BioMedBridges

Deliverable D4.8

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:	Report on Web Services based integration of BioMedBridges integration across all appropriate services
WP No.	4
Lead Beneficiary:	1: EMBL
WP Title	Technical Integration
Contractual delivery date:	31 December 2015
Actual delivery date:	21 December 2015
WP leader:	Ewan Birney and Helen Parkinson
Contributing partner(s):	1: EMBL

Partner	Infrastructure	Author (* = institutional technical lead)	Author email
EMBL	ELIXIR	Nick Juty* Julie McMurry Simon Jupp Tony Burdett Andy Jenkinson Helen Parkinson* Jon Chambers	juty@ebi.ac.uk jmcmurry@ebi.ac.uk jupp@ebi.ac.uk tburdett@ebi.ac.uk andy.jenkinson@ebi.ac.uk parkinson@ebi.ac.uk chambers@ebi.ac.uk
STFC	Instruct	Chris Morris* Martyn Winn	Chris.Morris@stfc.ac.uk martyn.winn@stfc.ac.uk
HMGU	Infrafrontier	Philipp Gormanns* Elida Schneltzer	philipp.gormanns@helmholtz-muenchen.de schneltzer@helmholtz-muenchen.de
TUM-MED	BBMRI	Raffael Bild	raffael.bild@tum.de
UDUS	ECRIN	Christian Krauth* Wolfgang Kuchinke	christian.krauth@med.uni-duesseldorf.de wolfgang.kuchinke@uni-duesseldorf.de
VUmc	EATRIS	Freek de Bruijn* Ward Blondé Jeroen Beliën	f.d.bruijn@nki.nl w.blonde@nki.nl jam.belien@vumc.nl
Erasmus MC	Eurobioimaging	Stefan Klein* Erwin Vast	s.klein@erasmusmc.nl e.vast@erasmusmc.nl
UMCG	BBMRI	Dennis Hendriksen Bart Charbon David van Enckevort Morris Swertz	d.hendriksen@umcg.nl bart.charbon@gmail.com d.j.van.enckevort@umcg.nl m.a.swertz@gmail.com

Contents

1 Executive Summary

European e-Infrastructure projects are increasingly turning to Semantic Web[1] (SemWeb) 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 three stages are reported in separate deliverables. This deliverable D4.8 reports on the strategy, implementation and lessons learned for the semantic web pilots for BioMedBridges.

Table A Overview of 3-Phased Semantic Web Pilot

	Del #	Due	Deliverable focus	Partners funded for deliverable	Nature of the activities
Prep	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.
Phase I	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.
Phase 2	D4.6	June 2015	Data integration	ErasmusMC (EuroBioimaging), EMBL-EBI (ELIXIR), HMGU (Infrafrontier), TUM-MED (BBMRI), STFC (Instruct), UDUS (ECRIN), VUmc (EATRIS) UMCG(BBMRI)	BMS-Research Infrastructures implement individual pilot projects in parallel, according to their respective roadmaps, using the technical guidelines and outcomes from phase I.

[1] http://www.w3.org/standards/semanticweb/

Phase 3	D4.8	Dec 2015	Data integration	ErasmusMC (EuroBioimaging), EMBL-EBI (ELIXIR), HMGU (Infrafrontier), TUM-MED (BBMRI), STFC (Instruct), UDUS (ECRIN), VUmc (EATRIS) UMCG(BBMRI)	Report on progress, sustainability and lessons learned in the semantic web pilot

By following this schedule and aligning partner-specific roadmaps to the blueprint delivered in Phase I (D4.4), the pilot projects were developed synchronously, allowing knowledge to be shared efficiently (Phases 2-3: D4.6-4.8). This enabled infrastructures to collaborate effectively and address common issues as they arose. In support of this effort, a knowledge-exchange workshop ran on the 29-30 April 2014 at TMF, Berlin, Germany. The programme and course materials are available at BioMedBridges website[2] and can be found in D4.7 Appendix 2 ('Resources'), item 3 ('SWAT4LS SemWeb training materials'). In December 2013 and December 2014, at SWAT4LS[3] and in May 2015 at an industry workshop, tutorials were delivered demonstrating the queries and analyses the RDF platform makes possible. A training course at the School of Computer Science, University of Manchester in December 2014 delivered a summary of the practicalities of working with and running an RDF platform, which summarised the technological approach and technology experience.

All of the SemWeb pilot work was informed by the work of WP3, with respect to the choice and use of ontologies, as well as provision and re-use of identifiers, reflecting the application of standards derived from the use case work packages (e.g. WP7 and WP10).

BioMedBridges played a major role at the CRI (Clinical Research Informatics) Day organised by ECRIN: "First CRI Solutions Day" (26-27 May 2014 at Heinrich-Heine University, Düsseldorf, Germany). The presentation about BioMedBridges (S. Suhr) was accompanied by presentations of software tools

[2] http://www.biomedbridges.eu/trainings/knowledge-exchange-workshop-resource-description-framework-rdf

[3] http://www.swat4ls.org/workshops/berlin2014/

and hands-on sessions with their developers. Tools developed or employed in BioMedBridges (BBMRI Catalogue, CTIM, tranSMART, MOLGENIS / BiobankConnect, XNAT) and the approaches for data sharing of BioMedBridges could be compared with those of many other Research Infrastructures and EU projects, such as EATRIS, BBMRI, ECRIN, BioSHaRE, TRANSFoRm, EHR4CR, and p-medicine.

The first impression of participants at the CRI Solutions Day was that EU projects often seem to cope with similar problems and have developed similar solutions. For example, all domains struggle with the same rigid conditions for data protection and the challenge of semantic interoperability. It was suggested that it might be better to bring projects together and to work jointly on software solutions to common problems. Especially in the area of semantic interoperability and legally compliant data sharing, BioMedBridges has developed generic solutions that may be of help for other projects.

A workshop on translational research infrastructure including tools like XNAT, tranSMART, MOLGENIS, OpenClinica and Galaxy and the interfaces between them, co-organised by Dutch representatives of BBMRI, EuroBioImaging, and EATRIS, was held at the OpenBridges symposium (see section 3.2.1.2, Pilots 1 through 3). Materials such as presentations, documentation on discussions relating to the OpenClinica-tranSMART and tranSMART-Galaxy connections, as well as XNAT at the Dutch DTL Programmers Meeting are available elsewhere[4].

Various other outreach and dissemination activities have taken place during the course of BioMedBridges WP4, and are detailed within this report, in the relevant sections.

While many of the most refined pilots (below) are centred on the use of RDF, it is not the only appropriate solution for data integration; our experience indicated that certain kinds of data are better suited to different distribution and integration mechanisms. Therefore, this report covers an array of solutions that achieve integration, including RDF-based, as well as those that provide integration

[4] https://wiki.dtls.nl/index.php/DTL_Programmers_Meeting

points for future efforts; D4.6 included both RDF-based and pilots using alternative methodologies, D4.7 was targeted specifically at judging the suitability and scalability of a purely RDF-centric data integration solution, with other deliverables providing various interfaces to data (REST, widgets, GUIs). Here we summarize our processes, products, and lessons learnt during the execution of this work package. We also describe the further work that has been conducted in the period intervening D4.6 and this report (D4.8), as well as detailing the sustainability of these pilot activities, and lessons learnt throughout the entirety of WP4.

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	
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

Many BioMedical Science (BMS) Research Infrastructures (RIs)5 and by extension the domains in which they operate, are impeded by issues around, data protection, data accessibility, and the lack of standardisation sufficient to enable of semantic interoperability. These common threads can be seen as an opportunity to bring projects together to work jointly, across infrastructures and

[5] http://ec.europa.eu/research/infrastructures/index_en.cfm?pg=landscape

domains, on software, standard operating protocols and infrastructure based solutions to shared issues.

BioMedBridges WP4 was tasked with implementing such solutions for RIs in the Life Sciences, and has particularly focused on semantic interoperability and improving the sharing of legally compliant data. In addition, BioMedBridges has developed many facilitating and generic solutions which provide ‘hooks’ into data in a standardised way, which may be built upon by existing or future workflows.

While many of the most refined pilots (below) are centred on the use of RDF, it is not the only appropriate solution for data integration; our experiences indicate that for each type of data (sequence based, image based, etc.) different distribution and integration mechanisms must be considered, and that each data type will usually have a particular type of solution. This report therefore summarises a set of solutions that achieve integration, covering a disparate set of data types, whose solutions include more than RDF-based ones. Many of the pilots also provide integration points for future efforts; D4.6 included both RDF-based and pilots using alternative methodologies, D4.7 was targeted specifically at judging the suitability and scalability of a purely RDF-centric data integration solution, with other deliverables providing various interfaces to data (REST, widgets, GUIs).

Here we summarize our processes, products, and lessons learnt during the execution of this work package. This deliverable (D4.8) summarises and extends the information from all previously reported pilot projects (D4.3 through D4.7), 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. It additionally describes updates for each pilot subsequent to D4.6, as well as reporting lessons learnt in WP4 and where available the sustainability of each pilot beyond BioMedBridges. We also describe the further work that has been conducted in the period intervening D4.6 and this report (D4.8), as well as detailing the sustainability of these pilot activities.

3.2 Integration pilots and activities

We report a variety of activities and pilots conducted during the course of the BioMedBridges programme, covering a diverse range of knowledge domains, and accomplished through a varied but informed choice of technologies.

The works detailed below are organised by the axis of integration chosen (Figure 1) and the integration level achieved; the pilots in section 3.2.1 (RDF and tranSMART) achieve specific integration targets in response to defined use cases, the REST pilots are described in 3.2.2 and provide the means to achieve integration through a programmatic interface, while those in 3.2.3 provide visualisation facilities, in particular for users, to view or access data. Sections 3.3 and 3.4 discuss sustainability and lessons learnt during the work package.

Axes of integration: These pilot projects provide complementary resources that are not all currently amenable to a single integrated query as no use case was identified that crossed all resources. Accordingly, three major “axes” of integration were identified:

- RDF (http://www.w3.org/RDF/)for public databases, both archival (e.g. BBMRI) and added value (e.g. Metabolomics)
- tranSMART (http://transmartfoundation.org/overview-of-platform/) to securely integrate private clinical datasets
- Interfaces
  - o BioJS (http://biojs.net/) javascript widgets to visualize and integrate existing web resources
  - o REST (https://en.wikipedia.org/wiki/Representational_state_transfer) interfaces to enable better performance and architecture maintenance
  - o User interfaces (https://en.wikipedia.org/wiki/User_interface) to provide accessibility to data was previously not available

Sustainability: To maximize the sustainability of the pilots and activities detailed below, each is:

- Use Case driven: Iteratively refined over the course of the project to be maximally useful
- Distributed: Individual pilots are maintained by the research infrastructure that expressed the need for the use case and developed these
- 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 were used to aid discoverability and collaboration and to ensure that implementation was not blocked by data access issues.

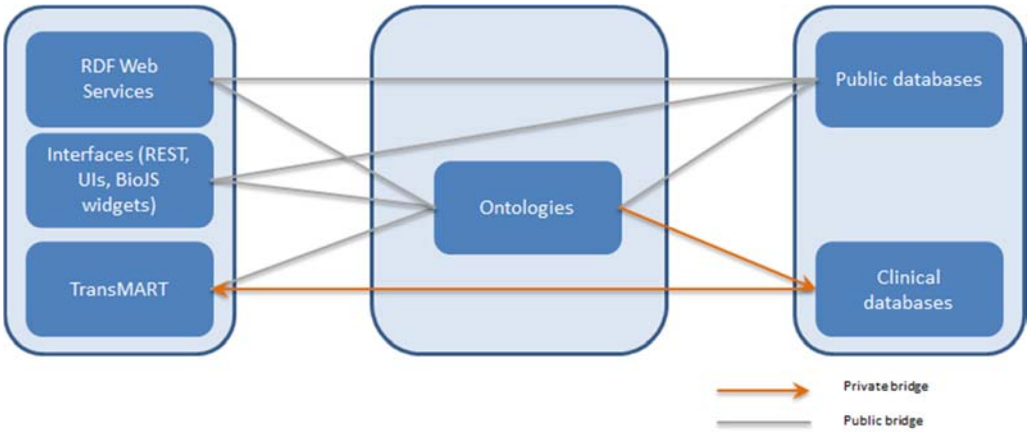


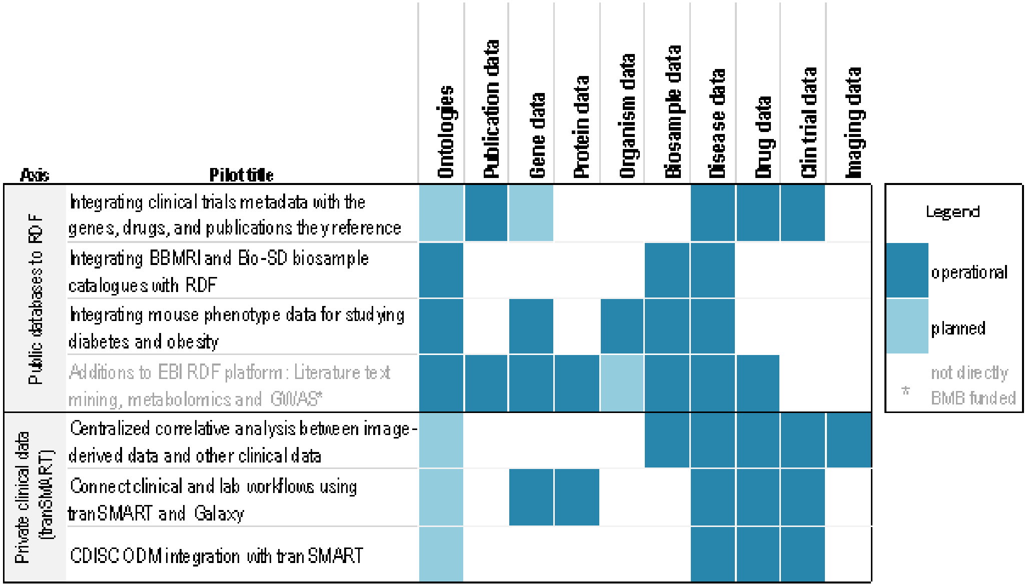
Figure 1 Technical integration axis

3.2.1 Overview of RDF and tranSMART pilot projects

Use Cases: D4.6 comprised seven of the nine individual pilot projects in this category (the remainder being BioJS visualisation widgets described in 3.2.3). Each of these pilots was:

- developed in response to a specific use case within the BMB community
- integrated diverse data types, spanning the domains within the BMB community (Figure 2)
- incorporates modern web standards including ontologies and Semantic Web technologies where appropriate

- planned for sustainability beyond the duration of BMB, and to continue development of new features in many cases (Figure 3)



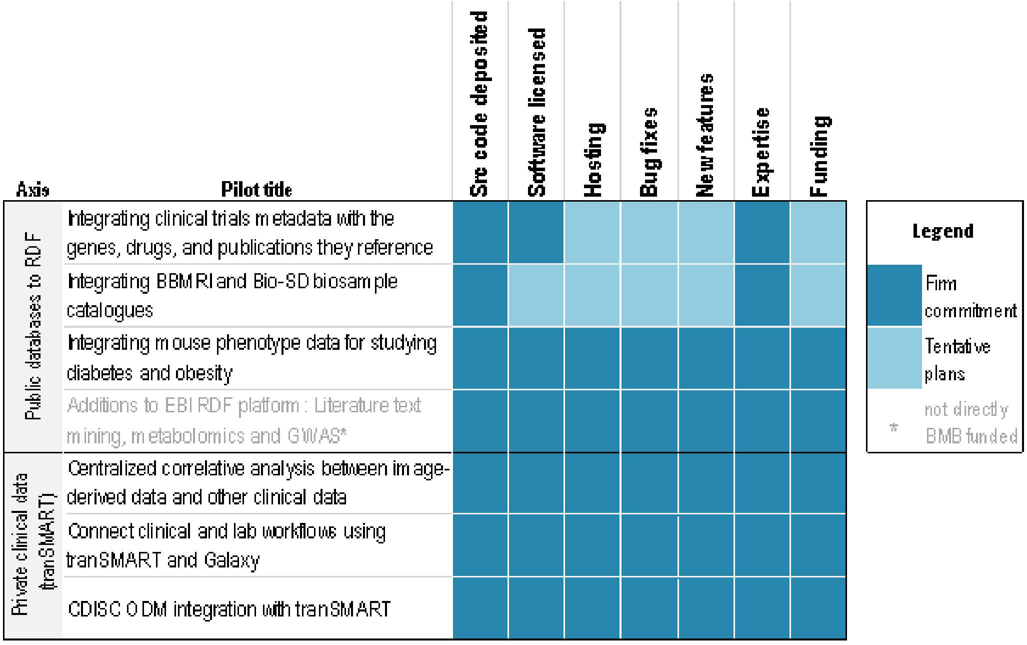
Axis	Pilot title	Ontologies	Publication data	Gene data	Protein data	Organism data	Biosample data	Disease data	Drug data	Clin trial data	Imaging data
Public databases to RDF	Integrating clinical trials metadata with the genes, drugs, and publications they reference	▨	■	▨				■	■	■	
	Integrating BBMRI and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

Legend

■	operational
▨	planned
*	not directly BMB funded

Figure 2 Data touching points across the RDF/tranSMART pilot projects

As an indicator of sustainability beyond BioMedBridges, information was captured pertaining to the availability of source code, whether the partner developing the pilot had undertaken to maintain it beyond the project, whether further funding was established and if new features would be implement. This information (Figure 3) demonstrates that the majority of these pilots have sustainability plans post-BioMedbridges.



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	■	■	□	□	□	■	□
	Integrating BBMRI and Bio-SD biosample catalogues	■	□	□	□	□	■	□
	Integrating mouse phenotype data for studying diabetes and obesity	■	■	■	■	■	■	■
	Additions to EBI RDF platform : Literature text mining, metabolomics and GWAS*	■	■	■	■	■	■	■
Private clinical data (tranSMART)	Centralized correlative analysis between image-derived data and other clinical data	■	■	■	■	■	■	■
	Connect clinical and lab workflows using tranSMART and Galaxy	■	■	■	■	■	■	■
	CDISC ODM integration with tranSMART	■	■	■	■	■	■	■

Legend	
■	Firm commitment
□	Tentative plans
*	not directly BMB funded

Figure 3 Partner plans for sustainability of RDF/tranSMART pilot projects beyond BMB

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

One potential solution to data integration is through the use of RDF based services and semantic technologies and standards. Through BioMedBridges, various datasets have been made available as RDF linked data, including mouse phenotype, clinical trials and biobank data from BBMRI (D4.6), as well as data from genome wide association studies (GWAS), literature from EuropePMC, and metabolic data from MetaboLights. This was through complementary funding but was informed by the activities of BioMedBridges such as following the roadmap and benchmarking (see Appendix A.4). An overview of the RDF integration efforts and technical details is given (Table 1 and Table 2, respectively).

