Coordinated Research Infrastructures Building Enduring Life-science services
- CORBEL -

Deliverable D6.3
Delivery of Toolkit supporting community ontology mapping and development

WP6 – Data access, management and integration
Lead Beneficiary: EMBL-EBI
WP leader: Helen Parkinson (EMBL-EBI) and Carole Goble (UNIMAN)
Contributing partner(s): EMBL-EBI, UNIMAN, Lygature, UMCG

Contractual delivery date: 31-12-2019
Actual delivery date: 07-02-2019

Authors of this deliverable: Simon Jupp (EMBL-EBI), Helen Parkinson (EMBL-EBI), Carole Goble (UNIMAN), Nick Juty (UNIMAN), Jan-Willem Boiten (LYGATURE), Morris Swertz (UMCG)

Grant agreement no. 654248
Horizon 2020
H2020-INFRADEV-1-2014
Type of action: RIA

Content

Executive Summary

This deliverable reports on the services that were developed for CORBEL to support access to ontologies and ontology mapping services. The service development has been driven by requirements from BMS ESFRIs in WP3 and 4 and international engagement. Tasks have included:

Developing a repository of biomedical ontologies that supports web based and API based access and visualisation of ontologies. This was undertaken within OLS, a recognised 'Recommended Interoperability Resource' in ELIXIR

- Extending and improving the Zooma repository of data to ontology mappings that can be used to predict ontology mappings based on prior curation knowledge
- Developing a new ontology cross-referencing service that allows users to map between ontologies and medical vocabularies from the UMLS
- Extending the BioBankUniverse service for granular cross biobank variable matching
- Development and refinement of tools to support the construction of new ontologies and application specific ontology views by domain experts that is being adopted by projects outside of CORBEL
- Improving APIs, query Interfaces and services comprising semantic infrastructure based on expert and user feedback from CORBEL partners and international users

Project objectives

This deliverable has describes delivery of WP6 objectives 3-6, described in Task 6.2 of the Description of Work:

- Deliver standards compliant web and programmatic based access to ontologies and linked open data from an ontology access service using semantic web ontologies.
- Deliver access to validated data-ontology maps with provenance between BMS ESFRI data and ontologies.
- Deliver ontology-ontology mappings with provenance supporting data integration across infrastructures.
- Deliver the semantic infrastructure supporting cross-domain use of data identifiers, minimal data standards, data types, tools, and domains required by BMS ESFRIs.

Detailed report on the deliverable

Background

WP6 has been focused on data access and interoperability through the development and provision of services to support data identity and semantics. In this deliverable we report on the services developed to provide access to terminology standards and the application of these to new, cross (inter-) infrastructure standards. These services meet a demand from RIs to access common

standards for describing data to ensure that data on equivalent entities or concepts can be readily discovered and integrated.

A number of ontology based standards have emerged that cover everything from small molecules, through to animal physiology, disease, biological systems and ecology. In some cases these ontologies may overlap, thereby providing an alternative view of a particular domain from a different scientific perspective. Integration of such information can be used to support different use cases. This is especially evident in the area of disease ontologies, where multiple representations of diseases exist, having been articulated for particular communities or developed to address specific use cases. The services developed in this work package aim to assist researches when accessing and adapting these ontologies to support their own use cases.

Use Cases

Access to ontologies

This use case focused on provision of access to ontologies. The Ontology Lookup Service, developed at EMBL-EBI, provides web based and REST API search services for approx 200 highly accessed public biomedical ontologies. Thus, OLS acts as a hub for biomedical ontologies and is a critical service infrastructure component for building a sustainable and interoperable data ecosystem in the life sciences. This work is aligned with global standards initiatives; we have worked with the ontology standards community (OBO Foundry[1]) to develop a standard metadata model for describing an ontology[2] that is used to register new ontologies into the OBO library and subsequently used to load these ontologies into OLS. This mechanism is also available for use by other registries, providing a common format for accessing ontology metadata along with a process for community authoring of this metadata, utilising the GitHub platform.

OLS undergoes a scheduled update nightly to ensure that each of its over 200 hosted open ontologies are accurate with respect to official ontology source. OLS thus ensures that the community is able to access the latest versions of each ontology, whilst also providing access to the log of the ontology history. This functionality allows users to track how and when terms have changed in the individual ontologies that they may use, between releases. OLS hosts many ontologies of relevance to the CORBEL project including the ontologies developed with EMBRC for WP4, and several public disease ontologies used by BBMRI and EATRIS.

OLS is also available to be deployable locally, allowing institutions to run their own instances of the services. This may be useful in scenarios where privacy or sensitive data is of concern. These private instances are implemented to allow access to additional terminologies and ontologies that are not hosted in OLS; it is possible to write custom loaders to host terminologies such as SNOMED for those users who have the appropriate licence.

