



## Deliverable D4.6

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:	Pilot integration of Web Services based simple object queries	
WP No.	4	
Lead Beneficiary:	1: EMBL	
WP Title	Technical integration	
Contractual delivery date:	30 June 2015	
Actual delivery date:	30 June 2015	
WP leader:	Ewan Birney	1: EMBL
Partner(s) contributing to this deliverable:	1: EMBL, 5: UDUS, 9: ErasmusMC	

*Julie McMurry, Simon Jupp, Tony Burdett, Andy Jenkinson, Helen Parkinson, Chris Morris, Martyn Winn, Philipp Gormanns, Elida Schneltzer, Raffael Bild, Christian Krauth, Freek de Bruijn, Ward Blondé, Jeroen Belien, Stefan Klein, Erwin Vast*



## Contents

1	Executive Summary .....	4
2	Project objectives .....	5
3	Detailed report on the deliverable .....	6
3.1	Background .....	6
3.2	Individual pilot projects.....	6
3.2.1	Overview of individual pilot projects .....	6
3.3	Overall conclusion D4.6 .....	7
3.4	Overall future work .....	7
3.4.1	Overall future work for D4.8 .....	7
3.4.2	Beyond BioMedBridges.....	8
4	Delivery and schedule .....	9
5	Adjustments made.....	9
6	Background information .....	10
Appendix A	RDF pilots.....	13
Appendix A.1	HMGU (INFRAFRONTIER): Integrating systemic mouse phenotype data from diverse sources .....	13
Appendix A.2	TUM-MED (BBMRI): Integrating human tissue biobanking data across Europe .....	16
Appendix A.3	EMBL-EBI (ELIXIR): additions to RDF platform.....	18
Appendix A.4	UDUS (ECRIN): Searching for clinical trials information and linking clinical trials to biosamples, drugs, genes, and publications .....	22
Appendix A.5:	RDF Training materials .....	27
Appendix B:	TranSMART pilots.....	28
Appendix B.1	VUMC (EATRIS) Integrating Galaxy workflows into tranSMART .....	28
Appendix B.2	VUMC (EATRIS): Integrating CDISC Operational Data Model into tranSMART i2b2 .....	31
Appendix B.3	ErasmusMC (Euro-BioImaging): Centralized correlative analysis between image-derived data and other clinical data. ....	34
Appendix C:	BioJS Widget pilots.....	40
Appendix C.1	STFC (INSTRUCT): Clinical Consequences of Protein Structure variation CCoPS .....	40
Appendix C.2	EMBL-EBI (ELIXIR) PLATO widget .....	43
Appendix C.2.1	Minimum Information for Accessing an Ontology (MIAO).....	1



## Figures

Figure 1 Axes of technical integration .....	7
Figure 2 Data touching points across the D4.6 pilot projects.....	7
Figure 3 Partner plans for sustainability of individual D4.6 pilot projects beyond BioMedBridges .....	8
Figure 4 Ontology use by Dataset.....	10
Figure 5 Overview of TransSMART Integration Pilots .....	4
Figure 6 Summary of BioJS widget-driven pilots .....	6
Figure 7 BioMedBridges software matrix shows pilots (rows) with links to the REST web service endpoints, visual web interface (GUI), RDF, source code, and BioJS documentation.....	8
Figure 8 Model for the semantic integration of mouse phenotype resources .....	14
Figure 9 RDF-ization of a UniProt protein name (CSF-1) mined from the full-text article PMC15040 with the OADM TextQuoteSelector.....	20
Figure 10 Displayed is the simple user interface of CTIM with a single search field. ...	23
Figure 11 Screenshot of a transSMART project with image-derived data imported from XNAT .....	36
Figure 12 schematic overview of the transSMART-XNAT link.....	36
Figure 13 Three dimensional viewing of brain MRI scan using web-based XNAT viewer .....	37
Figure 14 Screenshot of the Clinical Consequences of Protein Sequence variation (CCoPS) .....	41

## Tables

Table 1 Overview of 3-Phased Semantic Web Pilot .....	4
Table 2 Overview of new RDF data integration for D4.6.....	11
Table 3 Technical details of RDF implementation.....	1



# 1 Executive Summary

European e-Infrastructure projects are increasingly turning to Semantic Web<sup>1</sup> technologies to address data integration challenges. This approach is proving to be a solution to some of the emerging challenges in the life sciences. The BioMedBridges semantic web pilot spans deliverables 4.4, 4.6, 4.7, and 4.8; its goal is to test the suitability of a semantic web approach to the task of integrating research data and to report on our experience of running an RDF-based platform integrating multiple data resources.

In order to leverage experience where it exists and minimise the risks inherent in novel technology projects, a three-stage delivery was chosen for the pilot. As summarised below, these stages are reported on in separate deliverables.

**Table 1 Overview of 3-Phased Semantic Web Pilot**

	Del #	Due	Deliverable focus	Partners funded for deliverable	Nature of the activities
<b>Prep</b>	D4.4	2013	Planning	EMBL-EBI (ELIXIR), HMGU (Infrafrontier), TUM-MED (BBMRI), STFC (Instruct), UDUS (ECRIN)	Development of a roadmap for the semantic web pilot project overall.
<b>Phase I</b>	D4.7	Dec 2014	SemWeb scalability	EMBL-EBI (ELIXIR)	ELIXIR establishes mature semantic web services, basic best practices, and technical guidelines. Benchmarks technology and assesses scalability.
<b>Phase 2</b>	D4.6	June 2015	Data integration	ErasmusMC (Euro-BioImaging), EMBL-EBI (ELIXIR), HMGU (Infrafrontier), TUM-MED (BBMRI), STFC (Instruct), UDUS (ECRIN), VUMC (EATRIS)	Infrastructures implement individual pilot projects in parallel, according to their respective roadmaps, using the technical guidelines and outcomes from phase I.
<b>Phase 3</b>	D4.8	Dec 2015	Data integration	ErasmusMC (Euro-BioImaging), EMBL-EBI (ELIXIR), HMGU (Infrafrontier), TUM-MED (BBMRI), STFC (Instruct), UDUS (ECRIN), VUMC (EATRIS)	Report

<sup>1</sup> <http://www.w3.org/standards/semanticweb/>



By following this schedule and aligning partner-specific roadmaps to the blueprint delivered in Phase I (D4.4), the pilot projects can be developed synchronously and knowledge shared efficiently (Phases 2-3: D4.6-4.8). This enables infrastructures to collaborate effectively and address common issues. To support this effort, a knowledge-exchange workshop ran on the 29-30 April 2014 at TMF, Berlin, Germany. The programme and course materials are available on the BioMedBridges website<sup>2</sup>. In December 2013 and December 2014, at SWAT4LS3 and May 2015 at an industry workshop, we delivered tutorials during which we demonstrated the queries and analyses the RDF platform makes possible. We provided a training course to Computer Scientists in Manchester on the practicalities of working with, and running the RDF platform and summarised our technological approach and technology experience. All of the SemWeb pilot work is informed by the work of WP3 on choice and use of ontologies as well as provision and re-use of identifiers and reflects the application of standards in the use case work package.

However, RDF is not a one-size-fits-all technology; and our experience is that certain kinds of data are better suited to different distribution and integration mechanisms. D4.6 therefore includes some pilots that are not directly related to RDF. Here we summarize our processes, products, and lessons learned and identify future work for D4.8.

## 2 Project objectives

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

No.	Objective	Yes	No
1	Implement shared standards from work package 3 to allow for integration across the BioMedBridges project	x	
2	Expose the integration via use of REST based WebServices interfaces optimised for browsing information	x	
3	Expose the integration via use of REST based WebServices interfaces optimised for programmatic access	x	

<sup>2</sup> <http://www.biomedbridges.eu/trainings/knowledge-exchange-workshop-resource-description-framework-rdf>

<sup>3</sup> <http://www.swat4ls.org/workshops/berlin2014/>



4	Expose appropriate meta-data information via use of Semantic Web Technologies	x	
5	Pilot the use of semantic web technologies in high-data scale biological environments		x

## 3 Detailed report on the deliverable

### 3.1 Background

This 4.6 deliverable report cover the activities of the third phase of the Semantic Web Pilot, in which partners have implemented specific pilot projects in parallel, according to their respective roadmaps from D4.4, using the technical guidelines and outcomes from D4.7.

### 3.2 Individual pilot projects

#### 3.2.1 Overview of individual pilot projects

**Use Cases:** D4.6 comprises nine individual pilot projects, each of which:

1. Was developed in response to a specific use case within the BioMedBridges community
2. Integrates diverse kinds of data spanning the domains within the BioMedBridges community (Figure 1)
3. Incorporates modern web standards including ontologies and Semantic Web technologies where appropriate

**Axes of integration:** These pilot projects provide complementary resources that are not all currently amenable to a single integrated query. Accordingly, we chose three different “axes” of integration:

- [RDF](#) for public databases, both archival (e.g. BBMRI) and added value (e.g. Metabolomics)
- [tranSMART](#) to securely integrate private clinical datasets
- [BioJS JavaScript widgets](#) to visualize and integrate existing web resources

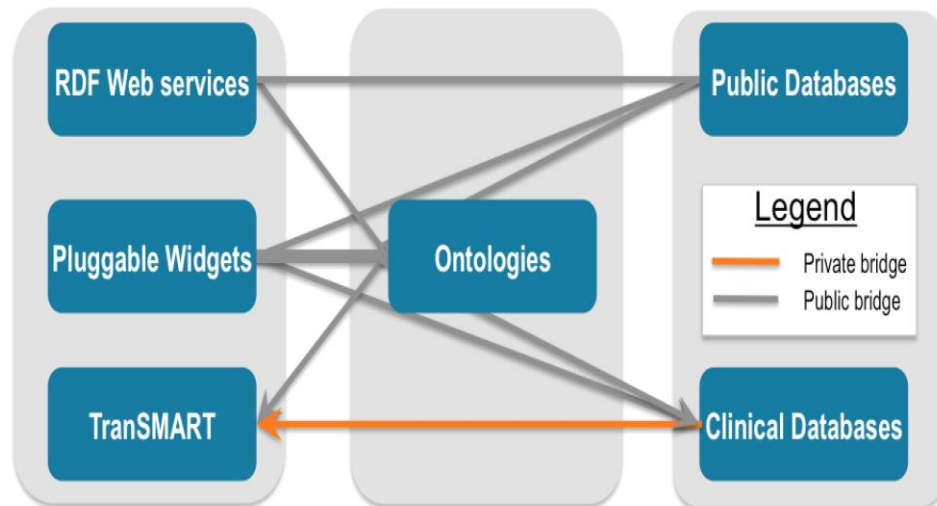


Figure 1 Axes of technical integration

Axis	Pilot title	Ontologies	Publication data	Gene data	Protein data	Organism data	Biosample data	Disease data	Drug data	Clin trial data	Imaging data
Public databases to RDF	Integrating clinical trials metadata with the genes, drugs, and publications they reference	planned	operational	planned				operational	operational	operational	
	Integrating BBMRI and Bio-SD biosample catalogues with RDF	operational					operational	operational			
	Integrating mouse phenotype data for studying diabetes and obesity	operational		operational		operational	operational	operational			
	Additions to EBI RDF platform: Literature text mining, metabolomics and GWAS*	operational	operational	operational	operational	planned	operational	operational	operational		
Private clinical data (tranSMART)	Centralized correlative analysis between image-derived data and other clinical data	planned					operational	operational	operational	operational	operational
	Connect clinical and lab workflows using tranSMART and Galaxy	planned		operational	operational			operational	operational	operational	operational
	CDISC ODM integration with tranSMART	planned						operational	operational	operational	operational
BioJS widgets	Clinical Consequences of Protein Sequence variation (CCOPS)	operational		planned	operational			operational			
	PLATO Plugin for Autocompletion on Ontologies	operational									

Figure 2 Data touching points across the D4.6 pilot projects

**Sustainability:** To maximize the sustainability of these pilots, each is:

- **User-driven:** Iteratively refined to maximise usefulness.
- **Distributed:** Maintained by the research infrastructure that developed it.
- **Open software:** Software developed is accessible, free, and when production ready registered in the ELIXIR tools and data services registry.



- **Open data:** Data, where possible, is open and accessible. High-level summaries of protected human datasets exist to aid discoverability and collaboration.

