Microbial Resource Research Infrastructure: status and perspectives on data integration and sharing

Paolo Romano[1], Giovanna Cristina Varese[2]

[1] IRCCS Ospedale Policlinico San Martino, Genoa, Italy

[2] Mycotheca Universitatis Turinensis, University of Turin, Turin, Italy

The Microbial Resource Research Infrastructure

The Microbial Resource Research Infrastructure (MIRRI) (http://www.mirri.org/) is being built in the sphere of the European Strategy Forum on Research Infrastructures (ESFRI) initiative. The MIRRI partnership comprises 16 main partners and 28 collaborating parties from 19 countries across Europe. Its mission is to overcome fragmentation in availability of resources and services, while focusing on needs and challenges facing both the microbial domain Biological Resource Centres (mBRCs) and the user of microorganisms from industry and research. It aims to provide a unique entry point to quality microbiological services and mBRC holdings.

With reference to Information Technologies (IT) issues, achieving MIRRI aims requires improvements in the interoperability between mBRCs, as well as an exceptional increment of data offers. Indeed, with the advent of high-throughput technologies in life sciences and of the consequent shift of the research focus from cellular to molecular data, essential information on molecular profiles of microorganisms must be provided in tight connection with the catalogue information. This implies the implementation of a smart and flexible architecture, able to properly cope with sequence information, phenotyping data, and images, which are inherently big data.

The status of BRC information systems

The vast majority of European mBRCs offers the catalogue information on-line for the benefit of interested academic and industrial stakeholders. Little is instead available in terms of integrated access to many catalogues. Moreover, BRCs has created a wide variety of information systems, largely heterogeneous, not interoperable, and far from the FAIR (Findable, Accessible, Interoperable, Reusable) approach.

A few successful examples in this regard refer to Common Access to Biological Information and Resources (CABRI, http://www.cabri.org/), StrainInfo (http://www.straininfo.net/) and the Global Catalogue of Microorganisms (GCMs, http://gcm.wfcc.info/). The CABRI network services offer integrated access to 25 catalogues including more than 130,000 microbial resources since 2000. This was enabled by the definition and adoption of common datasets and of a devoted data format for the catalogue contents. Indexing and querying catalogues was performed though an implementation of the Sequence Retrieval System (SRS, http://www.cabri.org/guidelines/catalogue/CPdata.html). The “Extended Query Form” of the standard SRS interface, as well as the custom “Simple Search” tool, allows searching all catalogues together in a variety of ways.

The StrainInfo database consists of metadata extracted from BRCs catalogues. This is used to create strain passport data to link the various collection numbers of a given strain to certified information, such as taxonomic name, sequence data, and bibliographic references. Thus, StrainInfo allows searches for strains belonging to a given species, or searches related to a given strain number, through more than 60 catalogues including ca. 700,000 collection numbers related to ca. 300,000 strains (see http://www.straininfo.net/stats).

A large number of strain details from many BRCs can also be accessed through GCMs, which likely is the most thorough integrated catalogue of microbial strains. However, a number of its features do not cater for the users' needs: the data of each catalogue must be manually transferred to the GCM, resulting in a number of out-of-date catalogues, and searches are quite basic and not flexible.

In this context, it is clear that a new information system is needed. Such information system must overcome the current limitations of CABRI, StrainInfo and GCM, while at the same time be able to extend its contents to high-throughput data and new software applications, which are not limited to the query of catalogues but are also able to perform some bioinformatics analyses.

The proposed IT architecture and some recent achievements

During its preparatory phase, a possible IT architecture for the MIRRI Information System was identified and described (see figure 1). It foresees the adoption of the following components: i) a standard format for exchanging data between BRCs, to be likely based on the Microbiological Common Language (MCL), ii) a Minimum Data Set for essential data, to evolve into Minimum Information about Biological Resources (MIaBRe), iii) a user-friendly interface to be included in a Collaborative Working Environment (CWE), and iv) state-of-the-art APIs and services/workflows for well-known and largely adopted integration software.

