BioMedBridges

Deliverable D3.4

Project Title:	Building data bridges between biological and medical infrastructures in Europe
Project Acronym:	BioMedBridges
Grant agreement no.:	284209
	Research Infrastructures, FP7 Capacities Specific Programme; [INFRA-2011-2.3.2.] “Implementation of common solutions for a cluster of ESFRI infrastructures in the field of "Life sciences"
Deliverable title:	Harmonisation of semantic standards supporting use cases from WP7 and WP10, report
WP No.	3
Lead Beneficiary:	1: EMBL
WP Title	ESFRI BMS Standards Description and Harmonization
Contractual delivery date:	31 December 2015
Actual delivery date:	21 December 2015
WP leader:	Helen Parkinson (EBI) and Morris Swertz (UMCG)
Contributing partner(s):	1: EMBL, 10: TMF, 11: HMGU, 22: VU-VUMC, 4: STFC

Partner	Infra-structure	Contact Name	Contact emails	Status
EMBL	ELIXIR	Tony Burdett Helen Parkinson Nick Juty Simon Jupp Thomas Liener Nathalie Conte Marco Brandizi	tburdett@ebi.ac.uk parkinson@ebi.ac.uk juty@ebi.ac.uk jupp@ebi.ac.uk tliener@ebi.ac.uk nconte@ebi.ac.uk brandizi@ebi.ac.uk	author author author author author author contributor
STFC	Instruct	Chris Morris	chris.morris@stfc.ac.uk	
HMGU	Infrafrontier	Christoph Lengger Philipp Gormanns	lengger@helmholtz-muenchen.de philipp.gormanns@helmholtz-muenchen.de	
TMF	EATRIS	Murat Sariyar	murat.sariyar@tmf-ev.de	
VU-VUMC/ NKI	EATRIS	Gerrit Meijer Janneke van Denderen	g.meijer@nki.nl j.v.denderen@nki.nl	

Contents

1 Executive Summary

BioMedBridges has performed ontology standardization for three use cases for Imaging Datasets, mouse datasets, and species neutral sample datasets from the BioSamples database hosted at EBI. We have developed an ontology access standard (MIAO) and deployed tooling for mapping annotations to ontologies in support of data integration activities. The results are available in public databases such as the EBI RDF platform and improve data queries and integration for the user community.

2 Project objectives

With this deliverable, the project has reached or the deliverable has contributed to the following:

No.	Objective	Yes	No
1	Provision and use of the ESFRI BMS common molecular identifiers (eCMI)		x
2	Identification, harmonization and integration of ESFRI BMS partner standards	x	
3	Provision of standards and harmonized elements in an accessible standards registry (eSTR)		x
4	Provision and population of the ESFRI BMS Service Registry (eSR)		x

3 Detailed report on the deliverable

3.1 Deliverable overview

This deliverable provides a summary of ontology standardization activities in BioMedBridges and their subsequent application for relevant Work packages

6, 7, and 10. It describes datasets integrated, tools generated, refined and tested, and standardization activities.

3.2 (Ontology) Background

The use of ontologies and controlled vocabularies is crucial in data harmonisation and integration; they are used not only to annotate biological entities or organise these into data structures such as directed acyclic graphs (DAGs) where nodes represent entities and edges represented defined relationships between nodes. Ontologies are often described as either ‘reference’ ontologies or ‘application’ ontologies, where the former are intended to be, by design, orthogonal, and the latter are usually multi-domain ontologies, which incorporate either specific terms or entire branches of other ontologies. The typical application ontology is therefore constructed specifically in response to use cases that are cross-domain in nature, where a single reference ontology would be insufficient, for example describing genes, cells or diseases and not the connections between these. Since ontologies are often ever-evolving, to reflect improving knowledge or through re-organisation of their terms or relationships between term, it is important to maintain cross references between application and reference ontologies, and to ensure their semantic/syntactic accuracy.

The number of ontologies available in the Life Sciences is growing rapidly, with ~300 present in the National Center for BioOntology’s BioPortal (http://bioportal.bioontology.org/ontologies). The ability to find not just the most appropriate term, but even the most appropriate ontology can itself be a daunting task for non-expert users. There are two main resources for ontology access - BioPortal (http://bioportal.bioontology.org/ontologies) and Ontology Lookup Service (http://www.ebi.ac.uk/ols/beta/) at the EBI. Each receives similar numbers of hits, but BioPortal accepts all ontologies, and OLS is limited to a subset of highly used ontologies. The aim of deliverable 3.4 was to explore access, use and standardization of ontologies in the context of BioMedBridges use cases. WP7 explores the cross species integration of mouse and human phenotype ontologies and WP10 explores sample data integration across domains in the context of the BioSamples database http://www.ebi.ac.uk/biosamples. In practice use of and access to ontologies has also been relevant to other tasks in WP3, such as D3.3 which used and extended the EDAM Ontology (http://edamontology.org/page) and implementation tasks in WP4 which have delivered ontology visualisation widgets.

