

Deliverable D4.5

Project Title:	Building data bridges between biological and medical infrastructures in Europe	
Project Acronym:	BioMedBridges	
Grant agreement no.:	284209	
	Research Infrastructures, FP7 Capacities Specific Programme; [INFRA-2011-2.3.2.] "Implementation of common solutions for a cluster of ESFRI infrastructures in the field of "Life sciences"	
Deliverable title:	Pilot integration of REST based vignette services for the second round BMS projects	
WP No.	4	
Lead Beneficiary:	1: EMBL	
WP Title	Technical integration	
Contractual delivery date:	31 December 2013	
Actual delivery date:	23 December 2013	
WP leader:	Ewan Birney	1: EMBL
Partner(s) contributing to this deliverable:	1: EMBL, 5: UDUS, 9: ErasmusMC	

Authors: Ewan Birney (EMBL-EBI), Benjamin Braasch (UDUS), Jon Chambers (EMBL-EBI), Jon Ison (EMBL-EBI), Töresin Karakoyun (UDUS), Julie McMurry (EMBL-EBI), John Overington (EMBL-EBI), Helen Parkinson (EMBL-EBI), Stefan Klein (Erasmus MC), Erwin Vast (Erasmus MC), Dani Welter (EMBL-EBI), Remo Sanges (SZN), Renzo Kottmann (Max Planck Institute for Marine Microbiology)



Contents

1	Executive summary	3
2	Project objectives	3
3	Detailed report on the deliverable	4
3.1	Background	4
3.2	Description of work	4
3.2.1	Collaborative activities	4
3.2.2	General Technical Strategy.....	5
3.2.3	General technical implementation - BioJS	5
3.2.4	Pilots.....	6
	Pilot 1: Sharing and integrating medical imaging data – Erasmus MC (Euro-BioImaging).....	6
	Pilot 2: Connectivity-Based Searching in UniChem – EMBL (EU OPENSCREEN).....	14
	Pilot 3: Sharing and visualising sequencing data from environmentally-derived biological samples – SZN/EMBL-EBI (EMBRC) and MPI-Bremen (Micro B3)..	18
	Pilot 4: Visualising and leveraging ontologies in queries – EMBL-EBI (Euro-BioImaging).....	23
	Pilot 5: Integrating gene and drug information with a clinical trials registry – UDUS (ECRIN)	29
	Future work	35
4	Delivery and schedule	35
5	Adjustments made.....	35
6	Background information	36



1 Executive summary

This deliverable consists of five pilot studies which integrate and visualise data from across the BioMedBridges domains and the biomedical sciences research infrastructures involved. It builds on the previous WP deliverables and activities (D4.1 to D4.4). The pilots described in this report have varied used cases and cover the following areas:

1. Sharing 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.

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 Web services interfaces optimised for browsing information	x	
3	Expose the integration via use of REST based Web services 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

The collection of REST ‘vignettes’ or visualisations presented here provides a foundation for the integration strategy of BioMedBridges to be expanded with pilots for Semantic Web integration (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 (D3.3). Each pilot is summarised in this report and is supported by technical documentation in appendices to this report and/or on the relevant website.

3.2 Description of work

3.2.1 Collaborative activities

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.

Collaborative activities involved WP4 contributors and other key BioMedBridges partners and included:

- discussion of D4.5 requirements during two meetings with EBI industry partners
- monthly consortium WP4 teleconferences
- biweekly consortium 4.5 pilot teleconferences
- Biweekly discussions with the ELIXIR technical coordinator (from November 2013)

Topics discussed during these activities include:

- 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 implementation strategy pertinent to a specific institution/service provision
- common understanding of the technology and other requirements to build the scientific bridges
- technical extensibility beyond the BioMedBridges domain to e.g. ELIXIR pilots, IMI projects, and generic data resources such as EUDAT.

3.2.2 General Technical Strategy

The collaborative activities ensure that developments fulfilled the relevant requirements identified by the D4.2 Technical workshop and points of interaction with other work packages, especially WP3; specifically:

- adopted technologies are interoperable
- use of common (or convertible) inputs and outputs
- use of common identifiers and accessions (WP3)
- 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

3.2.3 General technical implementation - BioJS

The technical strategy above informed the choice of technical platforms for implementation. Specifically, wherever possible and valuable, the BioJS framework was used. BioJS¹ 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

¹ J. Gómez et al. (2013) [BioJS: an open source JavaScript framework for biological data visualization](#). *Bioinformatics*, 10.1093/bioinformatics/btt100



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 project², whether as committers or adopters.

3.2.4 Pilots

Pilot 1: Sharing and integrating medical imaging data – Erasmus MC (Euro-Biolmaging)

Summary of work

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 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 atherosclerosis³. 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 Marketplace⁴”.

Background and scientific importance

There is an increasing demand to incorporate medical imaging into research, owing to the detailed information on anatomy, function, and pathology that images are able to provide. Examples of medical imaging modalities are magnetic resonance imaging (MRI), computed tomography (CT), and ultrasound. Often image-based information serves as surrogate endpoints for

² <http://www.ebi.ac.uk/Tools/biojs/registry/>

³ M.T.B. Truijman et al. (2013) [Plaque At RISK \(PARISK\): prospective multicenter study to improve diagnosis of high-risk carotid plaques](#). International Journal of Stroke, doi:10.1111/ijss.12167

⁴ <http://marketplace.xnat.org>



clinical research: compared to solely clinical endpoints, imaging can detect more subtle changes of pathologic developments that are related to specific treatments. Because surrogate endpoints improve efficiency in the discovery process, they can save costs by reducing study duration and patient group size.

Tools like XNAT are needed to better support multi-centre imaging research. Multi-centre projects pose particular challenges with regard to image storage, data sharing, and image analysis. Commonly, imaging data and imaging-derived results (e.g. organ volume measurements) are sent around between partners with ad-hoc methods, for example by sending CD-ROMs by post, using cloud storage services, or via temporary FTP file transfer servers. Analysis is done locally by the researchers in the different centres, often with different image processing tools. This way of working is often laborious, error-prone, and lacks provenance.

What is needed is a user-friendly data bridge for medical imaging trials that supports standardized image processing for multi-centre projects. Specifically, such a bridge will support the collection, analysis, and sharing of numerical and imaging data in a sustainable, standardized, validated manner. It will enable centralised analysis of imaging and image-derived data, and it will even support centralised correlative analysis between image-derived data, clinical data (e.g. disease status, age), and genetic data, and thus facilitate bridging between biomedical research domains.

The need for standardized and validated image analysis methods also means that representative imaging datasets should become available to the research community. This availability will enable (both academic and industrial) developers of image analysis software to validate their methods on these public datasets. For this, an infrastructure for sharing of medical imaging data among researchers is desired. Such infrastructure will also facilitate the sharing of research data underlying scientific publications and thus enable other researchers to reproduce and verify reported results.

To address these issues, web services were constructed that allow:

- collecting imaging data and image-derived data in multi-centre studies



- sharing data with collaborators in a controlled way, thanks to extensive authorization solutions for giving access rights to specific datasets and types
- web-based access and image viewing, DICOM-based access (for interfacing with radiological workstations), and programmatic access via a REST-API

The website <http://xnat.bigr.nl> links to the web services and explains how to use them, and also provides scripts for installation and maintenance.

Example scientific use cases

Use case 1 - Multi-centre clinical study with imaging

The Dutch CTMM Parisk project³ is a multi-centre clinical study on the vulnerability (risk of rupture) of atherosclerotic plaque in the carotid artery. Heterogeneous data from multiple sources is integrated: both imaging data (MRI, CT, Ultrasound) and non-imaging data (patient metadata) are collected, both raw (unprocessed imaging) data and derived data (annotations, automated image processing) are stored, data is produced and analysed by multiple institutions, and different data types will be related to each other. The institutions involved are Maastricht UMC, UMC Utrecht, Erasmus MC (data production and analysis), LUMC, Philips, and PieMedical Imaging (for data analysis).

The following four items describe the requirements for an image management e-infrastructure in this use case:

1. Images are acquired in multiple institutions, and should be stored centrally (after proper anonymisation).
2. Centrally stored image data needs to be accessed by multiple institutions, in three different ways:
 - a. via DICOM communication protocol (using a radiological workstation).
 - b. via a web-based viewer.
 - c. via an API, i.e. programmatic access, to facilitate automated image analysis.



3. Manual and automatic image analysis results (henceforth referred to as *annotations*) produced by multiple institutions need to be stored centrally.
4. Medical researchers from multiple institutions need to access these annotations and analyse these in relation to the clinical data.

Use case 2 - Sharing of medical imaging data with researchers worldwide

In the field of image analysis, there is an increasing trend to share research data. Examples from Erasmus MC are The BIGR Ergo Carotid Study⁵ and The Carotid Distensibility Study⁶. Building a new website for each project is time-consuming and unsustainable (e.g. who maintains and supports the site after the principal investigator leaves?). Moreover, the research data are scattered over multiple sites, which typically are not searchable. Repositories for sharing imaging data are therefore required.

The requirements for an e-infrastructure in this use case are:

1. A standardized data organization should be imposed.
2. The site should allow automated data upload and download, to avoid laborious manual procedures.
3. The data should be exposable via an API (REST or SOAP), to allow the construction of meta-registries searching across image databases of different groups, and to allow interfacing with external analysis and image viewing software.
4. Not only DICOM data should be supported, since in most research on image analysis other imaging formats, such as Nifti and MetalImage, are used as well, and even non-imaging data (annotations, patient information, image-derived results) in XML or CSV formats.