Table 1 Overview of new RDF data integration

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?	Mouse phenotype 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	What clinical trials about multiple myeloma exist and which publications exist that address specific topics of the identified clinical trials.	Clinical trials information from ClinicalTrials.gov, scientific publications from PubMed and data about associated biosamples from BioSamples	No RFD was employed, but linking and querying was done by Clinical Trial Metadata (search terms), PubMed Unique Identifier (PMID), ClinicalTrials.gov Identifier (NCT), BioSample Identifier

Table 2 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	1854	Apache Jena	N/A	BMB 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	BMB 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.	Europe PMC	ELIXIR

					Scaling and updating still outstanding.		
Metabolomics data	Metabolights, MassBank	free text to RDF	TBD	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	n.a.	n.a.	Solr, jQuery	n.a.	BMB WP4	ECRIN

* RDF was not employed, with metadata used for queries (PubMed Unique Identifier, ClinicalTrials.gov Identifier, BioSample Identifier). It was found that RDF was not the best solution as transformation of the complete ClinicalTrials.gov database resulted in an unmanageably large file, making any query slow and complicated.

The ELIXIR RDF platform was incrementally improved during the project (using BMB partner and feedback and some BMB resources for improvements e.g. in work on Ensembl integration). It is used by ELIXIR software applications, external groups including several industrial users (see below), and a variety of semantic web users. It is of growing relevance to the community, and will be used in development related to recently awarded EU grants including CORBEL and EXCELERATE. The platform has around 55 million hits monthly, executing 99% of queries in less than one second, and served over 115 million queries in 2014. Usage statistics for the period 2014-2015 are presented in Table 4; these represent the usage of the public-facing RDF platform components such as the documentation website and public SPARQL endpoints. It does not include dependent applications such as Zooma and GWAS. In addition, the statistics do not include the bulk download services offered through the platform, which encompass one of the major intended use cases the platform seeks to address. As such, these statistics are likely a significant underestimate of the actual usage statistics.

Table 3 ELIXIR-EBI RDF Platform usage statistics

Dataset	2014 hosts	2014 requests*	2015 hosts	2015 requests*
Reactome	683	67085437	1300	9420430
ChEMBL	2507	2301952	3580	8088882

atlas	2303	48576541	3041	1416592
biosamples	1050	834311	1955	953070
RDF web	10453	212994	15568	313950
BioModels	1102	108567	1339	117313
UniProt	339	704	564	1293
URI resolver	3058	231992	3355	104030
Total*	21,495	119,352,498	30,702	20,415,560

*Since some queries are federated across multiple endpoints, the 'requests' statistics are difficult to interpret directly, but do provide an indicator of usage levels

Linked Data initiatives are of growing relevance to the Life Science community, as well as for the pharmaceutical industry; members of the EMBL-EBI Industry Programme (Eli Lilly, UCB, Sanofi Aventis, Roche and Syngenta) have recently committed to Linked Data strategies and other pharma and agrifood companies are investigating this technology. The scale of investment is large, and may fuel coalescence behind these technologies as triple store suppliers seek to align to client requirements. Pilot studies within the Centre for Therapeutic Target Validation (http://www.targetvalidation.org/) 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 (http://www.generalbioinformatics.com/) 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[6] initiative for data interoperability and reuse and future work on RDF metadata profiles will contribute to the robustness and discoverability of RDF.

Central to the usefulness of this platform are the ontologies through which it allows data integration (Figure 4). Consequently, much of the integration work undertaken focused on exposing these metadata 'hooks' through which different dataset components could be related together. While individual datasets did contain this information prior to BioMedBridges, this was not exposed through an appropriate and standard interface. The RDF-ization of these data, and their

[6] http://www.datafairport.org/

consequent integration, is a major and significant output of BioMedBridges, which allows these exposed links to be further exploited through additional (future) dataset integration.

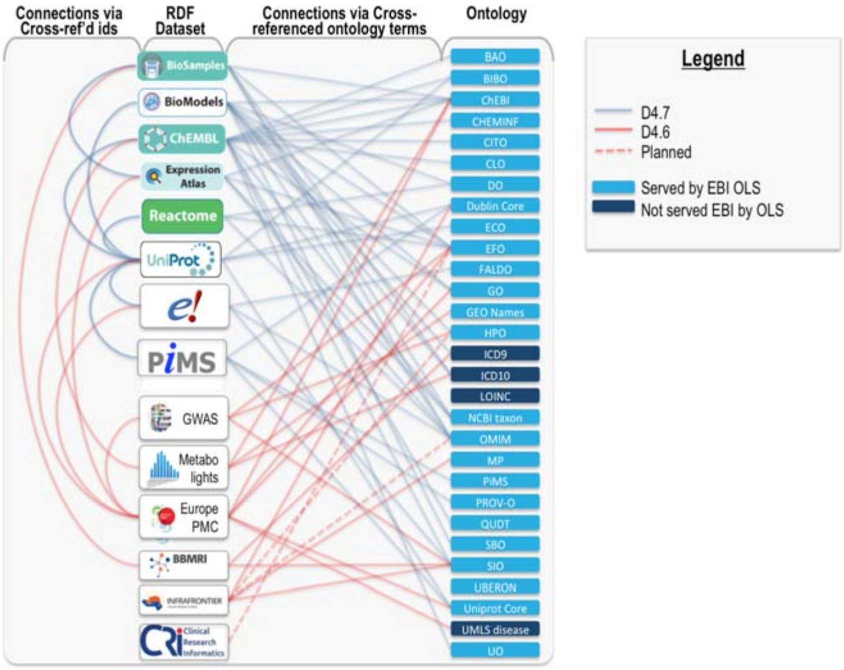


Figure 4 Ontology use by Dataset

Connecting lines show linkages between datasets (left) through cross-references, and ontologies (right). Grey connecting lines are those previously reported in D4.7 and red connecting lines were reported in 4.6. Dashed lines highlight planned extensions of current usage through incorporation of additional ontologies. 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. However, some ICD-9 and ICD-10 terms are cross-referenced through the EFO (Experimental Factor Ontology) and DO (Disease Ontology), allowing the mapping of ICD terms indirectly. This figure is not exhaustive; many links to other databases exist but are omitted for simplicity. See also Table 1.

Pilot 1: Integrating clinical trials metadata with the genes, drugs, and publications they reference (CTIM)

- **Background**: For researchers in biomedical sciences interested in clinical trial, it is indispensable to get access on clinical trials data and associated publications linked to their specific point of research. Not only accessing but in many cases finding the right publications is often a time consuming problem which is not easy to solve when the link to the research area is not completely clear.

- **Objective**: to develop a pilot for the extraction of disease related assertions from full text scientific literature related to acute myeloid leukemia. This includes the description of fitting algorithms for text mining and suitable software for the implementation and integration.
- **Key institutions**: University of Düsseldorf (UDUS, on behalf of ECRIN)

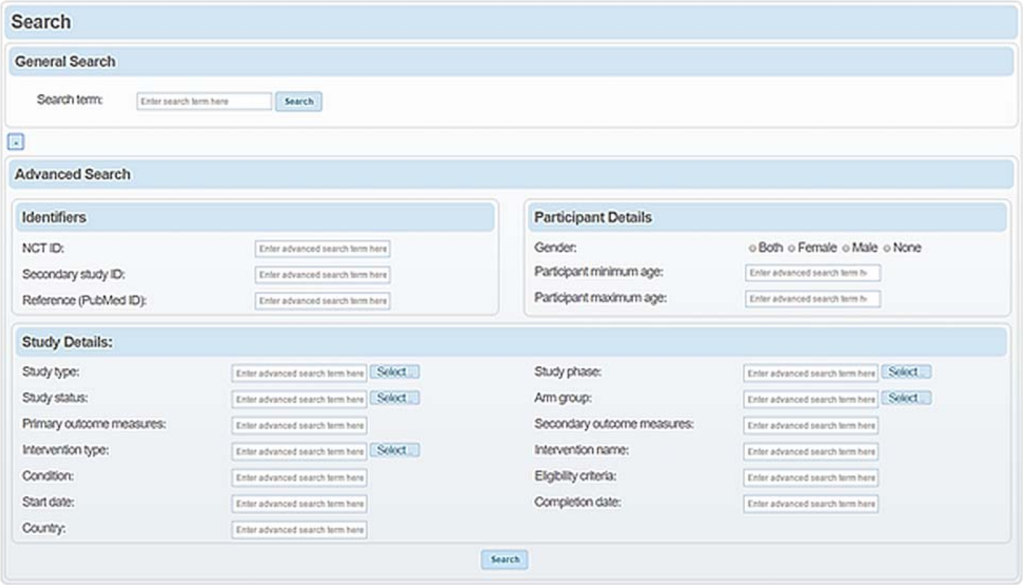
Introduction

In BioMedBridges the demand for a tool existed to improve the planning and assessment of clinical trials and to evaluate the effects of drugs by integrating clinical trials data with information of associated publications and biosamples / genomic information. This tool was required to provide an intuitive user interface for a simple query system that links searches in clinical trial registries with relevant publications and information about biosamples.

The Clinical Trial Information Mediator (CTIM) was developed to support researchers in finding and bringing together trial related information spread across different databases. It was designed to link clinical trials information to corresponding publications (e.g. trials and publications dealing with myeloma therapies) and to relevant biosamples or gene data. It has been tested in several use cases: search for information about clinical trials, search for eligibility criteria for participation in clinical trials, researching the area of high levels of cholesterol, finding biomarker genes associated with therapies tested in clinical trials, search for conditions of severe adverse effects due to chemotherapy in clinical trials,

Technically, Solr was implemented as search engine. Its database is filled with all study fields and result fields from ClinicalTrials.gov, all data are indexed and the field names are mapped to the appropriate values. So it is possible to search for all fields, just for specific ones or to search through all textual blocks of the clinical trials. This functionality exceeds the current search options offered by ClinicalTrials.gov. A trial is linked to an associated journal article through a unique identifier, or an unstructured trial-article link enabling the search for associated information between trials and publications. For example, a trial record in a repository may contain a free-text reference to a journal article or a journal article may mention a trial name or acronym but does not unambiguously identify a registered trial. Both links will be identified by CTIM. The web services of NCBI and EMBL-EBI were used to connect information for publications and biosamples to the clinical results. Both sources provide a programmatic interface with the ability to query for information and to get a XML response in return.

A Graphical User Interface to make the query process as easy as possible was developed. The user interface of CTIM is based on requirements of the research process of the BioMedBridges data bridges and simplifies the search for text information associated with clinical trials. The researcher has the option of either using a general term (single search field) or an advanced search (with choices between different search options). By offering the simple search option (Google paradigm; Figure 5) CTIM addresses the needs of clinical investigators to get an overview or receive fast, preliminary results that can be improved with further queries. Indeed, during presentations of the tool, most users preferred the simple search option over an advanced search. We therefore expect that the choice between simple and advanced search will increase its acceptance in the research community.



Search

General Search

Search term: Enter search term here

Advanced Search

Identifiers

NCT ID: Enter advanced search term here

Secondary study ID: Enter advanced search term here

Reference (PubMed ID): Enter advanced search term here

Participant Details

Gender: ○ Both ○ Female ○ Male ○ None

Participant minimum age: Enter advanced search term h

Participant maximum age: Enter advanced search term h

Study Details:

Study type: Enter advanced search term here

Study status: Enter advanced search term here

Primary outcome measures: Enter advanced search term here

Intervention type: Enter advanced search term here

Condition: Enter advanced search term here

Start date: Enter advanced search term here

Country: Enter advanced search term here

Study phase: Enter advanced search term here

Arm group: Enter advanced search term here

Secondary outcome measures: Enter advanced search term here

Intervention name: Enter advanced search term here

Eligibility criteria: Enter advanced search term here

Completion date: Enter advanced search term here

Figure 5 User Interface of CTIM. Simple search (top) and the advanced search (bottom) are shown

The CTIM prototype was developed against an Acute Myeloid Leukemia (AML) use case from WP8; initial work was reported in D4.3 (Solr instantiation and population with clinicaltrials.gov data), further developed in D4.5 (CTIM portlet within Liferay Portal CE) and refined in D4.6 (expanded data coverage, PubMed and BioSamples web services implemented, User Interface development). Further information is available in the appropriate deliverable report, and in Appendix A.1.

Subsequent to D4.6, the following improvements/enhancements have been made:

- the entire clinical trials database from CT.gov is being included, surpassing the original AML use case
- search algorithms have been optimised
- user interface optimisation performed, especially representation of results
- tool tested with different clinical use cases
- the performance of the tool was evaluated
- dissemination of the tool

The use of identifiers and web services for this CTIM pilot and the abandonment of RDF approach had several reasons which centred on providing improved simplicity and the usability of the tool for the common researcher. Using a SPARQL-Endpoint for supporting semantic web-based searches was discussed internally, but the structure and means of RDF and SPARQL to explore RDF databases do not easily serve as a search tool for human users. Rather, these web technologies are optimised for the gathering of links and locations to RDF to provide further data and information. For example, LinkedCT is the linked data version of ClinicalTrials.gov. It is an open semantic web data source for clinical trials data. For this purpose, clinical trials are transformed into RDF, links between records of trial data and several other data sources, such as Bio2RDF or PubMed exist. CTIM does not use semantic web technologies, but a flexible and quick string search. CTIM is easily usable by humans through intuitive interfaces. Nonetheless, for the automatic search and analysis of data semantic web standards may be preferable. But CTIM is aiming to gather rich information related to a certain topic, to provide an overview to the information available and a link to examine the resource further. In contrast, RDF aims to identify resources with strict properties with little or no information about the resource format, composition or content. To offer information to web contents, the web services of PubMed and BioSample were the better choice in terms of usability. Either automatically or by direct web search, a simple request to the web service provides information that correlates to a search term. Therefore, web services where used to feed CTIM searches for additional information regarding clinical trials. Partner sustainability statement: CTIM source code is open source.

Pilot 2: Integrating BBMRI and Bio-SD biosample catalogues

- **Background**: The BBMRI.eu catalogue provides an overview of the biobank landscape across Europe. BBMRI-LPC uses an advanced version of this catalogue for the Large Prospective Cohorts of the BBMRI-LPC project. Both catalogues have been linked to BioSD, and the LPC catalogue is being used in the pilot implementation of WP 5, Task 8.
- **Objective**: To develop a semantic service endpoint for the BBMRI.eu and BBMRI-LPC catalogue that allows queries for disease groups to get biobanks with sample types and number of samples.
- **Key institutions**: Technische Universität München (on behalf of BBMRI)
- **Key people**: Raffael Bild
- **Integration paths**: BBMRI-ERIC, BBMRI-LPC, ELIXIR

Biobank catalogues (BBMRI.eu, https://www.bbmriportal.eu/bbmri/ BBMRI-LPC, http://www.bbmri-lpc-biobanks.eu/catalogue.html) have been linked to BioSD (http://www.ebi.ac.uk/biosamples/). The LPC catalogue was also used in the pilot implementation of WP 5, Task 8. BBMRI use cases were developed during, and subsequent to, the preparatory phase of BBMRI, and continued by BBMRI-ERIC and –LPC (http://bbmri-eric.eu/reports). REST web services were developed by BMB (D4.3), and built upon (D4.6) to provide a data bridge between BBMRI and ELIXIR.

The work entailed the creation of an OWL ontology to represent biobank metadata and sample counts as RDF using OWL classes and properties through purpose-built Java code, and to expose this data through a SPARQL endpoint (https://www.bbmriportal.eu/bbmri2.0/sparql.html). Further information is available in the appropriate deliverable report, and in Appendix A.2.

Subsequently to D4.6, both the XML schema used by the REST web service and the OWL ontology have been extended, with particular focus on adding attributes specified by MIABIS 2.0 (https://github.com/MIABIS/miabis/wiki). Attributes which are newly exposed by the web services include access conditions, juristic person, and country of a biobank as well as the data categories provided (e.g. survey data, medical records, and biochemical measurements). Furthermore, biosample counts are now being provided for any combination of material type and storage temperature.

A tool, BiobankConnect[7], has additionally been developed in collaboration with WP3 and BioSHaRE. This allows semi-automated matching between desired and available data elements across biobank data through an intuitive user interface[8], significantly speeding up the harmonization process in pooling data from multiple locations. The software is available as an open source app, and may be useful for further integration efforts. The BiobankConnect tool will be taken forward within BBMRI-NL, BBMRI-ERIC, CORBEL and RD-connect.

Partner sustainability statement:

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.

Pilot 3: Integrating mouse phenotype data for studying diabetes and obesity

- **Background**: Systemic phenotyping data in mice is an invaluable resource for our understanding of human diseases. Large scale initiatives like the IMPC (International Mouse Phenotyping Consortium) aim to provide a comprehensive catalogue of mammalian gene function to foster research into human diseases. However, the current technical

[7] http://identifiers.org/pubmed/25361575
[8] demo available at http://www.biobankconnect.org

implementation does not fully exploit the potential of integrating these mouse phenotyping data with other resources. Enabling the semantic integration of the IMPC data will address this issue and will also allow straightforward technical application of the outcomes of the PhenoBridge work package (WP7).

- **Objective**: Development of a RDF scheme for systemic phenotyped mice data. Integrate mice data with the resources supported by outcomes of work package 7 (PhenoBridge).
- **Key institutions**: HMGU, EBI
- **Key people**: HMGU: Philipp Gormanns, Elida Schneltzer ; EBI: Terry Meehan, Helen Parkinson
- **Integration paths**: WP7

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).

The REST interface developed previously (D4.3) was built upon in (D4.6). 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. These user interfaces were subsequently enhanced to leverage the semantic richness made possible in D4.6, where a semantic data model was designed (using established identifiers and ontologies, to facilitate interoperability/integration between different murine resource providers. Over the course of this work, additional datasets have been transferred into the RDF triple store and exposed through a SPARQL endpoint (http://mousemodels.infrafrontier.eu/rdf/sparql), as well as providing a dynamic 'Phenomap' (http://mousemodels.infrafrontier.eu/tools/phenomap.jsf). Further information is available in the appropriate deliverable report, and in Appendix A.3.

