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

D8.4

Report on comparison of various integration software operating in Life Sciences

Deliverable Date:	July 2015
Actual Submission Date:	November 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:	1.0

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

Summary

This report has been produced as part of the activity of work package 8, task 8.4. It is linked to the previous reports from the same work package, which are cited, whenever appropriate. The activity of the work package is to define the desired IT architecture for the Microbial Resource Research Infrastructure (MIRRI), whose objectives are clearly defined in, e.g., the "MIRRI WP8 strategy paper about data resource management" and, more recently, the "Strategy for a FAIR (Findable, Accessible, Interoperable and Reusable) provision of data and information to meet MIRRI's data management and delivery needs" (part of D6.6); the architectural design is still under discussion.

The objective of this deliverable is the analysis of some data integration architectures and tools in Life Sciences to determine their respective advantages and disadvantages in meeting the objectives of the MIRRI infrastructure. It is noteworthy, however, that some WP8 partners have built and made available software demonstrators that show different aspects of the features that could be included in the MIRRI Information System (MIRRI-IS). Due to their relevance in the context of this deliverable, they are also briefly introduced here.

The following sections are included in this report:

- Data integration in Life Sciences: objectives, issues and expectations
- Introduction to information technology architectures for data integration
- Data integration platforms for molecular biology and microbiology
- Comparison of some data integration tools for Life Sciences
- The MIRRI Demonstrators
- Conclusions

Data integration in Life Sciences: objectives, issues and expectations

Main **goals of data integration in Life Sciences** are:

- the **achievement of a comprehensive view of available information**, which is usually spread in many different databases and sites,
- the **carrying out of complex analysis workflows** involving access to several databases and software tools in an automatic way,
- the **execution of large scale analysis**, that is the effective and efficient analysis of large volumes of data.

In general, data integration can be best achieved when the related information and desired analysis are stable in time. Additionally, it is made easier by deep and sound domain knowledge and by properly defined information and data. These conditions may lead to a standardization of data models and formats. Also essential is a clear definition of desired outputs. On the contrary, integration is hampered due to heterogeneous data and systems, uncertain domain knowledge, highly specialized and quickly evolving information, lack of predefined, clear goals and originality of procedures and processes.

Unfortunately, the latter is the case for Biology and Medicine and, therefore, **data integration in Life Sciences is a difficult task**. There are many factors impacting on this: from heterogeneity of data models and formats to distribution of databases in many different servers; from the variability of targets and aims of both databases and software tools to the database size, which is often impressively large; and the fast evolution of database contents and formats, which is required by new understandings of how living organisms work, just to mention a few.

Biological and medical data is currently available in many public, distributed, heterogeneous and uncoordinated systems. The Nucleic Acids Research (NAR) Molecular Biology Database Collection demonstrates this situation. In its 2015 edition (1), it includes 1,615 databases, grouped in 15 categories and 41 subcategories. The vast majority of these databases are hosted on different servers which only host one database each as shown in Table 1. This is the case for 1,216 databases out of 1,615, i.e. ca. 75%. The three main servers host 103 databases, which only represent about 6% of the total.

DB	Server	Coverage	Server names
1	1,216	75.29%	Too many to list
2	76	9.41%	Too many to list
3	15	2.79%	www.hiv.lanl.gov, www.glycosciences.de, www.gene-regulation.com, www.ddg-pharmfac.net, www.compbio.dundee.ac.uk, www.chem.qmul.ac.uk, www.cbil.upenn.edu, urgi.versailles.inra.fr, srv00.ibbe.cnr.it, projects.tcag.ca, lifecenter.sgst.cn, genome.ewha.ac.kr, datam.i2r.a-star.edu.sg, bioinformatics.psb.ugent.be, apps.sanbi.ac.za
4	10	2.48%	www.uniprot.org, www.megabionet.org, www.jcvi.org, www.imtech.res.in, www.genome.jp, www.cbs.dtu.dk, pbil.univ-lyon1.fr, genolist.pasteur.fr, genecards.weizmann.ac.il, cmbi.bjmu.edu.cn
5	2	0.62%	www.expasy.org, research.nhgri.nih.gov
6	4	1.49%	wwwmgs.bionet.nsc.ru, www.imgt.org, pir.georgetown.edu, mips.helmholtz-muenchen.de
7	1	0.43%	mips.gsf.de
8	1	0.50%	crdd.osdd.net
10	1	0.62%	caps.ncbs.res.in
16	1	0.99%	bioinformatics.charite.de
42	1	2.60%	www.ncbi.nlm.nih.gov
45	1	2.79%	www.ebi.ac.uk

Table 1: Distribution of databases by server host (from NAR Molecular Biology Databases Collection, last accessed August 3, 2015).

Of course, database size and number of accesses are not taken into account in this analysis. So, the above percentages do not represent **data usage**. Only databases which were described in a NAR paper and that are still available on-line are included in this count. There are many more databases than those listed.

This **size and information content** is increasing at an impressive rate. Release 124 of assembled/annotated sequences from the European Nucleotide Archive (ENA), issued in June 2015, contains 608,493,388 sequence entries comprising 1,326,648,168,352 ($1.3*10^{12}$) nucleotides. The numbers of entries grew by 14.22% since the previous release and by 37.80% in one year (see ftp://ftp.ebi.ac.uk/pub/databases/ena/sequence/release/doc/relnotes.txt, last accessed August 2, 2015). ArrayExpress, a microarray experiment database maintained by the European Bioinformatics Institute (EBI), included 58,583 experiments and 1,758,891 assays on August 3, 2015, which occupy 35.06 Tb of archived data (see http://www.ebi.ac.uk/arrayexpress/, last visited August 2, 2015). Data on scientific literature is also growing, Medline presently includes more than 22,000,000 bibliographic references with article abstract. PubMed Central (PMC), a free full-text archive of biomedical and life sciences journal literature which is tightly linked to Medline, includes more than 3.5 million full texts.

It is of interest to note that the nucleotide sequences databanks available at EBI, NCBI and the Japanese National Institute of Genetics (NIG) exchange data on a peer to peer basis under the framework of the International Nucleotide Sequence Database Collaboration (INSDC). Similarly, developments and management of databases related to microarray expression, molecular pathways and molecular interactions are **coordinated**. However, these are exceptions, the rule being that many databases with similar or even overlapping contents are managed independently, without sharing a common data structure.

Information in **secondary databases**, which are the vast majority of existing ones, is often of the highest quality. Data is derived from primary databases and “curated”, which involves a careful removal of errors and duplication, as well as extensive annotation. These often represent an essential resource for researchers since they focus on special research interests. These databanks are often created and maintained by small groups or even by single researchers. Their format is determined both by the type of information stored and user needs. This has led to a high number of heterogeneous databases, the majority of which are of a great interest to researchers.

As a result of this diffused and often uncoordinated development, **data is spread** over hundreds of Internet sites where it is stored using **heterogeneous** Database Management Systems and data structures. There are few common information sets and the semantics associated to data, i.e. the actual meaning associated by developers to each piece of data, can be different. Different names given to similar concepts (synonymy) and a unique term used to describe different concepts (homonymy) are two of the most frequent sources of semantic heterogeneity, eventually leading to potential confusion.

International standards for shared semantics and common formats of Life Sciences data are now emerging. One of the most important International efforts is the Minimum Information for Biological and Biomedical Investigations (MIBBI) (2), a community approach to developing standards for biomedical information. MIBBI presently lists 30 different projects, the best known are MIAME, for microarray data, MIAPE for proteomics experiment, CIMR for metabolomics, MIMIx for molecular interactions and MIxS for sequence data.

Efforts for enabling **semantic data integration** are also on-going. The Open Biomedical Ontologies (OBO) initiative (3) (http://www.obofoundry.org) aims at co-ordinating the development of a set of coherent and interoperating reference ontologies for Life Sciences. Among these, some have a clear relation to MIRRI, including Fungal Anatomy Ontology (FAO), NCBI organismal classification (NCBITaxon), Ontology for biomedical investigations (OBI), Biological Collections Ontology (BCO), Chemical entities of biological interest (CHEBI), Ascomycete phenotype ontology (APO). An on-going analysis being carried out by Alexander Vasilenko and colleagues has identified 173 terminologies and ontologies of interest to MIRRI. It is clear that much work is still required to reach a clear landscape of what is available and, especially, what is still inadequate or even missing. In this effort, a relevant role can be played by MIRRI on the basis of its expertise and long experience in the domain.

User interfaces and **query methods** of the existing databases are quite different. Consequently, searching, retrieving and integrating information may become very hard. Data is then often analysed by researchers accessing several servers through their web browsers and using the “cut & paste” technique to transfer data from one web resource to another.

The **adoption of** the eXtensible Markup Language (**XML**) to ensure well-structured data, that is manageable by software tools, is increasing and progressively becoming the standard way to exchange data. This, associated with Application Programming Interfaces (APIs), usually SOAP and REST based Web Services allowing software to interoperate for a semantic savvy data exchange, are opening the doors for a generation of flexible integration tools which are based on the definition of elaboration pipelines (or workflows) and their automatic execution (orchestration).