Axis	Pilot title	Src code deposited	Software licensed	Hosting	Bug fixes	New features	Expertise	Funding
Public databases to RDF	Integrating clinical trials metadata with the genes, drugs, and publications they reference	Firm commitment	Firm commitment	Tentative plans	Tentative plans	Tentative plans	Firm commitment	Tentative plans
	Integrating BBMRI and Bio-SD biosample catalogues	Firm commitment	Tentative plans	Tentative plans	Tentative plans	Tentative plans	Firm commitment	Tentative plans
	Integrating mouse phenotype data for studying diabetes and obesity	Firm commitment	Firm commitment	Firm commitment	Firm commitment	Firm commitment	Firm commitment	Firm commitment
	Additions to EBI RDF platform: Literature text mining, metabolomics and GWAS*	Firm commitment	Firm commitment	Firm commitment	Firm commitment	Firm commitment	Firm commitment	Firm commitment
Private clinical data (transSMART)	Centralized correlative analysis between image-derived data and other clinical data	Firm commitment	Firm commitment	Firm commitment	Firm commitment	Firm commitment	Firm commitment	Firm commitment
	Connect clinical and lab workflows using tranSMART and Galaxy	Firm commitment	Firm commitment	Firm commitment	Firm commitment	Firm commitment	Firm commitment	Firm commitment
	CDISC ODM integration with tranSMART	Firm commitment	Firm commitment	Firm commitment	Firm commitment	Firm commitment	Firm commitment	Firm commitment
BioJS widgets	Clinical Consequences of Protein Sequence variation (CCOPS)	Firm commitment	Firm commitment	Firm commitment	Firm commitment	Firm commitment	Firm commitment	Firm commitment
	PLATO Plugin for Autocompletion on Ontologies	Firm commitment	Firm commitment	Firm commitment	Firm commitment	Firm commitment	Firm commitment	Firm commitment

**Legend**

Firm commitment

Tentative plans

\* not directly BMB funded

**Figure 3 Partner plans for sustainability of individual D4.6 pilot projects beyond BioMedBridges**

**Challenges for the future beyond BioMedBridges:** Many valuable clinical datasets/studies remain undiscoverable because there is not a suitable repository with which to register them. Existing “long-tail” repositories such as Dryad and Zenodo are for depositing data rather than registering it per se. Biostudies is an emerging repository that may be useful in this context.

### 3.2.1.1 Integration of simple object queries using RDF-based web services

Two new BioMedBridges datasets (mouse phenotypes, and BBMRI) have now been made available as RDF expressly for this deliverable (D4.6). We also report on three additional datasets (GWAS, EuropePMC, and MetaboLights) that were added to the EBI RDF platform; these three additions were supported with complementary funding (see Table 1 and Appendix A.3) but





are mentioned here as they make use of the experience, training, and infrastructure that ELIXIR established in D4.7 with BioMedBridges support.

The ELIXIR RDF platform is now used by 15 ELIXIR software applications and 11 external dependent groups, including several industrial users (see below). It is also a dependency for recently awarded EU grants including CORBEL and EXCELERATE. There are around 55 million hits to the site monthly. The platform executes 99% of queries in less than one second. The platform served over 115 million queries in 2014, though a precise number of unique queries is difficult to obtain since a very large fraction arise from federation of queries across multiple endpoints. Moreover the RDF is often downloaded in bulk in order to execute queries in a private/secure fashion by pharmaceutical researchers. (Downloads are not currently tracked). Full usage data statistics are being compiled now for inclusion in the final deliverable, D4.8.

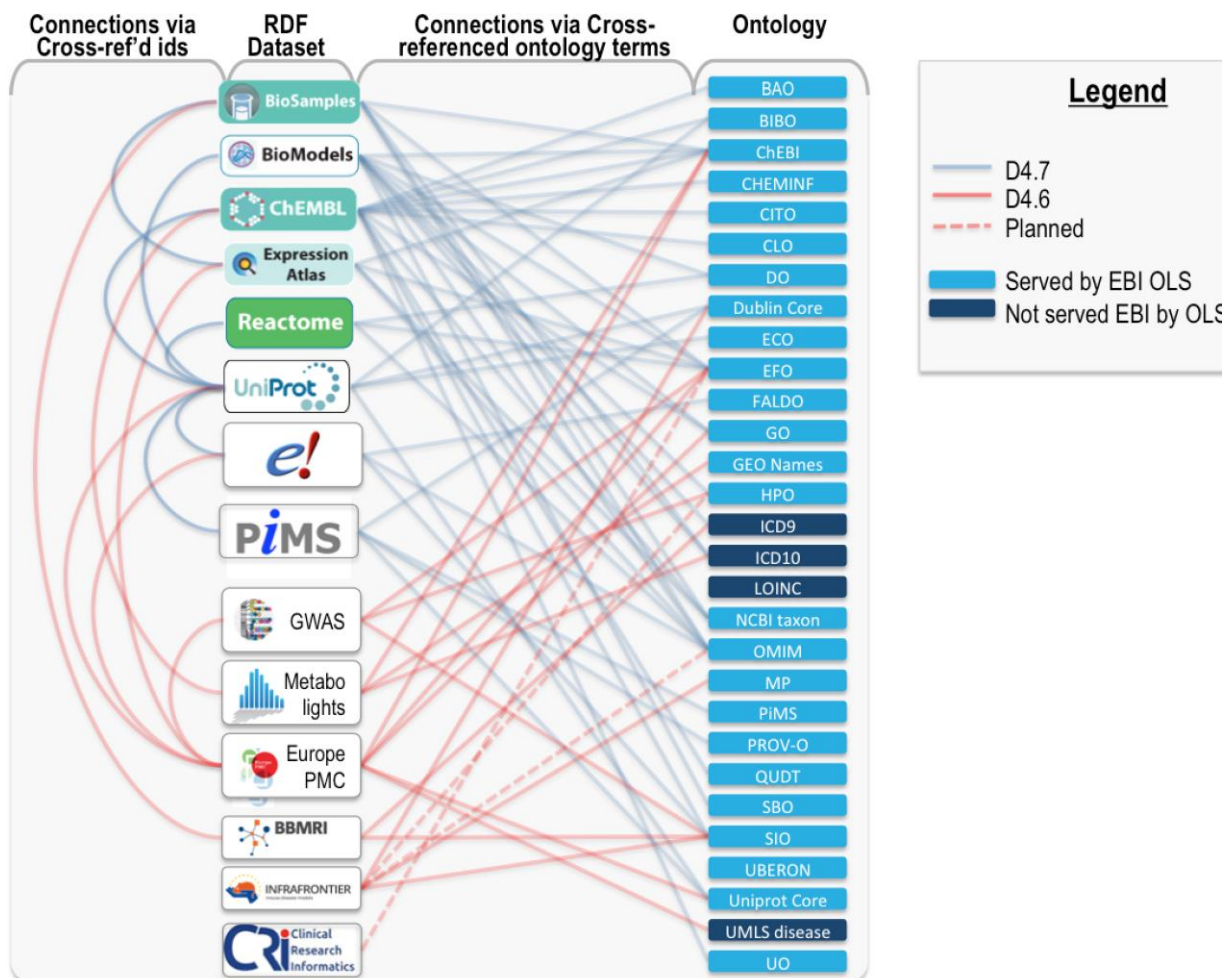
Members of the EMBL-EBI Industry Programme (Eli Lilly, UCB and Syngenta) have recently committed to Linked Data strategies for their global integrated data operations. The scale of investments is large, and may fuel a coalescence behind these technologies as triplestore suppliers seek to align to client requirements. Pilot studies within the Centre for Therapeutic Target Validation<sup>4</sup> are actively leveraging the RDF behind many of their core services. Other large global corporations such as Novartis, Novo Nordisk, AstraZeneca and GlaxoSmithKline are also making use of the technology. Small firms such as General Bioinformatics<sup>5</sup> also rely heavily on the platform. New European project proposals will require further development of resources in this area. RDF supports the recommendations by The Data FAIRport<sup>6</sup> initiative for data interoperability and re-use.

---

<sup>4</sup> <http://www.targetvalidation.org/>

<sup>5</sup> <http://www.generalbioinformatics.com/>

<sup>6</sup> <http://www.datafairport.org/>



**Figure 4 Ontology use by Dataset.** Connecting lines show linkages between datasets (left) and ontologies (right). Blue connecting lines are those previously reported in D4.7; red connecting lines are new as of 4.6. Dashed lines show planned uses. All ontologies shown are served by the EBI Ontology Lookup Service, with the exception of ICD-9, ICD-10, and UMLS which have redistribution restrictions. Some ICD-9 and ICD-10 terms are cross-referenced from the EFO application ontology and also the Disease Ontology can be used to map these. Direct links between datasets (as opposed to between datasets and ontologies) reflect cross-referenced identifiers. This figure is not exhaustive; many links to other databases exist but are omitted for simplicity. See **Table 2** and **Table 3** for details



**Table 2 Overview of new RDF data integration for D4.6**

Datatype	Database	Example query	Data integration	Ontologies
Biobank metadata and sample counts	BBMRI-LPC catalogue BBMRI.eu catalogue	Which biobanks focus on Neoplasm and contain at least 1000 tissue samples?	BioSD	SIO, ICD10
Mouse phenotype data	IMPC sample dataset	Which alleles are related to phenotypic alteration in the Diabetes relevant IPGTT procedure?	Mousephenotype data from IMPC, Mouse-Clinic and MGI curations. MGI marker and allele data. IMPRESS parameter, procedures and pipelines.	MGI,MP,DIAB,SIO, Dublin Core
Text-mined named entities in full text literature	Europe PMC	Show all the sentences in methods sections where PDB accession number 3NSS is mentioned.	ENA, RefSNP, PDB, UniProt, Pfam, ArrayExpress, RefSeq, data DOI, Ensembl, and InterPro	OA (Open Annotation) used to annotate to UniProt, ChEBI, GO, EFO, NCBI Taxonomy, OMIM, and UMLS Disease
Metabolomics data	Metabolights	Show all ChEBI compounds that have role herbicide	ChEBI, MassBank, DrugBank, ChEMBL	ChEBI, GeoNames
Genome wide association studies	GWAS	Show all GWAS traits for diabetes.	dbSNP and Ensembl	EFO
Clinical Trials	CTIM	(Not available; in development)	(Not available; in development)	(Not available; in development)



**Table 3 Technical details of RDF implementation**

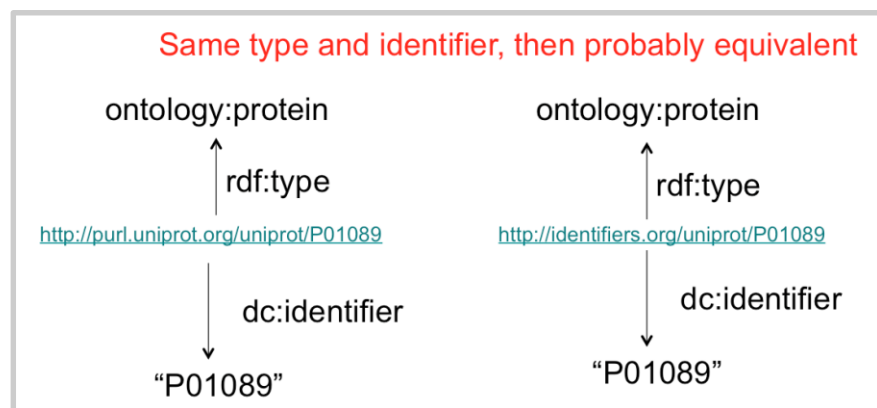
Datatype	Database	RDF-ization approach	Number of triples produced	General tools and resources used for conversion	Current challenges	RDF-ization supported primarily by	Infrastructure
Biobank metadata and sample counts	BBMRI-LPC catalogue BBMRI.eu catalogue	RDB conversion	1103	Apache Jena	N/A	BioMedBridges WP4	BBMRI-ERIC
Mouse phenotype data	IMPC sample dataset	RDB/Flatfile conversion	116,039	Virtuoso, Tomcat, Lodestar	Include all IMPC data; Mouse clinic data; MGI curation integration	BioMedBridges WP4	INFRAFRONTIER
Text-mined named entities in full text literature	Europe PMC	free text to RDF	1,563,241,810	Europe PMC text-mining pipeline	The current text-mining RDF store is a pilot with a static data set. It is not fully public. Scaling and updating still outstanding.	Europe PMC	ELIXIR
Metabolomics data	Metabolights, MassBank	free text to RDF	9233	Tomcat & LodeStar	The current RDF store is a pilot with a static data set. It is not fully public. Scaling and updating still outstanding.	COSMOS	ELIXIR
Genome wide association studies	GWAS	RDB conversion	TBD	OWL API	Scaling issues with OWL reasoner; Explore JSON-LD	NHGRI	ELIXIR
CTIM	ClinicalTrials.gov via CTIM	(Not available; in development)	(Not available; in development)	(Not available; in development)	(Not available; in development)	BioMedBridges WP4	ECRIN



### 3.2.1.2 Lessons learned with RDF knowledge representation

What we have collectively learned is that generating RDF is straightforward for mature resources where the data model is well understood, but that generating RDF that integrates well with other data is challenging. Below is a list of things that make RDF more easily integrated. Some of these lessons (noted with an asterisk below) are excerpted from D4.7 and updated here with recent experience. We recently conducted RDF training for an EBI industry workshop (materials available in Appendix A.2); among participants there was significant interest in learning what our practical experience had been with the RDF platform and what recommendations we could offer to implementers and users.