Subsequently to D4.6, the following features or functionality have been incorporated[9].

[9] while the integration is more an outcome of WP7, WP4 supported the interface development

1.) The Diabetes ontology (DIAB) from WP7 is now completely integrated and can be accessed via the SPARQL endpoint. It allows the query of mouse model for phenotypes which are known to occur at specific stages in Human Type 2 Diabetes.

2.) Together with WP7 the M3(Mining mouse models) tool has been developed; It provides rich user interfaces to analyse mouse phenotyping data alone and in the context of Diabetes. The tool is currently in beta test stage (mousemodels.infrafrontier.eu/demo/m3.jsf) and is planned to be published in beginning of 2016. A demo of the tool was shown at the Mouse phenotype resources workshop at the BMB symposium in November 2015.

Partner sustainability statement:

Our SPARQL endpoint and the M3 tool will be maintained by INFRAFRONTIER.

In 2016 it is planned to integrate these outcomes into the INFRAFRONTIER domain to access them via the INFRAFRONTIER main website (http://www.infrafrontier.eu). Since our mouse model RDF scheme was designed for further mouse resources integration, the incorporation of EMMA (European Mouse Mutant Archive) will be discussed at this time.

Pilot 4[10]: Additions to EBI RDF platform: Literature text mining, metabolomics and GWAS (through complementary funding)

- **Background**: Central to ELIXIR will be a solid platform for the integration and exchange of data between the ELIXIR hub, ELIXIR nodes, BioMedBridges partners and wider European stakeholders. Key to a successful infrastructure will be an open, inclusive, flexible and loose coupling model that does not impose technically complex or proprietary requirements.
- **Objective**: Develop a blueprint to aid partners in the provision of RDF infrastructure endpoints, lowering the barrier to entry and accelerating adoption. This will include guidelines for boilerplate components and

[10] Note: This pilot was not funded by BioMedBridges. No further information available at this time.

factors impacting on interoperability, an example architectural model and recommendations for software and/or hardware.

- **Key institutions**: EMBL-EBI, [Swiss Institute of Bioinformatics]
- **Key people**: Andy Jenkinson, Julie McMurry
- **Additional integration paths**: via UniProt to INSTRUCT

The ELIXIR/EBI RDF Platform hosts various RDF linked data for resources hosted at EBI. Three new datasets were added to the RDF platform (development version): MetaboLights (metabolomics data funded primarily by the Cosmos FP7 project), literature text mining (Europe PMC literature database[11]), and Genome Wide Association Studies (http://www.ebi.ac.uk/fgpt/gwas/). 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 BMB semantic web pilot has established.

Representation of MetaboLights as RDF highlighted the complexity of metabolite naming conventions and use; a name may refer to a specific molecule or sometimes a class of molecule, depending upon the context. Consequently, to ensure appropriate semantic mappings, it was necessary to revisit the data model, which required improvement. We anticipate a public release of the dataset in 2016. A text mining pipeline was implemented for Europe PMC literature database, which is currently being refined; we anticipate that the link will be advertised at the end of 2015. Further information is available in the appropriate deliverable report, and in Appendix A.4.

Subsequent to D4.6 the following improvements were made:

- BioSamples data updated (November 2015), number of triples 361635520. Documentation available at https://www.ebi.ac.uk/rdf/services/biosamples/
- Gene-transcript-protein representation was improved starting from the Ensembl data model and an RDF model was generated. Ensembl has

[11] 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

produced an RDF serialisation of genomic annotation on seventy model organisms. By doing this, Ensembl contributes a richly annotated and connected dataset to the Linked Data world. New opportunities for data integration have appeared immediately with other Linked Data providers, such as DisGeNet and the EBI text mining group allowing for the discovery of disease associations and research publications relating to genomic annotation.
At Biohackathon 2015 in Nagasaki (http://2015.biohackathon.org/) we collaborated with a large potential user group and drew on their collective insight to refine our data model. Subsequently we have integrated the production of Ensembl RDF into our regular release process so that the data is kept up to date with the latest annotation. Critically, Ensembl RDF is also published via the EBI RDF platform, where it fills a void in the available Linked Bioinformatics Data and embraces shared vocabulary and compatible standards. Additionally we have presented Ensembl RDF via the RDF platform as a component of a training course "Using the Semantic Web for faster (Bio-)Research" hosted at the Swiss Institute for Bioinformatics.

- Identifier use in the RDF platform was discussed in a submitted paper in collaboration with WP3.
- A collaborative workshop in Japan attended by relevant personnel and reviewed recent changes above. This was also used as a forum for user feedback

3.2.1.2 Integration of simple object queries using tranSMART

There are many challenges to creating bridges across medical research, particularly when clinical data is a domain to which a bridge would be desirable; two major concerns precluded the selection of RDF as a solution for integration across clinical data:

1) It would conflict with existing industry standards (e.g. DICOM for images) which are already heavily used (by pharma, academics, vendors of instruments and software),

2) The security concerns associated with graph-based data, since it is still very immature and patient confidentiality prohibits the public distribution of source data.

A suitable alternative was identified 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.[12] It is the platform of choice for the Netherlands premier translational research informatics program, CTMM-TraIT (http://www.ctmm-trait.nl/) in addition to that of the Innovative Medicines Initiative (http://www.imi.europa.eu/), "Europe's largest public-private initiative aiming to speed up the development of better and safer medicines for patients."[13]

Three tranSMART-related pilots were conducted. More detailed information is available in the deliverable report, D4.6 and 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.

An overview of the integration between the platforms and ontologies is shown in Figure 6. It should be noted that Galaxy supports ontologies served through BioPortal (italicised text), including UBERON, but not through OLS[14.]

[12] http://transmartfoundation.org/overview-of-platform/
[13] http://www.imi.europa.eu/
[14] http://www.ebi.ac.uk/ols/beta

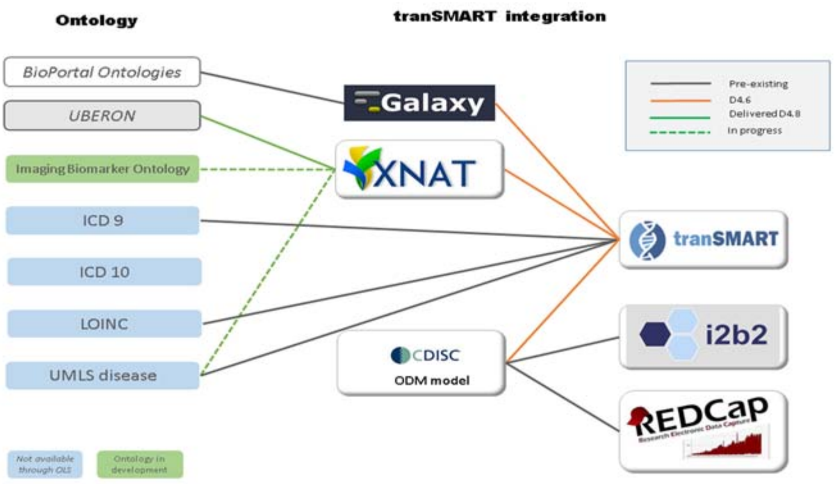


Figure 6 Overview of tranSMART Integration Pilots

Pilot 1: Centralized correlative analysis between image-derived data and other clinical data

- **Background**: Medical imaging (MRI, CT, Ultrasound, etc) is more and more included in multi-centre clinical studies. Infrastructure is needed to support centralised correlative analysis between image-derived data (i.e. organ volume measurements) and other clinical data (disease status, age, etc).
- **Objective**: Setup tranSMART web service which facilitates correlative analysis between image-derived data extracted from the XNAT service and clinical data extracted from an OpenClinica service.
- **Key institutions**: Erasmus MC (on behalf of EuroBioImaging); strong link with CTMM TraIT project (EATRIS/BBMRI)
- **Key people**: Stefan Klein (Erasmus MC), Erwin Vast (Erasmus MC), Marcel Koek (Erasmus MC)
- **Person months 4.4**: None
- **Person months 4.6**: 9PM
- **Integration paths**: tranSMART also links with research data on genetics, so this solution will also facilitate correlative analysis between genotypes and image-derived phenotypes.

This collaborative pilot involved CTMM TraIT (EATRIS/BBMRI) (http://www.ctmm-trait.nl/), who provide an IT infrastructure that collects, stores and analyses data generated in biomedical research projects, in which medical imaging data (MRI, CT, Ultrasound) plays an important role. The aim of this pilot was to extend the TraIT infrastructure to better support centralized statistical analysis of imaging biomarkers in relation with other research data. The solutions developed have been made open-source, and are not only applicable within the TraIT infrastructure, but also more widely.

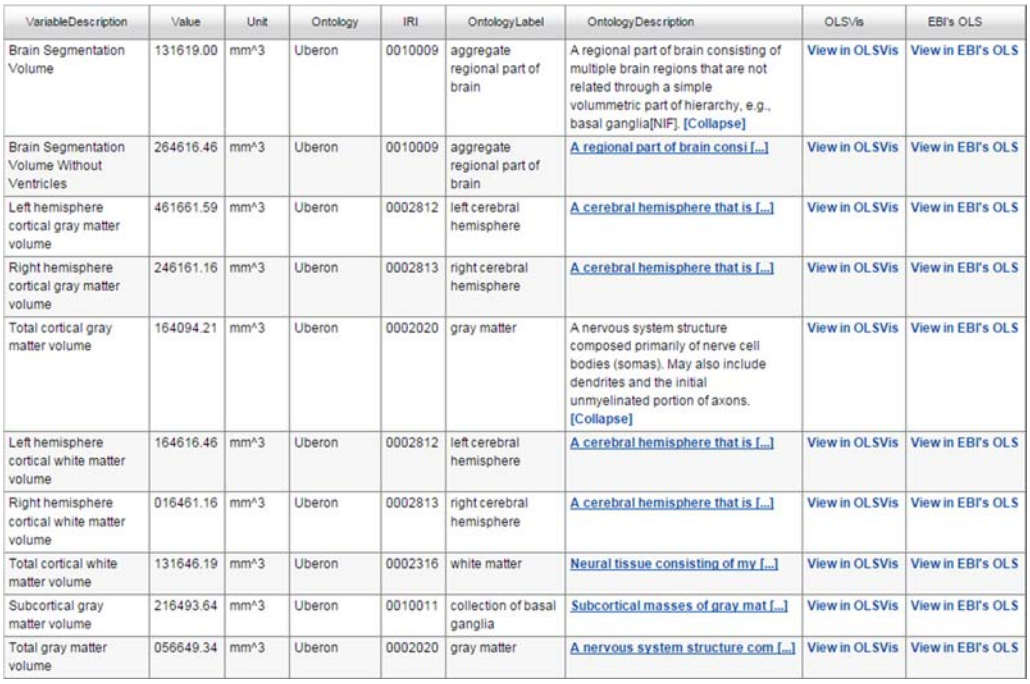
Work in this pilot required: 1) installation of a medical imaging platform (XNAT, D4.5) to store clinical imaging data (in DICOM format) and image-derived measurements; 2) development of a new open-source tranSMART plugin, to import image-derived data from XNAT to tranSMART (D4.6); 3) evaluate the XNAT-tranSMART bridge in the context of a study (Guyader et al[15)] that involved analysis of "apparent diffusion coefficients" (ADC[16)] imaging biomarkers (also D4.6). Also reported in D4.6 was an update of the XNAT installation to release 1.6.4; installation of a new XNAT image viewer; improving usability; and several updates to the Puppet configuration scripts (https://bitbucket.org/bigr_erasmusmc/puppet-xnat), which are available through the XNAT marketplace (http://marketplace.xnat.org/). Further information is available in the appropriate deliverable reports, and in Appendix B.1. For the REST federation of the BBMRI-NL biobank catalogues the RSQL REST API framework was reused (https://github.com/jirutka/rsql-parser), which has been positively evaluated and has now become the standard API for MOLGENIS (http://molgenis.github.io/documentation/) to be re-used for the BBMRI-ERIC biobank catalogue ('directory') federation and RD-Connect.

Subsequent to D4.6, during the last period of BioMedBridges, Erasmus MC has focused on activities aimed at maximising the adoption and usage of the tools previously developed in WP4 (and described in D4.5 and D4.6). The contributions were five-fold:

[15] 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.

[16] The ADC is a measure of the magnitude of diffusion (of water molecules) within tissue (http://radiopaedia.org/articles/apparent-diffusion-coefficient-1).

- Co-organisation of a workshop at the “Open Bridges For LifeScience Data” symposium on translational research infrastructure, in collaboration with EATRIS and BBMRI (MOLGENIS/BiobankConnect software from WP3). See the end of section 3.2.1.2 for more details. A press release on the workshop can be found in Appendix B.4.
- Organisation of a national workshop on XNAT, including a hackathon, conferencing sessions, and strategic meetings. An XNAT tutorial (developed for the workshop mentioned above) was offered here as well. Attendants ranged from experienced XNAT users/administrators, to novice and potential users, to people with a more strategic interest in infrastructures for medical research.
- Installation of the XNAT-tranSMART importer plugin (see D4.6) on the production environment of CTMM TraIT (EATRIS/BBMRI), such that it can be used by multi-centre biomarker studies. The XNAT-tranSMART importer plugin was presented at the Annual tranSMART meeting: http://lanyrd.com/2015/transmart-foundation-annual-meeting/sdrzyt.
- We performed a pilot study on the feasibility of linking to the UBERON anatomy ontology from within XNAT. By linking to the UBERON ontology, imaging biomarkers stored in XNAT can be annotated in a more standardized way than currently possible. We used the REST interface of the EBI Ontology Lookup Service (OLS) to implement this. Also, we included a link to the OLS Visualization browser (OLSVis http://ols.wordvis.com/). In principle, the implementation supports any ontology that is supported by the EBI OLS. The screenshot below (Figure 7) shows how the information is visualized in XNAT in our pilot implementation.
- Miscellaneous support actions, including improvement of documentation, maintenance of XNAT servers, minor improvements and bug fixes, collection of usage metrics, and user support.



VariableDescription	Value	Unit	Ontology	IRI	OntologyLabel	OntologyDescription	OLSVis	EBI's OLS
Brain Segmentation Volume	131619.00	mm^3	Uberon	0010009	aggregate regional part of brain	A regional part of brain consisting of multiple brain regions that are not related through a simple volummetric part of hierarchy, e.g., basal ganglia[NIF]. [Collapse]	View in OLSVis	View in EBI's OLS
Brain Segmentation Volume Without Ventricles	264616.46	mm^3	Uberon	0010009	aggregate regional part of brain	A regional part of brain consi [...]	View in OLSVis	View in EBI's OLS
Left hemisphere cortical gray matter volume	461661.59	mm^3	Uberon	0002812	left cerebral hemisphere	A cerebral hemisphere that is [...]	View in OLSVis	View in EBI's OLS
Right hemisphere cortical gray matter volume	246161.16	mm^3	Uberon	0002813	right cerebral hemisphere	A cerebral hemisphere that is [...]	View in OLSVis	View in EBI's OLS
Total cortical gray matter volume	164094.21	mm^3	Uberon	0002020	gray matter	A nervous system structure composed primarily of nerve cell bodies (somas). May also include dendrites and the initial unmyelinated portion of axons. [Collapse]	View in OLSVis	View in EBI's OLS
Left hemisphere cortical white matter volume	164616.46	mm^3	Uberon	0002812	left cerebral hemisphere	A cerebral hemisphere that is [...]	View in OLSVis	View in EBI's OLS
Right hemisphere cortical white matter volume	016461.16	mm^3	Uberon	0002813	right cerebral hemisphere	A cerebral hemisphere that is [...]	View in OLSVis	View in EBI's OLS
Total cortical white matter volume	131646.19	mm^3	Uberon	0002316	white matter	Neural tissue consisting of my [...]	View in OLSVis	View in EBI's OLS
Subcortical gray matter volume	216493.64	mm^3	Uberon	0010011	collection of basal ganglia	Subcortical masses of gray mat [...]	View in OLSVis	View in EBI's OLS
Total gray matter volume	056649.34	mm^3	Uberon	0002020	gray matter	A nervous system structure com [...]	View in OLSVis	View in EBI's OLS

Figure 7 A screenshot of image-derived data in XNAT, stored in a novel data type that supports links to ontologies like UBERON. Using the REST interface of the EBI Ontology Lookup Service, a description of the imaging biomarker is dynamically obtained and shown

Partner sustainability statement:

The developments on XNAT, tranSMART and MOLGENIS will be continued in the Horizon2020 CORBEL project, and in the BBMRI-NL2.0 project. Funding for the coming years is thus ensured. Efforts will focus on 1) user support, 2) evaluation of the tools in imaging-genetics studies, correlating features from clinical imaging data to genetic factors, 3) further use of ontologies to annotate imaging biomarkers in XNAT in a standardized way.

Pilot 2: Connect clinical and lab workflows using tranSMART and Galaxy

— **Background**: Within EATRIS/CTMM TraIT, several tools are used to store and process biomolecular data, such as: Galaxy, a popular bioinformatics workflow tool, and Molgenis (UMCG) and Phenotype Database. Both tools contain experimental data and study design for 'omics' and similar experiments. Those tools all expose REST APIs that deliver JSON data.

- **Objective**: In this deliverable, we will document these APIs for the specific public as well as private repositories available via EATRIS/CTMM-TraIT, and design a common data model that can be used to interrogate these repositories and link them with clinical, imaging and biobanking data. We will do this via tranSMART (used in the CTMM TraIT project, as well as other European funded projects like eTRIKS and EMIF) as a data-warehouse for translational research data. tranSMART already has an interface to do this in R, and we will expand on that to also deliver the specific biomolecular (raw, processed, catalogued) data that is present in the mentioned tools. An example client application will be built in R, and in Java/Groovy to facilitate easy programmatic access to this service by interested third parties.
- **Key institutions**: VUmc (representing EATRIS)
- **Key people**: Ruslan Forostianov (The Hyve), Ward Blondé (VUmc/NKI), Youri Hoogstrate (Erasmus MC), Freek de Bruijn (VUmc/NKI), Jeroen Beliën (VUmc)
- **Integration paths**: Clinical trials portal, Bioshare, tranSMART (4.6)

Both tranSMART and Galaxy are used heavily by the biomedical research community, with the latter enabling workflows. This pilot provided integration between tranSMART and Galaxy, allowing workflows to become newly available within tranSMART as well. This work built upon work done by the Galaxy team (and specifically by John Chilton on the blend4j library) and the tranSMART foundation. The integration itself (D4.6) is enabled through the following components:

- tranSMART (https://github.com/thehyve/Rmodules/tree/features/transmart-galaxy) plugin handling Galaxy workflows in tranSMART (by Ruslan Forostianov http://thehyve.nl/portfolio/ruslan-forostianov/)
- workflow runner (https://github.com/CTMM-TraIT/trait_workflow_runner) simplifies Galaxy API use from Java (by Freek de Bruijn https://github.com/FreekDB)
- blend4j (https://github.com/jmchilton/blend4j) library providing Galaxy API access (by John Chilton https://wiki.galaxyproject.org/JohnChilton)

- Galaxy API (http://galaxy-dist.readthedocs.org/en/latest/lib/galaxy.webapps.galaxy.api.html) REST API enables programmatic access (Galaxy Team https://wiki.galaxyproject.org/GalaxyTeam)

Further information is available in the appropriate deliverable report, and in Appendix B.2.

