## Deliverable D4.2

<table>
<thead>
<tr>
<th>Project Title:</th>
<th>Building data bridges between biological and medical infrastructures in Europe</th>
</tr>
</thead>
<tbody>
<tr>
<td>Project Acronym:</td>
<td>BioMedBridges</td>
</tr>
<tr>
<td>Grant agreement no.:</td>
<td>284209</td>
</tr>
<tr>
<td>Research Infrastructures, FP7 Capacities Specific Programme; [INFRA-2011-2.3.2.] &quot;Implementation of common solutions for a cluster of ESFRI infrastructures in the field of &quot;Life sciences&quot;</td>
<td></td>
</tr>
<tr>
<td>Deliverable title:</td>
<td>Assessment of feasible data integration paths in BioMedBridges databases</td>
</tr>
<tr>
<td>WP No.</td>
<td>4</td>
</tr>
<tr>
<td>Lead Beneficiary:</td>
<td>1: European Molecular Biology Laboratory-EBI (EMBL-EBI)</td>
</tr>
<tr>
<td>WP Title</td>
<td>Technical integration</td>
</tr>
<tr>
<td>Contractual delivery date:</td>
<td>30 June 2012</td>
</tr>
<tr>
<td>Actual delivery date:</td>
<td>7 October 2012</td>
</tr>
<tr>
<td>WP leader:</td>
<td>Ewan Birney 1: EMBL</td>
</tr>
<tr>
<td>Contributing partners:</td>
<td>N/A</td>
</tr>
</tbody>
</table>

Authors: Helen Parkinson, Nathalie Conte, Andy Jenkinson
# Contents

1. Executive summary........................................................................................................3
2. Project objectives .............................................................................................................3
3. Detailed report on the deliverable ..................................................................................4
   3.1 Background .................................................................................................................4
   3.3 Description of Work ....................................................................................................5
      3.3.1 Types of data in the project ...............................................................................5
      3.3.2 Technical workshop ............................................................................................7
   3.4 Next steps ....................................................................................................................9
4. Publications ......................................................................................................................9
5. Delivery and schedule ....................................................................................................10
6. Adjustments made ..........................................................................................................10
7. Efforts for this deliverable ...............................................................................................10
Appendices .........................................................................................................................11
Background information.......................................................................................................11
1 Executive summary

The aim of this deliverable is:

1. Understanding the technical feasibility of the WP4 integration strategy and refinement of the scientific use cases from the technical viewpoint
2. Understanding dependencies between WP3, 4 and 5 and
3. Validation and extension of the technical feasibility plan.

2 Project objectives

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

<table>
<thead>
<tr>
<th>No.</th>
<th>Objective</th>
<th>Yes</th>
<th>No</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Implement shared standards from work package 3 to allow for integration across the BioMedBridges project</td>
<td>x</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Expose the integration via use of REST based WebServices interfaces optimised for browsing information</td>
<td>x</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Expose the integration via use of REST based WebServices interfaces optimised for programmatic access</td>
<td>x</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>Expose appropriate meta-data information via use of Semantic Web Technologies</td>
<td>x</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>Pilot the use of semantic web technologies in high-data scale biological environments</td>
<td>x</td>
<td></td>
</tr>
</tbody>
</table>
3 Detailed report on the deliverable

3.1 Background

Deliverable 4.1 described summary scientific use cases for the federated web service strategy for integration by WP4. For the present deliverable, this work was expanded on by:

— examining the technical status of the WP4 contributors and other partners within BioMedBridges and assessing whether the technical plan is feasible

— determining which use cases can be delivered with existing technology and which will require de novo development of web services and

— clarifying external dependencies, such as e.g. authentication software for WP5 or limitations imposed by the RDF technology to be adopted in the latter stages of the project.

As there are dependencies between WP3 (standards) and WP4 (technical integration), a joint survey was circulated to all participating research infrastructures to assess the current use of standards, the scientific use cases, the technical approach within each infrastructure, and the current uses of research data. The survey questions, results and a summary are included in Appendix 1, 2 and 3 to this document and are analysed here.

In addition, a joint technical workshop was organised by WP3 and 4 that was also attended by WP5 members and members of the use cases. The results from this workshop, a report of which is included in Appendix 4 to this document, have flown into this deliverable and will make a significant contribution towards the further development of the technical integration strategy within the project.
3.3 Description of Work

3.3.1 Types of data in the project

All ten research infrastructures provided input to the WP3/4 data survey. In some cases the response by the infrastructure in question was collated by one partner before submission (i.e. INSTRUCT and Infrafrontier) while the responses of other infrastructures included the input of multiple institutes. In total, at least 21 responses were submitted, covering all research infrastructures.

The results were analysed and presented to partners at the monthly Technical Coordination Committee phone conference as well as to the technical personnel attending the WP3/4 workshop at the end of September. The survey was divided in two parts, one covering data types and standards (WP3) and the other covering more technical details (WP4).

Based on the survey, at least 80% of the respondents consume or produce/serve different types of data and at least 90% of them use standards when doing so. The responses are graphically represented in the tag clouds shown in Figures 1, 2 and 3.

![Figure 1 Types of data served within the project](image-url)
Briefly, the partners and research infrastructures can be divided into those producing/serving data (such as EBI/ELIXIR) and those consuming it, either because they are unable to share e.g. patient data or because the domain is not yet mature enough to have a portal or a central or federated repository. Research infrastructures that do not have portals typically do not have data models or provide APIs or web services, but do have experience in consuming data as well as preferred formats and standards to which they adhere.

The questions by WP4 revealed that 53% of the research infrastructures have an internal data model in place, 40% have expertise in REST-based service delivery and another 40% have expertise in RDF (semantic web) based data models (it should be noted here that the 40% in these two latter cases do not completely overlap, i.e. some research infrastructures have one but not the other type of expertise).

As would have been expected, technical expertise is concentrated in those research infrastructures that produce data and which include the compute heavy institutes such as EBI and STFC. Within these institutes, there is expertise in web
services and dynamic generation of RDF which will be shared with partners at the next technical meeting (see next steps below).

**3.3.2 Technical workshop**

The initial technical approach that emerged from and was subsequently adopted at the WP3/4 workshop consists of a loosely coupled web services approach.