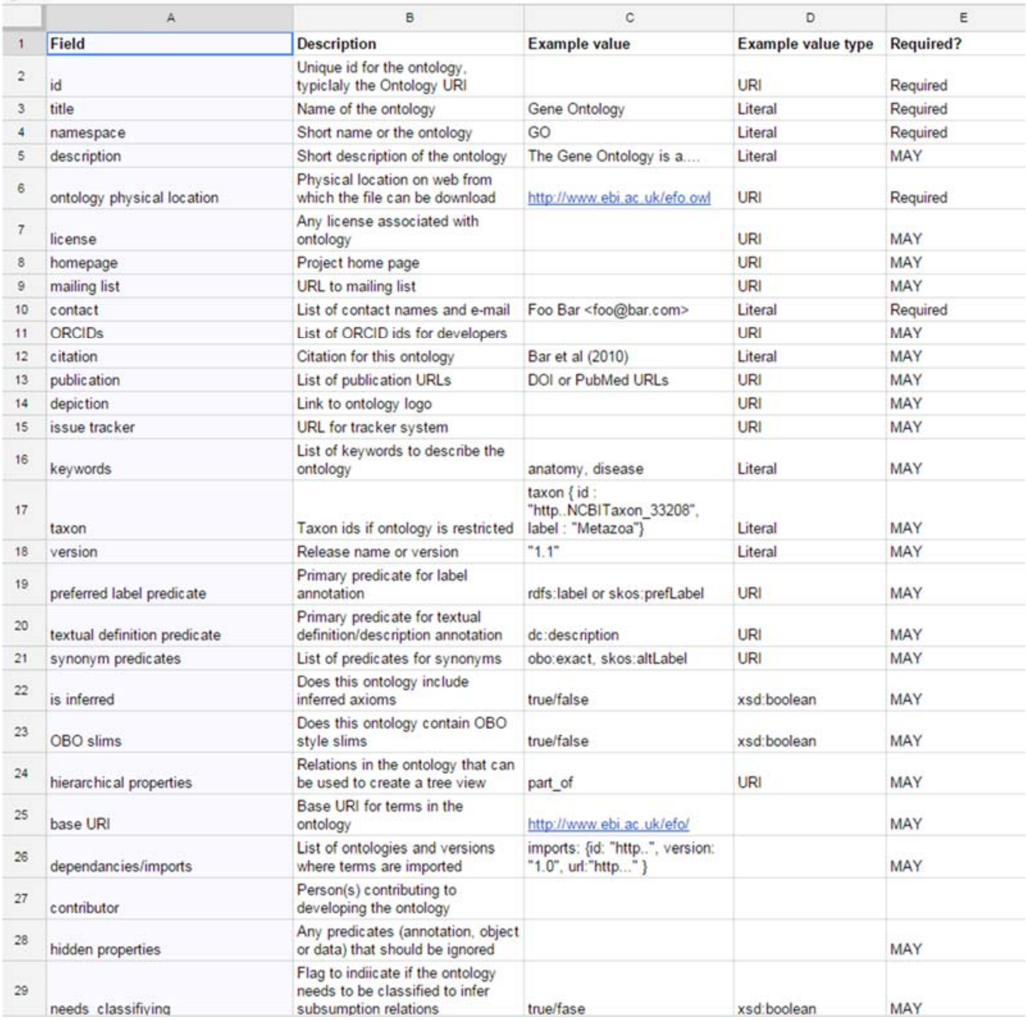
Here we describe:

The Minimum Information for Accessing Ontologies standard

The extension, evaluation and use of a tool for the data-ontology mapping, Zooma, used in standardizing data for WP6, WP7 and WP10, and a software implementation using the ontology integration.

3.2.1 The Minimum Information for Accessing Ontologies standard

Ontology documents are typically published on the Web as either OWL/RDF or OBO formatted files. Good practice dictates that every ontology should have a stable ontology URI that identifies the ontology and will resolve via the Web to the latest version of that ontology. Invariably ontology locations change, and ontology repositories such as BioPortal and OLS need to be regularly updated with details on how to resolve ontology file locations. Often, one ontology may release multiple different types or formats of files, for example, a version with full imports plus a simplified 'merged' version, or one version in OWL format and one version in OBO format. The publisher may wish to control which of these versions should be used by external applications. Other features of an ontology, such as a recommended reasoner or synonym predicate, are all useful information for applications that consume ontologies. In an attempt to standardise how ontology metadata is published, BioMedBridges, along with international collaborators, have been working towards adopting a common ontology metadata standard via the OBO foundry (See Figure 1).



	A	B	C	D	E
1	**Field**	**Description**	**Example value**	**Example value type**	**Required?**
2	id	Unique id for the ontology, typiclaly the Ontology URI		URI	Required
3	title	Name of the ontology	Gene Ontology	Literal	Required
4	namespace	Short name or the ontology	GO	Literal	Required
5	description	Short description of the ontology	The Gene Ontology is a....	Literal	MAY
6	ontology physical location	Physical location on web from which the file can be download	http://www.ebi.ac.uk/efo.owl	URI	Required
7	license	Any license associated with ontology		URI	MAY
8	homepage	Project home page		URI	MAY
9	mailing list	URL to mailing list		URI	MAY
10	contact	List of contact names and e-mail	Foo Bar <foo@bar.com>	Literal	Required
11	ORCIDs	List of ORCID ids for developers		URI	MAY
12	citation	Citation for this ontology	Bar et al (2010)	Literal	MAY
13	publication	List of publication URLs	DOI or PubMed URLs	URI	MAY
14	depiction	Link to ontology logo		URI	MAY
15	issue tracker	URL for tracker system		URI	MAY
16	keywords	List of keywords to describe the ontology	anatomy, disease	Literal	MAY
17	taxon	Taxon ids if ontology is restricted	taxon { id : "http..NCBITaxon_33208", label : "Metazoa"}	Literal	MAY
18	version	Release name or version	"1.1"	Literal	MAY
19	preferred label predicate	Primary predicate for label annotation	rdfs:label or skos:prefLabel	URI	MAY
20	textual definition predicate	Primary predicate for textual definition/description annotation	dc:description	URI	MAY
21	synonym predicates	List of predicates for synonyms	obo:exact, skos:altLabel	URI	MAY
22	is inferred	Does this ontology include inferred axioms	true/false	xsd:boolean	MAY
23	OBO slims	Does this ontology contain OBO style slims	true/false	xsd:boolean	MAY
24	hierarchical properties	Relations in the ontology that can be used to create a tree view	part_of	URI	MAY
25	base URI	Base URI for terms in the ontology	http://www.ebi.ac.uk/efo/		MAY
26	dependancies/imports	List of ontologies and versions where terms are imported	imports: {id: "http..", version: "1.0", url:"http..." }		MAY
27	contributor	Person(s) contributing to developing the ontology			
28	hidden properties	Any predicates (annotation, object or data) that should be ignored			MAY
29	needs_classifiying	Flag to indiicate if the ontology needs to be classified to infer subsumption relations	true/fase	xsd:boolean	MAY

Figure 1 The Draft MIAO Specification

The OBO foundry provide a central registry for many of the public biomedical domain ontologies and have recently moved to a YAML-LD based file format for publishing ontology metadata. The YAML schema defines various pieces of standard ontology metadata such as the common short name, title, description, stable ontology URL, along with other useful information such as, who the author of the ontology is and who to contact with questions or where to submit new term requests. Ontology repositories, such as OLS (www.ebi.ac.uk/ols/beta), are now adopting this YAML-LD schema and services such as OLS have been extending the schema to support additional metadata relating to how the ontologies should be visualised in applications. The OLS extensions include fields for preferred synonym and definition predicates, preferred reasoner for classification, and flags for object properties that are considered hierarchical (such as part of) for ontology visualisation. The OLS is now able to consume ontology configuration in the OBO YAML format, and has also been extended to identify certain ontology metadata information that is asserted directly in the ontology as OWL ontology annotations. The ontology standardization work will continue to be supported by OLS in future grant funded activities.

