



Grant Agreement Number: 312251

# MIRRI

## Microbial Resource Research Infrastructure

SEVENTH FRAMEWORK PROGRAMME  
SP4-Capacities  
Combination of CP & CSA  
PREPARATORY PHASES  
FP7-INFRASTRUCTURES-2012-1

**Start Date of Project:** 01.11.2012  
**Duration:** 36 Months

### Deliverable Number

## D8.6

### Report on human and programmatic access

**Deliverable Date:** July 2015  
**Actual Submission Date:** July 2015  
**Lead Beneficiary:** Partner 10 - JacobsUni  
**Authors (alphabetical order):** B. Bunk, D. Colobraro, P. Dawyndt, P. De Vos, F.O. Glöckner, A. Kopf, V. Robert, P. Romano, D. Smith, C. Söhngen, F. Van Hauwenhuyse, A. Vasilenko  
**Version:** 2.1

Dissemination Level		
<b>PU</b>	Public	<b>X</b>
<b>PP</b>	Restricted to other programme participants (including the Commission Services)	
<b>RE</b>	Restricted to a group specified by the consortium (including the Commission Services)	
<b>CO</b>	Confidential, only for members of the consortium (including the Commission Services)	

## Summary

This report has been produced as a result of the activity of MIRRI work package 8, Task 8.4, and it is tightly linked to the other reports from the same as well as other work packages of MIRRI. The report outlines the optimal approach following technology produced in the World Wide Web Consortium (W3C), as well as in the ELIXIR system and in the BioMedBriges project. It intends to get the maximum compliance with databases, ontologies and the standards for applications like human health, veterinary science, pharma, agriculture, food production, winemaking, brewing and life science in general.

## Contents

1. Introduction
2. Human and programmatic access
  - 2.1. Human access
  - 2.2. Access for content indexing by search engines
  - 2.3. Access restrictions
  - 2.4. Programmatic access
3. Documentation development
4. Contents annotation
5. Conclusion
- Annex 1. Short presentation of European Bioinformatics Institute (EMBL-EBI)
- Annex 2. Open PHACTS data sources
- Annex 3. Example of data request in programmatic access
- Abbreviations
- References

## 1. Introduction

In this report we describe solutions that give the optimal fitness to the needs of the potential users of the projected MIRRI information system (MIRRI-IS). From the very beginning our priority was the best use of mBRCs strains - in brewing, wine-making, biofuels, pharma, health system, agriculture, food industry, environmental protection as well as in academic research. To align between microbial data in these areas of application and currently available strains in mBRCs on the other side data integration is needed.

To find the potential partners in this data integration we inspected life science information systems and databases presented in Biosharing (<https://www.biosharing.org/>) and Open PHACTS (<https://www.openphacts.org/>) - 669 databases totally.

287 databases in this list present information on microbes (Bacteria, Archaea, Fungi or Viruses) and potentially could become MIRRI-IS partners, 61 are not clear yet - they are likely to use microbes, but we didn't find any evidence.

In the list of  $287+61=348$  databases we looked for the possible practical use areas:

all of them can be used for scientific research, 54 databases were also in health system (mostly human), 12 - in winemaking, baking, brewing, 11 - in pharma, 5 - in agriculture, 4 - biotechnology with no indication of specific field of activity.

Most of these databases operate with information systems designed for human access. Some of them, mostly from pharma and health, OpenPhacts project, EMBL-EBI and some other systems, also use webservice for programmatic access.

## 2. Human and programmatic access

### 2.1. Human access

Most of life-science information systems presented in BioSharing usually provide:

- (1) function of search (simple or advanced) - to provide specific data that the user needs,
- (2) function of browse - to show the content of the database.

Usual characteristics of these systems:

- nothing extra - presented material is only what is needed,
- help descriptions for all the functions, tools demo, tutorial videos, very easy lessons in clips format, FAQs, examples, Wiki on the subject, help links on most of pages, Webinars,
- search and browse menu systems,
- upload menu for the files produced by the WEB users for this database,
- links to the databases with additional service,
- contacts page, response of the users.
- MIRRI-IS should have these characteristics plus:
  - enable intuitive navigation and access to the portal by researchers,
  - Wiki for all the European mBRC IT infrastructure,
  - if response to the user request is very big it could be sent back as a file,

- additional links and comments for applications (curator job, health system, bioindustry, agriculture, etc.),
- desirable option - "to keep the subject" on all the links - if some fraction of the page needs additional comments and have a link at that point, the linked additional page should follow the subject of the previous page,
- request system should be user-friendly, by default it should follow the language and the coding of the requestor,
- natural language processing, \*
- voice processing. \*

\*Remark: The last two options have no reliable software tools at the moment, but they may exist in the future.

The special study had been conducted to evaluate potential effectiveness of a search system based aggregation of mBRCs catalogs [13]. Catalogue characteristics of one well-known microorganism (*Aspergillus brasiliensis* Varga et al. 2007) presented in ATCC, CBS, DSM and VKM were compared. In this comparison, 24 fields presented one strain in four collections, and only two fields in two collections were found to be the same. Potentially, it could be 144 coincidences, and formally for the computer this means that consistency level in this example was 1-2%. These four presentations were not so different and looked complimentary to the human eye, but in order to make them compatible and complimentary for a computer some ontologies should be constructed, annotated and used. For effectiveness of the search, MIRRI-IS also presupposes an additional mBRC catalogue annotation procedure (see pos. 2. in section "4. Contents annotation" of this report).

