



EMBL-EBI - Providing harmonized information about organisations in identifiers.org using ROR registry

Overview

European Bioinformatics Institute is Europe's largest provider of public biomolecular data resources. The institute is co-located with Elixir Hub and partnered with many relevant EU projects, including EOSC-Life, FREYA, and BY-COVID. In addition to supporting life sciences, EMBL-EBI is increasingly collaborating with other domains, e.g., social sciences in COVID-19 research.

EMBL-EBI provides consistent access to life science data by leveraging compact identifiers through the Identifiers.org resolution service. Anyone can use identifiers.org for consistent and globally unique references to objects in registered data collections. For example, researchers can use our link to their objects using our URIs and these will remain resolvable even if the object's online location changes.

This use case focuses on the addition of metadata artefacts to the identifiers.org registry. **This mainly consists of the use of the ROR registry as a source of truth for institution data.** Allowing the identifiers.org registry to correctly link its resources to the institution in charge of managing it and verify its own metadata on them based on RORs metadata.

Introduction

The ambition of EMBL-EBI in this use case **is to curate and update components of Identifiers.org.** The updates will be aligned with community standards and needs while following FAIR practices. Our tasks for this goal revolve around verifying the set of annotations available at the identifiers.org registry meets the community guidelines for PIDs. Thus including:

- Adding or removing attributes as necessary for our system to provide minimal and sufficient context to PIDs
- Curating the registry to verify that the correct metadata is associated with entries

For this means, **we added ROR identification to the institutions in our registry.** The ROR registry of institutions acts here as a metadata catalogue and it is a source of truth on information such as home pages, type of organization and other identifiers for organizations.

In the identifiers.org registry, we link institutions to their ROR ID when possible, the associated ROR ID of the institution in charge of a resource can be seen in our registry pages. **This association is performed manually by our curators.** This association provides us with the means to identify institutions through a persistent identifier but our registry also stores other metadata on them locally that overlaps with other metadata managed by ROR. For example, both ROR and identifiers.org contain institution name and home page. **This provides use with the opportunity of updating our registry by using the ROR registry as a source of truth.** However, considerable curation effort is needed to add and maintain ROR IDs in current and future entries. Additionally, ensuring that the metadata from ROR is equal to the one in our registry takes considerable time.

The addition of ROR IDs to our registry is a small change that leverages both the identity of institutions and allows for cross validation of other metadata values. **It is an example of a minimum change that makes our PIDs more useful, but even this small example considerably increases our curation efforts.** This displays the importance of carefully considering what additions make sense in our type of service.

In the future, we wish to consider to expand the registry with associations to other metadata registries such as FAIRsharing and Wikidata and allow users to resolve compact identifiers to their metadata at these providers and use metadata from these services in our search services.



Henning
Hermjakob
EMBL-EBI



Renato Juacaba
Neto
EMBL-EBI

Key topic

Interoperability Metadata & Ontologies PIDs

Scientific domain

Life science Physical and Technical Sciences

Leading organisation

EMBL -EBI

Challenges that need to be addressed

There is a need to find solutions to overcome some challenges related to PID practices and associated metadata, especially in cases where billions of data objects are included in a large number of resources, where:

- Repositories produce a large amount of entries,
- frequent data updates occur, and
- there is a high barrier to the adoption of global PID systems.

For these, the **identifiers.org registry** tries to keep the minimum amount of metadata to minimize the work to keep these up to date with repositories. Also, adding these associations will require our team to dedicate further efforts to keeping these up to date with providers. Such effort could make securing enough financing difficult. Thus, we aim for a solution where data providers are the ones in control of the associated metadata with zero or minimal maintenance from our team.

Our ROR ID association works illustrates this as the small addition of ROR IDs to institutions in our registry adds considerable curation effort to maintain overlapping metadata in sync and to ensure our registry entries have a valid ROR ID when available.

Expected impact of the Use Case

More efficient use of PID metadata will greatly improve interoperability and discoverability, two of the main FAIR principles, by using well-defined and harmonized data. For example, users can use ROR IDs from our registry to link entries with other resources in knowledge graphs that also use ROR IDs. **Co-designing the PID practices through EOSC alignment provides a research environment which is responsive to the needs of the various research communities.** EMBL-EBI is experienced in working with metadata standards, integration, discoverability, and display through its work with the Identifier.org service. **Furthermore, EMBL-EBI can bring its valuable expertise from its partnership with a large international consortium, the European Bioinformatics Institute, which is a public biomolecular data resource provider.** EMBL-EBI also has vast expertise gained from its partnership in many relevant EU projects, among others EOSC-Life, FREYA, and BY-COVID.

Expected impact of the Use Case

The use case will bring evolving Identifiers.org practices into a broader EOSC context and provide solutions to overcome some challenges related to PID practices. Making:

- Our registry more accessible through the use of semantic artifact catalogs
- Working towards determining minimum metadata for PID providers

2



Henning Hermjakob
EMBL-EBI



Renato Juacaba Neto
EMBL-EBI