Deliverable D3.3

Project Title:	Building data bridges between biological and medical infrastructures in Europe
Project Acronym:	BioMedBridges
Grant agreement no.:	284209
	Research Infrastructures, FP7 Capacities Specific Programme; [INFRA-2011-2.3.2.] “Implementation of common solutions for a cluster of ESFRI infrastructures in the field of "Life sciences"
Deliverable title:	Provision and population of the ESFRI BMS Service Registry (eSR)
WP No.	3
Lead Beneficiary:	1: EMBL
WP Title	ESFRI BMS Standards Description and Harmonization
Contractual delivery date:	31 December 2014
Actual delivery date:	19 December 2014
WP leader:	Helen Parkinson (EBI) and Morris Swertz (UMCG)
Contributing partner(s):	1: EMBL, 10: TMF, 11: HMGU, 13: VUMC, 15: UCPH, 16: UH, 4: STFC, 7: TUM-MED

Authors: Jon Ison, Julie McMurry, Helen Parkinson, Nathalie Conte, Janneke van Denderen, Jeroen Belien, Søren Brunak, Kristoffer Rapacki, Philipp Gormanns, Juha Muilu, Murat Sariyar, Raffael Bild, Chris Morris, Martyn Winn, Gergely Sipos

Contents

Figures

Tables

1 Executive Summary

The development of a prototype service registry is the objective of BioMedBridges Deliverable 3.3, with contributions from BioMedBridges partners and in collaboration with ELIXIR. The Tools and Data Services Registry is designed to make it easier for researchers to find, compare, and use biomedical software to address a scientific question or research support task. For instance: “What are all of the Gene Ontology tools? Which of these is most highly cited?”. By returning relevant, structured results, the registry aims to complement search engines like Google: The registry user can specify exactly what they need, using various search and filter options, and get a tailored list of suitable resources. From sequencing to structures, imaging to indexing, the registry’s domain scope is very broad; it also encompasses webservices, web GUIs, desktop GUIs, and commandline tools. This broad scope ensures coverage of a substantial portion of the tools and data services of use to Research Infrastructures represented in BioMedBridges. Information about tools includes crucial provenance details, links to relevant publications and grants and key contact information. Consistent with the overall mission of WP3, the service registry implemented and extended existing standards, formats, and ontologies wherever possible. To achieve the objectives, it was necessary to engage with the software community to develop a sustainable, scaleable minimum metadata model and formal schema to describe software; the model was purpose-built to be lightweight and flat in order to facilitate adoption by other software registries that may be looking to provide or aggregate software metadata in the future. The registry data model, software, and content (metadata describing the tools) are fully open access and open source in order to further encourage re-use and community participation. In the following report we summarise technical progress and outcomes of our work to date.

Future plans comprise a program of activities lead and coordinated by ELIXIR-DK (tools node) and it is expected that this will ensure long-term sustainability of the registry. The code, schema and data in the prototype tools registry has all been made available to the ELIXIR-DK node and to the wider community via Github[1]. Further contacts have been made with external projects e.g. the NIH funded Gene Ontology and the EC funded RD-Connect project to identify tools that they recommend and to include these in the service registry. This deliverable report provides detailed information on the work performed with BioMedBridges resources.

2 Project objectives

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

No.	Objective	Yes	No
1	Provision and use of the ESFRI BMS common molecular identifiers (eCMI)		x
2	Identification, harmonization and integration of ESFRI BMS partner standards	x	
3	Provision of standards and harmonized elements in an accessible standards registry (eSTR)		x
4	Provision and population of the ESFRI BMS Service Registry (eSR)	x	

[1] https://github.com/EBISPOT/biomedbridges/tree/master/query-interface

3 Detailed report on the deliverable

3.1 Registry overview

We have delivered four key components of the Tools and Data Services Registry

- **Software description model**, defining the information the registry must provide about each tool, service *etc.* to allow a user to find, understand, compare and select software as well as access and use it. This model has been transformed into a prototype schema.
- **Content** *i.e.* the actual metadata descriptions of the tools (with the structure following the model above)
- **Registration mechanisms** to enable content to be contributed by members of the community via simple spreadsheets or direct contributions by way of XML
- **Query interface** for viewing and searching for tools

The registry content is open access and can freely be repurposed; accordingly, members of the community may create their own interfaces tailored to their specific needs. For instance, RD Connect and the IMI EMIF projects have used the Spreadsheet mechanism within their project to manage tools and service information.

3.2 User-centered design process

From the outset and throughout the duration of implementation, an agile, user-centered design process was adopted to ensure the needs and desires of end users were satisfied. Through this process, we were able to identify the core purpose of the registry, founding principles (Appendix 1.1), use-cases, target audience, and technical focus. The agile approach did not only apply to implementation of the registry software, but to all aspects of the project from strategy and data gathering, through to the future upkeep strategy. The user-centered design process involved a combination of surveys (2), interviews (6), workshops and other events (12). The process is further described in

Appendix 2. From these mechanisms, all the identified needs and desires were also logged in our JIRA tracker so that features could be reviewed and prioritized on an ongoing basis. Much of the feedback has already been addressed in the current prototype (beta) version of the registry which was jointly developed by two BioMedBridges partners: EBI (ELIXIR UK) and the Danish Technical University (ELIXIR DK). DTU will further develop the registry software during the project period; thereafter they will sustain it in their capacity as ELIXIR DK and by engagement with the community to ensure content is up to date.

3.3 Software description model

A prototype information model was defined during a series of BioMedBridges workshops and meetings, in conjunction with a user survey that rated the value of each software attribute and prompted for missing attributes. The prototype model[2] was the basis for the interface and content that correspond to the BioMedBridges prototype registry[3]. Following iterative refinement, the mature model[4] is now formally under the aegis of ELIXIR DK who have used it to develop the next generation of interfaces for query and registration[5]. The mature data model is a structured list of attributes providing administrative, scientific and technical details about a given software. Whereas the prototype schema was flat, mature schema is organised into 10 blocks:

- summary *e.g.* name, short description
- operations *e.g.* operations, inputs and outputs
- usage *e.g.* interfaces, platforms
- documentation, *e.g.* link to REST API documentation or WSDL file
- support *e.g.* helpdesk, contact person
- restrictions *e.g.* license, terms of use
- credits, *e.g.* developer, grants
- literature, *e.g.* primary citation, relevant publications

---

[2] http://wwwdev.ebi.ac.uk/fgpt/toolui/schema.html
[3] http://wwwdev.ebi.ac.uk/fgpt/toolsui/
[44] http://bioregistry.cbs.dtu.dk/schema.html
[5] http://bioregistry.cbs.dtu.dk

- registration *e.g.* registrant name, last update
- see also *e.g.* URL of source registry or parent collection

Aside from a short textual description of the software, attribute values are free-text tags, ontology terms or URLs (for example links to terms of use). Each attribute is defined as mandatory, recommended or optional. The model core is minimal, mandating the minimum necessary and sufficient information to support the use cases.

3.3.1 XML schema / exchange format

For purposes of sharing software information to and from the prototype BioMedBridges registry[6], we provided a corresponding XML schema (XSD)/exchange format[7]. The prototype schema allows validation on the basis of completeness and correctness (legal values)[8]. In the time since this early modelling, we have done extensive further development of the schema. Comprehensive details for the mature schema including an example XML file are available on-line[9]. In accordance with the sustainability plan, this additional development is under the aegis of ELIXIR DK; future contributions to the registry are advised to adhere to the mature schema rather than the BioMedBridges prototype.

It is inevitable that many providers, integrators, and cataloguers will continue to use their preferred models, methods and formats for software descriptions. Therefore, our schema and especially the mandatory core is as simple as possible whilst retaining necessary utility. It is also technically extensible, and for all these reasons should be easily compatible to support future integration scenarios.