Subsequently to D4.6, the following features or functionality listed below have been incorporated. In the last period of BioMedBridges, BBMRI (UMCG), EuroBioImaging (Erasmus MC), and EATRIS (VUmc/NKI) have focused on activities aimed at maximising the adoption and usage of the tools previously developed in WP4 (and described in D4.5 and D4.6). The contributions were three-fold:

1. Co-organisation of a workshop at the “Open Bridges For LifeScience Data” symposium on translational research infrastructure, in collaboration with EuroBioImaging and BBMRI. See the end of section 3.2.1.2 for more details. A press release on the workshop can be found in Appendix B.4.
2. Presentation of the OpenClinica-tranSMART and tranSMART-Galaxy connections at the Dutch DTL Programmers Meeting (https://wiki.dtls.nl/index.php/DTL_Programmers_Meeting)
3. Installation of the OpenClinica-tranSMART and tranSMART-Galaxy connections on the production environment of CTMM TraIT (EATRIS/BBMRI).
4. Integration with the MOLGENIS/BiobankConnect open source software for the harmonization and integration of biobanks

Partner sustainability statement:

All tools/services described above are or will be deployed and maintained within TraIT (see http://www.ctmm-trait.nl/) as well as being used within BBMRI-NL (see http://www.bbmri.nl).

The workshop given at Hinxton will be tuned as a permanent training environment. Currently it is being investigated how to set-up such a, probably private cloud, environment in such a way that a “validated” user can start it in a user-friendly way anyplace, anywhere, anytime. More information on sustainability (and other topics) is available in the “BMB-specific software and progress” (https://docs.google.com/spreadsheets/d/19mKivAekDOz4qa0x_ScCa_491z2V18_g13L9vpis5CU/edit#gid=0) document.

Pilot 3: CDISC ODM integration with tranSMART

The standard format for Electronic Data Capture (EDC) systems such as OpenClinica and RedCap is CDISC’s Operational Data Model (ODM), encoded in XML. However, EDCs are not designed with data integration or analysis in mind; there is a need for exporting clinical data from EDCs.

A direct transformation from the ODM format to a tabular format that can be uploaded in tranSMART was the target for this pilot, and resulted in the ODM-to-i2b2 Java conversion tool (D4.6). This tool parses the input ODM/XML file and returns a tabular file of clinical data file, where each row represents the data of one patient, and columns represent data items (age, gender, weight, etc). It also provides meta-data on the columns of the tabular format, and a mapping file to interpret the data.

Five clinical studies, or examples of clinical studies, have been converted and loaded into tranSMART. Further information is available in the appropriate deliverable report, and in Appendix B.3.

Subsequently to D4.6, the following features or functionality have been incorporated:

- This integration was also demonstrated during the workshop at the “Open Bridges For LifeScience Data” symposium on translational research infrastructure; again see the end of section 3.2.1.2 for more details on how the workshop was prepared.
- Co-organisation of a workshop at the “Open Bridges For LifeScience Data” symposium on translational research infrastructure, in a collaboration by BBMRI, EuroBioImaging, and EATRIS

- Outcomes from Workshop on translational research infrastructure at final BioMedBridges symposium are as follows:

  a. The development of sandbox installations of XNAT, tranSMART, Galaxy, Molgenis/BiobankConnect, and OpenClinica on a virtual machine (VM) or in a Docker image that can be run in cloud-based environments.

  b. Development of a tutorial for the XNAT imaging platform, guiding novice users through a few basic operations, like upload, download, viewing, and processing of medical images.

  c. Similar exercises for the other platforms were developed: for the OpenClinica-tranSMART, tranSMART-Galaxy, and MOLGENIS-based connections.

  d. Prior to the workshop, several rehearsal sessions were organised to ensure the exercises were of good quality, did not take too much time, and were instructive for novice users.

  e. We also made sure that the tutorials can - as much as possible - be supported by the entire team and that the technical infrastructure was working as planned.

  f. Organisation and operation of the actual workshop. The Dutch SURFsara high-performance computing (HPC) cloud (https://userinfo.surfsara.nl/systems/hpc-cloud) was successfully used to host 40 instances of the VMs, such that each participant could have her/his own sandbox environment.

A press release on the workshop can be found in Appendix B.4. All tutorials can be found here: https://goo.gl/f14RKa.

Partner sustainability statement:

All tools/services described above are or will be deployed and maintained within TraIT (see http://www.ctmm-trait.nl/) as well as being used within BBMRI-NL (see http://www.bbmri.nl).

The workshop given at Hinxton will be tuned as a permanent training environment. Currently it is being investigated how to set-up such a, probably private cloud, environment in such a way that a “validated” user can start it in a user-friendly way anyplace, anywhere, anytime.

More information on sustainability (and other topics) is available in the “BMB-specific software and progress” document (https://docs.google.com/spreadsheets/d/19mKivAekDOz4qa0x_ScCa_491z2V18_g13L9vpis5CU/edit#gid=0).

tranSMART future work

The source code for the tranSMART platform and the plugins we have developed are already housed in open source repositories. An important next step is to deploy the platform for CTMM-TraIT researchers to use; during the intervening period since D4.6, the XNAT-tranSMART importer plugin has been successfully integrated into the CTMM TraIT platform. User experience testing of both the platform and plugins is being conducted to allow the system to be iteratively improved going forward. 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 6.

3.2.2 Overview of REST Web Service pilots

Six REST pilot integration studies (assessed in D4.2, delivered D4.3) were undertaken, spanning the BioMedBridges domains:

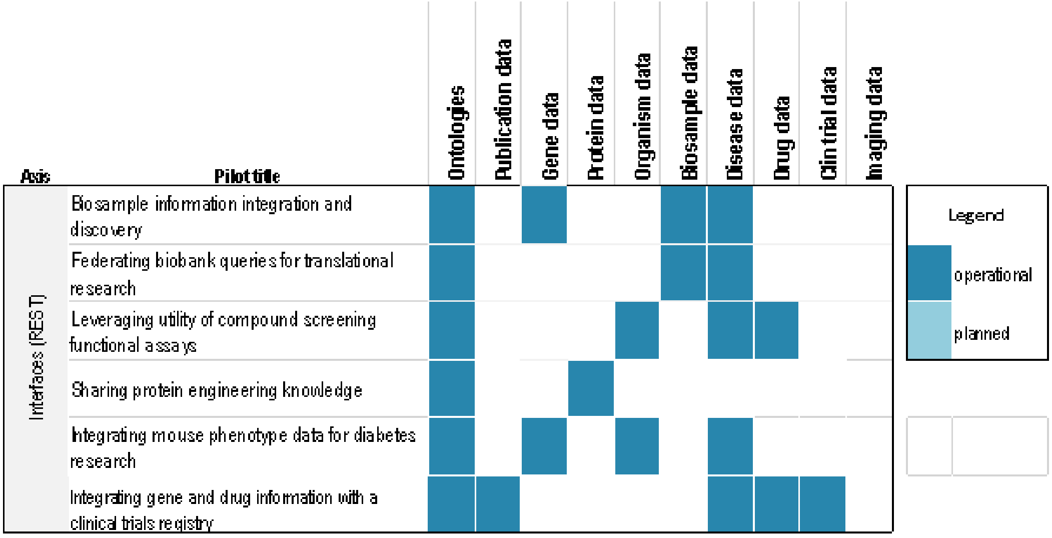
1. Biosample information integration and discovery
2. Federating biobank queries for translational research
3. Leveraging the utility of compound screening functional assays
4. Sharing protein engineering knowledge
5. Integrating mouse phenotype data for studying diabetes and obesity
6. Integrating gene and drug information with a clinical trials registry

These Web services provided a foundation for the integration strategy of BioMedBridges to be expanded with pilots for Semantic Web integration (D4.4. and D4.6), and refer to more detailed reports provided above. These pilots provide 'hooks' that are available for future integration opportunities as they arise.

General technical strategy

The collaborative activities helped to ensure that developments fulfilled the relevant requirements identified by the D4.2 Technical workshop, specifically:

- Adopted technologies are interoperable
- Use of common (or convertible) inputs and outputs
- Use of common identifiers and accessions
- Services are representative of the underlying resources and are extensible in the future
- Developments are sustainable in the context of BioMedBridges and beyond
- The pilots provide biologically meaningful and scientifically valid access to partner data



Axis	Pilot title	Ontologies	Publication data	Gene data	Protein data	Organism data	Biosample data	Disease data	Drug data	Clin trial data	Imaging data
Interfaces (REST)	Biosample information integration and discovery	🟦		🟦			🟦	🟦			
	Federating biobank queries for translational research	🟦					🟦	🟦			
	Leveraging utility of compound screening functional assays	🟦				🟦		🟦	🟦		
	Sharing protein engineering knowledge	🟦			🟦						
	Integrating mouse phenotype data for diabetes research	🟦		🟦		🟦		🟦			
	Integrating gene and drug information with a clinical trials registry	🟦	🟦					🟦	🟦	🟦	

Legend	
🟦	operational
▢	planned

Figure 8 REST data touching points

3.2.2.1 Summary of pilot activities

This section provides executive summaries of the work undertaken as part of deliverable D4.3[17] and their context with reference to subsequent deliverables.

[17] http://www.biomedbridges.eu/deliverables/43

Pilot 1: Biosample information integration and discovery

This pilot integrated two pivotal resources by transferring data from the BBMRI.eu and BBMRI-LPC catalogue to the BioSamples Database (BioSD). This fulfils the urgent requirement for integrated searches over larger sample collections in support of increasingly complex scientific questions, including query by relevant disease ontologies. For example, WP8 (Personalized Medicine) requires search over AML samples and will use heterogeneous AML data from biobanks. In this work, data from the following biobanks was 'pushed' from BBMRI to the BioSamples database:

- KORA- Cooperative health research in the Region of Augsburg
- Atrial Fibrillation Network Munich
- Atrial Fibrillation Network Munich - M4-Cluster-Biobank
- German National Cohort
- Cooperative Health Research in the region of Augsburg - specific studies

These are available via the BioSD search interface and modelled consistently with other similar data in the BioSD context. The pilot has therefore successfully combined multiple resources with complementary content that were not previously amenable to single, integrated queries. All other biobanks in the BBMRI catalogues which agree to publish their data are being seamlessly reflected into the BioSD, ensuring biobank data in the catalog and the BioSD grows simultaneously. The production services required to publish these data are available as web interfaces and through a RESTful API, facilitating future submissions. See also 3.2.1.1, Pilot 2 and Appendix A.2 detailing work done in D4.6.

Pilot 2: Federating biobank queries for translational research

This pilot achieves a seamless federated search of two biobank catalogues via new RESTful services. The pilot takes the first steps to enable integrated search of >200 biobank and translational databases in the BBMRI-NL catalogue; a scenario where it is not desirable to integrate the data by physical import. This pilot was implemented in MOLGENIS and used content from the Dutch CTMM project (CTMM-TraIT) catalog and proved the flexibility of both software and the Common Biorepository Model (CBM) format. A new scalable index built using

ElasticSearch indexing software and a REST API enables federated queries. Databases may also be searched via a convenient Web user interface, where users can perform ‘google’ type free text searches to retrieve merged results from the connected servers. This pilot has proved it is feasible to create fast and powerful federated search across two biobank catalogues created for different purposes. It also positively evaluated use of indexing technologies (ElasticSearch) to scale this federated search towards many biobank catalogues. The REST API was based on the RSQL REST API standard (https://github.com/jirutka/rsql-parser) which is now extensively used throughout all MOLGENIS applications (http://molgenis.github.io/documentation/).

Pilot 3: Leveraging the utility of compound screening functional assays

This pilot implemented a new vocabulary for functional assays, focusing on animal models of diabetes, and a RESTful Web service allowing external users to simply utilize these data. The pilot tackled the scenario where a large corpus of scientific documents exist (higher level functional/phenotypic assay descriptions in ChEMBL) but which are not currently amenable to normalisation, integration and efficient data mining because of a lack of consistent scientific descriptions and a lack of mechanisms to cluster and normalize the corpus (the assay descriptions are currently based on free text). For this pilot, approximately 14,000 textual description of distinct functional assays connected to compounds in ChEMBL classified as Diabetes drugs were curated and classified. Corresponding normalized diabetic animal models including terms, definitions and suggested hierarchy were submitted to the BioAssay Ontology (BAO). The service allows powerful new queries and data to be analysed in multiple ways. The service was tested on a new, large data set connecting drug classes to assay classes. Several scientifically interesting observations have been made, subject to verification by experts in the field. Future developments will ensure new terms are included into the official public release of BAO, and that tables of the terms are accessible via downloads and the ChEMBL interface. These improvements will allow better mining and (BioAssay) integration of ChEMBL data in the future, while the development of a suitable ontology would provide better annotation and integration opportunities for databases with similar data models (for instance EU-OPENSCREEN database, ECBD). These data can be queried via a REST service here: https://www.ebi.ac.uk/chembldb/index.php/ws

Pilot 4: Sharing protein engineering knowledge

A new RESTful interface over PiMS (http://pims.structuralbiology.eu/) was developed; a laboratory information management system that is widely used in recombinant protein production laboratories. This pilot was motivated by the need to support more complex and flexible queries than are currently possible, and to interpret and analyse data in light of other datasets. For this pilot, the Oxford Protein Production Facility (OPPF) agreed to share data from the Protein100 project and contributed 1062478 records for publication (a significant proportion of the known domain data). The new interface represents responses to searches in multiple Semantic Web formats (e.g. RDF/XML, N3), providing an opportunity for machine learning algorithms to make powerful inferences from the experimental data supplied. The development also augments the existing PiMS API, supporting new types of searches within the user interface. Crucially, it provides a step towards query facilities which should scale in the long term to the expected quantity and complexity of future queries and integration scenarios.

Pilot 5: Integrating mouse phenotype data for diabetes research

This pilot developed new RESTful Web services providing a gene-based integration of the Gene Expression Atlas with systemic phenotyped mice data, focusing on the relevant diseases for WP7 (diabetes and obesity). The pilot addresses limitations of EUROPHENOME and IMPC, whose Web user interfaces serve but a portion of the available data, limiting the potential usage and integration scenarios. For this pilot, a Web service was developed for 60 mouse lines analysed by the German Mouse Clinic (GMC). The service provides access to the expert phenotype annotations, allows co-querying of the Gene Expression Atlas for diabetes annotations for the gene related to a specific mouse line, and provides access to the systemic phenotyped experimental raw data. These developments were the foundation for further integration: RDF transformation of mouse data (D4.4/4.6) as well as for the outcomes of the use case for interspecies data mapping (WP7). This work was built upon for D4.6, using the basic REST framework established in D4.3. See also 3.2.2.1, Pilot 3, and Appendix A.3 detailing work done in D4.6.

Pilot 6: Integrating gene and drug information with a clinical trials registry

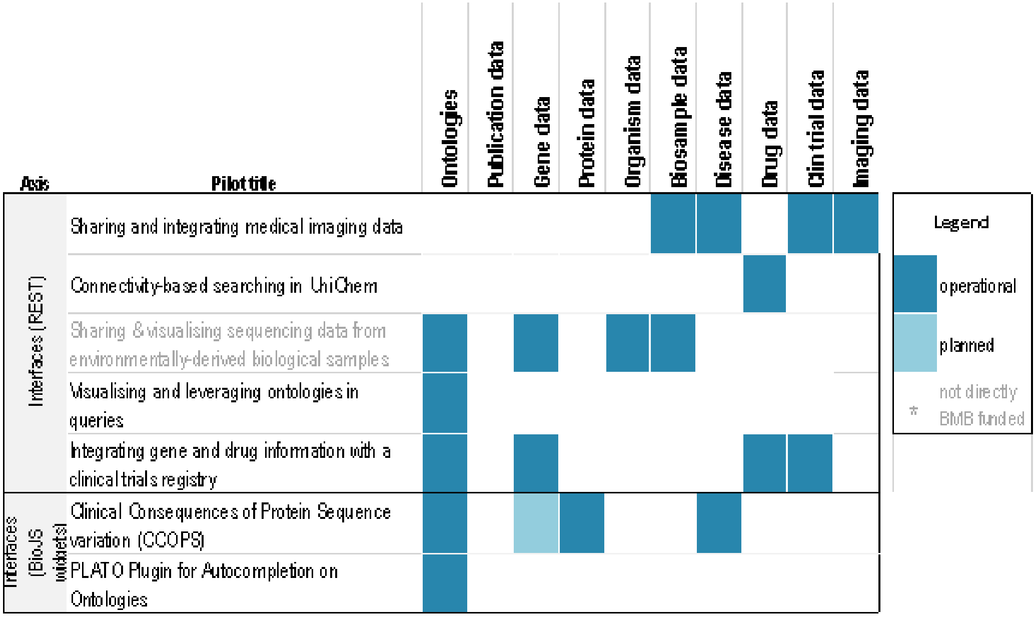
A clinical trials portal was implemented which links publications and information about genes and drugs to data from clinical trials registries. The pilot addresses the dearth of scientifically relevant annotations on clinical phenotypes and aims to provide more insight into the effect of drugs or genes on patients. The strategy is simply to cross-reference the data, then expose it via Web services to queries returning results that link out to these key resources. For this pilot, clinical trials metadata from ClinicalTrials.gov registry were augmented with metadata, to improve the semantics and open the bridge with MESH to publications data from PubMed (D4.6). A new Web service allows the user to do various types of queries over indexes of the data. See also 3.2.2.1, Pilot 1, and Appendix A.1 detailing work done in D4.6.