MIRRI-IS Wiki systems could be constructed on Semantic MediaWiki (<https://semantic-mediawiki.org/>).

Semantic MediaWiki (SMW) is a free, open-source extension to MediaWiki – the wiki software that powers Wikipedia – that lets you store and query data within the wiki's pages.

Semantic MediaWiki is also a fully-fledged framework, in conjunction with many spinoff extensions, which can turn a wiki into a powerful and flexible knowledge management system. All data created within SMW can easily be published via the Semantic Web, allowing other systems to use this data seamlessly.

Additional comment from Wikipedia ([https://en.wikipedia.org/wiki/Semantic\\_wiki](https://en.wikipedia.org/wiki/Semantic_wiki)):

A semantic wiki is a wiki that has an underlying model of the knowledge described in its pages. Regular, or syntactic, wikis have structured text and untyped hyperlinks. Semantic wikis, on the other hand, provide the ability to capture or identify information about the data within pages, and the relationships between pages, in ways that can be queried or exported like a database through semantic queries.

Semantic wikis were first proposed in the early 2000s, and began to be implemented seriously around 2005. As of 2013, the best-known semantic wiki software, and the only one with significant usage on public websites, is Semantic MediaWiki.

Popular mistakes in some information systems:

(1) response to one request is fragmented into a group of pages,

- (2) additional linked comment page lose the subject of initial page,
- (3) the meaning of some links is not clear before you click it.

## 2.2. Access for content indexing by search engines

Search engines, like Google, Yahoo Search, AOL Search, Ask.com and Bing, just to mention some of the best known, are nowadays essential components of the Internet: they allow users to find the contents of their interest in a quick and effective manner, even when users themselves do not know which site could provide them.

Although MIRRI-IS is a specialist system, not of a general interest for network users, biologists may strongly benefit from being able to retrieve contents provided by MIRRI-IS through a search engine, at least in an initial phase when the MIRRI-IS won't yet be universally known.

For this, it is essential that the MIRRI-IS interface is properly designed so that: i) all its contents are indexed by search engines, and ii) they are shown among the first results when they are searched for by end users.

To this aim, all MIRRI information should be accessible directly, through a series of pages inter-linked, without the need to perform any query and fill any query form. In other words, it is essential that access to data is only possible through some query forms. Information must also be accessible in pages which can be reached by browsing existing links, possibly available in some index pages.

A couple of examples may clarify this approach.

The Cell Line Data Base (CLDB) is a relational database including information on human and animal cell lines available at some European resource centers and some Italian laboratories. It is aimed to broadly disseminate information on characteristics of cell lines and to facilitate access to these important resources (PR1).

The CLDB is only accessible through its hypertext format, HyperCLDB, at <http://bioinformatics.hsanmartino.it/hypercldb/>. There are no query forms. Instead, HyperCLDB includes many interlinked index pages, mainly about reference vocabularies for given information. In these pages, a list of all relevant cell lines' descriptions are included.

So, every search engine would first index the contents of the home page, then the contents of each index page that is linked from the home page, and finally the contents of all cell line descriptions. If CLDB was only accessible through a query form, then search engines would probably be unable to index all its contents.

A similar case is represented by the CABRI web site. Here, many query forms are available. The so called "Simple Search" (<http://www.cabri.org/CABRI/srs-doc/index.html>) is the simplest to use by unskilled users. Standard SRS search interfaces (<http://www.cabri.org/CABRI/srs-bin/wgetz?-page+top+-newId>) allows for more advanced and precise queries. CABRI also hosts the CABRI HyperCatalogue (<http://www.cabri.org/HyperCat/index.html>), which includes various index pages leading, possibly in multiple steps, to each and all complete descriptions of CABRI resources (PR2).

In this case too, query forms would block the navigation of search engine while indexing contents and the hypertext allows them to index each page in the full HyperCatalogue, including both intermediate pages and all resource descriptions.

A second issue relates to the possibility of being classified "higher" in search engine results pages. In general, only the first two or three result pages are used by end users, and being listed in the first page is considered as a notable asset.

To achieve this, a simple solution does not exist, also because each engine adopts different ranking algorithms, variable in time. E.g., Google states: "Sites' positions in our search results are determined based on **hundreds of factors** designed to provide end-users with helpful, accurate search results." (See [https://www.google.com/intl/en\\_us/insidesearch/howsearchworks/index.html](https://www.google.com/intl/en_us/insidesearch/howsearchworks/index.html), last accessed June 29, 2015).

Some considerations can be done anyway. The first point is to adopt flexible and wide-spectrum strategies. For sure, all pages must be properly annotated by "metadata" (once more). Information enclosed in the HTML page headers, including content type, author, title, copyright, description and keywords, have a stronger impact on the scores than the actual page contents. The URL is also relevant: a web site devoted to some special information will rank higher when a clear term is included in the domain name than one with the same term in the following components (directory and file names). From this point of view, it could be advisable for MIRRI to have a domain name making explicit reference to microbial resources instead of the acronym of the infrastructure.