In D6.2 we reported on that the Ontology Lookup service (OLS) is the preferred registry for the listing of ontologies used in the CORBEL Open Call projects[3]. Within the projects instigated during those

[1] http://www.obofoundry.org/
[2] http://www.obofoundry.org/faq/how-do-i-edit-metadata.html
[3] https://www.corbel-project.eu/open-call.html

Calls, we noted that Gene Ontology was the most used vocabulary across projects (specifically stated in projects 2277, 2281, 2234, 2354).

All ontologies in OLS are also being made available as linked data through the EMBL-EBI RDF platform[4]. This allows for more fine grained access to the data held in the platform, which complies to ontological constraints (eg. organism or tissue). This public SPARQL endpoint is aimed to address the requirements of more advanced users who require access to the underlying RDF triples that constitute the ontology. By having the ontologies loaded in the RDF platform, users are able to query data held in the platform, such as gene expression data from a number of organisms, directly using the ontologies to integrate related datasets, for example.

Visualisation of ontologies

OLS has developed a number of visualisation tools for searching terms in ontologies, visualising ontologies as a tree (Figure 1) and a graph based visualisation (Figure 2). Providing good ontology visualisations that cater for different user needs is a challenge, as is presenting large or complex ontologies. To ensure the utility of these tools, the OLS team worked with 'User eXperience' (UX) teams to explore different ontology visualisation approaches. Consequently, the resulting visualisations have been well received by the community and featured in a recent article about cross resource approaches to ontology visualisation from an SME, indicating the relevance and high profile of the tools for the semantics and interoperability developer community as well as the user community and intra project specialists[5].

In order to promote a consistent user experience across the ontology user landscape, these visualisation tools have been made available as javascript widgets that can be embedded in external applications. This allows other providers and ontology resources to express ontological information in a consistent way, with respect to both the visual components, as well as exposing those metadata elements that are crucial to enable interoperability between ontology resources. Examples of the embedded ontology widgets can be seen in diverse internationally developed projects including: visualising the Environment Ontology (ENVO[6]), the data curation system for cell lines registered in the HiPSci project[7], the data submission systems for agricultural genomics data in the COPO portal (http://www.earlham.ac.uk/copo) and as a tool to support discovery of sickle cell data in the Sickle Cell Disease Portal[8]).

[4] https://academic.oup.com/bioinformatics/article/30/9/1338/234645/The-EBI-RDF-platform-linked-open-data-for-the-life
[5] https://www.scibite.com/exploring-ontology-visualisation-techniques-for-biological-data/
[6] http://environmentontology.org/Browse-EnvO
[7] http://www.hipsci.org
[8] http://scdontology.h3abionet.org

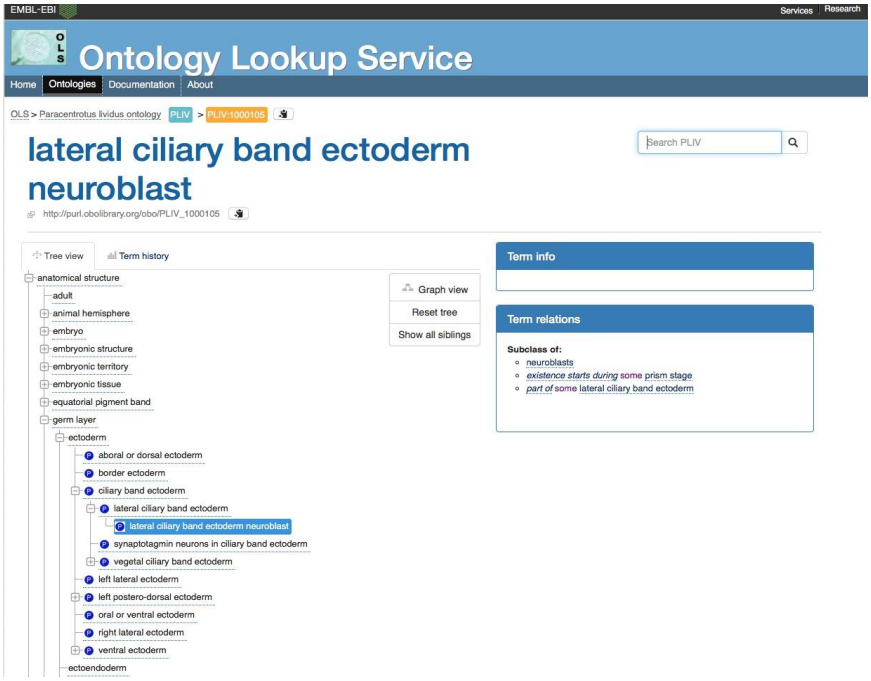

Figure 1. Screenshot of the Urichin Ontology (URCH) in the EMBL-EBI Ontology Lookup Service (OLS) developed for EMBRC. The image shows how OLS displays the classification of this term in a tree based visualisation.

