Building data bridges between biological and medical infrastructures in Europe

**Reporting**

<table>
<thead>
<tr>
<th>Project Information</th>
<th>Funded under</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>BIOMEDBRIDGES</strong></td>
<td>Specific Programme &quot;Capacities&quot;: Research infrastructures</td>
</tr>
<tr>
<td>Grant agreement ID: 284209</td>
<td></td>
</tr>
</tbody>
</table>

**Closed project**

<table>
<thead>
<tr>
<th>Start date</th>
<th>End date</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 January 2012</td>
<td>31 December 2015</td>
</tr>
</tbody>
</table>

**Total cost**

€ 13 729 287,62

**EU contribution**

€ 10 494 998,69

**Coordinated by**

EUROPEAN MOLECULAR BIOLOGY LABORATORY

Germany

This project is featured in...

RESEARCH*EU MAGAZINE
Executive Summary:

Building bridges

The gap between curiosity-driven, basic research and application-driven research is a long-standing challenge in the natural sciences. Basic research generates valuable new ideas that help scientists approach problems in new ways. In the life sciences, significant additional value and insights are gained through connecting knowledge from different disciplines. The main aim of BioMedBridges is to facilitate the translation of ideas into new medical and environmental applications by removing technical stumbling blocks related to interoperability of data from a variety of disciplines and scales.

Researchers from different scientific communities often describe things in very different ways. This can result in information on the same thing - as simple as high blood sugar levels in human patients or mouse disease models - not being fully compatible, so that different sources of information cannot easily be connected. To address this, BioMedBridges builds a shared data culture in the life sciences by linking up 12 of Europe’s new biological, biomedical and environmental research infrastructures (BMS RI).

Integrating the vast data resources in the life sciences, including data from genomics, biological and medical imaging, structural biology, mouse disease models, clinical trials, highly contagious agents and chemical biology, will enable new ways of analysing them to answer new, more complex scientific questions.

The central principle driving BioMedBridges is the development of necessary data infrastructure, including shared standards and semantic web technologies. While the work within BioMedBridges is done in the context of a specific set of “use cases” to demonstrate the power of such data integration, the project outcomes and the contribution it makes to data interoperability between the research infrastructures involved ultimately lays the foundation for the reuse, combination and analysis of data resources in many different contexts going forward, including future contexts that are not yet evident and that will only emerge with new challenges and directions of scientific inquiry.

By building these computational bridges, BioMedBridges progresses considerably beyond the state-of-the-art. It bridges data:
- across different spatial scales: from molecules through cells and organs to humans and the environment
- between different species: from bacteria through model organisms to humans
- between different technologies and the heterogeneous data they generate: from the nanotechnology of sequencing through the spectroscopy of cellular and whole organism imaging to the powerful synchrotrons for structure determination
- and across different research communities: from basic molecular biologists to clinicians and environmental researchers, who have not traditionally worked closely together.
The work of the project will lead to real and sustained improvement in the services offered by the BMS RIs to the research community. Data curation and sample description will be improved in all of them by the adoption of best practices and agreed standards. These efforts in turn will benefit society and boost the bioeconomy by speeding up the translation of knowledge from basic research to new drugs, treatments and products.

Project Context and Objectives:
Building blocks in the integration of life science data resources

The tools and resources developed in BioMedBridges are contributing greatly to the integration of the services provided by the participating BMS RI and lay the groundwork for future activities.

Identifying disease biomarkers by comparing images from different scales

Biomarkers indicate a specific state, such as a disease, in a given organism. They are measurable indicators such as certain molecules in the bloodstream or specific, visible changes in a tissue sample. Identifying new, useful biomarkers is a big and potentially costly effort as tremendous amounts of data need to be analysed and compared.

BioMedBridges partners have built the infrastructure to support the identification of new biomarkers for diseases such as cancer from imaging data. This is done by comparing the changes caused by certain genes in individual cells with tissue samples taken from mouse model organisms and human patients who have cancer.