3.2.2 Zooma: a data-ontology mapping tool

The use of ontologies to annotate data is an established method for adding semantics to metadata, facilitating integration and richer querying. By creating a repository of annotations and their mapped ontology terms, and scoring their quality (curated, predicted etc), a "smart" annotation and search service was created: Zooma[1] is a linked data repository of annotation knowledge, incorporating information from a variety of biological databases, providing an integrated resource that allows annotation searches and facilitates curation activities. It can be accessed through a REST-like API, a user interface and an endpoint for SPARQL querying. It provides a service that allows querying by plain text and returns possible annotations between matching properties and concepts identified by a URI. Zooma has been applied to multiple datasets during BioMedBridges and the content, design and ontology availability have all been extended to support the use cases described below. Zooma architecture and context is described in Figures 2 and 3.

The Zooma model is built using the Open Annotation Model proposed by the Open Annotation Community Group (http://www.w3.org/community/openannotation) and stores:

- The source of this annotation (e.g. the database)
- The "creator" of this annotation (a person or a script)
- The date the annotation was created
- An evidence code describing how this annotation was made

This information can be used in optimising a scoring algorithms for use during annotation searching, and data mining.

[1] http://www.ebi.ac.uk/fgpt/zooma

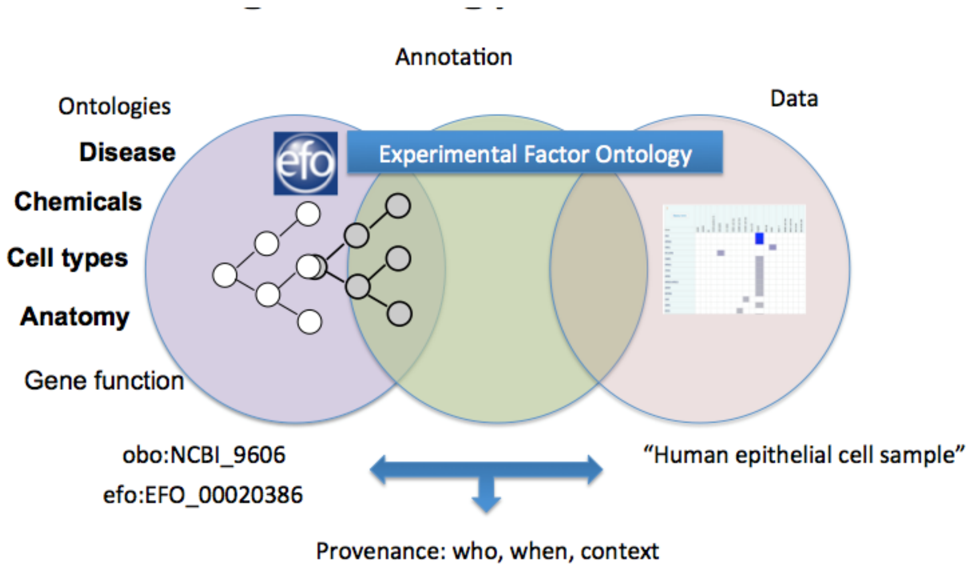
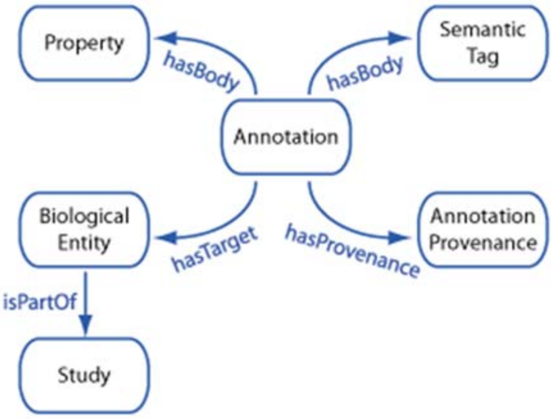


Figure 2 Interaction between the Ontology, Zooma and annotated data

The Zooma repository of annotations is currently around 17 million triples, and two of the most highly annotated datasets in the EBI RDF[2] platform, namely the Gene Expression Atlas and BioSamples database (WP10), combined contain over 700 million triples.

Zooma uses an RDF schema as its data model, and is composed of lightweight concepts such as Study, Biological Entity, Property, Annotation, Semantic Tag and Evidence.



A simple schematic representation of the ZOOMA annotation-centric data model

[2] http://www.ebi.ac.uk/rdf

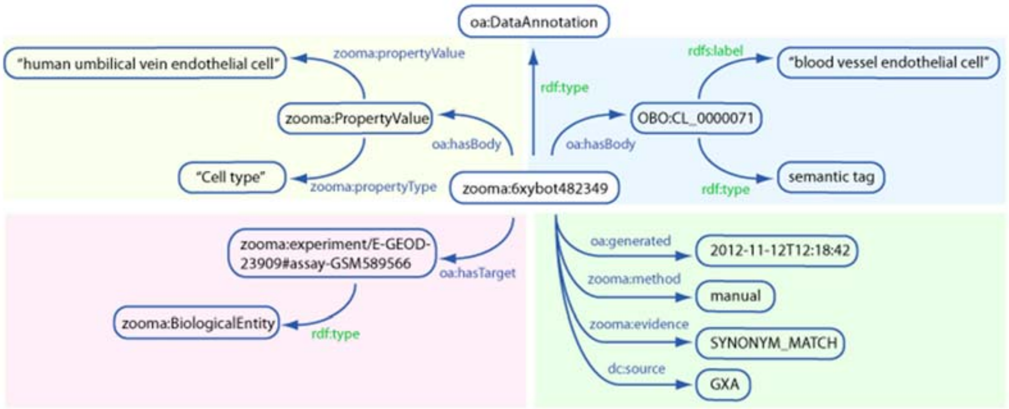


Figure 3 An example implementation of the ZOOMA data model for a real annotation

Zooma provides an easy-to-use toolkit for the annotation of data to ontologies, enabling a simple process for enrichment of datasets with no links to existing semantic standards enabling semantic standardization of free text data. Additionally, as the core component of Zooma is a linked data repository of annotation knowledge, annotation output can be shared between multiple use cases to ensure consistency of annotation and the sharing of semantic curation effort. In this deliverable we report the on the application of this principle to several BioMedBridges work packages where it was used by curators and bioinformaticians for data-ontology mapping activities via its user interface and programmatically.

