

OpenPIIR Product Scope and Principles

Recommendations of the ORG_ID Product Breakout Group

FINAL VERSION: 2017-July-17

Breakout Group Members: Geoffrey Bilder, Christopher Brown, Laura Cox, Trisha Cruse, Tim Devenport, Martin Fenner, Danny Goroff, Laure Haak (chair), Christina Hoppermann, Andres Mori, Andrew Pitts, Arthur Smith, Thomas Vestdam

Whereas:

- The research community¹ wants to use organization identifiers that are managed by a trusted non-profit member organization with a commitment to long-term availability of these identifiers.
- Institutions are asking for the opportunity to directly manage their own name in research information transactions, however it may not be possible to engage all institutions in managing records so we should endeavor to have a hybrid approach of both staff-curated and organization-managed records.
- There are existing organization identifier registries with specific service models, none of which involve a formal means for institution-managed records on a global scale
- There is a need for more comprehensive and non-manual mechanisms to facilitate cross-walking of existing organization identifier schemes.

We are proposing the following service components for an open Persistent Institution Identifier Registry (openPIIR). The principles and recommendations draw from conversations of the Product Breakout group about audience, use cases, personae, and curation processes. These are explored in the following documents:

- [Metadata](#)
- [Use cases](#)
- [Manual Curation](#)

Scope

The *Affiliation Use Case* should be the main focus of an open Registry: proper description of relationships between contributors, contributions, research sponsors, publishers, and employers. The complexity of describing affiliation information for researchers increases

¹ Research community includes any organization or individual associated with research, scholarship, or innovation. These include but are not limited to universities; non profit research institutes; government ministries, research facilities, and laboratories; trans-national research institutions; commercial research organizations; professional associations and scholarly societies; research sponsors and funders; publishers; vendors; researchers and research support staff.

dramatically when going further back in time, while at the same time the potential benefits of linking research outputs and their contributors to institutions decrease. For these reasons, the openPIIR should focus on active researchers and their affiliation at present and in the recent (~5 years) past.

What do we mean by affiliation?

Affiliate □ verb |əˈfɪliəteɪt| [with object] (usually be affiliated to/with) □ officially attach or connect (a subsidiary group or a person) to an organization

For the purposes of the open registry, “affiliation” describes any formal relationship between a researcher and an organization associated with researchers, including their employer, educator, funder, scholarly society, etc., and is not limited to employment.

Principle 1: The openPIIR is intended for use by the research community, for the purposes of increasing the use of organization identifiers in the community and enabling connections between organization records in various systems.

Who are the Users

There are several potential users of organization identifiers, for each sector in the research community. In addition to organizational authorities, there are organizational users, as well as researchers, software developers, and people who are researching organizations. To support these users, we propose a hybrid approach that blends staff-managed curation processes with an interface for organization authorities to view and securely manage their record. It should also provide means for public view of record data, both in a “home page” user interface and API. And, it should provide an administrative interface for customer service purposes. The support desk would work specifically with organizations on record use (and single-point cross walking), and assist with API usage. The openPIIR would not be tasked with e.g., bulk data syncing but it should work with providers to enable cross-walking.

Principle 2. The openPIIR will derive utility through a hybrid staff- and organization-curation service model, and will encourage cross-talk between existing registry providers.

Principle 3: Access to organizations for managing openPIIR records shall be via permission. openPIIR staff will be responsible for granting record management permission.

Product Principles

There are several existing organization identifier registries. For the openPIIR to be useful, it needs to augment the current offerings in a way that is open, trusted, complementary and collaborative, and not intentionally competitive. It needs to provide a service that the community finds helpful and not duplicative. The distinctive competencies of the openPIIR are to provide the research community a Registry with open governance, provide institutions with a means to manage their own organization record, and to serve as a central point for communicating with

current vendors to provide updates -- and in doing so to increase the use of organization identifiers in the community and enable connections between organization records in various systems.