This requires:

- cataloguing of the services provided and used by the partners (this work started with the WP3/4 data survey)
- understanding what technology is used by each partner (e.g. RESTful, SOAP, SPARQL endpoint)
- ensuring the technologies used are interoperable
- detecting dependencies between services/resources
- use of common (or convertible) inputs and outputs
- use of common identifiers and accessions (WP3)
- a monitoring framework able to detect service failure
- services representative of the resource, or extensible in future.

Challenges with this approach will occur, e.g.:

1. **A resource has data that cannot be accessed via a service** (e.g. human clinical data). WP4 will consider at least one as a special technical case and treat it as a proof-of-principle, working collaboratively with WP5. In this case, the existence of data at some summary level may suffice rather than full data access (see detailed descriptions of use cases within WP8 Personalised Medicine in Appendix 5).

BBMRI has already made progress in developing a tiered approach to this, where summary metadata, aggregated data and finally individual data are stratified to provide appropriate access to data in scenarios specified by the relevant data access committee.
2. **The institute supplying the data has no service.** This may mean that the institute requires assistance to develop a service; alternatively, it may adopt a service by providing data to another centre, e.g. via a repository which would then provide a service.

3. **Service drop out.** This may be mitigated through some degree of service duplication or up time commitment. Investigation of the EGI infrastructure is underway to explore whether this can be used to ensure up time.

4. **Semantics.** IDs can be resolved, services tested, semantic meaning of content can be encoded using ontologies. However, most web services neither use nor declare these explicitly. Initially, this does not affect data flow and is not limiting for the technical integration strategy, but it will become an issue when adopting an RDF-based data integration model.

A collection of web services itself is not an integration strategy or presentation paradigm, although it does represent a significant step in the process. BioMedBridges will also need one or all of the following components:

- centralized indexing of partner resources e.g. by Lucene/SOLR for service and data discovery, e.g. from the service registry for serving to users searching across many resources

- a robust update model supporting centralization of indexing, declared by resource

- a presentation layer for users unaccustomed to APIs

- a common provenance model to provide an audit trail of service ownership etc.

- usage statistics, to determine which are the most heavily used services, where these are compute heavy and where more resources may need to be provided

The RDF-based approach will require benchmarking, a process which is already underway for the OWLIM and Virtuoso platforms at the EBI, UMCG and FIMM. Such a benchmark, performed for the purposes of an EBI Linked Data Warehouse
pilot project, will assess load time, query performance and hardware requirements for specific platforms and deployment strategies. While we expect that technology will advance by the time we are ready to adopt these for BioMedBridges, in order to address RDF scalability issues, an understanding of these factors and the specific RDF platform that is selected for the pilot will inform the participants and make downstream integration easier. To aid this, several partners will exchange technical personnel to share expertise in this area. Specific challenges in performing this integration across BioMedBridges include:

— Use of RDF requires a common identification strategy for common resources, e.g. those provided by identifiers.org.

— Some resources that are critical for integration purposes for our use cases do not yet provide RDF dumps or SPARQL endpoints (e.g. Uniprot does, Ensembl does not). An inventory of existing services will be performed across the research infrastructures, starting with ELIXIR as a technology research infrastructures, to determine what is required to support the use cases.

3.4 Next steps

1. Follow-up WP3/4 technical workshop with technical presentations of semantic web technology and a strategy for cataloguing consortium services.

2. Close interaction with WP5 to determine details of the authentication mechanism WP4 will use.

4 Publications

N/A
5 Delivery and schedule

The delivery is delayed: ☑ Yes ☐ No

BioMedBridges is a multi-institute consortium, delivering a technical solution to data integration. Following the project kick-off meeting in March 2012, many institutes needed to employ new personnel. The WP4 technical lead at the EBI has now been hired, but will not formally start until 15 October 2012. In view of this and similar situations at other institutes, completion of the deliverable had to be delayed until enough technical personnel were in place to provide input on the detailed data survey and a technical workshop with participants from WPs 3, 4 and 5 could be held. The workshop took place on 26-27 September 2012; a report is supplied in Appendix 5 to this document.

6 Adjustments made

N/A

7 Efforts for this deliverable

<table>
<thead>
<tr>
<th>Institute</th>
<th>Person-months (PM)</th>
<th>Period</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>actual</td>
<td>estimated</td>
</tr>
<tr>
<td>1 EMBL-EBI</td>
<td>3 (in kind)</td>
<td>6</td>
</tr>
<tr>
<td>Total</td>
<td>3 (in kind)</td>
<td></td>
</tr>
</tbody>
</table>
Appendices

1. Results of the WP3/4 data survey
2. WP3/4 data survey questionnaire
3. Summary presentation of the WP3/4 survey results
4. Report from the WP3/4 technical workshop
5. Detailed personalised medicine plan exploring the issues around patient data (expanding on deliverable 4.1)

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 WP 11 will provide recommendations for other emerging technologies to use, aiming for the semantically richest integration.
<table>
<thead>
<tr>
<th>Work package number</th>
<th>WP4</th>
<th>Start date or starting event:</th>
<th>month 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Work package title</td>
<td>Technical Integration</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Activity Type</td>
<td>RTD</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Person-months per participant</td>
<td>58</td>
<td>40</td>
<td>38</td>
</tr>
</tbody>
</table>

**Objectives**

1. Implement shared standards from WP 3 to allow for integration across the BioMedBridges project
2. Expose the integration via use of REST based WebServices interfaces optimised for browsing information
3. Expose the integration via use of REST based WebServices interfaces optimised for programmatic access
4. Expose appropriate meta-data information via use of Semantic Web Technologies
5. Pilot the use of semantic web technologies in high-data scale biological environments.

**Description of work and role of participants**

We will provide a layered, distributed integration of BioMedBridges data using latest technologies. A key aspect to this integration will be the internal use of standards, developed in WP 3 which will provide the points of integration between the different data sources. The use of common sample ontologies (WP 3) will provide integration between biological sample properties, such as cell types, tissues and disease status, in particular bridging the EuroBioImaging, BBMRI, Elixir and Infrafrontier projects. The use of Phenotype based ontologies will provide individual and animal level characterisation which, when these can be associated with genetic variation, will provide common genotype to phenotypic links, and this will be used to bridge the ECRIN, EATRIS, INSTRUCT, BBMRI, Infrafrontier and Elixir Projects. The use of environmental sample descriptions and geolocation tags will bridge between EMBRC, ECRIN, ERINHA, EATRIS and Elixir. The use of chemical ontologies will help bridge between EU-OpenScreen, ECRIN, EuroBioImaging, 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 workpackage will be implemented in a collaborative manner across the BMSs, with frequent physical movement of individuals.