The first challenge is to match up (map) the different terms used to describe cells, human and mouse tumour tissue images so they can be directly compared. This has been achieved and implemented with a tool called CMPO, which supports the annotation of images from the different domains (cell-tissue-organism). Following this, the process was tested by annotating images from gene knock-out mice, human cancer samples and RNAi screens to try and recover preselected candidate biomarkers. Analysis of the resulting annotation data uncovered different annotation practices across RNAi screens, mouse and human samples that hamper comparisons across these imaging domains. These observations suggest a possible reorganization of CMPO and indicate that, in addition to annotation standards, annotation practices must also be standardised.

Bridging between humans and model organisms to boost research on diabetes and obesity

Diabetes and obesity are complex diseases that represent a major international public health threat. To date, the genetic factors underlying these complex conditions have not been identified. Understanding these factors better would greatly help the diagnosis, treatment, and prevention of diabetes and obesity.

In recent years, a large number of new genetically modified animal models including transgenic, generalized and/or tissue-specific knockout mice have been engineered for the study of diabetes and obesity. However, researchers working on mouse disease models and those working with human patients use different terms to describe the same type of data, resulting in a major challenge in translating
information from one to the other or even identifying a suitable mouse model to answer specific research questions. Working with experts from the different communities involved, BioMedBridges has employed large-scale datasets provided by INFRAFRONTIER, BBMRI and ELIXIR to develop infrastructure that enables “crossing the species bridge” between mouse models and human to enable research into the genetic factors underlying diabetes and obesity.

The resulting infrastructure consists of a number of components. First, a comprehensive ontology - an online representation which shows the relationship between different concepts - was developed to describe Type 2 diabetes and obesity phenotypes in mouse and human. Using this ontology, mouse and human datasets from a number of sources were annotated with specific terms used to describe Type 2 diabetes progression, enabling “translation” between mouse model and human data. Second, an online tool was developed to assists researchers in the identification of suitable mouse models based on diabetes and obesity-related phenotypes. Finally, a set of scripts integrates data from a number of sources to support the validation and prioritization of genes that could contribute to diabetes and obesity.

Integrating data sources to provide targeted patient treatment

Personalised medicine is an emerging practice in health care that tailors a patient’s treatment based on the results of specific diagnostic testing, including genetic profiling. As part of systematic clinical studies as well as in compassionate settings where all other treatment options have been exhausted, physicians treating cancer patients may use a personalised medicine approach involving the analysis of the tumour genome in order to find drivers of the disease and mutations that can be targeted with specific drugs. As physicians themselves normally have neither the time nor the expertise to analyse vast amounts of genomic data, BioMedBridges has built a repository for this sensitive information that provides summary reports and allows them to drill into the underlying detail as and when needed.

Relevant external data sources that are being tapped enable matching of patients whose tumours show a similar genetic make-up, identification of drugs that have originally been approved for a different anatomically-defined cancer that nevertheless shows the same mutations and so could be used in treatment, or prediction of the response to a particular drug treatment regime.

This infrastructure was tested in a use case involving haematologists at the Helsinki University Central Hospital Oncology Clinic treating relapsed acute myeloid leukaemia (AML) patients.

Organising, linking and discovering biomolecule structural data

Structural biology is producing unprecedented amounts of data that increase not only in number, but also in size and complexity and that span an ever-wider range of resolutions, including low-resolution data. While there are very good bioinformatics tools for the analysis, validation and comparison of detailed atomic structures, at present there are very few tools available to work with low-resolution data (i.e. volume or shape data). However, low-resolution data can have very high value and is gaining in importance. In Cryo-electron microscopy (cryoEM) for example, specimens are viewed in their physiological environment, preserving the natural conformations of molecules. This is necessary for example to check what structural changes might be induced by a drug binding to a target.
BioMedBridges partners have built the tools to support analysis of low-resolution data by developing a web-based service for searching the PDBe and EMDB public archives based on shape (volume data). By comparing the shape of a given biomolecule with the contents of the public archives, the service enables the identification of known structures that have a similar shape. This is further enhanced by the integration of software that enables the determination of the shape of various protein conformations sampled in solution in which they experience some kind of mobility, which presents a challenge to shape determination.