In biology, the **domain’s knowledge** changes very quickly. This, together with the complexity of involved information, make it difficult to design data models which are valid for different application domains and over time. The goals and the needs of researchers evolve very quickly, responding to new theories and discoveries that lead to new data, goals, and processes.

Flexibility of methods and systems, including the ability to support frequent changes of data models, software and objectives of the analysis, is then needed. Integrating biological information in a distributed, heterogeneous environment requires flexible, expandable and adaptable technologies and tools that allow the move towards automation of data analysis through systems that may automatically access remote sites, retrieve information from the databases of interest and/or use the appropriate software to achieve the desired analysis, while at the same time are able to cope with the heterogeneity of data sources and select and manage the right information (semantics) properly.

Introduction to information technology architectures for data integration

Different approaches are possible to tackle data integration issues in Life Sciences.

Some integration methods are based on **syntactical tools**, such as explicit cross-references, implicit links (e.g., by leveraging on shared names) and common contents (e.g., by adopting terms from shared vocabularies and lexicons). Data integration can be implemented by exploiting these links to provide hypertext-like navigation through cross references between sources. These methods, however, rely on the manual annotation of data, which is a long and costly task, prone to errors and very demanding. They also are unable to convey the semantics of the link: a connection between descriptions of a gene and a protein does not define why they are linked (e.g., expression, molecular interaction). Integration methods based on semantic links, such as those that can be derived by using reference ontologies, seem more adequate.

A **data warehouse** (DW or DWH), is a central repository of integrated data which is extracted from more disparate sources. The DW has its own data model and schema, which is usually aimed at optimizing the performance of the system towards a defined objective. Thus, the main tasks involved in the building and maintenance of a DW include extracting relevant data from each of the source data systems, transforming this information from the original data model to the one of the DW, and storing the resulting data in the production database.

Of course, these processes imply an adequate knowledge of the data models of the original data sources, which may vary over time, and a frequent update of the system, which may also involve a global reload of all information. The main advantage of the DWs is of course the improved performance that can be obtained by integrating all data in a unique, optimized system, while the main disadvantages consist in the requested knowledge of original databases and methods for extracting information and in the need for frequent updates. These difficulties can be overcome when a consensus schema, eventually associated with a shared exchange language or application interface, can be defined. **DWs are efficient and effective data integration tools when a common data model can be defined and data can be easily (and frequently) updated**.

A **federated database** (FD) is a collection of autonomous databases, which are transparently connected by means of a set of coherent "mappings". The original databases, which remain autonomous and decentralized, conserving its individual visibility, must be interconnected via a computer network. The resulting system is not a new database, but a system enabling integrated access to all participating databases. This provides a uniform user interface and enables users to retrieve and store data from multiple databases with a single query. A FD splits the original query into sub-queries for the relevant databases and later combines results into a unique dataset for the user. It usually submits sub-queries to original databases by adopting adequate query languages. Apart from the need to build proper query interfaces to remote databases its main advantage is a reduced administrative burden, since a central repository is not needed. Additionally, data is retrieved from original databases when it is requested, this heavily depends on networks and requires that all databases are constantly running optimally and remain accessible. It is therefore prone to network delays and faults, as well as to faults at the database web sites.

A flexible alternative to FD is represented by **Mashups**. These provide aggregate information extracted from remote sources without integrating them and usually adopting the same visualization format that is provided from the sources. As a result information is just displayed in the same graphic framework. This approach has therefore limited ability to extract a coherent subset of information from sources, but has the great advantages of limiting the needed knowledge on source data formats to a minimum, improving retrieval performances, and quickly adapting to evolution of remote sources; dependency on networks remain.

A **distributed system** (DS) is a system whose components (nodes) interact with each other to achieve a common goal. Nodes are autonomous systems, usually distinct servers, that communicate and coordinate their actions by message passing. DS main architectures are Client/Server and Peer-to-Peer (P2P): while in the former architecture, which may in part resemble a federated database, a single server provides a service to many clients; in the latter all components contribute processing power and memory to the aims of a distributed computation. The P2P schema, however, is not adequate for MIRRI needs.

Linked Data (LD) is a method for publishing structured data on the Web so that it can be found, accessed, retrieved, and interlinked, independently of its original dataset or database, by using standard query tools. It is noteworthy that, in this context, each single piece of information is published and can be accessed and retrieved, along with the database it comes from. LD is of special interest for data integration in Life Sciences when it is implemented by using standard Semantic Web technologies (RDF and URI) and properly annotated by means of shared ontologies. In this case, LD can be queried and integrated by using SPARQL queries.

The **Linked Open Data** (LOD) initiative aims at extending the Web by publishing various open data sets as RDF and sets RDF links between data items from different data sources. A Linked Open Data cloud diagram is available on-line [http://lod-cloud.net/]. This shows that a significant number of biomedical data is already available on the LOD: much data derives from a unique data set (Bio2RDF, see next section), but there also are some that are independently built, e.g. Diseasome, a dataset extracted from OMIM that includes information on disorders and disease-related genes linked by known associations, UniProt, BioModels and BioSamples.

Workflows can be defined as automations of processes, in whole or part. Their goal is the execution of data elaboration processes in standardized environments, where standardization supports automation. In the design phase, all steps in a complex process are ordered in the proper way and interlinked so that in the execution phase the process can be carried out by enacting each task when all its prerequisites are fulfilled and by transferring data as required. Workflows' advantages relate to effectiveness (automation of repetitive procedures), reproducibility (an analysis encoded in a workflow can be reproduced easily), reusability (of both procedures and intermediate results) and traceability (through the analysis of intermediate results and associated metadata).

A **Workflow Management System** (WMS) is a system that supports the definition and creation of workflows and controls their execution. Its main components usually are: i) a graphical interface for composing workflows, entering data, displaying results, ii) an archive to store workflow descriptions and results of their executions, iii) a registry of available services for task execution, both database and analysis software, local and remote, iv) a scheduler, able to orchestrate the execution of the workflow, v) a monitor for the execution of the workflows, vi) adequate visualization capabilities for displaying different types of results.

A methodology for the deployment of **data retrieval and analysis** processes by means of workflow systems have been proposed (4). It foresees: i) the adoption of standard languages for data storage, representation and exchange, often in XML or JSON, ii) the availability of programmatic interfaces (usually SOAP and REST Web Services) for software interoperability, iii) the deployment of ontologies to support Web Services discovery, selection and interoperation, as well as data type characterization, iv) the creation of workflows for frequently performed analysis, to be made publicly available through user-friendly portals. This methodology limits all changes due to the evolution of databases and software to the interface level, thus facilitating maintenance of the workflows.

Data integration platforms for molecular biology and microbiology

In this section, various approaches to data integration in Life Sciences and microbiology, which had, or currently have, proven to be successful, are briefly introduced. These were selected from the many existing approaches and tools because they are the most relevant to the MIRRI conrext. In particular, they are meant to manage data from multiple sources, make it available through a unique access point, and are open to programmatic access. These are all desirable features for MIRRI.

Sequence Retrieval Software (SRS)

The Sequence Retrieval Software (SRS) was first developed in the 90's at the European Molecular Biology Laboratory (EMBL), in Heidelberg, and later in Cambridge, at the European Bioinformatics

Institute (EBI), by Thure Etzold and his group (5). It became a commercial product and was further developed until ca. 2010.

For SRS, data integration may be achieved by identifying links among data sets (usually identifiers or terms from widely used vocabularies), defining common formats for data interchange, and enabling data exchange among software. Data must be stored in a unique server as “flat files” (text only files) with predefined syntaxes. Flat files are among the most widely used format for distributing biological information and can easily be created from any database management system. Integration with external software for special data analyses is facilitated: in this case results are then treated as new databases which can then be further linked to.

SRS is not a data management system itself and it relies on other software, usually relational databases offered by the data owner, to manage the information and to create the flat files that can then be indexed by SRS. Although various add-ons have been developed to extend its functionality, it essentially remains a tool for integrating data from multiple sources.

Bio2RDF

Bio2RDF (6) is a tool developed at the University of Laval with the goal of giving integrated access to a vast number of biomedical databases through Semantic Web technologies, i.e. RDF for data archiving and SPARQL (SPARQL Protocol and RDF Query Language) for queries. To this aim, many databases have been converted to RDF by special scripts, called RDFizers, while some information systems that were already offering a viable format and interface where directly linked to the system.

This conversion was based on a unified ontology, taking care of properties included in the information resources already available in RDF. Moreover, the system provided a unified URI schema, overcoming heterogeneity of URI already provided by other systems. All major genomics, proteomics, networks and pathways, and nomenclatures databases were included in the system, as well as some clinical, e.g. Online Mendelian Inheritance in Man (OMIM), and bibliographic ones, e.g. PubMed, and the Gene Ontology.

Common Access to Biological Resources and Information (CABRI)

Common Access to Biological Resources and Information (CABRI, http://www.cabri.org/), a demonstration project funded by the European Union from 1996 to 1999, implemented a unified access to culture collection catalogues of participating collections (7). Through its online 'one-stop-shop' for biological resources, researchers can search, analyse, identify, select and pre-order strains of their interest. CABRI services, which are still maintained with more than 130,000 high quality resources representing seven distinct material types (human and animal cell lines, archaea and bacteria, filamentous fungi and yeasts, plasmids, phages, plant cells and plant viruses) from

28 collections meeting the specifications laid down in the quality control guidelines, allows the user to check for the availability of a particular biological resource by interrogating one or more catalogues at the same time.