The proposed technologies are:

1. REST-based “vignette” integration, allowing presentation of information from specific databases in a human readable form. An example is shown in Figure 1. These resources allow other web sites to “embed” live data links with key information into other websites. This infrastructure would then be used to provide browsers that, on demand, bridge between the different BioMedBridges groups – for example, information which can be organised around a gene or a chemical compound would be presented across the BioMedBridges project.

2. Web service based “query” integration, where simple object queries across distributed information resources can be used to explore a set of linked objects using the dictionaries and ontologies present. Each request will return a structured XML document.

3. Scaleable semantic web based technology. We are confident that semantic based technology can work for the rich but low data volume meta data (eg, sample information) which we will expose using semantic web technologies such as RDF and SPARQL. However, it is unclear whether this scales to the very large number of data items or numerical terms in the BioMedBridges databases (such as SNP sets or numerical results from Clinical trials) We will pilot a number of semantic web based integration of datasets, using RDF based structuring of datasets In the latter phases of the project we will look to align these solutions to other broader standards in the eScience community, taking input from the Technology Watch (WP11) group; we hope in many cases our technology choice which has been already informed by alignment to future eScience technology (eg. 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.
<table>
<thead>
<tr>
<th>No.</th>
<th>Name</th>
<th>Due month</th>
</tr>
</thead>
<tbody>
<tr>
<td>D4.1</td>
<td>A brief collation of existing use cases to start the agile software iteration</td>
<td>3</td>
</tr>
<tr>
<td>D4.2</td>
<td>Assessment of feasible data integration paths in BioMedBridges databases</td>
<td>6</td>
</tr>
<tr>
<td>D4.3</td>
<td>Pilot integration using REST Web Services</td>
<td>18</td>
</tr>
<tr>
<td>D4.4</td>
<td>Identification of feasible BioMedBridges pilots for semantic web integration</td>
<td>18</td>
</tr>
<tr>
<td>D4.5</td>
<td>Pilot integration of REST based vignette services for the second round BMS projects</td>
<td>24</td>
</tr>
<tr>
<td>D4.6</td>
<td>Pilot integration of Web Services based simple object queries</td>
<td>36</td>
</tr>
<tr>
<td>D4.7</td>
<td>Report on the scaleability of semantic web integration in BioMedBridges</td>
<td>36</td>
</tr>
<tr>
<td>D4.8</td>
<td>Report on Web Services based integration of BioMedBridges integration across all appropriate services</td>
<td>48</td>
</tr>
</tbody>
</table>
REPORT

BioMedBridges Technical Workshop (WP3/4)

25-27 September 2012

European Bioinformatics Institute, Hinxton
<table>
<thead>
<tr>
<th>Contents</th>
</tr>
</thead>
<tbody>
<tr>
<td>List of participants .................. 3</td>
</tr>
<tr>
<td>Introduction and aims of the workshop ........................................ 4</td>
</tr>
<tr>
<td>Presentations by WP contributors ................................................. 4</td>
</tr>
<tr>
<td>WP3 Work plan proposals .................. 4</td>
</tr>
<tr>
<td>Research Infrastructure needs ....................................................... 5</td>
</tr>
<tr>
<td>Biobank catalogue levels and sharing ............................................... 5</td>
</tr>
<tr>
<td>PIMS for BioMedBridges, a perspective ........................................... 6</td>
</tr>
<tr>
<td>Infrafrontier resources ................................................................. 7</td>
</tr>
<tr>
<td>BioMedBridges user-centered design – introduction ................................ 7</td>
</tr>
<tr>
<td>WP5 progress update ................................................................. 8</td>
</tr>
<tr>
<td>Types of data in the project: Brief presentation of WP3/4 data survey ........ 8</td>
</tr>
<tr>
<td>Brainstorming/icebreaker sessions .................................................. 8</td>
</tr>
<tr>
<td>Introduction ................................................................. 8</td>
</tr>
<tr>
<td>Reports from group discussions ....................................................... 9</td>
</tr>
<tr>
<td>Overall project status and progress ................................................ 9</td>
</tr>
<tr>
<td>Introduction of eagle-i ................................................................. 10</td>
</tr>
<tr>
<td>Focus groups ................................................................. 10</td>
</tr>
<tr>
<td>Data producers ................................................................. 10</td>
</tr>
<tr>
<td>WP4/5 overlap ................................................................. 12</td>
</tr>
<tr>
<td>Data consumers ................................................................. 13</td>
</tr>
<tr>
<td>WP8 Use case: Personalized Medicine .............................................. 13</td>
</tr>
</tbody>
</table>
List of participants

- Deborah Alferez, WP2, VUMC, EATRIS
- Ewan Birney, WP4, WP8, WP11, EBI, ELIXIR
- Jan Willem Boiten, WP3, WP4, WP2, CTMM, EATRIS
- Benjamin Braasch, WP8, WP4, WP5, WP2, UDUS
- Alvis Brazma, WP10, WP5, EBI, ELIXIR
- Cath Brooksbank, WP12, WP2, EBI, ELIXIR
- Jon Chambers, WP5, EBI, EU-OPENSSCREEN
- Nathalie Conte, WP3, WP7, EBI, ELIXIR
- Chuck Cook, WP3, EBI, EMBRC
- Martin Eckert, WP2, WP3, WP4, WP5, WP8, UDUS, ECRIN
- Henrik Edgren, WP8, FIMM, EATRIS
- Adam Faulconbridge, WP3, WP4, WP5, WP10, EBI, ELIXIR
- Martin Fransson, WP3, KI, BBMRI
- Philipp Gormanns, WP3, WP4, WP5, WP6, WP7, HMGU, Infrafrontier
- Krister Helbing, WP5, WP3, TMF, EU-OPENSSCREEN, EATRIS, BBMRI
- Jon Ison, WP3, WP7, EBI, ELIXIR
- Andy Jenkinson, WP4, EBI, ELIXIR
- Töresin Karakoyun, WP8, WP4, WP5, WP2, UDUS, ECRIN
- Stefan Klein, WP3, WP4, WP5, ErasmusMC, Euro-Bioloimaging
- Roland Krause, WP5, WP3, TMF, EU-OPENSSCREEN, EATRIS, BBMRI
- Narayanan Krishnan, WP4, STFC, INSTRUCT
- Klaus Kuhn, WP5, WP3, WP8, WP4, TUM-MED, BBMRI
- Julie McMurry, WP4, EBI, ELIXIR
- Chris Morris, WP3, WP4, WP5, WP9, STFC, INSTRUCT
- Juha Muilu, WP3, UH-FIMM, EATRIS, BBMRI
- Helen Parkinson, WP3, WP6, WP7, EBI, ELIXIR
- Ardan Patwardhan, WP9, EBI, ELIXIR
- Francis Rowland, WP12, EBI, ELIXIR
- Gabriella Rustici, WP6, EBI, ELIXIR
- Ugis Sarkans, WP5, WP10, EBI, ELIXIR
- Sebastian Seufert, WP5, UDUS, ECRIN
- Stephanie Suhr, WP1, EBI, ELIXIR
- Morris Swertz, WP3, UMCG, BBMRI
- Imre Västrik, WP8, WP3, UH-FIMM, EATRIS
- Martyn Winn, WP9, WP5, WP3, WP4, STFC, INSTRUCT
**Introduction and aims of the workshop**