Another web-based service enables the decomposition of volume data into its components (e.g. various proteins, possibly RNA or DNA etc.), which supports the interpretation of complex volume data in terms of possible and plausible structures of components of that data (e.g. when annotating particles in a tomogram).

Finally, BioMedBridges has provided a new RESTful interface for PiMS, a laboratory information management system for use in protein production laboratories. This interface enables the analysis of PiMS data in light of other datasets, such as protein sequence and structural information, and enables a future vision of tools which track the provenance of published biomedical results to the datasets and samples involved in the research project.

Linking diseases with biosamples and genes

BioMedBridges partners have mapped existing healthcare sector (disease) terminologies together with their phenotypic descriptions used to characterize patient conditions (symptoms)—and biobank samples (BBMRI) to molecular-level data (ELIXIR) and developed a query mechanism to identify samples with specific characteristics held in BBMRI biobanks.

Facilitating interoperability between these different types of data enables analyses to help find the underlying causes of disease and, for example, the molecular links between diseases and their comorbidities. Put together with information from genotyping of patients, the work supports the identification of groups at risk for certain diseases and other risk factors for patients with existing diseases, as well as supporting research into different diseases and the identification of disease biomarkers.

Supporting data sharing

Standards and data harmonization: prerequisites to data integration

The provision and use of common and unambiguous identifiers for bio-molecules such as genes, proteins and bioactive compounds is key to supporting the information flow from basic science, model organism biology, bioinformatics and structural biology through to translational research and clinical care. This issue, which lies at the core of the BioMedBridges mission, has been addressed through a wide variety of initiatives and resources, in particular the harmonisation of existing standards and identifiers and the identification of new identifiers where needed, the provision of a registry for users to be able to find suitable standards, and a registry of tools and data resources. In addition, the project has co-organised a series of workshops establishing longer-lasting important collaborations and contributing to the
harmonisation of the life science data interoperability landscape.

The accumulated knowledge and expertise across all BioMedBridges partners and domains, as well as external experts, has been made available to the wider community in key publications.

Nuts and bolts: the building blocks of data integration

Biomedical research requires increasingly sophisticated infrastructure. While early studies were completed without any formal infrastructure (single use) or only using basic, local infrastructure, research that is done now and in the future requires robust infrastructure that is large-scale, non-trivial, and interconnected across disciplines, data types and resources, and countries. BioMedBridges is building these specific information infrastructure bridges.

An information infrastructure has “layers”, just like biology: DNA is translated to RNA, which is translated to protein. Similarly, machine-friendly representations of data must be translated step by step to human-friendly ones that are accessible also to the end user without detailed bioinformatics knowledge.

Data interoperability in BioMedBridges services is achieved via a technology stack that initially involves using established REST-based technology and ultimately aims to achieve more sophisticated semantic interoperability. Using this stepwise, layered approach has contributed to systematically bringing all research infrastructures and data resources to a higher level of integration while creating the necessary expertise to further advance data interoperability in future efforts.

The construction of this information infrastructure ultimately supports the development of Web Services-based simple object queries, which enable researchers to find in one simple step information most relevant to a scientific question or related to a specific disease in a very large number of data resources with millions of data entries.

BioMedBridges partners have connected resources from the different BMS RIs in a series of pilot projects. For example, the BioMedBridges contribution to UniChem supports the integration of high-throughput screening data on small molecules with chemical libraries and chemical resources for drug discovery and optimization. A new search function enables researchers to find not only “same” compounds across different resources, but also “similar” compounds, allowing discovery of a much wider range of results from a larger number of resources, ultimately making a significant contribution towards drug discovery also by users from industry.