CABRI can be seen as a pioneer model for an integrated SRS based database for distributed collections whose catalogues are made searchable together through a common gate. Structure and contents of CABRI catalogues are compiled according to well defined data sets, specific for each biological material: the Minimum Data Set (MDS), which consists of information needed to identify a unique resource in a catalogue, and the Recommended Data Set (RDS), which includes supplementary information that is useful to achieve an improved description of the characteristics, functions and properties of the material.

Guidelines for the creation of CABRI catalogues were designed with the aim of improving their structure and content similarity, so that searches involving more than one catalogue at the same time are facilitated. Data input procedures define each field of the MDS and RDS by providing a detailed textual description of its contents and by specifying the input process for the corresponding values which can be selected from reference lists of agreed values or vocabularies.

CABRI catalogues can be searched either by a simplified interface or by an almost standard SRS interface, which was adapted from the original query forms. The simple search interface simplifies the search for those users which are not proficient with SRS and it is able to manage searches involving synonyms and previous species names in an adequate way. This interface is also particularly useful for a quick retrieval of strains for which the collection number is known. It also allows a free text search on all the catalogues' contents.

BioloMICS Software

BioloMICS was first created almost 25 years ago to manage yeast collections and perform batch morphological and physiological identifications. BioloMICS is now a complete software solution for storage, management, analysis and publication of biological data and it is of choice for research or industrial laboratories, culture collections and many more (8). Many data types can be stored and handled in BioloMICS, from morphology, physiology, biochemistry, chemistry, chromatography, electrophoresis, molecular to bibliography, taxonomy, geography, ecology or administrative data. The data structure is flexible. One can easily create tables and fields (24 different field types are available) of interest on the fly. The system keeps track of all the changes made in the database. The system currently can use MySQL, MSSQL, or PostgreSQL for the underlying database. It can also use MongoDB for very large datasets.

BioloMICS offers a large number of tools to analyse morphological, physiological or sequence data. Polyphasic or multi-locus identifications are available as well as clustering tools that can produce hierarchical trees or three dimensional structures. The software also includes a Laboratory Information Management System for the management and analysis of 1st generation sequencing and associated data.

The software allows writing and fully integrates scripts using Visual Basic or C# languages. Scripts can be integrated in the existing interface allowing the extension of the functionalities of software to fit the needs of the end-users. Recently a debugger and a form designer have also been integrated. On a similar level, workflows can be created by and integrated in BioloMICS to manage or analyse data.

This software is now sold as a commercial product to a number of collections world-wide, such as CBS-KNAW, CABI, Pasteur Institute, CDC, University of California, and several more. It also includes a web publication interface that is used by a number of large international initiatives, such as MycoBank, Q-bank or the European Barcoding Database Mirror. The web interface allows basic and advanced queries on any sections of the database; a number of online data analyses tools are also available together with REST and SOAP based web services that can be consumed by specialised third parties. The latter web services are dynamic which means that they can be created on the fly by the curators of the database without any programming knowledge.

Hosting facilities are available where BioloMICS databases, websites and web services can be stored or published, removing important IT burden from the culture collections since curators, technicians or researchers only have to take care of data entries.

GCM

The Global Catalogue of Microorganisms (GCM) (http://gcm.wfcc.info/) was created in the sphere of the World Data Centre for Microorganisms (WDCM) with the aim of supporting small collections to raise to a standard data management level, according to Word Federation for Culture Collections (WFCC) requirements (9). It has now developed much further and it incorporates not only the catalogues of many CCs around the world, but it also includes many additional information elements on literature and patents, nucleotide and protein sequences, protein 3D structures. Extensive features have also been added to offer a species centered view (opposed to a strain centered one), as well as taxonomic and geographic origin views.

GCM offers an original interface, tailored on strain data and centered on user needs. The basic search can be carried out by scientific name and by strain number. An advanced search form is also available which allows queries in all GCM data fields individually. Strains can be searched all together or individually by collection.

Results are first displayed in a summary page where they can be listed by strain name and number, by collection or by isolation source. Moreover, in the same page the search can be refined by using a “facet” system, that limits the output by some characteristics, e.g. temperature range and organism type. Single strains and species are displayed in particularly information rich pages, where additional information, extracted as previously specified by external databases, is also displayed, thus forming a comprehensive image of data available on a given strain or species.

Further features include a Species Tree viewer and a Map Viewer. A taxonomic tree, built starting with data from the Species2000 International information system (http://www.sp2000.org/) is available where the number of strains for each genus is displayed. The GCM Map viewer exploits Google maps API to offer a view of the geographic distribution of strain origins. To this aim, the imprecise geographic information that is usually associated to strains is automatically translated into precise longitude and latitude data. Manual annotation has been used when automatic identification of coordinates failed. Thanks to the inclusion of nucleotide and protein sequences in GCM, sequence alignments between various strains, generally at species level, are possible. To this aim, GCM also offers a BLAST-based sequence alignment feature.

For information on strains, GCM implements the WDCM Minimum and Recommended Data Sets, which are derived from widely adopted standards such as the OECD Best Practice Guidelines for Biological Resource Centres, MINE and CABRI dataset definitions. The information included in the GCM data set are defined in general terms, with some syntactic constraints and adoption of a few reference lists, e.g. for types of organisms, and ontologies, e.g. the Metagenome and Microbes Environmental Ontology (MEO) for isolation sources.

As of September 2015, GCM included descriptions of 337,578 strains from 44,607 species, held in 74 CCs from 35 countries, mainly European and Asian (see up-to-date statistics at http://gcm.wfcc.info/StatisticgraphServlet). Catalogues can be uploaded as excel or XML files and GCM staff may manually correct the catalogue information for uniformity reasons.

StrainInfo

Data on microorganisms are scattered over data bases that contain different aspects of the microorganisms such as i) taxonomic information, ii) phylogenetic information (DNA and rRNA sequences) iii), cultivation characteristics, iv) availability of the resources, and v) published information. Furthermore, this distributed information is fragmentarily present in the various mBRCs that are internationally spread in Europe but also globally. These mBRCs offer this information on line in different formats and because descendants of a given organism (culture) can be found in more than one mBRC, data on the same biological material can be found under different strain numbers that correspond to the represented subcultures of the same material (so-called equivalent material) of the different mBRCs that have this equivalent material in their holdings. The integration of these data that is envisaged by StrainInfo applies the Nuckels-and-Nodes Approach (10).

Cumulating new data as the consequential result of scientific research in various microbial fields, are appearing in the different databases mentioned above in a dynamic process and the integration process therefore needs regular updates. In practice the participating mBRCs in StrainInfo are encouraged to share their own updates via a simple, though efficient format (MCL-Microbial Common Language) (11) that can deliver the updated data set in a semi-automated way to the StrainInfo portal that runs the integration of the information in the background.

Other International repositories of data (e.g. Genbank - http://www.ncbi.nlm.nih.gov/genbank/ -, LPSN - http://www.bacterio.net/ - List on prokaryotic names with standing in nomenclature = database on bacterial taxonomy) are linked back and forth to StrainInfo. This approach dynamically updates the cumulating information at the strain level and it also supports the visibility of the participating mBRCs that - although subsidized by national funding bodies - need to support their activities partly by their offered services.

The StrainInfo team developed the so-called StrainInfo passport (12) as a search result presenting on the fly an overview of integrated data at the strain level.

In bacterial research, the 16S rRNA gene sequence is regarded as a cornerstone for taxonomy and phylogeny of the various taxa at least at the genus and higher taxonomic levels. The lack of curation resulting in a decreased reliability of the 16S sequences of the type strains (name bearers of the bacterial species taxon) both in relation to technical aspects of the sequence and in relation to the subcultures that have been used to generate the sequences, is at the origin of many mistakes of interpretation of identifications that are generated via BLAST and FASTA algorithms that are offered as identification tools via International databases such as Genbank etc.

The SILVA database (http://www.arb-silva.de/) offers manually selected 16S sequence of high quality for all type strains of bacteria based on technical aspects of the sequence, but not on the relation between the sequence and the biological material of which genetically identical copies are deposited in various mBRCs.

StrainInfo has taken the opportunity to screen all available 16S rRNA gene sequences of the type material and select the number of ‘trustful sequences’ taking into account technical aspects, as length, ambiguities and homopolymers, following the so-called POSET approach (13). The calculated similarities per type strain offers a workable tool for checking the 16S rRNA gene sequence variability that is expected to be within the range accepted for intra species variation (14). In all other cases, the question must be raised on i) the authenticity of the biological material from which the sequences have been generated, and ii) mistakes in the information provided at the moment of deposition of the 16S rRNA gene sequence in the International databases, iii) mistakes in catalogues of mBRCs or iv) a combination of i), ii) and/or iii). This StrainInfo tool can thus be used by the mBRCs to spot confused or inaccurate information in International databases and mBRC catalogues.