3.3 Ontology Standardization Use cases

3.3.1 WP6 - Interoperability of large scale image datasets from different biological scales

The Cellular Microscopy Phenotype Ontology (CMPO) was developed as part of WP6 to support integration of imaging data annotated with cellular microscopic phenotypes (this represents an extension of the original plan to support WP7 and 10. This was pursued as WP6 had data and use cases early in the project and provided complementary data to WP7 and 10, resulting in more integration with EuroBioImaging. To support the annotation of images to this new standard, we made the CMPO ontology available to Zooma, CMP) was developed in WP6 to describe phenotypes detected using cellular imaging technologies e.g. delayed mitosis. New datasets annotations between previously unseen images and the CMPO ontology could therefore be generated by relying on lexical matches between the free-text descriptions of the phenotypes in the image and the labels of terms in the ontology. However, this provides sub-optimal coverage of many phenotypes. As such, 10 canonical datasets were curated to determine new mappings to CMPO which could not be predicted by lexical matching alone. To accomplish this, the curator made use of mappings predicted by Zooma, performed new searches for possible synonymous terms using the curators background scientific knowledge, and browsed the CMPO ontology using tools such as BioPortal and OLS. This time-consuming process resulted in the creation of 132 new data-ontology annotations that were formatted and stored in the Zooma annotation knowledgebase. These annotations provide the spine for semantic alignment of future cellular phenotype images and will continue to be used in future projects such as CORBEL and continue to be used by image data curators.

We loaded the CMPO ontology into Zooma to assess its coverage. Zooma reports three types of mapping. An ‘automated annotation’ is where the phenotype can be automatically annotated to a CMPO term with very high confidence. Typically an automatic annotation is only possible when Zooma has seen a manually verified example before. The second category is ‘requires curation’, which reflects the scenario where multiple potential annotations scored equally high and Zooma is unable to make an automatic annotation. The final category is where Zooma has no suggested annotation. Querying Zooma with the original 201 free text phenotype descriptions, we find 116 matches to the ontology, but all require curation. That is, Zooma has no evidence other than a label match to validate the annotation.

In order to demonstrate the utility of Zooma 132 manually verified CMPO annotations were loaded into Zooma. Table 3 shows the results of querying Zooma with the original 201 free text phenotypes using CMPO alone, or using a combination of CMPO and the manual annotations. The manually verified annotations provide Zooma with evidence for certain mappings so that it can predict an automated annotation with higher confidence.

Table 1 Zooma results for mapping CAMPO datasets

Tool	Automated annotation	Requires curation	No suggested annotation	Total % annotated
Zooma with CMPO ontology only	0	116	85	58%
Zooma with CMPO ontology and manually curated annotations	76	67	58	72%

3.3.2 WP7 - Phenobridge - crossing the species bridge between mouse and human

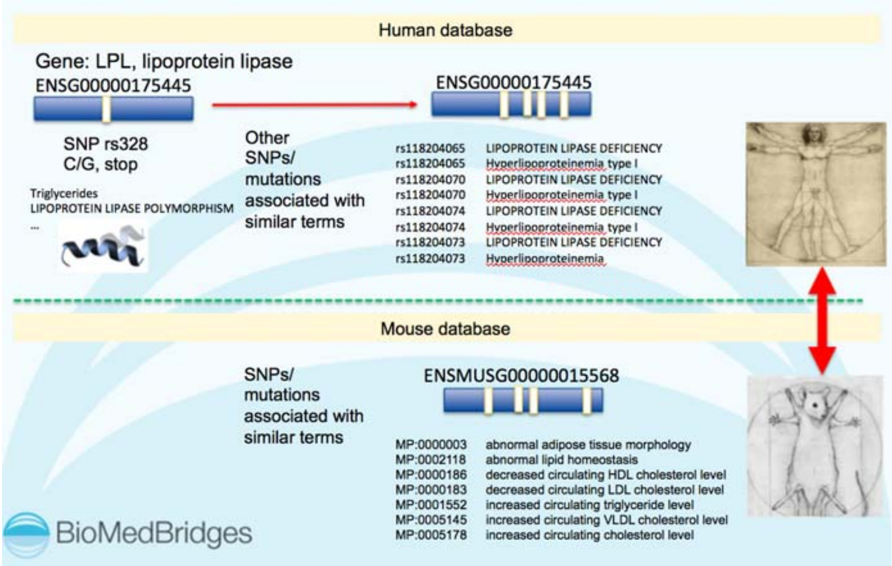
Phenobridge (WP7) aims to deliver a semantic and most recently a genomic bridge between human and mouse datasets. This involves mapping human and mouse ontologies together, designing ontology interoperability strategies and acquiring and mapping available datasets from partners to explore data annotations required to perform analyses. Further we have integrated ontological standardization with genomic analyses to provide a service which leverages the ontological standardization work and deploys a user facing component integrating human GWAS data and mouse phenotype data for the first time (see Section 3.3.2.1 below).