In general, the openPIIR should have the following features and services:

- Unique and persistent IDs for organizations in the research community
- IDs resolve to information about the entity: human and machine readable
- Open API/content negotiation [tie to business model?]
- Record editing interface for authorized entity representative
- 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

The openPIIR data model should focus on organization levels that are pertinent for the affiliation use case. The openPIIR data model should have a minimal required metadata set to describe an organization, such as physical location (headquarters and/or mailing address city+country), electronic address (URL), persistent ID, organization name in official language and if available in Latin characters, and provenance (who made what assertion, when, and from what source). In addition to the required data elements, the registry should include optional fields such as physical address (lat/long, street address), crosswalks to other IDs, variant names, organization sector, and [copyright requirements](#).

openPIIR data should be both human and machine readable, available through web interface and APIs. Documentation about data models and APIs should be freely available to the community. Required openPIIR data must be openly reusable.

Principle 4. The openPIIR will focus on the organization levels that are most pertinent for the affiliation use case (who employs, who educates, who funds, etc).

Principle 5. The openPIIR will require data elements for each record sufficient to uniquely identify the organization.

Principle 6. openPIIR documentation and required data will be available for use under a recommended [Open Definition conformant](#) license, in human and machine readable formats.

Data Management

Building the openPIIR requires community governance as well as deep community involvement in data quality. The openPIIR should be started using seed data, available from existing providers without license restrictions and meeting minimum metadata requirements for unique identification. Data may be obtained using an API and/or via bulk donation.

Principle 7. The openPIIR will seek seed data from organization identifier providers who serve the research community, whose data meet the metadata requirements, and which data are available under a recommended [Open Definition conformant](#) license.

The openPIIR should have dedicated staff to manage records and a means for organizations to provide input for record creation and updates. The initial focus of the openPIIR will be staff curation of seed data; organizations may provide input via the PIIR support desk. After the initial seeding, the openPIIR will build out tools to support record self-management by organizations. This means the openPIIR will need to build hybrid record management processes - and engage with the community to encourage involvement. This raises specific questions of transparency, control, data quality, and effort. Without transparent record management criteria and accountability of curators, the quality of the openPIIR will suffer.

There is an inevitable need for processes to deal with record duplication, erroneous records, and relationships between records. Curators may have to make decisions that not only affect the metadata contained in the record or record creation but also establish relationships between these, through record redirection and record obsolescence. In cases of duplication or merger, records should be redirected to the relevant active record. In the rare cases where erroneous records have been created, there should be an obsolescence mechanism that identifies these records as such. When possible, the openPIIR will engage with the entity itself to make record changes, and all changes (assertions) will be marked with provenance indicating who made the change (organizational authority, openPIIR staff) and a timestamp.

Principle 8. There will be open criteria and documented processes for inclusion/exclusion, creating, merging, and deprecating an openPIIR record. openPIIR staff will have authority to enforce record management criteria.

Principle 9. Record changes will be tracked and recorded using an open provenance model. openPIIR records may be deprecated, but no assigned identifier will be deleted. .

Principle 10. The openPIIR will maintain a customer support ticketing system and an open knowledge base.

Appendix. Summary of BG tasks from OrgID website:

1. What registry data are open? What does “open” mean in this context?

Principle 6. openPIIR documentation and required data will be available for use under a recommended [Open Definition conformant](#) license, in human and machine readable formats.

2. What are the specific components of the registry MVP? (MVP = Minimum Viable Product is a product with just enough features to gather validated learning about the product and its continued development.)

In general, the openPIIR should have the following features and services:

- Unique and persistent IDs for organizations in the research community
- IDs resolve to information about the entity: human and machine readable
- Open API/content negotiation
- Record editing interface for authorized entity representative
- 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

The initial focus of the openPIIR will be staff curation of seed data; organizations may provide input via the PIIR support desk.

3. How is this different from existing registries? What responsibilities does the registry have with respect to current stakeholders? Does it need to be a new registry? Do existing providers offer the product that meets these aims? How would they need to change in order to solve these problems?