In the mature schema (now under ELIXIR DK), each software has at least one function (Figure 2), with one or more inputs and outputs (Figure 3), each of a specified type and supported format(s). Function, data type and format are

---

[6] http://wwwdev.ebi.ac.uk/fgpt/toolsui/
[7] http://wwwdev.ebi.ac.uk/fgpt/toolui/schema.html
[8] this is achievable by XSD/XML validation
[9] http://bioregistry.cbs.dtu.dk/schema.html

described in terms from the EDAM controlled vocabulary[10]. There is also a schema element (`FunctionDescription`) for a more verbose, free-text description of the function. An optional element (`FunctionHandle`) provides a programmatic hook to the implementation of the function, such as the name of an operation of a SOAP service, the URL scheme for a REST service endpoint, or a command-line flag. A handle on the data (`DataHandle`), provides a programmatic identifier of the input or output, such as a URL parameter name or a command-line flag. Each registry entry may belong to one or more collections, for example, a named software suite, database or REST API. Other contextual details, such as interfaces, are also provided to allow smart navigation within the Query Interface.

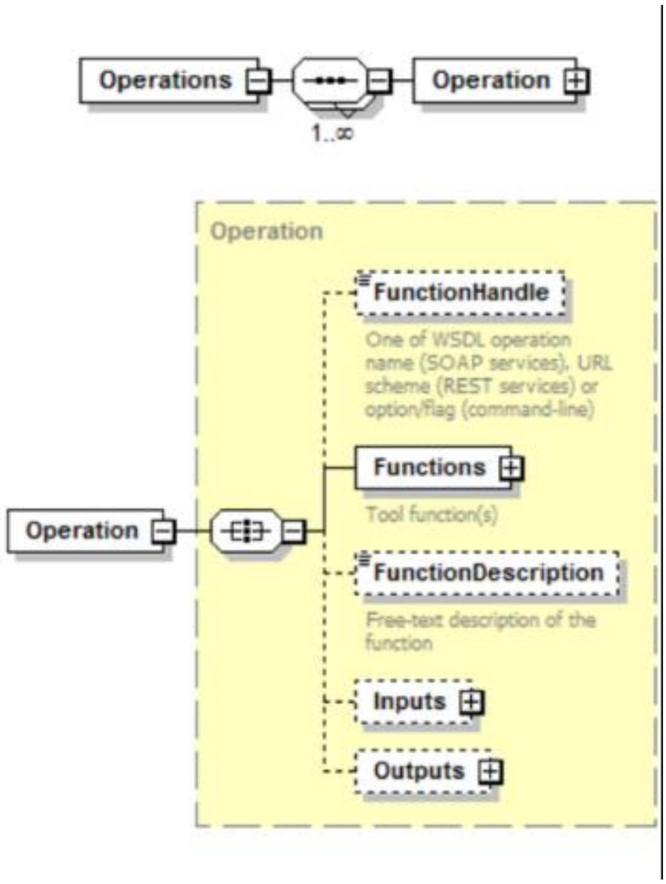

Figure 1 Model of tool function

[10] http://bioportal.bioontology.org/ontologies/EDAM/?p=classes

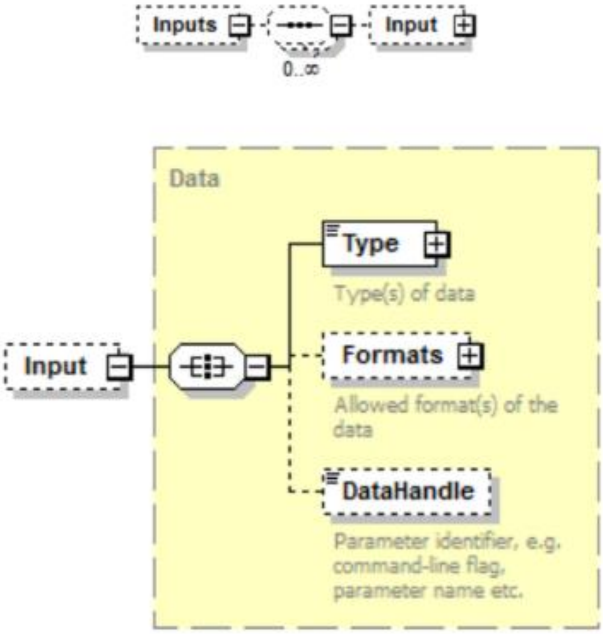

Figure 2 Model of tool I/O

3.4 Registry content

3.4.1 Overview of content and export

The BioMedBridges prototype tools and data services registry currently includes over 1,900 entries comprising over 22,000 annotations, including manual entries by BioMedBridges partners and content from institutional partners (EMBL-EBI, CBS-DTU), the EMBOSS sequence analysis suite and the GO Tools collection, all of which are imported into the registry. A breakdown of the content by source is below (

Table 1).

Table 1 Summary of registry data (by source)

Source	Type	Entries	Annotations	Import mechanism
BioMedBridges partners / workshops	Various types	318	3227	Google spreadsheet
EMBL-EBI	Tools available as Web applications	225	4280	Google spreadsheet Authorized access[11]
EMBL-EBI	Databases	58	734	Google spreadsheet Authorized access
EMBL-EBI	Web services & downloadable packages	80	1201	Google spreadsheet Authorized access
EMBL-EBI	Other software	28	2500	Google spreadsheet Authorized access
RD Connect[12]	Various types	24	379	Google spreadsheet
EMIF[13]	Various types	16	206	Google spreadsheet
CBS-DTU	Tools available as downloadable packages, Web applications and Web services)	60	545	XML
DRCAT	Databases	675	2860	XML
EMBOSS	Command-line tools	397	8193	XML
GO Tools	Various types	130	1092	XML

The entire contents of the registry are available to be downloaded from http://wwwdev.ebi.ac.uk/fgpt/toolsui/data/tools.xml.

[11] Authorized access: all public information from EBI's registration spreadsheets is viewable/downloadable within the BioMedBridges prototype Query Interface; however these spreadsheets are open authorised users only. External view access can be requested as needed at the links provided in the above table.
[12] Rare Disease Connect http://rd-connect.eu/
[13] European Medical Information Framework http://www.emif.eu/emif/about-emif

3.4.2 Content registration, storage, and transformation

3.4.2.1 Registration via Google Spreadsheets

Seven of the eleven datasets were imported via Google spreadsheets. The spreadsheet approach is designed to be convenient for anyone wishing to maintain a software collection in this format and especially for use in curation workshops where collaborative editing of shared document on the Web is required. It has proved to be a successful means of quickly acquiring and curating data. Each spreadsheet includes a single row per tool, and a single column per attribute. A java-based Google Spreadsheet Parser, was developed in order to programmatically fetch, validate, combine, and parse the data directly from the spreadsheets into the XML format required by the Query Interface.

The BioMedBridges Google Spreadsheet parser has the following features:

- *Optional attributes* - not every software attribute (column) must be included, providing flexibility to those who wish to maintain only a subset of fields (as in the case of EMBL-EBI tools, and the expectation generally)
- *Customisable column headers* - allows user-supplied column headers to be mapped to software attributes, to support alternative semantic schemes
- *Configurable XML binding* - allows for evolution of the output XML format
- *Flexible annotation* - allows free-text tagging of any software attributes which in the schema is defined as requiring an ontology term. This is a practical requirement, for example to allow annotation with terms which do not yet exist in the ontology, or to allow mapping of tags to ontology terms *post hoc*.
- *Semantic validation* - of user-supplied tags to ontology terms. Tags for attributes which require an ontology term, such as Function, Input *etc.* are checked against an enumeration of terms from the corresponding branches of EDAM and SWO and a report generated. For example, a tag for a tool input will be checked against the EDAM “data” branch with a report on matches to EDAM labels and synonyms.

The codebase for the Google Spreadsheet Parser is available on GitHub[14].

3.4.2.2 Registration via Standalone XML

Four of the eleven datasets were imported via standalone XML. This approach is more appropriate when the data already exists in a curated external database and can be easily transformed. The prototype schema can be used to validate the contributed files. See future work section for developments to the schema and registration interface.

3.4.3 Content annotation with ontologies

The software description model includes terms from two ontologies (Figure 4): EDAM for topic, function, data types (including identifiers) and formats specific for the life sciences, and the Software Ontology (SWO)[15] for general software attributes. In order to provide semantic coverage of the BioMedBridges tools, dozens of new terms have been added to EDAM and SWO including a major revision and extension of the EDAM Topic (scientific domain) branch. This resulted in new versions of EDAM (1.3) and SWO (0.4).