3.3.2.1 Bridging Mouse and Human Genomes

The fundamental genetic similarity between mice and humans allows researchers to infer a human gene's function based on studies with laboratory mice. Model organisms are providing fundamental clues on gene function has proved to be a powerful approach to gain knowledge on both human disease and mammalian biology. Large-scale pan genomic community projects using mice like the International Mouse Phenotyping Consortium (IMPC, http://www.mousephenotype.org/) aim to generate and phenotypically characterize knockout mutant strains for every protein-coding gene in the mouse. Other resources like the Mouse Genome Informatics (MGI, http://www.informatics.jax.org) collects and integrates literature curated information about genotypes and phenotypes in the laboratory mouse. In collaboration with WP7 we have developed a tool to map genetic information between Mouse and Human and which leverages the data-ontology mapping tools described in 3.3.2. above. These resources provide access to mouse and human syntenic mapping, variation and phenotypes and also mouse resources providing integrated genetic and functional/phenotypic data. Resources and data types integrated:

- Ensembl (http://www.ensembl.org/): Genome databases for vertebrates and other eukaryotic species.
- GWAS catalogue (http://www.ebi.ac.uk/gwas/): Curated, literature-derived collection of all published genome-wide association studies.
- MGI and IMPC: Mouse genotype/phenotype databases.
- Human/Mouse variation
- Human/Mouse variation associated overlapping gene or nearest gene
- Human/Mouse gene associated annotation (genomic coordinates, gene ID, gene symbol...)
- Human/Mouse phenotypic annotations

This systematic mapping between syntenic regions allows discovery of functional conservation and can help with the validation/prioritization of candidate disease genes or regions. A schematic example is shown in Figure 4.



Human database

Gene: LPL, lipoprotein lipase
ENSG00000175445

ENSG00000175445

SNP rs328
C/G, stop

Triglycerides
LIPOPROTEIN LIPASE POLYMORPHISM
...

Other SNPs/ mutations associated with similar terms

rs118204065	LIPOPROTEIN LIPASE DEFICIENCY
rs118204065	Hyperlipoproteinemia type I
rs118204070	LIPOPROTEIN LIPASE DEFICIENCY
rs118204070	Hyperlipoproteinemia type I
rs118204074	LIPOPROTEIN LIPASE DEFICIENCY
rs118204074	Hyperlipoproteinemia type I
rs118204073	LIPOPROTEIN LIPASE DEFICIENCY
rs118204073	Hyperlipoproteinemia

Mouse database

SNPs/ mutations associated with similar terms

ENSMUSG00000015568

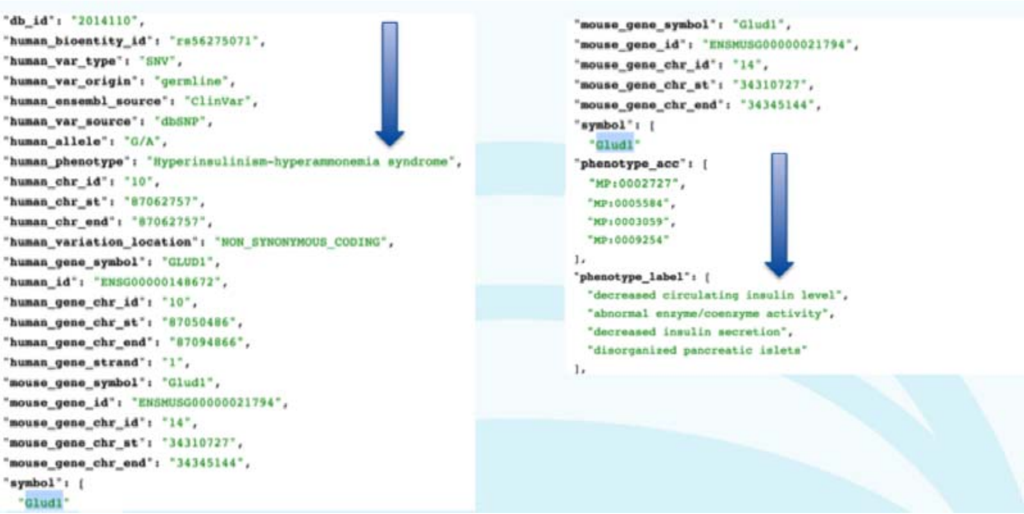
MP:0000003	abnormal adipose tissue morphology
MP:0002118	abnormal lipid homeostasis
MP:0000186	decreased circulating HDL cholesterol level
MP:0000183	decreased circulating LDL cholesterol level
MP:0001552	increased circulating triglyceride level
MP:0005145	increased circulating VLDL cholesterol level
MP:0005178	increased circulating cholesterol level

BioMedBridges

Figure 4 Mapping Human and Mouse genes and phenotypes. Here we are querying LPL, lipoprotein lipase gene, this gene has several distinct variants causing triglyceride metabolism phenotypes in humans. The automatic mapping of this gene in the Mouse shows that the function is conserved making it a potential suitable model for further investigation. The annotations are standardized using Zooma and manually refined.

GenoBridge was developed using Perl and Mysql scripts deposited in SVN integrating data from several main databases (ensemble http://www.ensembl.org/, MGI http://www.informatics.jax.org/, GWAS http://www.genome.gov/gwastudies/) in a data warehouse (Mysql). Apache Solr (http://lucene.apache.org/solr/) was used for database integration, faceted search, dynamic clustering and indexing. We generated approximately 3.8 million documents in SolR allowing dynamic search of mouse and human genotype/phenotype data (see Figure 5).

This screenshot is an example of a search of documents containing the word "insulin" in phenotype fields in human and mouse (shown in the blue arrows). The search allow the discovery of a document containing information on Glud1 gene. This document contains information of genomic location, variation, phenotype in both species (all fields in black) and show that this gene has variants involved in Human congenital hyperinsulinism and Mouse models display phenotypes linked to insulin metabolism. Database versions: A data freeze was done in May 2015 using latest version of MGI, GWAS and Ensembl 79_38 at this period.



```
"db_id": "2014110",
"human_bioentity_id": "rs56275071",
"human_var_type": "SNV",
"human_var_origin": "germline",
"human_ensembl_source": "ClinVar",
"human_var_source": "dbSNP",
"human_allele": "G/A",
"human_phenotype": "Hyperinsulinism-hyperammonemia syndrome",
"human_chr_id": "10",
"human_chr_st": "87062757",
"human_chr_end": "87062757",
"human_variation_location": "NON_SYNONYMOUS_CODING",
"human_gene_symbol": "GLUD1",
"human_id": "ENSG00000148672",
"human_gene_chr_id": "10",
"human_gene_chr_st": "87050486",
"human_gene_chr_end": "87094866",
"human_gene_strand": "1",
"mouse_gene_symbol": "Glud1",
"mouse_gene_id": "ENSMUSG00000021794",
"mouse_gene_chr_id": "14",
"mouse_gene_chr_st": "34310727",
"mouse_gene_chr_end": "34345144",
"symbol": [
  "Glud1"
```