- Use common URIs for things\* (see also our recently submitted manuscript on identifier design, provision, and reuse<sup>7</sup>)
- Shared design patterns / predictable ways to integrate identifiers, e.g.:



- When creating full URIs, the safest option is for data providers to create full URIs in their own domain and cross ref to identifiers.org for supporting integration. (See detail in Appendix A.2)
- Common schemas/ontologies for typing resources<sup>8</sup>
- Well defined predicates to relate resources\*
- When modelling the data, there are trade-offs between fitting use cases and purely representing knowledge. There are two main advantages to the former: a) it is otherwise quite difficult to know when the model is good/complete enough to publish and

<sup>7</sup> <http://zenodo.org/record/18003>

<sup>8</sup> Previously described in D4.7



- b) it is otherwise possible to end up with a model that is over-engineered or not fit for any purpose
- When modelling the data, there are trade-offs between lightweight schemas or heavyweight ontologies. The choice is use-case dependent.
  - RDF is not a good fit for JavaScript widgets until the JSON-LD standard is more adapted
  - RDF providers should support SPARQL queries that result in human readable names as well as opaque identifiers, otherwise it is much less useful for generating hypotheses or debugging
  - RDF providers should not use excessively general properties as it makes queries more complex to write and more demanding to process. Instead, links of convenience can be added.
  - Semantically strong typing is the main advantage of RDF, so we should focus on this aspect when producing the data. Ontologies should also be a consideration even if RDF is not an immediate aspiration as they enable semantic typing which enables data integration using other technologies.
  - When database entries correspond to real-world entities, modelling identifiers has important trade-offs. There are three basic approaches. 1) Use entries as proxy for the real world entity, 2) model identity using OWL 3) model identity using strong typing (e.g. EDAM). The trade-offs are semantic richness, LOD friendliness, and triple bloat. (See detail in Appendix A.2)

### **RDF future work**

The lessons in 4.7 and here above will serve as a starting point for a white paper to be developed in the coming year. Such a paper has been a frequent request from industry. We will continue to seek out and collect key performance metrics for the RDF provided and will continue to address work on documenting the data, including clear data licensing / terms of use.



### 3.2.1.2 Integration of simple object queries using tranSMART

Creating bridges for medical research remains one of the project's core objectives, but one which comes with many challenges. We considered RDF as an exchange format for clinical data but decided against it for two reasons:

1. This would contradict existing industry standards (e.g. DICOM for images) already heavily used (by pharma, academics, vendors of instruments and software),
2. Security of graph-based data is still very immature and patient confidentiality prohibits the public distribution of source data.

We found a suitable alternative in TranSMART, an “open-source”, community-driven knowledge management platform for translational medicine. TranSMART is a project that is being collaboratively developed by more than 100 computer scientists and physicians from more than 20 organizations from around the world. It is open to all and can enable pre-competitive and private data sharing within and across organizations<sup>9</sup>. It is the platform of choice for the Netherlands premier translational research informatics program, CTMM-TRAIT<sup>10</sup> in addition to that of the Innovative Medicines Initiative<sup>11</sup>, “Europe's largest public-private initiative aiming to speed up the development of better and safer medicines for patients.”

Three tranSMART-driven pilots were pursued. They are described in detail in **Appendix B**:

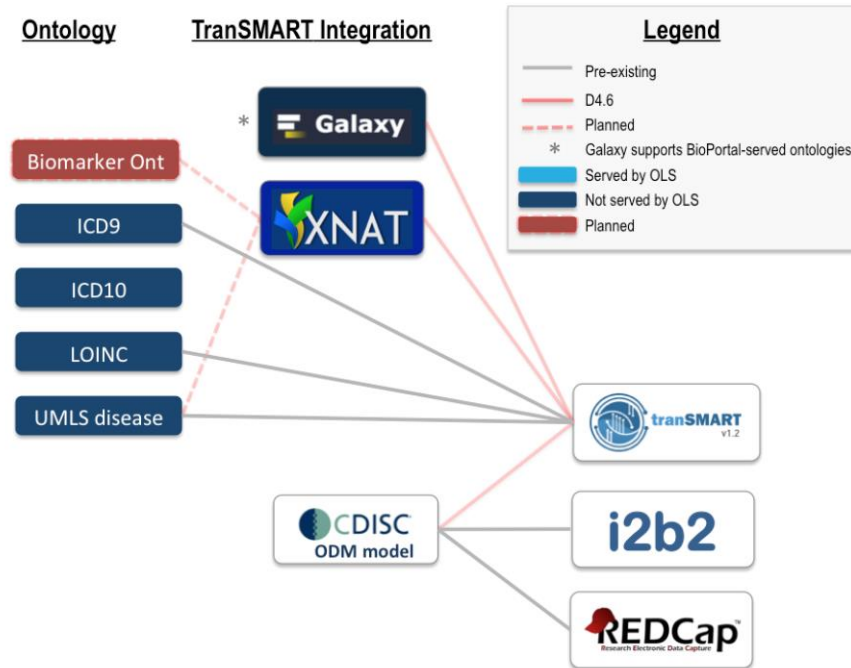
- Centralized correlative analysis between image-derived data and other clinical data
- Connect clinical and lab workflows using tranSMART and Galaxy
- CDISC Operational Data Model (ODM) integration with tranSMART

---

<sup>9</sup> <http://transmartfoundation.org/overview-of-platform/>

<sup>10</sup> <http://www.ctmm-trait.nl/>

<sup>11</sup> <http://www.imi.europa.eu/>



**Figure 5 Overview of TransSMART Integration Pilots**

### TransSMART lessons learned

TransSMART proved to be a powerful and flexible framework for data integration in translational medicine. To improve the usability of tranSMART in multi-center clinical studies, we suggest the following directions for future developments by the tranSMART community:

1. Import of data into tranSMART via the web interface should be implemented for other data sources than imaging data as well, in a similar fashion as worked out in the pilot on “Centralized correlative analysis between image-derived data and other clinical data”. This would greatly streamline central correlative data analysis in multi-center studies.
2. Functionality of tranSMART in comparing multiple biomarkers of the same subject (or baseline and follow-up measures of the same patient) should be improved. Current functionality seems rather designed for comparing measures across patients.

Finally, In our experience, the Grails plugin framework provides a convenient “modular” approach of adding extensions to tranSMART.



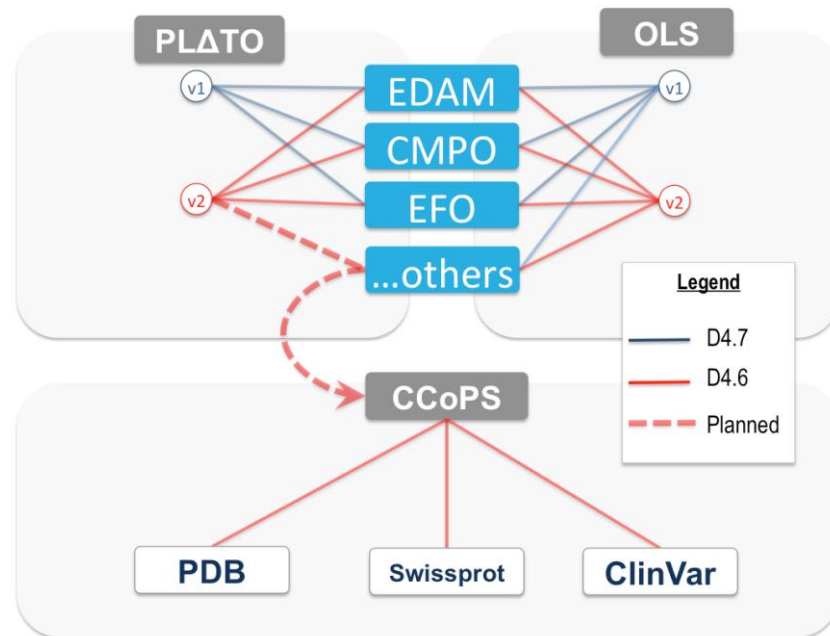


## TranSMART future work

The source code for the TranSMART platform and the plugins we have developed are already housed in open source repositories; however, an important next step is to deploy the platform for CTMM-TRAIT researchers to use. We anticipate that the beta release for the hosted system is expected to occur in late 2015 or early 2016. At that point we will conduct formal and informal user experience testing of both the platform and plugins. The system will be iteratively improved accordingly. Please see **Appendix B** for additional details on these pilots. An overview of the integration between the platforms and ontologies is shown above in Figure 5.

### 3.2.1.3 Integration of simple object queries using BioJS widgets

There are two BioJS widget pilots that were part of D4.6: The Plugin for Autocomplete on Ontologies (**PLATO**), and Clinical Consequences of Protein Sequence variation (**CCoPS**). PLATO is a multipurpose widget for leveraging ontologies when annotating data or when searching across ontology-annotated data. CCoPS integrates sequence variation data with data about the clinical consequences of that variation; it layers this information within a reference structure visualisation for the protein within PDB. See **Appendix C** for details on both of these pilots; a summary is below.



**Figure 6 Summary of BioJS widget-driven pilots.** Version 1 of the PLATO widget was deployed to work with version 1 of OLS (EBI Ontology Lookup Service). The PLATO widget (Version 2) was deployed to work a) depend on a newer and better-supported software library (Select2) and also b) consume services from the newly redeveloped Ontology Lookup Service (see Appendix C.2 for details). CCoPS integrates two webservices: 1) protein structure from PDB and 2) clinical impact of protein sequence modifications (ClinVar). CCoPS then implements a BioJS plugin (SwissProt) to visualize these two data sources together

### Widget lessons learned

It is ideal to build on existing services, rather than to re-develop them; however we encountered several problems from a variety of providers.

- 1) Identifiers that are not persistently resolvable
- 2) Documentation that does not correspond to the correct data release
- 3) Failure to provide a Last-Modified header, although the date is present in human-readable form
- 4) PURLs that do not support an [HTTP HEAD](#) requests (needed to efficiently determine whether a page has changed without downloading the entire body of the response)
- 5) Many services provide HTML only; no machine readable representations, not even RSS
- 6) Malformed RSS, XML
- 7) HTTPS certificate issues makes documentation link unusable in some browsers



- 8) CORS (Access-Control-Allow-Origin) is not allowed, or is allowed but not consistently
  - a) Some that allow CORS do so only for certain kinds of requests (eg 200, but not 303)
  - b) One major provider refuses to allow CORS on security grounds (It is our opinion that this concern can be addressed through issuance of secure keys to consumers of the service).
- 9) Finally, help desks from several major providers responded slowly to the reported problems above, some of which remain unsolved due to a variety of reasons including their level of available resources.

### **Widget future work**

Recently-added PDB webservice will now allow PDB protein structures to be queried using Gene Ontology terms corresponding to annotations of function, compartment and biological process and taxonomies (among others). The PLATO widget may therefore be deployed within CCoPS in order to make effective use of this new PDB query availability.

## **3.3 Overall conclusion D4.6**

For D4.6, we have built working implementations of nine pilot projects, we note that RDF is well suited to some use cases and not others. However, ontologies and identifiers have proven to be important considerations, no matter what the technology choices.

## **3.4 Overall future work**

### **3.4.1 Overall future work for D4.8**

To facilitate discoverability of the various pilot projects, we have developed a summary website here: <http://www.biomedbridges.eu/bmb-software-matrix>



## BMB Software Matrix

Pilot title (short name for the work done for BMB)	REST	GUI	RDF	Source	BioJS
Visualising and leveraging ontologies in queries: PLATO Plugin for Autocompletion on Ontologies					
Sharing and integrating medical imaging data					
Integrating mouse phenotype data for studying diabetes and obesity					
Sharing protein engineering knowledge					
Pilot federated biobank catalogue					
Biosample information integration and discovery					
Leveraging the utility of compound screening functional assays					
Connectivity-Based Searching in UniChem					
Connect clinical and lab workflows using tranSMART and Galaxy					
Sharing and visualising sequencing data from environmentally-derived biological samples					
Integrating clinical trials metadata with the genes, drugs, and publications they reference					

**Figure 7 BioMedBridges software matrix shows pilots (rows) with links to REST web service endpoints, visual web interface (GUI), RDF, source code, and BioJS documentation**

This currently includes only a subset of the pilots, but will be completed in time for D4.8. See the individual pilot reports (Appendices) for specific future work on each pilot.

### 3.4.2 Beyond BioMedBridges

We have identified efforts that would be valuable to the BioMedBridges community, but which are outside of the scope or resourcing of this current grant:

**Best practices for data services:** In the process of developing CCoPS and for other web service-related efforts in BioMedBridges more broadly, we encountered many challenges, some of which were related to the provision of identifiers; we recently submitted a manuscript<sup>12</sup> that speaks to those. However, we encountered several other problems that extend beyond that

<sup>12</sup> <https://zenodo.org/record/18003>



paper's scope; this experience has prompted us to consider writing an additional paper specifically about the provision of data services on the web.

