Deliverable D4.3

Project Title:	Building data bridges between biological and medical infrastructures in Europe	
Project Acronym:	BioMedBridges	
Grant agreement no.:	284209	
	Research Infrastructures, FP7 Capacities Specific Programme; [INFRA-2011-2.3.2.] “Implementation of common solutions for a cluster of ESFRI infrastructures in the field of "Life sciences"	
Deliverable title:	Pilot integration using REST web services	
WP No.	4	
Lead Beneficiary:	1: EMBL	
WP Title	Technical integration	
Contractual delivery date:	30 June 2013	
Actual delivery date:	17 July 2013	
WP leader:	Ewan Birney	1: EMBL

Contributing partner(s):	4: STFC 5: UDUS 6: FVB 7: TUM-MED 9: ErasmusMC 11: HMGU 13: VUMC

Authors: Jeroen Belien (VUMC), Ewan Birney (EMBL-EBI), Jan Willem Boiten (VUMC), Benjamin Braasch (UDUS), Sabine Brunner (TUM-MED), Jon Chambers (EMBL-EBI), Adam Faulconbridge (EMBL-EBI), Philipp Gormanns (HMGU), Safey Halim (TUM-MED), Dennis Hendriksen (UMCG), Jon Ison (EMBL-EBI), Töresin Karakoyun (UDUS), Klaus Kuhn (TUM-MED), Narayanan Krishnan (STFC), Christoph Lengger (HMGU), Holger Maier (HMGU), Willi Mann (TUM-MED), Julie McMurry (EMBL-EBI), Chris Morris (STFC), Christian Ohmann (UDUS), John Overington (EMBL-EBI), Helen Parkinson (EMBL-EBI), Ralph Steinkamp (HMGU), Morris Swertz (UMCG), Martyn Winn (STFC)

Contents

1 Executive summary

Deliverable D4.3 implements technical plans devised in deliverable D4.2. The work is implemented within six pilot integration studies which span the BioMedBridges domains and which will act as drivers for future WP4 activities. The following infrastructures: BBMRI, EATRIS, ECRIN, ELIXIR, EU-OPENSCREEN, Euro-BioImaging, INFRAFRONTIER and INSTRUCT contributed to this deliverable on behalf of the partner institutes listed in section 6. Throughout the text, the contributing partners should be understood to represent the respective research infrastructures.

Pilot 1 (Biosample information integration and discovery) integrated two pivotal resources by transferring data from the BBMRI.eu 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. Future developments will ensure that all other biobanks in the BBMRI catalogue, which agree to publish their data, will be seamlessly reflected into the BioSD.

Pilot 2 (Federating biobank queries for translational research) achieves a seamless federated search of two biobank catalogues via new RESTful services and demonstrates the feasibility of a search across two biobank catalogues created for different purposes. Future developments will federate searches across ELIXIR, BBMRI, and EATRIS databases, resulting in the first instance with an integrated search across 6 countries.

Pilot 3 (Leveraging the utility of compound screening functional assays) implemented a new vocabulary for functional assays, focusing on animal models of diabesity, and a RESTful Web service allowing external users to utilize these data in powerful new ways. Future developments will ensure new terms are included into the official public release of BioAssay Ontology, and that tables of the terms are accessible via downloads and the ChEMBL interface.

Pilot 4 (Sharing protein engineering knowledge) developed a new RESTful interface over PiMS; 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. Planned integration of PiMS with UniProt will enable scientists using the PiMS system to seamlessly traverse between PiMS data and UniProt data and deliver a major enhancement in the usability of the whole experimental pipeline.

Pilot 5 (Integrating mouse phenotype data for diabetes research) developed new RESTful Web services providing a gene-based integration of the Gene Expression Atlas with systemic phenotyped mice data, focussing on the relevant diseases for WP7 (diabetes and obesity)*.* These developments are the groundwork for further integration, which will be enriched by RDF transformation of the mice data (D4.4/4.6) as well as by the outcomes of the PhenoBridge use case for interspecies data mapping (WP7).

Pilot 6 (Integrating gene and drug information with a clinical trials registry) implemented a clinical trials portal 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. In future developments, the Web service will be enhanced with additional information relevant to a clinical trials query.

These services and the underlying data lay the foundations for the integration strategy of BioMedBridges: the developments are extensible in the future and will act as drivers for future WP4 activities and will be built upon in subsequent deliverables. The developments are also sustainable in the context of BioMedBridges and beyond.

2 Project objectives

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

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

Deliverable 4.2 examined the technical status of WP4 contributors and assessed the feasibility of the technical plans: which use cases could be delivered with existing technology and which required development of new programmatic access in the form of Web services. This deliverable implements these technical plans within six pilot integration studies which span the BioMedBridges domains and which will act as drivers and examples for future WP4 activities:

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

A series of collaborative activities including on-site meetings, face-to-face discussions at project workshops, teleconferences and phone-calls were conducted throughout the reporting period to ensure that the requirements and recommendations identified in the D4.2 Technical workshop (technical integration strategy) relevant to this deliverable were fully addressed.

The collection of Web services presented here provides a foundation for the integration strategy of BioMedBridges to be expanded with pilots for Semantic Web integration (D4.4. and D4.6) and 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.

3.2 Description of work

3.2.1 Collaborative activities

These involved WP4 contributors and other key BioMedBridges partners and included:

- 9 on-site meetings hosted at partner institutes including 3 at EMBL-EBI
- 6 group teleconferences
- Many teleconferences

The collaborative activities identified:

- Current data provision and specific plans for data/metadata sharing between sites (i.e. the scientific bridges)
- Specific plans for service implementations reflecting dependencies between services and data resources
- The technology used by each partner, their expertise and technical limitations pertinent to a specific institution / service provision

- Common understanding of the technology and other requirements to build the scientific bridges

Partners and research infrastructures are broadly either data producers or consumers, a minority are both consumers and producers. At the onset of this deliverable, around half the research infrastructures had internal data models in place and expertise in REST-based service delivery or RDF, with expertise concentrated in data provider institutes. To facilitate and promote developments, services provided by or used by the partners have been described at a coarse-grained level and catalogued in the emerging BioMedBridges Tools Catalogue (http://wwwdev.ebi.ac.uk/fgpt/toolsui/), providing an integration point with WP3's service registry task 3.3.

3.2.2 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 (aligned with WP3 activity)
- 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

The work towards this deliverable addressed the following challenges. These are documented as they provide context generally for WP4, and specifically for the services delivered.

- An institute supplying data does not provide access via Web services, either because such access has not been required or there are technical limitations or other constraints. Some institutes required assistance to develop their services. Alternatively, the strategy is to sidestep the problem by one centre providing data to another centre or repository which then provides the service by proxy.

- An institute has data that cannot, for reasons of data privacy and security, be accessed via a Web service. An appropriate strategy is to consider the provision of summary level data rather than full data access. For example, as a proof-of-principle, BBMRI takes a tiered approach whereby summary metadata, aggregated data and individual data are served according to the constraints of the relevant data access committee.
- Multiple existing resources provide complementary content which is not currently amenable to single, integrated queries. An appropriate integration strategy is to import the underlying data from one resource to another & harmonise the content.
- Large, distributed data sets are available and are partially exposed via human-usable Web user interfaces (GUIs), but not in a form that exposes all the available data to programmatic access, limiting the integration scenarios in which it can be used. In this scenario it is not possible or desirable to integrate the data by physical import. An appropriate strategy is to expose all available content to a Web service providing access over the federated content.
- A large corpus of scientific documents exists but which are not currently amenable to normalisation, integration and efficient data mining because of a lack of consistent scientific descriptions. An appropriate strategy is to create a common, harmonising vocabulary with which to curate and power searches over the resource. (Activity aligned with WP3)
- A data resource exists with well established query facilities which are not, however, expected to scale in the long term to the expected quantity and complexity of future queries and integration scenarios. An appropriate strategy is, as a proof of principle, to convert and serve the data in a format suitable for use on the Semantic Web.
- A critical data resource exists whose value is not being fully exploited because of barriers to interpreting the data in light of the broader scientific context. An appropriate strategy is to cross-reference the

data, then expose it via Web services to queries returning results that link out to these key resources.