**Ewan Birney, Helen Parkinson**

- BioMedBridges is a pan-ESFRI project that will serve a community of at least ½ million researchers in Europe
- The intention is to create links between the biomedical sciences research infrastructures – to create a web-based service infrastructure
- To achieve this, focus will be on REST-based services, serving XML of the delivering infrastructure’s choice
- Will move towards RDF-based coordination in the future
- Researchers will be able to use the registry from ELIXIR and EBI
- Need detailed use case descriptions as e.g. from WP8, then determine feasibility and tools needed:
  - Is a gene that we find mutated a known cancer gene?
    - doable: REST call to Ensembl on COSMIC
  - What was the disease course of other patients with the same mutation in gene X?
    - This is hard – is it EATRIS or BBMRI?
  - Is the gene directly druggable?
    - doable: chEMBL REST call
- Build on WP3/4 data survey responses; explore where either no or not the expected responses were given
- Develop a detailed deliverable plan for WP3 and 4.

**Presentations by WP contributors**

**WP3 Work plan proposals**

**Martin Fransson**

- Presented a suggestion for structuring the work of WP3 (4 tasks, 4 deliverables).
Research Infrastructure needs

Juha Muilu

— Nordic countries have a long tradition of large-scale biobanking and comprehensive, population-based health data registries linkable on unique personal identifiers, enabling follow-up studies spanning many decades

— National Institute for Health and Welfare:
  - 170k samples
  - GWAS and sequence data (50k+2K)

— Single access point for biobanking research: biobanks, hospital clinical standards

— Data and sample discovery, want feedback on how the samples are used, report usage to the funders - need to know what is done to samples

— Data access challenges:
  - identification of the applicant
  - handling of applicants
  - control data usage

— Existing solutions
  - REMs - resource entitlement management system – CSC
  - Existing IdP policy federation fw (HAKA - Finnish universities)
  - DataShield - University of Leicester
  - BioShare - federated meta-analysis
  - iRODs – rule based sharing platform
  - ORCID – researcher-based id program

Biobank catalogue levels and sharing

Morris Swertz

— What level of biobank data is needed?
  - only the name (MIADIS) - list of cohort and basic info
  - metadata on the variable - e.g. questionnaires
- some aggregate data - is data available, how many, counts (SAIL)
- individual anonymized data (OPAL)
- pseudonymised links, e.g. cancer registry, links between biobanks

— Importance to harmonize data

- Building catalogue for biobanks, who has what; lifelines
- Harmonization model for these 5 levels; many come from hospitals.
- Observ-OM - Observ-TAB - represented generically the model and some of the process
- can be used as a data exchange process

— Can we project all these data levels on to a unified syntax solution that clinical researchers can understand?

- Simple tabular format (researcher cannot do XML, RDF, etc)
- Mapping to more complex models, formats, SW - want to share

— User experience for searching these data

- Molgenis - national catalogue for the biobank, basic, minimal info that people were willing to share
- Lifelines - better one, research platform
- Bio-flavoured RDF

**PiMS for BioMedBridges, a perspective**

*Narayanan Krishnan*

— Presentation of PiMS:

- Management of records for targets, experiment, samples in molecular biology
- Exposing RDF on a web service
- XML based representation format
- Resource and relationship as triplets
- Help in easier data interoperability and data evolution
- Technical details: Jena API
• RDF schema representing relational database - in house tool
• RDF/XML method – REST-based model for service
• Linked to Uniprot core ontology
• RESTful RDF call; no SPARQL backend
• Expose on a request basis

Infrafrontier resources

Philipp Gormanns

— Use case presentation from Infrafrontier - EMMA, systemic phenotyping
— Mouse clinic - Systemic phenotyping - IMPC

BioMedBridges user-centered design – introduction

Francis Rowland, External Services, EBI

— Apply user-centred design during development of services in BioMedBridges
— Apply Design Council design framework - design process to discover, define, develop, deliver
  • Discover: who are your users what do they do? How does BMB fit into a workflow, where does it fit in?
  • Define: converge idea: from fuzzy to definite ideas
  • Develop: coders and developers: make and test (usability testing, validate your assumptions)
  • Deliver: new service, collect feedback from people
— Discovery starts now for BioMedBridges
— Short agile iterations help this process
— Personas: based on in-depth interviews with real people (users) with varying skills and interest. Then assembles into a schematic persona which is used during the design process – possible to develop freely without e.g. discussing about real people. Loop is closed when prototypes are tested with real users again. Reiterative process.
**WP5 progress update**

*Klaus Kuhn*

- WP5 survey
- Developed master plan WP5 for 2012-2013
  - 12 months, 18 months for legal and regulatory aspects; WP5 survey relates to this
  - Security requirements analysis - usage scenarios -> phone interviews will be held later
  - Domain scenario - can use this here
  - Encryption etc.
  - What systems exist now - requirements for linking - implication will support what is going on for WP3/4
  - Microdata - individual level, stats data = aggregate (age groups, male, female) - linkage for these

**Types of data in the project: brief presentation of WP3/4 data survey**

*Nathalie Conte*

- Got responses from all 10 research infrastructures
- 80% or respondents serve data, 86% use data
- Various types of data served, standards and formats used (serve and consume)
- Barriers in three areas: sample/data standards used and availability of data, technical issues, people/ethical issues.

