0f34> a org:FormalOrganization ;
  rdfs:label "RWTH Aachen University" ;
  org:hasUnit <https://ror.org/04xfq0f34#ORG-78HXG> ;
  org:identifier "https://login.rz.rwth-aachen.de/shibboleth" .
```

As a second instance within the knowledge graph it is now possible to create the organizational units and link them respectively. As internal IDs we appended a fragment to the existing ROR ID allowing to at lease resolve the top level organization automatically using the resolver mechanisms provided by ROR. We add additional, internal identifiers using custom schemes using `org:identifier`.

```
<https://ror.org/04xfq0f34#ORG-42NHW> a org:OrganizationalUnit ;
  rdfs:label "IT Center" ;
  org:identifier <org-id:ORG-42NHW> ;
  org:identifier <ikz:022000> .
```

For each employee we create an instance containing personal information like the name, that could be extended with additional properties form the FOAF vocabulary. As previously discussed we use `foaf:openid` with a custom scheme referring to the logon protocol that provides that ID. In the case of eduGAIN this is mostly done according to the SAML standard [2].

```
<https://ror.org/04xfq0f34#TrU...qp> a foaf:Person ;
  foaf:name "Marius Politze" ;
  foaf:openId <SAML:TrU...qp@rwth-aachen.de> .
```

Finally, we can add a Membership node that makes the person a member of the previously defined organizational unit.