3.2.3 Pilot 1: Biosample information integration and discovery

Background

Advances in drug discovery especially in the context of personalized medicine require access to larger arrays of biosamples in support of increasingly complex scientific questions. For example, WP8 (Personalized Medicine) requires search functionality for acute myeloid leukaemia (AML) samples and related attributes. The aim of WP8 is to enrich the personalised data set of FIMM AML patients to improve patient treatment. Therefore heterogeneous AML data from a variety of data sources will be used. One source is biobanks with samples from AML patients. Integrating data from the BBMRI.eu catalogue to BioSamples Database (BioSD) facilitates the search for these samples, generates traffic from translational and biomolecular researchers accessing ELIXIR resources, exposes the data for query by relevant disease ontologies, and ensures the data are shared and preserved.

As the diversity, complexity and connections between data grows, it is more important to organize the data such that users can more easily find and query data that are relevant to their interests. In the case of biosample repositories, developing facility for integrated searches will help maximize, for patient benefit, the impact of domain-specific solutions in translational research that need to be expanded to different categories of users. The existing pilot helps address this aim by integrating two pivotal resources:

1. The **BBMRI.eu catalogue** provides a comprehensive overview of the European Biobanking landscape. It is based on questionnaires that cover many aspects of biobanking (topics of interest, disease groups, origin and use of samples, study design and recruitment, data sources, confidentiality, consent, access, principal variables of interest, sample storage conditions and so on.) Pertinent to this pilot, it also contains data about the number of collected samples and their material types. The BBMRI.eu catalogue was originally developed during the BBMRI preparatory phase.

2. **The BioSamples Database (BioSD)** aggregates sample information for reference samples e.g. Coriell Cell lines and samples for which data exist in one of the EBI's assay databases such as ArrayExpress, the European Nucleotide archive, or Pride. It provides links to assays for specific samples, and accepts direct submissions of samples. Samples are described using attributes of key/value pairs, with optional ontology or controlled vocabulary references, as well as relationships to other samples.

A common objective of the BBMRI.eu catalogue and BioSD is the collection of data about available biological samples, though the context and query profile of the respective user communities are different. A BBMRI user may query for common samples for the purpose of ordering samples, or constructing a study, a BioSamples user may query for a sample in order to identify related biomolecular data such as RNA-seq, or exome sequences performed on the samples. In order to facilitate the search for biosamples that might be relevant for a particular research, sample data contained in the BBMRI.eu catalogue should be available via the BioSD search interface and represented consistently with other similar data in the BioSD context.

Technical implementation

Early discussions concluded that the best way to achieve the required integration was for the BBMRI.eu catalogue to acquire data in SampleTab format to BioSD, via a new submission tool on the side of the BBMRI.eu catalogue using a BioSD submission service that expects SampleTab files. SampleTab is a tab-delimited file format that can be created using a spreadsheet editing software package. Details about the format are available at (http://www.ebi.ac.uk/microarray-srv/biosd/static/st.html). The BioSD provides Web services for use by BioMedBridges and the wider biomedical community, for programmatically sending sample information to its database. Three Web services are available:

1. **Submission service:** for submitting SampleTab files
2. **Validation service:** for validating SampleTab files without submitting them

3. **Accessioning service:** for assigning accessions to samples and groups

The three services are available via web interfaces and run as production services for use in this pilot. Using the submission service web interface, the end user can upload a SampleTab file to the BioSD, or check the validity of its content by uploading it to the validation service without inserting the data into the database. The Web services are also available for programmatic use via a RESTful API. Our integration solution for pushing data from the BBMRI catalogue to the BioSD makes use of the BioSD's RESTful submission service. The BioSD's RESTful submission service takes the SampleTab file content converted into a JSON message and encapsulated in a RESTful POST request.

On the BBMRI catalogue side, a standalone tool (the BioSD Push Service) was developed. This service runs as an overnight job whose role is to query the catalogue database for biobank data changes. If data have been updated, or new data have been inserted, they are exported into SampleTab format, and that exported file is then uploaded to the BioSD by calling the BioSD submission service. The Push Service keeps track of the latest version of the data that have been pushed to the BioSD in order not to push redundant data and to be able to handle submission failures correctly. For example: if there was a submission failure, and data have not been changed till the next submission time, then the previous SampleTab file has to be resubmitted although there was no change. Data from the catalogue are submitted to the BioSD one biobank at a time (one JSON message contains information about one biobank). When a JSON message is successfully submitted, the BioSD submission service returns a "Submission Identifier". The catalogue's Push Service then stores that identifier in an SQLite database. Each "Submission Identifier" is stored with its corresponding catalogue's biobank ID along with the submission timestamp, and the submission status (success or failure). For each biobank, those four pieces of information ("Submission Identifier", "biobank ID", "submission timestamp", and "submission status") are used to decide whether data for the same biobank should be submitted again or not by comparing the biobank's "last modified" date to the last submission timestamp, and by checking the last submission status.

The BBMRI catalogue exposes biobank data as an XML Web service. That Web service is called by the BioSD Push Service to query biobank data from the catalogue. After querying the biobank data, the Push Service converts the XML document into SampleTab using XSLT. The resulting SampleTab file is then encapsulated in a JSON message and sent to the BioSD submission service. The catalogue data was converted in two steps: XML and then SampleTab, to keep the BBMRI catalogue client-agnostic: biobank XML data can be retrieved by any client, or can be converted by that client to any other format.

Data from the following biobanks have been automatically pushed from the BBMRI catalogue to the BioSample 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 the catalogue's biobanks which agreed to publish their data, the mechanism is easily extensible should other biobanks also wish to publish data in future.

Ongoing and future developments

By realizing aforementioned integration scenario between the BBMRI catalogue and the BioSD, all data in the catalogue from biobanks that agreed to publish their data will be available in the BioSD. Having the catalogue's BioSD Push Service running as an overnight job ensures that any updates to the catalogue data will be automatically and seamlessly reflected into the BioSD, therefore biobank data in the catalogue and the BioSD will grow simultaneously.

3.2.4 Pilot 2: Federating biobank queries for translational research

Background

Biobanking underpins research advances in a multitude of clinical disciplines and is critical to worldwide efforts into translational research, or the effort to translate clinical, pathological and other information into medical practice and meaningful health outcomes. To realise the full potential of the biobanks that are already available, it is essential to provide integrated searches over the resources in a rational way.

BBMRI (UMCG) and EATRIS (VUMC) aim to improve discoverability of biobank and translational datasets, using the Netherlands community as testbed. In this pilot project, the first steps are made to enable federated search of biobank and translational databases in NL (>200 in the BBMRI-NL biobank catalogue). This pilot has proven it is feasible to create a federated search across two biobank catalogues created for different purposes. It also positively evaluated use of indexing technologies (ElasticSearch) as additional method to scale this federated search towards many biobank catalogues.

Technical implementation

The pilot achieves a seamless federated search of two biobank catalogues, implemented using RESTful services built on the MOLGENIS database platform (http://www.molgenis.org) and following the Common Biorepository Model (CBM, https://wiki.nci.nih.gov/display/TBPT/Common+Biorepository+Model+(CBM)).

Two biobank catalogues were populated with a (disjunct) public CBM data available at:

http://molgenis46.target.rug.nl/ (user name/password: admin/admin)

http://molgenis47.target.rug.nl/ (user name/password: admin/admin)

The content in the demonstrator is from the Dutch CTMM project (CTMM-TraIT, http://www.ctmm-trait.nl/) where a catalogue of all translational medicine sets in the Netherlands is compiled. The purpose of this data set was to evaluate the CBM format and tailor it for the Dutch situation (CBM-NL).