Technical implementation

In this WP4 pilot study, Erasmus MC successfully adopted and deployed the XNAT platform⁷ to facilitate storage and management of raw imaging data (in DICOM format, or Nifti, Analyze, NRRD) and image-derived annotations. XNAT enables access to data in various ways: via the DICOM protocol, via the web-interface, and via a REST API.

⁵ <http://ergocar.bigr.nl> [Hameeteman et al, 2013, Phys. Med. Biol]

⁶ <http://ctadist.bigr.nl> [Hameeteman et al, 2013, MedIA]

⁷ www.xnat.org



The focus of our technical developments for BioMedBridges was threefold:

- A. simplify installation and maintenance of XNAT-based web services
- B. develop documentation to assist researchers in using XNAT services, allowing them to upload, download and view images in multiple ways
- C. set up XNAT web services and apply these services in four projects.

A) Simplify installation and maintenance of XNAT services

To enable research centres to install their own XNAT service, a script was developed that installs XNAT on a new system. This script makes it easier to set up new data sharing services compared to manual installations and lays a foundation to include more steps in the script to install additional software. The script is based on the Puppet software⁸, making it more convenient to use in existing systems and to use the script on a cluster of machines, as it supports a master/client architecture. The script can install both a new virtual machine (VM) with the XNAT software and it can install the XNAT software on an existing (virtual) machine. It has been extensively tested with the operating systems Ubuntu, Redhat and Fedora Linux. The VM creator is available on: https://bitbucket.org/bigr_erasmusmc/xnat-vm-creator and the Puppet script on: https://bitbucket.org/bigr_erasmusmc/puppet-xnat/. To make the existence of the scripts known, they were also uploaded to the XNAT Marketplace⁹, which is a collection of publicly available tools and plugins for XNAT.

A test script is provided that uploads and downloads the software and checks to ensure that it works. A cron-job from a remote machine can be used to nightly test the availability of the system. This supports the sustainability of the system as network or firewall problems can be detected before users become aware of this. The test script can be found on: https://bitbucket.org/bigr_erasmusmc/xnat/.

B) Using the XNAT services

The automation of data uploads as required for Use Case 2 is enabled by using the existing Pyxnat library (<https://github.com/pyxnat/pyxnat>). Pyxnat is a

⁸ <http://puppetlabs.com/>

⁹ <http://marketplace.xnat.org>



wrapper, written in the programming language Python, around the XNAT REST interface to increase the usability of XNAT for researchers. With Pyxnat, one can browse the repository, perform queries, and download and upload imaging data. This enables researchers to automate data uploads and downloads, create data analysis pipelines that use the data from XNAT (or another source) and store the resulting image-derived data in XNAT. To get new users started with data uploads, an example Pyxnat script was provided at <http://xnat.bigr.nl>. Furthermore, this website also provides a quick-start guide for new users, including information about how to upload data and manage access rights.

The popular 3DSlicer package¹⁰ for medical image visualization and analysis also provides a module that interfaces with XNAT via the REST interface. This demonstrates how the REST interface enables other applications to visualize data from the web-service. We installed the XNATSlicer module (<https://github.com/skumar221/XNATSlicer>), tested it and provided feedback for the developer to increase the usability of this module. The result of the feedback from Erasmus MC (and other research institutions) was that the XNATSlicer module now requires fewer actions for the user to import the data in 3DSlicer.

The XNAT DICOM gateway also makes DICOM data available via a DICOM interface, such that any existing DICOM viewers can be used to inspect the data. This has been tested and works satisfactorily. This enables radiologists using workstations that work specifically with DICOM to exchange scan session data with the XNAT service.

C) Providing XNAT services

Using the new installation script it is easy to configure a new XNAT service for various projects that require their own XNAT service. Currently, several instances of XNAT are hosted within Erasmus MC for testing purposes and internal projects. For BioMedBridges, an XNAT service was set up in the Demilitarised Zone (DMZ) of Erasmus MC (outside the hospital firewall), which is accessible from everywhere (<https://bigr-xnat.erasmusmc.nl>, Figure 1). For this service, nightly tests (see item A above) and backups ensure that no data

¹⁰ www.slicer.org



is lost in case of hardware problems. Imaging data of three research projects was uploaded using Pyxnat (see item B above).

In collaboration with CTMM TraIT (EATRIS/BBMRI), an instance has been hosted outside Erasmus MC, accessible from everywhere (<https://xnat.bmia.nl>). For CTMM TraIT, Vancis is the company that provides hosting. Their technician executed our installation script (see item A above). This demonstrates that the script makes it possible to deploy the software anywhere, and that it improves sustainability since it automatically documents the exact installation procedure, and it is written such that any skilled IT technician could operate it.

The screenshot displays the XNAT web interface. The main content area shows the 'MR Session: case001' details, including the subject name 'digmnt_E00004', date '10/15/2013', and scanner 'OE MEDICAL SYSTEMS GENERIS_BIONA'. Below this is a table of scans with columns for Scan ID, Type, Series Desc, Usability, Files, and Note. A 'History' link is visible below the table. On the right side, there is a 'Session Information' panel with fields for Session ID, LAD ID, SEX, AGE, SEX, ADDRESS, ACQ. DATE, SCANNER, STABILIZATION, REF. NUMBER, and INVESTIGATOR. At the bottom right, a small window shows a brain scan image with a coordinate system overlay.

Scan	Type	Series Desc	Usability	Files	Note
#1	SAO LOCALIZER	SAO LOCALIZER	usable	Show Counts	
#2	T1 SPIN ECHO	T1 SPIN ECHO	usable	Show Counts	
#3	SCAN 1 T2 SCAN 2 PD	SCAN 1 T2 SCAN 2 PD	usable	Show Counts	
#4	SAO LOCALIZER	SAO LOCALIZER	usable	Show Counts	
#5	3D SPOR VOLUME	3D SPOR VOLUME	usable	Show Counts	
#6	3D SPOR VOLUME	3D SPOR VOLUME	usable	Show Counts	
#7	3D SPOR VOLUME	3D SPOR VOLUME	usable	Show Counts	
#8	T1 SPIN ECHO	T1 SPIN ECHO	usable	Show Counts	
#9	HEME SUSCEPTIBILITY	HEME SUSCEPTIBILITY	usable	Show Counts	
#10	T1 SPIN ECHO	T1 SPIN ECHO	usable	Show Counts	
				Total Counts	

Figure 1 Screenshot of the <https://bigr-xnat.erasmusmc.nl> interface

Example data analysis

Figure 1 shows the interface of the installed service with an example brain scan dataset.

Regarding Use Case 1, all currently collected CTMM Parisk data (~500GB) have been uploaded to the XNAT instance¹¹. To expose the Parisk dataset to researchers outside the CTMM Parisk project, an entry has been created in the BioSamples database¹². External collaborations may thereby be proposed;

¹¹ <https://xnat.bmia.nl>

¹² www.ebi.ac.uk/biosamples



all such proposals would be discussed with the Parisk project coordinators on a case-by-case basis. Once the request is approved, the researcher will get an account and the correct authorization options will be set in the XNAT software to enable download of the requested data.

Regarding Use Case 2, three large datasets were uploaded to the <https://bigr-xnat.erasmusmc.nl> service. These three datasets were also registered in the BioSamples database¹³. Below, a short description of each dataset is given:

CTADIST - Carotid artery 4D CTA:

This data was used in the following publication: "K. Hameeteman, S. Rozie, C.T. Metz, R. Manniesing, T. van Walsum, A. van der Lugt, W.J. Niessen and S. Klein, Automatic carotid artery distensibility measurements from CTA using nonrigid registration, Medical Image Analysis, July 2013, vol. 17, pp. 515-524". It includes 75 four-dimensional CT angiography (CTA) scans, with information about several risk factors of the patients.

ErgoCar - Carotid Artery MRI:

This data belongs to the following publication: "K. Hameeteman, R. van 't Klooster, M Selwaness, A. van der Lugt, J.C.M. Witteman, W.J. Niessen and S. Klein, Carotid wall volume quantification from magnetic resonance images using deformable model fitting and learning-based correction of systematic errors, Physics in Medicine and Biology, 2013, vol. 58, pp. 1605". It includes 38 scans of the carotid artery, including manual annotations of the lumen and outer vessel wall.

Prostate MRI:

Image data and segmentations that were used in the following paper: "S. Klein, U.A. van der Heide, I.M. Lips, M. van Vulpen, M. Staring and J.P.W. Pluim, Automatic segmentation of the prostate in 3D MR images by atlas matching using localized mutual information, Medical Physics, April 2008, vol. 35, pp. 1407-1417." Patients 101 to 150 correspond to section III.A.2 of the paper. For these datasets, three manual segmentations are available, and one combined segmentation (by majority vote). Volunteers 1 to 8 correspond to section III.A.1 of the paper. Each volunteer is scanned 5 times. For these datasets, 1 manual segmentation is available per image. <http://ergocar.bigr.nl>

All imaging data was anonymised before uploading. Researchers worldwide are invited to create an account, request access to any of these datasets, download the data, and use the data to verify our findings, to develop and evaluate new image processing methods, and for further applied research.