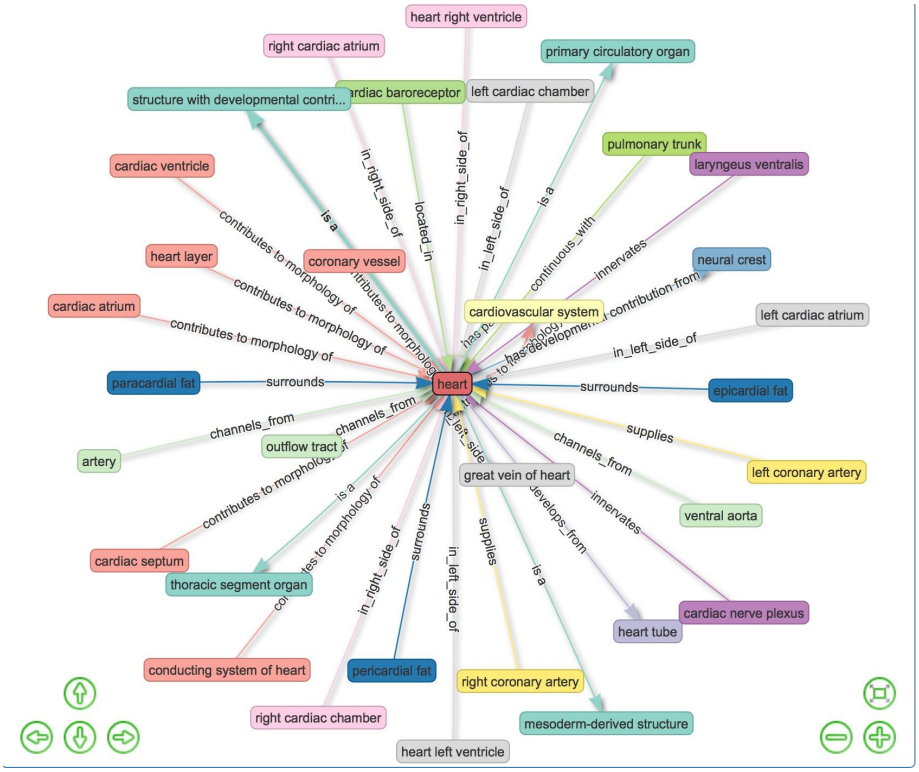

Figure 2. Screenshot of a graph based ontology view in EMBL-EBI's OLS. This shows the term for "heart" in the UBERON anatomy ontology along with all the relationships to other terms in that ontology.

Annotating data with ontologies

Tagging or annotating structured data with ontology terms is a common curatorial activity that is often a crucial part of the data FAIRification process[9]. One of the challenges in this process is how to ensure data is consistently curated with the appropriate ontologies and how to apply this process at scale to both newly generated data, and how to retrofit this curation as the large amounts of data sitting in existing data archives. This challenge is recognised by BBMRI, ELIXIR core data resources and other efforts which are generating text based curation. Increasing amounts of biological data are being deposited in public repositories, but generating high quality annotations is costly and time consuming which leads to an annotation gap. In order to obtain maximum value from the coverage offered by current ontologies, it is important to reduce this annotation gap by making it easier and more efficient for data providers to annotate using ontology terms, and to automate as much of the curation process as possible. This will free expert curators to concentrate on producing annotations for previously unseen data, rather than continually repeating the same manual annotation processes. By capturing the knowledge about data annotations, decoupled from the underlying databases, it becomes possible to integrate annotation data from a variety of different sources, and reuse this knowledge earlier in the submission or curation processes, and to time stamp different annotation versions. The Zooma and SORTA system have been developed to address this issue and provide access to annotation knowledge provide by expert curation. This data can be be used to predict annotation of new unseen data.

Over time the curators can draw from experience from prior annotations to inform future annotations. For example, the property values 'M' or 'F' can be automatically annotated to ontology terms for 'male' [EFO 0001266] or 'female' [EFO 0001265] respectively when they are accompanied by a property type of either 'sex' or 'gender'. Similarly, the property value 'arm' is found with a property type of 'organism part or 'genotype', requiring a different ontology annotation in each case. There are other cases where the annotation is less obvious, for example, 'non diabetic' is frequently annotated to the EFO term 'control' [EFO 0001461], usually based on the context of the property within an experiment. Over time the annotators build a rich collection of cases where an annotation can hold.

The property type / property value pattern for describing data is prevalent across many databases in the life sciences. This is evident from the BioSamples database that brings together sample data from multiple databases. The annotation of data on this scale (2 million BioSamples at present, growing to 12 million in the near future) presents a major challenge which requires solutions that help spread the cost and burden of annotation across resources. The Zooma knowledge base is integrated with OLS and users can query Zooma for pre-existing annotations and can detect ontology terms in OLS where Zooma doesn't find a curated match. Zooma has both a graphical user interface (GUI) and a programmatic access via API. Results are tuned for precision rather than recall, so matches are likely to be accurate when matched, allowing curators to have high confidence in the results.