There are several existing organization identifier registries. For the openPIIR to be useful, it needs to augment the current offerings in a way that is open, trusted, complementary and collaborative, and not intentionally competitive. It needs to provide a service that the community finds helpful and not duplicative. The distinctive competencies of the openPIIR are to provide the research community a Registry with open governance, provide institutions with a means to manage their own organization record, and to serve as a central point for communicating with current vendors to provide updates -- and in doing so to increase the use of organization identifiers in the community and enable connections between organization records in various systems.

4. What features can/should be part of a subsequent version of the registry?

After the initial seeding, the openPIIR will build out tools to support record self-management by organizations. This means the openPIIR will need to build hybrid record management processes - and engage with the community to encourage involvement. This raises specific questions of transparency, control, data quality, and effort. Without transparent record management criteria and accountability of curators, the quality of the openPIIR will suffer.

5. What are the technical considerations for curation, and for updating of records? Given that not all organizations will engage, what solutions do we need to have in place (curation vs. self-management)?

There is an inevitable need for processes to deal with record duplication, erroneous records, and relationships between records. Curators may have to make decisions that not only affect the metadata contained in the record or record creation but also establish relationships between these, through record redirection and record obsolescence. In cases of duplication or merger, records should be redirected to the relevant active record. In the rare cases where erroneous records have been created, there should be an obsolescence mechanism that identifies these records as such. When possible, the openPIIR will engage with the entity itself to make record changes, and all changes (assertions) will be marked with provenance indicating who made the change (organizational authority, openPIIR staff) and a timestamp.

6. How does the registry enforce uniformity of data, e.g. data structure and validation for business rules vs. open contribution?

openPIIR data are managed by staff and permissioned contributions (see #7) by organizations. The openPIIR data model should focus on organization levels that are pertinent for the affiliation use case. The openPIIR data model should have a minimal required metadata set to describe an organization, such as physical location (headquarters and/or mailing address city+country), electronic address (URL), persistent ID, organization name in official language and if available in Latin characters, and provenance (who made what assertion, when, and from what source). In addition to the required data elements, the registry should include optional fields such as physical address (lat/long, street address), crosswalks to other IDs, variant names, organization sector, and [copyright requirements](#).

7. What types of roles/permissions are needed for the members of the organizations in the registry?

Principle 2. The openPIIR will derive utility through a hybrid staff- and organization-curation service model, and will encourage cross-talk between existing registry providers.

Principle 3: Access to organizations for managing openPIIR records shall be via permission. openPIIR staff will be responsible for granting record management permission.

8. Who is the person/people that can assert relationships between organizations in the registry? What sources of official information are available? How do we resolve conflicts?

Principle 8. There will be open criteria and documented processes for inclusion/exclusion, creating, merging, and deprecating an openPIIR record. openPIIR staff will have authority to enforce record management criteria.

Principle 9. Record changes will be tracked and recorded using an open provenance model. openPIIR records may be deprecated, but no assigned identifier will be deleted. .

Principle 10. The openPIIR will maintain a customer support ticketing system and an open knowledge base.

9. How do we account for the subtlety of organizations with regard to permissions and change history (merges, splits, etc.)? Where do we start? Do we build a historical record for what has already happened in the past or do we start from this point forward?

The *Affiliation Use Case* should be the main focus of an open Registry: proper description of relationships between contributors, contributions, research sponsors, publishers, and employers. The complexity of describing affiliation information for researchers increases dramatically when going further back in time, while at the same time the potential benefits of linking research outputs and their contributors to institutions decrease. For these reasons, the openPIIR should focus on active researchers and their affiliation at present and in the recent (~5 years) past.

10. What data sources can pre-populate or seed the registry if it is to be a new one?

Building the openPIIR requires community governance as well as deep community involvement in data quality. The openPIIR should be started using seed data, available from existing providers without license restrictions and meeting minimum metadata requirements for unique identification. Data may be obtained using an API and/or via bulk donation.

Principle 7. The openPIIR will seek seed data from organization identifier providers who serve the research community, whose data meet the metadata requirements, and which data are available under a recommended [Open Definition conformant](#) license.