The number of times a queried term appears in the contents of a page is also relevant. It's usually "normalized" with respect the contents length. That is, two occurrences of a queried term in a document of length N give the same score than one occurrence of the same term in a page of length N/2.

Specific for Google search is the PageRank algorithm(PR3).

This iterative algorithm ranks web pages based on the number and PageRank of other web sites and pages that link there, on the premise that good or desirable pages are linked to more than others. So, a page that is linked from many high-ranked pages in other web sites receives a high score.

Although direct access to all MIRRI-IS contents and proper, extended annotation of pages are prerequisite conditions for improving access for indexing by search engine, it is clear that an interface built on these conditions is far from optimized for both access of experts and for programmatic access. It is therefore essential that an interface having these characteristics is additional, not alternative, to other more effective interfaces.

### 2.3. Access restrictions

244 BioSharing information systems reported open access to their databases, 217 systems - restricted access with some or other format of security.

MIRRI-IS access restrictions:

1. Restricted access for special content to provide custom-tailored "data packages".
2. Restricted access due to legal restrictions based on IPR e.g. patents or personal data e.g. by hosting cell lines where the information on the origin e.g. cancer cells is protected by privacy legislations.
3. Limiting access to information on organisms that could be misused in biosecurity related activity but this is not likely to include the straightforward strain information; it would probably be specific information related to risk assessment. If biorisk is assessed as described in OECD best practice guidelines on biosecurity for mBRCs then information provided for example on "aerosol" formation or lethal dose etc. may be misused.

In fact, for both types of sensitive and regulated information, it is extremely unlikely that mBRCs will pass on this kind of data to a second or third party (MIRRI) portal for data storage or any kind of publication.

Usually patent data are not publicly available/accessible at all. The only exception might be inquiries that are made by the verified depositor of the respective material by contacting the patent office.

## 2.4. Programmatic access

The main technology for programmatic access in ELIXIR, EMBL-EBI and BioMedBridges is Semantic WEB, general description in <http://www.w3.org/2001/sw/>.

Resource Description Framework (RDF) is the central point of this technology (Introductory guide <http://www.ebi.ac.uk/rdf/about-technology>).

The RDF is a family of web standards maintained by the World Wide Web Consortium (W3C). It can be used as a way to represent, share and interact with data on the web. RDF encompasses a number of different technologies including a data model, syntax schemas, serialization formats, and a query language (SPARQL).

From a technical point of view, the Semantic Web consists primarily of three technical standards:

- (1) **RDF**: The data modeling language for the Semantic Web. All Semantic Web information is stored and represented in the RDF.
- (2) **SPARQL** (SPARQL Protocol and RDF Query Language): The query language of the Semantic Web. It is specifically designed to query data across various systems. An example of request in SPARQL is presented in Annex 3 to show the power of the language.
- (3) **OWL** (Web Ontology Language) The schema language, or knowledge representation (KR) language, of the Semantic Web. OWL enables to define concepts composable so that these concepts can be reused as much and as often as possible. Composability means that each concept is carefully defined so that it can be selected and assembled in various combinations with other concepts as needed for many different applications and purposes.

For implementation of Semantic WEB technology MIRRI-IS will need an RDF platform.

An RDF platform comprises five main components ([J14] page 39 and <http://www.Ebi.ac.uk/rdf/platform>):

- RDF Resolver - a simple Apache VM
- RDF Website - a Drupal site (<http://www.ebi.ac.uk/rdf/platform>)
- FTP site
- SPARQL endpoint/RDF browser (<http://rdf.ebi.ac.uk/dataset/biomodels/28>)
- Triple store databases

Semantic Web technology stack includes standards explicitly developed to help map data in legacy systems to RDF:

- RDB to RDF Mapping Language (R2RML) is a markup language that allows to specify how to map data from a relational database schema to RDF, so that a relational database can be queried by using SPARQL, possibly in association with other RDF datasets.
- Gleaning Resource Descriptions from Dialects of Languages (GRDDL) is a standard for associating XML documents with transformations that can be automatically run to convert XML into RDF.

Databases in RDF and ontologies are combined into the datasets. The datasets make nodes on Linked Data cloud (see <http://linkeddata.org/> and [http://lod-cloud.net/versions/2014-08-30/lod-cloud\\_colored.svg](http://lod-cloud.net/versions/2014-08-30/lod-cloud_colored.svg)). The Linked Data best practices propose that every dataset should provide provenance and licensing information, dataset-level metadata, and information about additional access methods (see also <http://linkeddata.org/guides-and-tutorials>).