Comparison of data integration tools for Life Sciences

As already stated, the issue of data integration in Life Sciences is a complex one, still largely unsolved. The first integration software aimed at coping with this issue were proposed in the 1990s'. Many systems have been proposed since then and more are still being proposed. New information and communication technologies and methodologies are driving this development effort. For this reason, it is difficult, or even impossible, to identify a tool or a technology that can singly solve the data integration issue for the aim of the MIRRI overarching infrastructure.

As a consequence, the data integration tools that are being presented and compared in this section cannot represent all possible solutions or even the kinds of solutions. However, they are considered the software that currently are the most likely to offer the best, and most widely adopted solutions for integrated access to biomedical data. As stated later, in the last section of this document, the proposed MIRRI-IS architecture should largely rely on a new system, purposely developed for the microbial resource infrastructure. Nevertheless, it will have to interoperate with existing data integration software. For this, the following presentation is of relevance for the design of the MIRRI-IS.

BioMart is a federated database including both the user interface and needed tools for interfacing, and thus federating, a generic relational database, so that a certain view of its contents can be queried and then displayed in the user portal. Alternatively, both **Taverna and Galaxy** are **bioinformatics workflow management systems**, i.e. WMSs aimed at designing, implementing and executing a complex data analysis workflow in the biomedical domain.

BioMart

BioMart is an open source federated database (15, 16). As such, it provides a unified access to distributed data sources. Due to its architectural design, which is independent from the platform and data model of the data sources, any existing database can be incorporated into its framework.

It has two alternative implementation models: server/slave and peer-to-peer. In the first model, one BioMart implementation (the portal) is the reference access point for all databases, while the other ones simply provide to the portal data for a given database. In the second model, all BioMart implementations act as portals for users and provide access to all databases' information, which is exchanged among them.

BioMart allows databases hosted on different servers to be presented seamlessly to users, facilitating collaborative projects between research groups. There is no need for the databases to be similar, neither in contents, nor in format. BioMart includes query optimization tools to efficiently manage large data sets and provides both graphical end user interfaces and APIs.

BioMart adopts the data agnostic modelling, which simplifies the time consuming data modelling task for developers. To this aim, a predefined relational schema able to represent any kind of data is used. Through the data federation approach, multiple, different and distributed databases appear to the end user as a single database, with the ability to give access to data and cross reference it from the original sources through a single user interface, without the need for building and maintaining a data warehouse.

The MartConfigurator is an additional tool, distributed with BioMart, aimed at allowing the configuration of the user interface, as well as the definition of the relationships among data sources. RESTful, SOAP and Java APIs are also available with BioMart semantic queries SPARQL.

Development and maintenance of individual databases is left to the data providers. Developers however can decide which data are exposed and which do not deliver. This introduces a strong dependency of the integrated platform both on skills and efforts devoted locally

Among the foreseen developments for BioMart are specialized ‘pre-packaged’ data portals. These would include preconfigured access to data sources and analysis tools that are useful for a given research domain. There are plans to build such portals for different research areas, such as oncology, and model organisms.

BioMart also has a large community of users and developers. It has been adopted both by academic and by industry research groups. The BioMart Central Portal is a community-driven effort able to provide unified access to many biological databases. Another 27 BioMart servers are currently available, providing more than 40 databases, none of which is directly related to microorganisms information, through a uniform web interface and standardized APIs (http://www.biomart.org/community.html).

Taverna Workbench and server

Taverna is an open source workflow management system (17), developed initially at the University of Manchester as part of the myGrid platform which is now a project within the Apache incubator.

Taverna Workbench is a standalone software for the design and enactment of scientific workflows. The Taverna workflow engine, able to enable the execution of predefined workflows, is also available, both as a command line tool and as a server.

Taverna tools are able to interoperate with a great variety of programming interfaces, including WSDL SOAP or REST Web services, BioMart, SoapLab, BioMOBY SADI, and EMBOSS, giving access to analysis tools and databases. Taverna workflows can also invoke R scripts for statistical analysis, perform various text manipulations, import spreadsheets and run local Java-like scripts.

Taverna, therefore, does not itself include any databases (information is retrieved from remote services) and does not perform any bioinformatics elaboration (analysis are carried out by remote servers). It is instead able to manage the execution of remote processes and to orchestrate them according to the description given in the designed workflows. Users can add further processes to the workflows, including their own developments, to be executed locally, but this is the exception, not the rule.

Taverna Workbench includes three main windows, devoted respectively to the workflow, which is described both as a series of components and as a graphic of interconnected nodes, to the available services, that can be selected and included in the workflow by connecting to the services already available, and the results windows, where both intermediate and final results are shown, in various formats. Taverna can monitor running workflows and allow the user to examine the provenance of the data.

Taverna also allows “pipelining and streaming” of data: services can elaborate lists of data as soon as their initial tokens are available, without the need to wait for the whole list to be downloaded. Moreover, services may be executed in parallel and looping is possible.

BioCatalogue is a public curated registry of Life Science Web services developed by implementing some of the typical features of social networks (18). Taverna workflows can be shared through the myExperiment social web site (19). BioCatalogue and myExperiment have also been developed in the context of the myGrid platform. Taverna has been tightly connected to both BioCatalogue and myExperiment. It is able to search for services described in BioCatalogue and for workflows on myExperiment. It can download, modify, upload and run workflows discovered on myExperiment.

Galaxy

Galaxy is an open-source workflow management system written in Python (20, 21). It continues to be developeded at Penn State and Johns Hopkins University, with support from the Galaxy Community, which includes users, developers, and organizations.

Galaxy simple graphical interface provides users the ability to create even complex workflows for typical data analyses (although it does not support looping constructs, contrary to Taverna) that are built by properly connecting the many available analysis tools. Galaxy also supports data uploads from various sources, including the user's computer for “personal” data and URLs for public information, and directly from some online resources, including BioMart servers. It supports many widely used data formats, including BAM, FASTA, and VCF. Galaxy also provides text manipulation utilities, allowing users to effortlessly reformat data according to their needs.

Galaxy supports reproducibility by archiving metadata information on input, intermediate and final data sets, along with parameters related to every step of the analysis, so that this can easily be repeated, in full or in part, e.g. starting from on step of the analysis. It allows researchers to share datasets, tools, and workflows either publicly or with specified users. Shared items can be copied and modified, as well as used by other users.

Unique to Galaxy is the ability to manage “histories”, that is workflow runs including input datasets, elaboration steps and related parameters, along with intermediate and output datasets. On the contrary, workflows are specifications, all steps and related parameters in the analysis, without data. So, histories allow the replication of the same experiment, while workflows allow the reuse of the same procedure. All Galaxy components can include annotations and the Galaxy Pages tool allows the creation of a sort of virtual paper that may describe in detail an experiment by combining them.

The main public implementation of Galaxy is available at https://usegalaxy.org/. This server is devoted to genomic data analysis and includes various related bioinformatics tools. Users can register and create save and share histories, workflows, and datasets on the server. But Galaxy software can also be downloaded, installed, both locally, on a server or in a cloud environment, and customized to match special needs or to serve special communities. These newly developed servers may also be made public. Galaxy is intrinsically modular and easily extensible, and new tools can be developed, integrated and possibly shared within the Galaxy ToolShed. Although it was initially developed for genomics research, it is now used in many different domains.

As a consequence of its modular and extendible architecture, many Galaxy tools have been developed and made available by different research groups. The Galaxy Tool Shed alone contains about 3,500 tools. Being possible and relatively easy to implement autonomous Galaxy servers, about 70 servers have been implemented and made available. Nowadays Galaxy has been adopted in a variety of life science domains, including genome assembly, epigenomics, transcriptomics, gene expression and proteomics.

The Galaxy Community has intense activity, suitable support and help tools, and it holds annual meetings.

Adoptability by MIRRI

Each of the three data integration tools briefly introduced above offer valuable features and present distinct peculiarities. In the context of the MIRRI preparatory project, these must be evaluated in the perspective of the MIRRI Information System.

The main characteristic of MIRRI is the presence of a great number of heterogeneous databases from mBRCs and collections. These catalogues need integration in order to present a comprehensive landscape of microorganisms held along with their biological properties; thus offering the most appropriate organism to meet the needs of users and customers of the infrastructure. At the same time, the ambition of MIRRI is to offer an integrated view of information related to microorganisms, that is included in many databases external to the mBRCs and managed by other organizations.

The perspective of the adoption of BioMart for the MIRRI-IS appears to be limited by the need to develop new *ad-hoc* **wrappers** essentially for all catalogue and for all databases of interest. This can only be achieved provided that all desired features and involved databases are identified from the beginning of the development. This would reduce the flexibility of the system and strongly limit the possible uses of the platform. Moreover, a huge effort from both a central MIRRI IT core unit and from the mBRCs IT staff would be required to build BioMart wrappers for all distinct mBRC information systems.

Both Taverna and Galaxy essentially are workflow management systems and do not represent a viable solution for the MIRRI-IS. Taverna's most interesting tool still is the workbench, i.e. the standalone version for the needs of single researchers. Taverna server does not presently offer a viable tool that could constitute the basis for the MIRRI-IS. It has recently made a great advancement by entering the Apache incubator initiative. This should increase its base of users and developers and facilitate the maintenance of its current components.

Galaxy server provides a viable option for the platform. However, it does not allow the inclusion of all mBRC catalogues and associated information systems (like BacDive, BioloMICS and StrainInfo) in a ready-to-use data repository.