```
[] a org:Membership ;
  org:member <https://ror.org/04xfq0f34#TrU...qp> ;
  org:organization <https://ror.org/04xfq0f34#ORG-42NHW> .
```

4 Just the Begin of the Journey

Bringing together ROR and eduGAIN Entity IDs seems like a viable choice for connecting both directories. However, right now this is a manual process of looking up the correct entity ID from the Entity Database. An initial brief test showed that aligning both directories using names and institutional websites can yield high quality results. For authoritative records this would, however, need a more decent quality control.

4.1 Automation of the Matching

The general matching approach can be described as follows: Based on each of the entries in the Entity Database and their available information, try to find the most suitable ROR id within the ROR data dump. In more detail we took four different approaches:

Name uses the Entities English name and the name and aliases of ROR entries. If the names or the name and an alias are exactly the same this is considered a match.

URL uses the home pages supplied for the organization. Since some organizations provide deep links to sub pages, only the host name is considered. If they are exactly the same this is considered a match.

Wikidata uses information about public API endpoints available in Wikidata. Wikidata entries often list both, the ROR ID and the respective eduGAIN entity ID. If this pairing can be found it is considered a match.

The approaches were first run and evaluated separately and then merged using two different strategies:

Union All recorded matches are merged together equally.

Score The matches are scored based on their quality. For matching based on the name we score 2, for matching based on the URL we score 1. Since Wikidata is the most trustworthy source their information takes precedence and scores 10.

Finally for evaluation we considered the following scenarios:

Unique An entity ID matches exactly one ROR ID and the ROR ID is matched exactly once; for the scored strategies the highest score is only considered if there is no tie.

Ambiguous An entity ID that matches multiple ROR IDs or a ROR ID that is matched more than once; for scored strategy only in case of a tie.

No Match An entity ID that does not have a matching ROR ID at all.

The result of the evaluation is displayed in Figure 2. It is clearly visible that the matching approaches by name and by URL are less stable than the one involving Wikidata as a source. Especially for URLs we observe a high ratio of ambiguous matches around 18%. The Wikidata approach shows the best ratio at about 0.1% but also has the lowest unique matching rate.

With the simple union approach it becomes visible that the approaches can profit from one another as the unique matches significantly increase. At the same time, however, ambiguity also increases if approaches produce different matches for the same entity ID. Adding scores according to the quality of the sole performance of the algorithms then allows to resolve some of this additional uncertainty resulting in a ratio of about 8% while providing a match for about 68% of the Entity IDs.

For purposes of future reuse the matching application is made available as an open source project [5].

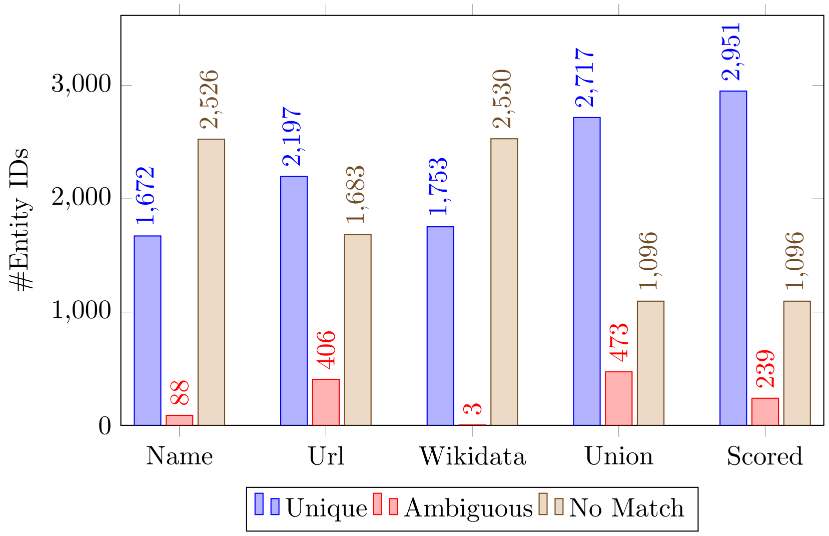


Figure 2: Results of matching eduGAIN Entity IDs to ROR IDs using different traits.

4.2 Outlook

Within Coscine we were able to bring both systems, eduGAIN Entity IDs and ROR, together and add internal organizational structures and organization memberships in a W3C conform information model using ORG. The matching algorithm allows to automatically produce an alignment of both IDs with a high accuracy.

The presented matching algorithm serves only as a proof of concept. For more productive use the matches presented have to be reviewed and verified on a sample basis. Also, the matching algorithm could be extended to include other facets in the approaches like other languages, abbreviations, country or more fuzzy string matching to account for diacritics.

Our future work in this area will entail to build a repository of machine-readable organization information. Allowing academic and research organizations to maintain their own internal structure and reuse information from others using an open source inspired, collaborative model. This repository should then be used by Coscine in order to allow respective login and organization assignment for the users.

5 Acknowledgments

The work was supported with resources granted by NFDI4Ing funded by Deutsche Forschungsgemeinschaft (DFG) under project number 442146713.

References

[1] Dan Brickley and Libby Miller. Foaf vocabulary specification 0.99.

[2] Scott Cantor, John Kemp, Rob Philpott, and Eve Maler, editors. *Security Assertion Markup Language (SAML) V2.0*. OASIS Standard, 2005.

[3] Laurence Lannom. Handle system overview. In *66th IFLA Council and General Conference*, 2000.

[4] Norman Paskin. Digital object identifier (doi ®) system. In Marcia J. Bates and Mary Niles Maack, editors, *Encyclopedia of Library and Information Sciences, Third Edition*, volume 6, pages 1586–1592. CRC Press, 2009.

[5] Marius Politze. mpolitze/matchrortoedugrain: Version 1.0.0, 2021.

[6] Marius Politze, Florian Claus, Bela Brenger, M. Amin Yazdi, Benedikt Heinrichs, and Annett Schwarz. How to manage it resources in research projects? towards a collaborative scientific integration environment. In *European Journal of Higher Education IT 2020-1*. Paris, France, 2020.

[7] W3C. The organization ontology.

[8] W3C. Rdf 1.1 concepts and abstract syntax.

6 Authors' Biography

Dr. Marius Politze is head of the group "Process and Application Development for Research" at the IT Center of RWTH Aachen University. Before that he held various posts at the IT Center as a software developer, software architect and as a teacher for scripting and programming languages. His research focuses on Semantic Web and Linked Data architectures for distributed and service-oriented systems in the area of research data management.