General principles to follow in order to process and supply data:

### I. Linked data principles

- ★ make your stuff available on the Web (whatever format) under an open license
- ★★ make it available as structured data (e.g., Excel instead of image scan of a table)
- ★★★ use non-proprietary formats (e.g., CSV instead of Excel)
- ★★★★ use URIs to identify things, so that people can point at your stuff
- ★★★★★ link your data to other data to provide context

### II. Correct use of Unique Identifiers.

According to [J9] the General recommendations are:

- Use any currently-available identifier scheme that is “machine actionable, globally unique, widely (and currently) used by a community, and that has a long-term commitment to persistence (for example, see the W3C persistence policy). Best practice is to choose a scheme that is cross discipline.”
- The primary identifier of an entity must be unique and unambiguous: i.e. a 1:1 relationship of identifier: entity, and designed so that it never has to be changed, retired, or reassigned.
- We recognize the need for more formal specifications of identifier formats, and/or alignment between existing specifications. Some key considerations for identifier format are below.
  - An identifier may be used in more than one format (e.g. a database accession number and URI), but it must be possible to transform one format to the other.
  - Identifiers should adhere to an unambiguous format, ideally one definable by a regular pattern and whose prefix is unique with respect to other identifier schemes .
  - Consider the format `http://{domain}/{dataset}/{identifier}` for URL-based identifiers where {domain} is a stable domain name (e.g. [www.uniprot.org](http://www.uniprot.org)), {dataset} is a descriptive tag for the type of entity for which the URL will return data, and {identifier} is the primary entity identifier (typically a database accession number) For instance:  
<http://www.informatics.jax.org/allele/MGI:3845668>.
  - Regardless of how the entity record will be accessed, it should be comprised solely of web-friendly and printable ASCII characters without whitespace.
  - For database accessions:
    - Consider a fixed alphabetical prefix that intuitively conveys the identifier type and authority, and a numerical suffix that confers uniqueness, whilst keeping overall length as short as is practicable.
      - The alphabetical characters should be (non-accented) English letters, preferably not mixed case.



- Consider omitting a delimiter, or using an underscore if a delimiter is needed.
  - Consider using check digits or similar scheme to guard against typos. Check digits is rarely implemented in bioinformatics because doing so is harder and lengthens the identifier.
- For ontology accessions, consider following established best practice

### Creating identifiers

- Where an entity is already well identified, re-use the existing canonical identifier. If multiple identifiers already exist for an entity, and none has broader adoption, consider using the identifier that has the best-maintained mappings to the others. Otherwise, it is acceptable to create a new identifier, and maintain and publish the mappings to the others.
- Work with established authorities, e.g. major databases, on assignment of new identifiers, especially where they are expected to eventually host your dataset
- Work with dedicated and operationally independent services, e.g. identifiers.org , on issuance of new URL-based identifiers for database accessions
- Management policy of identifiers must be well defined and documented. Documentation should be publicly available and describe how ids are assigned and maintained.
- Versioning policy must be documented i.e. what kind of changes in data triggers creation of a new version number and how to obtain the current version. See versioning section for details.
- Register identifier types, tooling, and related implementations
  - Register new identifier types in the EDAM ontology (EDAM is an ontology of well established, familiar concepts that are prevalent within bioinformatics, including types of data and data identifiers, data formats, operations and topics, see [J22] and <https://github.com/edamontology/edamontology/issues>)
  - Register identifier-related services (e.g. a resolver or mapping service) in the BMB/ELIXIR Tools and Data Services Registry
  - Register in the BioSharing registry (<http://www.biosharing.org>, <https://rd-alliance.org/group/biosharing-registry-connecting-data-policies-standards-databases-life-sciences/case-statement> and <http://metadatascenter.org>) any public systems (such as databases and content standards) that make use of public identifiers.

These registers are in the process of being connected (cross-referencing records) under the ELIXIR umbrella.

### III. FAIR

According to (<https://www.force11.org/group/fairgroup/fairprinciples>) MIRRI-IS will follow FAIR Guiding Principles:

1. To be **Findable** any Data Object should be uniquely and persistently identifiable [4]
  - 1.1. The same Data Object should be re-findable at any point in time, thus Data Objects should be **persistent**, with emphasis on their metadata, [4 and [JDDCP 4](#) and [JDDCP 6](#)
  - 1.2. A Data Object should minimally contain basic machine readable metadata that allows it to be distinguished from other Data Objects [see [JDDCP 5](#)

1.3. Identifiers for any concept used in Data Objects should therefore be **Unique** and **Persistent** [5] and [JDDCP 4](#) and [JDDCP 6](#)].

2. Data is **Accessible** in that it can be always obtained by machines and humans

2.1 Upon appropriate authorization [6]

2.2 Through a well-defined protocol [7 and JDDCP 5]

2.3 Thus, machines and humans alike will be able to judge the actual accessibility of each Data Object.

3. Data Objects can be **Interoperable** only if:

3.1. (Meta) data is machine-readable [8]

3.2. (Meta) data formats utilize shared vocabularies and/or ontologies [9]

3.3 (Meta) data within the Data Object should thus be both syntactically parseable and semantically machine-accessible [10]

4. For Data Objects to be **Re-usable** additional criteria are:

4.1 Data Objects should be compliant with principles 1-3

4.2 (Meta) data should be sufficiently well-described and rich that it can be automatically (or with minimal human effort) linked or integrated, like-with-like, with other data sources [11 and JDDCP 7 and JDDCP 8]

4.3 Published Data Objects should refer to their sources with rich enough metadata and provenance to enable proper citation (ref to JDDCP 1-3).