The technology used for the implementation is the open source MOLGENIS database platform (http://www.molgenis.org and http://github.com/molgenis) and using the Observ-OM (Observation Object) model which was co-developed by UMCG and EBI in a previous project[1]. Data loading went very smoothly proving the flexibility of both software and model. On top of this a new scalable index was developed as part of BioMedBrides using ElasticSearch indexing software and a REST API to enable federated queries. The result is an unbelievably fast and powerful search capability which convinces us to further develop this in the next iteration.

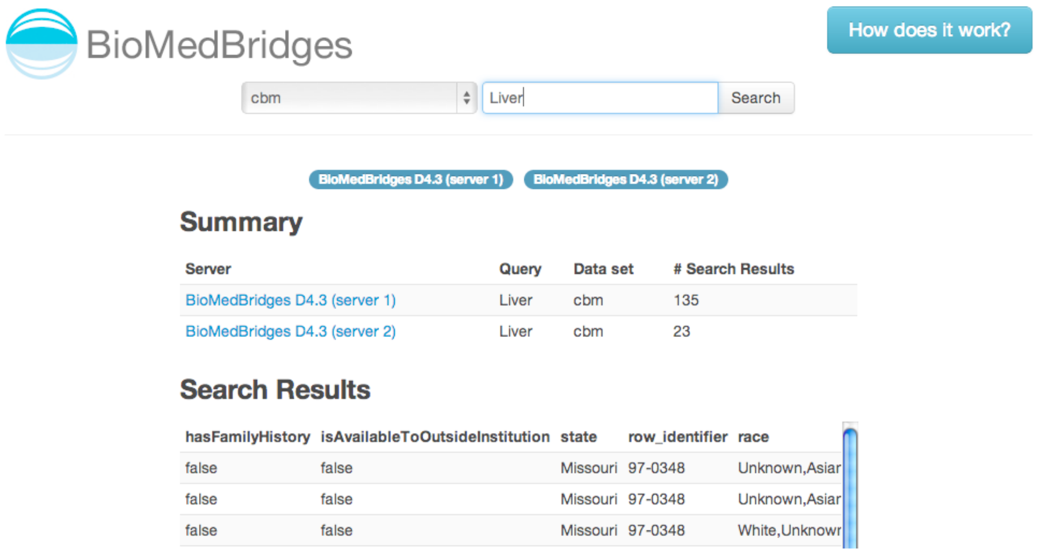
The pilot services enables search of these databases via a Web user interface (see Figure 1) at:

http://molgenis12.target.rug.nl/js/bmb_d43.html

(Works with Chrome 27 or other modern browser)

Within the pilot system users can use a ‘Google’ type search via a user interface. At each search two REST requests are posted to the underlying biobank catalogues, in this example ‘server 1’ and ‘server 2’. These REST request body specifies ‘query rules’ in JSON format and can be expanded to an item level search in the next iteration. One can log-in to one of the two source servers and there see how that can work in the ‘data explorer’ tab. The RESTful services return a simple list of results also in JSON format which are rendered in a user readable form in the demo. This REST API can easily be implemented on other servers.

[1] Adamusiak T, Parkinson H, Muilu J, Roos E, van der Velde KJ, Thorisson GA, Byrne M, Pang C, Gollapudi S, Ferretti V, Hillege H, Brookes AJ, Swertz MA. Observ-OM and Observ-TAB: Universal syntax solutions for the integration, search, and exchange of phenotype and genotype information. Hum Mutat. 2012 May;33(5):867-73. doi: 10.1002/humu.22070



Summary

Server	Query	Data set	# Search Results
BioMedBridges D4.3 (server 1)	Liver	cbm	135
BioMedBridges D4.3 (server 2)	Liver	cbm	23

Search Results

hasFamilyHistory	isAvailableToOutsideInstitution	state	row_identifier	race
false	false	Missouri	97-0348	Unknown,Asiar
false	false	Missouri	97-0348	Unknown,Asiar
false	false	Missouri	97-0348	White,Unknowr

Figure 1. Pilot services enable federated search of two test biobanks via a Web user interface. User interface of the BBMRI/EATRIS federated search pilot as available on http://molgenis12.target.rug.nl/js/bmb_d43.html. Users can type free text in the search box on top and the keywords are then sent to connected servers as listed in 'Summary'. On each search a query is sent to the connected servers and the returning results are merged and presented.

Ongoing and future developments

The next step will be to implement federated search across ELIXIR, BBMRI, and EATRIS databases, based on collaboration in BioMedBridges between ELIXIR, BBMRI-Nordic and BBMRI-NL. The first milestone of this effort would be integrated search across 6 countries and using MIABIS as a standard, as agreed upon at the Catalogue Workshop (Copenhagen, Spring 2013). The adoption of content standards in this domain and subsequent mapping of these integrates this deliverable with D3.2.

3.2.5 Pilot 3: Leveraging the utility of compound screening functional assays

Background

Within the ChEMBL database, assays with defined molecular targets are relatively simple to analyse and integrate with external data sets, as they are 'keyed' on robust identifiers, such as UniProt IDs. However, a large number of the assays in ChEMBL are higher level functional/phenotypic assays, where assay descriptions are free text, and mechanisms to cluster/normalize these assays within ChEMBL, for the purposes of mining ChEMBL data, and integration with external sources, do not yet exist. The ability to flexibly search and organise these functional screens would be applicable to other databases with similar data models, such as the prototype EU-OPENSCREEN database: ECBD.

The objective is to leverage the utility of functional assays in the ChEMBL database by:

1. Create normalized terms to describe a subset of functional assays in ChEMBL (animal models of diabesity were chosen as a suitable subset).
2. Create a means of storing and updating these terms in the ChEMBL database.
3. If possible, create the means for external users to simply utilize these data, ideally via a RESTful Web service.

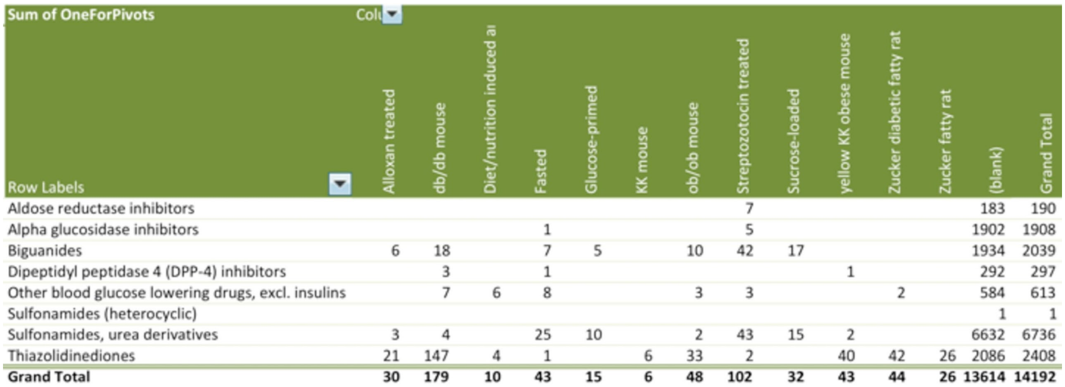
In addition to permitting Bioassay integration between ChEMBL and other resources, these developments will allow users to ask questions of ChEMBL data, such as... “Historically, what sort of functional assays and animal models have been used to develop different classes of drugs for Diabetes?”

Technical implementation

Approximately 14,000 distinct functional assays (textual descriptions), that are connected to compounds in ChEMBL classified as Diabetes drugs by the WHO ATC classification, were analysed. A large number were in vitro assays, but a significant number were in vivo assays utilizing an animal model of diabetes or obesity. These were curated and classified. 12 categories were clearly defined (see Figure). Following discussions between ChEMBL and BioAssay Ontology (BAO), it was concluded that the most maintainable and robust mechanism for managing and utilizing these normalized terms would be to submit them to the BAO. This would also be the most effective means of allowing further integration with other resources, as the BAO is rapidly becoming the integration hub for BioAssay annotation in this domain. PubChem have begun to use this, and EU-OPENSCREEN plan to use it too. Accordingly, the normalized diabesity animal models defined in the current curation process, together with some others not present in the current data set (but commonly occurring in the literature), have been submitted to the BAO, together with definitions and a suggested hierarchical assignment.