3.2.3 Overview of visualisation pilots

The five pilot studies selected, focused on the integration and visualisation of data from across the BioMedBridges domains were undertaken. These were built upon early work package deliverables and activities (D4.1-D4.4). The pilots cover a variety of use cases over the following areas:

1. Sharing and integrating medical imaging data
2. Visualising and leveraging ontologies in queries
3. Connectivity-based searching for drugs in UniChem
4. Integrating gene and drug information with a clinical trials registry
5. Sharing and visualising sequencing data from environmentally-derived biological samples

The collection of REST ‘vignettes’ or visualisations presented here provided a foundation for the integration strategy of BioMedBridges, and are available for other integration/visualisation efforts. These were expanded with pilots for Semantic Web integration (D4.6), and were subsequently built upon with a centralised registry for service / data discovery, federated provision of service metadata including provenance, a presentation layer, and service monitoring for example of service availability and usage (D3.3).



Axis	Pilot title	Ontologies	Publication data	Gene data	Protein data	Organism data	Biosample data	Disease data	Drug data	Clin trial data	Imaging data
Interfaces (REST)	Sharing and integrating medical imaging data						■	■		■	■
	Connectivity-based searching in UniChem								■		
	Sharing & visualising sequencing data from environmentally-derived biological samples	■		■		■	■				
	Visualising and leveraging ontologies in queries	■									
	Integrating gene and drug information with a clinical trials registry	■		■					■	■	
Interfaces (BioJS widgets)	Clinical Consequences of Protein Sequence variation (CCOPS)	■		□	■			■			
	PLATO Plugin for Autocompletion on Ontologies	■									

Legend

■ operational

□ planned

* not directly BMB funded

Figure 9 ‘Visualisation’ pilot data touching points

Wherever possible and valuable, the BioJS framework was used. BioJS[18] is an open-source project whose main objective is the visualization of biological data in JavaScript. BioJS provides an easy-to-use consistent framework for bioinformatics application programmers. It follows a community-driven standard specification that includes a collection of components purposely designed to require a very simple configuration and installation. In addition to the programming framework, BioJS provides a centralized repository of components available for reutilization by the bioinformatics community. Over twenty institutions are involved in the BioJS project, whether as committers or adopters.

Each of the pilot REST vignettes developed here were iteratively refined according to ongoing user feedback and testing. Some of the work done also contributed to the Semantic Web Pilot project (D4.6), leading to greater

[18] Gómez J, García LJ, Salazar GA, Villaveces J, Gore S, García A, Martín MJ, Launay G, Alcántara R, Del-Toro N, Dumousseau M, Orchard S, Velankar S, Hermjakob H, Zong C, Ping P, Corpas M, Jiménez RC. BioJS: an open source JavaScript framework for biological data visualization. Bioinformatics. 2013 Apr 15;29(8):1103-4.

semantic alignment of the data, and facilitating tighter integration across databases and resources. The adoption of content standards in this domain and subsequent mapping of these integrates this deliverable with D3.2 and the ontology widget can be used to deliver semantic integration results from D3.4.

3.2.3.1 Summary of visualisation pilots

This section provides executive summaries of the work undertaken as part of D4.5, and was built upon in subsequent deliverables. Further information on these pilots is available in D4.5. An additional activity is also reported here, conducted as part of D4.8, concerned with visualisation of ontologies through OLS

Pilot 1: Sharing and integrating medical imaging data, ERASMUS MC (Euro-BioImaging)

XNAT is not only the most widely-used informatics platform for imaging research, it is also open-source and highly extensible. Building on XNAT in accordance with the BioMedBridges WP4 objectives, Erasmus MC has developed and deployed REST web services that allow other applications (both web-services and desktop applications) to embed information/data from medical imaging databases. Via these new web-services, three large datasets have been made available to a wide audience for research purposes, and, in collaboration with EATRIS/BBMRI, the web-services were also applied in a multi-centre clinical study on imaging atherosclerosis[19]. Besides just developing the web-services for these specific imaging datasets, a generic script was made available for deploying the web-service for any imaging dataset, thereby enabling other researchers, technicians, and companies to offer similar services to the community. The availability of the datasets has been registered in the BioSamples database, and the availability of the new script has been registered in the “XNAT Marketplace” (http://marketplace.xnat.org/). Web Services and

[19] Truijman, M. T. B., Kooi, M. E., van Dijk, A. C., de Rotte, A. A. J., van der Kolk, A. G., Liem, M. I., Schreuder, F. H. B. M., Boersma, E., Mess, W. H., van Oostenbrugge, R. J., Koudstaal, P. J., Kappelle, L. J., Nederkoorn, P. J., Nederveen, A. J., Hendrikse, J., van der Steen, A. F. W., Daemen, M. J. A. P. and van der Lugt, A. (2013), Plaque At RISK (PARISK): prospective multicenter study to improve diagnosis of high-risk carotid plaques. International Journal of Stroke. doi: 10.1111/ijs.12167

explanatory documentation can be found here: http://xnat.bigr.nl. See also 3.2.1.2, Pilot 1, and Appendix B.1 detailing work done in D4.5 and D4.6.

Pilot 2: Connectivity-Based Searching in UniChem - EMBL (EU OPENSCREEN)

Unichem (https://www.ebi.ac.uk/unichem/) serves as the chemistry data integration layer for EU-OPENSCREEN, ChEMBL, ChEBI and other chemically aware databases. UniChem achieves this integration by mapping identifiers from these different resources to their corresponding standard InChI identifier. However, different resources may depict the same molecule in a slightly different way (e.g.: with different stereochemistry), and at different levels of granularity. EU Openscreen have developed a user-friendly web interface for UniChem's refined "Connectivity-based searching" functionality which overcomes differences in chemical nomenclature across databases. This work builds on the REST web service developed in D4.3, and will further serve as a bridge to align diverse resources according to the chemical compounds they reference. A web application for non-programmatic access and REST service methods for programmatic access have been developed, and have now been released and accessible outside of the EBI (https://www.ebi.ac.uk/unichem/info/widesearchInfo), and a full publication now exists describing this service. The web interface was not developed into a BioJS widget because it was clear from user feedback that the existing web interface and REST web services were better suited to their needs. However, a BioJS widget remains an option if users request this in future. UniChem Connectivity maintenance has secure funding through a ChEMBL award from the Wellcome Trust till approx. 2019.

Pilot 3: Sharing and visualising sequencing data from environmentally-derived biological samples: In-Kind contribution from SZN/EMBL-EBI (EMBRC) and MPI-Bremen (Micro B3)

The European Marine Biological Resource Centre provides access to marine organisms, techniques, and data to the scientific community at large, including universities and industry. Accordingly, EMBRC contributed in-kind to BMB D4.5 by a) submitting to the European Nucleotide Archive (ENA) sequencing data from marine-derived bacteria and by b) extending their geologic mapping widget to make it maximally useful to other research groups. The megx.net (http://www.megx.net/) portal for Marine Ecological GenomiX is a web site for specialized georeferenced databases and tools for the analysis of marine bacterial, archaeal, and phage genomes and metagenomes. One of the key elements of megx.net is the Genes Mapserver (originally developed within the frame of the EU-funded project MetaFunctions and currently further developed within in the frame of the EU-funded project Micro B3 http://www.microb3.eu/), which facilitates the interpretation of the sequence in its environmental context via a browsable world map. In accordance with WP4 objectives, Megx.net was enhanced in two important ways. It has been developed as an embeddable javascript widget that is going to conform to BioJS specification, and it now accepts an array of ENA accession numbers and renders them as pins on a layered geologic map. A beta version of the widget, populated with ENA data has been deployed at (http://mb3is.megx.net/megx-embrc-widget.html) and has been submitted for consideration by the BioJS project (http://www.ebi.ac.uk/Tools/biojs/registry/). This widget is of general use and can be embedded in any web site to render any ENA accessions, and can be extended in future to support a set of identifiers from any relevant service.

Pilot 4: Visualising and leveraging ontologies in queries - EMBL-EBI (Euro-BioImaging)

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 (http://www.ebi.ac.uk/ols). This widget was developed to be generically configurable for use with multiple ontologies. For demonstration and evaluation purposes, it has been deployed for use with three different ontologies: the

Cellular Microscopy Phenotype Ontology (CMPO), the Experimental Factor Ontology (EFO), and EDAM (Embrace Data and Methods), each of which is being actively used within BioMedBridges. Features of the widget include the ability to visualise matching ontology terms in their rightful place in the tree, and to expand and collapse nodes as desired. To maximise its reuse, the widget conforms to BioJS specification and has been committed to the BioJS project (http://www.ebi.ac.uk/Tools/biojs/registry/). See also 3.2.3.1, BioJS pilots (PLATO), and Appendix C.2 detailing work done in D4.5 and D4.6.

Pilot 5: Integrating gene and drug information with a clinical trials registry - UDUS (ECRIN)

Building on the foundation of the REST interface developed in D4.3, a web application was developed to facilitate the discovery of clinical trials for Acute Myeloid Leukemia. This development was informed by the personalised medicine use case (WP8) and its stated goal of searching trials by gene and by drug while providing links to the corresponding publications.

This work was further developed in D4.6 in which text mining and semantic alignment are used in order to enable the user to search trials according to the compounds and genes they reference. Those features will open the bridge to other biomedical infrastructures of the BioMedBridges project. See also 3.2.1.1, Pilot 1, and Appendix A.1 detailing work done in D4.6, and D4.3.

3.2.3.2 Integration of simple object queries using BioJS widgets

Two BioJS widget pilots were undertaken as part of D4.6: The Plugin for Autocomplete on Ontologies (PLATO), and Clinical Consequences of Protein Sequence variation (CCoPS). PLATO is a multipurpose widget that is particularly useful 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.

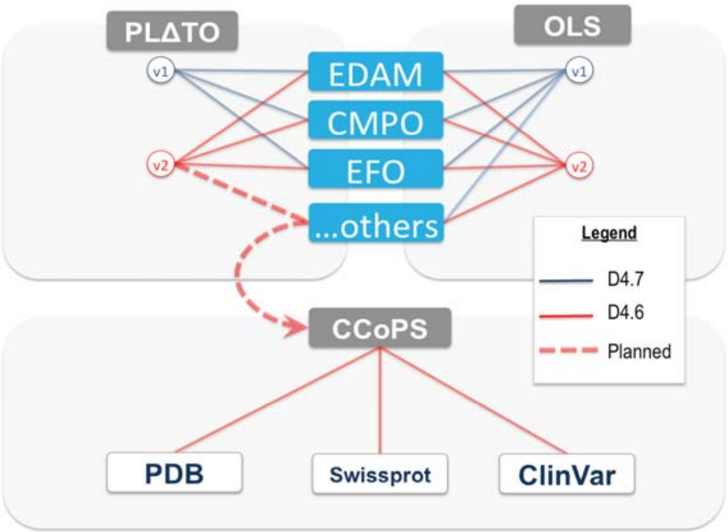


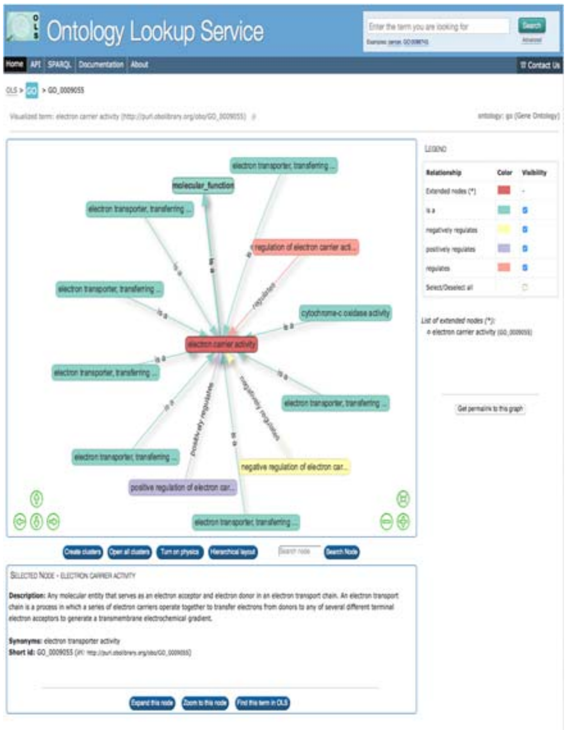
Figure 10 Summary of BioJS widget-driven pilots

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 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 for details). CCoPS integrates two web services: 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

3.2.3.3 Ontology visualisation

The javascript visualisation that was developed as part of the WP offers a new and innovative way to explore a given ontology. Its primary purpose is to offer an alternative viewpoint for ontologies stored in EBI's new Ontology lookup service[20]. A generic approach was necessary to process the different ontologies and their structure - more than 100 diverse ontologies are stored in OLS and it is constantly growing. The new visualisation is interactive, enabling users to dynamically expand the number of nodes, zoom in/out, search for nodes, change the display layout, or hide specific relationships. Besides allowing exploration of an ontology and retrieval of term information, the tool offers a convenient way to take screenshots for presentations.

[20] http://www.ebi.ac.uk/ols/beta/



Visualized term: electron carrier activity (http://purl.obolibrary.org/obo/GO_0009055)

ontology: go (Gene Ontology)

LEGEND

Relationship	Color	Visibility
Extended nodes (*)	🟥	-
is a	🟩	☑
negatively regulates	🟨	☑
positively regulates	🟪	☑
regulates	🟥	☑
Select/Deselect all		□

List of extended nodes (*):
- electron carrier activity (GO_0009055)

SELECTED NODE - ELECTRON CARRIER ACTIVITY

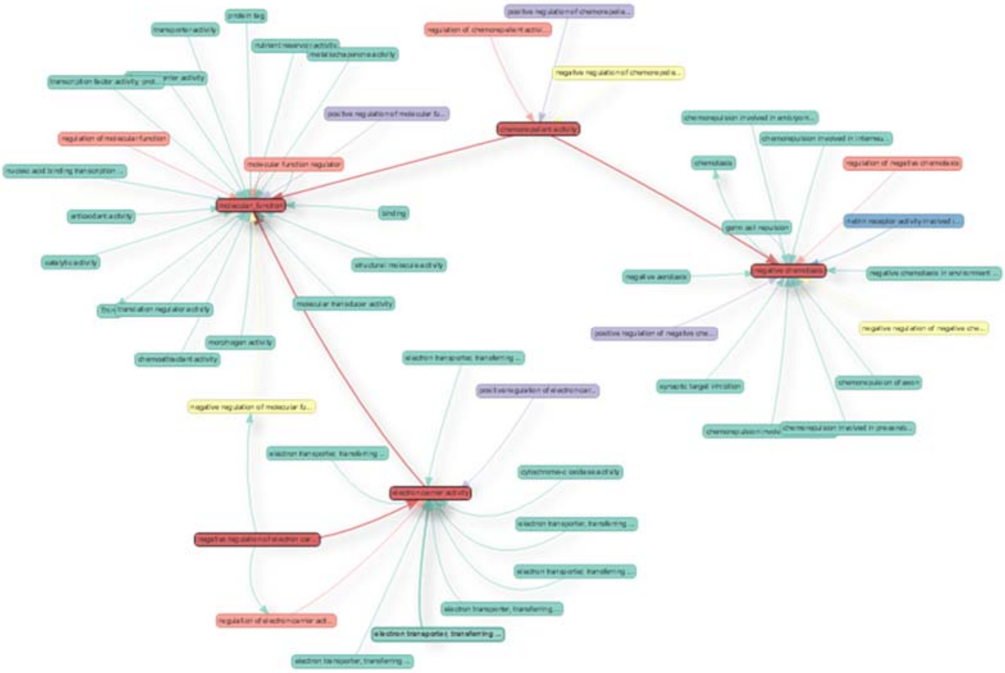
Description: Any molecular entity that serves as an electron acceptor and electron donor in an electron transport chain. An electron transport chain is a process in which a series of electron carriers operate together to transfer electrons from donors to any of several different terminal electron acceptors to generate a transmembrane electrochemical gradient.

Synonyms: electron transporter activity
Short id: GO_0009055 (IRI: http://purl.obolibrary.org/obo/GO_0009055)

Screenshot of the visualisation plugin embedded in the new OLS lookup Service

The visualisation is heavily based on the javascript framework visjs[21], an open source network visualisation framework. The functionality of this framework was extended to satisfy the requirements needed for ontology visualisation. The tool was developed as a standalone plugin and is available to the community through public repositories (e.g. github). In the near future, the ontology visualisation might be incorporated within other EBI resources (e.g. CTTV, GO Browser), or external to the EBI domain. To support the different use cases and designs of alternative websites, the plugin offers a multitude of easily adjustable options such as size of the widgets, colour scheme, and the ability to modify the parameters of the physical engine being used. Other more sophisticated behaviours can also be modified, for instance it is possible to override event callback functions.

[21] http://visjs.org/


protein tag
transporter activity
nutrient reservoir activity
metallochaperone activity
transcription factor activity, prot...
regulation of molecular function
positive regulation of molecular fu...
molecular function regulator
nucleic acid binding transcription ...
molecular function
binding
antioxidant activity
catalytic activity
structural molecule activity
translation regulator activity
molecular transducer activity
morphogen activity
chemoattractant activity
negative regulation of molecular fu...
positive regulation of chemorepella...
regulation of chemorepellent activi...
negative regulation of chemorepella...
chemorepellent activity
chemorepulsion involved in embryoni...
chemorepulsion involved in interneu...
chemotaxis
regulation of negative chemotaxis
germ cell repulsion
netrin receptor activity involved i...
negative aerotaxis
negative chemotaxis
negative chemotaxis in environment ...
positive regulation of negative che...
negative regulation of negative che...
synaptic target inhibition
chemorepulsion of axon
chemorepulsion involved in presenti...
electron transporter, transferring ...
positive regulation of electron car...
cytochrome-c oxidase activity
electron carrier activity
negative regulation of electron car...
regulation of electron carrier act...