The ontologies may be browsed within BioPortal:

EDAM	http://bioportal.bioontology.org/ontologies/EDAM?p=classes
SWO[16]	http://bioportal.bioontology.org/ontologies/SWO?p=classes

[14] https://github.com/EBISPOT/biomedbridges/tree/master/GoogleSpreadsheetParser
[15] http://bioportal.bioontology.org/ontologies/SWO
[16] EDAM is imported into SWO and is therefore visible in the BioPortal view of SWO

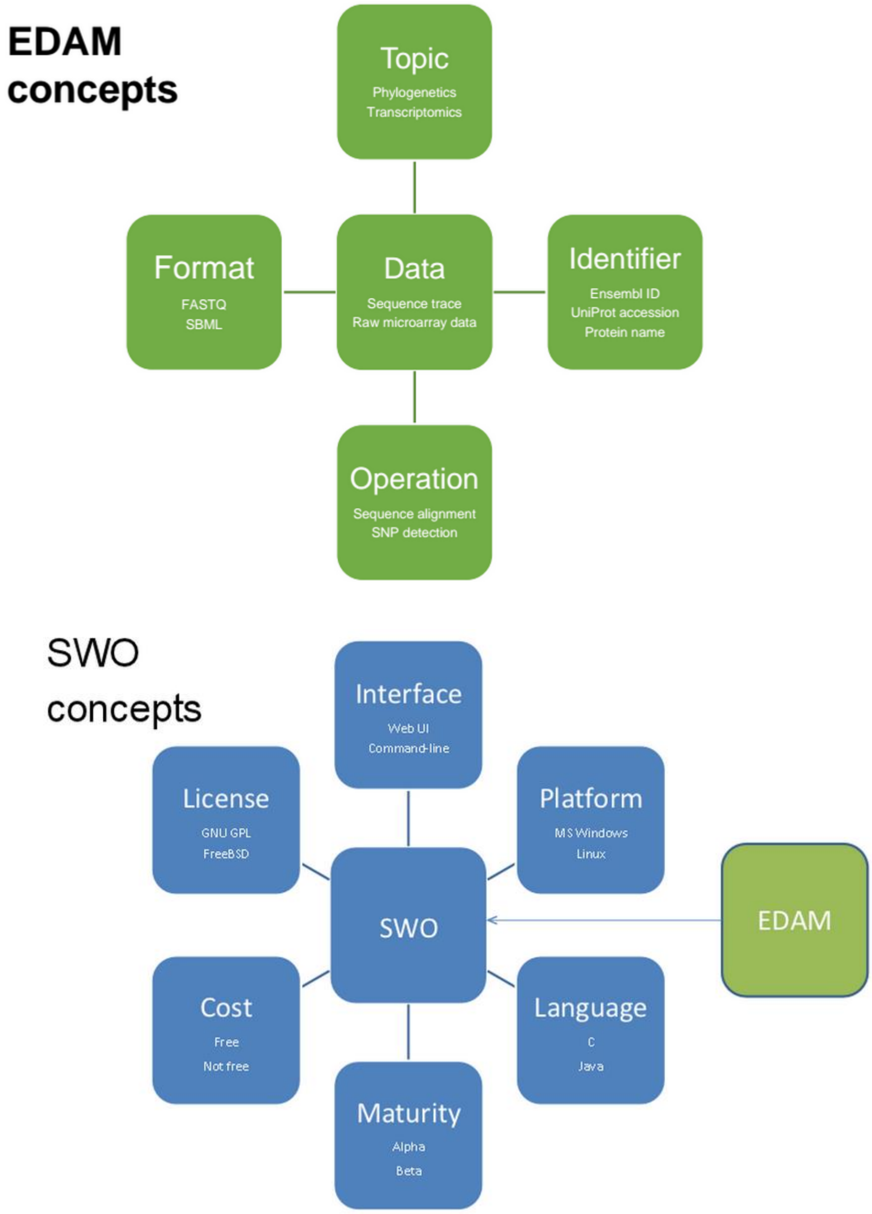

Figure 3 The upper level classes of the EDAM and SWO ontologies[17]

3.5 Sustainability

3.5.1 Content upkeep strategy

Registry content maintenance is based on a federated curation model which seeks to minimise the content that must be centrally maintained: Several

---

[17] SWO imports EDAM as a module

major providers and cataloguers have agreed to make their own software information available in a format that can be shared and imported into the registry; so far these include institutes such as EBI, registries such as BioCatalogue and projects such as BioConductor, SeqWIKI, and CCP4. Upkeep will entail basic housekeeping duties and specific activities to build and support the community to curate the registry. The federated curation model and registration mechanisms are summarised below. Resources has been invested in first phase population with BioMedBridges personnel both supervising collection of data and curating the data collected.

3.5.2 Federated curation model

The federated curation model is predicated upon the principles of *responsibility*, *sustainability* and *collaborativeness* (Appendix 1.1) and reflects the huge scale and diversity of the software landscape, confounded by the fact that there are very many registries, catalogues and lists of software out there. Key providers will assume responsibility for their software descriptions. The registry then merely imports a snapshot of the available information from these providers.

3.6 Prototype Query Interface

A prototype Query Interface (http://wwwdev.ebi.ac.uk/fgpt/toolsui/) was developed to allow users to browse, search and query information through following a spreadsheet-style grid; it supports free-text search and filtering over all available information fields.

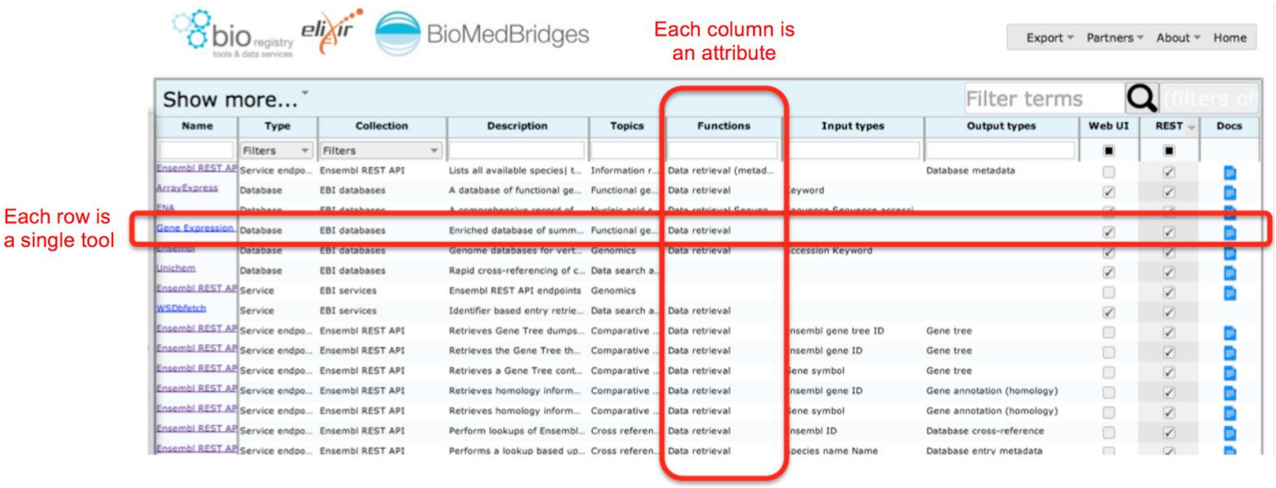

Figure 4 The Query Interface

Each row in the grid represents a single software entity, *e.g.* a tool or database, whereas each column represents a single software attribute. The Figure shows the software short description ('Description' column), the broad domain the tool belongs to ('Topic') and the specific function it performs ('Functions'). The values given in these columns are terms from EDAM.