**Brainstorming/icebreaker sessions**

**Introduction**

*Helen Parkinson*

- What tools should we build – from where to where
- Issues with the technology plan – challenges, blockers
- Standardization tasks - what first, prioritization
— Dependencies – internal/external e.g. WP5/WP4 – service-based authentication
— What tools exist which we can use, repurpose, add a service to

**Reports from group discussions (Group A, B, C, D)**
*Rapporteurs: Philipp Gormanns, Martyn Winn, Imre Västrik, Julie McMurry*

— All groups learn about the process of building successful bridges
— Create a BioMedBridges data inventory/registry
— Interacting with LIM systems - possible for some projects
— Must overcome paranoia of sharing data, data not organised well, data hard to share
— Infrastructures consume and produce at different levels (some only produce, others only consume)
— Personalised medicine use case has data security challenge

**Overall project status and progress**

*Stephanie Suhr, BMB Project manager*

— Brief overview of main points in draft ethical governance framework – overall pragmatic approach, use expertise of local ethics or data access committees to the furthest possible extent
— Data providers must ensure that legal and ethical requirements for data they serve within BioMedBridges have been addressed
— Project overview: BioMedBridges is a cluster project that is intended to bring the biomedical sciences research infrastructures together and start building technical bridges
— The project structure reflects this, with the partners being represented by the Research Infrastructure leads
— Internal communication in the project:
  - Research Infrastructure to Research Infrastructure (Executive Steering Committee; RI coordinators represent partners)
  - Each Research Infrastructure to the BioMedBridges partners it represents
Within the Research Infrastructure, including RI members that are not BioMedBridges members

— Progress is now needed in the project, use cases need to be very clearly defined – also see comment by Scientific Advisory Board during kick-off meeting.

Introduction of **eagle-i**

*Julie McMurry, WP4 technical lead*

— US NIH-funded project
— Challenge to overcome: resources can be invisible to users/the scientific community
— 9 participating US-based institutions
— User feedback positive: users found what they were looking for, also found resources they did not know existed
— Feature of eagle-i suite: semantic web entry to create new resource
— Ontology-driven architecture is powerful and flexible

Focus groups

**Data producers**

*Rapporteur: Helen Parkinson*

— what is the level at which data can be shared:
  ▪ descriptive data
  ▪ aggregate data
  ▪ everything is open - sample level/individual-level data

— BioMedBridges could facilitate access via a data access committee

**Develop a registry of services**

1. Service - where use of data has to be approved by a data access committee, need a tool to apply for data access. This exists within the [EGA](#) already, [iRODS](#) could also be used for this
2. Data sharing between e.g. two groups – a place where data can be shared securely
   — Resource catalogue should include information on how to get access to protected data or services
   — Resource catalogue should include data, tools and services
   — Resources could be shown on a geographical map with everything, e.g. methods, services, data, in one representation but tailored for different queries
   — BioSD could link from there and then link to everything else, the provider will like to consume it
   — WP5 will need to provide data authentication – some data providers may not be willing to place things in a central index -> would need a central index with some federation
   — For data, BioCatalog could be used
   — Include:
     - providers of analytical procedures e.g. barc - list of contact for companies and institutes for information on specific experiments
     - services and tools
     - resources

**ECRIN**

—Could provide meta registry for clinical trials
—Data is pseudonymised or anonymised
—Could list the name of a clinical trial and link to a contact person or the study owner; average time to get access to data, process and conditions to get access

**INSTRUCT**

— Data sharing tool – part of PiMS – could catalogue the data later
—Could work on many levels – researchers can be in control of what is included – could register anything if asked.

**BBMRI**

— Include the BBMRI biobank catalogue (EU biobanks), list of catalogues of BBMRI Sweden, but limit the granularity

**General**
— Most RIs probably provide data/services online; in case of Euro-BioImaging this could be the front end for Euro-BioImaging tools
— Upload could be via a mixed model: upload part of the info and then check with owner whether info is correct
— Can start right away with registering tools, need permissions for datasets
— Linking up data is a different bridge/layer of complexity – necessary level of granularity not yet available
— Need to coordinate different ontologies used by partners as there may be clashes
— In cases where services already exist, just add a link (instead of duplicating the resource)
— Produce a document describing the resource and what it is for - add a link to a publication and the documentation, support for uploading
— Service registry: build a form for uploading info
— Service registry: include an Excel upload

**WP4/5 overlap**

*Rapporteur: Ugis Sarkans*

— Federated identity: who is the user (involves identity providers); federated authorisation: what privileges does the user have (involves service providers)
— Well-studied technical solutions exist for communications between identity providers and service providers
— Find a practical way to get input/feedback from “local” ethics- or data access committees during the development of the BioMedBridges authentication framework
— The main security threats to data integration in the BMB context were identified as:
  - leakage of personal data due to lack of care with data handling
  - combining datasets in ways that enable re-identification
  - leaks of IP related to drug candidates
  - research data falsification
— Options to address these threats (especially the first two) include minimizing the need to download data by issuing timed certificates that
enable data access through a virtual machine running on the server, or homomorphic encryption, which enables certain operations on encrypted data without prior decryption

— Weighing complexity vs. robustness of proposed security solutions: need a pragmatic, agile rollout strategy so avoid various corner cases delaying overall progress

Data consumers

Rapporteur: Gabriella Rustici

— EATRIS: case matching – molecular profile, find similar patients, follow-up
— INSTRUCT: EMDB link to PDBe links
— Data provider map of ICGC as an example: who has data, what do they have and which tools
— People are often willing to share data, but they don’t have time - need a communications plan; first contributions are the hardest in an effort

WP8 Use case: Personalized Medicine

Rapporteur: Nathalie Conte

— Who is the tool for? What bridges do we want?

  ▪ Bridges: omics data, generate a list of targets, e.g. mutated genes:
    → translate mutated gene into protein
    → integrate data from structure
    → integrate data from chemical compounds?