Screenshot of the javascript plugin with a couple of expanded nodes, leading to a network of terms displaying a part of the structure of an ontology

3.3 Sustainability of pilots

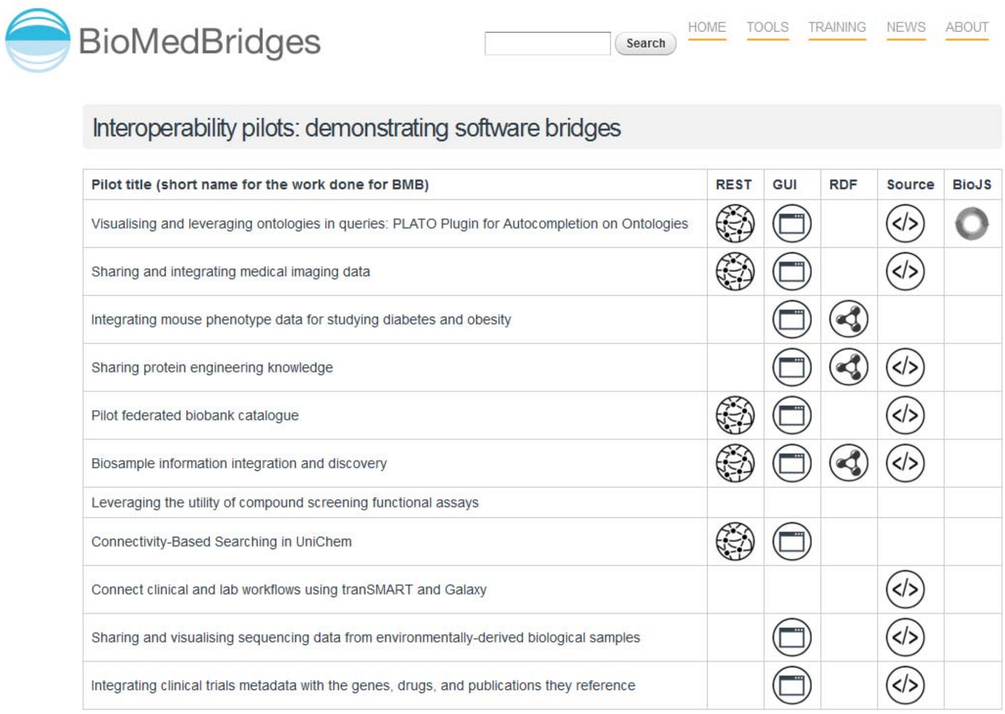
Over the course of WP4, a variety of pilots were undertaken, each of which was directly in response to the urgent needs of least one BioMedBridges partner. Each pilot and use case were iteratively refined over the course of the work, in response to both intra- and inter-pilot (and intra- /inter-institutional) feedback, ensuring they would be optimally fit for purpose. These premeditated considerations, as well as the use of standard technologies and open standards and (largely) open data in implementation, will minimise the effort required in their maintenance by the host partner/institute.

During WP4, working implementations of nine pilot integration projects have been delivered. The level at which integration is implemented was dependent upon early exploratory work and subsequent refinement of project and use case. Such work suggested that RDF is well suited to some use cases and that all implementation technology choices had pros and cons. However, ontologies and identifiers have proven to be important considerations, no matter what the technology choices.

The sustainability of the pilots in the medium and long term is indicated above (see section 3.2 and pilot subsections therein), in a sustainability statement from each partner. Each of the integration pilots is currently accessible through a combination of technological and user centric interfaces, and is available for the community to exploit further by building upon.

To facilitate discoverability of the interfaces to these pilot projects, a summary website has been provided (Figure 11): http://www.biomedbridges.eu/interoperability-pilots-demonstrating-software-bridges

This matrix provides information on the most advanced pilots. See the individual pilot reports above (and Appendices) for specific future work on each pilot.



Interoperability pilots: demonstrating software bridges

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 11 BioMedBridges software matrix. The most refined final pilots (rows) are listed with the interfaces that are available for each (columns). Each cell is linked to a web location where the REST web service endpoints, visual web interface (GUI), RDF SPARQL endpoint, source code, and BioJS documentation can be found, respectively

3.4 Overall conclusions and future work

To continue the long tradition of European excellence is research and innovation, it is crucial that there is closer communication and better knowledge exchange between Research Infrastructures.

BioMedBridges was tasked with building and facilitating such processes, allowing bridges to be built across RIs. It has been a complex project, crossing scientific domains, connecting large RIs and navigating solutions through workflows based on different technologies, software and expertise. WP3 was tasked with exploration and recommendation of standards and semantics, while WP4 was tasked with integration through an informed technological implementation. One of the most important outcomes of this work are the lessons learnt, which will inform future efforts in cross-RI communication/exchange and provide feedback to existing RIs on how they can better facilitate access to their data.

3.4.1 Lessons learnt

From an overarching perspective, many of the learnings detailed below can be assigned to one of three types of barrier: scientific (domain knowledge associated), sociological (community/attitude associated) and technological (software/tool/expertise associated). Some examples of the differentia encountered include that each RI deals with different domain knowledge, which may comply with or use different standards, be represented (in knowledge modelling terms) in different ways, and may deal with open or restricted data sets, and be made available (or not) through different formats, using open or proprietary software and tools, employing one of a plethora of interfaces to the data.

Some general lessons learnt on technology:

- technological solution for a particular use case defined after potential solutions evaluated
- no one technology works for all use cases
- ‘evaluating fitness for purpose’ of technology is important

- technological advancement and long term impact should be part of project plans
- personnel and expertise matter

Some general lessons learnt on data:

- interoperability is different in each use case, its meaning must be defined in each case
- data to support use case needed up front
- revise/revisit use case based on data available, revise use case if necessary
- try to use open data for prototyping
- crediting sources matter (sociological / scientific impact)

Some general lessons learnt on engagement and buy in:

- consider carefully if a new tool is needed: extend existing tool or inclusion in established tool set gains immediate user base
- user centric design, user testing are important, and need to be repeated frequently
- tools need to be reliable (accessible, consistent)
- provide training material, courses and documentation

It was clear during this work that more lessons were to be learnt about emerging or developing standards and technologies (RDF/SPARQL), than for well-established ones (REST). More specific lessons are presented below, with links to further documentation as necessary. They are categorised by the general integration axis used in the implementation, but do not include REST interfaces, as these involve well established procedures, standards and interfaces.

3.4.1.1 RDF knowledge representation

Over the course of WP4, given the extensive integration efforts that were RDF-centric, a number of lessons have been learnt which will be informative to future works in this area. As a general overview, the lessons below can be envisaged as steps in a pathway to linked data, where each incrementally closes on the objective of ‘Linked Data’:

- collect use cases
- define scope of work
- develop and publish an ontology (schema) to describe the data
- link your data items to other data sets using URIs and define the URLs carefully
- may your data using community standard ontologies and semantic web ontologies
- provide your data as RDF
- expose your RDF through a SPARQL endpoint
- publish information about what your data is about, how often released, license information, etc. (metadata)
- have considered sustainability throughout and at each stage of the project
- model biology and not data

Lessons were reviewed at a recent RDF training course (EBI industry workshop; materials available in Appendix A.5) where participants voiced significant interest in the practical experience in dealing with the RDF platform and what recommendations we could offer to implementers and users. As a follow up, these experiences, as well as those of other RDF/linked data providers such as Open PHACTS, were shared during the BioMedBridges Symposium (Appendix A.6). This information is being collated currently to generate a ‘white paper’ on RDF best practices and will be submitted in 2016. RDF is not always an appropriate solution, as exemplified by the CTIM pilot (see 3.2.1.1 Pilot 1). Specifically, while originally envisaged as an RDF integration, it was decided following further evaluation and feedback, that RDF was not ideally suited in this instance. RDF Linked Data technologies are optimised for the gathering of links and locations to RDF to provide further data and information. Large datasets, such as ClinicalTrials.gov converted to RDF were found to be non-performant for the desired analysis/integration tasks required.

3.4.1.2 tranSMART integration

In dealing with ethically sensitive data (see pilots in section 3.2.1.2), the solutions implemented centred on the use of tranSMART, an open source knowledge management platform for translational medicine. It is used extensively by CTMM-TraIT (Netherlands translational research informatics program http://www.ctmm-trait.nl/) and the Innovative Medicines Initiative (http://www.imi.europa.eu/).

Partner feedback on lessons learnt includes:

- Building bridges between existing systems allows for greater use as well as re-use. It does often require a high level of knowledge of the systems being bridged: which functionality is already available and how can missing functionality best be added (to ensure adoption by the development community of an existing system).
- Test data is often difficult to get and to share with collaborators. CTMM TraIT has tackled this issue by creating a test dataset based on commonly available cell lines that can be shared without restrictions. This test dataset and others like it will be useful in future integration efforts.
- Besides offering a web service, provide scripts that automate the deployment of the web service, allowing others to offer the same web service. This improves sustainability of the work and adoption by the community. This is especially relevant if it is unlikely that a single instance of the web service will serve the needs of the entire community, and when a distributed system of web services is more appropriate. As an example, deliverable D4.5 (http://dx.doi.org/10.5281/zenodo.14119) describes the Puppet scripts that facilitate installation of the XNAT imaging platform.

3.4.1.3 Widgets

Widgets potentially provide user friendly interfaces, built upon existing services (of any type) from a resource, according to a standardised procedure. There are competing ‘standards’[22] and repositories for such widgets, though this project employed BioJS. Lessons focus on issues that are generated by the data provider, rather than the BioJS technical solution:

[22] http://www.w3.org/TR/2013/REC-widgets-apis-20131031/

- service issues at source reflected in widget e.g. identifiers used in dataset not persistently resolvable
- documentation not up to date
- failure to update a Last-Modified header (often present in human-readable form only)
- many services are format specific, with no machine-readable representation
- many formats not valid by their own specification (e.g. malformed XML)
- inconsistent behaviour with CORS (Access-Control-Allow-Origin)
- help desk support highly variable, but consistently slow

3.4.2 Challenges beyond BioMedBridges

Many valuable clinical datasets/studies remain undiscoverable because there is not a suitable repository within which to register them. Existing "long-tail" repositories such as Dryad and Zenodo are for depositing data rather than registering it per se. There is a requirement to bridge between literature, supplemental materials and unstructured data repositories to archives and knowledge bases which will allow better data discovery, improved provenance and data access. 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: Many of these will be addressed in future projects.

Best practices for data services: In the process of developing CCoPS and for other web service-related efforts in BMB more broadly, we encountered many challenges, some of which were related to the provision of identifiers; we recently submitted a manuscript (https://zenodo.org/record/18003) that speaks to those. However, we encountered several other problems that extend beyond that 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 impactful 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 metadata repositories such as the newly launched "BioStudies" platform, from the EBI. For BMB 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 the awareness of partners to a higher level with respect to data integration readiness so that, in the future, pre-clinical findings can more readily provide biological insights.

EUDAT infrastructure offers the use of important data services and a Collaborative Infrastructure. EUDAT has developed a number of core services, which can be used as a generic interchange layer for more specific services developed in different EU projects, like BioMedBridges. All services should be embedded into the Collaborative Data Infrastructure (CDI) of EUDAT. Following services are offered:

1. B2SAFE, Replicate research data safely
2. B2STAGE, Get data to computation, transferring data to high-performance computing
3. B2SHARE, Store, share and publish research data, for publishing / storing of smaller data sets
4. B2FIND, metadata harvesting and cataloguing, find research data
5. B2DROP, Synchronisation and exchange of research data, exchanging data with colleagues

In addition, a set of EUDAT core operational services that are essential for the management of the CDI have been defined:

PID (Persistent Identification), register handles to data objects and retrieve data objects; RCT (Resource Coordination Tool), allows community managers and service providers to advertise their resources and to make resource requests and queries; SSR Site and Service Registry, contains information about the EUDAT sites and services; Monitoring, a service providing health information for the EUDAT resources; B2HOST, a Service Hosting Framework which allows communities to deploy and operate their own applications and data-oriented services; Authentication, Authorization and Identity layer for data security and data access control.

Joining EUDAT requires the installation and configuration of a minimum software stack (the EUDAT CDI). Furthermore, it requires the community to dedicate some resources (storage, compute and man-power) to the project. But to use EUDAT services and embed own services into the EUDAT infrastructure will further data sharing, collaboration and data security enormously, because all services operate as part of a single high-quality structure.

4 Delivery and schedule

The delivery is delayed: ☐ Yes ☑ No

5 Adjustments made

No adjustments were made to the deliverable.

6 Background information

This deliverable relates to WP 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, STFC, UDUS, FVB, TUM-MED, ErasmusMC, HMGU, VU-VUMC

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 WP 11 will provide recommendations for other emerging technologies to use, aiming for the semantically richest integration.

Work package number	WP4	**Start date or starting event:**	month 1
Work package title	Technical Integration		
Activity Type	RTD		

Participant number	1:EMBL	4:STFC	5:UDUS	6:FVB	7:TUM-MED	9:ErasmusMC	11:HMGU	13:VUMC
Person-months per participant	69	40	38	0	37	15	32	3 7

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 WebServices interfaces optimised for browsing information
3. Expose the integration via use of REST based WebServices 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-BioImaging, 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-OPENSCREEN, ECRIN, Euro-BioImaging, 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. Scaleable 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.

Appendix A: RDF pilots

A.1 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 keyword (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 employing an 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.

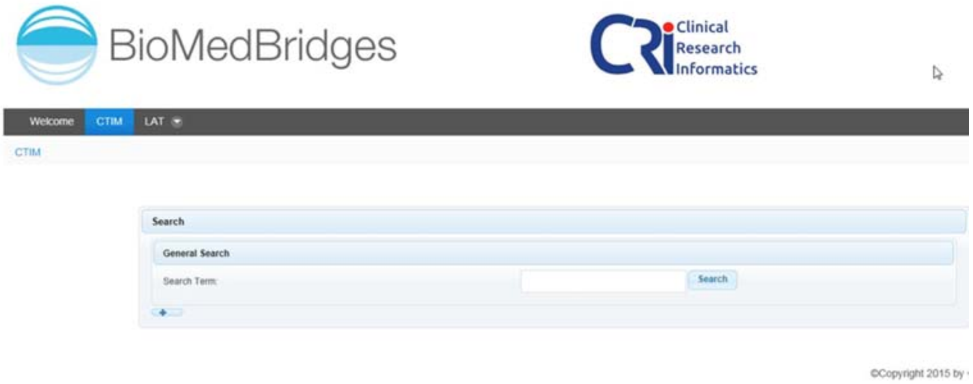


Figure A1 Displayed is the simple user interface of CTIM with a single search field. As an option, expert search is also provided

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 Solr as database was established; it is a NoSQL enterprise search server on the basis of the Apache Lucene back-end. Solr is known to be highly scalable and one of the most popular open source enterprise search engines. Solr is used to search for clinical trial data and was therefore filled with all study fields and result fields from ClinicalTrials.gov. Solr also automatically indexed all data and maps field names to the appropriate values. In this way, it becomes possible to search for all fields or can decide to search just for specific ones. Search can be done through all textual blocks of the clinical trial data, which is a feature not offered by ClinicalTrials.gov's own search functionality. The data core for clinical trials got expanded so that now all trials from clinicaltrials.gov are involved in CTIM searches. Furthermore the web services for BioSample and PubMed were implemented as additional search interfaces (Fig. A2). Additionally further filtering possibilities were developed and a user friendly interface created (Fig. A3).

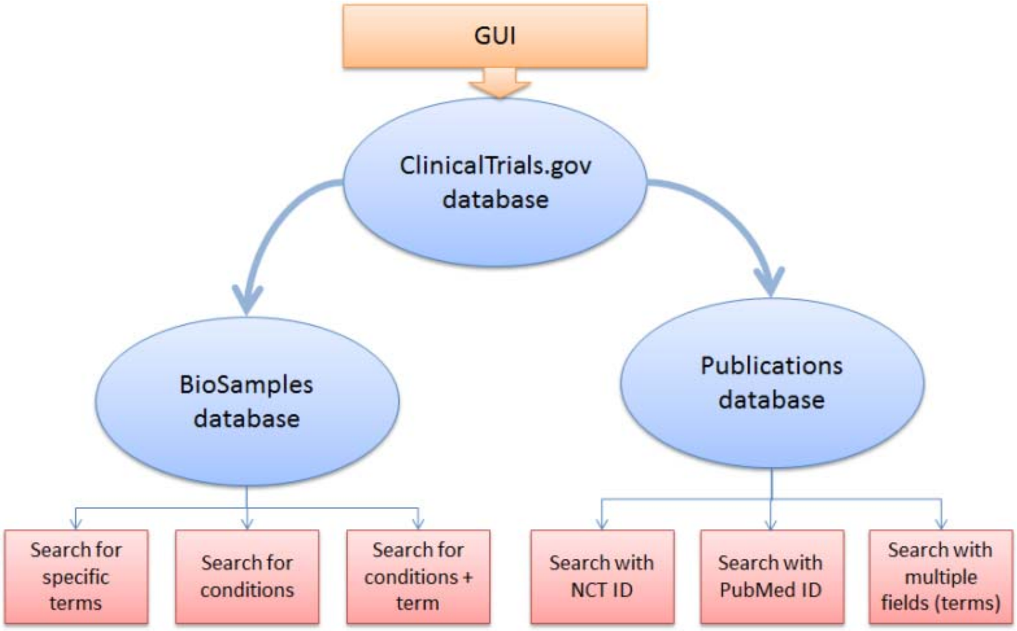

GUI
ClinicalTrials.gov database
BioSamples database
Publications database
Search for specific terms
Search for conditions
Search for conditions + term
Search with NCT ID
Search with PubMed ID
Search with multiple fields (terms)

Figure A2 CTIM workflow to connect the different data sources and the kind of data search and linking methods. GUI=user interface (see Fig. A1), NCT=Clinical trials ID

RDF and clinical trials

The RDF data format is primarily build to genuinely identify contents / entities of digital information. The data stored in a RDF-model is metadata concerning this one identifiable entity, which can be any form of electronically stored data (e.g. web page, user, locations, multimedia files, document files, biosample data, etc.). In the context of clinical trials two approaches are of relevance: LinkedCT and Bio2RDF.

LinkedCT is the RDF version of the ClinicalTrials.gov database; it represents an open semantic web data source for clinical trials data. The data of LinkedCT is generated by transforming the existing data source of clinical trials into RDF. Links exist between records of the trials and several other data sources, such as Bio2RDF or PubMed. Several demonstrators and prototypes have been developed; covering links between LinkedCT and IARC TP database, a pacemaker register and a WHO health observatory dataset. This approach is primarily based on a ontology based semantic matching method, but can also be combined with techniques of string matching. CTIM does not use semantic web technologies, but a flexible and quick string search method.

The Bio2RDF has its main focus on biological and pharmaceutical data. The Dumontier Lab has proven that it can be quite easy to unite data using semantic web technologies. They provide a large number of data including PubMed as SPARQL-Endpoints. Bio2RDF however does not provide a research-based user interface. Though one can access the data stored and see the attribute-based links, but for an untrained researcher will have difficulties successfully querying Bio2RDF.