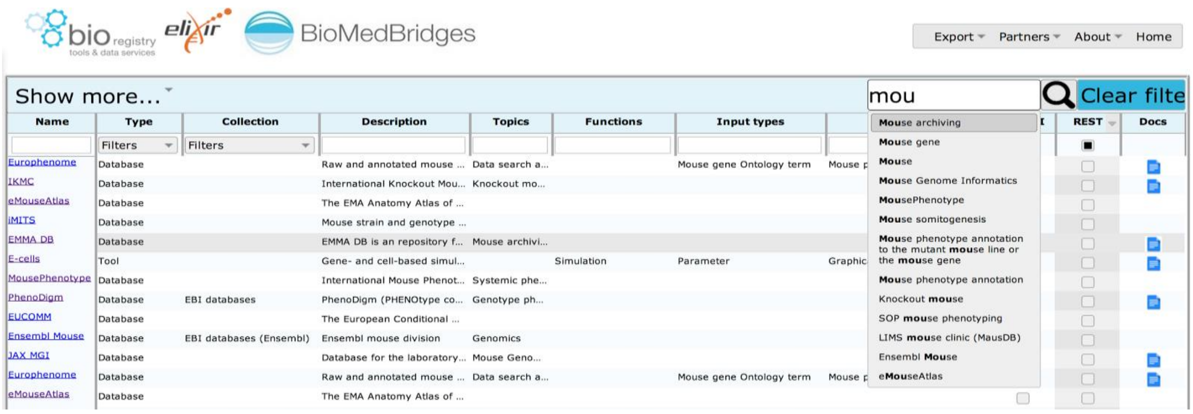

Figure 5 Search box with autosuggest

As you type into this box, you get a drop down with suggestions of search terms, and corresponding results which change in real-time as you type. For example searching for returns entries with the string "mouse" in the description or some other field of information. Additional advanced features of the query interface are outlined in Appendix 4.

3.7.1 Query Interface Use Cases

The Query Interface was designed to answer the following questions determined by user experience analysis:

- I have this task in mind, what tools are there for it ?
- I have this data, what can I do with it?
- I need to generate some data, what tools are there ?
- I need a tool that reads or writes data in a specific format?
- I need mass access for a specific type of data, what databases are available?
- What resources are available in a general area, e.g. proteomics?
- What resources arose from a particular grant?
- What are the outputs of a particular research institute or infrastructure?

The benefits to the community include, amongst others:

- Save time (money) finding the right tools for the job
- Understand /compare tools
- Coordinate developments / identify collaborations
- Avoid wasteful duplication of coding efforts
- Developers, institutions, infrastructures *etc.* get credit and visibility for tools
- Tool providers *etc.* get traffic to their websites

For more specific examples of scientific questions that the registry can answer, see Appendix A.3.

3.8 Future Work

Mission of this project is to make it easier for users to find, compare and select bioinformatics tools. Consistent with this mission, we aim to do the following:

3.8.1 Enable end-users to contribute tool metadata

Currently only system administrators can upload XML into the Tools and Data Services Registry or register a new spreadsheet source. In the coming year, ELIXIR-DK will provide a data registration interface so that tool providers can upload their own tool collection metadata and/or can edit it manually.

3.8.2 Content: Refine annotations, transform to be compliant to new schema

While more than 1900 tools have been registered, annotations vary in richness. Consistent with the federated curation model, annotations within the BioMedBridges datasets will continue to be refined. The datasets will also be transformed to be compliant with the mature (ELIXIR-DK) schema.

3.8.4 Evaluate and Collect Metrics

To focus future development we plan to implement extensive user evaluation into the next generation of user interface (ELIXIR-DK). Workshops will help to add to the contents as well as to pinpoint shortcomings in the user interface design. In addition to using Google Analytics (page impressions) we will log and analyse terms/phrases used in queries.

This enables us to focus the future collection of new metamodel information as well as learn what components are particularly popular and would be good to develop further. Finally, we have implemented a feedback component to enable users to easily report issues with the registry.

3.8.5 Integrate with other registries

Finally, we aim to further the integration of the tools and data services registry with the other registries in the standards domain (Eg the Meta Models and Mapping Registry, and Identifiers.org/EDAM). We will add bi-directional cross links with identifiers.org to enable users to find how the models/formats map onto identifiable records in (public) databases and what models/formats are used within the various services and tools. Moreover, in collaboration with biosharing.org we want to add more references on the usage of the tools by adding URIs refering to instances where the model is used.

By collaboration with ELIXIR partners, ELIXIR-DK in particular, we expect that the work we have completed will be sustainable and maintained going forward.

4 Delivery and schedule

The delivery is delayed: ☐ Yes ☑ No

5 Adjustments made

No adjustments were made to the deliverable.

6 Background information

This deliverable relates to WP 3; background information on this WP as originally indicated in the description of work (DoW) is included below.

WP 3 Title: ESFRI BMS Standards Description and Harmonization

Lead: Helen Parkinson (EMBL-EBI, Morris Swertz (UMCG)

Participants: EMBL, KI, STFC, UDUS, TUM-MED, ErasmusMC, TMF, HMGU, VU-VUMC, UCPH, UH, UMCG, CIRMMP

Standardization is necessary to ensure infrastructures can work together (syntactic interoperability: data models, data formats, API's, services descriptions, registration and discovery of services), understand each other data (semantic interoperability: ontologies, vocabularies, coding systems, common identifiers), have analysis and supporting tools that complement each other and can be combined in a pipeline (process interoperability) and allow multiple data sets from different origins (including public resources) to be analysed together.

This work package (WP) requires close collaboration with domain experts, research infrastructures, WP4 which will provide implementation based on standardization deliverables described here, and WP5 which will address security issues and use case work packages 6-10. In order to work efficiently a nominated individual from each ESFRI BMS expert area will be responsible

both for tasks in this WP, registration of standards, representation of, and correspondence with, relevant domain specific external standardization parties and to represent their community requirements in this WP. WP3 partners are also represented in the use case work packages and will ensure their requirements are supported here. This WP involves the majority of partners, and exchange of information, registry of services and meta mapping activities will require a diverse set of personnel. The design of this WP therefore includes an allowance for exchange of personnel between this WP and others to facilitate the implementation of deliverables in other WPs and to support interaction with external experts at meetings and workshops where necessary. This will ensure that relevant experts have the opportunity to actively solve problems by working closely with individuals from work packages to which they have not been assigned. We have also allowed developer time for the creation of training materials and delivery of training at BioMedBridges workshops, as described in WP12.

Work package number	WP3	**Start date or starting event:**	month 1
Work package title	ESFRI BMS Standards Description and Harmonization		
Activity Type	RTD		

Participant	1: EMBL	3: KI	4: STFC	5: UDUS	7: TUM-MED	9: ErasmusMC	10: TMF	11: HMGU	22: VU-VUMC	15: UCPH	16: UH	19: UMCG	20: CIRMMP
Person months	42	21	6	28	4	5	16	30	16	8	11	32	14

Objectives

Addition of scientific value and support for the integration of data between the ESFRI BMS domains by catalogue, review, modification, harmonization, registration and implementation of existing identifier, content, syntactic and semantic standards across the ESFRI BMS projects to support data exchange, integration and infrastructure development. 1. Provision and use of the ESFRI BMS common molecular identifiers (eCMI) 2. Identification, harmonization and integration of ESFRI BMS partner standards 3. Provision of standards and harmonized elements in an accessible standards registry (eSTR) 4. Provision and population of the ESFRI BMS Service Registry (eSR)
Description of work and role of participants The standardization task is large as ESFRI BMS projects have been active in this area evaluating intra-domain standards, bottlenecks and solutions and there are numerous external standards efforts corresponding to content, data format, semantic and identifier standardization in this domain in which many project partners are involved. Examples include the gene ontology (GO) as an example of a semantic standard, DICOM as an imaging format standard, MIMPP as a content standard from EUROPHENOME, the LCF/MTZ file format, and the CCPN data model for macromolecular NMR. WP will address the following tasks to provide focus: 1. Common identifiers (Task Lead ELIXIR) The provision and use of common identifiers to determine unambiguous molecular identity 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. The ESFRI BMS project partners will work together to determine a ‘Molecular Dictionary’ of identifier types and their attributes for use in this project which will constitute best practice for cross domain integration. Where no authoritative identifier standard exists, we will

work with the respective community to determine one to support the activities of WP4 and use cases. Relevant identifiers include those for samples (Task 2), small molecules, macromolecular assemblies, genes, drugs and proteins especially where these relate to clinical scenarios.