Future work

¹³ <http://www.ebi.ac.uk/biosamples/>



First, we will further extend the link between XNAT and the 3DSlicer software. With this, we aim to streamline the typical analyses performed by human readers on imaging data. The envisioned workflow is: 1) user selects and retrieves image from XNAT database (made possible thanks to the REST interface), 2) user performs analysis with the tools provided by 3DSlicer, 3) the resulting image-derived data is directly uploaded to the XNAT database (again, via the REST interface), and stored such that the link with the original data is preserved.

Second, we will develop a solution for centralised correlative analysis between image-derived data stored in the XNAT database and clinical data stored in an OpenClinica¹⁴ database. To this end, we will explore the use of TransSMART¹⁵ and exploit the REST interfaces of XNAT and OpenClinica.

Pilot 2: Connectivity-Based Searching in UniChem – EMBL (EU OPENSREEN

Summary of work

EU-OPENSREEN developed a user-friendly web interface for UniChem's refined "Connectivity-Based Search" 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.

Background and scientific importance

Unichem¹⁶ serves as the chemistry data integration layer for EU-OPENSREEN, ChEMBL, ChEBI and other chemically aware databases. It achieves this integration by mapping identifiers from these different resources to their corresponding standard InChI identifier.

A weakness of the standard InChI mechanism is that, very often, different resources will depict the same molecule in a slightly different way (e.g. with different stereochemistry), or as a different salt form. Sometimes, these differences are due to curation errors, but sometimes they are genuinely different molecules, albeit with the same covalent connectivity. Therefore,

¹⁴ www.openclinica.com

¹⁵ <http://transmartfoundation.org>

¹⁶ www.ebi.ac.uk/unichem/



mapping between the similar molecules in different resources required additional reasoning—which was the basis of the work done as part of this deliverable: the Standard InChI has been cleverly designed to allow the depiction of molecules at different levels of granularity. These levels are expressed as ‘layers’ in the InChI string itself. By effectively ‘peeling back’ the layers of the InChI, and then comparing the shorter forms of the InChI strings, one is able to map between compounds with identical connectivity, but different stereochemistry and isotopic composition. The InChI strings of mixtures and salts also lends itself to be easily parsed into separate InChI strings, and thus allow mapping between different salt forms of the same parent molecule.

As a result of this work, UniChem is now able to offer a service which permits users to generate (on the fly, if required) bridges or links between *similar molecules* as well as between *identical* molecules. The InChI mechanism was not itself changed; rather, its richness was leveraged in order to overcome some of its inherent limitations.

Example scientific use case and data analysis

As with the InChI mechanism overall, the Connectivity-Based Search application was developed at the direct request of industry partners and has been iteratively refined in accordance with industry feedback and therefore connects BioMedBridges to Pharma requirements.

Suppose a user of UniChem wishes to know whether other resources contain information on a particular molecule. Using ChEMBL25 (Aspirin) as an example, the structure for ChEMBL25 within the ChEMBL database may be represented by the InChIKey BSYNRYMUTXBXSQ-UHFFFAOYSA-N. Querying with this InChIKey, or the ChEMBL Id itself, on the main UniChem page¹⁷ returns records showing where else this *identical* structure may be found in other sources¹⁸ (Figure 2).

¹⁷ <https://www.ebi.ac.uk/unichem/>

¹⁸ https://www.ebi.ac.uk/unichem/frontpage/results?queryText=ChEMBL25&kind=src_compound_id&sources=1&incl=exclude



src_id	Source Name	src_compound_id	Currently Assigned	LR *	UCI **	Standard InChIKey
1	chembl	CHEMBL25	Yes		161671	BSYNRYMUTXBXSQ-UHFFFAOYSA-N
2	drugbank	DB00945	Yes		161671	BSYNRYMUTXBXSQ-UHFFFAOYSA-N
3	pdb	AIN	Yes		161671	BSYNRYMUTXBXSQ-UHFFFAOYSA-N
5	pubchem_doff	24714725	Yes		161671	BSYNRYMUTXBXSQ-UHFFFAOYSA-N
6	kegg_ligand	C01405	Yes		161671	BSYNRYMUTXBXSQ-UHFFFAOYSA-N
7	chebi	15385	Yes		161671	BSYNRYMUTXBXSQ-UHFFFAOYSA-N
10	emolecules	474821	Yes		161671	BSYNRYMUTXBXSQ-UHFFFAOYSA-N
11	ibm	DB938244616517D179E20E04720971A1	Yes		161671	BSYNRYMUTXBXSQ-UHFFFAOYSA-N
12	atlas	aspirin	Yes		161671	BSYNRYMUTXBXSQ-UHFFFAOYSA-N
12	atlas	acetylsalicylic acid	Yes		161671	BSYNRYMUTXBXSQ-UHFFFAOYSA-N
14	fdasrs	R16CO5Y76E	Yes		161671	BSYNRYMUTXBXSQ-UHFFFAOYSA-N
17	pharmgkb	PA448497	Yes		161671	BSYNRYMUTXBXSQ-UHFFFAOYSA-N
18	hmdb	HMDB01879	Yes		161671	BSYNRYMUTXBXSQ-UHFFFAOYSA-N
20	selleck	aspirin-acetylsalicylic-acid	Yes		161671	BSYNRYMUTXBXSQ-UHFFFAOYSA-N
21	pubchem_tpharma	15195166	Yes		161671	BSYNRYMUTXBXSQ-UHFFFAOYSA-N

Figure 2 The InChI resolver widget returns a list of all compounds with identical structure to Aspirin (CHEMBL25)

This is how users have been using UniChem since its exposure to the external community at the start of this year. However, most users are interested to know if variant forms of Aspirin also exist. Thus, querying with these same terms using the interface developed for BioMedBridges¹⁹ (and setting the 'C' query option to '4' returns over 200 records showing where else *identical and similar* molecules exist²⁰ (Figure 3), and the data structure returned includes information which allows the user to sort/filter/parse the data easily to discover in what specific ways the similar molecules differ from the query (e.g. sorting on the 'i' column allows users to easily identify that an isotopic form of Aspirin exists in the 'eMolecules' source).

Relationship. Query InChI...	src_id	Source	src_compound_id	Asn	label	p	b	t	m	s	i	En	src_compound_id InChIKey	#
...matches...	1	CHEMBL	CHEMBL25	1		0	0	0	0	0	0		BSYNRYMUTXBXSQ-UHFFFAOYSA-N	1
...matches...	1	CHEMBL	CHEMBL447221	1		0	0	0	0	0	0		BSYNRYMUTXBXSQ-UHFFFAOYSA-M	1
...matches...	2	DrugBank	DB00945	1		0	0	0	0	0	0		BSYNRYMUTXBXSQ-UHFFFAOYSA-N	1
...matches...	3	PDBe	AIN	1		0	0	0	0	0	0		BSYNRYMUTXBXSQ-UHFFFAOYSA-N	1
...matches...	5	PubChem: Drugs of the Future	24714725	1		0	0	0	0	0	0		BSYNRYMUTXBXSQ-UHFFFAOYSA-N	1
...matches...	6	KEGG Ligand	C01405	1		0	0	0	0	0	0		BSYNRYMUTXBXSQ-UHFFFAOYSA-N	1
...matches...	7	CHEBI	13719	1		0	0	0	0	0	0		BSYNRYMUTXBXSQ-UHFFFAOYSA-M	1
...matches...	7	CHEBI	15385	1		0	0	0	0	0	0		BSYNRYMUTXBXSQ-UHFFFAOYSA-N	1
...matches...	9	ZINC	ZINC00000053	1		0	0	0	0	0	0		BSYNRYMUTXBXSQ-UHFFFAOYSA-M	1

Figure 3 The InChI resolver widget returns a list of all compounds with similar structure to Aspirin (CHEMBL25)

Technical implementation

Implementation has required some modification to the UniChem data model (extra tables to 'cache' mappings between InChIKey connectivity layers of mixture/salt forms and the separate components). In addition, extra components to the application layer and data loaders have been introduced

¹⁹ <https://www.ebi.ac.uk/unicem/widesearch/widesearch>

²⁰ [Example connectivity search in UniChem for 'Aspirin'](#)



which permit InChIs to be parsed and compared very quickly. A web application for non-programmatic users and REST service methods for programmatic users have both been developed. Beta versions of both can be accessed and used outside of the EBI, as can the documentation²¹ (Figure 4). 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.

EMBL-EBI

UniChem

- Home / Search
- Web Services
- Whole source mapping
- Sources
- Auxiliary data mapping
- General Info...
 - Background
 - Getting in touch
 - FAQ
 - Stats
 - + Other

EBI > Databases > Small Molecules > UniChem

Connectivity-Based Searching Documentation.