**Data licensing:** Another problem that we recognize needs to be addressed is the lack of clear terms of use / licenses for data, not just for software. This is an area that is actively being examined.

**Discoverable pre-clinical studies:** Formal clinical trials are very expensive, high profile and generally discoverable in repositories such as ClinicalTrials.gov. However, there is a long tail of important pre-clinical and observational studies that are discoverable primarily by the published literature and word of mouth / social media. Since opportunities for collaboration are not always discoverable in a timely manner, the impactfulness of these studies is potentially hampered. Addressing this gap is important, but there remain social, technical, and legal challenges. A first important step has been made by initiatives like CTMM-TraIT in The Netherlands by establishing well-maintained central repositories (OpenClinica, XNAT, etc.) for these smaller, investigator-driven studies, and linking them together using TranSMART. The next step to make these studies truly discoverable would be inclusion of selected high-level study data in meta data repositories such as the newly launched "[BioStudies](#)" platform, from the EBI. For BioMedBridges however, we have focussed our development efforts on facilitating multi-centre, multimodal translational studies, incorporating biological knowledge and ontologies where possible. The work being done in BioMedBridges raises partners to a greater level of data integration readiness whereby in the future, pre-clinical findings can more readily inform biological insights.

## 4 Delivery and schedule

The delivery is delayed:  Yes  No

## 5 Adjustments made

None.



## 6 Background information

This deliverable relates to WP 4; background information on this WP as originally indicated in the description of work (DoW) is included below.

WP 4 Title: Technical Integration

Lead: Ewan Birney (EMBL)

Participants: EMBL

In work package 4 we will implement a federated access system to the diverse data sources in BioMedBridges. This will focus on providing access to data or metadata items which utilise the standards outlined in WP 3. Experience across the BioMedBridges partners is that executing a federated access system, in particular a federated query system, is complex for both technological and social reasons. Therefore we will be using an escalating alignment/engagement strategy where we focus on technically easier and semantically poorer integration at first and then progressively increase the sophistication of the services. In each iteration, we will be using biological use cases which are aligned to the capabilities of the proposed service, thus providing progressive sophistication to the suite of federated services.

Our first iteration involves using established REST based technology to provide userbrowsable visual integration of information. This will be useful for both summaries of data rich resources (such as Elixir) and summaries of ethically restricted datasets where only certain meta-data items are public (such as BBMRI, ECRIN and EATRIS). We will then progress towards lightweight distributed document and query lookups, where the access for ethically restricted data will incorporate the results of WP 5. Finally at the outset of the project we will explore exposure of in particular meta-data sets via RDF compatible technology, such as SPARQL, and the presence of the technology watch WP11 will provide recommendations for other emerging technologies to use, aiming for the semantically richest integration.

<b>Work package number</b>	WP 4	<b>Start date or starting event:</b>							month 1
<b>Work package title</b>	Technical Integration								
<b>Activity Type</b>	RTD								
<b>Participant number</b>	1:EMBL	4:STFC	5:UDUS	6:FVB	7:TUM-MED	9:ErasmusMC	11:HMGU	13:VUMC	
<b>Person-months per participant</b>	69	40	38	0	37	15	32	37	

### Objectives

1. Implement shared standards from WP 3 to allow for integration across the BioMedBridges project
2. Expose the integration via use of REST based Web services interfaces optimised for browsing information
3. Expose the integration via use of REST based Web services



- interfaces optimised for programmatic access
4. Expose appropriate meta-data information via use of Semantic Web Technologies
  5. Pilot the use of semantic web technologies in high-data scale biological environments.

### **Description of work and role of participants**

We will provide a layered, distributed integration of BioMedBridges data using latest technologies. A key aspect to this integration will be the internal use of standards, developed in WP 3 which will provide the points of integration between the different data sources. The use of common sample ontologies (WP 3) will provide integration between biological sample properties, such as cell types, tissues and disease status, in particular bridging the Euro-Biolmaging, BBMRI, Elixir and Infrafrontier projects. The use of Phenotype based ontologies will provide individual and animal level characterisation which, when these can be associated with genetic variation, will provide common genotype to phenotypic links, and this will be used to bridge the ECRIN, EATRIS, INSTRUCT, BBMRI, Infrafrontier and Elixir Projects. The use of environmental sample descriptions and geolocation tags will bridge between EMBRC, ECRIN, ERINHA, EATRIS and Elixir. The use of chemical ontologies will help bridge between EU-OPENSOURCE, ECRIN, Euro-Biolmaging, INSTRUCT and Elixir. By applying these standards in the member databases (themselves often internally federated) we will create a data landscape that theoretically can be traversed, data-mined and exploited. To expose this data landscape for easy use, we will deploy a variety of different distributed integration technologies; these technologies are organised in a hierarchy where the lowest levels are the semantically poorest, but easiest to implement, whereas the highest levels potentially expose all information in databases which are both permitted for integration (some are restricted for ethical reasons, see WP 5) and can be described using common standards. We will develop software with aspects appropriate for the distributed nature of this project taken from agile engineering practices, such as rapid iterations between use cases and partial implementation. In particular we will be using the enablement/alignment strategy (Krcmar H., Informationsmanagement, Springer) to ensure that the use cases that drive the project are aligned to feasible capabilities that can be delivered. The work package will be implemented in a collaborative manner across the BMSs, with frequent physical movement of individuals.

The proposed technologies are:

1. REST-based “vignette” integration, allowing presentation of information from specific databases in a human readable form. An example is shown in Figure 1. These resources allow other web sites to “embed” live data links with key information into other websites. This infrastructure would then be used to provide browsers that, on demand, bridge between the different BioMedBridges groups – for example, information which can be organised around a gene or a chemical compound would be presented across the BioMedBridges project.
2. Web service based “query” integration, where simple object queries across distributed information resources can be used to explore a set of linked objects using the dictionaries and ontologies present. Each request will return a structured XML document.
3. Scalable semantic web based technology. We are confident that



semantic based technology can work for the rich but low data volume meta data (eg, sample information) which we will expose using semantic web technologies such as RDF and SPARQL. However, it is unclear whether this scales to the very large number of data items or numerical terms in the BioMedBridges databases (such as SNP sets or numerical results from Clinical trials) We will pilot a number of semantic web based integration of datasets, using RDF based structuring of datasets In the latter phases of the project we will look to align these solutions to other broader standards in the eScience community, taking input from the Technology Watch (WP11) group; we hope in many cases our technology choice which has been already informed by alignment to future eScience technology (e.g. RDF/SPARQL) so this may only require appropriate registration/publication of our resources. Where unforeseen but useful technologies are developed we will build systematic connections from these BioMedBridges federation technologies to other federation technologies.

#### Deliverables

No.	Name	Due month
D4.1	A brief collation of existing use cases to start the agile software iteration	3
D4.2	Assessment of feasible data integration paths in BioMedBridges databases	6
D4.3	Pilot integration using REST Web Services	18
D4.4	Identification of feasible BioMedBridges pilots for semantic web integration	18
D4.5	Pilot integration of REST based vignette services for the second round BMS projects	24
D4.6	Pilot integration of Web Services based simple object queries	36
D4.7	Report on the scalability of semantic web integration in BioMedBridges	36
D4.8	Report on Web Services based integration of BioMedBridges integration across all appropriate services	48





## Appendix A RDF pilots

### Appendix A.1 HMGU (INFRAFRONTIER): Integrating systemic mouse phenotype data from diverse sources

#### Background

Mouse phenotype data makes an important contribution to the study of human diseases. Data is generated in single phenotyping centres (e.g. German Mouse Clinic(GMC)), large-scale phenotyping projects (e.g. IMPC), and through the manual curation of publication data (e.g. as available in the MGI database).

To integrate this data we focussed on developing a semantic web model which is capable of integrating and analyzing data from these different fields. Moreover, it enables comprehensive integration of mouse phenotype data with emerging SPARQL endpoints from the various research infrastructures on the ESFRI roadmap, to enable new human-mouse phenotype bridges.

#### Scientific use case

Mouse phenotype data can aid the development and testing of hypotheses in various scientific fields. This data is made even more impactful due to its rich annotations which allow it to be mapped to annotated human phenotype data (via HPO mappings, as shown in WP7 DIAB ontology). Moreover, the measured parameter sets (e.g. blood glucose) correspond to assays on the human side.

Here, we employed semantic web technologies to enable integration of systemic phenotype mouse data from IMPC, MGI together with data from single mouse clinics. The work done in 4.6 makes it possible for researchers to ask questions such as “Which alleles are related to phenotypic alteration in the Diabetes relevant IPGTT procedure and has been validated by mouse phenotyping experts and statistical analysis?”<sup>13</sup>.

---

<sup>13</sup> See other sample queries at <http://mousemodels.infrafrontier.eu/rdf/sparql>



## Previous work

The work done for 4.6 builds on the availability of the datasets, and on the basic REST framework established in D4.3. However, neither the semantic modeling nor the technical implementation of RDF below has been previously reported. The BioMedBridges WP7 PhenoBridge developed interfaces that allow users unfamiliar with mouse models to filter and display mouse phenotyping information according to their specific research interests (e.g. all relevant blood parameters). These user interfaces may be enhanced in the future to leverage the semantic richness made possible in D4.6. (See future work section).

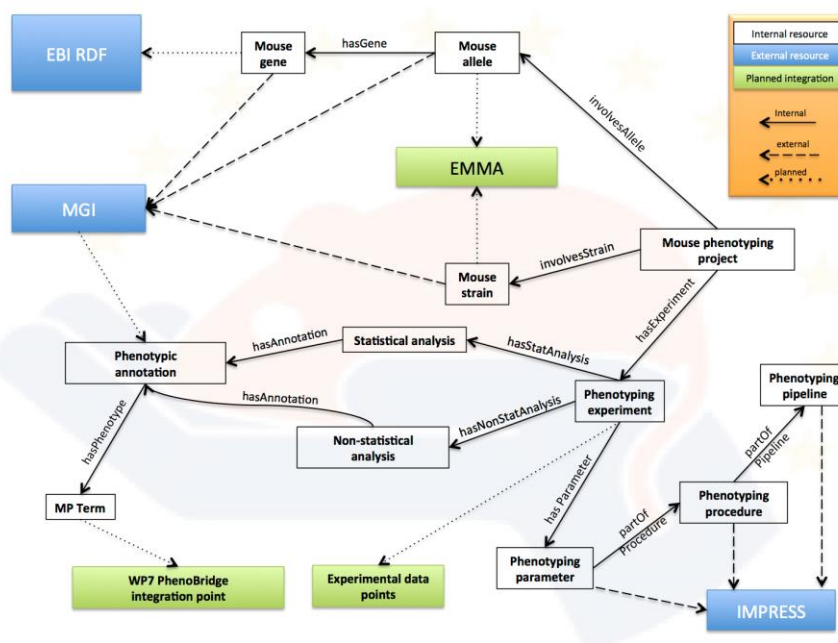


Figure 8 Model for the semantic integration of mouse phenotype resources

## Work done for 4.6

### Technical implementation

We designed our semantic data model to reuse existing identifiers/ontologies (e.g. MGI identifiers) while also maximizing future interoperability with other mouse resources (e.g. raw phenotyping data and European Mutant mouse archive (EMMA) data). The main integration point is the "phenotypic annotation" resource that is designed to combine data from various mouse



phenotype databases (e.g. it can be directly mapped to MGI annotations). During the first iteration, we extracted from IMPC a sample dataset containing all significant genotype-phenotype relationships. Moreover, a set of mouse lines from the GMC were transferred into the RDF triple store. In the second iteration, which is currently in process, all mouse lines from IMPC will be integrated, including legacy data from Europhenome. Finally, MGI curation data will be added. A Virtuoso server implementation is used as a data repository and the Lodestar plugin was integrated for advanced linked-data browsing. The dynamic Phenomap was developed using primefaces and JavaScript. To achieve all these goals we participated in RDF training courses provided by the EBI which and adapted suggestions and developed software(e.g. Lodestar plugin) from D4.7.

- RDF: <http://mousemodels.infrafrontier.eu/rdf/sparql>
- Phenomap GUI: <http://mousemodels.infrafrontier.eu/tools/phenomap.jsf>

### **Metrics**

Usage of SPARQL endpoint and interfaces will be monitored. User experience testing is currently performed internally within the GMC; it will be repeated with external users in July/August 2015.