This short analysis demonstrates that none of the analysed data integration tools discussed here can represent a reference system for the MIRRI-IS alone. However, these software tools are the best known and most widely adopted integration tools. So, the MIRRI-IS must be interoperable with them in order to take advantage of the large communities that already know these systems and use them in their bioinformatic analysis. An effective interoperation of the MIRRI-IS with these software would likely open the microorganisms data universe of mBRCs catalogues to these communities in a very short time.

The key for this is the development of adequate interfaces of MIRRI-IS for each of these environments. A BioMart portal for MIRRI data could easily be implemented. For the community of Taverna users, proper Web Services and valid analysis workflows can also be developed and published in the BioCatalogue.org and myExperiment.org social networks. A Galaxy server, mainly for demonstration and promotion purposes, but also as a way to support validation of catalogue data, and an adequate number of Galaxy tools, able to export mBRCs catalogue and associated data in a viable format, ready to be imported in Galaxy servers according to their various specializations, could contribute to this aim too.

The MIRRI Demonstrators

Three data integration demonstrators have been developed in the MIRRI preparatory phase. These tools have different approaches and aims to data integration. Overall, they depict a good landscape of opportunities that can be offered by MIRRI-IS and ways to reach these.

The three demonstrators have been developed to cope with distinct issues, all equally essential towards the creation of a MIRRI Information System. A short presentation of each demonstrator is reported below. For a first approach, it is essential to consider that BacDive aims at extending the contents of catalogues with a greater number of better defined data, while StrainInfo is targeted towards a better integration among collection catalogues through the identification of common strains, and the USMI Galaxy demonstrator is aimed at both supporting data curation and integrating catalogues with external resources.

The MIRRI-IS should profit from all three improvements. 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 beginning with the most recent/interesting strains.

The StrainInfo demonstrator 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 demonstrator 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.

The DSMZ Bacteria “deep characterization”

Knowledge gaps in environmental microbiology

Though usually invisible to the human eye, bacteria and archaea are omnipresent in soils, water and even our bodies. Nevertheless, in comparison to their great species variety, the information about their individual functions, interaction with higher taxa as well as their importance for ecosystem functioning is still poorly understood. Therefore it is important to mobilize, harmonize and match, scattered scientific information of environmental relevant strains distributed amongst different mBRCs and literature. This process ideally leads to seamlessly structured database content. Information can be extracted from databases and publications investigating a biological resource.

BacDive demonstrators for environmental research – towards an environmental knowledge base

Comprehensive data mobilization of five selected bacterial strains demonstrates the valuable data sources represented by mBRCs database content, publications and related research results (22). In this attempt, an average of 276 (180-319) data points per strain could be mobilized. The taxon-related descriptive information has been manually annotated from species descriptions and from results of recently completed and ongoing environmental research projects carried out at the DSMZ. The ecological and environmental relevance of these strains is given by their participation in nitrogen and carbon cycles (significant exoenzyme activity) within their particular habitat. The detailed documented metabolic and physiologic profile of each strain indicates their significant adaptation to the soils of different climates. Hence, these strains are important biotic components of their environment.

The detailed descriptions can be retrieved from the BacDive - The Bacterial Diversity Metadatabase portal (http://bacdive.dsmz.de):

Aridibacter kavangonensis Huber et al. 2014 Ac_23_E3
(http://bacdive.dsmz.de/index.php?search=24777)

Aridibacter famidurans Huber et al. 2014 A22_HD_4H
(http://bacdive.dsmz.de/index.php?site=search&rd=24776)

Blastocatella fastidiosa Fösel et al. 2013 A2_16
(http://bacdive.dsmz.de/index.php?site=search&rd=23454)

Edaphobacter aggregans Koch et al. 2008 emend. Dedysh et al. 2012 WBG 1
(http://bacdive.dsmz.de/index.php?site=search&rd=133)

Edaphobacter modestus Koch et al. 2008 JBG-1
(http://bacdive.dsmz.de/index.php?site=search&rd=132)

Although each bacterial strain might have its own significant importance within a certain microbial community, it would be too laborious to screen all its physiological, environmental and molecular biology properties on one site. This even holds true if the strain is already isolated, cultivated and deposited within mBRCs and therefore is already a target of active research.

Altogether, information remains scattered in publications and mBRCs databases around the world. A centralized MIRRI IS might facilitate locating and identifying biological resources of environmental impact. Information on associated biological resources (e.g. fungi, phages, algae, small eukaryotes etc.) in an integrative MIRRI IS would significantly increase its value as an environmental knowledge base. Finally, this would enable in-silico screening for candidates of soil rehabilitation or endangered microbial symbiotic interactions.

StrainInfo

In microbiology, the cornerstones of scientific research are the microbial strains or, even more specific, isolates. The system of mBRCs that keeps track of the microbial resources and their accompanying data has been developed over several decades. This has grown in an organic way although attempts were undertaken to harmonize the information by European initiatives such as MINE and CABRI (http://www.cabri.org/). The distribution of microbial strains, although controlled by the individual mBRCs, has led to a proliferation of research and distributed data linked to (known) subcultures of a particular microbial strain. Integration of these dispersed data is a huge task only reachable via informatics. The development of the tools themselves is one hurdle to be taken, the curation of the information that is distilled from publically available data and literature is even more troublesome.

In its so-called ‘strain passport’ (12), StrainInfo offers a single page of strain numbers relating to the various subcultures (also called equivalent strains) of a particular strain. Each strain number is linked back to the source of information, namely the mBRC catalogue that offers it (see figure 1). During the development of StrainInfo, it became apparent that users would benefit from access to further strain related information, the main categories of which are its taxonomic identification, molecular sequence information and publication information, later joined by genome project information, all of which competed for place on this single strain information page.

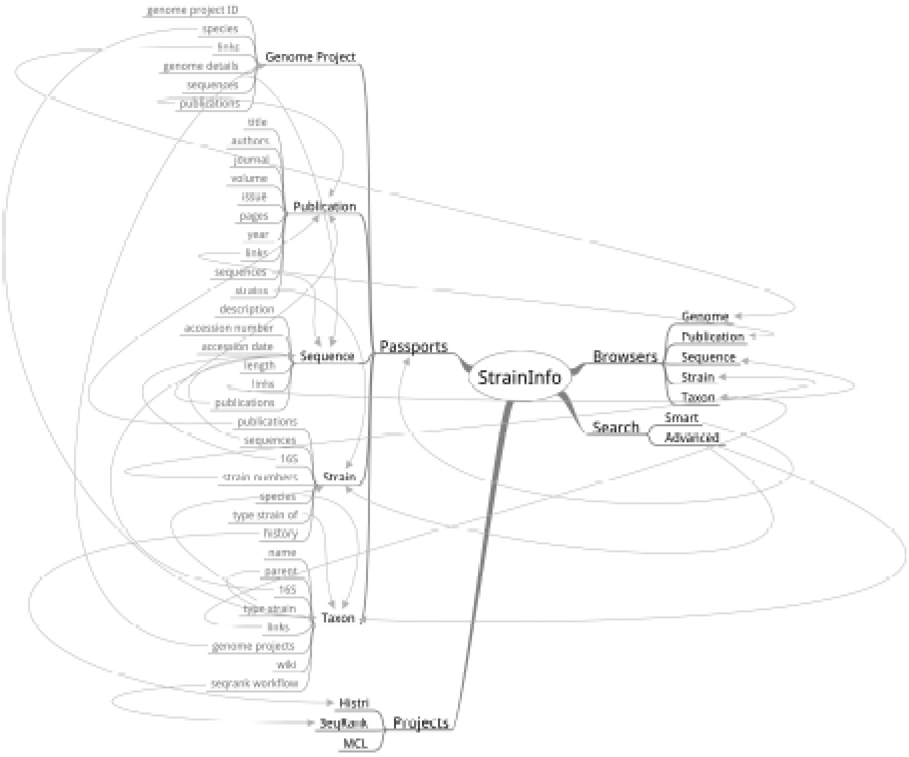
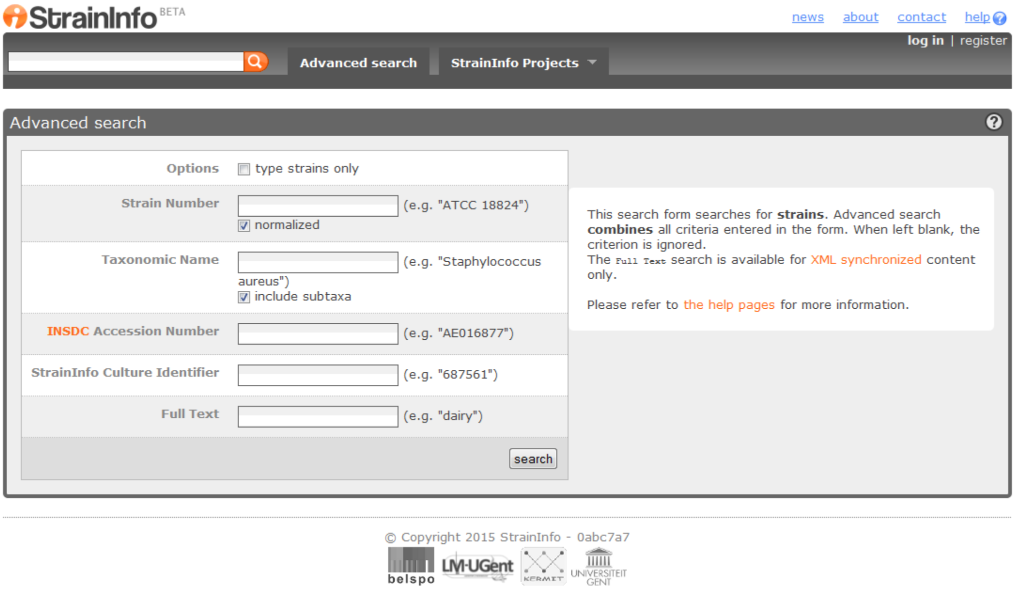


Figure 1 demonstrates the lay out of the StrainInfo portal and the extensive cross-linking between passports; each node refers to either a page in StrainInfo or an information item on that page.