Contents

- 1 [Introduction](#)
- 2 [Query Construction](#)
- 3 [Search Terms](#)
 - 3.1 [General](#)
 - 3.2 ['key_search' Search Terms](#)
 - 3.3 ['cpd_search' Search Terms](#)
- 4 [Search Criteria A-H](#)
 - 4.1 [General](#)
 - 4.2 [A. Sources](#)
 - 4.3 [B. Pattern](#)
 - 4.4 [C. Relationships](#)
 - 4.5 [D. Frequency Block](#)
 - 4.6 [E. InChI Length Block](#)
 - 4.7 [F. UniChem Labels](#)
 - 4.8 [G. Assignment Status](#)
 - 4.9 [H. Data Structure](#)
- 5 [How UniChem Avoids the Problem of Returning Unwanted Data](#)
- 6 [Further Background Information on Connectivity Searching](#)
 - 6.1 [General](#)
 - 6.2 ['Keys Only' Sources](#)
 - 6.3 [Multiply assigned src_compound_ids](#)
 - 6.4 [Querying with InChIKeys that do not exist in UniChem](#)
 - 6.5 [Working within the constraints of InChI](#)

Figure 4 UniChem Connectivity-Based Search Documentation page

Future work

The existence of the Connectivity-Based Search functionality, both via REST and via a friendly web application, makes it possible for users to better align data from different resources on the basis of shared chemical compounds. One such plan is to identify drugs within the clinical trials metadata from ClinicalTrials.gov. Planned enhancements to the web application include

²¹ <https://www.ebi.ac.uk/unicem/info/widesearchInfo>



images corresponding to the chemical structures. The work will also be submitted for publication.

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

Summary of work

The role of EMBRC is to provide access to marine organisms, techniques, and data to the scientific community at large, including universities and industry. Accordingly, EMBRC (SZN/EMBL-EBI and MPI Bremen via the Micro B3 project²²) contributed *in-kind* to this deliverable by

- a) submitting to the European Nucleotide Archive (ENA) sequencing data from marine-derived bacteria and
- b) extending their geologic mapping widget to make it maximally useful to other research groups.

Background and Scientific Importance

The growing use of Next Generation sequencing technology has made metagenomes and metatranscriptomes widely available, but has also presented new bioinformatics challenges. Numerous emerging analytical techniques are being extended beyond their traditional use in diversity studies (i.e. in biogeochemical studies and climate change manipulations). The Stazione Zoologica Anton Dohrn (SZN) in Naples is conducting an investigation of the heterotrophic bacterial communities of different sites of the Gulf of Naples, encompassing different ecological conditions, in order to find microbiological indicators of ecological status. It has submitted sequencing data for the first two such samples. In order to view these and other environmentally-derived samples in their full geological context, mapping software is especially helpful. For example, of the two samples submitted to ENA, one was taken from the Sarno River closest to Naples and the other was taken from the middle of the Gulf of Naples.

The megx.net portal for Marine Ecological GenomiX is a web site for specialized georeferenced databases and tools for the analysis of marine

²² <http://www.microb3.eu/>



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 B3²²), which facilitates the interpretation of the sequence in its environmental context via a browsable world map. The unique strength of the Genes Mapserver is its Geographic Information System (GIS), which integrates environmental data layers extracted from the World Ocean Atlas (WOA) and World Ocean Database (WOD), among others. Currently, 'on the fly' interpolated data include temperature, salinity, dissolved oxygen, apparent oxygen utilization, percent oxygen saturation, phosphate, silicate, and nitrate at standard depths, averaged over annual, seasonal, and monthly periods for any location in the marine system.

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 project²³. 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.

Example scientific use case

In a pilot analysis, bacteria were sampled at two locations in the Gulf of Naples and DNaseq of their metagenomes was performed. Data (1 billion raw reads) were uploaded to the EMBL-EBI SRA archive²⁴ using the FTP. Once in SRA, the data entered the Metagenomic pipeline in order to be analyzed.

Technical implementation

The EMBL-EBI Metagenomics service identifies rRNA sequences, using rRNASelector, and performs taxonomic analysis upon 16S rRNAs using Qiime. The remaining reads are submitted for functional analysis of predicted

²³ <http://www.ebi.ac.uk/Tools/biojs/registry/>

²⁴ http://www.ebi.ac.uk/ena/about/sra_submissions



protein coding sequences using the InterPro sequence analysis resource²⁵. InterPro uses diagnostic models to classify sequences into families and to predict the presence of functionally important domains and sites. Data submitted to the EBI Metagenomics service is automatically archived in the SRA, which is part of the European Nucleotide Archive (ENA). Accession numbers are supplied for sequence data as part of the archiving process; the data embargo will be lifted following publication.

Example data analysis

Once the data were uploaded and checked, they entered the EMBL-EBI Metagenomics pipeline described above. At the end of the analyses, the results were present in the EMBL-EBI Metagenomics page where they could be evaluated and downloaded (Figure 5, Figure 6 and Figure 7).

The screenshot shows the EBI Metagenomics interface for sample MICGEN7. The page is titled 'Sample (ERS255957) Tyrrhenian Sea surface water sample MICGEN7'. It features several sections:

- Description:** Tyrrhenian Sea surface water sample MICGEN7. Classification: Environmental > Aquatic > Marine > Saline > Sea water. Project: Marine microbial indicators in coastal areas: a metagenomic approach (ERP003408).
- Environmental conditions:** Biome: Dilution basin mediterranean. Experimental feature: Sea env:0000024. Material: Surface water env:00002042.
- Localisation:** Latitude/Longitude: 40.583, 13.9497. Geographic location: Tyrrhenian Sea. A map shows the location in the Tyrrhenian Sea, near Italy.
- Other information:** Project name: MICGEN. Geographic location (depth): 0 m. Instrument model: Illumina HiSeq 2000.

Figure 5 Overview of the sample MICGEN7 reporting experimental information about the experiment as well as the geographic location with the origin of the sample

²⁵ <http://www.ebi.ac.uk/interpro/>



Sample (ERS255956)

Tyrrhenian Sea surface water sample MICGEN6

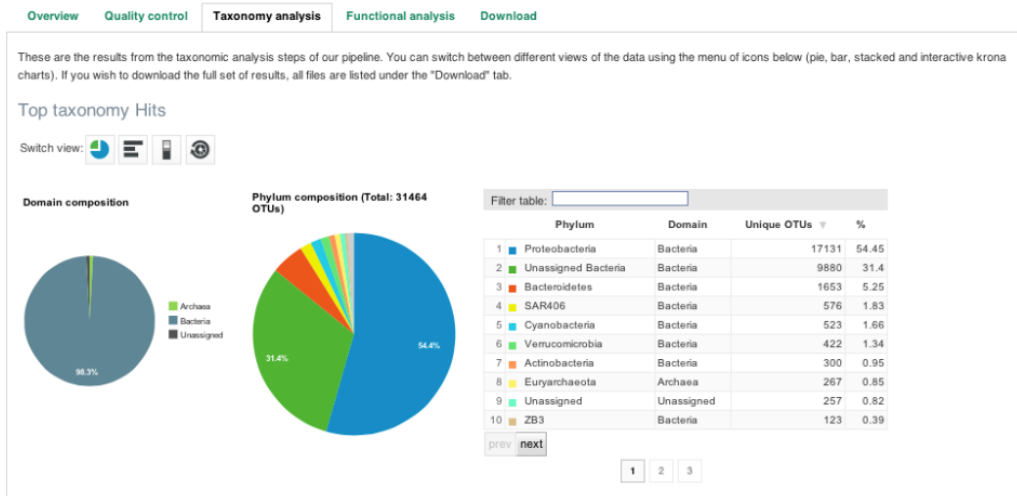


Figure 6 Example of the Taxonomy analysis from the sample MICGEN6 indicating the relative proportion of the different species identified inside the sample by the EBI Metagenomic pipeline

EMBL-EBI Services Research Training Industry About us

EBI Metagenomics

Home Submit data Projects **Samples** About Metagenomics Contact Cecilia Balestra (edit) logout

EBI Metagenomics > Project: Marine microbial indicators in coastal areas: a ... > Sample: Tyrrhenian Sea surface water sample MICGEN6

Sample (ERS255956)

Tyrrhenian Sea surface water sample MICGEN6

Overview Quality control **Taxonomy analysis** Functional analysis Download

You can download in this section the full set of analysis results files and the original raw sequence reads.

Sequence data

- Submitted nucleotide reads (ENA website)
- Processed nucleotide reads (FASTA) - 42266 MB
- Processed reads with pCDS (FASTA) - 40196 MB
- Processed reads with InterPro matches (FASTA) - 8111 MB
- Processed reads without InterPro match (FASTA) - 32084 MB
- Predicted CDS (FASTA) - 23009 MB

Functional Analysis

- InterPro matches (TSV) - 15016 MB
- Complete GO annotation (CSV) - 148 KB
- GO slim annotation (CSV) - 7 KB

Taxonomic Analysis



- Reads encoding 5S rRNA (FASTA) - 1 MB
- Reads encoding 16S rRNA (FASTA) - 20 MB
- Reads encoding 23S rRNA (FASTA) - 30 MB
- OTUs and taxonomic assignments (BIOM)  - 3 MB
- Phylogenetic tree (Newick format)  - 186 KB
- OTUs and taxonomic assignments (TSV) - 3 MB

Figure 7 Download page of the MICGEN6 sample. From this page it is possible to download all the results produced by the EBI Metagenomics pipeline

It should be mentioned that the submission of sequencing data with an unprecedented half-billion reads caused several ENA processing jobs to fail, mainly because of high memory requirements. The ENA software team therefore made substantial changes to ensure that memory was managed



adequately and the submission was successful. This has therefore been useful in benchmarking existing processes to support future submissions by EMBRC, and also represents an in-kind contribution.

Proof-of-concept for the megx mapping tool has been done with geocoded data that is openly available in ENA (as the newly submitted EMBRC data is currently embargoed, preventing its use before publication).