Another tool enables the sharing, integration and centralised analysis of medical imaging data. With medical imaging techniques such as magnetic resonance imaging (MRI), computed tomography (CT) and ultrasound being used intensively in large, distributed studies, there is an increasing demand for a platform to exchange such medical imaging data and the derived results. BioMedBridges has customised the open source platform XNAT so it enables sharing and analysis of imaging and image-derived data and supports analysis between image-derived data and other types of data such as clinical data (e.g. disease status, age) and genetic data.
As a third and final example of the services produced within the project, the MIABIS Connect service enables sample discovery across a large number of biobanks. It is an Open-source software framework that can be easily adopted by the biobank and research communities. By sharing information via MIABIS the software allows biobanks to retain their own, individual sample descriptions, saving tremendous efforts that would otherwise be required for annotating samples with different terms. Supporting easy adoption, MIABIS Connect allows biobank data to stay at the biobank - only the results of queries are fetched from the biobanks to be viewed through the common query interface. Finally, the tool requires minimum IT support at the biobanks while enabling biobanks to visualise summary information on the samples they hold.

Sharing sensitive data securely and ethically

Which legal and ethical requirements do researchers have to consider when using data from external sources? What about data protection, privacy, etc.? For an individual researcher it is almost impossible to be informed and knowledgeable about all of the issues involved, especially when this is not their daily business. At the same time, consulting experts in order to get this information can be a time consuming process. To support researchers in identifying legal and ethical requirements connected to sensitive data BioMedBridges partners have developed an online tool, which is based on a thorough analysis of the regulatory landscape in Europe conducted in the first project year.

Informed by the identified formal requirements for secure and ethical data sharing and based on a number of use cases within the project, including personalised medicine data, BioMedBridges is putting appropriate safeguarding (data protection and data security) mechanisms in place.

Project Results:
The below provides a summary of the work package activities and results. Detailed reports are included in the attachments.

1. Construction work packages (WP3, 4, 5)

WP3 Standards (M1-M48)
The detailed report from work package 3 is available in Annex 2 to the periodic report.

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

WP4 Technical Integration (M1-M48)
The detailed report from work package 4 is available in Annex 2 to the periodic report.

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

During the last reporting period, in May 2015, at an industry workshop, tutorials were delivered demonstrating the queries and analyses the RDF platform makes possible. During a joint SWAT4LS / BioMedBridges hackathon in November 2015, participants applied semantic web technologies supported by BioMedBridges, such as the development of RDF converters for some existing biological data, development of SPARQL queries over existing data that targets specific use cases, or the developing of applications that are built on semantics web technologies.

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

The final WP4 deliverable D4.8 reports on the strategy, implementation and lessons learned for the semantic web pilots for BioMedBridges.

WP5 Secure Access (M1-48)

The detailed report from work package 5 is available in Annex 2 to the periodic report.

During the last reporting period, the work of WP5 focussed mainly on the tasks of WT8, the implementation of a pilot application for the security framework. This pilot was based on the preparatory work of all previous WTs, but especially on WT7 (Design of a security architecture and framework). The implementation was done in collaboration with WP4 and WP3 and with the different WT members of WP5.

The pilot implementation connects different software applications to allow users controlled access to restricted biomedical data (controlled access). It is realised in a modular way to make it easy to integrate external or additional components. The pilot demonstrates the feasibility of the security architecture developed in D5.3; it expands an existing data bridge between the research infrastructures BBMRI and ELIXIR including the integration of public biobank metadata with data about biosamples. In this context, the security requirements for an e-infrastructure (WT5), the threat and risk analysis for sharing data or biomaterials (WT6) and the design of the security architecture and framework (WT7) formed a preparatory basis for the pilot. The LAT (Legal Assessment Tool) developed in WT4 became part of the implementation and was further refined for it. The work of WT8 resulted in Deliverable 5.4: Implementation of a pilot for the security framework.