— Determine what data is there and whether it can be used – ask partners what suitable data they can provide
— Is it possible to get data from outside the consortium, e.g. from ICGC?
— build a data provider map, show participating RI or institute and the available tool.
<table>
<thead>
<tr>
<th>Infrastructure Name(s)</th>
<th>WP3</th>
<th>WP4</th>
<th>WP5</th>
<th>WP6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Center for Translational Molecular Institute</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>CSC - IT Center for Science</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Heinrich-Heine-University Düsseldorf</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Erasmus MC</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>STFC Daresbury Laboratory</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Stazione Zoologica Anton Dohrn</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Helmholtz Zentrum Muenchen - German TUM-MED</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>EBI University Duesseldorf</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>EuroBioImaging</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
</tbody>
</table>

**Results of the WP3/4 Data Survey**

<table>
<thead>
<tr>
<th>Question</th>
<th>Yes</th>
<th>No</th>
<th>Unknown</th>
</tr>
</thead>
<tbody>
<tr>
<td>Do you serve data?</td>
<td>3/3</td>
<td>2/2</td>
<td>2/2</td>
</tr>
<tr>
<td>Do you use standards when serving data?</td>
<td>3/3</td>
<td>No</td>
<td>2/2</td>
</tr>
<tr>
<td>Which standards do you support for serving data? Please provide a name, a url/citation and type (e.g. content, messaging, or ontology) for each standard.</td>
<td>OBI, <a href="http://obi-ontology.org">http://obi-ontology.org</a></td>
<td>ChEBI, <a href="http://www.ebi.ac.uk/chebi/">http://www.ebi.ac.uk/chebi/</a></td>
<td>AIM, <a href="">ftp://ftp.wwpdb.org/pub/emdb/doc/AIM</a></td>
</tr>
<tr>
<td>What format(s) do you serve data in? Please be specific, if XML, which XML?</td>
<td>Tabular (MAGE-TAB, XGAP, Observ-OM)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Do you use data from databases?</td>
<td>2/3</td>
<td>Yes</td>
<td>1/2</td>
</tr>
<tr>
<td>Which kinds of data do you use and where do these data come from (e.g. genotype data from the EGA)?</td>
<td>Internal to consortia</td>
<td>Macromolecular structures from wwPDB, <a href="http://www.pangaea.de/">http://www.pangaea.de/</a></td>
<td>Small molecule structures from ChEMBL, <a href="http://www.ebi.ac.uk/chembl/">http://www.ebi.ac.uk/chembl/</a>, Chemical databases, coming from pharmacy,</td>
</tr>
<tr>
<td>Do you use standards (formats, content or ontologies) when using or acquiring data?</td>
<td>2/3</td>
<td>Yes</td>
<td>1/2</td>
</tr>
<tr>
<td>What formats do you acquire data in? E.g. GFF</td>
<td></td>
<td>1/2</td>
<td></td>
</tr>
<tr>
<td>Attachment 2: Results of the WP3/4 data survey</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>------------------------------------------------</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Does your technical group have an internal data model?</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Yes (2/3)</strong></td>
<td><strong>Yes (1/2)</strong></td>
<td><strong>Yes (1/2)</strong></td>
<td><strong>No (2/2)</strong></td>
</tr>
<tr>
<td><strong>If you have an internal data model, is there web accessible documentation of this?</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(1/2)</td>
<td>No</td>
<td>(3)</td>
<td>No</td>
</tr>
<tr>
<td>Insert a link to web accessible documentation for your data model here:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>List the 3 most important datasets which your group can share openly that do not have restrictions on use relating to national law, donor consent or ethics committee approvals. Use a pubmed id and/or a database accession/URL if available.</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>List the 3 most important datasets which your group can share, e.g. via managed access mechanisms, that have some form of restriction on use relating to national law, donor consent or ethics committee approval.</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Does your technical group have expertise in REST-based service delivery?</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(2/3)</td>
<td>(1/2)</td>
<td>(1/2)</td>
<td>(1/2)</td>
</tr>
<tr>
<td><strong>Please provide up to 3 example servers.</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Does your technical group have expertise in RDF (Semantic Web) based data models?</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(2/3)</td>
<td>(1/2)</td>
<td>(1/2)</td>
<td>(1/2)</td>
</tr>
<tr>
<td><strong>Is there an example RDF dump or SPARQL endpoint for this?</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td><strong>If available, please insert download link:</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EM-Madrid: Not accessible from outside, but Scipion database is queried using SPARQL</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Work package 3, who are you?

These questions address the standards, ontologies and formats needed to integrate the various infrastructures. They are designed to address these issues for using and serving project and external data.

1. Your Name
   
2. Your email address
   
3. Institute Name
   
4. Infrastructure Name (s)
   - [ ] BBMRI
   - [ ] ELIXIR
   - [ ] EATRIS
   - [ ] ECRIN
   - [ ] EMBRC
   - [ ] ERINHA
   - [ ] EU-OPENSESCREEN
   - [ ] EuroBioImaging
   - [ ] Infrafrontier
   - [ ] INSTRUCT
   
Other (please specify)
Work package 3, serving data 1

5. Do you serve data?

☐ Yes  ☐ No
6. What kind of data do you serve? E.g. Protein structures, gene expression?

1. 
2. 
3. 
4. 

*7. Do you use standards when serving data?

- [ ] Yes
- [ ] No
8. Which standards do you support for serving data? Please provide a name, a URL/citation and a type e.g. content, messaging, or ontology for each standard.

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td></td>
</tr>
<tr>
<td>2.</td>
<td></td>
</tr>
<tr>
<td>3.</td>
<td></td>
</tr>
<tr>
<td>4.</td>
<td></td>
</tr>
<tr>
<td>5.</td>
<td></td>
</tr>
<tr>
<td>6.</td>
<td></td>
</tr>
</tbody>
</table>

9. What format(s) do you serve data in? Please be specific, if XML, which XML?

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td></td>
</tr>
<tr>
<td>2.</td>
<td></td>
</tr>
<tr>
<td>3.</td>
<td></td>
</tr>
<tr>
<td>4.</td>
<td></td>
</tr>
<tr>
<td>5.</td>
<td></td>
</tr>
<tr>
<td>6.</td>
<td></td>
</tr>
</tbody>
</table>
10. Do you use data from databases?

☐ Yes

☐ No
Work package 3, using and acquiring data

11. Which kinds of data do you use and where do these data come from (e.g. genotype data from the EGA)?

1. 
2. 
3. 
4. 
5. 
6. 
12. Do you use standards (formats, content or ontologies) when using or acquiring data?