### **Future work**

- Integration of further datasets from all main sources (IMPC, GMC, MGI)(continuously)
- Diabetes ontology(DIAB) import to allow human phenotype queries. (June/July 2015)
- Definition of further disease-parameter presets for enhanced Phenomap browsing (continuously).
- Extend interfaces for network driven analysis. (September 2015)
- Enrich user interface with other RDF data (from D4.6)



## Appendix A.2 TUM-MED (BBMRI): Integrating human tissue biobanking data across Europe

### BBMRI RDF Background

The BBMRI.eu catalogue<sup>14</sup> provides an overview of the human tissue biobank landscape across Europe. BBMRI-LPC uses an advanced version of this catalogue<sup>15</sup> for the Large Prospective Cohorts of the BBMRI-LPC project. Both catalogues have been linked to BioSD<sup>16</sup>, and the LPC catalogue is being used in the pilot implementation of WP 5, Task 8.

### Scientific use case

Beginning during the preparatory phase of BBMRI, and continued by BBMRI-ERIC and –LPC, use cases have been identified. Some of them are documented in the preparatory phase deliverable 5.4<sup>17</sup>. A typical group of questions can be summarized by: “I am looking for biobanks focusing on disease group x, containing at least y samples of material type z”. Further differentiation includes age groups and sex.

### Previous work

The work in D4.6 further builds on the REST web service implemented in D4.3<sup>18</sup> in order to establish a data bridge between BBMRI and ELIXIR.

### Work done for 4.6

### Technical implementation

To support the above type of query, we specified an OWL ontology defining the classes Biobank, SampleCollection, MaterialType, and ICD10CodeGroup (using the ICD10 ontology<sup>19</sup>) along with various properties. In alignment with the lessons learned documented in D4.7 the ontology is simple, driven by the scientific use case, and focusing on the relevant parts of the data. Triples

---

<sup>14</sup> <https://www.bbmriportal.eu/bbmri/>

<sup>15</sup> <https://www.bbmriportal.eu/lpc/>

<sup>16</sup> <http://www.ebi.ac.uk/biosamples/>

<sup>17</sup> <http://bbmri-eric.eu/reports>

<sup>18</sup> <https://www.bbmriportal.eu/bbmri2.0/bbmri/bbmri.xml?u=biosd&p=biosdpass>

<sup>19</sup> <http://bioportal.bioontology.org/ontologies/ICD10/>



establishing relations to the Semantic science Integrated Ontology (SIO)<sup>20</sup> have been added to the OWL ontology, as recommended in D4.7. An RDF representation of the ontology in Turtle syntax is available online<sup>21</sup>. A Java program has been written which transforms biobank metadata and sample counts into triples according to the OWL ontology using the Apache Jena framework. This program takes data exported by the REST web service realized for D4.3 as its input and represents them in RDF<sup>22</sup>. The triple store is automatically updated nightly. For querying the RDF content, a SPARQL endpoint based on the Fuseki server has been installed and a graphical user interface to the SPARQL endpoint is provided<sup>23</sup>.

### **Metrics**

Usage statistics of the catalogue, including the webservices, are being automatically generated using the software AWStats and used for internal feedback.

### **Future work**

The work done for WP 4 will be further developed and sustained in alignment with BBMRI-LPC and BBMRI-ERIC. Specifically, BBMRI-ERIC will integrate the LPC catalogue in its system landscape.

---

<sup>20</sup> <https://code.google.com/p/semanticscience/wiki/SIO>

<sup>21</sup> <https://www.bbMRIportal.eu/bbmri2.0/bbmri2rdf.ttl>

<sup>22</sup> See <https://www.bbMRIportal.eu/bbmri2.0/rdfexport.n3>

<sup>23</sup> <https://www.bbMRIportal.eu/bbmri2.0/sparql.html>



## Appendix A.3 EMBL-EBI (ELIXIR): additions to RDF platform

Three new datasets are in the process of being added to the EBI RDF platform: metabolomics, literature text mining, and Genome Wide Association Studies. While primary support for the modelling and transformation of the datasets has come from sources other than BioMedBridges, these new datasets are mentioned because they are making effective use of the infrastructure, expertise, and the best practices that the BioMedBridges semantic web pilot has established.

### Metabolomics

Metabolomics experiments measure unique chemical outputs in order to better understand cellular physiology and pathology. The complexity of the relationships makes RDF a potentially capable platform to represent them. Through interactions with MetaboLights users and stakeholders, we collected a set of sample queries that a SPARQL endpoint should be able to answer. These queries encompass a range of granularity and level of integration; for instance “Show metabolites intensities measured from the same samples, but in different assays (positive/negative mode, MS/NMR, using mzMine/XCMS, ...)” or “What are the differentially expressed **genes** for which both **pathway** data and **metabolite profiles** exist”.

In converting MetaboLights to RDF, we discovered that mapping metabolite names to compound identifiers is tricky: while some names clearly refer to a single compound, others may represent a whole class of compounds. For ambiguous cases, we had to find a semantic mapping that reflects this ambiguity. We have generated a pilot RDF dataset that is currently running on a development server. We will continue to refine the model and service iteratively in response to user feedback. We anticipate a public release of the dataset in 2016. Pilot modelling and provision of the MetaboLights dataset was funded primarily by the Cosmos FP7 project; the dataset will be hosted on the EBI RDF platform has been informed by best practices (4.7).

### Literature text mining results

Named Entity Recognition (NER) is one of the main tasks in text mining, and its goal is to extract names of entities (e.g., persons, genes, proteins,



chemicals, etc.) from unstructured free text. Once names are identified, they are linked to ontologies or databases. Publishing these links using RDF enriches the publication as well as the text-mined entities. For example, we can easily enrich these mined named entities with additional information from other RDF resources (e.g. UniProt). As another example, we can share these mined entities with their URIs on Europe PMC articles and provide a tool for readers to make comments on them.

To produce RDF triples from Europe PMC literature database<sup>24</sup>, first we applied our text-mining pipeline, which mainly consists of named entity taggers<sup>25</sup>, accession number tagger<sup>26</sup> and section tagger<sup>27</sup> to Open Access full-text articles<sup>28</sup>, and then we converted the text-mined results into triples based on the Open Annotation Data Model (OADM)<sup>29</sup>. The OADM treats annotations as primary resources and provides a standard description mechanism for these annotations them between systems.

---

<sup>24</sup> Europe PMC: a full-text literature database for the life sciences and platform for innovation. Europe PMC Consortium. *Nucleic Acids Res* Volume 43 (2015) p.d1042-8

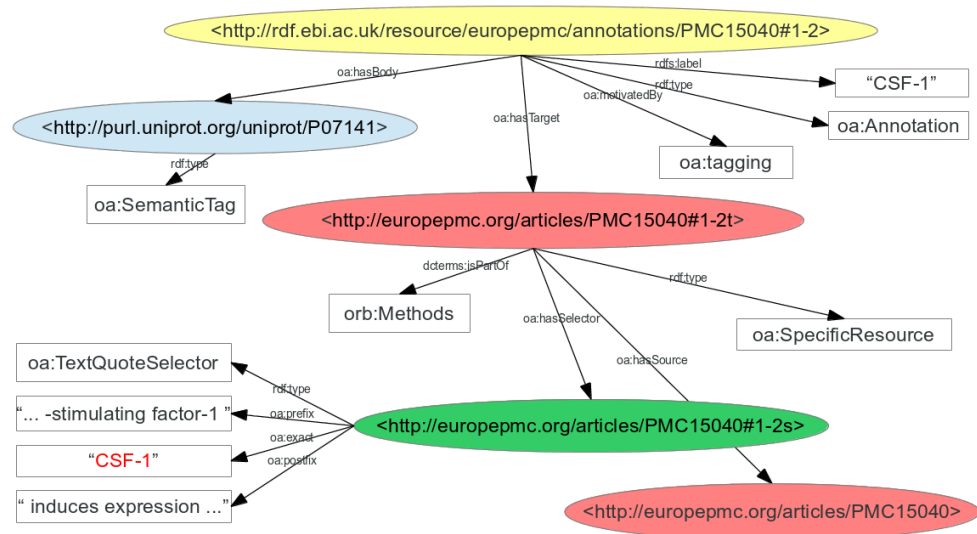
<sup>25</sup> Text processing through Web services: calling Whatizit. Rebholz-Schuhmann D, Arregui M, Gaudan S, Kirsch H, Jimeno A. *Bioinformatics*. 2008 Jan 15;24(2):296-8. Epub 2007 Nov 15

<sup>26</sup> Database citation in full text biomedical articles. Kafkas Ş, Kim JH, McEntyre JR. *PLoS One* Volume 8 (2013) p.e63184

<sup>27</sup> Section level search functionality in Europe PMC. Kafkas Ş, Pi X, Marinos N, Talo' F, Morrison A, McEntyre JR. *J Biomed Semantics* Volume 6 (2015) p.7

<sup>28</sup> Europe PMC open access articles <http://europepmc.org/ftp/archive/v.2014.09/oa/>

<sup>29</sup> <http://www.w3.org/TR/2014/WD-annotation-model-20141211/>



**Figure 9** RDF-ization of a UniProt protein name (CSF-1) mined from the full-text article PMC15040 with the OADM TextQuoteSelector, whose role is to specify the location of the mined text within its original context. The link from the protein name CSF-1 to the protein identifier (<http://purl.uniprot.org/uniprot/P07041>) is achieved by using the text-mining pipeline

Currently, our text-mining RDF service is running on a development server at EBI; it stores 1,563,241,810 triples text-mined from 400,746 Open Access articles in Europe PubMed Central. Modelling and provision of the dataset was funded by Europe PMC. We are still refining our text-mining pipeline and modelling the text-mined results with a better URI scheme, and we anticipate that the link will be advertised at the end of 2015.

One thing we learned from this process is, modelling text-mined results automatically produced in a large scale is a challenging task and requires careful thoughts on:

- capacity / stability of a RDF store,
- design of better URIs<sup>30</sup>, and
- interoperability within text-mining community.

### Genome-Wide Association Studies:

“One of the challenges for a successful GWA study in the future will be to apply the findings in a way that accelerates drug and diagnostics

<sup>30</sup> e.g. <http://europepmc.org/articles/PMC15040/methods/genes/CSF-1> instead of using hashing





development, including better integration of genetic studies into the drug-development process and a focus on the role of genetic variation in maintaining health as a blueprint for designing new drugs and diagnostics.”<sup>31</sup>

Until recently, the GWAS diagram<sup>32</sup> is driven by an RDF representation of the GWAS catalogue data which is run through an OWL reasoner. We discovered that this approach simply does not scale with large numbers of triples. We therefore changed the implementation to instead query the RDF using SPARQL over a virtuoso instance. Although some of the power of the OWL reasoner is lost when using SPARQL, this is offset by how much more readily the queries can scale. However, since the OWL reasoner is no longer being used, we are now exploring whether to use JSON-LD instead of RDF. Although not as powerful as RDF is, JSON-LD offers important advantages: 1) JSON-LD can be easily and cheaply generated 2) it does not require setup or maintenance of a triplestore. JSON-LD is not the right choice for all datasets (for instance where transitive graph-based queries are routinely required). However JSON-LD specifications are still young and evolving. We have flagged this technology as one for the Technology Watch work package (WP11) to follow.

---

<sup>31</sup> Iadonato SP; Katze MG (September 2009). "Genomics: Hepatitis C virus gets personal". *Nature* 461 (7262): 357–8. doi:10.1038/461357a. PMID 19759611. As referenced in [http://en.wikipedia.org/wiki/Genome-wide\\_association\\_study](http://en.wikipedia.org/wiki/Genome-wide_association_study)

<sup>32</sup> <http://www.ebi.ac.uk/fgpt/gwas/>



## **Appendix A.4 UDUS (ECRIN): Searching for clinical trials information and linking clinical trials to biosamples, drugs, genes, and publications**

### **Background**

For researcher in life sciences, cross-domain searches through different databases is often a time consuming and complicated process, because the databases have to be queried separately. Especially researchers interested in clinical trials and who want to design new studies, finding trial related information in different biomedical databases is an essential, but often tedious step of their research. The Clinical Trial Information Mediator (CTIM) was developed to support researcher in their searches. It was designed to address this problem by linking clinical trials information to corresponding publications and information about biosamples /genes.

In summary, it bridges the gap between different databases and is providing a solution that links research databases. Thereby it enables researchers to conduct a search from a unified front-end.

The ultimate aim of the CTIM is to enable the design of new research questions and of new clinical trials that provide more insight into the interplay of genes, drugs and adverse events on patients based on real clinical trials information and on suitable publications. The important aspect of CTIM is that it does the linking between clinical trials and publications not through an ID or a key word (code item), but through information content that provides the basis for the queries.