[9] https://rd-connect.eu/what-we-do/data-linkage/fairification/

Mapping between ontologies

Through working with the use-case partners on WP3 and WP4, and with the wider semantics and interoperability community, we identified a need to access rare-disease ontologies and ontology mappings to support data search in the BBMRI-ERIC directory. These include a need for mappings between ICD-8, ICD-9, ICD-10, SNOMED (various versions), HPO, OMIM, and Orphanet to provide access to community curated cross-ontology mappings. This need stems from the fact that clinically oriented data will often be coded to different ontologies such as SNOMED or ICD, which come with license restrictions and as some ontologies, e.g. ICD, are is not well integrated to the wider open ontology landscape. These limitations necessitate that these ontologies be translated or mapped to publically developed open ontology standards in order to integrate datasets annotated to different (clinical) standards. For instance, the UK Biobank data contains health information on participants that is annotated using ICD-10 codes, which are currently being remapped to pubic ontologies using tools developed within CORBEL.

A new service 'Ontology Cross Reference' or' OxO' was designed and implemented during this project. Development was informed by community requirements, as well as by participation by the OLS developer team in the Pistoia Alliance Ontology Mapping Challenge[10] which was designed to map between several human disease and phenotype ontologies in use by pharma. As industrial organisations are known users of OLS (we are aware of several OLS installations in pharma/agrifood companies) we participated in this challenge to ensure that the services developed during the project met industrial as well as academic needs, for example in translation of biological knowledge relevant to drug discovery. Results of the Pistoia Challenge are available[11] and have informed both the design and implementation of OxO, as well as benchmarking with several datasets which has delivered new requirements, improving the service for all its users.

We report on ontology mapping performed using OxO for rare and common disease ontologies needed by the BBMRI-ERIC directory, which are now made available through the OxO service (Table 1). The total number of mappings and reciprocal mappings are reported, ie. bi-directional mappings are significant and need to be recorded, since in some cases a single term may map to multiple terms in another ontology. The mappings also include OxO inferred mappings of 'path' distance 2, where a mapping from one ontology to another can be made through a 3rd party or intermediary ontology e.g. if an OMIM and an Orphanet term have been mapped to the same SNOMED term, then we can infer a mapping between the OMIM and Orphanet term. By mapping across multiple ontologies we are able to discover potentially new and interesting mappings that have previously lain undiscovered. This technique has demonstrable utility; several pharma companies, through the Pistoia Alliance ontologies mapping project, have used this strategy to predict ontology mappings for internal (proprietary) vocabularies to those that are publicly available. Figure 3 shows a screenshot from OxO for the available mapping to ICD-10.

[10] https://pistoiaalliance.atlassian.net/wiki/spaces/PUB/pages/43089928/Ontologies+Mapping+Resources
[11] http://www.pistoiaalliance.org/projects/ontologies-mapping/

	ICD-9	ICD-10	SNOMED-CT	HPO	Orphanet	OMIM
ICD-9		5,265	10,357	1,069	1,058	766
ICD-10	5,613		7,220	1,252	2,588	1,582
SNOMED-CT	21,577	16,345		8,315	6,068	4,537
HPO	948	1,154	3,570		849	526
Orphanet	1,690	7,239	3,185	849		8,734
OMIM	1,576	6,262	2,997	567	10,478	

Table 1. Ontology mappings between human medical and rare disease ontologies required by the BBMRI-ERIC directory. Mappings accessed January 2019.