[JDDCP \(Joint Declaration of Data Citation Principles\)](#)

[FAIR Principles Working Detailed Document](#)

For more detailed description of these tools also look at [J1] - [J15].

### 3. Documentation development

Document development is a key deliverable in Culture Collections and mBRC operations: Standard Operation Procedures, MAA, MTA, PIC, articles and regular reports. According to our experience the most efficient technology in development document contents is Top-Down schema [12]:

"Top-down design is a recursive heuristic for solving problems by writing functions: starting with a big-picture view of the problem; break it into a few big sub-problems; figure out how to integrate the solutions to each sub-problem; and then repeat for each part. The big-picture view: resources (mostly arguments), requirements (mostly return values), the steps which transform the one into the other. Breaking into parts: try not to use more than 5 sub-problems, each one a well-defined and nearly-independent calculation; this leads to code which is easy to understand and to modify. Synthesis: assume that a function can be written for each sub-problem; write code which integrates their outputs. Recursive step: repeat for each sub-problem, until you hit something which can be solved using the built-in functions alone. Top-down design forces you to think not just about the problem, but also about the method of solution, i.e., it forces you to think algorithmically; this is why it deserves to be part of your education in the liberal arts. Exemplification: how we could write the lm function for linear regression, if it did not exist and it were necessary to invent it."

More detailed descriptions can be found in:

- <http://www.egr.msu.edu/classes/ece410/mason/files/TopDownDesign.pdf>

- <https://software.intel.com/en-us/articles/how-to-tune-applications-using-a-top-down-characterization-of-microarchitectural-issues>
- <http://search.bwh.harvard.edu/new/pubs/ChangingYourMind.pdf>

In Information Processing technologies Top-Down was reported approximately 45 years ago for preparation of computer programs free of bugs. In fact it is not only the method of programming but the very efficient method of thinking. Just from the beginning and until now it is popular in many branches of the human activity.

Wikipedia ([https://en.wikipedia.org/wiki/Top-down\\_and\\_bottom-up\\_design](https://en.wikipedia.org/wiki/Top-down_and_bottom-up_design)) also recommends it for many application domains, including: product design and development, computer science, software development, programming, parsing, nanotechnology, psychology, management and organization, state organization, public health, architecture, neurosciences and ecology.

Examples of convenient tools for common access to the documents constructed in the team are Dropbox and Google drive.

Dropbox is a very convenient tool for a fast construction of common documents on seminars, workshops and videoconferences (<https://www.dropbox.com/home?preview=Getting+Started.pdf>). Several participants can browse/edit the document simultaneously.

Free accounts come with 2GB of space. The Dropbox is also available for iPhone, iPad, Android, and Blackberry.

Google drive provides deployment of the catalogues for a big library of documents with sophisticated passwords system.

For each of these tools good Internet access is necessary.

A good example of the content, structure and the style are documents produced in BioMedBriges on the subject of this report ([J1]-[J22]). BioMedBriges team participates in ELIXIR. EMBL-EBI is a leader in Information Technology Work Packages 3 and 4. MIRRI-IS construction team may have to follow them.

## 4. Contents annotation

MIRRI-IS supposes three annotation procedures.

1. Some ontologies used by potential partners in MIRRI-IS data integration will need additional refinements. For example, Environmental ontology (<http://purl.bioontology.org/ontology/ENVO>) presents only 77 types of soil (Full Id [http://purl.org/obo/owl/ENVO#ENVO\\_00001998](http://purl.org/obo/owl/ENVO#ENVO_00001998)). International soil classification [I4] presents much more detailed schema.

2. Comparison of mBRC catalogue strain properties with properties of the same strain in other mBRCs is highly desirable. Three options could be envisaged:

- (a) comparison inside aggregated MIRRI-IS catalogue on mBRC request,
- (b) protocol of differences discovered when mBRC uploads its catalogue,
- (c) protocol of new differences when other mBRC upload their catalogues.

Strains algorithm similar to StrainInfo is also necessary for this comparison.

3. Special algorithm could calculate the correlation between species genes (or the groups of genes) and phenotype properties. This is quite mathematical task. In fact, any MIRRI catalogue fields may be compared for correlation, the only requirement - field values must be comparable. In

this more general format this content annotation procedure looks like proposal in MIRRI D8.5 report [11] page 9.

## Conclusions

According to analysis dated June 2015, BioSharing (<https://www.biosharing.org>) presents 666 databases. This is only a fraction of the databases used in Life science. BioMedBriges uses 814 databases (<http://wwwdev.ebi.ac.uk/fgpt/toolsui/index.html> Filters=Database), MetaBase (MB) (<http://MetaDatabase.Org>) presents approximately 1800 databases.

If MIRRI-IS will do data integration with fraction of BioSharing databases that refer to microorganisms, this could mean that MIRRI-IS makes an agreement with each of these database producers. In our analysis this is 176 institutions and looks impossible.

We propose to start the data integration procedure with the most valuable database producers. Initial proposal - with EMBL-EBI and OpenPhacts (see Annex 1 and 2). The systems are the practical areas of biomedicine and pharmacology, both use Semantic WEB technology.