2. Sample meta data standards (Task Leads BBMRI)

The ability to identify samples and describe their attributes, so data relating to them can be integrated and analysed is common to all ESFRI BMS domains. Content standards which determine exist for given experimental scenarios which data should be collected e.g. age, sex, phenotype, disease state, sampling time, processing state, etc. These are typically determined based on requirements for analysis, data sharing needs and regulations within a research or technology based domain. For example, the MIAME standard determines which information should be stored about a gene expression experiment performed on a microarray. This is not necessarily consistent with core information about the same sample stored in a BioBank which may include sample processing state, disease and tissue, a sample used to determine a protein structure, or a live animal sampled from the ocean. Where processing states, provenance, storage conditions, or other experimental context are important for a domain e.g. INSTRUCT or for re-use of data relating to samples across domains, these will also be explored with respect to the use cases. The clinical data community have specific requirements relating to integration of Electronic Health Records (EHR), use of clinical terminologies such as SNOMED-CT, description of medical imaging procedures and provision of molecular data in clinical context with appropriate quality control data and translation across these domains is relevant to this task, Task 4 and WP10. Standards in use within the ESFRI BMS projects for data content and semantics will be documented in a public interactive matrix consisting of project, standard and individual elements of standards. Comparable elements across standards will be identified by a harmonization and mapping process across partners. For example BBMRI has produced a lexicon which defines important concepts for the bio-banking domain and EATRIS has analysed standards relating to inter and intra operability between organisations. Standards in use by partners relating to samples will be meta-mapped;

common elements e.g. from BBMRI will be cross referenced to relevant concepts from ELIXIR, ECRIN and EATRIS. Where standards are in development e.g. from 2008 roadmap ESFRI BMS projects these will be added and harmonized once they are determined to be stable and valid within a domain, e.g. imaging standards are under development by EuroBioImaging. We do not expect all standards to be fully interoperable and the process of meta-mapping and presentation of these data in an interactive and updated form will inform partners and focus use cases. We will pay specific attention to widely adopted standards, and supporting integration rather than development of standards de novo.

3. Service registration and annotation (Task Lead ELIXIR)

The description of where data and services exist, and by what mechanism these are accessible is key to integrating and exchanging data and has been identified by ELIXIR, EATRIS and others as a blocker to integration especially across domains. Therefore we will develop the Meta-Services Registry comprising tools and terminology for annotation of services (eSR) to catalogue services across partners, domains allowing partners to self register their own and others services. This will build on previous work in the Bioinformatics domain (EMBRACE, BioCatalogue) and will be extended this with the 2008 roadmap ESFRI BMS partners and throughout the grant as services appear and are used. This will promote the use of domain specific services across partners and also internationally.

4. Semantic standards – ontologies and annotation (Task Lead ELIXIR)

Content standards define what data about a sample in a context or domain. However the meaning of data can be made explicit only by the use of defined terminologies. The use, standardization and mapping of terminologies across domain and species will be explored in the context of use case Work Packages 7 and 10. WP7 explores the semantic integration between mouse models of disease, phenotype and WP10 explores integration of sample data of different types. In order to make these tasks feasible prioritized dataset(s) will be identified with WP7/10 by means of integration criteria which will be developed jointly with these work packages. For example – availability of data

in the public domain and /or focus on a key disease type which is well represented in the terminologies to be integrated and available datasets.

Appendix 1: Summary of outcomes of user-centered design

1 Founding principles

The design and long-term success of the registry is predicated upon 16 principles that were identified during the user-centered design process (Section 2):

Purposefulness - The registry has two primary purposes; to help people discover software and use it. By “discovery” we mean to find, understand, compare and select. However, the registry does not aim to support “interoperability” of tools; the reason for this is that doing so would require careful and extensive modeling of dependencies and such modeling is out of scope.

Use-case driven - The registry should provide an “on-ramp” for the working bioinformatician and be a staging post for websites where tools and services may be researched in greater detail, used or downloaded. That said, the registry should be useful generally and support other roles identified by our market research, for example a scientist browsing the available offerings, a manager surveying the output of a grant, institute or infrastructure, a developer reviewing the state of art, and so on.

Generality - Many types of software and software interfaces will remain in use for the foreseeable future: downloadable analytical software with a command-line interface or desktop GUI, remote-access REST and SOAP-based Web services and Web UIs, covering all the bioscience domains and providing access to a vast array of biological databases. The registry - especially as a tool inventory for BioMedBridges - must cover this diversity, but focus on the prevalent types, notably Web UIs and data services.

Detailed - Considering the broad scope, the registry must concentrate upon the principal and common software features that are necessary to support the use-cases, without descending into excessively fine-grained details. To fulfill it's purpose however and especially to allow comparison, the registry must provide information in sufficient detail and structure, which cannot conveniently be obtained from a Google search or cursory inspection of a provider's website.

Practicality - Software manifests in a dizzying complexity of deployments, interfaces, wrappers, suites, libraries, packages *etc.* The registry must be practical and therefore expose software in a useful way, i.e. in terms of readily understood, consistent functional units. It must shield people from superfluous details whilst providing enough information to place a tool or service in context of the deployments, interfaces and collections in which it appears.

Comprehensiveness - In order to fulfil its purpose the registry *must* be comprehensive, include all prevalent databases, tools and services and obtain the widest coverage of the scientific domains as resources allow. A given resource must be fairly comprehensively detailed, as described above. This effort has to be distributed and involve the major tool providers, integrators and cataloguers, as well as individual developers and scientists.

Quality - The registry will face fierce competition. Success depends on goodwill engendered by solid Quality Assurance: it is of utmost importance that the registry stay relevant with continuous upkeep. The information must not be wrong or go stale and therefore must be validated, for example to ensure there are no broken links, invalid annotations or missing fields. Again, this can only be accomplished by the creation and support of a community and distributing the effort among its members.

Accessibility - Registry users will have high expectations and demand key information is put immediately at their fingertips, in an interface that is very simple and intuitive to use. The registry query interface must therefore be immediately accessible, highly streamlined and practical. The search and browse/query functions must be powerful, convenient to perform and quickly yield concise, relevant and meaningful results.

Consistency - The query interface must have a consistent look and feel and also return consistent and therefore comparable information. To that end we envisage all registry entries to be annotated with controlled vocabularies: EDAM for annotation of topic, function, data types and formats specific for the life sciences area, SWO for general software attributes, and others, whenever appropriate. This will allow for precise searching and consistent search results, taking the impact of the registry beyond merely finding tools: comparison, evaluation and interoperability will be greatly facilitated.

Compatibility - There are a multitude of relevant technologies and developments, including *ad hoc* catalogues and lists of tools, formal registries, service description models and exchange formats, controlled vocabularies *etc.* The registry should, where possible and desirable, use, build upon and operate harmoniously with all of it. For example, we expect a variety of models, methods and formats for describing and documenting software will remain in use, and aim therefore to be compatible with these. Similarly for databases, our information model must for example be compatible with the core attributes of biological databases as defined by BioDBCore[18].

Responsibility - We believe that tool and service data, as part of software documentation, is rightly the “property” of the developer or provider, or at least a responsible cataloguer. In so far as these enterprises are publicly funded, there is a responsibility to share information. We are therefore promoting this notion and encourage providers *etc.* to publish such data alongside their tools and services. In this spirit, we share the registry data and code, and open their development to anyone. Further, where software data are registered with us that are in scope of some other catalogue, we will forward the data and collaborate to ensure the catalogue incorporates it.

Sustainability - The registry effort must be sustainable in the long term with limited resources. Digital curation is, however, time consuming and costly. To minimise future maintenance costs, we are promoting a federated curation model which envisions key providers, integrators and cataloguers as the “primary citizens” that maintain and share information about the resources within their scope: curation responsibilities are federated. Rather than aspiring

---

[18] http://biodbcore.org/

to maintain software information centrally, the registry serves as a collation or "snapshot" of the available information distributed on the Web. For example, the registry will import Web service descriptions from BioCatalogue[19] and those of GRID-enabled software from EGI AppDB[20].

Collaborativeness - Success is predicated ultimately on the good will of enthusiastic individuals, working within the key institutes and infrastructures, to assume responsibility for the resources that they provide or care about. The primary challenge in the long term is therefore a social one: building and supporting a community to maintain a federated curation of information. We cannot do everything and so must encourage others to do "heavy lifting" in their specialised areas, for example BioCatalogue to process machine-readable service descriptions (WADL, WSDL, Swagger *etc.*)

Focus - The registry must remain focused on community requirements as gauged by workshops, mailing lists and other activities. "Mission creep", even into important areas, must be avoided given the limited resources. For example, the registry will not, in the first instance, provide a tools forum with user ratings and comments, a community wiki, or become a software repository. Similarly, while service versioning and monitoring are important to the user the registry will not, in the first instance, implement a fully-fledged service monitoring framework, or anything else that is properly the provider's responsibility.