2. Use case work packages (WP6, 7, 8, 9, 10)

WP6 Interoperability of large scale image data sets from different biological scales (M13-M48)

The detailed report from work package 6 is available in Annex 2 to the periodic report.

WP6 addressed interoperability of large-scale image data sets, which is a prerequisite for reusing and analysing data sets generated at different scales and in different biomedical models. The aim was to facilitate interoperability of different image data sets (cellular, mouse and human tissue) and to create the
tools that facilitate the comparison of cellular phenotypes observed in different samples: cell lines showing phenotypes associated with individual gene knockdowns and imaging data from diseased tissue specimens (both human and mouse tissues).

With the decreasing cost of sequencing technologies, genome sequences of many tumors are becoming available. While mutations can be identified in these genomes, determining their functional consequences and relevance to the disease remains a challenge. Linking phenotypic data specific to individual genes to morphological imaging data from diseased tissue specimens (both human and mouse tissues) could be used to infer functional consequences of somatic tumor mutations and therefore help in adding functional annotations to cancer genome data. For example, when a certain cellular phenotype, like ‘mitotic delay’ or ‘multinucleated cells’, observed in cells after gene knockdown experiments, is also observed in cells of a cancer tissue, we may infer that the knocked down gene is involved in the aetiology of the disease, in that specific tissue.

This information could also be used to guide clinical intervention, for example to design more targeted drug therapies or to identify new diagnostic or prognostic markers. In the first half of the BMB project, WP6 had identified aspects of image data sets that significantly influence their interoperability: i) Standards used in different file formats; ii) accessibility of the data; and iii) image annotation and use of ontologies. In deliverables 6.1 6.2 and 6.3 WP6 had created the lists of the most frequently used image file formats for cellular, mouse and human image data sets; and ontologies for image annotation. Based on this, WP6 developed the Cellular Microscopy Phenotype Ontology (CMPO). In the scope of deliverable 6.4 WP6 further worked on improvement of CMPO and its integration into workflows used by scientists. This new software tool CMPO is now publicly available (http://www.ebi.ac.uk/cmpo/) a species neutral ontology for describing general phenotypic observations relating to the whole cell, cellular components, cellular processes and cell populations. It allows scientists to harmonize the annotation of cellular phenotypes, making them interoperable.

Testing of CMPO for the annotation of imaging datasets derived from different biological domains (cell lines, mouse and human tissues) had demonstrated that CMPO is suitable to annotate cellular phenotypes observed in such images, consequently making the data interoperable.

In the framework of WP6 Task 2, the BMB partners selected images for genes reported to be mutated in cancer and with cellular phenotypes in the RNAi screens and for which images of orthologous genes knock-out mice were available. They accessed biomedical data at The Cancer Genome Atlas (TCGA, http://cancergenome.nih.gov/) searching this archive for human cancer samples whose genetic information indicated a likely loss-of-function mutation in one of the selected genes making these genes good biomarker candidates. The detailed procedure is described in the Annex. In summary, for this particular example they uncovered different annotation practices in different domains:

- Each domain relies on a different branch of the ontology
- The human cancer images are annotated with 2-3 times more terms than the mouse and RNAi screen images
- The mouse images are annotated with terms that are more generic (lower information content) than those used for the human cancer and RNAi screen images
For the future, these observations suggest two avenues to make cross-domain annotations useful: One is to standardize the annotation practices beyond standardizing the annotations, for example, using and comparing only annotations for phenotypes observable across all domains. Typically, this would mean using annotations from the cellular component domain as cell processes usually are only accessible in RNAi screens.

The second avenue concerns modification of the ontology, in particular the cell population branch. Although this branch describes more quantitative aspects of the cellular component branch, the two branches don't seem to be semantically connected. Although this could be due to the current sparseness of the ontology, this points to a possible reorganization of the ontology.

WP7 Use case: PhenoBridge - crossing the species bridge between mouse and human (M13-48)
The detailed report from work package 7 is available in Annex 2 to the periodic report.