MIRRI-IS minimal solution could look like a dataset presented in Linked Data cloud (see <http://linkeddata.org/> and [http://lod-cloud.net/versions/2014-08-30/lod-cloud\\_colored.svg](http://lod-cloud.net/versions/2014-08-30/lod-cloud_colored.svg)) with:

- aggregate mBRC catalogue database in RDF format,
- ontology of microbial names for Bacteria, Archaea and Fungi in OWL format.

Practically this also means participation in:

- MOLGENIS (<http://www.molgenis.org/wiki/WikiStart>),
- MIRIAM Registry (<http://www.ebi.ac.uk/miriam/main/>),
- MMR (BioMedBridges Metadata Model and Mapping registry <https://molgenis08.target.rug.nl/>),
- BioMedBriges Tools registry (<http://wwwdev.ebi.ac.uk/fgpt/toolsui/>),
- BioSharing Standards registry (<https://www.biosharing.org/standards/>),
- as well as presentation of mBRC catalogues, CABRI, StrainInfo and WDCM/GCM in biosharing.org and in MetaBase (MB)

## Annex 1. Short presentation of EBI from [J21]

The European Bioinformatics Institute (EBI) is the largest bioinformatics resource provider in Europe. Our databases are accessible via dedicated interfaces, web services, data download and (in a few cases) direct database access. Modern research in the life sciences necessitates an understanding of data at many different levels: multi-omics, from cells to biological systems, across many different species and studying many different experimental conditions. The biology underpinning these research questions is intrinsically connected, yet data are often collected and stored in technology or domain-specific repositories.

Efforts in the Semantic Web community are already beginning to invest in technology that enables data to be readily integrated (Belleau et al., 2008; Katayama et al., 2010; Marshall et al., 2008). One method used among the Semantic Web community is using the W3C's resource description framework (RDF) model to represent data. RDF provides a common mechanism for describing data and querying data using SPARQL.

To better serve complex research questions across resources, and to meet an increased demand on the EBI to produce RDF, we have developed an RDF platform. The aim of such a platform is to offer users the ability to ask questions using multiple connected resources that share common identifiers and have a common format (RDF) and query interface (SPARQL). This platform complements other existing data access modes such as our Web site and RESTful web services, but additionally contains explicit links between the different data resources. This enables a single query to be asked across multiple distributed datasets and across a range of biological domains. This approach has been applied for the following EBI resources: Gene Expression Atlas [EBI1], ChEMBL [EBI2], BioModels [EBI3], Reactome [EBI4], BioSamples [EBI5] and also includes a collaboration with the UniProt Consortium to deliver UniProt RDF [EBI6].

## Annex 2. Open PHACTS (<https://www.openphacts.org/>) data sources include:

- \* ChEBI (<http://www.ebi.ac.uk/chebi/>),
- \* ChEMBL (<https://www.ebi.ac.uk/chembl/>),
- \* ChemSpider (<http://www.chemspider.com/>),
- \* ConceptWiki (<http://www.conceptwiki.org/>),
- \* DisGeNET (<http://www.disgenet.org/web/DisGeNET/menu>),
- \* DrugBank (<http://www.drugbank.ca/>),
- \* Gene Ontology (<http://geneontology.org/>) Gene database - search system,
- \* neXtProt (<http://www.nextprot.org/>) - human proteins,
- \* UniProt (<http://www.uniprot.org/>) - accessible resource of protein sequence and functional information,
- WikiPathways (<http://www.wikipathways.org/index.php/WikiPathways>).

Comment: \* - pharma, Semantic WEB

Open PHACTS databases not presented in BioSharing: ChEBI and ConceptWiki

Annex 3. Example of data request in programmatic access (<http://www.ebi.ac.uk/rdf/example-sparql-queries>).

How are the protein targets of the gleevec drug differentially expressed, which pathways are they involved in?

### SPARQL

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX cco: <http://rdf.ebi.ac.uk/terms/chembl#>
PREFIX chembl_molecule: <http://rdf.ebi.ac.uk/resource/chembl/molecule/>
PREFIX biopax3: <http://www.biopax.org/release/biopax-level3.owl#>
PREFIX atlasterms: <http://rdf.ebi.ac.uk/terms/atlas/>
PREFIX sio: <http://semanticscience.org/resource/>
PREFIX dcterms: <http://purl.org/dc/terms/>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>

SELECT distinct ?dbXref (str(?pathwayname) as ?pathname) ?factorLabel
WHERE {

  # query chembl for gleevec (CHEMBL941) protein targets
  ?act a cco:Activity;
    cco:hasMolecule chembl_molecule:CHEMBL941 ;
    cco:hasAssay ?assay .
  ?assay cco:hasTarget ?target .
  ?target cco:hasTargetComponent ?targetcmpt .
  ?targetcmpt cco:targetCmptXref ?dbXref .
  ?targetcmpt cco:taxonomy .
  ?dbXref a cco:UniprotRef

  # query for pathways by those protein targets
  SERVICE <http://www.ebi.ac.uk/rdf/services/reactome/sparql> {
    ?protein rdf:type biopax3:Protein .
    ?protein biopax3:memberPhysicalEntity
      [biopax3:entityReference ?dbXref] .
    ?pathway biopax3:displayName ?pathwayname .
    ?pathway biopax3:pathwayComponent ?reaction .
    ?reaction ?rel ?protein .
  }