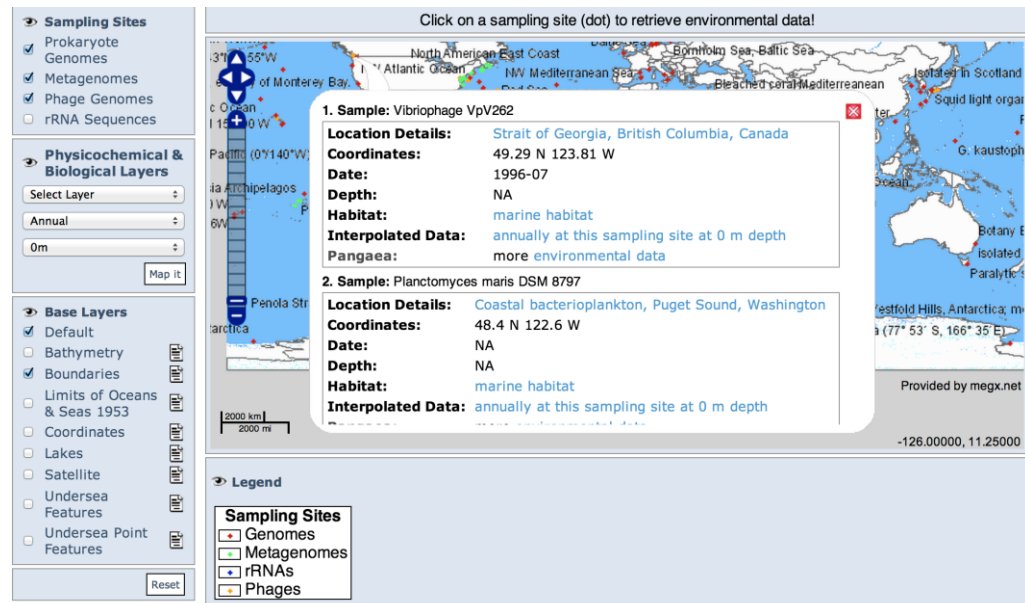


Figure 8 Megx' Genes Mapserver view, showing the environmental context for two marine-derived specimens that were sequenced



Figure 9 Locations for all geo-tagged samples within the European Nucleotide Archive overlaid for demonstration purposes with World Ocean Atlas Temperatures



Future plans

The Megx BioJS widget will be further developed so that it will accept accession numbers and data from other relevant databases beyond those in the ENA. The layered architecture of Megx could be leveraged for different applications; we plan to explore adaptations of the Megx widget to map samples within the BioSamples Database (building on work from deliverable 4.3). As an additional point of integration, ten tools were listed on the Megx tool page; these have now been registered in the BioMedBridges tools and data services registry²⁶ and will undergo additional annotation in 2014.

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

Summary of work

For this deliverable, EMBL-EBI (for Euro-Biolmaging) developed PLATO (Plugin for Autocomplete on Ontologies), which is an embeddable JavaScript widget for browsing and searching ontologies. This widget was developed to be generically configurable for use with multiple ontologies. For demonstration and evaluation purposes, it is being deployed at wwwdev.ebi.ac.uk/fqpt/PLATOdemo for use with three different ontologies: the Cellular Microscopy Phenotype Ontology (CMPO²⁷), the Experimental Factor Ontology (EFO²⁸), and EDAM (Embrace Data and Methods²⁹), each of which is being actively used within BioMedBridges.

PLATO's features 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 project³⁰.

Background and scientific importance

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

²⁶ www.tinyurl.com/bmbtoolsui

²⁷ <http://www.ebi.ac.uk/ontology-lookup/browse.do?ontName=CMPO>

²⁸ <http://www.ebi.ac.uk/efo/>

²⁹ <http://edamontology.org/page>

³⁰ <http://www.ebi.ac.uk/Tools/biojs/registry/>



broadly: of the twenty institutions within the BioJS community, virtually all have expressed some interest in re-using 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, PLATO and its accompanying ontology REST service backend were therefore developed using an API from the Ontology Lookup Service³¹ at the EBI.

Example scientific use cases

Use case 1: WP6/research imaging)

The cellular microscopy phenotype ontology³², now released, was purpose-built for integration of phenotypes generated for image data. WP6 partners have extended the Zooma³³ autocuration tool in order to match CMPO ontology terms to existing name-value pair descriptions provided by experimentalists with their data. While WP6 focus is on data annotation and semantic data integration, WP4 has built tools that address the complementary challenge of semantically supported data search and retrieval (for example via a query portal). The embeddable JavaScript widget developed for this task provides autocomplete and expandable/collapsible nodes on matched ontology terms. Together, this harmonization across image datasets from different biological scales—when paired with semantically-aware search technologies—will make it possible for users perform flexible queries and more comprehensive searches.

³¹ www.ebi.ac.uk/ols

³² <http://systemsmicroscopy.eu/first-release-cellular-microscopy-phenotype-ontology-cmpo>

³³ <http://www.ebi.ac.uk/fgpt/zooma/>

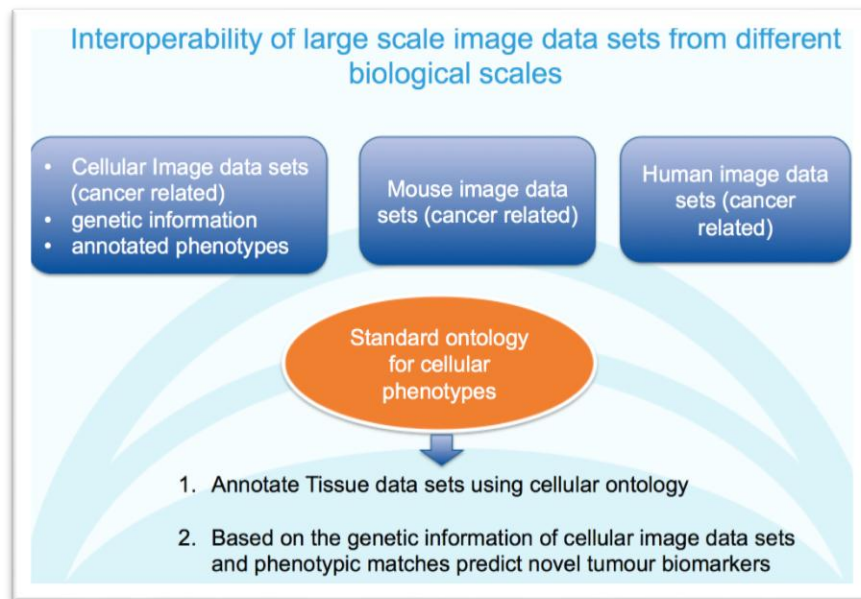


Figure 10 Data workflow for integrating large-scale image datasets from different biological scales

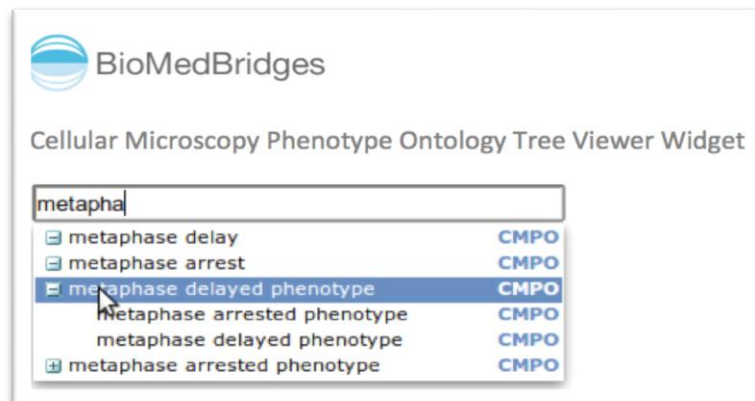


Figure 11 PLATO deployed with the CMPO ontology

Use case 2: Human-mouse data bridge

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. One of the data sources used in WP7 is the NHGRI GWAS Catalogue of curated SNP-trait associations. In this project, free text trait associations were mapped to ontology terms, work which is in progress for other WP7 datasets as part of BioMedBridges. As a proof of principle for WP7 usability and technical feasibility, PLATO is being deployed for use in the GWAS Catalogue (Figure 12), which uses the Experimental Factor Ontology²⁸



available from OLS. From the deployed catalogue³⁴, we will gather feedback from users and curators from the GWAS community to assess the utility of the widget for this data type. It will also serve to demonstrate how this widget can be used within a semantic web technologydriven visualisation which we expect to deliver for WP7 later in the project. For example, the widget allows the researcher to choose very specific terms (e.g. type II diabetes mellitus) or more general parent terms that will return more data (i.e. selection of metabolic disease will pull all GWAS terms associated with by an abnormal metabolic process, including several other diseases/conditions).

In addition to PLATO-EFO being used in the GWAS viewer, PLATO-EFO could also be used separately within GenoBridge, a tool developed within WP7. GenoBridge will systematically map human-mouse syntenic regions, thus allowing discovery of functional conservation (gene, variation, regulation and phenotype) and assisting candidate disease genes/region discovery/validation. This data will be annotated with relevant ontologies for phenotypes, disease and/or gene functions. In turn, PLATO could be incorporated into the GenoBridge interface in order to provide semantic searches over the annotated data.