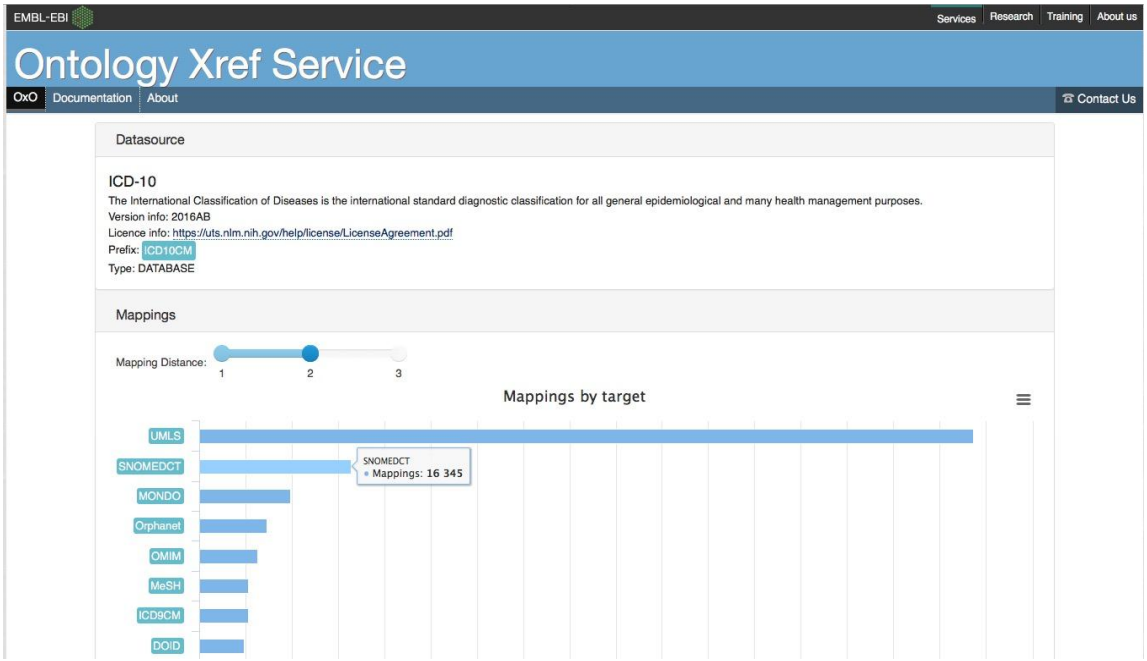

Figure 3. Screenshot of the ICD-10 page in OxO, providing an overview of the number of mappings within OxO from ICD-10 to other ontologies. The mappings distance 'path' can be modified to allow 'transitive inference' across ontological space through 3rd party ontologies (see above). In this image the mapping distance is set to 2 and OxO displays that there are 16,345 mappings from ICD codes to SNOMED-CT codes.

Building ontologies

This use case is for supporting communities in building and extending ontologies for their own specific needs. In some cases this may mean providing a specific view over existing ontologies to support a particular application or community perspective. For example, the UBERON anatomy ontology provides a reference terminology for all animals, while the Human Cell Atlas (HCA) project only requires a human view of anatomy. Hence, to optimise a generic tool for use within a specific

community or for a particular application, tools are needed to 'hide' the irrelevant terms (non-human in this example) from HCA curators.

Developing or extending ontologies is a specialised task with specialist tools that are often a (technical) barrier for domain (scientific) experts and user communities, hindering their ability to participate in the process, and potentially slowing the addition of scientific knowledge. In CORBEL we looked to develop simpler tools and processes that allow domain experts to contribute more easily and efficiently to ontology development and evolution. This has been demonstrated through de-novo ontology construction for EMBRC to support the annotation of genomic data for three marine metazoan organisms.

The Marine Invertebrate Models Database (http://marimba.obs-vlfr.fr) being developed by EMBRC partners in CORBEL WP4, required a number of new ontologies to describe morphological features for three animals (jellyfish, sea urchin and amphioxus). The ontologies developed for each organism capture specific terminology with stable identifiers for anatomy, cell types and developmental stages. The anatomical terms are organised in a partonomy and cover the entire life-cycle stage of the organism from embryogenesis to adulthood. Terms for developmental stages are provided and additional relationships have been added that capture how the stages typically progress and how the anatomical structures emerge in early development. These terms are being used to provide consistent annotation of gene expression data and images in the MARIMBA database and the ontology will provide new opportunities for querying, categorising and visualising the data (Figure 5).

WP6 provided support and tooling for developing these ontologies, as well as hosting them through the OLS when released. An ontology development pipeline was created that allowed domain experts to construct the ontologies using spreadsheet formats, with which they were more familiar (See Figure 4). Two spreadsheet templates were provide for each ontology, one for capturing information about the development stages, and one for the anatomy terms. The experts were asked to provide labels, synonyms and definitions for each term along with details of how the terms were related to each other. The data collected in the spreadsheets was then put into a pipeline developed in WP6 that utilised the ROBOT (http://robot.obolibrary.org) software to convert spreadsheet data into an ontology format (versions of the ontology were generated in both the OBO and OWL format).

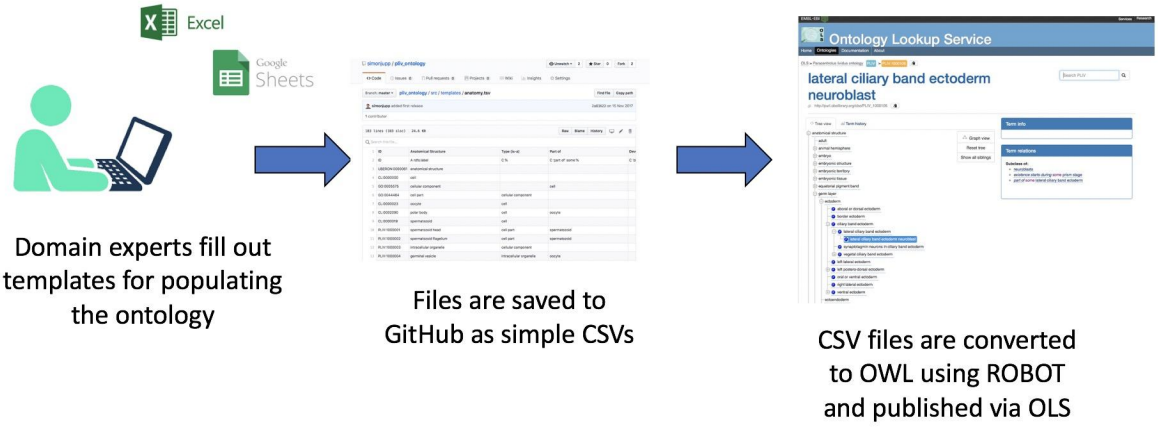

Figure 4. Overview of ontology building pipeline where domain experts fill out ontology templates, that are stored on GitHub and converted to OWL files using ROBOT. Once the OWL file is published via GitHub they are picked up and indexed in OLS through an automated process.