Extensibility - The technical components, for example the registry query interface, underlying schema *etc.* should be extensible and adaptable by others for their own purposes. More broadly, the technical outputs of this initiative are, in principle, applicable to non-software entities. One could envisage for example an adaptation of the registry software to federation of information about community events or job announcements. Extensibility was therefore borne in mind in the design process.

---

[19] https://www.biocatalogue.org/
[20] https://appdb.egi.eu/

Inclusivity - The requirement for a tool and data service registry is universal. Duplicated, wasteful efforts should be avoided. The registry - as a key component of the bioinformatics and biomedical infrastructure - therefore embraces all relevant individuals, projects, institutions and infrastructures from the outset, from individual developers to major projects and providers, regardless of whether they are officially associated with the research infrastructures within BioMedBridges.

2 Software attributes

Table A 1 Attributes in the prototype software description model

attributeName	description	representation	required
RegistrationURL	Stable unique URL in the biotoolsregistry.net domain identifying a registration	URL in the biotoolsregistry.net domain	No (calculated*)
RegistrantName	Name of person who registered the software	Text	Optional
RegistrantEmail	Email address of person who registered the software	Text	Recommended
RegistrationDate	Date the collection was first registered	Date	No (calculated)
UpdateDate	Date the collection was last updated	Date	No (calculated)
DescriptionFileURL	URL locating the software description file for the registration (if available)	URL	No (calculated*)
EntryURL	Stable unique accession (URL) identifying the software in the registry	URL in the biotoolsregistry.net domain	No (calculated*)
Name	The canonical name of the software (tool, service, package etc.)	Text	Mandatory
Homepage	Software homepage (URL)	URL	Mandatory
Type	Basic software type: one of "Framework", "Package", "Database", "Tool", "Service" or "Other"	SWO concept (term + URI)	Mandatory
Description	Short textual description of the software	Text	Mandatory
Topics	General domain the software serves	EDAM Topic (term + URI)	Mandatory

attributeName	description	representation	required
Tags	Miscellaneous semantic annotations not covered by EDAM Topics	Ontology concept (term + URI)	Optional
Cost	Cost of purchase: one of "free" or "not free"	SWO concept (term + URI)	Optional
Version	Version of the software e.g. version number	Text	Recommended
Maturity	Software maturity: one of "alpha", "beta" or "production"	SWO concept (term + URI)	Optional
Functions	Name(s) of the software function(s)	EDAM Operation (term + URI)	Recommended
FunctionDescription	Concise textual description of the function(s)	Text	Optional
FunctionHandle	One of WSDL operation name (SOAP services), URL scheme (REST services) or option/flag (command-line)	Text	Recommended
InputTypes	Type of data (primary input)	EDAM Data (term + URI)	Recommended
InputFormats	Allowed format(s) of the data (primary input)	XSD primitive type or EDAM Fomat (term + URI)	Recommended
InputHandle	Parameter identifier, e.g. command-line flag, parameter name etc. (primary input)	Free text	Optional
OutputTypes	Type of data (primary output)	EDAM Data (term + URI)	Recommended
OutputFormats	Allowed format(s) of the data (primary output)	XSD primitive type or EDAM Fomat (term + URI)	Recommended
OutputHandle	Parameter identifier, e.g. command-line flag, parameter name etc. (primary output)	Free text	Optional
Interfaces	Software interface type: "REST API", "SOAP API", "Web UI", "Command-line" or "Desktop GUI"	SWO concept (term + URI)	Mandatory
License	Software or data usage license	SWO concept (term + URI)	Recommended
Platforms	Platforms (OS and chipset combination) supported by a downloadable software package	SWO concept (term + URI)	Optional*
Languages	Languages (for APIs etc.) or technologies (for Web applications, applets etc.)	SWO concept (term + URI)	Optional
Download	Software or data downloads page (URL)	URL	Optional
WSDL	Location of WSDL (URL)	URL	Optional*

attributeName	description	representation	required
Availability	Whether a Web service is available for use	"Yes" or "No"	No (calculated)
Downtime	The percentage of time a Web service has been unavailable	%	No (calculated)
DocsHome	Software documentation main page (URL)	URL	Recommended
TermsOfUse	Link to license text or terms of use (URL)	URL	Recommended
DocsCommandLine	Command-line documentation (URL)	URL	Optional
DocsREST	REST service documentation (URL)	URL	Optional
DocsSOAP	SOAP service documentation (URL)	URL	Optional
DocsSPARQL	SPARQL service documentation (URL)	URL	Optional
ContactPage	URL of helpdesk or page with general contact details	URL	Optional
Helpdesk	Email of helpdesk	Email address	Recommended
ContactName	Name of contact person	Text	Optional
ContactEmail	Email address of contact person	Email address	Recommended
ContactTel	Telephone number of contact person	Telephone number	Optional
Developer	Name of person that developed the software	Text	Optional
DeveloperInterface	Name of person that developed the software interface	Text	Optional
Contributors	Name of person contributing to the software	Text	Optional
Institutions	Name of the institution that developed or provide the software	OpenAIRE term?	Optional
Infrastrutures	Research infrastructure in which the tool was developed or provided	OpenAIRE term?	Optional
Funding	Details of grant funding supporting the software	Grant number	Optional
WorkPackages	Work package in which the software was developed	Text	Optional
PublicationsPrimary	PMCID, PMID or DOI of the primary publication	Text	Recommended
PublicationsOther	PMCID, PMID or DOI of the primary publication	Text	Recommended

attributeName	description	representation	required
CitationOtherID	PMCID, PMID or DOI of other relevant publications	PMCID, PMID or DOI	Optional
citationURL	Citation instructions (URL)	URL	Optional
EntryURL	Link to an entry for the software in the registry from which it was imported.	URL	Optional
Collection	Registry URL of a collection that the software has been developed or maintained as part of, e.g.a suite, toolkit, library, framework, project, portal, workbench etc.	URL in the biotoolsregistry.net domain	Recommended
providesInterfaceTo	Registry URL of a collection that the software has been developed or maintained as part of, e.g.a suite, toolkit, library, framework, project, portal, workbench etc.	URL in the biotoolsregistry.net domain	Optional
uses	Registry URL of another tool (in the registry) that this tool uses	URL in the biotoolsregistry.net domain	Optional

3 Query Interface usability testing tasks

A selection of tasks were created for use in one-on-one software usability tests, to help test the registry Query Interface:

Task 1 - You have a large set of molecular sequences you need to compare in order to identify conserved sites, and have been asked to generate a multiple sequence alignment from them. Can you identify appropriate tools?

Task 2 - You must also compare all pairs of sequences from this set in turn. Again, what tools are available for this?

Task 3 - A collaborator has requested your sequences and alignments but wants them in a different format to what you have. Can you find tools to do this reformatting?

Task 4 - Having completed your basic sequence comparison, you've been asked to perform a range of phylogenetic analyses on your sequences. Can you find appropriate tools, but ideally a package or suite, that does this?

Task 5 - You've just got a new job at EMBL-EBI and want to bring yourself up to speed with their available offerings. Can you find all the tools credited to the EBI? How about all the databases?

Task 5a - What other types of software does EMBL-EBI provide?

Task 6 - You need to do some text mining using EMBL-EBI tools; can you find contact details of the person to speak to about it?

Task 7 - You need to retrieve the gene trees corresponding to a list of Ensembl gene IDs, but you have to do this programmatically (i.e. using some API). Can you find an appropriate service?

Task 8 - When working with your list of Ensembl IDs, you realise they are from an older version of Ensembl. Can you find a tool that will read these IDs and convert them to their current equivalents?

Task 9 - A big project is starting around using the Gene Ontology. Can you bring up a list of all the software in the GO tools collection? What types of software are available ?

Task 10 - What is the most cited software in GO Tools? Can you retrieve the relevant paper?