  # get Atlas experiment plus experimental factor where protein is expressed
  SERVICE <http://www.ebi.ac.uk/rdf/services/atlas/sparql> {
    ?probe atlasterms:dbXref ?dbXref .
    ?value atlasterms:isMeasurementOf ?probe .
    ?value atlasterms:hasFactorValue ?factor .
    ?value rdfs:label ?factorLabel .
  }
}
```

## Abbreviations

CABRI	Common Access to Biological Resources and Information
CLDB	Cell Line Data Base
ELIXIR	European Life Science Infrastructure for Biological Information
EMBL-EBI	European Bioinformatics Institute
ENVO	Community ontology for the concise, controlled description of environments
MB	MetaBase
mBRC	microbial Biological Resource Centre
MIRRI	Microbial Resource Research Infrastructure
MMMR	BioMedBridges Metadata Model and Mapping registry
OWL	Web Ontology Language
R2RML	RDB to RDF Mapping Language
RDF	Resource Description Framework
SMW	Semantic MediaWiki
SPARQL	SPARQL Protocol and RDF Query Language
VM	Virtual Machine
W3C	World Wide Web Consortium
WDCM	World Data Center for Microorganisms



## References

- I1. MIRRI Microbial Resource Research Infrastructure. Deliverable Number D8.5. Report on users requests, desired features, and meta-analyses of the integrated platform. October 2013
- I2. "Top-Down Design (Introduction to Statistical Computing)". Masi.cscs.lsa.umich.edu. September 19, 2011. Retrieved September 18, 2012.
- I3. A. Vasilenko, S. Ozerskaya, O. Stupar. Current WFCC-CC catalogues as a starting ground for networking efforts. WORLD FEDERATION FOR CULTURE COLLECTIONS Newsletter (No.50)– JULY 2011. pp. 5-15
- I4. World reference base for soil resources 2014 International soil classification system for naming soils and creating legends for soil maps. FOOD AND AGRICULTURE ORGANIZATION OF THE UNITED NATIONS. Rome, 2014, 191pp. ISSN 0532-0488
- J1. BioMedBridges. Data strategies for research infrastructures. Considerations for infrastructure management based on conclusions from BioMedBridges workshops. Stephanie Suhr, Rafael Jimenez, Cath Brooksbank, Steven Newhouse, Tom Hancocks
- J2. BioMedBridges. E-Infrastructur Infrastructur Infrastructur e support for the life sciences – Preparing for the data deluge for the data deluge. A BioMedBridges knowledge exchange workshop hosted by ELIXIR, 15-16 May 2014, Hinxton, UK. Stephanie Suhr, Guy Cochrane, Nathalie Stanford, Jan-Willem Boiten, Jason Swedlow, Chris Morris, Rafael Jimenez, Pieter Neerincx
- J3. BioMedBridges. Principles of data management and sharing at European Research Infrastructures. Version 1.0, January 2014
- J4. BioMedBridges Identifier Best Practice and Supporting Tools.
- J5. BioMedBridges. SUMMARY: BioMedBridges resource integration workshop. 1 October 2014, EMBL –EBI, Hinxton
- J6. BioMedBriges. REPORT: BioMedBridges standards workshop. Nathalie Conte, Tom Hancocks, Stephanie Suhr
- J7. BioMedBridges. WORKSHOP REPORT: A common vocabulary to classify life science. 15 October 2014, Science Forum, Brussels.
- J8. BioMedBridges. Deliverable D2.3. Definition of a user-centered design process with the technical work packages. Francis Rowland and Cath Brooksbank
- J9. BioMedBridges. Deliverable D3.1. ESFRI BMS Online Dictionary of common molecular identifiers (eCMI). Jon Ison, Julie McMurry, Helen Parkinson, Nathalie Conte, Philipp Gormanns, Murat Sariyar, Gergely Sipos, Søren Brunak, Kristoffer Rapacki
- J10. BioMedBridges. Deliverable D3.2. Mapping and registry of ESFRI BMS standards (eSTR). Julie McMurry, Helen Parkinson, Philip Gormanns, Juha Muilu, Murat Sariyar, Morris Swertz, Dennis Hendriksen, Fleur Kelpin, Jonathan Jetten, Chao Pang
- J11. BioMedBridges. Deliverable D3.3. Provision and population of the ESFRI BMS Service Registry (eSR). 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
- J12. BioMedBridges. Deliverable D4.2. Assessment of feasible data integration paths in BioMedBridges databases. Helen Parkinson, Nathalie Conte, Andy Jenkinson
- J13. BioMedBridges. Deliverable D4.4. Identification of feasible BioMedBridges pilots for semantic web integration.



- J13a. BioMedBridges. Deliverable D4.6. Pilot integration of Web Services based simple object queries. Julie McMurry, Simon Jupp, Tony Burdett, Andy Jenkinson, Helen Parkinson, Chris Morris, Martyn Winn, Philipp Gormanns, Elida Schneltzer, Raffael Bild, Christian Krauth, Freek de Bruijn, Ward Blondé, Jeroen Belien, Stefan Klein, Erwin Vast