During the last reporting period, the PhenoBridge use case has completed the development of a comprehensive ontology to describe type 2 diabetes and obesity phenotypes in mouse and human. The work was achieved in three steps: (1) identification of the disease-phenotype relationship using text mining, (2) expert review of the resulting terms, and (3) development of the DIAB Ontology Model using the Web Ontology Language (OWL). The resulting DIAB ontology was published in BioPortal.

Using the DIAB ontology, mouse and human datasets from a number of sources were annotated specific terminology representing Type 2 Diabetes progression. Co-annotation was completed for sample datasets from PhenoBridge partners and the MetaboLights and BioSamples public data resources.

Last but not least, significant progress has been made on the development of two tools, the M3 (Mining Mouse Models) tool and the GenoBridge tool. M3 assists researchers in the identification of suitable mouse models based on diabetes and obesity-related phenotypes, while GenoBridge integrates data from a number of sources to enable systematic mapping between syntenic regions in the mouse and human genome, enabling the discovery of functional conservation to validate/prioritize candidate disease genes or regions.

WP8 Use case: Personalized Medicine - integrating complex data sets to understand disease pathogenesis and improve biomarker and treatment selection (M1-M48)
The detailed report from work package 8 is available in Annex 2 to the periodic report.

The final deliverable of the work package consists of a report on and demonstration of interoperability between different types of personalised medicine (PM) data. A prototype IT system was developed for storing, structuring and making available of controlled access data and annotation/filtering of data used for PM. The installation at UH/FIMM supports the decision making of hematologists at the University of Helsinki Central Hospital in treating leukemia patients while the installation at UMCG supports genome-first differential diagnostics of new-born babies with acute metabolic disease.

WP9 Use case: From cells to molecules - integrating structural data (M13-M48)
The detailed report from work package 9 is available in Annex 2 to the periodic report.
A new web service PDBeShape has been released which gives the general biologist access to a new and growing class of experimental data, namely structural volume data (Deliverable 9.2). The service provides access to a database containing high quality volume data for prokaryotic and eukaryotic ribosomes, and class I and II chaperonins, taken from the Electron Microscopy and Protein Data Banks (EMDB and PDB). Entries are annotated with links to external databases such as Pfam/Rfam and NCBI taxonomy. The volume database contains a set of pre-calculated volume/shape alignments determined by a software pipeline SMaSB. The user can also upload their own volumes, and align against the contents of the volume database.

3D segmentation of molecular volume data is used to identify individual components in a larger complex (Deliverable 9.3). Identification of individual macromolecular components facilitates more precise annotation with molecular identifiers. For segmentation of single particle volume data, we use Chimera-Segger in the volume pre-processing pipeline of SMaSB. The pre-processing pipeline is applied in a completely automated fashion to all volumes in the database, and the segmentation step results in a .seg file containing details of the segments found. In addition, we have manually segmented a selection of entries from the PDBeShape volume database, giving an alternative interpretation of each volume. The manual segmentation is recorded in a Chimera-Segger .seg file. A separate Excel spreadsheet holds further annotation of each manual segmentation.

We have developed the MAXOCC software for quantifying the range of accessible protein conformations, and extended it to evaluate MaxOR/MinOR for regions of conformational space. Based on a case study of the protein calmodulin, a protocol for including MaxOcc results has been decided (Deliverable 9.4). The PDBeShape volume database contains a link to a file of MaxOcc results. This represents structural annotation of volumes, which supplements the sequence-based annotations considered in other tasks. The main results of the 3rd reporting period are:

1. Maintenance and minor updates of the SMaSB pipeline for volume matching
2. Extension of the volume database to include class I and II chaperonins
3. Public release of the PDBeShape web service
4. Incorporation of automated segmentation into the SMaSB pipeline, and inclusion of selected manual segmentations into the volume database
5. Extension of the MaxOcc software for quantifying MaxOR and MinOR for a region of conformational space
6. Development of a protocol for annotating volumes with putative conformations.