Appendix 2: User-centred design process

The design process included practical activities of the following types:

- Scoping
- Surveys
- Interviews
- Workshops

The activities, summarised below, resulted in:

- a list of key collections of tools
- 45 responses to a general survey on requirements for the registry
- 27 responses to a survey which prioritised attributes in the software description model
- 17 software usability tests, including 8 intense, one-on-one sessions
- 191 feature requests or detailed comments, suggestions or requirements
- features and requirements prioritised using agile methodology
- extensive and detailed critique of the Query Interface
- detailed insight into user personas, information models, exchange formats, the federation of registry curation, software architecture, interaction with Google *etc.,* from technical discussions
- common strategic and technical plans discused within the BioMedBridges project and also with community stakeholders in this domain, e.g. BioCatalogue, ELIXIR-DK, ELIXIR-NO, BioSharing *etc.*)

To kick-start the registry with content of importance to BioMedBridgespartners and to provide a proof of principle for the Query Interface, workshops were held in which BioMedBridges partners manually curated with an initial 584 tools and 6,784 corresponding annotations. This seed content has now been supplemented with an additional 1,359 entries (see Section 4).

1 Scoping

Scoping of the registry development included identifying specific technical requirements, challenges and options for software data sharing and federation of curation / content, including standards for REST service description and software data exchange format. In addition, three lists were compiled of key technology and identify key providers, integrators, cataloguers and contacts:

- The existing key collections of tools including data services such as software registries/catalogues (*e.g.* BioCatalogue, AppDB, DRCAT, Databib), institutional providers of tools and databases (*e.g.* EMBL-EBI tools and data services, CBS tools, SIB tools), software packages/distributions (*e.g.* Debian Med, EMBOSS and BioConductor), WIKIs (*e.g.* SEQanswers WIKI, Wikipedia), Web portals and *ad hoc* lists on the Web etc (*e.g.* Bioinformatics.org) *etc.*
- The existing key technologies, controlled vocabularies (*e.g.* EDAM, SWO), standards initiatives (*e.g.* BioSiteMaps, RDA, SciencePad, re3data.org, BioDBCore/BioSharing) *etc.*
- Contact point for tools/services at each infrastructure, their national nodes, key tools providers within the nations and other relevant individuals

Ongoing revision and extension of these lists, by consultation with key individuals, projects, institutes and infrastructures and via a series of workshops, is part of the sustainability plan now in development.

2 Surveys

General user survey

http://tinyurl.com/bmbsurvey-general

This asked 25 general questions about the personal profile of the respondent, user requirements including goals and use cases, registry scope especially in terms of what types of software should be included, appropriate level of details of registered tools and services, responsibilities for software description and registration, preferences for software curation, requirements for service monitoring etc. The survey results informed the founding principles of the registry (Section 1.3) and the design and priorities of all aspects of implementation including software description model, curation, interfaces *etc.*

Software attributes survey

http://tinyurl.com/bmbsurvey-attributes

This asked 47 questions which rated each attribute in the software description model (Section 3) on a scale from "1 (Not useful)" to "5 (Very useful"). The survey results optimised the model, in terms of which attributes to include and whether an attribute should be mandatory, recommended or optional, but also helped to set the curation priorities and the prominence given to the attributes in the Query Interface, for example, which attributes should be visible in the default view.

3 Interviews

We conducted formal user experience sessions as well as informal user interviews. The formal sessions tested the usability of the Query Interface (Section 6.1). They involved a user, and an observer who recorded the actions and issues of the user as they worked through a set of ten realistic, biologically relevant exercises (Appendix 1-3), over a 15 minute session. The informal interviews were conducted by email, phone, or in person, and gauged the experiences of users who were tasked to perform usability testing on the Query Interface and report their experiences, as well as any other general suggestions. There were also *ad hoc* usability tests on each new software release. In addition, bilateral meetings with providers, integrators and cataloguers of software developed the general registry strategy.

4 Workshops

Here we list the trainer-led workshops, hackathons etc. in which BioMedBridges personnel participated AllBio workshop - "Web services for improved interoperability in bioinformatics" (Oct 2-5 2012, Munich)

- Ontology hackathon (October 9-13 2012, EMBL-EBI, UK)
- BioCatalogue meeting (December 7 2012, Manchester UK)

- BioMedBridges AGM Registry Workshop (March 11-12 2013, Dusseldorf, 14 attendees)
- AllBio / EMBRACE Continuity Workshop (March 18-20 2013, Amsterdam, 10 attendees)
- Imperial Registry Workshop (May 8 2013, Imperial College, UK, 15 attendees)
- BioMedBridges Registry Working Call (July 11 2013, Hinxton UK, 10 attendees)
- ELIXIR/BioMedBridges Workshop on Tool Registries (October 16-18 2013, CBS-DTU, Denmark, 14 attendees)
- Debian Med/Bio-Linux Sprint / hackathon (January 31 - February 3 2014, Aberdeen)
- BioMedBridges AGM Tools Workshop (Mar 9-12 2014)
- ALLBIO Workshop - “Metagenomics & interoperability” (April 10-12 2014, Amsterdam, 17 attendees)
- Mobyle, EDAM and Service Registry hackathon (June 17-18 2014, Paris)

These events brought together registry end-users including bioinformaticians, scientists and managers of all levels of experience, key providers of software, integrators, cataloguers and other key individuals representing relevant projects, packages, infrastructures, ontologies *etc.* The format of the workshops was optimised as time progressed, and included focussed presentations and discussions, but mostly mostly hands-on sessions where agile design and software development methodologies were applied to various aspects of the project. The purposes included:

- identify and refine elements of strategy for resource discovery and (inter)operability, including software description, cataloguing and data sharing
- identify common software attributes and project-specific ones
- develop the common software description model
- identify practical methods and best practice for software annotation
- develop related technology notably controlled vocabularies
- identify / tackle technical challenges for software data sharing / import

- identify modes and methods for federation of registry curation
- improve the usability of the Query Interface
- investigate applicability to Interoperability - software use and interconnection

5 Inreach and outreach

Inreach and outreach is ongoing and has included regular teleconferences with BioMedBridges partners, ELIXIR Tools Taskforce and ELIXIR Tools Node, regular EMBL-EBI Tools Committee meetings, and numerous other face-to-face meetings, calls and emails with BioMedBridges partners, ELIXIR nodes (*e.g.* ELIXIR-DK, CBS-DTU; ELIXIR-NO, Bergen University; ELIXIR-UK, Manchester University *etc.*) and other key partners without the BioMedBridges network (e.g. European Grid Infrastructure, AllBio, Neuroscience Information Framework *etc.*). In addition to these general activities there have been 8 specific outreach events to date, where the registry initiative was presented by poster or oral presentation:

- EBI RDF Meeting / BioMedBridges Technical Workshop, September 25-26, 2012)
- Gen2Phen 9th General Assembly Meeting (October 23-24, 2012, Toulouse, France)
- BioMedBridges AGM (March 11-12 2013, Dusseldorf, Germany)
- EGI / EMI Conference (April 9 2013, Manchester University)
- eIRG Meeting (May 22-23, 2013, Trinity College, Dublin, ~140 delegates)
- BioMedBridges AGM (March 9-12 2014, Florence, Italy)
- Wellcome Trust Genome Campus Resources Day (May 15 2014, Hinxton UK)
- ISMB Conference (July 2014, Boston, USA)

6 Feature requests / bug tracker

All requests, suggestions and comments are tracked using JIRA:

— Ontologies component: https://www.ebi.ac.uk/panda/jira/browse/BioMedBridges/component/10905
— Registries component: https://www.ebi.ac.uk/panda/jira/browse/BioMedBridges/component/10906

7 Mailing lists

There are two mailing lists:

— bioregistries-discuss: registry-related discussions
— bioregistries-announce: low traffic announcements by key providers, integrators and catalogues

To join a list, visit:

— http://elixirmail.cbs.dtu.dk/mailman/listinfo/bioregistries-announce
— http://elixirmail.cbs.dtu.dk/mailman/listinfo/bioregistries-discuss

To post to the list, mail:

— bioregistries-discuss@elixirmail.cbs.dtu.dk
— bioregistries-announce@elixirmail.cbs.dtu.dk (authority required)

Appendix 3: Registry Content

BioMedBridges partners were asked (during workshops *etc.*) to highlight which resources in the registry that they or their institute use routinely, or which are otherwise key to their work, and add to the registry any resources that were missing. This identified a total of 584 tools *etc.* and 6784 corresponding annotations. A breakdown of these data by entry type, interface type and annotation type is shown below (Table 2). There is an overwhelming interest in tools (application software) and databases, with surprisingly few (4) Web services proper listed. This reflects the typical BioMedBridges user phenotype; a scientists or biologist who does not require programmatic access. This is confirmed by the interface type preference, which is strongly “Web UI” or “REST API”, both highly accessible types of interface. “Web UI” refers to the typical graphical user interface on the Web, whereas “REST API” in practice refers to any tool that provides URL-based access. There was a significant number of other interface types too, confirming the requirement for broad scope.