Example data analysis

By querying ChEMBL for all drugs annotated as ‘Diabetes drugs’ by the ATC WHO (but excluding multiple forms of insulin, i.e. A10 category, but excluding A10A), then retrieving all functional bioassays for these drugs, then retrieving, where they exist, all animal model annotation terms for these assays, a large data set connecting drug classes to assay classes was assembled. These data may be analysed in multiple ways, but as an example below (see Figure 2), data were pivoted on ‘ATC level4 description’ vs. ‘Animal model category’. Several observations may then be made, such as the use of historical Streptozotocin (first used in the 1960’s, and considered a IDDM model) appear to have been used less frequently to assay more recently developed drug classes such as the thiazolidinediones, where the more recently introduced NIDDM genetic models such as ob/ob and Zucker may have been chosen, presumably because they may more closely mimic the human target condition (although this rationale would need to be verified by an expert in the field).



Sum of OneForPivots	Colu▾												
Row Labels ▾	Alloxan treated	db/db mouse	Diet/nutrition induced a	Fasted	Glucose-primed	KK mouse	ob/ob mouse	Streptozotocin treated	Sucrose-loaded	yellow KK obese mouse	Zucker diabetic fatty rat	Zucker fatty rat	(blank)	Grand Total
Aldose reductase inhibitors								7					183	190
Alpha glucosidase inhibitors				1				5					1902	1908
Biguanides	6	18		7	5		10	42	17				1934	2039
Dipeptidyl peptidase 4 (DPP-4) inhibitors		3		1						1			292	297
Other blood glucose lowering drugs, excl. insulins		7	6	8			3	3			2		584	613
Sulfonamides (heterocyclic)													1	1
Sulfonamides, urea derivatives	3	4		25	10		2	43	15	2			6632	6736
Thiazolidinediones	21	147	4	1		6	33	2		40	42	26	2086	2408
Grand Total	**30**	**179**	**10**	**43**	**15**	**6**	**48**	**102**	**32**	**43**	**44**	**26**	**13614**	**14192**

Figure 2. Example analysis of functional bioassays for diabetes drugs pivoted on ‘ATC level4 description’ vs. ‘Animal model category’.

Ongoing and future developments

Collaboration with the BAO will ensure the submitted terms are officially included into the public release of BAO. In addition, these data are also stored in tables within the production ChEMBL database. Currently, these data are only accessible by direct querying of the production ChEMBL database. In future releases, the tables, or modified forms of them, may be accessible via downloads and the ChEMBL interface. It is anticipated that RESTful Web service methods to query for these data may be best implemented using BAO identifiers (when assigned).

A REST service method for querying ChEMBL for these data is documented together with existing methods at https://www.ebi.ac.uk/chembldb/index.php/ws.

3.2.6 Pilot 4: Sharing protein engineering knowledge

Background

PiMS (http://pims.structuralbiology.eu/) is a laboratory information management system for use in recombinant protein production laboratories; to manage the stages from the selected target protein to the production of soluble protein. PiMS development is part of a larger vision to provide a unified and extensible set of software tools for structural biology, offering seamless data transfer and a consistent user experience, from target selection to the interpretation of the structure.

The previous system of manual record keeping is stressed by the increasing productivity of protein scientists, as protocols become more effective and high throughput methods spread. Concomitantly, other scientists and end-users are demanding increasingly complex, deeper and more flexible queries of the data and especially to interpret and analyse data in light of other datasets, for example, protein sequence and structural information. The desire to support such queries and other advanced query scenarios motivated this pilot, which provides a new RESTful interface over PiMS serving data in Semantic Web formats.

The PiMS REST interface provides an opportunity for life scientists to examine their experimental data and derive more information that was lost/went unnoticed at earlier instances. The interface represents responses to searches in multiple semantic-web formats. This presents an opportunity for machine learning algorithms to use this information and make inferences from the experimental data supplied.

Technical implementation

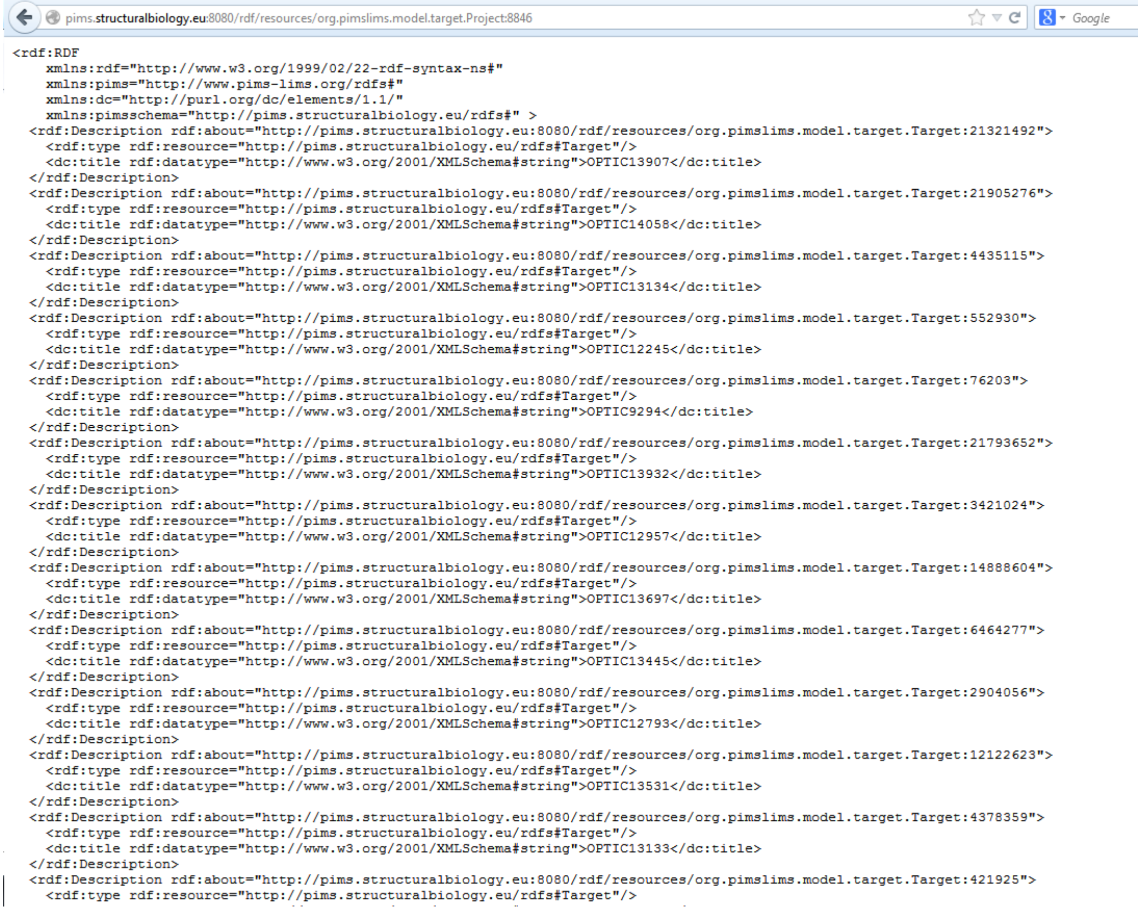
The data: One of the major hurdles in the implementation of the system was to identify a group of PiMS users to make available an actual set of data to be used in the REST service developed. It was a concern as few users were willing to release their experimental data for fear of data exploitation as the REST service would be public. Fortunately, the Oxford Protein Production Facility (OPPF), one of the Instruct centres using PiMS, was willing to share data acquired during a project they were involved in: Protein100. The dataset that was shared consisted of 1062478 records that could be published, and represents a significant quantity of data in this domain.