WP10 Use case: Integrating disease related data and terminology from samples of different types (M13-M48)

The detailed report from work package 6 is available in Annex 2 to the periodic report.

Work on WP10 proceeded along two main tracks that demonstrated bridging from BBMRI and EATRIS to ELIXIR: mapping data in BBMRI biobanks to the ELIXIR infrastructure (Tasks 1 and 3), and linking disease nomenclature to gene databases (Task 2).

For the BBMRI-ELIXIR bridge we concentrated on mapping the common set of biobank descriptors, “MIABIS”, into the SampleTab format used for populating the BioSamples database, an ELIXIR resource. This work resulted in deliverable D10.1 “Mapping between data elements” that was described in the first
periodic report and delivered in January 2014. The developed mapping enabled the work towards a prototype federated query interface. There are two activities that are relevant for this prototype. First, WP10 worked together with WP5 on the emerging pilot for the security framework (D5.4). This pilot demonstrates an easy to use bridge between components of BBMRI and ELIXIR that will enable users to discover data in the public BioSamples database, look up more detail in the BBMRI Hub catalogue, request individual level data access in the Resource Entitlement Management System, and access granular data, all within the same space of federated user authentication and authorization.

The second line of work leading to the prototype was defined in 2014, and the work was performed in 2015. A common query interface for mapping and accessing biobank data using the MIABIS standard was specified, implemented and deployed for 3 biobanks.

The work on linking health terminologies, in particular, the widely used ICD10, to associated genes, resulted in D10.2 “A prototype linking ICD10/SNOMED CT concepts to Ensembl gene identifiers” that was delivered in August 2014. In 2015 we developed a new approach creating associations to microRNAs. Since their discovery in 1994, miRNAs as a target are gaining attraction in the research community, and, to our knowledge, no-one has worked on associating miRNAs to ICD10 diagnostic codes before. Both the prototype biobank federation system and the novel work on linking disease terminology to miRNAs is described in more detail in D10.3 “A prototype federated query interface for information on biosamples, and linking of biosamples and disease terminology to genome”.

3. Supporting work packages (WP2, 11, 12)

WP2 Inreach/Outreach (M1-M48)

The detailed report from work package 2 is available in Annex 2 to the periodic report.

In this reporting period, the main aim was to communicate the BioMedBridges outcome to the internal and external stakeholders. To this aim the BioMedBridges Newsletter has been published each quarter. In each newsletter we have, next to the project highlights, some fixed items i.e. “tool in the spotlight” and upcoming events. To gain insight in the outreach of the project, we list the outreach activities of the project members on a regular basis.

The BioMedBridges website has been made more attractive and more open to the general public. To this aim a user persona analysis has been made. Furthermore, the tools developed by BioMedBridges have been described in a user friendly language using less tech speak.

To get insight in the communication channels between the BMS RIs we have listed the communication tools of the infrastructures.

WP11 Technology watch/e-infrastructure advisory board (M1-M48)

The third periodic report of the e-infrastructures advisory board is project deliverable 11.3 Final report by the e-infrastructure advisory board.

WP12 Training

The detailed report from work package 12 is available in Annex 2 to the periodic report.
During the last period WP12 has undertaken the following work:

- Planning for a knowledge exchange workshop that contribute towards deliverable 12.2 “Documentation from workshops 3 & 4”.
- Assisting in the writing of a follow-on report on activities at this event.
- Data strategies for research infrastructures
- Development of e-learning training modules as described in deliverable 12.3 “User training plan”.
- Delivery of webinars with project partners to disseminate information on software and tools.
- Presentation of a poster on “Tips for good end-user engagement” at the “Open bridges for life science data” symposium in Hinxton, UK (November 2015).
- Assisting in the delivery of the “Open bridges for life science data” symposium in Hinxton, UK (November 2015).