This pipeline serves as a prototype or template for how future ontologies could be developed for other domains. Importantly, this pipeline ensures that the ontologies being developed are compliant to community standards established by the Open Biological Ontology (OBO) foundry. Additionally, each ontology will be submitted for inclusion in the OBO library (http://obofoundry.org), as standard practice, enable uptake by other users and communities. Terms in each ontology are also accompanied by a canonical URI identifier, providing a means for standardised access to term information, as well as facilitating mapping (other ontologies) where necessary through additional OLS tools.These tools and services thereby facilitate ontological interoperability and mapping in a consistent and community approved manner.

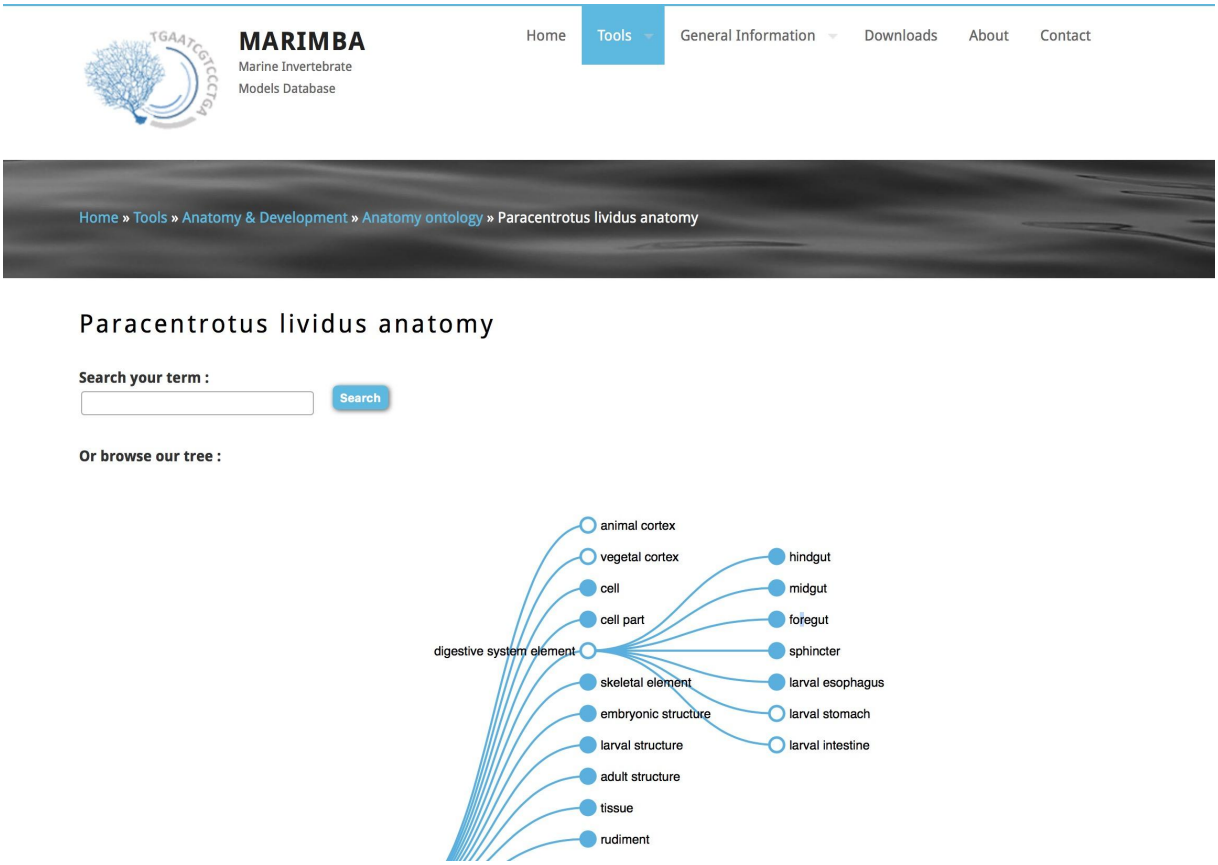

Figure 5. Browsing the URCH anatomy ontology in the MARIMA database. Further developments are planned to enable the viewing of gene expression data and images annotated with selected terms from the URCH ontology.

Ontology name	Organism	Source repository	Registry url (Not yet public)
CHEM	*Clytia hemisphaerica*	https://github.com/EBISPOT/chem_ontology	https://www.ebi.ac.uk/ols/ontologies/chem
URCH	*Paracentrotus*	https://github.com/EBISPOT/pli	https://www.ebi.ac.uk/ols/ontolo

	lividus (urchin)	v_ontology	gies/pliv
AMPH	*Branchiostoma lanceolatum* (amphioxus)	https://github.com/EBISPOT/amph_ontology	https://www.ebi.ac.uk/ols/ontologies/amph

Table 2. This table summarises the ontologies developed for the Marine Metazoan Model Organism databases. It includes the URL to the source ontology file and the link to the ontologies in the EMBL-EBI OLS.