The API: A new API for retrieving data from the data model and render them as RDF serves data on an on-demand basis. The API, together with the extensively complex PiMS API, can help the UI to do many kinds of searches. A simple test of the extent of the API search is to locate a single resource by its URL. This is made possible by a design choice that was made during the creation of PiMS, which allowed every accessible resource in the data model to be identified by a hook.

Example:

http://pims.structuralbiology.eu:8080/rdf/resources/org.pimslims.model.target.Project:8846

The above link will fetch the resource Project:8846 from the data model and represent it as RDF/XML. Screenshot is shown in Figure 3.


pims.structuralbiology.eu:8080/rdf/resources/org.pimslims.model.target.Project:8846
Google

```
<rdf:RDF
    xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
    xmlns:pims="http://www.pims-lims.org/rdfs#"
    xmlns:dc="http://purl.org/dc/elements/1.1/"
    xmlns:pimsschema="http://pims.structuralbiology.eu/rdfs#" >
  <rdf:Description rdf:about="http://pims.structuralbiology.eu:8080/rdf/resources/org.pimslims.model.target.Target:21321492">
    <rdf:type rdf:resource="http://pims.structuralbiology.eu/rdfs#Target"/>
    <dc:title rdf:datatype="http://www.w3.org/2001/XMLSchema#string">OPTIC13907</dc:title>
  </rdf:Description>
  <rdf:Description rdf:about="http://pims.structuralbiology.eu:8080/rdf/resources/org.pimslims.model.target.Target:21905276">
    <rdf:type rdf:resource="http://pims.structuralbiology.eu/rdfs#Target"/>
    <dc:title rdf:datatype="http://www.w3.org/2001/XMLSchema#string">OPTIC14058</dc:title>
  </rdf:Description>
  <rdf:Description rdf:about="http://pims.structuralbiology.eu:8080/rdf/resources/org.pimslims.model.target.Target:4435115">
    <rdf:type rdf:resource="http://pims.structuralbiology.eu/rdfs#Target"/>
    <dc:title rdf:datatype="http://www.w3.org/2001/XMLSchema#string">OPTIC13134</dc:title>
  </rdf:Description>
  <rdf:Description rdf:about="http://pims.structuralbiology.eu:8080/rdf/resources/org.pimslims.model.target.Target:552930">
    <rdf:type rdf:resource="http://pims.structuralbiology.eu/rdfs#Target"/>
    <dc:title rdf:datatype="http://www.w3.org/2001/XMLSchema#string">OPTIC12245</dc:title>
  </rdf:Description>
  <rdf:Description rdf:about="http://pims.structuralbiology.eu:8080/rdf/resources/org.pimslims.model.target.Target:76203">
    <rdf:type rdf:resource="http://pims.structuralbiology.eu/rdfs#Target"/>
    <dc:title rdf:datatype="http://www.w3.org/2001/XMLSchema#string">OPTIC9294</dc:title>
  </rdf:Description>
  <rdf:Description rdf:about="http://pims.structuralbiology.eu:8080/rdf/resources/org.pimslims.model.target.Target:21793652">
    <rdf:type rdf:resource="http://pims.structuralbiology.eu/rdfs#Target"/>
    <dc:title rdf:datatype="http://www.w3.org/2001/XMLSchema#string">OPTIC13932</dc:title>
  </rdf:Description>
  <rdf:Description rdf:about="http://pims.structuralbiology.eu:8080/rdf/resources/org.pimslims.model.target.Target:3421024">
    <rdf:type rdf:resource="http://pims.structuralbiology.eu/rdfs#Target"/>
    <dc:title rdf:datatype="http://www.w3.org/2001/XMLSchema#string">OPTIC12957</dc:title>
  </rdf:Description>
  <rdf:Description rdf:about="http://pims.structuralbiology.eu:8080/rdf/resources/org.pimslims.model.target.Target:14888604">
    <rdf:type rdf:resource="http://pims.structuralbiology.eu/rdfs#Target"/>
    <dc:title rdf:datatype="http://www.w3.org/2001/XMLSchema#string">OPTIC13697</dc:title>
  </rdf:Description>
  <rdf:Description rdf:about="http://pims.structuralbiology.eu:8080/rdf/resources/org.pimslims.model.target.Target:6464277">
    <rdf:type rdf:resource="http://pims.structuralbiology.eu/rdfs#Target"/>
    <dc:title rdf:datatype="http://www.w3.org/2001/XMLSchema#string">OPTIC13445</dc:title>
  </rdf:Description>
  <rdf:Description rdf:about="http://pims.structuralbiology.eu:8080/rdf/resources/org.pimslims.model.target.Target:2904056">
    <rdf:type rdf:resource="http://pims.structuralbiology.eu/rdfs#Target"/>
    <dc:title rdf:datatype="http://www.w3.org/2001/XMLSchema#string">OPTIC12793</dc:title>
  </rdf:Description>
  <rdf:Description rdf:about="http://pims.structuralbiology.eu:8080/rdf/resources/org.pimslims.model.target.Target:12122623">
    <rdf:type rdf:resource="http://pims.structuralbiology.eu/rdfs#Target"/>
    <dc:title rdf:datatype="http://www.w3.org/2001/XMLSchema#string">OPTIC13531</dc:title>
  </rdf:Description>
  <rdf:Description rdf:about="http://pims.structuralbiology.eu:8080/rdf/resources/org.pimslims.model.target.Target:4378359">
    <rdf:type rdf:resource="http://pims.structuralbiology.eu/rdfs#Target"/>
    <dc:title rdf:datatype="http://www.w3.org/2001/XMLSchema#string">OPTIC13133</dc:title>
  </rdf:Description>
  <rdf:Description rdf:about="http://pims.structuralbiology.eu:8080/rdf/resources/org.pimslims.model.target.Target:421925">
    <rdf:type rdf:resource="http://pims.structuralbiology.eu/rdfs#Target"/>
```

Figure 3. A new API for retrieving data from the data model and render them as RDF serves data on an on-demand basis. Note: Some browsers (like Internet Explorer) have interpreters of RDF embedded in them. In such cases the whole structure of RDF can be viewed by viewing the source.

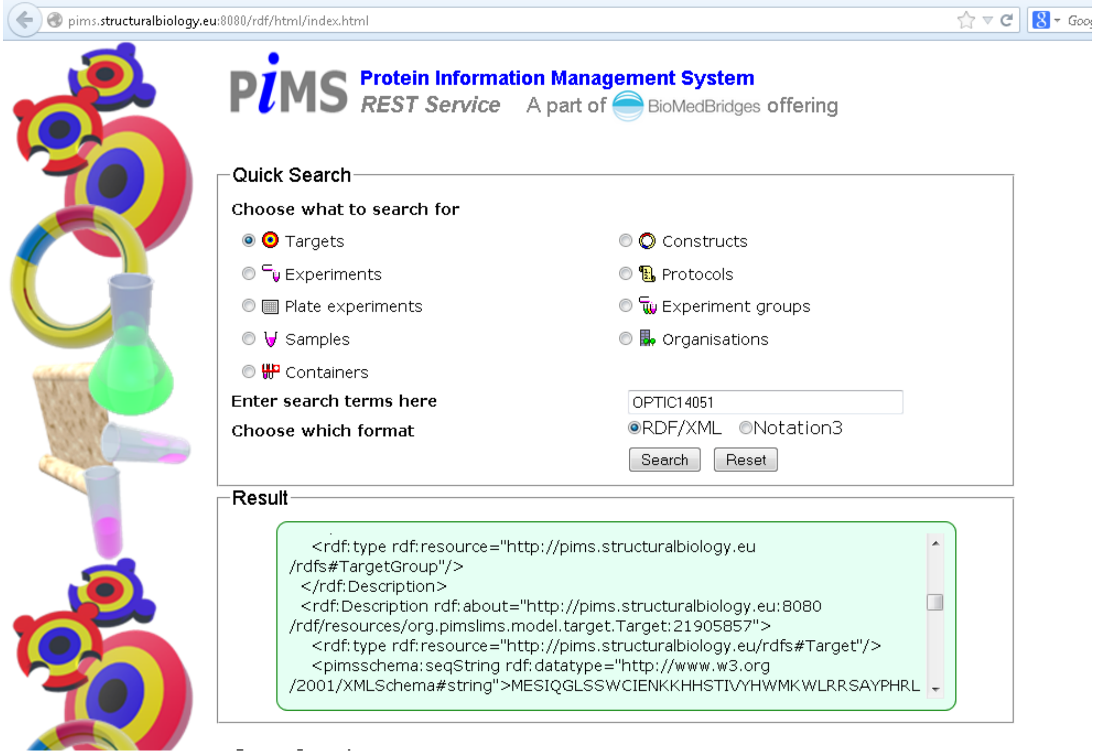
The interface can be accessed at:

http://pims.structuralbiology.eu:8080/rdf/

Documentation of the interface is available at:

http://pims.structuralbiology.eu:8080/pims/functions/Help.jsp

The UI: The user interface (Figure 4) of the service is simple and intuitive for any first time user. The interface has query sections and results sections. The interface also provides a choice of RDF/XML format (less human readable, better for machines and algorithms) and Notation3 format (better for human reading). The Quick Search interface helps in quickly obtaining information on the different types of most used searchable classes. Advanced searches below the Quick Search section helps to identify resources based on various parameters. All searches are free text searches.



pims.structuralbiology.eu:8080/rdf/html/index.html

PiMS Protein Information Management System
REST Service A part of BioMedBridges offering

Quick Search

Choose what to search for

● Targets
○ Experiments
○ Plate experiments
○ Samples
○ Containers
○ Constructs
○ Protocols
○ Experiment groups
○ Organisations

Enter search terms here: OPTIC14051

Choose which format: ● RDF/XML ○ Notation3

Search Reset

Result

```
    <rdf:type rdf:resource="http://pims.structuralbiology.eu
/rdfs#TargetGroup"/>
  </rdf:Description>
  <rdf:Description rdf:about="http://pims.structuralbiology.eu:8080
/rdf/resources/org.pimslims.model.target.Target:21905857">
    <rdf:type rdf:resource="http://pims.structuralbiology.eu/rdfs#Target"/>
    <pimsschema:seqString rdf:datatype="http://www.w3.org
/2001/XMLSchema#string">MESIQGLSSWCIENKKHHSTIVYHWMKWLRRSAYPHRL
```

Figure 4. The user interface provides convenient basic and advanced search functionality and returns results in a variety of formats.

Validation and reasoning: An in-house tool was also written to represent the data model as an RDF schema in the expectation it would be useful along with the generated RDF for any reasoning programs. Reasoning was also done to make sure that the generated RDFs can be used to make sensible and predictable inferences when used by reasoning systems. A thorough and comprehensive test-driven approach was adopted through the development process for the stability and accountability of the program.

Technology choice: The underlying API used to generate RDF is Apache Jena API. However, a wrapper programming interface is written around the Jena API in order to make it disjoint from the developed application and if need be, replace it with another API in the future.

Ongoing and future developments

Using formats like Resource Description Format (RDF-XML) helps in integrating with data from disparate sources. One such example is the planned integration of PiMS with UniProt which will enable scientists using PiMS system to seamlessly traverse between PiMS data and UniProt data. Such integration has been reported by scientists using the system as a major enhancement in the usability of the whole experimental pipeline.

3.2.7 Pilot 5: Integrating mouse phenotype data for diabetes research

Background

Result data from systemic phenotyping of mutant mouse lines obtained at the EUROPHENOME and IMPC (International Mouse Phenotyping Consortium) project are currently only accessible through web front-end portals and not all details are publicly accessible. This limits the potential usage of the data as well as its integration with other data sources, e.g. Gene Expression Atlas. An integration of mouse phenotyping results with other resources is currently performed manually as a pilot in WP 7 “PhenoBridge”. A technical framework providing the basis to utilize the fruits of the PhenoBridge would allow users to exploit the outcomes immediately. Accordingly, RESTful pilot Web services were developed by INFRAFRONTIER and ELIXIR. These are mandatory for data access and they provide a first gene-based integration of the Gene Expression Atlas with mutant mouse phenotyping results focusing on diseases chosen as examples for WP7 (diabetes and obesity).

Technical implementation

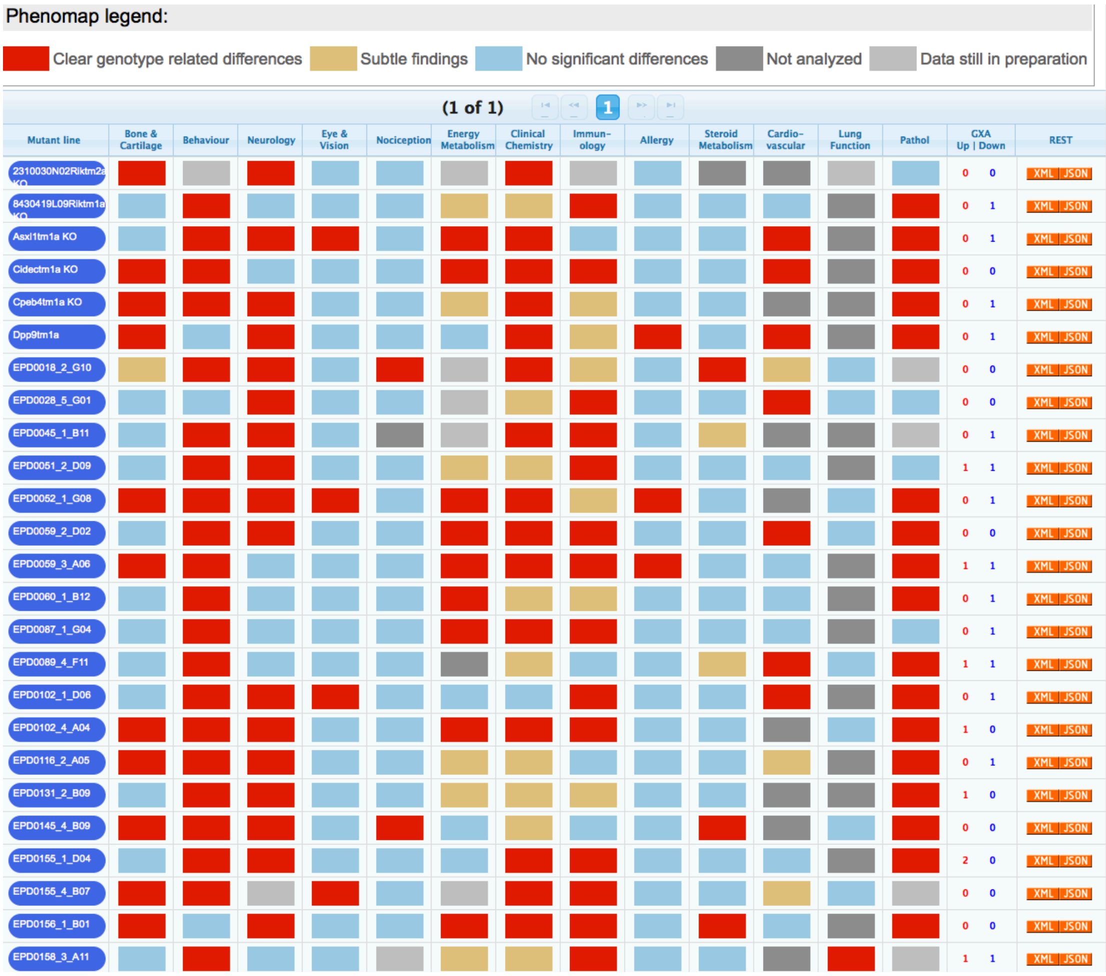
A RESTful Web service was developed for 60 mutant mouse lines examined by the German Mouse Clinic (GMC). The service permits RESTful access to the phenotype annotations which were added by the mouse phenotyping experts of the GMC. Moreover, this service also allows co-querying the Gene Expression Atlas for diabetes annotations for the gene which is mutated in the related mouse line. Also the service provides RESTful access to the phenotyping raw data of these mouse lines retrieved from the