StrainInfo has been set up in such a way that a push and pull system of data allows relatively easily the introduction of new or updated data from mBRCs via the MCL format (Microbiological Common Language): a standard for electronic information exchange in the Microbial Commons (11).



Advanced search

Options	□ type strains only
Strain Number	(e.g. "ATCC 18824") ☑ normalized
Taxonomic Name	(e.g. "Staphylococcus aureus") ☑ include subtaxa
INSDC Accession Number	(e.g. "AE016877")
StrainInfo Culture Identifier	(e.g. "687561")
Full Text	(e.g. "dairy")
	search

This search form searches for **strains**. Advanced search **combines** all criteria entered in the form. When left blank, the criterion is ignored.
The Full Text search is available for XML synchronized content only.

Please refer to the help pages for more information.

© Copyright 2015 StrainInfo - 0abc7a7

Figure 2: The StrainInfo advanced search tool interface.

The advanced search tool (see figure 2), although only partially developed, allows the retrieval of answers to questions such as:

- Which strains have a link with Phosphate?
- Which strains have a link with “dairy”?
- Which strains have a link with “dairy” and “cow’?

The initial examples show that further development is needed. A standardized information format as the basis of MCL is a prerequisite for an efficient search tool that is of practical use for both the academic and the bio-industry.

Furthermore, as StrainInfo provides integrated information at the strain level, the history of strain exchange between culture collections can be reconstructed. The integration of data allows tracing of discrepancies in data between records of subcultures distributed in various mBRCs (23).The principle of curation of sequence data, particularly their links with the so-called equivalent strains deserves our attention. In this context StrainInfo already allows the tracing of incompatibilities of 16S sequences of type strains that can be used by the mBRCs to verify the authenticity of their holdings (24).

The USMI Galaxy Demonstrator

Integration of mBRCs catalogues information with data from taxonomic, genomics, proteomics, metabolomics and literature information is fundamental. Also important for catalogue curators is the availability of tools supporting accurate and effective annotation of strains. The USMI Galaxy Demonstrator (UGD) (25) aims to support both mBRC staff and researchers to perform tasks that:

i) find, query and merge catalogues of microbial resources, thus offering a comprehensive vision of available strains and associated data,

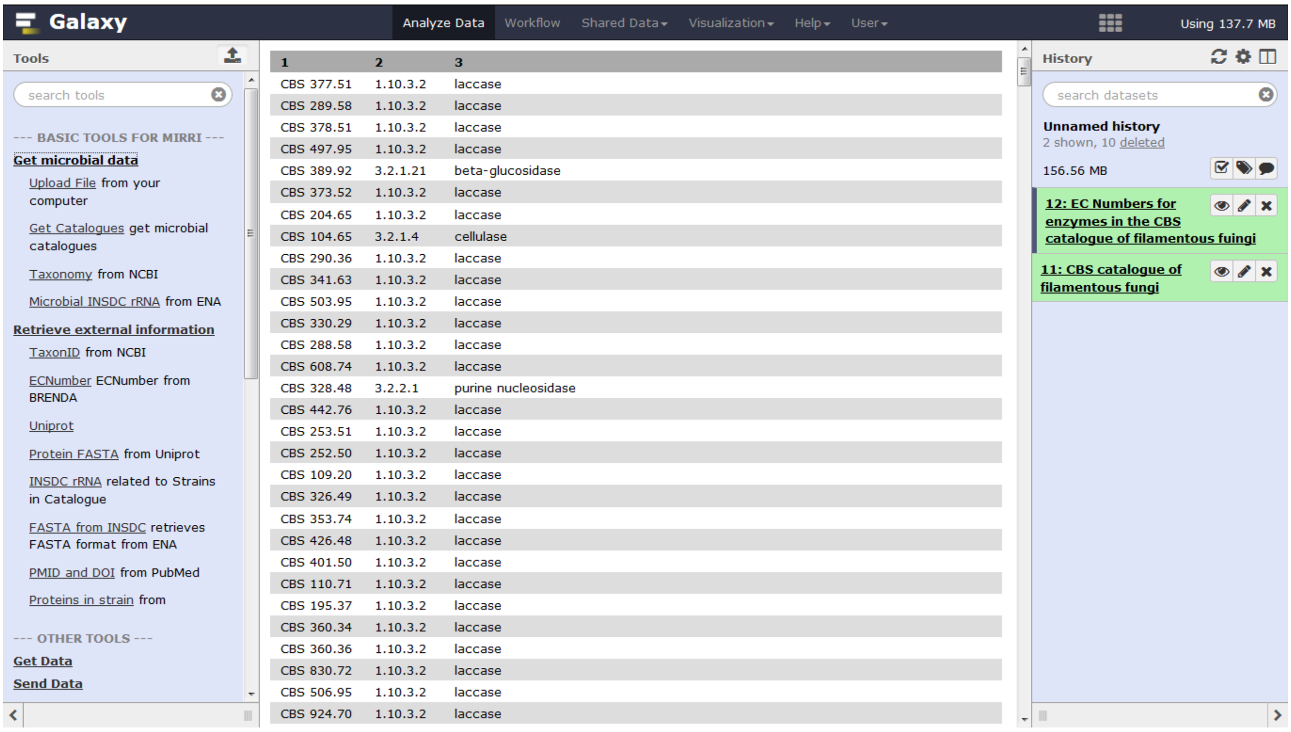
ii) support insertion and update of information in catalogues according to agreed procedures,

iii) support curators to improve catalogue data by revising contents, identifying relevant data in remote databases and inserting their links into the catalogues,

iv) make available catalogue data to bioinformatic data analysis workflows.

As described in a previous section, Galaxy is a public web-based data and analysis integration framework, available for installation in a workstation for personal use as well as in a web server for use by a community of researchers. Galaxy has been installed in the USMI bioinformatics web server and is available at http://bioinformatics.hsanmartino.it:8080/. It presently includes various tools (i.e., in this context, Galaxy scripts providing some sort of elaboration) provided by Galaxy developers, some tools related to mBRC catalogue data that were purposely developed, and a few data analysis workflows for demonstration purposes.

The graphical interface of Galaxy has two lateral panels (see figure 3): the 'Tools panel' on the left, where all tools are sorted and grouped by topics, and the 'History panel' on the right, where all performed tasks are listed. The central part of the interface is reserved for the interaction with the user: it is used both for displaying results and for inserting query related parameters.


Galaxy
Analyze Data
Workflow
Shared Data
Visualization
Help
User
Using 137.7 MB
Tools
search tools
--- BASIC TOOLS FOR MIRRI ---
Get microbial data
Upload File from your computer
Get Catalogues get microbial catalogues
Taxonomy from NCBI
Microbial INSDC rRNA from ENA
Retrieve external information
TaxonID from NCBI
ECNumber ECNumber from BRENDA
Uniprot
Protein FASTA from Uniprot
INSDC rRNA related to Strains in Catalogue
FASTA from INSDC retrieves FASTA format from ENA
PMID and DOI from PubMed
Proteins in strain from
--- OTHER TOOLS ---
Get Data
Send Data
History
search datasets
Unnamed history
2 shown, 10 deleted
156.56 MB
12: EC Numbers for enzymes in the CBS catalogue of filamentous fuingi
11: CBS catalogue of filamentous fungi