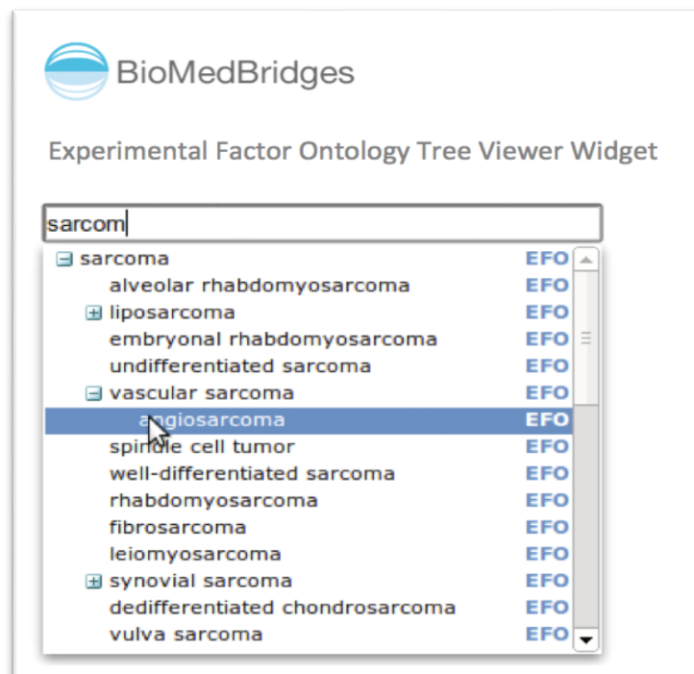


Figure 12 PLATO deployed with the Experimental Factor Ontology

³⁴ <http://wwwdev.ebi.ac.uk/fgpt/PLATOdemo>



Use case 3: WP3 / BioMedBridges Tools and Data Services Registry

A BioMedBridges tools and data services registry³⁵ is being developed as part of WP3. Each of the infrastructures has contributed metadata to the effort and has annotated the metadata using the EDAM ontology. To browse the EDAM ontology and search for matching tools, a widget is needed.

data	Description	Topics	Functions	Input Types	Output Types	Web UI	REST API
Data imputation							
Data retrieval							
Data handling	Speed multiple sequence...	Sequence align...	Multiple sequence align...	Protein or nucleotide sequenc...	Multiple sequence alignment	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
Data submission	ral purpose global multi...	Sequence align...	Multiple sequence align...	Protein or nucleotide sequenc...	Multiple sequence alignment	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
Data anonymisation	genetic tree generation ...	Phylogenetics	Phylogenetic tree const...	Protein or nucleotide sequenc...	Phylogenetic tree	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
Data management	l multiple sequence align...	Sequence align...	Multiple sequence align...	Protein BLAST result and corr...	Multiple sequence alignment	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
Data processing and validation	nd accurate multiple se...	Sequence align...		Protein or nucleotide sequenc...	Multiple sequence alignment	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
Data visualisation	nternal duplications by c...	Sequence align...	Pairwise sequence align...	Two protein or nucleotide seq...	Pairwise sequence alignment	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
Data search and retrieval	nce alignment using th...	Sequence align...	Multiple sequence align...	Protein or nucleotide sequenc...	Multiple sequence alignment	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
Data retrieval (metadata and documentation)	ved version of the Nee...	Sequence align...	Pairwise sequence align...	Two protein or nucleotide seq...	Pairwise sequence alignment	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
Data security	are two DNA sequences...	Sequence align...	Pairwise sequence align...	Two nucleotide sequences	Pairwise sequence alignment	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
Data retrieval (database metadata)	le sequence alignment t...	Sequence align...	Multiple sequence align...	Protein or nucleotide sequenc...	Multiple sequence alignment	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
Data management	two sequences assumin...	Sequence align...	Pairwise sequence align...	Two nucleotide sequences	Sequence alignment (nucleic aci...	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
Protein NMR data	le sequence alignment ...	Sequence align...	Multiple sequence align...	Protein or nucleotide sequenc...	Multiple sequence alignment	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
TMF library for data protection concept	man-Eggert local align...	Sequence align...	Pairwise sequence align...	Two protein or nucleotide seq...	Sequence alignment (protein pai...	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
Existing metadata linked	erman-Wunsch global all...	Sequence align...	Pairwise sequence align...	Two protein or nucleotide seq...	Pairwise sequence alignment	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
EG Bacteria data slicer	-Waterman local pairwi...	Sequence align...	Pairwise sequence align...	Two protein or nucleotide seq...	Pairwise sequence alignment	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
EG Fungi data slicer	v reformats the results ...	Data processin...	Sequence alignment ref...	Protein or nucleotide sequenc...	Multiple sequence alignment	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
Protein Binding Microarray data	le sequence alignment ...	Sequence align...		protein or nucleotide sequenc...	multiple sequence alignment (vl...	<input type="checkbox"/>	<input type="checkbox"/>
1000Genomes data slicer	nce search with Blast	Sequence align...		protein or nucleotide sequenc...	multiple sequence alignment (vl...	<input type="checkbox"/>	<input type="checkbox"/>
	h and alignment of prot...	Sequence align...		Protein sequence	Ranked list of hits and pairwise ...	<input type="checkbox"/>	<input type="checkbox"/>
	ral sequence alignment ...	Sequence align...				<input type="checkbox"/>	<input type="checkbox"/>
	mbilistic multiple alignment	Sequence align...	Multiple sequence align...	Sequence	Sequence alignment	<input checked="" type="checkbox"/>	<input type="checkbox"/>
	Tool	Fast and accurate multiple all...	Multiple sequence align...	Sequence	Sequence alignment	<input checked="" type="checkbox"/>	<input type="checkbox"/>
	Database	A large collection of protein f...				<input type="checkbox"/>	<input type="checkbox"/>

Figure 13 The BioMedBridges registry query interface (without PLATO)

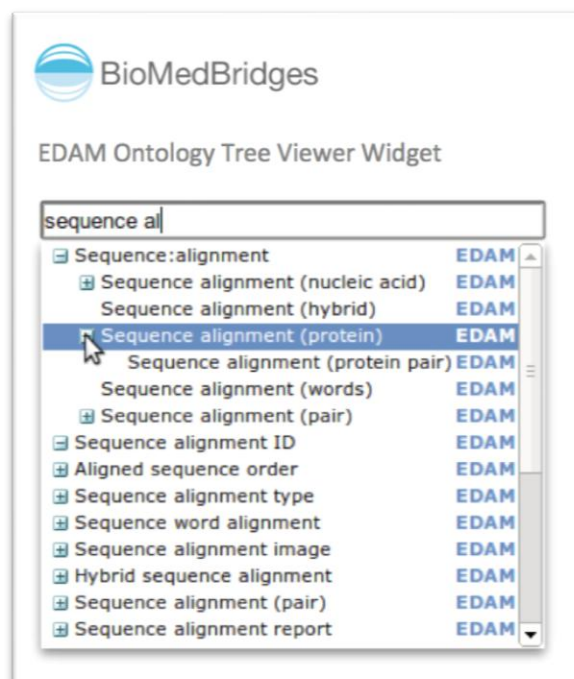


Figure 14 PLATO deployed with the EDAM ontology

³⁵ www.tinyurl.com/bmbtoolsui



Technical implementation

The core desired features of the widget were:

1. autocompletion on ontology terms
2. autocompletion on a configurable set of free-text terms
3. performant centralised query interface (via webservice)
4. visualisation of matching terms within their immediate tree context
5. ability to expand / collapse tree nodes
6. subsumption queries
7. configurability with any ontology and dataset
8. highlight of results as search term, child term or synonym

Back end: Ontology Lookup Service

The Ontology Lookup Service (OLS) provides a centralized query interface for ontology and controlled vocabulary lookup. The OLS provides 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 PLATO. The web service interface of OLS was extended to provide a new REST service was developed which provides results in a standard format (JSON). This will also allow the OLS service to be of use to a broad array of applications beyond even PLATO.

Front end: Javascript BioJS ontology viewer widget

Our requirements analysis identified seven desirable features; existing open source applications in this space were reviewed and found to offer only a subset of these features. To speed the development process, we identified one of these to modify and extend into a generic and re-usable BioJS widget. First, it was adapted to accept JSON served up by the new OLS web service described above.

Future work

This widget is undergoing testing and feedback by BioMedBridges partners and other user groups via its three intra- and extra-project deployments



described above. The code has also been submitted and is now pending review for incorporation into the open-source BioJS project. The anticipated broad use of PLATO is likely to spur more community contributions to the code, thereby making it even more extensible, robust, and feature-rich with time. We also plan to update the OLS service to support OWL format ontologies, which we expect to improve adoption of PLATO.

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

Summary of work

Building on the foundation of the REST interface developed in 4.3, 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.

Background and scientific importance

For biomedical researchers, finding related entries in different databases is often a time consuming task. In the future, this pilot will support the researcher by linking publications, genes, and drugs to clinical trials data. This aims to provide more insight into the effect of drugs or genes on patients based on real clinical trials. The nature of personalized medicine use cases is to have a data analysis bridge which enriches patient data with various other relevant data. To support the Acute Myeloid Leukaemia (AML) use case of WP8, the clinical trials data stored in the database behind the CTIM is limited to AML-related trials. This reduces the number of false positives in the results. Another advantage is that this opens the possibility to adapt the interface completely to the AML use case; this adaptation will be executed in the scope of deliverable 4.6. The researcher is not limited to the graphical user interface introduced in this deliverable, but also can also make use of the REST web service for programmatic access. Multiple output formats like csv, xml or json support this latter approach.