The pipeline developed for EMBRC was successfully deployed in other project settings including the development of the sickle cell disease ontology (SCDO), an integrative and universal knowledge representation system for SCD, providing a SCD common vocabulary for the disease verified by international SCD experts. This is an inter- and multi-disciplinary collaborative project in which researchers and practitioners from different backgrounds with expertise in SCD, as well as in ontology design, are contributing. An initial set of concepts for the ontology were generated using existing biological databases and ontologies, supplemented with literature-based text mining, together with SCD management guidelines and standards of care from diverse countries. In three workshops organized by the Sickle Africa Data Coordinating Center (SADaCC), experts agreed on SCD concepts and their properties, as well as on relationships between these concepts and other axioms necessary for translating current SCD knowledge into an explicit computer-interpretable format to enable effective semantic integration of heterogeneous data, interoperability within and across areas, as well as SCD knowledge transfer. The data collected was transformed into the Web Ontology Language (OWL) using the ontology building pipeline developed (see above). This ontology has been published and is hosted through the OLS[12] run by EMBL-EBI.

The code

OLS is developed as a Java application and the source code is available under a open Apache 2.0 licence from GitHub[13]. Documentation on how to use OLS and setup a local installation is available[14]. The OLS service can also be deployed as a container using Docker. The OLS visualisation widgets are available as standalone javascript libraries that are published through the node package manager (npm). These include the autocomplete widget[15], the tree view widget[16], and the graph view[17].

The OxO service is also Java based and the source code is available under an open Apache 2.0 licence from GitHub[18]. Documentation on how to use the OxO REST API are online[19]. OxO includes a number of python libraries for loading new mappings into OxO and for predicting new mappings. These scripts can be run locally to set up a local OxO instance or generate new mappings that can be shared back to the community by submitting them back the OxO service running at EMBL-EBI. The goal is to

[12] https://www.ebi.ac.uk/ols/ontologies/scdo
[13] https://github.com/EBISPOT/OLS
[14] https://www.ebi.ac.uk/ols/docs/index
[15] https://www.npmjs.com/package/ols-autocomplete
[16] https://www.npmjs.com/package/ols-treeview
[17] https://www.npmjs.com/package/ols-graphview
[18] https://github.com/EBISPOT/oxo
[19] https://www.ebi.ac.uk/spot/oxo/docs

foster a community where curated mappings can be shared through the OxO service and to avoid duplication of mapping effort by multiple groups.

The Zooma code is hosted on GitHub[20] and further documentation on how to use Zooma and the API, as well as hosting Zooma locally, can also be found online[21] (see above for description).

The recent publication of UMCG's BiobankUniverse[22] semantic mapping tool provided a dissemination opportunity and promoted take up of the tools by several new EU projects (LifeCycle, BBMRI, RD-Connect) and contributed to future sustainability and generalisation of the tools. These tools use comparable approaches to the EBI hosted tools and together these tools are tuned for different use cases. For example, an implementation and consolidation of biobank/cohort data harmonization services in the MOLGENIS platform provided user access to annotated data using BiobankUniverse. In addition CORBEL supported Implementation of Bioschema into MOLGENIS to promote systematic findability of patient-mutation database in rare disease and to promote findability of sample collections in the BBMRI-ERIC Directory of biobanks (collaboration with ELIXIR's work on BioSchemas, this is an example of a CORBEL implementation). BioBankUniverse addresses granular and specific matching challenges at the cross biobank variable level and therefore operates at a deeper level of granularity than the EBI hosted tools. It now has been adopted as part of newly funded EU project EUCAN-Connect to be applied as tool to aid data harmonisation, i.e., to map epidemiological phenotype data items from a large number of cohorts onto one standard to enable pooled analysis of all these cohorts to reach statistical power to elucidate subtle factors such as what environmental factors positively influence child development. This is a good example of CORBEL's infrastructural development usage by specialist projects and will contribute to the infrastructure's sustainability. Additionally this extension makes the data resource compliant with the EOSC guidelines for dataset description published by EOSCpilot ("EDMI"). As a bonus users can now find the MOLGENIS data via Google dataset search[23] improving project visibility and findability in support of the FAIR principles.