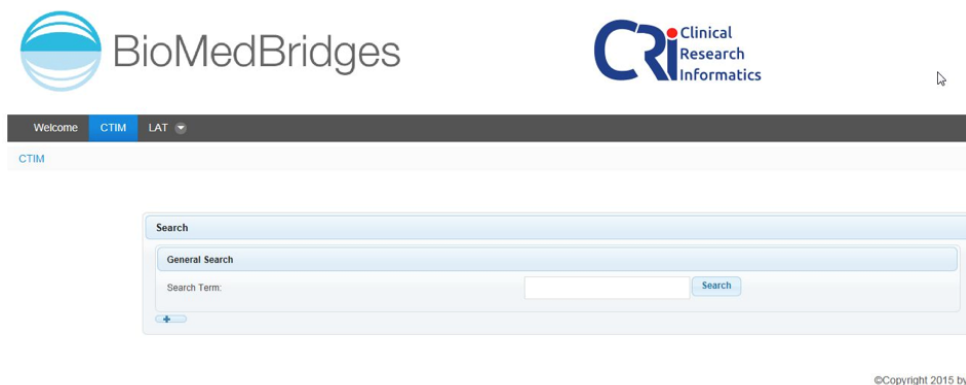
The knowledge base for the clinical trials is based on the CT.gov database, the largest repository of clinical trials in the world. In this way, CTIM opens clinical trials information to the biomedical researcher who doesn't have to search in CT.gov and PubMed separately. Trial registers like the CT.gov database are an important resource for research, physicians and even the general public. Registered trial information is a resource to make research available to a much wider audience. By searching a trial, it is possible avoid duplicating research and wasting valuable resources. Clinicians can use trials



information to find detailed and accurate data about trials involving new therapies, allowing them to make an informed choice about treatments.

### ***Desired features of CTIM***

- Provide concrete benefits to the user by enabling joint queries in different databases: The user of the tool can employ a unified front-end
- High degree of user-friendliness by providing a one-field search like Google and an expert search option (Figure 1). The tool raises awareness of the dependency of clinical trials registration and publication of clinical trials.
- The tool should be easier to use and the received results should be more relevant than the ones for separate searches in the databases
- The tool should allow the continuous updating of the knowledge base.
- The tool should be extensible so that new databases or repositories can be entered.



**Figure 10 Displayed is the simple user interface of CTIM with a single search field. As an option, expert search is also provided**

### **Scientific use case**

For usability testing CTIM is currently tailored to the WP8 use case of leukemia, in particular, Acute Myeloid Leukemia (AML). During testing research questions dealing with chemotherapy, specific drugs and availability of biosamples with specific mutations were employed. Specifically designed search fields for this use case enable scientists to identify only those clinical trials that may be relevant to solve his/her research question.



#### Use Case Examples:

- Find clinical trials with results involving drugs X or Y
- Find publications involving clinical trial Z
- Find bio samples involving clinical trial Z and mutation A

#### **Technical realisation**

As previously reported in D4.3, previous work covered an Apache Solr server that was installed and filled with data from clinicaltrials.gov. It uses the Lucene Java search library and features full-text search, near real-time indexing and database integration. Solr has REST-like HTTP/XML and JSON APIs.

In D4.5 we further developed CTIM as a portlet within Liferay Portal CE. The user interface was realised with the PrimeFaces Framework which is based on JavaServer Faces (JSF). The interfaces use programmatic web services provided by PubMed and BioSamples to get publications and biosample data.

#### **Work done for 4.6**

For 4.6 the data core for clinical trials got expanded so that now all trials from clinicaltrials.gov are involved. Furthermore the web services for BioSample and PubMed were implemented. Additionally further filtering possibilities were developed and a user friendly interface was created.

#### **RDF transformation**

The RDF data format is primarily build to genuinely identify contents / entities of electronic stored data. The data stored in a RDF-model is meta-data concerning this one identifiable entity, which can be any form of electronic stored data (e.g. webpage, user, locations, multimedia files, document files, biosample data, etc.).

A RDF about one biosample can include the information about the form the data of the biosample is stored (e.g. picture, chip-analysis or other forms experimental data), the biological origin, location where the biosample was taken and examined, when the biosample was taken and more.



RDFs can build “relationships” to other RDF-data, interlinking different data or referencing data between RDFs. For that reason a RDF can be described by already existing RDFs.

A SparQL Endpoint gives access to the data stored in RDF form. To query a SparQL-Endpoint it is necessary to implement all the referencing RDF language terms or vocabulary that was used describing the RDFs stored, which in case of the BioSamples RDF database are 12 different so called prefix libraries. Five of those prefix libraries are basic to the RDF data format, four are resource location libraries and three are for the description of the biosample data.

The query itself combines all the library terms to question for different properties of the BioSample RDF, resulting in a collection of URIs (uniform resource identifier) for the biosamples that match the properties in question.

Implementing those properties in the SparQL-query it is necessary to follow the relations of the describing RDFs. To implement, for example, a search parameter like “Homo Sapiens”, “Homo Sapiens” must be declared as a label for bio-characteristics and “organism” as the type for this bio-characteristic label. Further both type and label have to be linked via semantic science resource RDF-database for use in the search. This structure has to be declared a “derivedFrom” search via resource location libraries, which has to be declared as RDF label class by RDF basic format library. The result of this search is as previously mentioned a collection of uniform resource identifiers, which hold the links to the RDF form of the BioSample data. For any other information other than the URI, a query has to be built with the same structured definitions as described above.

With BioSamples as an example, the structure and means of RDFs and SparQL to explore RDF databases is not meant to serve as a researching tool rather than an identification tool. CTIM aims to gather information related to a certain topic, to provide an overview to the information available and the link to examine the resource. RDFs structure and means aims to identify resources with strict properties with little or no information about the resources format, composition or content thereof.



As CTIM is to offer basic information to web contents, like BioSamples, and the link to the source of that information, the web service of BioSample offers a better solution. Either programmatically or by direct web search, a simple request to the BioSamples web service provides information regarding the biosamples, which correlate to a simple search term, more than just the link to the resource. Therefore the web services of BioSamples and PubMed were used to feed CTIM searches for additional information regarding clinical trials.

### **Metrics**

No

### **Future work**

It is planned to improve the usability (display of results, listing according to data, status, etc.). Furthermore it is contemplated to integrate lexEVS (terminology server) into the query engine to enable search by terminology and with synonyms. Both of these improvements will be done in the context of the BioMedBridges project. In the end, it would be worthwhile to extend the data bridges with more significant databases, e.g. ArrayExpress, Genbank or DrugBank. An extension of data sources would be advantageous but out of scope for the remaining project period.



## Appendix A.5: RDF Training materials

<http://tinyurl.com/ebirdftraining2015>



## Appendix B: TranSMART pilots

### Appendix B.1 VUMC (EATRIS) Integrating Galaxy workflows into tranSMART

#### Background

Medical researchers use a lot of software to do their work. New tools come along all the time and it is often difficult to predict whether an investment in learning yet another tool will be worth it. Sometimes one tool can be used to access another tool, thereby lowering the mental load for new users.

#### Scientific use case

Both tranSMART and Galaxy can provide interesting functionality to medical researchers, but learning these systems is quite a burden for new users. The integration that we have created between tranSMART and Galaxy allows workflows that were thus far only usable in Galaxy to become available within tranSMART as well. Because users can already run R scripts in tranSMART, the Galaxy workflow functionality can be added quite naturally to the system. For a user it does not matter whether an analysis runs as an R script on the tranSMART server or a workflow on the Galaxy server.

#### Previous work

Our group had not done previous work in this specific area. We have built this integration using the work done by the Galaxy team (and specifically by John Chilton on the blend4j library) and the tranSMART foundation.

#### Work done for 4.6

The integration between tranSMART and Galaxy is built using the following components:

- tranSMART plugin<sup>33</sup>: this plugin handles Galaxy workflows in tranSMART and was written by Ruslan Forostianov (The Hyve)<sup>34</sup>;

---

<sup>33</sup> <https://github.com/thehyve/Rmodules/tree/features/transmart-galaxy>

<sup>34</sup> <http://thehyve.nl/portfolio/ruslan-forostianov/>





- workflow runner<sup>35</sup>: this component simplifies using the Galaxy API from Java and was written by Freek de Bruijn (VUmc)<sup>36</sup>;
- blend4j: this existing library provides access to the Galaxy API (and other systems) from Java and was written by John Chilton (Penn State University)<sup>37</sup>;
- Galaxy API<sup>38</sup>: the existing REST API that is part of Galaxy and enables developers to interact with Galaxy programmatically; it was written by members of the Galaxy Team (Penn State University and Johns Hopkins University)<sup>39</sup>.

### Technical implementation

The tranSMART plugin is written in Groovy (which is the programming language used for tranSMART and it works seamlessly with Java and other [JVM languages](#)<sup>40</sup>). The workflow runner and blend4j are both written in Java. The Galaxy API is written in Python, but since the other components communicate via HTTP and JSON with the API, the interface is language independent.

By making a group of small components, we have created building blocks that can be reused for other projects.

### Metrics

We have not collected any usage metrics yet; see future work below.

### Future work

The following work is currently planned:

One improvement that our users could ask for is adding the possibility to permanently identify and store the results of a Galaxy workflow run in tranSMART using FAIR principles.

---

<sup>35</sup> [https://github.com/CTMM-TraIT/traIT\\_workflow\\_runner](https://github.com/CTMM-TraIT/traIT_workflow_runner)

<sup>36</sup> <https://github.com/FreekDB>

<sup>37</sup> <https://wiki.galaxyproject.org/JohnChilton>

<sup>38</sup> <http://galaxy-dist.readthedocs.org/en/latest/lib/galaxy.webapps.galaxy.api.html>

<sup>39</sup> <https://wiki.galaxyproject.org/GalaxyTeam>

<sup>40</sup> [http://en.wikipedia.org/wiki/List\\_of\\_JVM\\_languages](http://en.wikipedia.org/wiki/List_of_JVM_languages)



In the current version, all input data has to come from tranSMART. We want to allow users to be able to select references to data outside of tranSMART to be used by a Galaxy workflow. These references could be based on EPIC PIDs (persistent identifiers)<sup>41</sup>. Once this is possible, we can support analyzing very large data files that have to be stored outside of tranSMART.

For each Galaxy workflow that is added to a tranSMART server, a small change has to be made to the code of the tranSMART plugin that handles the Galaxy workflows. We want to investigate ways to simplify this process, for example by using configuration instead of programming.

---

<sup>41</sup> <http://www.pidconsortium.eu/>



## Appendix B.2 VUMC (EATRIS): Integrating CDISC Operational Data Model into tranSMART i2b2

### Background

Electronic Data Capture (EDC) systems like OpenClinica and RedCap capture clinical data, like data about patients and clinical studies. They organize data in terms of events (like a doctor visit), Case Report Forms (CRFs, like questionnaires and health statistics), item groups (like medicine, brand, dose and unit) and items (like a measured value). The standardized format to export such data is CDISC's Operational Data Model (ODM), which is encoded in XML. However, EDCs are not designed to analyze the clinical data, nor to integrate it with other types of data, like genome data. Analysis platforms like i2b2 (informatics for integrating biology and the bedside) and tranSMART, which is based on i2b2, are specifically designed for analysis and integration of clinical data. Apart from these two platforms, also the generic statistical tool SPSS is still popular to analyze clinical data in the form of tabular files, like Excel files or tab-separated text files.

### Scientific use case

There is a need for exporting clinical data from EDCs towards a data format that is ready to be imported in analysis and integration platforms. The tabular format seems very suitable for this task, since it enables further transformation to i2b2, tranSMART and SPSS. An R-script that transforms and loads tabular files into the i2b2 database tables below tranSMART does exist already. For this reason a direct transformation from the ODM format to a tabular format that can be uploaded in tranSMART is very desirable.

### Previous work

No work for this pilot has been previously reported in BioMedBridges. The ODM-to-i2b2 Java conversion tool is a fork of the RedCap-to-i2b2 project<sup>42</sup>, which loads ODM files directly in i2b2 database tables. The RedCap-to-i2b2 project transforms the XSD description of the ODM/XML format into automatically generated Java classes. ODM-to-i2b2 is also dependent on the

---

<sup>42</sup> <https://community.i2b2.org/wiki/display/ODM2i2b2/Home>



availability of data in EDCs. It is the next Extract-Transform-Load (ETL) step for the OCDataImporter tool<sup>43</sup>.

### **Work done for 4.6**

ODM-to-i2b2 builds upon the Java classes that were automatically generated from the XSD description of ODM. It has copied and modified the Java class of RedCap-to-i2b2 that crawls through the ODM file in a systematic manner. Extra Java classes were built to export to tabular files. A series of functionality improvements were made, such as creating a tree-structure, choosing human readable names, handling different studies, translating embedded HTML code, adding configuration abilities for special characters and the maximal length of entries, and designing and implementing a suitable format to write repeated measurements of data into the tabular format.

### **Technical implementation**

The project uses Java SE (Standard Edition) 7 and Maven 3. It was developed and tested in IntelliJ on Windows and tested on Linux/Unix. The conversion tool only executes file operations. It parses the input ODM/XML file and returns output as three tabular text files. The main tabular file is the clinical data file, which contains all the clinical data. In the absence of repeated measurements of the same data, each row represents the data of one patient. The columns represent items of data about the patient, like age, gender, weight and questions. Repeated measurements of the same data are written in different rows. The values of five of the seven first columns form a key together that identify a data observation through patient, event type, event number, item group type and item group number. The other two columns provide human readable names for event types and item group types. The second tabular file, the columns file, provides meta-data in the form of a tree-structure about the columns in the clinical data file. The third tabular file is the wordmap file, which maintains a mapping between human readable data values and natural numbers, like 1 for yes and 2 for no. The natural numbers are used in the clinical data file to decrease the size of the data. Data values that have such a mapping use categorical values, which are treated different from numerical values in the analysis tools.