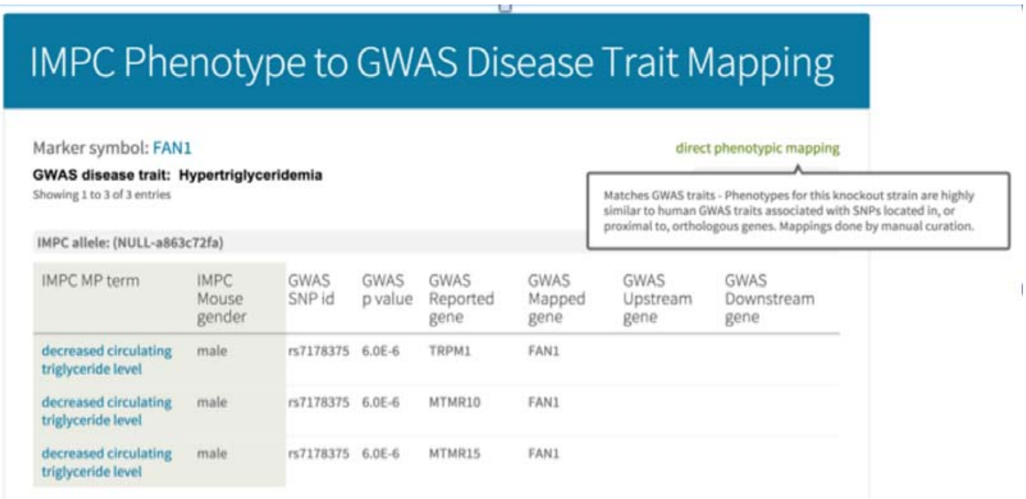
```
"mouse_gene_symbol": "Glud1",
"mouse_gene_id": "ENSMUSG00000021794",
"mouse_gene_chr_id": "14",
"mouse_gene_chr_st": "34310727",
"mouse_gene_chr_end": "34345144",
"symbol": [
  "Glud1"
"phenotype_acc": [
  "MP:0002727",
  "MP:0005584",
  "MP:0003059",
  "MP:0009254"
],
"phenotype_label": [
  "decreased circulating insulin level",
  "abnormal enzyme/coenzyme activity",
  "decreased insulin secretion",
  "disorganized pancreatic islets"
],
```

Figure 5 Example of SolR search result

This dataset contains variation coming from a wide range of sources like GWAS catalogue, ClinVar (Variants of clinical significance imported from ClinVar) or Orphanet (The portal for rare diseases and drugs) and many more. A complete list of sources is available with the link below (http://www.ensembl.org/info/genome/variation/sources_documentation.html). The phenotypic data associated with variants associated with protein coding genes was standardized using Zooma (described in 3.1.2.2.)

We chose to integrate a GenoBridge sub dataset corresponding to variation data from GWAS catalogue in IMPC portal. This choice was made as GWAS is not represented in IMPC portal. GWAS studies identify loci and phenotypes used as diagnostic and prognostic indicators, integration of this type of data in IMPC can help to validate those indicators in Human if the phenotype is conserved in both species. We used manual curation to select matching orthologous genes and phenotypes for integration in IMPC website after standardization and experience gained using Zooma on human variation phenotypes. An example query[3] showing links to GWAS data is shown in Figure 6.

[3] http://dev.mousephenotype.org/data/genes/MGI:105083



IMPC Phenotype to GWAS Disease Trait Mapping

Marker symbol: FAN1

direct phenotypic mapping

GWAS disease trait: Hypertriglyceridemia

Showing 1 to 3 of 3 entries

Matches GWAS traits - Phenotypes for this knockout strain are highly similar to human GWAS traits associated with SNPs located in, or proximal to, orthologous genes. Mappings done by manual curation.

IMPC allele: (NULL-a863c72fa)

IMPC MP term	IMPC Mouse gender	GWAS SNP id	GWAS p value	GWAS Reported gene	GWAS Mapped gene	GWAS Upstream gene	GWAS Downstream gene
decreased circulating triglyceride level	male	rs7178375	6.0E-6	TRPM1	FAN1		
decreased circulating triglyceride level	male	rs7178375	6.0E-6	MTMR10	FAN1		
decreased circulating triglyceride level	male	rs7178375	6.0E-6	MTMR15	FAN1		

Figure 6 GWAS-IMPC integration example showing an example of integration of human GWAS data within IMPC portal. Mice deficient in Fan1 gene display abnormal triglyceride levels (grey shaded), interestingly integration with GWAS data show that the orthologous gene in Human has a variant (rs7178375) believed to cause hypertriglyceridemia. The phenotype in both species are mapping directly making Fan1 a strong candidate for triglyceride metabolism.

3.3.2.2 Ontological standardization of human phenotypes using Zooma

Section 3.2.2 describes the Zooma architecture. We tested the Zooma annotation on the human variation data from Ensembl described above. Our aim was to standardize the free text annotation of variation contained within Ensembl. For example, these three terms (DIABETES MELLITUS, TYPE II; DIABETES MELLITUS, NON INSULIN-DEPENDENT; and DIABETES MELLITUS, INSULIN-RESISTANT) are phenotypic annotations associated with human variations found in Ensembl. Each term is referring to the same disease and therefore should be mapped with a single ontology term identifier: http://purl.obolibrary.org/obo/HP_0005978 (which is labelled “Type II diabetes mellitus”).