Usage

OLS has an international presence and is one of two major international services delivering ontologies to the biomedical community, the other is BioPortal[24] developed by Stanford University in the US. BioPortal differs in allowing anyone to upload an ontology and as such contains nearly 800 ontologies whereas OLS contains a smaller curated set (just over 200) ontologies. Whilst BioPortal is comprehensive, OLS aims to provide fewer of the more widely used ontologies, for example those that conform to the OBO Foundry principles are prioritised and focuses on those that are relevant for data and collaborators within the ELIXIR and EMBL-EBI user community. OLS is an ELIXIR Recommended Interoperability Resource (RIR)[25] and is used by several ELIXIR Core Data Resources and Core Data Deposition Databases and has a crucial role in delivering semantic interoperability for FAIR data access, for example by automated ontological mapping of the BioSamples database free text annotation. For instance, it is embedded in resources providing public interfaces containing

[20] https://github.com/EBISPOT/zooma
[21] https://www.ebi.ac.uk/spot/zooma/docs
[22] https://biobankuniverse.com/
[23] https://toolbox.google.com/datasetsearch/search?query=chd7%20patients&docid=ZIpknfBQY6jmZxjYAAAAAA%3D%3D
[24] https://bioportal.bioontology.org
[25] https://www.elixir-europe.org/platforms/interoperability/rir-selection

ontology references such as Expression Atlas[26], hosts the terminologies used by resources such as Europe PubMed Central[27], and is working with other RIRs such as identifiers.org to harmonise and expose canonical URIs for the Life Science ontology domain, and FAIRsharing[28] for use in marking up content standards.

Besides OLS itself, many services at EMBL-EBI rely on the semantic toolkit developed within CORBEL. The OLS, Zooma and OxO service are part of the curation pipelines for consistent annotations of public data at the EBI with ontology terms. The result of this curation can be best presented through the OpenTargets portal. Open Targets is a collaboration between EMBL-EBI, the Sanger institute and a number of pharma companies to provide an open biomarker/drug target discovery platform. It contains over 3 million gene-disease associations, where each disease is mapped to over 9000 ontology terms from the Experimental Factor Ontology (EFO). Both the curation pipelines and the development of EFO have benefited from the tools developed within CORBEL.

OLS is also integrated into a number of third party applications and curation interfaces e.g. the COPO portal (http://www.earlham.ac.uk/copo) that assists scientists in describing their plant experimental metadata before submitting to data archives. As part of the integration of OLS into COPO, an automated pipeline for importing the Crop Ontologies (http://www.cropontology.org) into OLS was developed to support researchers in the agronomy domain. OLS now contains over 40 agronomy specific ontologies as a result of this integration.

OLS and OxO have also attracted a strong interest from industry with local installations of OLS running in Nestle and Bayer crop, who use the OLS technology as their internal terminology management system. The OxO service is also deployed at the Fraunhofer-Institute for Algorithms and Scientific Computing (SCAI), where it is being used in the Human Brain Project to provide mappings across public and private terminologies. Other collaborations include Global Alliance for Global Health[29], where OLS is used as a term resolution service by the Clinical and Phenotype Working group. It is also the preferred resolver for a number of OBO library ontologies such as EDAM[30], ENVO[31], Uberon[32] and the Monarch Disease Ontology (MONDO[33]).

OLS usage by country lists the United States, United Kingdom, Germany, India and China as the top 5 countries with requests from 164 countries in total with ~40% of our overall usage coming from European countries. OLS has on average 6,000 unique visitors per month and serves over 50 million page or API requests per month.

Next steps

- Increase the mapping coverage in OxO through integration of curated and predicted mappings generated by the Pistoia Alliance Project.

[26] https://www.ebi.ac.uk/gxa/home
[27] http://www.europepmc.org
[28] http://www.fairsharing.org/
[29] https://www.ga4gh.org/
[30] http://edamontology.org/
[31] https://www.ebi.ac.uk/ols/ontologies/envo
[32] https://www.ebi.ac.uk/ols/ontologies/uberon
[33] https://www.ebi.ac.uk/ols/ontologies/mondo

- We have, to date, been unable to support a need from the BBMRI-ERIC registry to host the ICD and SNOMED terminologies publically through OLS. This is due to the licence restrictions[34] imposed by SNOMED and the absence of a canonical OWL representation of ICD[35]. We will work to establish a mechanism for hosting additional UMLS vocabularies, such as MeSH, ICD and SNOMED through the public OLS service so that automated mapping tools can take advantage of the rich terminology (in the form of labels and synonyms) from these ontologies to cover gaps in the existing public disease ontologies. These activities are relevant to all ESFRIs handling human data, and are extensible conceptually to non human data as the mapping principles and infrastructure are domain and species neutral. We will also explore the feasibility and desirability of service restriction for users with a Snomed-CT licence.
- Formally publish the EMBRC marine ontologies with the OBO library to increase the presence and awareness of these resources. This will be subject to review by the OBO community as a community standard.
- Publish a report on the evaluation of the OLS and Zooma service for automatically mapping the UK BioBank data indicating utility for data coded with ICD10.
- Work with CORBEL WP3 to test WP6 tools on cohort harmonisation

Publications

N/A

References

N/A

Delivery and schedule

The delivery is on schedule as planned

Adjustments made

No adjustments were made

Appendices

N/A

[34] https://www.snomed.org/snomed-ct/get-snomed
[35] https://icd.who.int/