Table A 2 Registry entries by type

Entry Type	Entry Type Description	Number of Entries
Service	A software system designed to support programmatic access and interoperable machine-to-machine interaction over a network. It has an interface described in a machine-processable format21.	4
Service end-points	A URL that provides a functional interface to a service.	34
Tools	Any application software that performs one or more specific operations to achieve a task22.	364
Database	Any organised collection of data served to the user via tools that allow interaction	111

[21] Based on the W3C definition of Web service.
[22] A service is a special type of tool.

Entry Type	Entry Type Description	Number of Entries
	with the user.	
Framework	Any software platform intended to help people to develop software applications, including code libraries, support tools such as compilers, tool sets, and application programming interfaces (APIs).	16
Package	A collection of related application software.	29
Other	Library, Macros, Script, Catalogue, Plugin, Virtual machine, Widget	11
Not specified		15

Table A 3 Registry entries by interface type

Interface Type	Interface Type Description	Number of Tools Entered
Web UI	A graphical user interface (GUI) to a tool provided by a Web site.	210
Command-line	A textual interface to a tool whereby commands to a program are typed in at the computer console.	63
REST API	An application programming interface (API) that follows a representational state transfer (REST) architectural style, in which a service manipulates representations of Web resources via a set of stateless operations23.	95
SOAP API	An API of Web services proper, i.e. an interface described in machine-processable format (WSDL) defining how other systems interact with the service using SOAP messages, typically serialised	53

[23] In practice many so-called REST services and REST APIs only loosely follow the REST paradigm

Interface Type	Interface Type Description	Number of Tools Entered
	in XML and conveyed using HTTP.	
Desktop GUI	A GUI that runs in the context of an operating system / window manager that is installed on a user's personal computer.	31
Not specified		132

Table A 4 Registry annotations by type

Registry annotation type	Number of annotations
Free-text	3272
Tags or ontology terms	1728
URL	1784

Table A 5 Summary of ontology development and annotations

Ontology/ branch	Software attribute	Description	Number of Terms
EDAM Topic	topic24	General domain the software serves	169
EDAM Operation	functionName9	Name of a software function	571
EDAM Data	dataType9	Type of data (primary inputs or outputs)	112625
EDAM Format	dataFormat	Allowed format of the data (primary	364

[24] Mandatory attributes
[25] This includes 595 terms for data proper and 531 for identifiers

Ontology/ branch	Software attribute	Description	Number of Terms
		inputs or outputs)	
SWO	softwareType9	Basic entity type: one of "Framework", "Package", "Database", "Tool", "Service" or "Other"	
SWO	interface9	Software interface type: "REST API", "SOAP API", "Web UI", "Command-line" or "Desktop GUI"	10
SWO	platform	Platform (OS and chipset combination) supported by a downloadable software package	
SWO	language	Language (for APIs etc.) or technologies (for Web applications, applets etc.)	46
SWO	license	Software or data usage license	40
SWO	maturity	Software maturity: one of "alpha", "beta" or "production"	1026
SWO	cost	Cost of purchase: one of "free" or "not free"	3

[26] Only three of these (“alpha”, “beta” and “production”) are used in the registry

Appendix 4: Query interface advanced features

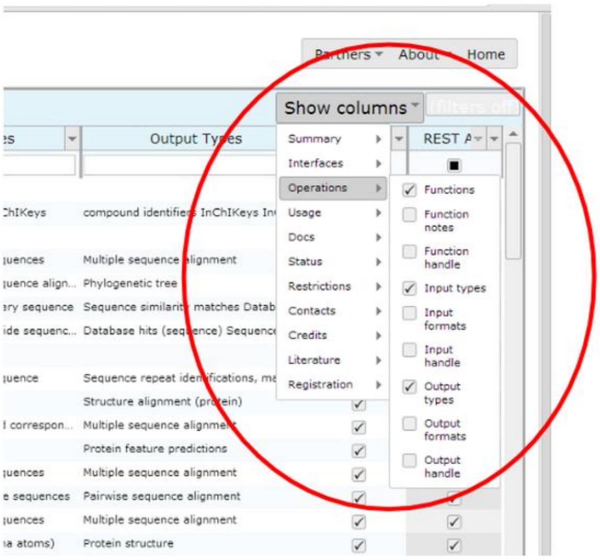

Figure A 1 Customisable columns

The registry workshops identified that, while most users enjoyed the simple grid-based navigation, approximately half of users expected a simple search box as the main entry point for searches. Accordingly, there is a search box (Figure 9) that lets you search over all the available fields of information.

In addition to the main search box, there are search boxes on each column allowing you to search directly within a particular field. In Figure 11 for example we are specifying we are looking for something with the specific function of "sequence alignment" that takes as an input a "protein". The grid displays a list of the available tools.

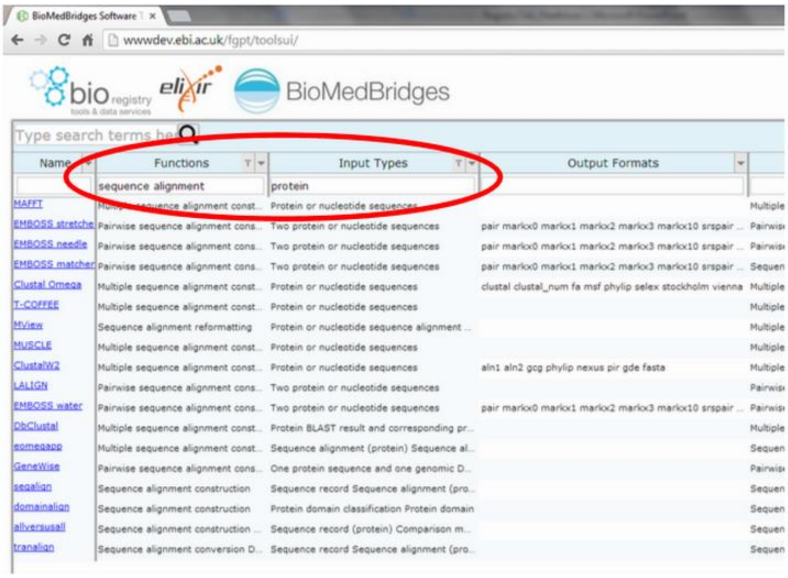

Figure A 2 Field-specific searches

Where appropriate, columns have a checkbox (Figure 12), which performs in a similar way as the search boxes to filter the results. Any number of column filters may be applied and the results are additive.

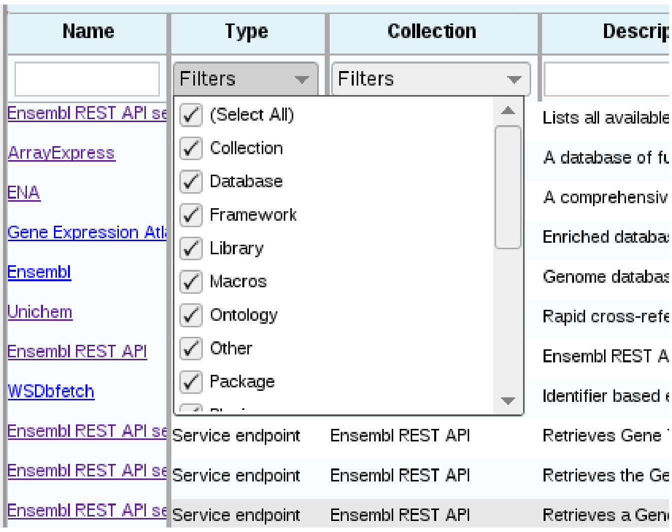

Figure A 3 Checkbox column filters