Using a manually curated process, we mapped free text annotation of data to phenotype ontologies, tested and extended the Zooma application and provided a mapped and annotated dataset for future use in Zooma. The user interface was improved, and extended with BioMedBridges specific datasets. Table 2 provides a summary of 4089 free text annotations corresponding to protein coding gene variation associated phenotypes extracted from Ensembl, and mapped to ontologies before and after BioMedBridges specific extensions to Zooma. In successive tests of the system we improved the mapping capability for this datasets, verified mappings manually and then stored these in the Zooma knowledge base for future use. The standardization of this data allowed us to integrate the phenotypic annotations across the various aggregated datasets in Ensembl, and to extract the subset of these integrated with IMPC data above (Section 3.3.2.1) providing a new cross species data resource in combination with genomic data.

Table 2 Zooma results in mapping human phenotypes associated with variation data before and after improvements in selection of ontologies and addition of curated datasets

	Count (before)	Percentage (before)	Count (after)	Percentage (after)
Variation Phenotype Terms	4089	100%	4089	100%
Zooma “automatic” mappings	194	5%	1150	28%
Zooma suggestions	1223	30%	1225	30%
Unmapped	2672	65%	1714	42%

3.3.3 WP10 - Integrating disease related data and terminology from samples of different types

WP10 demonstrates the feasibility and provides a prototype for linking disease to molecular information of two levels - terminology and data. One key objective was to link data in selected BBMRI biobanks to samples in the EBI Biosamples Database.

The Biosamples Database contains nearly four and a half million samples from a variety of different sources, including direct submissions, exchange with biobanks, and samples brokered from assay databases such as the European Nucleotide Archive. Biosamples supports rich semantic metadata descriptions of the samples it contains, although the quality of this metadata can be variable depending on the source.

To support the harmonisation of sample metadata generated within BioMedBridges with samples inside the Biosamples database, we identified a core set of 869,917 submitter provided samples that include least one well-described, well formatted annotation to a resolvable ontology term, typically for disease, tissue, phenotype or experimental treatment such as drug or other compound. In total, there were 1.12 million annotations to 1,715 terms, indicating that are large number of these annotations represent the same concepts and are therefore already well aligned in Biosamples. This set of 1.12 million semantic annotations have been loaded into Zooma to ensure that these annotations are available to users wishing to annotate sample data.

We have also developed a pipeline using Zooma that enables automatic annotation of the remaining samples in the Biosamples database (i.e. all those not included in the 1.12 million above) to ontologies. These samples had metadata that were previously annotated only as free text and could not be queried using an ontology, and therefore could not be integrated semantically. This requires processing of 3.3 million samples, comprising in total nearly 20 million annotations, against the Zooma API. The resulting ontology-annotated BioSamples are included in the EBI's RDF platform and at each release the Zooma pipeline is used to annotate exported data. As of November 2015, the Biosamples RDF dataset contained 361,635,520 triples. Developing the software to create and update these Biosamples feature annotations in an efficient way has been particularly challenging and we have been reporting about the experience gathered on that in dissemination activities[4]. Both such experience and the software we produced[5] can be useful in dealing with similar tasks. We have prototyped a software tool[6] to update Zooma annotations in an incremental way and directly at the Oracle database level (the primary storage backend for the Biosamples database), rather than re-generating them from scratch upon RDF export.

3.3.4 Summary of Use Cases

We have shown that we have utilised Zooma in three separate BioMedBridges use cases (extending our original remit), and that in each case we have made use of Zooma to annotate data with ontologies and pushed the results of curation and review of this annotation process back into the Zooma

[4] https://prezi.com/vxox0pgra6d7/biosd-linked-data-lessons-learned/
[5] https://github.com/EBIBioSamples/biosd2rdf
[6] https://github.com/EBIBioSamples/biosd_feature_annotator

knowledgebase. In doing so, each of the three work packages has taken advantage of curation performed by the others, and this has helped to ensure compatibility and harmonisation across each work package. The semantic integration coupled with the genomic bridging in collaboration with WP7 has allowed us to deploy a new component for the IMPC portal, and expose data in new contexts.

As data across three separate work packages has been annotated to the core set of ontologies contained within Zooma, and has been annotated using the same tooling and methodology, we can be confident that these datasets share discoverability criteria using ontology-enabled search, and also are readily integratable along the semantic metadata axis, we have demonstrated the utility in the mousephenotype.org portal, and continue to use Zooma in support of curatorial activities.

3.4 Sustainability

Zooma is a component that was extended by BioMedBridges to three new use cases, warranting the addition of multiple new datasets and ontologies. The Zooma tool was developed by the Samples, Phenotypes and Ontologies Team at EBI, who have an ongoing commitment to its maintenance, supported by several other grants. Curation data that was specifically generated for BioMedBridges will be stored for the foreseeable future in the Zooma knowledgebase, and is clearly date stamped.

3.5 Future Work

Zooma continues to be used at the EBI, and a service to the wider community. As part of work on CORBEL we will integrate Zooma with the Ontology Lookup Service, continue to update new datasets and improve the user interface for curators. The MIAO standard will be refined, and will be made available from appropriate standards registries. The deployment for GWAS data integration with IMPC is currently hosted on a development server, after completing user acceptance testing it will be deployed to a production server.

4 Delivery and schedule

The delivery is delayed: ☐ Yes ☑ No

5 Adjustments made

No adjustments were made to the deliverable.

6 Background information

This deliverable relates to WP 3; background information on this WP as originally indicated in the description of work (DoW) is included below.

WP 3 Title: ESFRI BMS Standards Description and Harmonization

Lead: Helen Parkinson (EMBL-EBI, Morris Swertz (UMCG)

Participants: EMBL, KI, STFC, UDUS, TUM-MED, ErasmusMC, TMF, HMGU, VU-VUMC, UCPH, UH, UMCG, CIRMMP

Standardization is necessary to ensure infrastructures can work together (syntactic interoperability: data models, data formats, API's, services descriptions, registration and discovery of services), understand each other data (semantic interoperability: ontologies, vocabularies, coding systems, common identifiers), have analysis and supporting tools that complement each other and can be combined in a pipeline (process interoperability) and allow multiple data sets from different origins (including public resources) to be analysed together.

This work package (WP) requires close collaboration with domain experts, research infrastructures, WP4 which will provide implementation based on standardization deliverables described here, and WP5 which will address security issues and use case work packages 6-10. In order to work efficiently a nominated individual from each ESFRI BMS expert area will be responsible both for tasks in this WP, registration of standards, representation of, and correspondence with, relevant domain specific external standardization parties and to represent their community requirements in this WP. WP3 partners are also represented in the use case work packages and will ensure their requirements are supported here.