- J14. BioMedBridges. Deliverable D4.7. Report on the scalability of semantic web integration in BioMedBridges. Julie McMurry, Simon Jupp, James Malone, Tony Burdett, Andy Jenkinson, Helen Parkinson, Mark Davies, Marco Brandizi, Sarala Wimalaratne, Nicole Redaschi, Chris Morris, Martyn Winn
- J15. BioMedBridges. Deliverable D11.2. Second periodic report on current developments in the ICT e-infrastructures. Neil Geddes, John Chevers, Dai Davies, Bob Jones, Kimmo Koski, Steven Newhouse, Gergely Sipos
- J16. Data Strategies for Research Infrastructures: Financial Requirements. Steven Newhouse. Head of Technical Services, EMBL-EBI
- J17. Christopher J. Mungall, David B. Emmert, and The FlyBase Consortium. A Chado case study: an ontology-based modular schema for representing genome-associated biological information. *BIOINFORMATICS*. Vol. 23 ISMB/ECCB 2007, pages i337–i346, doi:10.1093/bioinformatics/btm189
- J18. Dan M. Bolser, Pierre-Yves Chibon, Nicolas Palopoli, Sungsam Gong, Daniel Jacob, Victoria Dominguez Del Angel, Dan Swan, Sebastian Bassi, Virginia González, Prashanth Suravajhala, Seungwoo Hwang, Paolo Romano, Rob Edwards, Bryan Bishop, John Eargle, Timur Shtatland, Nicholas J. Provart, Dave Clements, Daniel P. Renfro, Daeui Bhak, and Jong Bhak. MetaBase—the wiki-database of biological databases. *Nucleic Acids Res.* Jan 2012; 40(Database issue): D1250–D1254.
- J19. Nick Juty, Nicolas Le Novère, and Camille Laibe. Identifiers.org and MIRIAM Registry: community resources to provide persistent identification *Nucleic Acids Res.* Jan 2012; 40(Database issue): D580–D586. Published online Dec 1, 2011. doi: 10.1093/nar/gkr1097. PMID: PMC3245029
- J20. Hans-Werner Hilse and Jochen Kothe. Implementing Persistent Identifiers. Overview of concepts, guidelines and recommendations. Consortium of European Research Libraries. European Commission on Preservation and Access. 2006
- J21. Simon Jupp, James Malone, Jerven Bolleman, Marco Brandizi, Mark Davies, Leyla Garcia, Anna Gaulton, Sebastien Gehant, Camille Laibe, Nicole Redaschi, Sarala M. Wimalaratne, Maria Martin, Nicolas Le Novère, Helen Parkinson, Ewan Birney, and Andrew M. Jenkinson. The EBI RDF platform: linked open data for the life sciences. *Bioinformatics*. 2014 May 1; 30(9): 1338–1339. Published online 2014 January 11. doi: 10.1093/bioinformatics/btt765. PMID: PMC3998127
- J22. Jon Ison, Kalaš, Inge Jonassen, Dan Bolser, Mahmut Uludag, Hamish McWilliam, James Malone, Rodrigo Lopez, Steve Pettifer and Peter Rice. EDAM: an ontology of bioinformatics operations, types of data and identifiers, topics and formats. *Bioinformatics*, 2013, Volume 29, Issue 10, Pp. 1325-1332
- EBI1. Kapushesky M, et al. Gene Expression Atlas update—a value-added database of microarray and sequencing-based functional genomics experiments. *Nucleic Acids Res.* 2012;40:D1077–D1081. [[PMC free article](#)] [[PubMed](#)]
- EBI2. Gaulton A, et al. ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res.* 2011;40:D1100–D1107. [[PMC free article](#)] [[PubMed](#)]
- EBI3. Li C, et al. BioModels Database: an enhanced, curated and annotated resource for published quantitative kinetic models. *BMC. Syst. Biol.* 2010;4:92. [[PMC free article](#)] [[PubMed](#)]

EBI4. Matthews L, et al. Reactome knowledgebase of human biological pathways and processes. *Nucleic Acids Res.* 2008;37(Suppl. 1):D619–D622. [\[PMC free article\]](#) [\[PubMed\]](#)

EBI5. Gostev M, et al. The BioSample Database (BioSD) at the European Bioinformatics Institute. *Nucleic Acids Res.* 2012;40:D64–D70. [\[PMC free article\]](#) [\[PubMed\]](#)

EBI6. Redaschi N. Consortium, UniProt. UniProt in RDF: Tackling data integration and distributed annotation with the semantic web. 2009 Available from Nature Precedings <http://dx.doi.org/10.1038/npre.2009.3193.1>.

PR1 P Romano, A Manniello, O Aresu, M Armento, M Cesaro, B Parodi. Cell Line Data Base: structure and recent improvements towards molecular authentication of human cell lines. *Nucleic Acids Research* 37 (suppl 1), D925-D932

PR2 P Romano, M Kracht, MA Manniello, G Stegehuis, D Fritze. The role of informatics in the coordinated management of biological resources collections. *Applied Bioinformatics* 4 (3), 175-186

PR3 Brin S, Page L. The anatomy of a large-scale hypertextual Web search engine. *Computer Networks and ISDN Systems* 1998, 30: 107–117.