☐ Yes

☐ No
13. What data standards do you use when consuming/acquiring data? Please provide a name, a url/citation and type (e.g. content, messaging, or ontology) for each standard.

1. 
2. 
3. 
4. 
5. 
6. 

14. What formats do you acquire data in? E.g. GFF

1. 
2. 
3. 
4. 
5. 
6. 
<table>
<thead>
<tr>
<th>Work package 3: Your use cases and blockers</th>
</tr>
</thead>
<tbody>
<tr>
<td>15. What are your main use cases for BioMedBridges? Please be specific. E.g. integration of GWAS data with gene expression data in human and mouse.</td>
</tr>
<tr>
<td>16. What are your main barriers to data integration? E.g. Lack of available data in the public domain, lack of clinical annotation, need to harmonise across studies.</td>
</tr>
</tbody>
</table>
17. Does your technical group have an internal data model?

- Yes
- No
18. If you have an internal data model, is there web accessible documentation of this?

☐ Yes  ☐ No
19. Insert a link to web accessible documentation for your data model here:
20. List the 3 most important datasets which your group can share openly that do not have restrictions on use relating to national law, donor consent or ethics committee approvals. Use a pubmed id and/or a database accession/URL if available.

1. 
2. 
3. 

21. List the 3 most important datasets which your group can share, e.g. via managed access mechanisms, that have some form of restriction on use relating to national law, donor consent or ethics committee approval.

1. 
2. 
3. 

22. List the 3 most important datasets which your group would like to consume. Use a pubmed id/text description/accession+URL.

1. 
2. 
3. 
23. Does your technical group have expertise in REST-based service delivery?

- [ ] Yes
- [ ] No
24. Please provide up to 3 example servers.

1. 

2. 

3. 
25. Does your technical group have expertise in RDF (Semantic Web) based data models?

- Yes
- No
26. Is there an example RDF dump or SPARQL endpoint for this?

☐ Yes  ☐ No
27. If available, please insert download link:

1.

2.

3.
28. Please provide the name of your technical contact.

29. Please provide the Email address of your technical contact.
WP3/4 Survey Summary

Nathalie Conte, Ewan Birney and Helen Parkinson
WP3/WP4 survey: Overall response summary

- Responses from 10 out of 10 RIs:
  - BBMRI
  - ELIXIR
  - Euro-BioImaging
  - ECRIN
  - Infrafrontier
  - EATRIS
  - EMBRC
  - INSTRUCT
  - ERINHA
  - EU-OPENSSCREEN

- INSTRUCT and Infrafrontier provided a summary response

- Other RIs provided multiple responses

- At least 21 individual responses in total
WP3 questions: Response summary

- 80% respondents serve data
  - 91% use standards when serving data
- 86% use data from databases
- 100% use standards when acquiring data
What data served?

Tag cloud, limits 50 words

- annotations
- assays
- bioactivity
- biobanking
- biological
- biosamples
- cellular
- chemical
- clinical
- commercial
- data
- diffraction
- electron
- em
- environmental
- est
- expression
- gene
- generic
- genomics
- genotypes
- geo-referenced
- imaging
- maps
- medical
- metadata
- metatranscriptomics
- micrographs
- mirrors
- mouse
- phenotypes
- physical preparation
- protein
- purification
- qtl
- records
- registries
- related
- research
- ribosomal
- rna
- sector
- sequence
- statistical
- strain
- targets
- tomograms
- transcriptomics
- treatment
Standards for consuming and acquiring data?

Tag cloud, limits 50 words

bam bao bedgraph biomart bridg cdash cdisc chebi csv dicom fasta gff gsc hl7 html icd10 inchi jpeg jpg json loinc meddra miabe mmcf mp mzdata mzml mzxml obo observ-om odm oracle owl pcrom pdb pdf php plink sdf sdtx send snomed sq tiff tsv vcf xgap xml xsd

BioMedBridges
Formats
(acquire and consume)
Use cases

- **Personalised medicine and Integration of different data sources from different centers** WP8-WP10
  - Using "omics" data for understanding disease pathogenesis, identification of biomarkers and improving treatment
  - Using biosampling and imaging data together with clinical data for human personalised medicine trials

- **Environmental data integration**
  - Integration of sequence data with environmental parameters for ecosystems biology

- **Integration of structural data** WP9
  - Integration of EM data generated by Instruct with services provided by Elixir
  - Interpretation of complexes/viruses in EM: protein-protein interactions, and domain prediction
  - Mapping protein structures from Elixir onto experimentally-obtained cross-sections of complexes
  - Integration of structural data with gene variation (e.g. SNPs)

- **Crossing the species bridges between Human and Mouse** WP7
  - Translation of genotype/phenotype relations from mouse to man in the field of diabetes and obesity

- **Demonstrate utility of interoperability of large scale image data sets from different biological scales** WP6
  - Integrated access to systematic imaging data of disease gene function in cultured cells (Euro-BioImaging)
  - WP6 Tissue microarrays of diseased tissue from human patients (EATRIS, BBMRI) and mouse models (Infrafrontier)

- **Sharing data in a secure fashion** WP5

- **Harmonization of Standards/Ontology** WP3/WP4
Barriers to integration

- Barriers mentioned in survey responses can be sorted into three different categories:
  - Sample/data standards and availability issues
  - Technical issues
  - People/ethical issues
Barriers to integration

- **Sample/data standards and availability issues**
  - Lack of semantic interoperability: standards are poorly adopted in practice
  - Lack of a common disease ontology
  - Lack of common sample/data model (with joint annotation metadata, ontologies)
  - Lack of consistent data sets across various data domains relevant to translational research
  - Complementary and well-documented public phenotyping data from mouse and man still need to be identified and brought together
  - Lack of well-organised public domain data on drugs-target pairs
  - Lack of structural data for many proteins, i.e. much required data simply doesn’t exist
  - Lack of comprehensive genome variant disease association information in the public domain in easily usable manner

- **Technical issues**
  - Lack of suitable portals / data displays / queries standardized method to define data mappings for pooling data from different resources
  - Lack of adoption of standards among software packages
  - Lack of user friendly tools for organising, managing and sharing data
  - Lack of privacy/security