CTIM links with the biosamples database BioSamples. The BioSamples database aggregates sample information for reference samples and samples for which data exist in one of the EBI's assay databases, such as ArrayExpress, the European Nucleotide Archive or Proteomics Identificates Database. It is an effective resource to store information about samples that are referred by data in multiple repositories. The RDF version of the Biosamples data reflects the data that can be accessed by end user from the web interface and its structure is largely influenced by the SampleTab submission format. Making Biosamples available in RDF gives advantages like the possibility of exploiting the ontologies and a method to combine data from multiple repositories. The SPARQL endpoint is the main interface to the Biosamples RDF dataset.

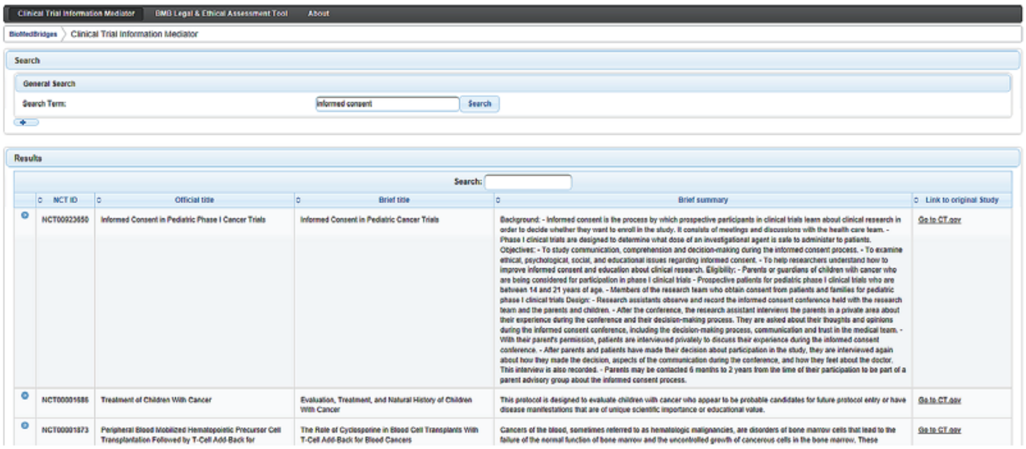
A RDF about one biosample can include the information about the form the data of the biosample is stored (e.g. picture, chip-analysis or other forms experimental data), the biological origin, location where the biosample was taken and examined, when the biosample was taken and more.

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.


Clinical Trial Information Mediator
BMB Legal & Ethical Assessment Tool
About
BioMedBridges
Clinical Trial Information Mediator
Search
General Search
Search
Results
Search:

Search Term: informed consent

	NCT ID	Official title	Brief title	Brief summary	Link to original Study
●	NCT00923650	Informed Consent in Pediatric Phase I Cancer Trials	Informed Consent in Pediatric Cancer Trials	Background: - Informed consent is the process by which prospective participants in clinical trials learn about clinical research in order to decide whether they want to enroll in the study. It consists of meetings and discussions with the health care team. - Phase I clinical trials are designed to determine what dose of an investigational agent is safe to administer to patients. Objectives: - To study communication, comprehension and decision-making during the informed consent process. - To examine ethical, psychological, social, and educational issues regarding informed consent. - To help researchers understand how to improve informed consent and education about clinical research. Eligibility: - Parents or guardians of children with cancer who are being considered for participation in phase I clinical trials - Prospective patients for pediatric phase I clinical trials who are between 14 and 21 years of age. - Members of the research team who obtain consent from patients and families for pediatric phase I clinical trials Design: - Research assistants observe and record the informed consent conference held with the research team and the parents and children. - After the conference, the research assistant interviews the parents in a private area about their experience during the conference and their decision-making process. They are asked about their thoughts and opinions during the informed consent conference, including the decision-making process, communication and trust in the medical team. - With their parent's permission, patients are interviewed privately to discuss their experience during the informed consent conference. - After parents and patients have made their decision about participation in the study, they are interviewed again about how they made the decision, aspects of the communication during the conference, and how they feel about the doctor. This interview is also recorded. - Parents may be contacted 6 months to 2 years from the time of their participation to be part of a parent advisory group about the informed consent process.	Go to CT.gov
●	NCT00001586	Treatment of Children With Cancer	Evaluation, Treatment, and Natural History of Children With Cancer	This protocol is designed to evaluate children with cancer who appear to be probable candidates for future protocol entry or have disease manifestations that are of unique scientific importance or educational value.	Go to CT.gov
●	NCT00001873	Peripheral Blood Mobilized Hematopoietic Precursor Cell Transplantation Followed by T-Cell Add-Back for	The Role of Cyclosporine in Blood Cell Transplants With T-Cell Add-Back for Blood Cancers	Cancers of the blood, sometimes referred to as hematologic malignancies, are disorders of bone marrow cells that lead to the failure of the normal function of bone marrow and the uncontrolled growth of cancerous cells in the bone marrow. These	Go to CT.gov

Figure A3 CTIM results of a search for clinical trials

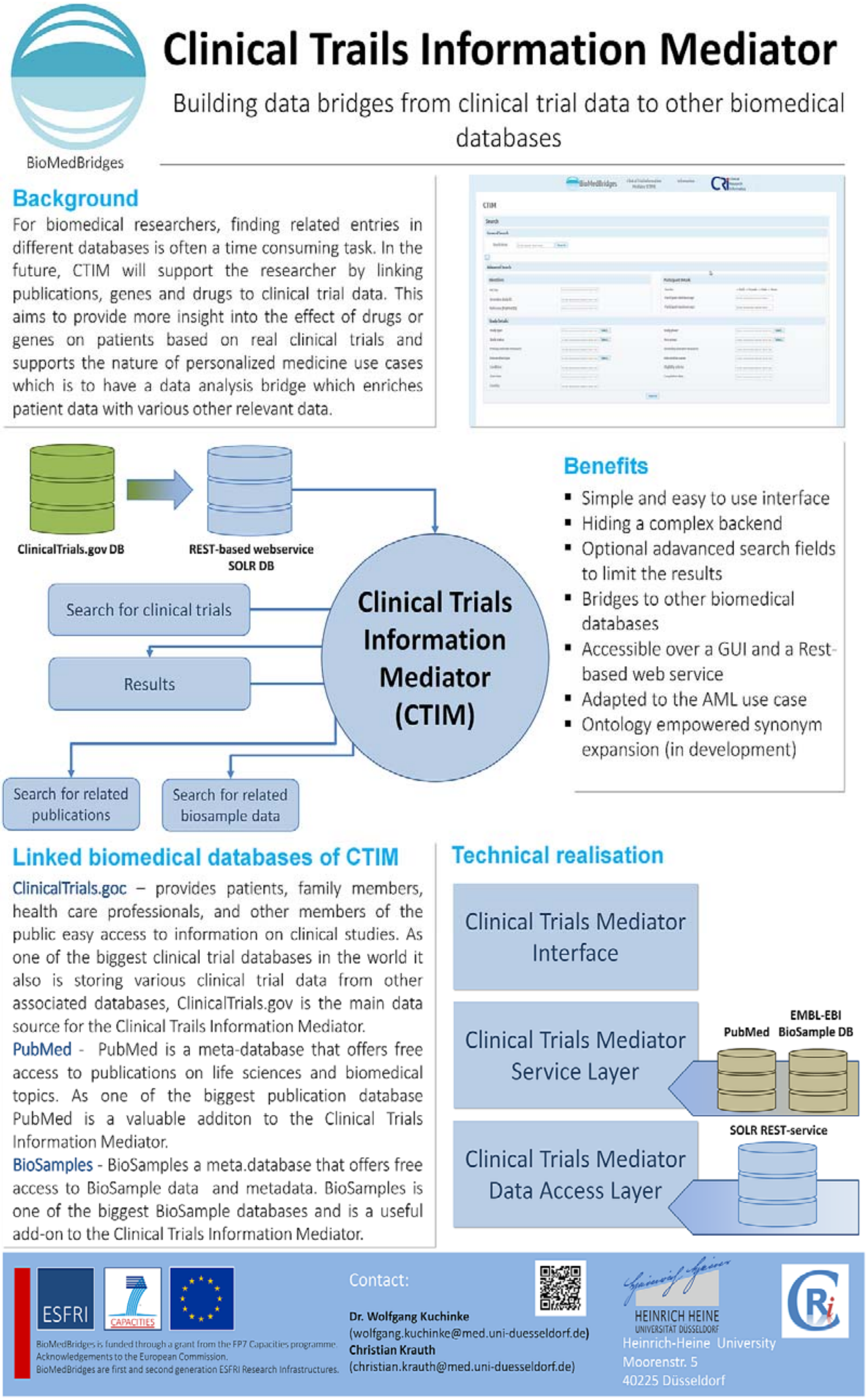
Because CTIM offers basic information to web content, like BioSamples, and the link to the source of this information, we found that not RDF but the web service of BioSample offered the better solution for CTIM. The reason for this decision was that any querying by employing web services is far easier and faster that a query method based on RDF. 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 implemented to feed CTIM searches with additional information regarding clinical trials (Fig. A4).

Technical implementation:

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 BMB project. In the end, it would be worthwhile to extend CTIM with additional databases, e.g. ArrayExpress, Genbank or DrugBank. An extension of data sources would be advantageous but should not compromise the easy usability of the tool.



Clinical Trails Information Mediator

Building data bridges from clinical trial data to other biomedical databases

BioMedBridges

Background

For biomedical researchers, finding related entries in different databases is often a time consuming task. In the future, CTIM will support the researcher by linking publications, genes and drugs to clinical trial data. This aims to provide more insight into the effect of drugs or genes on patients based on real clinical trials and supports the nature of personalized medicine use cases which is to have a data analysis bridge which enriches patient data with various other relevant data.

ClinicalTrials.gov DB
REST-based webservice SOLR DB
Search for clinical trials
Results
Clinical Trials Information Mediator (CTIM)
Search for related publications
Search for related biosample data

Benefits

- Simple and easy to use interface
- Hiding a complex backend
- Optional adavanced search fields to limit the results
- Bridges to other biomedical databases
- Accessible over a GUI and a Rest-based web service
- Adapted to the AML use case
- Ontology empowered synonym expansion (in development)

Linked biomedical databases of CTIM

ClinicalTrials.goc – provides patients, family members, health care professionals, and other members of the public easy access to information on clinical studies. As one of the biggest clinical trial databases in the world it also is storing various clinical trial data from other associated databases, ClinicalTrials.gov is the main data source for the Clinical Trails Information Mediator.

PubMed - PubMed is a meta-database that offers free access to publications on life sciences and biomedical topics. As one of the biggest publication database PubMed is a valuable additon to the Clinical Trials Information Mediator.

BioSamples - BioSamples a meta.database that offers free access to BioSample data and metadata. BioSamples is one of the biggest BioSample databases and is a useful add-on to the Clinical Trials Information Mediator.

Technical realisation

Clinical Trials Mediator Interface
Clinical Trials Mediator Service Layer
Clinical Trials Mediator Data Access Layer
PubMed
EMBL-EBI BioSample DB
SOLR REST-service

ESFRI
CAPACITIES
BioMedBridges is funded through a grant from the FP7 Capacities programme.
Acknowledgements to the European Commission.
BioMedBridges are first and second generation ESFRI Research Infrastructures.

Contact:
Dr. Wolfgang Kuchinke
(wolfgang.kuchinke@med.uni-duesseldorf.de)
Christian Krauth
(christian.krauth@med.uni-duesseldorf.de)

HEINRICH HEINE
UNIVERSITÄT DÜSSELDORF
Heinrich-Heine University
Moorenstr. 5
40225 Düsseldorf

Figure A4 Poster explaining the architecture and use of CTIM

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

BBMRI RDF Background:

The BBMRI.eu catalogue (https://www.bbmriportal.eu/bbmri/) provides an overview of the human tissue biobank landscape across Europe. BBMRI-LPC uses an advanced version of this catalogue (https://www.bbmriportal.eu/lpc/) for the Large Prospective Cohorts of the BBMRI-LPC project. Both catalogues have been linked to BioSD (http://www.ebi.ac.uk/biosamples/), and the LPC catalogue is being used in the pilot implementation of WP 5, Task 8.

BBMRI RDF 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 which can be found at http://bbmri-eric.eu/reports. 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.

BBMRI RDF Previous work:

The work in D4.6 further builds on the REST web service implemented in D4.3 (https://www.bbmriportal.eu/bbmri2.0/bbmri/bbmri.xml?u=biosd&p=biosdpass) in order to establish a data bridge between BBMRI and ELIXIR.

Work done for 4.6:

BBMRI RDF 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 http://bioportal.bioontology.org/ontologies/ICD10/) 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 establishing relations to the Semantic science Integrated Ontology (SIO) (https://code.google.com/p/semanticscience/wiki/SIO) have been added to the OWL ontology, as recommended in D4.7. An RDF representation of the ontology in Turtle syntax is available at https://www.bbmriportal.eu/bbmri2.0/bbmri2rdf.ttl. 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 as shown in https://www.bbmriportal.eu/bbmri2.0/rdfexport.n3. The triple store is automatically updated nightly. For querying the RDF content, a SPARQL endpoint based on the Fuseki server has been installed. A graphical user interface to the SPARQL endpoint is provided at https://www.bbmriportal.eu/bbmri2.0/sparql.html.

BBMRI RDF Metrics:

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

BBMRI RDF 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.

A.3 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 analysing 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?" 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 modelling nor the technical implementation of RDF below has been previously reported. The BMB WP7 (PhenoBridges) 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).

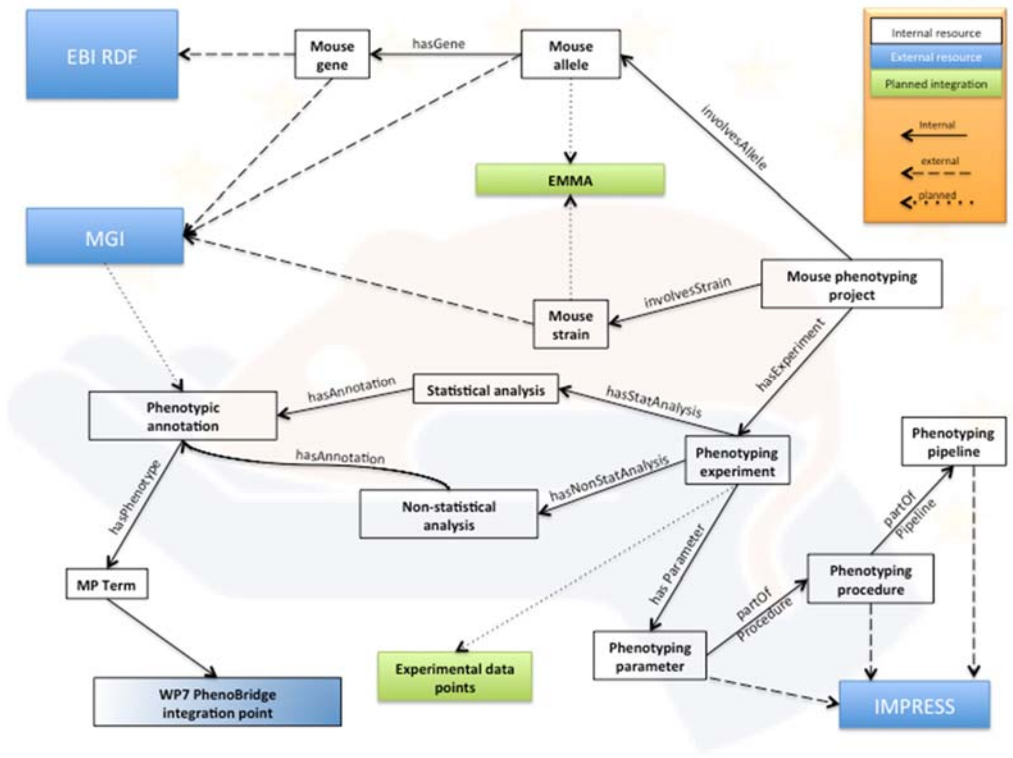


Figure A5 Model for the semantic integration of mouse phenotype resources

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 once the services are hosted in the Infrafrontier domain. User experience testing is currently performed internally within the GMC; it will be repeated with external users at a later date.

Future work:

- Integration of further datasets from all main sources (IMPC, GMC, MGI)(continuously)
- Diabetes ontology(DIAB) import to allow human phenotype queries.
- Definition of further disease-parameter presets for enhanced phenomap browsing (continuously).
- Enrich user interface with other RDF data (from D4.6)

A.4 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 BMB 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[23] , first we applied our text-mining pipeline, which mainly consists of named entity taggers[24], accession number tagger[25] and section tagger[26] to Open Access full-text articles[27], and then we converted the text-mined results into triples based on the Open Annotation Data Model (OADM). The OADM treats annotations as primary resources and provides a standard description mechanism for these annotations them between systems.

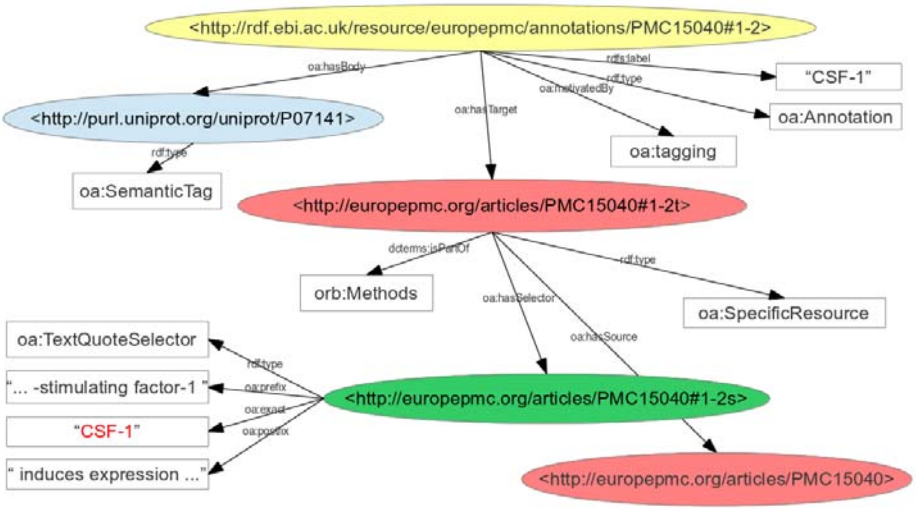

<http://rdf.ebi.ac.uk/resource/europepmc/annotations/PMC15040#1-2>
oa:hasBody
rdfs:label
"CSF-1"
rdf:type
oa:motivatedBy
oa:Annotation
oa:hasTarget
<http://purl.uniprot.org/uniprot/P07141>
oa:tagging
rdf:type
oa:SemanticTag
<http://europepmc.org/articles/PMC15040#1-2t>
dcterms:isPartOf
rdf:type
orb:Methods
oa:hasSelector
oa:SpecificResource
oa:hasSource
oa:TextQuoteSelector
rdf:type
"... -stimulating factor-1 "
oa:prefix
<http://europepmc.org/articles/PMC15040#1-2s>
oa:exact
"CSF-1"
oa:postfix
" induces expression ..."
<http://europepmc.org/articles/PMC15040>

Figure A6 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

[23] 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
[24] 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.
[25] Database citation in full text biomedical articles. Kafkas Ş, Kim JH, McEntyre JR. PLoS One Volume 8 (2013) p.e63184
[26] 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
[27] Europe PMC open access articles http://europepmc.org/ftp/archive/v.2014.09/oa/

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 (e.g., http://europepmc.org/articles/PMC15040/methods/genes/CSF-1 instead of using hashing), 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 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.”[28]

Until recently, the GWAS diagram (http://www.ebi.ac.uk/fgpt/gwas/) 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 triple store. JSON-LD is not the right choice for all datasets (for instance where transitive graph-based queries are routinely

[28] 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

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.