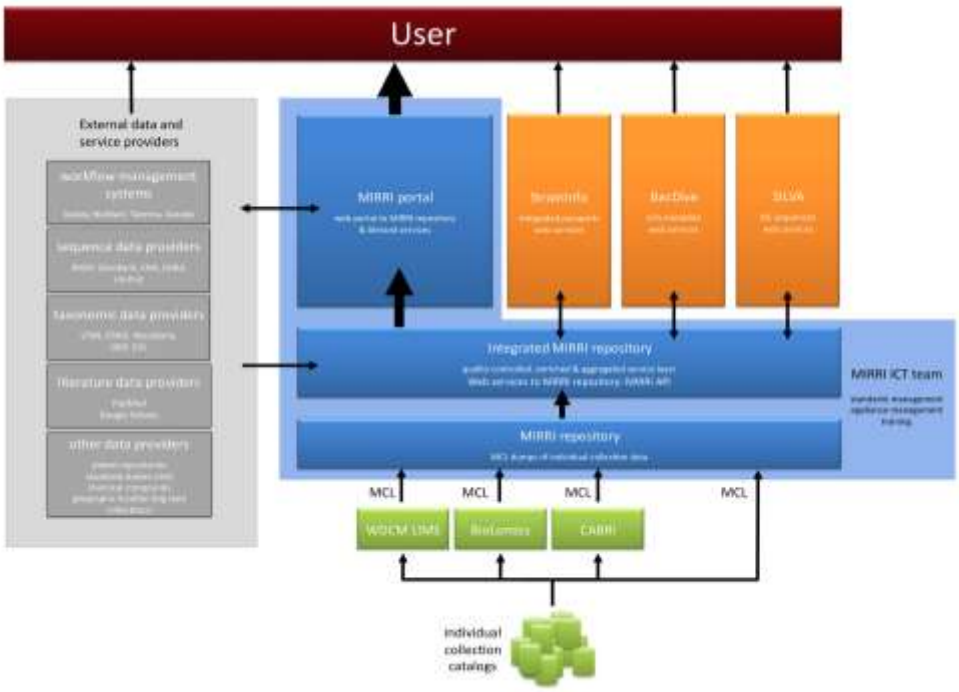

Figure 1: Architectural design of the MIRRI Information System

On the way to achieve the desired outcome, three systems, the so-called "MIRRI demonstrators", were developed in the MIRRI preparatory phase. The three demonstrators have been developed to cope with distinct issues, all equally essential towards the creation of a MIRRI Information System. The BacDive demonstrator aims at extending the contents of catalogues with a greater number of better-defined data. The effort put into the building of BacDive is greater than what can usually be done by a generic culture collection. However, it demonstrates which information could be useful (for a given organism type) and designs a way to manage all this information. Then, it shows how this "content extension" can be achieved progressively, by selecting subdomains of interest starting from the most recent/interesting strains.

The StrainInfo demonstrator is targeted towards a better integration among collection catalogues through the identification of common strains. It makes some order in strains available in various collections and makes possible the re-organization of collections and the sharing of data between catalogues.

The USMI Galaxy demonstrator is aimed at supporting data curation and at integrating catalogues with external resources. It makes it possible to integrate collections' data with other bioinformatics databases by leveraging on existing tools, like Galaxy, well known and with little development requirements. Moreover, it allows the improvement of a collection's data by automating links to external databases that may help the adoption of standardized, and shareable, terminologies and data values.

Needs and perspectives at the Italian level

As per any European Research Infrastructure defined in the ESFRI context, the actual realization of MIRRI must go through the involvement of interested countries, which are called to both support the implementation of the coordinating node and the needs of the national community. While the former is clearly defined at the European level by an agreement among interested counties, the latter may follow various forms, according to the initial conditions that may vary from country to country.

In Italy, there are many collections of microbial strains, although only a few of them have a clear mission for providing services. Moreover, coordination among these collections is still limited. However, there are great potentials for a positive follow up of the creation of an effective network in many different fields including biotech research and industry, health and many others. For these reasons, we believe that Italy should also support the creation of a network of mBRC at the national level.

A Joint Research Unit for the implementation of the Italian node of MIRRI (MIRRI-IT JRU) has recently been formed with the contribution of the Universities of Turin, Perugia, and Modena and Reggio Emilia, the National Research Council, and the University Hospital San Martino. Among the aims of the MIRRI-IT JRU is the development of a tight network of Italian collections of microbial resources. For its implementation along the lines of the MIRRI Information System architecture, important IT facilities are required. Due to the public nature of the JRU, which gather public research institutes connected to GARR, the natural partner for all IT facilities is the public GARR network.