Table 1 *Example scientific use case*

Use Case ID	UC_Example_01
Use Case Name	Enter Query
Summary	The Scientist (User) enters the query concerning his research area.
Priority	Essential Expected Desired Optional
Use Frequency	Always Often Sometimes Rarely Once
Direct Actors	Scientist (User)
Main Success Use Case Scenario	Scientific Use Case related results displayed after submitting the query

Technical implementation

Software and Tools

For the software development process, several existing libraries were used:

Liferay Portal (<http://www.liferay.com>) is open source software that is used primarily in businesses as an employer and business process oriented enterprise portal. There are three segments in Liferay Portal: Liferay CMS is a content management system building on Liferay Portal. Liferay Collaboration is used for web-based teamwork and social networks. Liferay Portal is based on a service oriented architecture (SOA). This enables further Liferay or self-written components to be added via portlets or access to existing applications respectively.

Primefaces (<http://www.primefaces.org>) is a Java Framework based on Java Server Faces (JSF). JSF is a Java specification for building user interfaces for web applications.

Apache Lucene (<http://lucene.apache.org>) is an open source library for full text indexing and search functionalities by the Apache Open Source Foundation. It is easily scalable and its open source character allows for easy customization. The social network Twitter and the Wikipedia search function are both built on Apache Lucene.



Apache Solr (<http://lucene.apache.org>) is an open source search platform from the Apache Lucene Project. Solr and Lucene were merged in 2010 and are now produced by the same development team. 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.

Databases

ClinicalTrials.gov is a web site that provides patients, family members, health care professionals, and other members of the public easy access to information on clinical studies on a wide range of diseases and conditions (<http://www.clinicaltrials.gov>). Information is provided and updated by the sponsor or principal investigator of the clinical study and the web site is maintained by the U.S. National Library of Medicine (NLM) at the National Institutes of Health (NIH). As one of the biggest clinical trials databases in the world and also storing various clinical trials data from other comparable databases, ClinicalTrials.gov is an excellent data source for the Clinical Trials Information Mediator.

Ontologies

The Experimental Factor Ontology (EFO³⁶) provides a systematic description of many experimental variables available in EBI databases, and for external projects such as the NHGRI GWAS catalogue. It is an application ontology comprising imports of several biological ontologies, such as anatomy, disease and chemical compounds and can be used to define disease types, tissues etc. It was used in this case in order to query the clinicaltrials.gov database for all 59 known synonyms of AML³⁷.

Software developed in this deliverable

The Clinical Trials Information Mediator (short CTIM) is the perfect place for a scientist to search for clinical trials and linked publications. The software architecture of CTIM is shown in Figure 15. The graphical user interface, which is the main part of this deliverable, strongly supports usability for the scientist.

³⁶ <http://www.ebi.ac.uk/efo/>

³⁷ http://www.ebi.ac.uk/efo/EFO_0000222

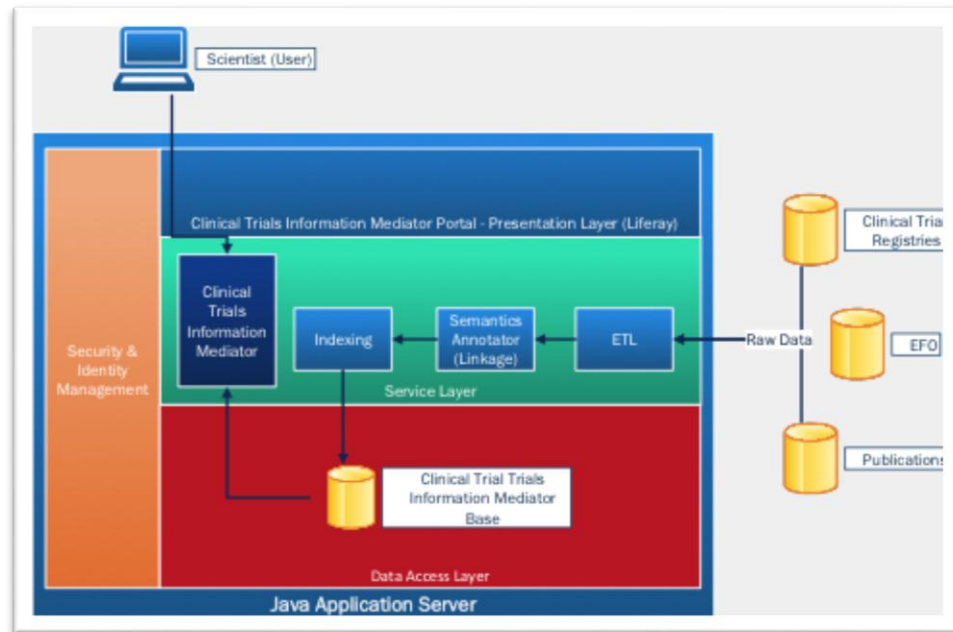


Figure 15 Architecture of the CTIM (Semantics and Publications not yet integrated)

The main aim of the interface is to grant the user easy but functional access to clinical trials associated data. The workflow of a user searching for clinical trials data is visualized in Figure 16.

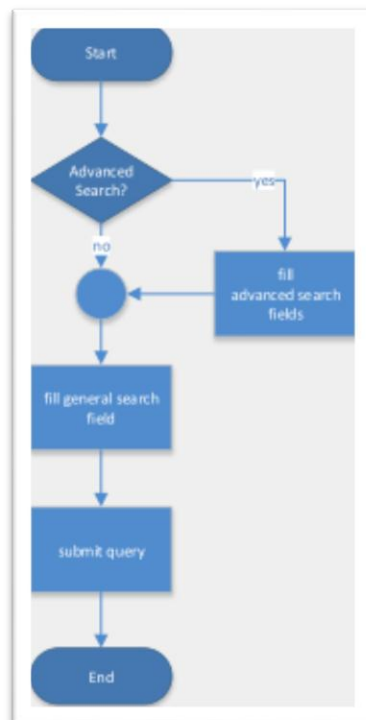


Figure 16 Query the database: workflow



Figure 17 Start screen – general search term

Figure 17 shows the first screen a user sees when he uses the CTIM interface. The simplest way to enter a query is to use the general search field. Pressing the Go-button performs a search over every data field of the internal clinical trials database. Different search terms can be entered in the general search field; wildcard searches are also supported. For example:

- Search string: haemoglobin: search for the exact word haemoglobin
- Search string: *globin: search for words, ending with globin
- Search string: *cardio*: search for words including the substring cardio
- Search string: cardiogram haemoglobin: search for both words, cardiogram and haemoglobin

Identifiers		Participant Details	
NCT ID:	<input type="text"/>	Gender:	<input type="text"/>
Secondary study ID:	<input type="text"/>	Minimum age (years):	<input type="text"/>
		Maximum age (years):	<input type="text"/>
Study Details			
Official title:	<input type="text"/>	Brief title:	<input type="text"/>
Brief summary:	<input type="text"/>	Condition:	<input type="text"/>
Primary outcome measures:	<input type="text"/>	Secondary outcome measures:	<input type="text"/>
Detailed description:	<input type="text"/>	Eligibility criteria:	<input type="text"/>
Country:	<input type="text"/>	Overall status:	<input type="text"/>
Start date (MMMM JJJJ):	<input type="text"/>	Study design:	<input type="text"/>
Study type:	<input type="text"/>		

Figure 18 Advanced search fields

The advanced search window (Figure 18) enables the user to specify his query. Only those search fields filled with text will be considered in the query.



Search					
General Search					
Search Term: <input type="text" value="cytarabine"/>					<input type="button" value="Search"/>
Results					
Search: <input type="text"/>					
NCT ID	Official title	Brief title	Brief summary	CLink to original Study	
NCT01338974	Proteomic Signature Associated With Clinical Response to Cytarabine Based Induction Therapy in Patients With AML 56 Years and Older	S9031-S9333-S0112-S0301-A Biomarkers Associated With Response to Cytarabine in Samples From Older Patients With Acute Myeloid Leukemia	RATIONALE: Studying samples of blood and tissue from patients with cancer treated with cytarabine in the laboratory may help doctors learn more about the effects of cytarabine on cells. It may also help doctors understand how well patients respond to treatment. PURPOSE: This research trial studies biomarkers associated with response to cytarabine in samples from older patients with acute myeloid leukemia.	Go to CT.gov	
NCT00549999	Phase I Study Evaluating the Chemosensitizing Effect of Everolimus Administered With Cytarabine and Daunorubicin in Patients With Acute Myeloid Leukemia in Relapse	Everolimus, Cytarabine, and Daunorubicin in Treating Patients With Relapsed Acute Myeloid Leukemia	RATIONALE: Drugs used in chemotherapy, such as cytarabine and daunorubicin, work in different ways to stop the growth of cancer cells, either by killing the cells or by stopping them from dividing. Everolimus may help cytarabine and daunorubicin work better by making cancer cells more sensitive to chemotherapy. Giving everolimus together with cytarabine and daunorubicin may kill more cancer cells. PURPOSE: This phase I trial is studying the side effects and best dose of everolimus when given together with cytarabine and daunorubicin in treating patients with relapsed acute myeloid leukemia.	Go to CT.gov	
NCT01635296	A Multicenter, Open Label, Phase 1B Study of Escalating Doses of RO5045337 Administered Orally, With Cytarabine Administered A) Subcutaneously, or B) Intravenously, in Patients With Acute Myelogenous Leukemia (AML)	A Study of RO5045337 in Combination With Cytarabine in Patients With Acute Myelogenous Leukemia	This multi-center, open-label, Phase 1b study will evaluate the safety, pharmacokinetics and efficacy of RO5045337 in combination with cytarabine in patients with acute myelogenous leukemia. In Arm A, cohorts of previously untreated patients deemed unsuitable for standard induction therapy will receive escalating oral doses of RO5045377 and cytarabine 20 mg/m ² subcutaneously daily for Days 1 to 10 of each 28-day cycle. In Arm B, cohorts of patients who have relapsed or are refractory after at least one cytarabine/anthracycline containing regimen will receive escalating oral doses of RO5045377 on Days 1 to 5 and cytarabine 1 gm/m ² intravenously on Days 1 to 6 of each 28-day cycle. Patients will receive up to 4 cycles of therapy, patients in Arm A who achieve hematologic response may continue additional cycles until disease progression.	Go to CT.gov	