The main results achieved during this period are:

- Delivery of the final workshop for deliverable 12.2 “Documentation from workshops 3 & 4”:
  -- Data strategies for research infrastructures (February 2015)
- Completion of deliverable 12.2 “Documentation from workshops 3 & 4”.
- Delivery of e-learning modules described in deliverable 12.3 “User training plan”:
  -- UniChem: quick tour
  -- Cellular Microscopy Phenotype Ontology: quick tour
  -- Biomedical data: ethical, legal and societal implications
  -- User experience design
  -- Structural volume data
- Coordination of webinars from project partners:
  -- UniChem: EMBL-EBI’s mapping tool for small molecule database identifiers: webinar
  -- BioSamples RDF Platform: webinar
  -- Licensing Web Services: webinar
- Completion of deliverable 12.4 “Report of user training performed in months 36-48”.

Potential Impact:
In a joint statement under the BioMedBridges umbrella, the biomedical sciences research infrastructures have highlighted key points concerning life science data management and sharing. Confirming the importance of making data available and accessible so it can be widely reused and repurposed to answer new and different scientific questions, the RIs point out that some data may only be shared under certain conditions and with appropriate safekeeping mechanisms in place. For example, such data may include personally identifiable data and data subject to ethical or legal restrictions such as personalized medicine data, which requires comparison to cancer and human genetics reference data. Technical expertise and high standards of security and traceability, based on insight into specific legal and ethical requirements, are also essential to ensure trust by data providers or depositors. In addition, in order for its full value to be realised, data that is made available in public repositories has to meet specific standards and formats, and it has to be curated by highly trained experts in the respective scientific discipline. Finally, to support data sharing, the hurdles for data depositors must be minimised by providing systems, services and resources to facilitate straightforward data deposition, including support with necessary data use agreements and software licenses, and consent forms for sensitive data.
A central message that is reinforced by the experience of the BioMedBridges project is that data sharing must be approached in a systematic and sustainable way. A workshop on data strategies for research infrastructures has captured the high-level points to consider in this effort, and the extended community contributed to intensive discussions on various issues around life-science data integration and sharing during a final open symposium in November 2016. BioMedBridges has shown that, while there is much room for generic tools and services that can be used across many different scientific disciplines, there are also strong domain-specific requirements for tools that are tied to scientific questions which require the work of domain experts to integrate data from different life-science disciplines.

Life science research is changing fast. There will be an ongoing need to continue the integration of life science data resources across disciplines as new lines of enquiry open up that may not yet be apparent. The central role played by initiatives such as BioMedBridges in particular in channelling the common requirements of life-science researchers and infrastructures in particular was highlighted also by the final report of the project’s e-infrastructure and scientific advisory boards (Annex 3 and Annex 4).

BioMedBridges empowers world-class research in Europe in biology and medicine by facilitating new discoveries that would not have been possible without the data bridges. Research is made more efficient by the provision of easier access to current knowledge. More challenging problems with higher levels of complexity can be tackled because of improved interdisciplinarity and merging of heterogeneous data. Improved communication and integration between different communities, especially basic biologists and clinical researchers and the biomedical sciences and ICT communities ultimately leads to better translation between basic biology and clinical research, for example through improved target discovery and validation and improved design of new (better and safer) therapeutics. The better exploitation of data and added value for research introduced by BioMedBridges adds value also to existing national investments: past investments in the creation of data are leveraged. The definition of common standards and protocols for data access and sharing within the project and across the vast BMS RI communities leads to improved data security and protection of sensitive data.

The BioMedBridges project contributes significantly to strengthening the European Research Area by coordinating and linking the BMS RIs and their attached scientific communities and by combining expertise across national boundaries and scientific domains. Addressing the “Grand Challenges” in the biomedical sciences domain will only be possible through coordination and sharing of knowledge between different scientific communities at the interface of biology and medicine, as this is the area most likely to achieve major advances in the next decade. BioMedBridges makes a significant contribution towards addressing this need.

List of Websites:
http://www.biomedbridges.eu/

Last update: 10 August 2016

Permalink: https://cordis.europa.eu/project/id/284209/reporting