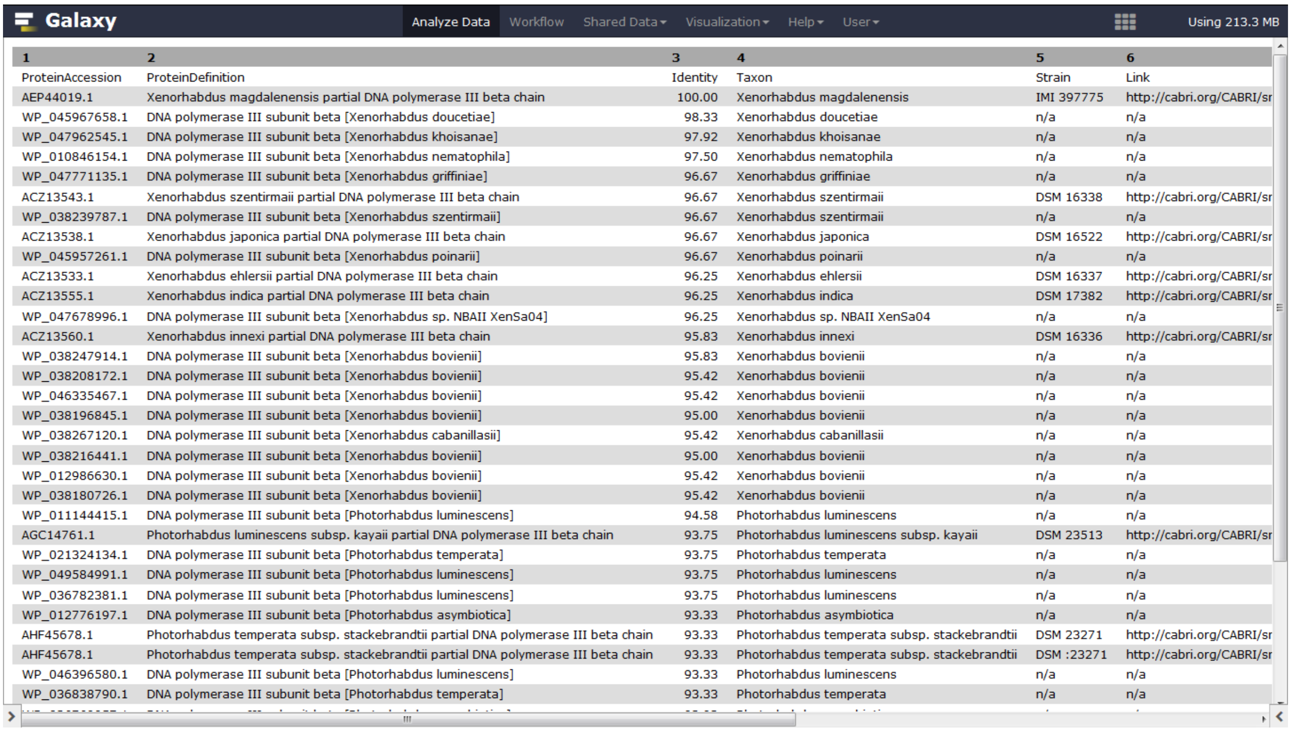
1	2	3
CBS 377.51	1.10.3.2	laccase
CBS 289.58	1.10.3.2	laccase
CBS 378.51	1.10.3.2	laccase
CBS 497.95	1.10.3.2	laccase
CBS 389.92	3.2.1.21	beta-glucosidase
CBS 373.52	1.10.3.2	laccase
CBS 204.65	1.10.3.2	laccase
CBS 104.65	3.2.1.4	cellulase
CBS 290.36	1.10.3.2	laccase
CBS 341.63	1.10.3.2	laccase
CBS 503.95	1.10.3.2	laccase
CBS 330.29	1.10.3.2	laccase
CBS 288.58	1.10.3.2	laccase
CBS 608.74	1.10.3.2	laccase
CBS 328.48	3.2.2.1	purine nucleosidase
CBS 442.76	1.10.3.2	laccase
CBS 253.51	1.10.3.2	laccase
CBS 252.50	1.10.3.2	laccase
CBS 109.20	1.10.3.2	laccase
CBS 326.49	1.10.3.2	laccase
CBS 353.74	1.10.3.2	laccase
CBS 426.48	1.10.3.2	laccase
CBS 401.50	1.10.3.2	laccase
CBS 110.71	1.10.3.2	laccase
CBS 195.37	1.10.3.2	laccase
CBS 360.34	1.10.3.2	laccase
CBS 360.36	1.10.3.2	laccase
CBS 830.72	1.10.3.2	laccase
CBS 506.95	1.10.3.2	laccase
CBS 924.70	1.10.3.2	laccase

Figure 3: Output of the UGD EC Number tool, showing EC numbers of enzymes reported in the CBS catalogue of filamentous fungi.

In UGD, tools related to microbial collections is available under the label 'Basic Tool for MIRRI', which includes two sections: 'Get microbial data' and 'Retrieve external information'. The first section includes tools, which import datasets, both catalogues and subsets of external databases, for later analysis in the session. Among them, the 'Get Catalogues' tool allows the import of catalogues. These are stored in MCL in a purpose web repository (presently, about 30 catalogues are included). Other tools allow retrieving information from the NCBI Taxonomy database and from ENA.

The second section, named 'Retrieve external information', includes some tools which gather microorganism related information from external databases, thus supporting insertion of proper identifiers and codes in the catalogues. E.g., the 'ECNumber' tool gathers EC numbers from BRENDA (BRaunschweig ENzyme DAtabase) (26) (http://www.brenda-enzymes.org) for enzyme names found in a microbial catalogue.

All available tools may be combined to set up a data analysis workflow. For demonstration purposes, one tool and one workflow are presented here. The tool is meant to support mBRC staff in the retrieval of EC numbers for enzymes produced by strains in the collection. Once a catalogue is imported, the tool reads all strain descriptions for enzyme production in the catalogue and then connects to BRENDA and retrieves the corresponding identifiers. This is shown in the figure 3.



1	2	3	4	5	6
ProteinAccession	ProteinDefinition	Identity	Taxon	Strain	Link
AEP44019.1	Xenorhabdus magdalenensis partial DNA polymerase III beta chain	100.00	Xenorhabdus magdalenensis	IMI 397775	http://cabri.org/CABRI/sr
WP_045967658.1	DNA polymerase III subunit beta [Xenorhabdus doucetiae]	98.33	Xenorhabdus doucetiae	n/a	n/a
WP_047962545.1	DNA polymerase III subunit beta [Xenorhabdus khoisanae]	97.92	Xenorhabdus khoisanae	n/a	n/a
WP_010846154.1	DNA polymerase III subunit beta [Xenorhabdus nematophila]	97.50	Xenorhabdus nematophila	n/a	n/a
WP_047771135.1	DNA polymerase III subunit beta [Xenorhabdus griffiniae]	96.67	Xenorhabdus griffiniae	n/a	n/a
ACZ13543.1	Xenorhabdus szentirmaii partial DNA polymerase III beta chain	96.67	Xenorhabdus szentirmaii	DSM 16338	http://cabri.org/CABRI/sr
WP_038239787.1	DNA polymerase III subunit beta [Xenorhabdus szentirmaii]	96.67	Xenorhabdus szentirmaii	n/a	n/a
ACZ13538.1	Xenorhabdus japonica partial DNA polymerase III beta chain	96.67	Xenorhabdus japonica	DSM 16522	http://cabri.org/CABRI/sr
WP_045957261.1	DNA polymerase III subunit beta [Xenorhabdus poinarii]	96.67	Xenorhabdus poinarii	n/a	n/a
ACZ13533.1	Xenorhabdus ehlersii partial DNA polymerase III beta chain	96.25	Xenorhabdus ehlersii	DSM 16337	http://cabri.org/CABRI/sr
ACZ13555.1	Xenorhabdus indica partial DNA polymerase III beta chain	96.25	Xenorhabdus indica	DSM 17382	http://cabri.org/CABRI/sr
WP_047678996.1	DNA polymerase III subunit beta [Xenorhabdus sp. NBAII XenSa04]	96.25	Xenorhabdus sp. NBAII XenSa04	n/a	n/a
ACZ13560.1	Xenorhabdus innexi partial DNA polymerase III beta chain	95.83	Xenorhabdus innexi	DSM 16336	http://cabri.org/CABRI/sr
WP_038247914.1	DNA polymerase III subunit beta [Xenorhabdus bovienii]	95.83	Xenorhabdus bovienii	n/a	n/a
WP_038208172.1	DNA polymerase III subunit beta [Xenorhabdus bovienii]	95.42	Xenorhabdus bovienii	n/a	n/a
WP_046335467.1	DNA polymerase III subunit beta [Xenorhabdus bovienii]	95.42	Xenorhabdus bovienii	n/a	n/a
WP_038196845.1	DNA polymerase III subunit beta [Xenorhabdus bovienii]	95.00	Xenorhabdus bovienii	n/a	n/a
WP_038267120.1	DNA polymerase III subunit beta [Xenorhabdus cabanillasii]	95.42	Xenorhabdus cabanillasii	n/a	n/a
WP_038216441.1	DNA polymerase III subunit beta [Xenorhabdus bovienii]	95.00	Xenorhabdus bovienii	n/a	n/a
WP_012986630.1	DNA polymerase III subunit beta [Xenorhabdus bovienii]	95.42	Xenorhabdus bovienii	n/a	n/a
WP_038180726.1	DNA polymerase III subunit beta [Xenorhabdus bovienii]	95.42	Xenorhabdus bovienii	n/a	n/a
WP_011144415.1	DNA polymerase III subunit beta [Photorhabdus luminescens]	94.58	Photorhabdus luminescens	n/a	n/a
AGC14761.1	Photorhabdus luminescens subsp. kayaii partial DNA polymerase III beta chain	93.75	Photorhabdus luminescens subsp. kayaii	DSM 23513	http://cabri.org/CABRI/sr
WP_021324134.1	DNA polymerase III subunit beta [Photorhabdus temperata]	93.75	Photorhabdus temperata	n/a	n/a
WP_049584991.1	DNA polymerase III subunit beta [Photorhabdus luminescens]	93.75	Photorhabdus luminescens	n/a	n/a
WP_036782381.1	DNA polymerase III subunit beta [Photorhabdus luminescens]	93.75	Photorhabdus luminescens	n/a	n/a
WP_012776197.1	DNA polymerase III subunit beta [Photorhabdus asymbiotica]	93.33	Photorhabdus asymbiotica	n/a	n/a
AHF45678.1	Photorhabdus temperata subsp. stackebrandtii partial DNA polymerase III beta chain	93.33	Photorhabdus temperata subsp. stackebrandtii	DSM 23271	http://cabri.org/CABRI/sr
AHF45678.1	Photorhabdus temperata subsp. stackebrandtii partial DNA polymerase III beta chain	93.33	Photorhabdus temperata subsp. stackebrandtii	DSM :23271	http://cabri.org/CABRI/sr
WP_046396580.1	DNA polymerase III subunit beta [Photorhabdus luminescens]	93.33	Photorhabdus luminescens	n/a	n/a
WP_036838790.1	DNA polymerase III subunit beta [Photorhabdus temperata]	93.33	Photorhabdus temperata	n/a	n/a