---

<sup>43</sup> <https://community.openclinica.com/extension/ocdataimporter>

**Metrics**

Five clinical studies, or examples of clinical studies, have been converted and loaded in a test instance of tranSMART. A usability test is scheduled in June 2015.

**Future work**

Support for ontology-annotated clinical data is planned as future work. The URI of an ontology like SNOMED CT or an ontology from the OBO Foundry would be written down in the columns file, thereby clarifying the meaning of the columns in the clinical data file. Since ontology URIs are not yet supported by EDC systems like OpenClinica, the URIs would be mapped to CRFs in some library that is maintained outside the EDC.



## Appendix B.3 ErasmusMC (Euro-Biolmaging): Centralized correlative analysis between image-derived data and other clinical data.

### Background

Medical imaging (MRI, CT, Ultrasound) is becoming a more integral part of multi-center clinical studies. Quantitative analysis of image-derived biomarkers can be useful in diagnosing individuals and in studying patient populations; it can also be used as a surrogate endpoint for clinical trials themselves. Often, imaging biomarkers are analyzed in relation with other clinical data such as disease status, genetics or age, in order to provide more accurate diagnoses or to verify hypotheses. Therefore, what is needed is a user-friendly data infrastructure to support centralized correlative analysis between image-derived data (e.g. organ volume measurements) and clinical data.

### Scientific use case

We are closely collaborating with the CTMM TraIT project (EATRIS/BBMRI) (<http://www.ctmm-trait.nl>), which provides an IT infrastructure that facilitates the collection, storage, analysis, and archiving of data generated in biomedical research projects, with a particular focus on the needs of translational, multidisciplinary research in multi-center settings. Medical imaging data plays an important role in many of these projects. Our work in WP4 aims at extending the TraIT infrastructure to better support centralized statistical analysis of imaging biomarkers in relation with other research data.

For D4.6 specifically, we have selected the study described in Guyader et al<sup>44</sup> as a very concrete test case to guide the development. The study by Guyader et al was originally performed in the context of the Quic-Concept project ([www.quic-concept.eu](http://www.quic-concept.eu)). In 5 volunteers, diffusion-weighted magnetic resonance images (MRI) were collected at two time points. From these

---

<sup>44</sup> J.-M. Guyader, L. Bernardin, N.H.M. Douglas, D.H.J. Poot, W.J. Niessen and S. Klein, Influence of image registration on apparent diffusion coefficient images computed from free-breathing diffusion MR images of the abdomen, *Journal of Magnetic Resonance Imaging*, <http://www.ncbi.nlm.nih.gov/pubmed/25407766>, in press.



images, we computed apparent diffusion coefficients (ADC)<sup>45</sup> for regions of interest in the abdomen and investigated the reproducibility using various image processing schemes. The aforementioned infrastructure should be able to visualize these results in charts in a similar way as depicted in the paper. For D4.6, the specific part of this use case that we focussed on was the storage and analysis of the ADC imaging biomarkers.

### **Previous work**

As previously described in Deliverable D4.5, we started by installing XNAT (<https://bigr-xnat.erasmusmc.nl>), a platform for sharing medical imaging data. Beside imaging data, XNAT can also store image-derived analysis results. For example, the volume of white matter in the brain can be stored for each patient scan, or like in Guyader et al, the ADC values in regions of interest. XNAT has a programmatic interface (REST API) which enables us to retrieve these results from other applications. We build on previous WP4 work as follows.

### **Work done for 4.6**

#### **Technical implementation**

As presented at the BioMedBridges AGM, but not previously reported, we chose the tranSMART (<http://transmartfoundation.org>) data integration and browsing platform as the platform of choice for central correlative analysis. tranSMART is a key informatics platform within the CTMM-TRAIT project, the Innovative Medicines Initiative, and others.

We created a new open-source tranSMART plugin to import clinical image-derived data from XNAT to tranSMART. By storing image-derived data in tranSMART, its relation with other medical data can be further analyzed. It should be noted that we do *not* aim to import the original images into tranSMART. We only focus on *image-derived* biomarkers, such as organ volumes or mean ADC values for regions of interest, as only these quantitative measurements will be used for statistical analysis. Figure 11 shows an example of a current tranSMART project that includes image-derived data from XNAT. Note also the “Goto XNAT” link here, which brings the user

---

<sup>45</sup> The ADC is a measure of the magnitude of diffusion (of water molecules) within tissue (<http://radiopaedia.org/articles/apparent-diffusion-coefficient-1>).

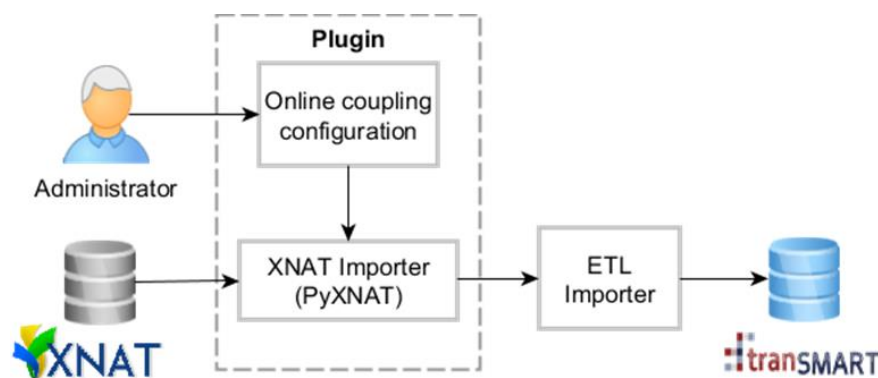


directly to a page in the XNAT system where the original images can be inspected, if desired.

Subject	Patient	Subset	Trial	Age	roi1_registration_region...	roi2_no_processing_re...	Im
1000384743	BIGRXNAT_S01628	subset1	VOLUNTEERSQC37	62	0.02828	0.018	Go
1000384744	BIGRXNAT_S01627	subset1	VOLUNTEERSQC37	63	0.0495	0.013	Go
1000384745	BIGRXNAT_S01624	subset1	VOLUNTEERSQC37	30	0.09192	0.009	Go
1000384746	BIGRXNAT_S01625	subset1	VOLUNTEERSQC37	35	0.01414	0.012	Go
1000384747	BIGRXNAT_S01626	subset1	VOLUNTEERSQC37	64	0.02828	0.007	Go

**Figure 11 Screenshot of a tranSMART project with image-derived data imported from XNAT. The column “roi1\_registration\_region...” contains the value of an imaging biomarker computed in “region of interest 1”**

A schematic overview of the tranSMART-XNAT linking mechanism is shown below. First, to configure which XNAT image-derived data is imported in tranSMART, an administrator should create a “coupling configuration” which defines a mapping between the XNAT data structure and the tranSMART data structure. Then, the administrator can trigger the import process, upon which the image-derived data is retrieved from XNAT via its REST API, which is implemented using Pyxnat (a Python library that simplifies the use of XNAT REST API calls). The plugin subsequently converts the data obtained from XNAT to tranSMART’s data format, and uses the ETL<sup>46</sup> importing system to import the data.



**Figure 12 schematic overview of the tranSMART-XNAT link**

<sup>46</sup> The ETL importer (<https://wiki.transmartfoundation.org/display/TSMTGPL/Data+ETL>) for Clinical Data is used.

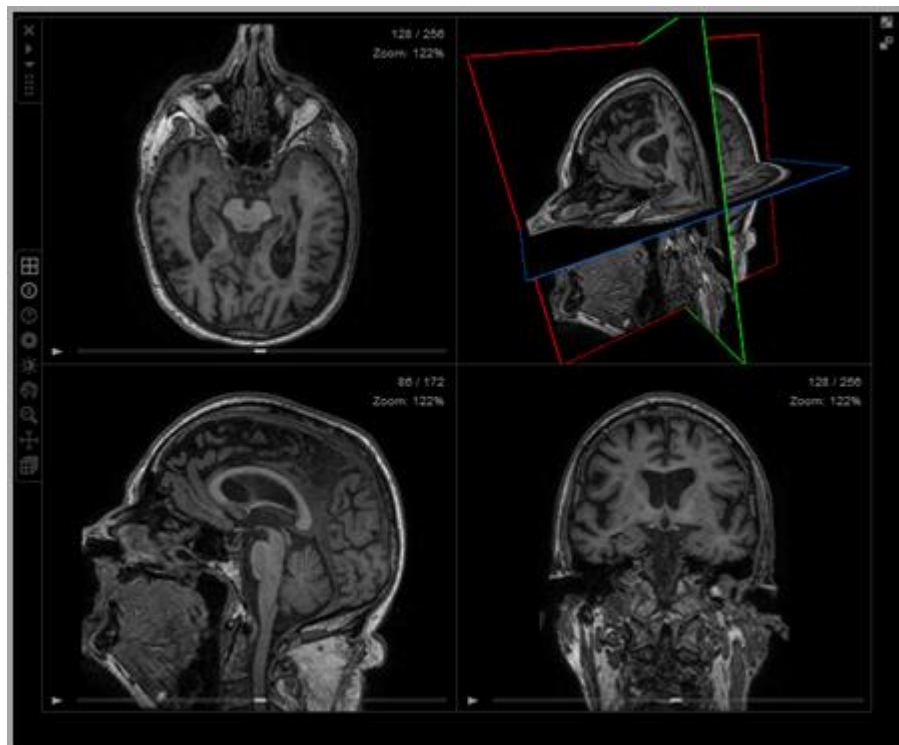




The plugin is setup such that after installation, the configuration and import process can all be managed from the tranSMART web interface, using a new administration panel that is added by the plugin. The plugin therefore greatly streamlines the import and conversion of image-derived data from XNAT to tranSMART.

The plugin source code is available on GitHub<sup>47</sup>. We have made available both a user guide for data managers and technical documentation for developers<sup>48</sup>.

Besides developing the tranSMART plugin, we updated our XNAT installation to the latest version (release 1.6.4), installed a new XNAT image viewer which greatly enhances the user-friendliness (see Figure 13 below), and we have prepared a formal user agreement. To support administrators, we have updated and streamlined the Puppet configuration scripts<sup>49</sup>, and made these available as a new release on the XNAT marketplace<sup>50</sup>.



**Figure 13 Three dimensional viewing of brain MRI scan using web-based XNAT viewer**

<sup>47</sup> on <https://github.com/evast/transmart-xnat-importer-plugin>

<sup>48</sup> <https://github.com/evast/transmart-xnat-importer-plugin/blob/master/docs/>

<sup>49</sup> [https://bitbucket.org/bigr\\_erasmusmc/puppet-xnat](https://bitbucket.org/bigr_erasmusmc/puppet-xnat)

<sup>50</sup> <http://marketplace.xnat.org/>



## Metrics

The plugin we built for 4.6 will actively be put to use now within CTMM-TRAIT biomarker projects. We already automatically store the amount of user-logins on XNAT. Furthermore, we inspect the number of users, projects, patients, imaging sessions and the data size of each project.

## Future work and sustainability

In order to standardize the storage of image-derived biomarkers, which will further simplify the transfer between XNAT and tranSMART and facilitate querying across databases, we plan to create a new imaging biomarker ontology. Such an ontology would help to manage the endless variety of biomarkers that can be derived from images. Biomarkers are context-specific, depending on anatomical regions of interest, measures of interest, and the software tool used to compute them. Each imaging biomarker will be represented in the ontology as an object, which includes a description of its meaning, a protocol how to compute it (for example a link to the software tool), the physical unit, possible relations to other biomarkers, etc. Given the endless variety of imaging biomarkers that have been described in the literature, the ontology is not meant to be exhaustive; only imaging biomarkers that are actually used in clinical/population studies need to be included. The ontology of imaging biomarkers will be hosted at <http://BioPortal.bioontology.org>. Subsequently, we will create a standardized data type in XNAT that links in an RDF fashion to the online ontology, in order to annotate the meaning of the image-derived values stored in XNAT based on the information stored in the ontology. The ontology will contain references to the UBERON anatomy ontology to indicate the anatomical region of interest the biomarker is related to. This would in turn make it possible to perform powerful searches that leverage relationships such as “X is part of Y”. For instance, a volume measurement tagged “hippocampus” would be among the search results whether a user queried for “brain”, or for “mediotemporal lobe”. We intend to have a first prototype of the online imaging biomarker ontology, including a limited number of often used imaging biomarkers, at the end of 2015.