- **People/ethical issues**
  - Language barrier
  - Reluctance to share biological/clinical data
  - Crossing a skill boundary - few people can interpret different sorts of data
WP4 questions:
Response summary

- Internal data model exists?
  - 53% yes, 1 skipped

- Accessible documentation?
  - 37% yes, 1 skipped, 4 actually included a link

- REST interface experience?
  - 40%, 7 skipped, 4 provided example URLs

- RDF expertise?
  - 40%, 1 skipped
  - Instruct has an RDF dump SPARQL end point
Attachment 5

Authors: Imre Västrik, Henrik Edgren

WP8 Personalised medicine: detailed use case

To determine services that would be useful from an end user view point, we describe the example personalized medicine use case of an acute myeloid leukemia patient for which at least some of the following types of data is available:

— Exome-seq data for both germline and one or more cancer samples. Coverage 30 to 40x for germline and 60-80x for cancer
— Low coverage whole genome sequencing (wgs) data for germline (8x) and cancer (15x)
— Targeted sequencing data for predefined panels of cancer genes, both for germline and cancer
— Amplicon sequencing for somatic mutations, used for validation and quantification of mutation frequencies
— RNA-seq data with ~80 million read pairs for the cancer sample
— Protein array data for the cancer sample: protein abundance as well as phosphorylation state data for 100-200 central signalling proteins
— Drug screening data: cancer cell dose response data for a collection of 250 cancer drugs as well as some signaling pathway inhibitors not in clinical use
— DNA copy number data, either from exome-seq, low coverage wgs or array-CGH

From these, at least the following types of data have been derived:

— Somatic mutations in the cancer sample
— Somatic mutation frequencies
— Germline variants (at least for select cases)
— DNA copy number status for all genes
— Genomic rearrangements (inversions, duplications, translocations)
— Gene expression data
— Data on whether a somatic mutation is expressed
— Dose response data (cell viability) for 250 chemical entities

Questions / use cases

The answers to the questions presented here fall into one of these two possible categories:

1. The answer is available directly (exists precomputed)
2. Data can be provided that allows computation of the answer

This is the difference between (1) checking the answer from a list of validated somatic mutations (i.e. what ends up in a table in a publication) and (2) providing access to e.g. the necessary parts of .bam files for a set of cases, allowing us to call variants ourselves. The latter is likely to be important to ensure that variants from our own data and those provided by research infrastructures are called the same way. Or, more generally, providing the "building blocks" for users to calculate the answer themselves, even when the answer has not been precomputed.

In the questions below, to keep the text simple, the term "mutated gene" is used throughout, but the same questions apply also for deleted, amplified and otherwise rearranged genes.

**Is a gene that we find mutated a known cancer gene?**

In other words, integration with the Cancer Gene Census gene list as well as the mutation data in Cosmic. The former is easy to do locally, but there would also be value in doing it programmatically from the primary data source, i.e. always having the latest version.

**If the mutated gene is not a "recognized" known cancer gene (Cancer Gene Census), is it known to be mutated all the same?**

The primary curated data set we know of is Cosmic. Basically, answering the question whether the gene is recurrently mutated and therefore likely a driver
mutation. This would also be very useful to calculate directly from various studies that have not yet been curated to e.g. Cosmic.

For non-curated data, the data deposited in e.g. EGA would be highly useful, for example studies like EGAS00001000288. Any solutions that would allow easy queries against e.g. the EGA data sets one has already been granted access to would be very useful. Otherwise handling the access to numerous studies, each of which may require a separate permission to access, quickly becomes onerous. Alternatively, the data would need to be downloaded, defeating the purpose of bridging across research infrastructures and resulting in the same data being duplicated across many institutions.

Is the gene likely to be an oncogene or tumor suppressor?
This might best be seen from our own data (single or double hit mutations), but for a recurrently mutated gene with sufficient amounts of mutations in e.g. Cosmic, it might be possible to "guess" whether it is an oncogene or tumour suppressor.

Is the mutation we see in our patient located in a mutational hotspot?
Trying to answer the question "yes, the mutation lies in a cancer gene, but how likely is it that this specific mutation is a driver vs. just a random passenger event in an important gene".

Are mutations in the gene expressed in other cases?
When we find a somatic expressed mutation in gene X and gene X is mutated in other cases, do those cases express it too? If it is a driver oncogene, they should express it.

We observe mutation of two or more genes in the same patient. Are these same genes mutated together in other patients?

Is the gene directly druggable?
Is there an existing approved drug that targets this protein? Or a drug far enough in development that access to it might be possible through clinical trials?
Is the gene indirectly druggable?
If the gene is not directly targetable with an existing drug, can we establish another link that would suggest a certain drug? For instance an upstream kinase?

Do mutations in the gene have clinical consequences?
Beyond being direct targets of treatment, are mutations predictors of e.g. good or poor prognosis? What is known in the literature on the clinical links of a specific mutation or mutations in gene X in general? Basically, are mutations or expression changes in the gene biomarkers for anything clinically useful? (See next question.)

Are mutations, the type of genomic rearrangements observed, expression profile etc. biomarkers for anything clinically useful?
Examples include e.g. the sample having an expression profile that predicts response or resistance to a specific drug, or a pattern of genomic changes (amplifications, LOH extending to telomeres etc.) that predicts response to a certain drug.

What was the disease course of other patients with mutations in gene X or, more generally, patients that resemble our case?
It is hard to define "resemble" in this case, and the answer depends on whether there is data on e.g. mutations and the clinical history of the same cancer cases.

Based on gene expression data, what kind of cancer or what kinds of patients does our case resemble?
The first question is relevant for identifying cancers of unknown primary origin, i.e. hard to classify leukaemias in our case. The second tries to find "similar" patients, the treatment history of which would hopefully give us clues on treatment options or prognosis.

The solution could again either be querying existing normalized data sets or getting e.g. an ExpressionSet from something like ArrayExpress, after which we compute similarity ourselves.
What are the molecular targets of the drugs that are effective in our drug screen?

For many this is known, but most kinase inhibitors inhibit a larger number of targets than the "known" ones. This would be useful when several drugs in the screen are active and we would like to create a combination. We might e.g. like to pick a combination of drugs that target several different pathways. Alternatively, if drugs that target the same protein are chosen, at least ensure that they have modes of action as dissimilar as possible (e.g. binding different pockets or having clearly different chemical structures) to reduce the likelihood of a resistance developing.