EUROPHENOME database (www.europhenome.org). Documentation of the service and interface is available here:

http://146.107.35.38/biomedbridges_hmgu/documentation.jsp

All RESTful Web services are also visually accessible via a phenomap which provides an entry point for scientist interested in the analysis of mouse phenotyping results (Figure 5):

http://146.107.35.38/biomedbridges_hmgu_phenomap/jsp/annotation/bmb/phenomap.jsf



Phenomap legend:

🟥 Clear genotype related differences 🟨 Subtle findings 🟦 No significant differences ⬛ Not analyzed ⬜ Data still in preparation

(1 of 1) 1

Mutant line	Bone & Cartilage	Behaviour	Neurology	Eye & Vision	Nociception	Energy Metabolism	Clinical Chemistry	Immun-ology	Allergy	Steroid Metabolism	Cardio-vascular	Lung Function	Pathol	GXA		REST
														Up	Down	
2310030N02Riktm2a KO	🟥	⬜	🟥	🟦	🟦	⬜	🟥	⬜	🟦	⬛	⬛	⬜	🟦	0	0	XML JSON
8430419L09Riktm1a KO	🟦	🟥	🟦	🟦	🟦	🟨	🟨	🟥	🟦	🟦	🟦	⬛	🟥	0	1	XML JSON
Asxl1tm1a KO	🟦	🟥	🟥	🟥	🟦	🟥	🟥	🟦	🟦	🟦	🟥	⬛	🟥	0	1	XML JSON
Cidectm1a KO	🟥	🟥	🟦	🟦	🟦	🟥	🟥	🟥	🟦	🟦	🟥	⬛	🟥	0	0	XML JSON
Cpeb4tm1a KO	🟥	🟥	🟥	🟦	🟦	🟨	🟥	🟨	🟦	🟦	⬛	⬛	🟥	0	1	XML JSON
Dpp9tm1a	🟥	🟦	🟥	🟦	🟦	🟦	🟥	🟨	🟥	🟦	🟥	⬛	🟥	0	1	XML JSON
EPD0018_2_G10	🟨	🟥	🟥	🟦	🟥	⬜	🟥	🟨	🟦	🟥	🟨	🟦	⬜	0	0	XML JSON
EPD0028_5_G01	🟦	🟦	🟥	🟦	🟦	⬜	🟨	🟥	🟦	🟦	🟥	🟦	🟦	0	0	XML JSON
EPD0045_1_B11	🟦	🟥	🟥	🟦	⬛	⬜	🟥	🟥	🟦	🟨	⬛	⬛	⬜	0	1	XML JSON
EPD0051_2_D09	🟦	🟥	🟥	🟦	🟦	🟨	🟨	🟥	🟦	🟦	🟦	⬛	🟦	1	1	XML JSON
EPD0052_1_G08	🟥	🟥	🟥	🟥	🟦	🟥	🟥	🟨	🟥	🟦	⬛	🟦	🟥	0	1	XML JSON
EPD0059_2_D02	🟦	🟥	🟥	🟦	🟦	🟥	🟥	🟥	🟦	🟦	🟥	🟦	🟥	0	0	XML JSON
EPD0059_3_A06	🟥	🟥	🟦	🟦	🟦	🟥	🟥	🟥	🟥	🟦	🟦	⬛	🟥	1	1	XML JSON
EPD0060_1_B12	🟦	🟥	🟦	🟦	🟦	🟥	🟨	🟨	🟦	🟦	🟦	⬛	🟥	0	1	XML JSON
EPD0087_1_G04	🟦	🟥	🟦	🟦	🟦	🟥	🟥	🟥	🟦	🟦	🟦	⬛	🟦	0	1	XML JSON
EPD0089_4_F11	🟦	🟥	🟦	🟦	🟦	⬛	🟨	🟦	🟦	🟨	🟥	🟦	🟥	1	1	XML JSON
EPD0102_1_D06	🟦	🟥	🟥	🟥	🟦	🟦	🟦	🟥	🟦	🟦	🟥	⬛	🟥	0	1	XML JSON
EPD0102_4_A04	🟥	🟥	🟥	🟦	🟦	🟥	🟥	🟥	🟦	🟦	⬛	🟦	🟥	1	0	XML JSON
EPD0116_2_A05	🟥	🟥	🟥	🟦	🟦	🟨	🟨	🟦	🟦	🟦	🟨	⬛	🟥	0	1	XML JSON
EPD0131_2_B09	🟦	🟥	🟥	🟦	🟦	🟨	🟨	🟨	🟦	🟦	⬛	⬛	🟥	1	0	XML JSON
EPD0145_4_B09	🟥	🟥	🟥	🟦	🟥	🟦	🟨	🟦	🟦	🟥	⬛	🟦	🟥	0	0	XML JSON
EPD0155_1_D04	🟦	🟥	🟥	🟦	🟦	🟦	🟥	🟥	🟦	🟦	🟦	⬛	🟥	2	0	XML JSON
EPD0155_4_B07	🟥	🟥	⬜	🟥	🟦	⬜	🟥	🟥	🟦	🟦	🟨	🟦	⬜	0	0	XML JSON
EPD0156_1_B01	🟥	🟦	🟥	🟦	🟦	🟥	🟥	🟥	🟦	🟥	🟦	⬛	🟥	0	0	XML JSON
EPD0158_3_A11	🟦	🟥	🟦	🟦	⬜	🟨	🟨	🟥	🟦	🟦	⬛	🟥	⬜	1	1	XML JSON

Figure 5. Web services are visually accessible via a phenomap which provides an entry point for scientist interested in the analysis of mouse phenotyping results.

Technology Choice

Web services provided by INFRAFRONTIER are developed in JAVA using Maven/Spring/Hibernate and run on a Tomcat Server. Visualisation and frontend development is done with JSF and PrimeFaces. The integration of Gene Expression Atlas is performed through the RESTful GXA interface provided by ELIXIR.

Sample queries

To illustrate the utility of the Web service sample queries are shown below.

1. Expert phenotypic information for mouse line derived from the EUCOMM clone ‚EPD0215_1_F05'

http://146.107.35.38/biomedbridges_hmgu/rest/line=EPD0215_1_F05&format=xml

2. Experimental raw data for mouse line derived from the EUCOMM clone ‚EPD0215_1_F05' in the phenotypic screen category ‚Clinical_Chemistry and Hematology'

http://146.107.35.38/biomedbridges_hmgu/rest/line=EPD0215_1_F05&screen=8&format=xml

3. Phenotypic data extraction for mouse line derived from the EUCOMM clone ‚EPD0164_4_E12' enriched by expression related 'diabetes' information for the involved gene extracted from the Gene Expression Atlas.

http://146.107.35.38/biomedbridges_hmgu/rest/line=EPD0164_4_E12&disease=diabetes&format=xml

Ongoing and future developments

These pilot Web services are the groundwork for further integration, which will be enriched by RDF transformation of the mice data (D4.4/4.6) as well as by the outcomes of the PhenoBridge use case for interspecies data mapping (WP7). The service that provides access to the experimental raw data of EUROPHENOME mouse lines is temporary and will be replaced by a RESTful service provided by ELIXIR which allows querying a combined resource of all EUROPHENOME, IMPC and MGP datasets.

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

Background

Advances in understanding the genetic basis of disease, gene-based disease management and the efficacy of drug treatments relies critically upon systematic study of clinical phenotypes in light of information about implicated genes, drugs and the available scientific literature. For researchers in biomedical sciences, finding and collating all the related entries from different biomedical databases is often a time consuming problem. The implementation of this pilot solves this problem by linking publications and information about genes and drugs to data from clinical trials registries. The aim is to include fitting publications for queries of clinical trials results and provide more insight into the effect of drugs or genes on patients. For example to support queries of the type “Find clinical trials with results involving drugs *X* or *Y*" or "Find publications involving clinical trial *Z*".