In collaboration with the CTMM TraIT project<sup>51</sup> (EATRIS/BBMRI) we are currently integrating our tranSMART-XNAT module into the TraIT infrastructure. In this way, we hope to facilitate central correlative analysis between image-derived data and clinical data in many multi-center studies.

---

<sup>51</sup> <http://www.ctmm-trait.nl>



## Appendix C: BioJS Widget pilots

### Appendix C.1 STFC (INSTRUCT): Clinical Consequences of Protein Structure variation CCoPS

#### Background

It is a routine research technique to make knockout mice to investigate a gene of interest. Clinical genetics databases provide information about a natural experiment, observation of knockout humans. They also provide an extra level of detail, on the clinical consequences of SNPs. Because of the interdisciplinary nature of this approach, this information is underexploited by structural biologists and drug developers.

#### Scientific use case

The clinical consequences of SNPs are a probe of the relationship between structure and function. This information is now accessible to structural biologists as annotation on a visualization of protein structure.

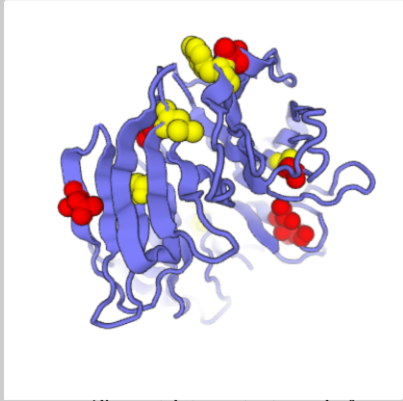
#### Previous work

STFC's contribution to prior WP4 deliverables centred around the Protein Information Management System (PiMS). Unfortunately, core funding for PiMS ended in March 2015 and applications for renewal were unsuccessful. The code itself is in GitHub and there are several active users/contributors. However, the lack of continued funding for PiMS prompted us to pursue a platform-independent offering for D4.6 as it was more likely to be sustainable and impactful in the longer term. For details regarding the RDF transformation previously done for PiMS, please see the deliverable report for D4.3.



**Consequence of Variation of a Protein**

PDB ID code: 3eu7 [Get Structure](#)



Alignments between structure and reference sequences

Chain	Structure Start	Structure End	Protein	Protein Start	Protein End
A	5	356	PALB2_HUMAN	835	1186
X	1	19	BRCA2_HUMAN	21	39

▼ Details

Clinical variation of PALB2\_HUMAN [ClinVar](#) [SwissVar](#)

**Progress**

- Structure Displayed
- PALB2 Checking 619 variants ...
- BRCA2 Checking 4291 variants ...
- decorated: A:Pathogenic atoms: 0
- decorated: A:Benign atoms: 0
- decorated: A:Mixed atoms: 0
- Found 325 variant residues for gene: PALB2\_HUMAN
- decorated: X:Pathogenic atoms: 43
- decorated: X:Benign atoms: 0
- decorated: X:Mixed atoms: 80
- Found 2561 variant residues for gene: BRCA2\_HUMAN

**Figure 14 Screenshot of the Clinical Consequences of Protein Sequence variation (CCoPS).** Pathogenic mutations are displayed in red; residues of unknown or mixed impact are rendered in yellow. The protein itself can be visually rotated in space and zoomed

## Work done for 4.6

### Technical implementation

The implementation is a web page, which integrates two web services:

- The reference structure is provided by PDB
- Clinical variations are provided by NIH (ClinVar)

A Javascript program running in the user's web browser fetches and integrates these data. It then implements a [PV, BioJS widget developed at SwissProt](#) to visualize the protein structure and overlay the residues of interest. This illustrates the fact that the spread of web services using standard interfaces and BioJS widgets makes it easier to implement new composed services: "bridges". On the other hand, we encountered frequent challenges due to unreliable or unsuitable services (see overall lessons learned, section 3.2.1.3 Integration of simple object queries using BioJS widgets).



## **Metrics**

Recently announced, no metrics yet. Visits to the page are logged, and will be analysed. Two rounds of user testing and feedback were part of the development process for this tool.

## **Future work**

Dissemination and training in this approach can begin to crowdsource novel bioinformatics services. This tool will become part of the Structural Biology Work Bench to be developed by the West-Life project. In future iterations we will incorporate a newly available [PDBe REST webservice](#) that will enable proteins to be queried by parameters other than their exact PDB ID, for instance Gene Ontology terms and Taxonomic terms. To this end, we will explore with users whether the PLATO widget (below) this would be useful in this context.



## Appendix C.2 EMBL-EBI (ELIXIR) PLATO widget

### Background

Ontologies can play a fundamental role in the organisation, retrieval, and integration of data; accordingly, they feature prominently within the BioMedBridges work and indeed within the biomedical community more broadly: dozens of institutions have expressed interest in such a widget. Visualising ontologies and locating terms (e.g. within a query interface) is a common problem that benefits from being solved in a general way. Currently, groups interested in incorporating ontology visualisation into their web applications must essentially start over since existing tools are not easily configured, scalable, or benchmarked. Furthermore, existing tools typically rely on local copies of ontology files and can easily get out of sync with their live ontology counterparts. To address this general challenge, an embeddable javascript widget and an accompanying ontology REST service backend were therefore developed using an API from the Ontology Lookup Service at the EBI<sup>52</sup>.

### Scientific use case

Although many potential applications for this widget exist, three specific scientific use cases drove the development of the widget. They are summarized below and described in more detail in deliverable 4.5 report:

#### ***CMPO (WP6)***

The cellular microscopy phenotype ontology is a purpose-built ontology for integration of phenotypes generated for image data (WP6). Annotating images easily using ontology terms is important to the integrity of the data.

#### ***PhenoBridge (WP7)***

PhenoBridge aims to deliver a semantic 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.

---

<sup>52</sup> [www.ebi.ac.uk/ols](http://www.ebi.ac.uk/ols)



### ***BioMedBridges tools and data registry (WP3)***

The Tools and data registry was developed to aid the discovery, comparison, and selection of tools and services. To browse the EDAM ontology and search for matching tools, a widget is needed.

#### **Technical Implementation**

##### ***Back end: Ontology Lookup Service***

The Ontology Lookup Service (OLS) provides, among other things, a web service interface to query multiple ontologies from a single location with a unified output format. The OLS supports any ontology available in the Open Biomedical Ontology (OBO) format; thus it provided a natural starting point for the backend required by the ontology viewer widget. As the ontology community moves to adopt the W3C Web Ontology Language (OWL) format, the backend for OLS has been rebuilt to support both OWL and OBO ontologies. The web services have also been redeveloped to replace the old SOAP/XML API in favour of a modern REST/JSON API that will better support developer access to the OLS services. A specification for a Minimum Information for Accessing Ontologies (MIAO)<sup>53</sup>, **Appendix C.2.1** was developed that described how an ontology should be accessed. MIAO is currently being aligned with similar efforts from the the Gene Ontology Consortium and OBO community to provide a standard that could be used by ontology registries like OLS, BioPortal, and BioSharing. The MIAO specification is currently in draft form and anticipated to be published later this year.

The specific BioMedBridges 4.6 contribution to the OLS back-end development was to create a custom [CORS](#)-enabled REST API over the Solr-Lucene JSON service. This additional layer exposes the JSON content in a way that javascript widgets like PLATO and webpages can consume it in accordance with modern best practices for the web.

##### ***Front end: Javascript BioJS ontology viewer widget***

Our requirements analysis identified seven desirable features for a widget:

---

<sup>53</sup> <http://tinyurl.com/miaospecification>





- 1) Autocompletion on ontology terms
- 2) Autocompletion on a configurable set of free-text terms
- 3) Performant centralised query interface (via webservice)
- 4) Visualisation of matching terms within their immediate tree context
- 5) Ability to expand / collapse tree nodes
- 6) Subsumption queries
- 7) Configurability with any ontology and dataset
- 8) Highlight of results as search term, child term or synonym

Existing open source applications in this space were reviewed and found to offer only a subset of the above features, or performance of the features was poor. To speed the development process we identified one open source application to modify and extend into a generic and re-usable BioJS widget. First, the widget was adapted to accept JSON served up by the new OLS web service described above. The first-generation widget required hard-coded modifications to a jquery library. Since the first generation widget, additional libraries have come along that are better supported and more feature-rich.

### **Future work**

The first-generation PLATO widget was reported in deliverable 4.5. This widget, is undergoing testing and feedback by BioMedBridges partners and other user groups via its three intra- and extra-project deployments described above.

Prioritization of the features has changed in response to user feedback from the prototypes: for instance, the immediate tree context of a term was found to be underutilized and a bit confusing to novice users when embedded directly within the autocomplete context. We have therefore re-developed the widget in collaboration with the Centre for Therapeutic Target Validation (CTTV)<sup>54</sup>. They have developed a widget to display detail about an ontology term (e.g. definition, synonyms etc.). Future work for D4.8 will combine the PLATO widget with the CTTV's Ontology Term Overview Widget.

The original version of the code is now in the open-source BioJS project. The anticipated broad use of the widget is likely to spur more community

---

<sup>54</sup> <http://www.targetvalidation.org/>



contributions to the code, thereby making it even more extensible, robust, and feature-rich with time.



## Appendix C.2.1 Minimum Information for Accessing an Ontology (MIAO)

[tinyurl.com/miaospecification](http://tinyurl.com/miaospecification)

Field	Description	Example value	Example value type	Required ?	Possible predicates	...	BioPortal OMV
id	Unique id for the ontology, typically the Ontology URI		URI	Required	dcterms:identifier		
title	Name of the ontology	Gene Ontology	Literal	Required	dcterms:title		
namespace	Short name or the ontology	GO	Literal	Required	idot:preferredPrefix		
description	Short description of the ontology	The Gene Ontology is a....	Literal	MAY	dce:description		
ontology physical location	Physical location on web from which the file can be downloaded	<a href="http://www.ebi.ac.uk/efo.owl">http://www.ebi.ac.uk/efo.owl</a>	URI	Required	dcat:accessURL	dcat:downloadURL	<a href="http://omv.ontoware.org/2005/05/ontology#Location">http://omv.ontoware.org/2005/05/ontology#Location</a>
license	Any license associated with ontology		URI	MAY	dct:license		<a href="http://omv.ontoware.org/2005/05/ontology#LicenseModel">http://omv.ontoware.org/2005/05/ontology#LicenseModel</a>
homepage	Project home page		URI	MAY	foaf:page		
mailing list	URL to mailing list		URI	MAY	doap:mailing-list		
contact	List of contact names and e-mail	Foo Bar <foo@bar.com>	Literal	Required			<a href="http://omv.ontoware.org/2005/05/ontology#Person">http://omv.ontoware.org/2005/05/ontology#Person</a>
ORCIDs	List of ORCID ids for developers		URI	MAY	dcterms:creator		
citation	Citation for this ontology	Bar et al (2010)	Literal	MAY	dcterms:bibliographicCitation		
publication	List of publication URLs	DOI or PubMed URLs	URI	MAY			
depiction	Link to ontology logo		URI	MAY	foaf:logo		
issue tracker	URL for tracker system		URI	MAY			
keywords	List of keywords to describe the ontology	anatomy, disease	Literal	MAY	dcat:keyword		<a href="http://omv.ontoware.org/2005/05/ontology#OntologyDomain">http://omv.ontoware.org/2005/05/ontology#OntologyDomain</a>
taxon	Taxon ids if ontology is restricted	taxon { id : "http://NCBITaxon_33208", label : "Metazoa" }	Literal	MAY			
version	Release name or version	"1.1"	Literal	MAY	owl:versionInfo	pav:version	
preferred label predicate	Primary predicate for label annotation	rdfs:label or skos:prefLabel	URI	MAY			



Field	Description	Example value	Example value type	Required?	Possible predicates	...	BioPortal OMV
textual definition predicate	Primary predicate for textual definition/description annotation	dc:description	URI	MAY			
synonym predicates	List of predicates for synonyms	obo:exact, skos:altLabel	URI	MAY			
is inferred	Does this ontology include inferred axioms	true/false	xsd:boolean	MAY			
OBO slims	Does this ontology contain OBO style slims	true/false	xsd:boolean	MAY			
hierarchical properties	Relations in the ontology that can be used to create a tree view	part_of	URI	MAY			
base URI	Base URI for terms in the ontology	<a href="http://www.ebi.ac.uk/efo/">http://www.ebi.ac.uk/efo/</a>		MAY			
dependencies/imports	List of ontologies and versions where terms are imported	imports: {id: "http..", version: "1.0", url:"http..." }		MAY			
contributor	Person(s) contributing to developing the ontology						
hidden properties	Any predicates (annotation, object or data) that should be ignored			MAY			
needs_classifying	Flag to indicate if the ontology needs to be classified to infer subsumption relations	true/false	xsd:boolean	MAY			
expressivity properties	If classification is required should we use a DL or EL reasoner	{EL, OWL2}					
	All relationships						