This WP involves the majority of partners, and exchange of information, registry of services and meta mapping activities will require a diverse set of personnel. The design of this WP therefore includes an allowance for exchange of personnel between this WP and others to facilitate the implementation of deliverables in other WPs and to support interaction with external experts at meetings and workshops where necessary. This will ensure that relevant experts have the opportunity to actively solve problems by working closely with individuals from work packages to which they have not been assigned. We have also allowed developer time for the creation of training materials and delivery of training at BioMedBridges workshops, as described in WP12.

Work package number	WP3	**Start date or starting event:**	month 1
Work package title	ESFRI BMS Standards Description and Harmonization		
Activity Type	RTD		

Participant	1: EMBL	3: KI	4: STFC	5: UDUS	7: TUM-MED	9: ErasmusMC	10: TMF	11: HMGU	22: VU-VUMC	15: UCPH	16: UH	19: UMCG	20: CIRMMP
Person months	42	21	6	28	4	5	16	30	16	8	11	32	14

Objectives

Addition of scientific value and support for the integration of data between the ESFRI BMS domains by catalogue, review, modification, harmonization, registration and implementation of existing identifier, content, syntactic and semantic standards across the ESFRI BMS projects to support data exchange, integration and infrastructure development.

1. Provision and use of the ESFRI BMS common molecular identifiers (eCMI)
2. Identification, harmonization and integration of ESFRI BMS partner standards
3. Provision of standards and harmonized elements in an accessible standards registry (eSTR)

4. Provision and population of the ESFRI BMS Service Registry (eSR)

Description of work and role of participants

The standardization task is large as ESFRI BMS projects have been active in this area evaluating intra-domain standards, bottlenecks and solutions and there are numerous external standards efforts corresponding to content, data format, semantic and identifier standardization in this domain in which many project partners are involved. Examples include the gene ontology (GO) as an example of a semantic standard, DICOM as an imaging format standard, MIMPP as a content standard from EUROPHENOME, the LCF/MTZ file format, and the CCPN data model for macromolecular NMR. WP will address the following tasks to provide focus:

1. Common identifiers (Task Lead ELIXIR)

The provision and use of common identifiers to determine unambiguous molecular identity for bio-molecules such as genes, proteins and bioactive compounds is key to supporting the information flow from basic science, model organism biology, bioinformatics and structural biology through to translational research and clinical care. The ESFRI BMS project partners will work together to determine a ‘Molecular Dictionary’ of identifier types and their attributes for use in this project which will constitute best practice for cross domain integration. Where no authoritative identifier standard exists, we will work with the respective community to determine one to support the activities of WP4 and use cases. Relevant identifiers include those for samples (Task 2), small molecules, macromolecular assemblies, genes, drugs and proteins especially where these relate to clinical scenarios.

2. Sample meta data standards (Task Leads BBMRI)

The ability to identify samples and describe their attributes, so data relating to them can be integrated and analysed is common to all ESFRI BMS domains. Content standards which determine exist for given experimental scenarios which data should be collected e.g. age, sex, phenotype, disease state, sampling time, processing state, etc. These are typically determined based on requirements for analysis, data sharing needs and regulations within a research or technology based domain. For example, the MIAME standard determines which information should be stored about a gene expression experiment performed on a microarray. This is not necessarily consistent with core information about the same sample stored in a BioBank which may include sample processing state, disease and tissue, a sample used to determine a protein structure, or a live animal sampled from the ocean. Where processing states, provenance, storage conditions, or other experimental context are important for a domain e.g. INSTRUCT or for re-use of data relating to samples across domains, these will also be explored with respect to the use cases. The clinical data community have specific requirements relating to integration of Electronic Health Records (EHR), use of clinical terminologies such as SNOMED-CT, description of medical imaging procedures and provision of molecular data in clinical context with appropriate quality control data and translation across these domains is relevant to this task, Task 4 and WP10. Standards in use within the ESFRI BMS projects for data content and semantics will be documented in a public interactive matrix consisting of project, standard and individual elements of standards. Comparable elements across standards will be identified by a harmonization and mapping process across partners. For example BBMRI has produced a lexicon which defines important concepts for the bio-banking domain and EATRIS has analysed standards relating to inter and intra operability between organisations. Standards in use by partners relating to samples will be meta-mapped; common elements e.g. from BBMRI will be cross referenced to relevant concepts from ELIXIR, ECRIN and EATRIS. Where standards are in development e.g. from 2008 roadmap ESFRI BMS projects these will be added and harmonized once they are determined to be stable and valid within a domain, e.g. imaging standards are under development by EuroBioImaging. We do not expect all standards to be fully interoperable and the process of meta-mapping and presentation of these data in an interactive and updated form will inform partners and focus use cases. We will pay specific attention to widely adopted standards, and supporting integration rather than development of standards de novo.

3. Service registration and annotation (Task Lead ELIXIR)

The description of where data and services exist, and by what mechanism these are accessible is key to integrating and exchanging data and has been identified by ELIXIR, EATRIS and others as a blocker to integration especially across domains. Therefore we will develop the Meta-Services Registry comprising tools and terminology for annotation of services (eSR) to catalogue services across partners, domains allowing partners to self register their own and others services. This will build on previous work in the Bioinformatics domain (EMBRACE, BioCatalogue) and will be extended this with the 2008 roadmap ESFRI BMS partners and throughout the grant as services appear and are used. This will promote the use of domain specific services across partners and also internationally.

4. Semantic standards – ontologies and annotation (Task Lead ELIXIR)

Content standards define what data about a sample in a context or domain. However the meaning of data can be made explicit only by the use of defined terminologies. The use, standardization and mapping of terminologies across domain and species will be explored in the context of use case Work Packages 7 and 10. WP7 explores the semantic integration between mouse models of disease, phenotype and WP10 explores integration of sample data of different types. In order to make these tasks feasible prioritized dataset(s) will be identified with WP7/10 by means of integration criteria which will be developed jointly with these work packages. For example – availability of data in the public domain and /or focus on a key disease type which is well represented in the terminologies to be integrated and available datasets.