Technical implementation

A clinical trials portal was implemented using several different state of the art technologies and applications.

Clinical trials data: Clinical trials metadata was harvested from ClinicalTrials.gov registry, which holds the results of publicly and privately supported clinical studies of humans from around the world. Registry data was collated as XML files and then processed with new Perl and Java-based applications to transform it into Java clinical trial objects in preparation for input for JAX-B in Java and Apache Solr. During the transformation, further enhancements and especially to improve the metadata semantics are conducted. These are necessary to open the bridge with MESH between publications data from PubMed (D4.6) and clinical trials data. In D4.6 more modifications will be necessary to bridge clinical trials data to data from genes or drugs. Indexing of the data in support of the required search functionality was added with Apache Lucene, a highly efficient and open-source search library provided by the Apache foundation.

Service interface: An appropriate RESTful interface was implemented using Apache Solr. The code is successfully deployed on the EMBL Cloud in a virtual machine.

The deployed service is accessible over the following URL: http://193.62.52.249:8983/solr/ct/

The service allows the user to do various types of queries, including add, delete and update functionality. Some examples are shown below:

Querying:

- select?q=leukemia&fl=nct_id (return only id fields)
- select?q=leukemia&fl=nct_id,title (return id and title fields)
- select?q=leukemia&wt=json (return response in JSON format)

Sorting:

- select?q=leukemia&sort=nct_id desc (sort in descending order)
- select?q=leukemia&sort=nct_id asc (sort in ascending order)

Deleting:

- update?stream.body=
  <delete><query>id:123</query></delete>&commit=true

Documentation of the interface is available here: http://193.62.52.249.

Ongoing and future developments

As a Clinical Trials Information Mediator (CTIM) the Web service is planned to become the perfect place to go for various use cases involving clinical trials. Access to results of publications and genetic or drug information in other databases that are relevant to a clinical trials query will be simplified, saving the researcher a lot of time. An easy to use and intuitive graphical user interface (GUI) will support the researcher to get the information he is searching for.

The developed REST Web service will be enhanced with linking information so that the bridges between other data can be created. More data and an automated update mechanism will create the perfect basis for the next deliverables, which include a GUI for the service and the semantic integration. Security mechanisms are not yet implemented, but are warranted partly owing to firewall rules of the EMBL Cloud Service.

This deliverable implements the technical plans established in Deliverable 4.2 within six pilot integration studies that span the BioMedBridges domains. The studies deliver a collection of Web services that are representative of the underlying resources and provide biologically meaningful and scientifically valid access to partner data. Recommendations identified in the D4.2 Technical workshop (technical integration strategy) relevant to this deliverable were reflected within the various strategies that were deployed to overcome a range of technical challenges and other constraints.

3.3 Conclusions

Pilot 1 (Biosample information integration and discovery) integrated two pivotal resources by transferring data from the BBMRI.eu catalogue to the BioSamples Database (BioSD). This fulfils the requirement for integrated searches over larger sample collections in support of increasingly complex scientific questions, including query by relevant disease ontologies. For this pilot, five of the BBMRI.eu catalogues agreed to share their data, all of which 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. Future developments will ensure that all other biobanks in the BBMRI catalogue, which agree to publish their data, will be seamlessly reflected into the BioSD, ensuring biobank data in the catalogue and the BioSD grows simultaneously.

Pilot 2 (Federating biobank queries for translational research) 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 used content from the Dutch CTMM project (CTMM-TraIT) catalogue 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 enable 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. Future developments will federate searches across ELIXIR, BBMRI, and EATRIS databases, resulting in the first instance with an integrated search across 6 countries.

Pilot 3 (Leveraging the utility of compound screening functional assays) implemented a new vocabulary for functional assays, focusing on animal models of diabesity, 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 descriptions of distinct functional assays connected to compounds in ChEMBL classified as Diabetes drugs were curated and classified. Corresponding normalized diabesity animal models including terms, definitions and suggested hierarchy were submitted to the 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.

Pilot 4 (Sharing protein engineering knowledge) developed a new RESTful interface over PiMS; 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, 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. Planned integration of PiMS with Uniprot will enable scientists using the PiMS system to seamlessly traverse between PiMS data and Uniprot data and deliver a major enhancement in the usability of the whole experimental pipeline.

Pilot 5 (Integrating mouse phenotype data for diabetes research) 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. All services are also visually accessible via the phenomap on the GMC website. These developments are the groundwork for further integration, which will be enriched by RDF transformation of the mice data (D4.4/4.6) as well as by the outcomes of the PhenoBridge use case for interspecies data mapping (WP7).

Pilot 6 (Integrating gene and drug information with a clinical trials registry) implemented a clinical trials portal 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. In future developments, the Web service will be enhanced with additional information relevant to a clinical trials query. An automated update mechanism will create the basis for the next deliverables, which include a graphical interface for the service and the semantic integration.

These services and the underlying data lay the foundations for the integration strategy of BioMedBridges: the developments are extensible in the future and will act as drivers for future WP4 activities and will be built upon in subsequent deliverables. The developments are also sustainable in the context of BioMedBridges and beyond.

4 Delivery and schedule

The delivery is delayed: ☑ Yes ☐ No

The work for deliverable D4.3 was completed in time; however, the submission of this report was slightly delayed due to the temporary absence (maternity leave) of key personnel.

5 Adjustments made

The biomedical sciences research infrastructures participating in the project are at various stages of maturity, both technically and logistically. To address this variability within the BioMedBridges project in the most productive way possible, resources were shifted within WP4 (from FVB to EMBL) and from this deliverable to other WP4 deliverables (STFC, UDUS, TUM-MED, ErasmusMC, HMGU, VUMC) as well as WP7 (HMGU). The shift was part of an amendment to the Grant Agreement.

6 Efforts for this deliverable

Institute	**Person-months (PM)**		**Period**
	actual	**estimated (after GA adjustment)**	
1: EMBL	13	13	June 2012 - June 2013
4: STFC	17	17	January 2012 - June 2013
5: UDUS	10	10	September 2012 - June 2013
6: FVB	0	0	-

7: TUM-MED	5	5	November 2012-June 2013
9: ErasmusMC	0	0	-
11: HMGU	6	6	January 2013 - June 2013
13: VUMC	3	3	January 2013 - June 2013
Total	54		

7 Background information

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

WP 4 Title: Technical Integration
Lead: Ewan Birney (EMBL)
Participants: EMBL

In work package 4 we will implement a federated access system to the diverse data sources in BioMedBridges. This will focus on providing access to data or metadata items which utilise the standards outlined in WP 3. Experience across the BioMedBridges partners is that executing a federated access system, in particular a federated query system, is complex for both technological and social reasons. Therefore we will be using an escalating alignment/engagement strategy where we focus on technically easier and semantically poorer integration at first and then progressively increase the sophistication of the services. In each iteration, we will be using biological use cases which are aligned to the capabilities of the proposed service, thus providing progressive sophistication to the suite of federated services.
Our first iteration involves using established REST based technology to provide userbrowsable visual integration of information. This will be useful for both summaries of data rich resources (such as Elixir) and summaries of ethically restricted datasets where only certain meta-data items are public (such as BBMRI, ECRIN and EATRIS). We will then progress towards lightweight distributed document and query lookups, where the access for ethically restricted data will incorporate the results of WP 5. Finally at the outset of the project we will explore exposure of in particular meta-data sets via RDF compatible technology, such as SPARQL, and the presence of the technology watch WP11 will provide recommendations for other emerging technologies to use, aiming for the semantically richest integration.

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	37

Objectives

1. Implement shared standards from WP 3 to allow for integration across the BioMedBridges project
2. Expose the integration via use of REST based 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.