A.5 RDF Training materials

http://tinyurl.com/ebirdftraining2015

A.6 RDF BioMedBridges Symposium materials

Background:

A symposium organised by BioMedBridges, entitled 'Open bridges for life science data' , took place at the Wellcome Trust Conference Centre on the 17th and 18th November, 2015; http://biomedbridges.eu/news/symposium-open-bridges-life-science-data. This incorporated a number of parallel workshops, focused on specific issues. One such session communicated the lessons learnt during the course of WP4 in particular: "Challenges in interoperability and semantic data integration - Lessons learned from BioMedBridges and OpenPhacts" (http://www.biomedbridges.eu/workshop-challenges-interoperability-and-semantic-data-integration-lessons-learned-biomedbridges-and). The materials from this workshop are collected here: https://drive.google.com/open?id=0B0PYIHPze5SXS0x4QXJCWXFudG8.

Appendix B: tranSMART pilots

B.1 ErasmusMC (Euro-Bioimaging): 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-centre 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 analysed 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-centre 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[29] as a very concrete test case to guide the development. The study by Guyader

[29] 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.

et al was originally performed in the context of the Quic-Concept project (http://www.quic-concept.eu). In 5 volunteers, diffusion-weighted magnetic resonance images (MRI) were collected at two time points. From these images, we computed apparent diffusion coefficients (ADC)[30] 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. Besides 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 intend to better integrate the data using ontologies (see future work section). We build on previous WP4 work as follows.

Work done for 4.6:

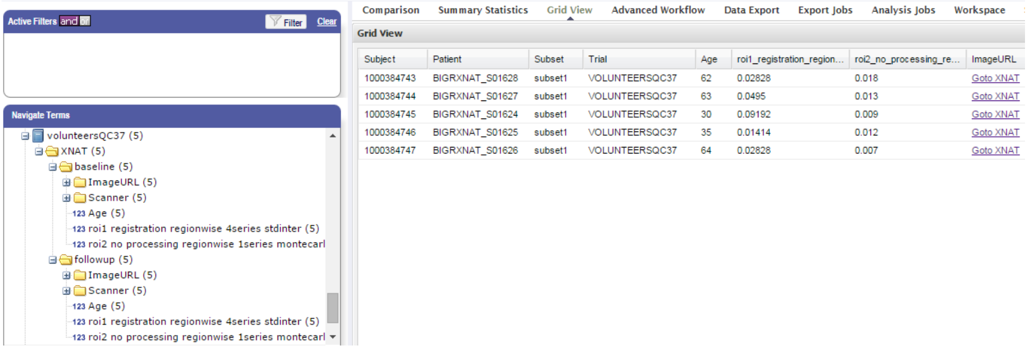
Technical implementation:

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

[30] The ADC is a measure of the magnitude of diffusion (of water molecules) within tissue (http://radiopaedia.org/articles/apparent-diffusion-coefficient-1).

volumes or mean ADC values for regions of interest, as only these quantitative measurements will be used for statistical analysis. The figure B.1 below shows an example of a current tranSMART project that includes image-derived data from XNAT. Note also the “Go to XNAT” link here, which brings the user directly to a page in the XNAT system where the original images can be inspected, if desired.



Grid View

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

Figure B1 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[31] importing system to import the data.

[31] The ETL importer (https://wiki.transmartfoundation.org/display/TSMTGPL/Data+ETL) for Clinical Data is used.

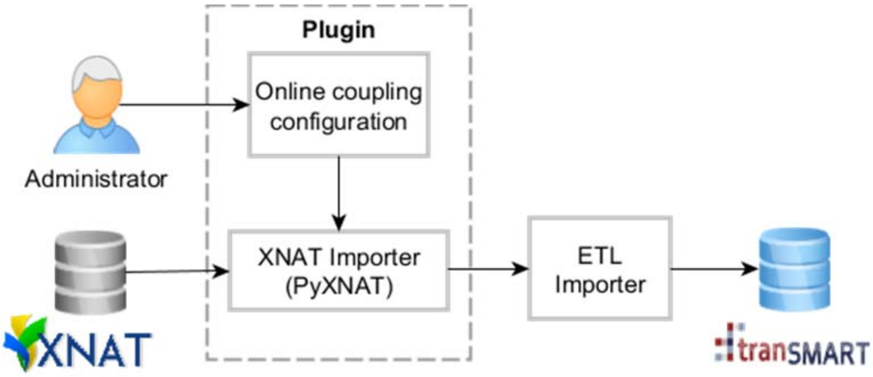


Figure B2 schematic overview of the tranSMART-XNAT link

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 can be found on https://github.com/evast/transmart-xnat-importer-plugin. We have written both a user guide for data managers as technical documentation for developers, available on: https://github.com/evast/transmart-xnat-importer-plugin/blob/master/docs/.

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 B.1.3 below), and we have written a formal user agreement. To support administrators, we have updated and streamlined the Puppet configuration scripts (https://bitbucket.org/bigr_erasmusmc/puppet-xnat), and made these available as a new release on the XNAT marketplace (http://marketplace.xnat.org/).

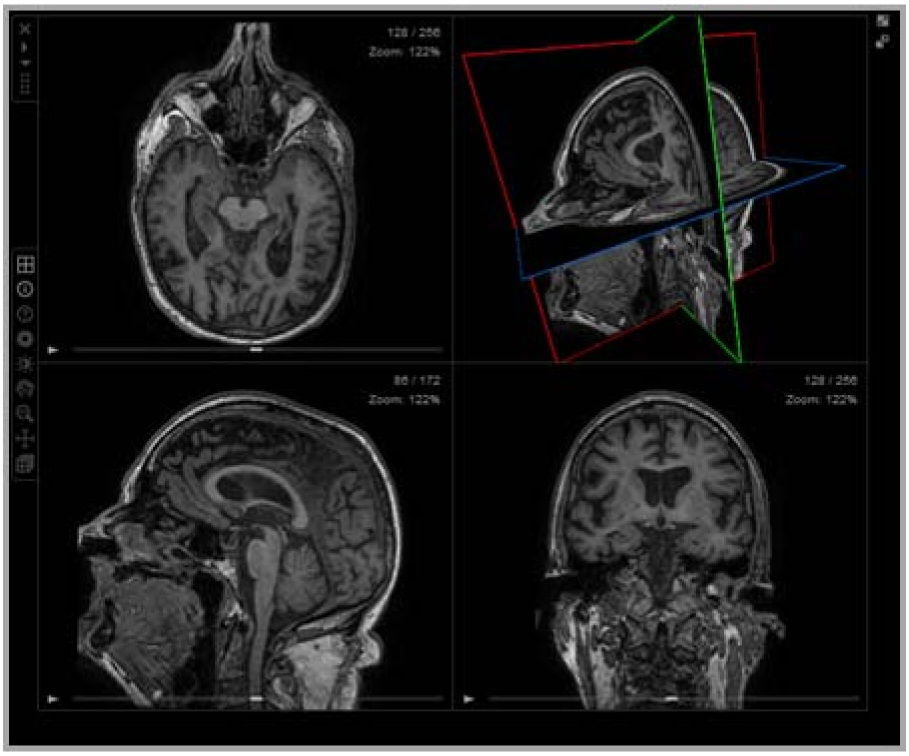


Figure B3 Three dimensional viewing of brain MRI scan using web-based XNAT viewer

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.

Further work after D4.6

Subsequently to D4.6, during the last period of BioMedBridges, Erasmus MC has focused on activities aimed at maximising the adoption and usage of the tools previously developed. See Section 3.2.1.2 of the current deliverable for more details.

B.2 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 (https://github.com/thehyve/Rmodules/tree/features/transmart-galaxy) plugin: this plugin handles Galaxy workflows in tranSMART and was written by Ruslan Forostianov (The Hyve) (http://thehyve.nl/portfolio/ruslan-forostianov/);
- workflow runner (https://github.com/CTMM-TraIT/trait_workflow_runner): this component simplifies using the Galaxy API from Java and was written by Freek de Bruijn (VUmc) (https://github.com/FreekDB);

- blend4j (https://github.com/jmchilton/blend4j): this existing library provides access to the Galaxy API (and other systems) from Java and was written by John Chilton (Penn State University) (https://wiki.galaxyproject.org/JohnChilton);
- Galaxy API (http://galaxy-dist.readthedocs.org/en/latest/lib/galaxy.webapps.galaxy.api.html): 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) (https://wiki.galaxyproject.org/GalaxyTeam).

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 http://en.wikipedia.org/wiki/List_of_JVM_languages). 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.
- 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

http://www.pidconsortium.eu/). Once this is possible, we can support analysing 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.

B.3 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 analyse the clinical data, or 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 analyse 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 (https://community.i2b2.org/wiki/display/ODM2i2b2/Home), 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 availability of data in EDCs. It is the next Extract-Transform-Load (ETL) step for the OCDataImporter tool (https://community.openclinica.com/extension/ocdataimporter).

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.

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.

B.4 Workshop on Translational Research Infrastructure at OpenBridges symposium hosted on the SURFsara HPC Cloud

At the BioMedBridges symposium titled Open bridges for life science data (http://www.biomedbridges.eu/news/symposium-open-bridges-life-science-data) which was held on 17 and 18 November 2015 at Wellcome Conference Centre in Hinxton, United Kingdom a workshop was organised on Translational Research Infrastructures. The workshop was co-chaired by Stefan Klein (Euro-

BioImaging, Erasmus MC), Morris Swertz (BBMRI, UMCG), and Freek de Bruijn (EATRIS, VUmc and NKI).

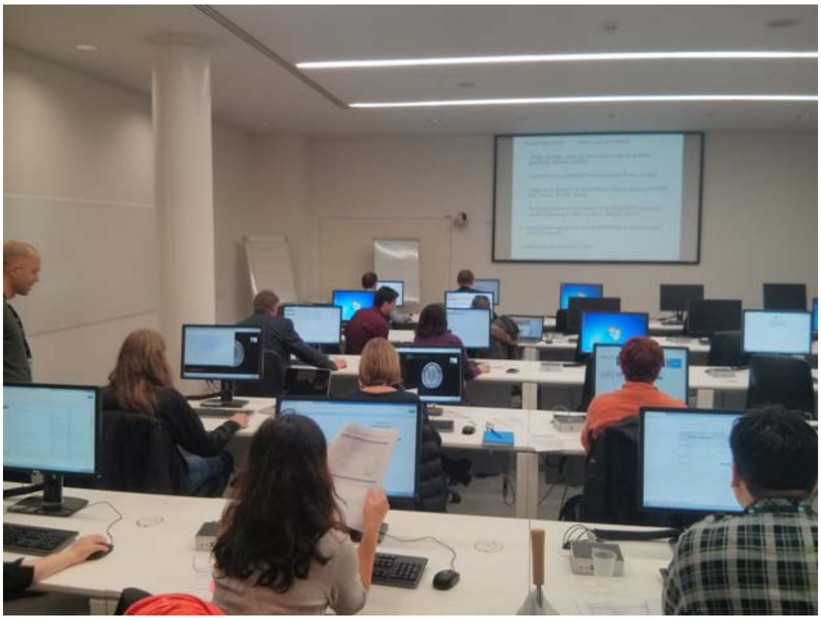
During this workshop a hands-on introduction to IT infrastructure for translational biomedical research is given, including the "bridges" that were jointly built in the BioMedBridges project. These bridges allow medical researchers to submit, use, manage and combine imaging data, molecular data, and clinical data (including biosamples: query biobanks to find samples and related data that are relevant for their/your research). Five tutorials could be followed to see some of the bridges in action.

The workshop described the functionality offered by the systems and the connections at a high level and was mainly targeted at researchers interested in using the infrastructure. Participants were given the opportunity to get hands-on experience with several systems/services/bridges:

1. Medical image storage, viewing, and processing using XNAT and Fastr.
2. Integrative analysis of multi-domain data using tranSMART and Galaxy.
3. Analysing clinical data using OpenClinica and tranSMART.
4. Pooling data from different biobanks using metadata model mapping & registry in MOLGENIS.
5. Using the MOLGENIS Personalized Genomics portal for NGS patient annotation.

The more tech-savvy were most welcome as well, because the people who implemented the bridges and some of the used systems were available at the workshop.

The capacity for the workshop was 40 people. We used the HPC Cloud (https://userinfo.surfsara.nl/systems/hpc-cloud) of SURFsara to provide access to private virtual machines for each participant and to be flexible regarding the number of participants showing up. Having their own virtual machine gave complete freedom to the participant to play around with the systems. Running the machines on the HPC Cloud enabled us to make highly demanding tutorials in terms of memory and CPU usage.



Acknowledgements

For this workshop, important contributions were made by the following people besides the organisers:

- Jeroen Beliën (VUmc) (contact for more information mailto:jam.belien@vumc.nl or http://www.ctmm-trait.nl/)
- Ward Blondé (VUmc, NKI)
- Dennis Hendriksen (UMCG)
- Bart Charbon (UMCG)
- Erwin Vast (Erasmus MC)
- Marcel Koek (Erasmus MC)

- Mattias Hansson (Erasmus MC)

Appendix C: BioJS Widget pilots

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.

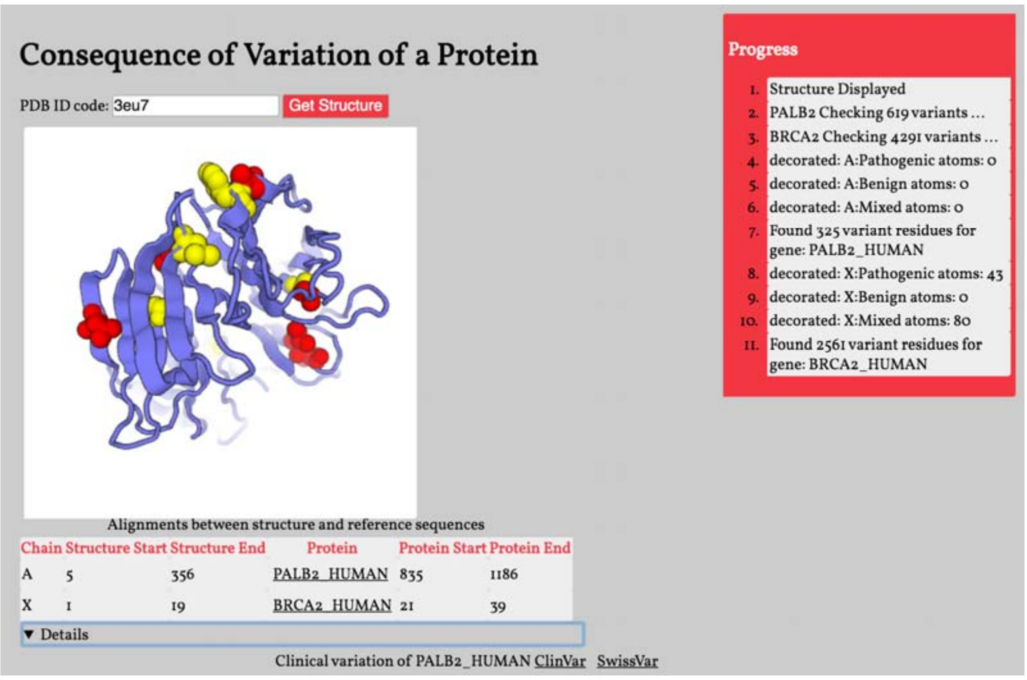
Work done for 4.6:

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 (https://github.com/biasmv/pv) 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 lessons learned).



Consequence of Variation of a Protein

PDB ID code: 3eu7

Progress

1. Structure Displayed
2. PALB2 Checking 619 variants ...
3. BRCA2 Checking 4291 variants ...
4. decorated: A:Pathogenic atoms: 0
5. decorated: A:Benign atoms: 0
6. decorated: A:Mixed atoms: 0
7. Found 325 variant residues for gene: PALB2_HUMAN
8. decorated: X:Pathogenic atoms: 43
9. decorated: X:Benign atoms: 0
10. decorated: X:Mixed atoms: 80
11. Found 2561 variant residues for gene: BRCA2_HUMAN

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

Clinical variation of PALB2_HUMAN ClinVar SwissVar

Fig C1 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

Metrics:

Recently announced, no metrics available currently. 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 web service (http://www.ebi.ac.uk/pdbe/api/doc/search.html) 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.

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 (http://www.ebi.ac.uk/ols).

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) 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.

BioMedBridges tools and data registry (WP3) 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 http://tinyurl.com/miaospecification), Appendix C.2.1 was developed that described how an ontology should be accessed. MIAO is currently being aligned with similar efforts from 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 BMB 4.6 contribution to the OLS back-end development was to create a custom CORS-enabled (http://en.wikipedia.org/wiki/Cross-origin_resource_sharing) 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:

1. Auto completion on ontology terms
2. Auto completion on a configurable set of free-text terms
3. Performant centralised query interface (via web service)
4. Visualisation of matching terms within their immediate tree context
    a. Ability to expand / collapse tree nodes
5. Subsumption queries
6. Configurability with any ontology and dataset
7. 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 http://www.targetvalidation.org/). 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 contributions to the code, thereby making it even more extensible, robust, and feature-rich with time.

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

The representation of the specification for this ontology is available here: http://tinyurl.com/miaospecification (https://docs.google.com/spreadsheets/d/1pTdRsCM9terVYS7biAw5mvdNwKBKONa4dro2ry_wCqI/edit#gid=0).