Figure 19 Results screen (without details)

Study details	
Identifiers	Participant Details
Secondary study ID:	Gender: Both
	Minimum age: 56 Years
	Maximum age: N/A
Study Details	Eligibility criteria:
<p>Detailed description: OBJECTIVES: - Refinement and testing of a multiparameter flow cytometry-based cell-signaling signature (FC classifier) associated with in vivo likelihood of complete response (CR) to cytarabine-based induction chemotherapy in elderly patients (56 years and older) newly diagnosed with non-M3 acute myeloid leukemia (AML). - Identification of cell-signaling signature(s) associated with continuous CR to cytarabine-based induction chemotherapy at one year (CCR1) in adult patients 56 years and older with a newly diagnosed non-M3 AML. - Identification of cell-signaling signature(s) associated with relapse-free survival at one year (RFS1) in adult patients 56 years and older with a newly diagnosed non-M3 AML who received cytarabine-based induction chemotherapy and achieved CR. - To investigate changes in cell-signaling signature(s) between matched pre- and post-treatment specimens of relapsed/refractory patients. OUTLINE: This is a multicenter study. Cryopreserved specimens are incubated with cytokines (e.g., interleukins), growth factors (e.g., saquinavir or flgastin), and chemotherapeutic agent (e.g., cytarabine, etoposide) and other modulators. Cells are then fixed, permeabilized, and stained with antibodies that recognize extracellular markers (for example, surface phenotypic markers such as clusters of differentiation, drug transporters, and receptors) in conjunction with intracellular activation-state-specific epitopes of designated signaling molecules. Subsequently, cell are analyzed by phospho-flow cytometry (FC) in a random manner (without knowledge of clinical variables and outcomes) to a training set (complete pre-specified FC) versus a testing set. Cells are also analyzed by proteomic assays. Results are then compared with individual patient scores, including predicted clinical outcomes.</p>	<p>DISEASE CHARACTERISTICS: - Newly diagnosed with non-M3 acute myeloid leukemia (AML) - Pretreatment and relapsed/refractory cryopreserved marrow and circulating mononuclear cells (MC) from patients who meet the following criteria: - Eligible and evaluable patients on study SWOG-9031, SWOG-9333 (Ara-C/IDR induction arm only), SWOG-9012, or SWOG-S0301 - Did not refuse consent for this use of specimens - Have 2+ vials of pretreatment marrow cells and/or 2+ vials of pretreatment peripheral blood cells in the Southwest Oncology Group (SWOG) AML/MDS Repository PATIENT CHARACTERISTICS: - See Disease Characteristics PRIOR CONCURRENT THERAPY: - See Disease Characteristics</p>
Primary outcome measures: Response to induction chemotherapy: complete response (CR) vs non-complete response (NR)	Secondary outcome measures:
Start date: March 2011	Study type: Observational
Study design: Time Perspective: Retrospective	Condition: Leukemia
Location Countries:	Intervention type:
Intervention name: Other	Recruitment status: Completed

Figure 20 Results screen (detailed view)

The results window (Figure 19) opens when the user commits a query. Another search field on the top of the frame allows the user to filter the results. The small arrow on the left of every results row opens detailed information (Figure 20).

Future work



This work will be further developed in 4.6, in which we plan to perform text mining and semantic alignment in order to enable the user to search trials according to the compounds and genes they reference. Additional text mining will be done to enable showing the trials with their published results. Those features will open the bridge to other biomedical infrastructures involved in the BioMedBridges project.

Future work

Each of the pilot REST vignettes developed here will be refined according to ongoing user feedback and testing. Some of the work done here as part of deliverable 4.5 will also provide a basis for work in the Semantic Web Pilot project (D4.6), which will lead to greater semantic alignment of the data, thereby facilitating tighter integration across databases. 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.

Additionally, further BioJS widgets of relevance to BioMedBridges are planned for development in 2014. These include widgets for the BioSamples Database, the Zooma curation tool, and the BioMedBridges Software and Data Services Registry. These may be used by the consortium, and a BioMedBridges knowledge exchange event is planned for early 2014 which will provide training on how to use, design and deploy these tools (WP12).

4 Delivery and schedule

The delivery is delayed: Yes No

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 towards this deliverable within WP4 (UDUS, ErasmusMC). The shift was part of an amendment to the Grant Agreement.



Erasmus MC proposes to further shift resources from D4.5 to D4.6, since for D4.5 they could build on existing XNAT software, whereas the developments for D4.6 are expected to be more laborious than previously anticipated.

6 Background information

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

WP 4 Title: Technical Integration

Lead: Ewan Birney (EMBL)

Participants: EMBL

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

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

Work number	package	WP 4	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
<p>Objectives</p> <ol style="list-style-type: none"> 1. Implement shared standards from WP 3 to allow for integration across the BioMedBridges project 2. Expose the integration via use of REST based Web services interfaces optimised for browsing information 3. Expose the integration via use of REST based Web services interfaces optimised for programmatic access 4. Expose appropriate meta-data information via use of Semantic Web Technologies 5. Pilot the use of semantic web technologies in high-data scale biological environments. 								
<p>Description of work and role of participants</p> <p>We will provide a layered, distributed integration of BioMedBridges data using latest technologies. A key aspect to this integration will be the internal use of standards, developed in WP 3 which will provide the points of integration between the different data sources. The use of common sample ontologies (WP 3) will provide integration between biological sample properties, such as cell types, tissues and disease status, in particular bridging the Euro-Biolmaging, BBMRI, Elixir and Infrafrontier projects. The use of Phenotype based ontologies will provide individual and animal level characterisation which, when these can be associated with genetic variation, will provide common genotype to phenotypic links, and this will be used to bridge the ECRIN, EATRIS, INSTRUCT, BBMRI, Infrafrontier and Elixir Projects. The use of environmental sample descriptions and geolocation tags will bridge between EMBRC, ECRIN, ERINHA, EATRIS and Elixir. The use of chemical ontologies will help bridge between EU-OPENSREEN, ECRIN, Euro-Biolmaging, INSTRUCT and Elixir. By applying these standards in the member databases (themselves often internally federated) we will create a data landscape that theoretically can be traversed, data-mined and exploited. To expose this data landscape for easy use, we will deploy a variety of different distributed integration technologies; these technologies are organised in a hierarchy where the lowest levels are the semantically poorest, but easiest to implement, whereas the highest levels potentially expose all information in databases which are both permitted for integration (some are restricted for ethical reasons, see WP 5) and can be described using common standards. We will develop software with aspects appropriate for the distributed nature of this project taken from agile engineering practices, such as rapid iterations between use cases and partial implementation. In particular we will be using the enablement/alignment strategy (Krcmar H., Informationsmanagement, Springer) to ensure that the use cases that drive the project are aligned to feasible capabilities that can be delivered. The work package will be implemented in a collaborative manner across the BMSs, with frequent physical movement of individuals.</p> <p>The proposed technologies are:</p> <ol style="list-style-type: none"> 1. REST-based “vignette” integration, allowing presentation of information from specific databases in a human readable form. An example is shown in Figure 1. These resources allow other web sites to “embed” 								



- live data links with key information into other websites. This infrastructure would then be used to provide browsers that, on demand, bridge between the different BioMedBridges groups – for example, information which can be organised around a gene or a chemical compound would be presented across the BioMedBridges project.
2. Web service based “query” integration, where simple object queries across distributed information resources can be used to explore a set of linked objects using the dictionaries and ontologies present. Each request will return a structured XML document.
 3. Scalable semantic web based technology. We are confident that semantic based technology can work for the rich but low data volume meta data (eg, sample information) which we will expose using semantic web technologies such as RDF and SPARQL. However, it is unclear whether this scales to the very large number of data items or numerical terms in the BioMedBridges databases (such as SNP sets or numerical results from Clinical trials) We will pilot a number of semantic web based integration of datasets, using RDF based structuring of datasets In the latter phases of the project we will look to align these solutions to other broader standards in the eScience community, taking input from the Technology Watch (WP11) group; we hope in many cases our technology choice which has been already informed by alignment to future eScience technology (e.g. RDF/SPARQL) so this may only require appropriate registration/publication of our resources. Where unforeseen but useful technologies are developed we will build systematic connections from these BioMedBridges federation technologies to other federation technologies.

Deliverables

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