Figure 4: Output of the workflow “From Protein sequence to Strain Number”, showing identifiers of microbial resources with information on similar proteins. The last column reports a link to CABRI network services for extended strain information and pre-order. Lateral panels have been hidden in this figure.

The workflow “From Protein sequence to Strain Number” is targeted to the needs of a researcher willing to retrieve information on strains, which are the source for some proteins and have shown similarity to a given polypeptide. The idea is that after a similarity search the researcher maybe interested in retrieving information on the source strains of proteins. To this aim, the user is asked to submit a short protein sequence. This is used to find similar proteins by using a Galaxy basic tool (various parameters can be used to tune this step according to the user needs). The returned list of proteins is then exploited to check if the related source material is available from the mBRC catalogues and in this case the relative information is also returned. This is shown in the figure 4.

Conclusion

In this report, we focused on the analysis of some of the fundamental issues related to the development of the MIRRI-IS. This platform will allow the exploitation of all available information from mBRC catalogues, together with various databases and tools which are maintained outside the MIRRI infrastructure. It will enable an innovative downstream data analysis of microbiological information in various application domains. The analysis of data integration methodologies and tools for life sciences allows us to identify some of the main ideas for the creation of the MIRRI-IS.

The MIRRI Information System should represent a portal for accessing all mBRC catalogues. This information should be represented in a uniform format. It should also include extended annotations on strain characteristics, beyond information that is usually included in catalogues, as they can only be provided by specialized information systems. This information should be of high-quality and it should be linked to many relevant data and service providers, external to MIRRI.

Information exchange between mBRCs and MIRRI should be based on a standard data exchange language. An extended version of the Microbiological Common Language (MCL) able to represent the whole richness and complexity of mBRCs catalogues should be adopted both for data exchange and repository. An initial proposal in this direction has been presented in MIRRI deliverable D8.3.

The information from mBRCs should only constitute a central core of the MIRRI-IS. As shown in figure 5 (27), information from other specialized systems, like StrainInfo, BacDive, BioloMICS and SILVA (the European reference database for ribosomal RNA gene sequences), can constitute an important add-on for the MIRRI-IS, providing highly characterized information relevant for various life science domains. Similarly, information from prominent external data providers, like NCBI and EBI for sequence, literature, and taxonomic data, should be linked in order to provide both static data, to be included in the MIRRI-IS, and on-the-fly data, according to users requests.

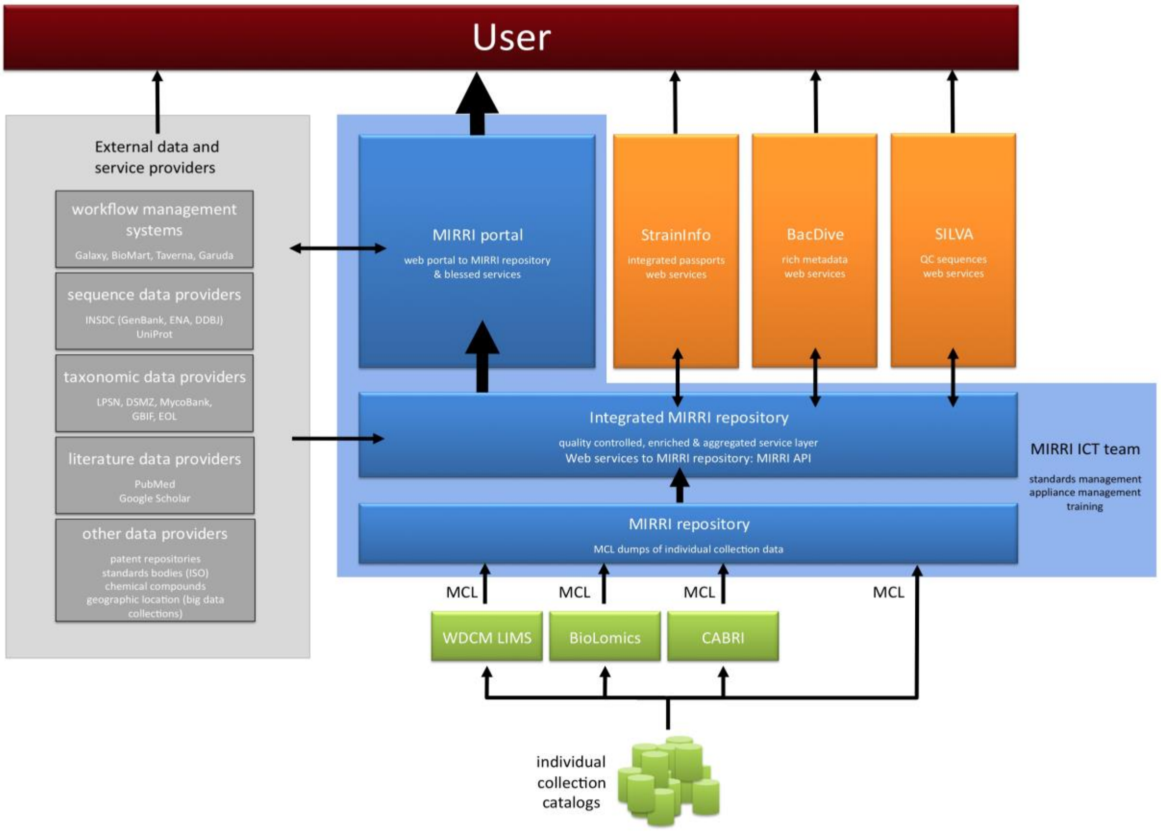


Figure 5: A possible architecture for MIRRI-IS.

The web interface of MIRRI-IS will be essential for the success of the MIRRI portal, but Application Programming Interfaces (APIs) will be even more important, also in the view of the planned MIRRI Collaborative Working Environment (CWE).

The CWE will enable an integrated view of, and access to, MIRRI services, including not only strain and associated data from the MIRRI-IS, but also expert clusters, events, training, and various services. The best way to implement such a interlinked system is to build independent systems and an umbrella software able to query all of them and report to the user the requested information. MIRRI-IS APIs will then constitute the natural way to make access and retrieve data from the MIRRI-IS in an effective, yet flexible, way.

The development and implementation of adequate APIs for software platforms like BioMart, Taverna and Galaxy could also allow a huge increase both in the awareness of researchers regarding the availability of high quality information on microorganisms, and in the effective utilization of the integrated mBRC catalogues.

For this, appropriate additional servers should be built. A first hypothesis may relate to a devoted Galaxy server, a BioMart portal, and the submission of a number of Galaxy tools and workflows, and of Taverna workflows in the relative repositories, like the Galaxy ToolShed, Biocatalogue.org, and myExperiment.org.

Semantic Web technologies, which probably are not yet mature enough to constitute a “first choice” option for MIRRI-IS, are however developing rapidly. Initiatives like Bio2RDF and the EBI RDF library are showing that there is an opportunity for the creation of a uniform, semantic aware, linked data approach for Life Sciences data. For this reason, the MIRRI APIs should also be ready to provide RDF datasets in line with the most updated ontological definitions.

References

1. Michael Y. Galperin MY, Rigden DJ, Fernández-Suárez XM. The 2015 Nucleic Acids Research Database Issue and Molecular Biology Database Collection. Nucl. Acids Res. 2015, 43(D1):D1-D5. (doi: 10.1093/nar/gku1241)
2. Taylor CF, Field D, Sansone S-A, et al. Promoting coherent minimum reporting guidelines for biological and biomedical investigations: the MIBBI project. Nat Biotech 2008, 26(8):889-896. (doi: 10.1038/nbt.1411)
3. Smith B, Ashburner M, Rosse C, et al. The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. Nat Biotech 2007, 25(11):1251-1255. (doi:10.1038/nbt1346)
4. Romano P. Automation of in-silico data analysis processes through workflow management systems. Briefings in Bioinformatics 2008 9(1):57-68. (PMID: 18056132; doi:10.1093/bib/bbm056)
5. Etzold T, Ulyanov A, Argos P. SRS: information retrieval system for molecular biology data banks. Methods Enzymol 1996;266:114-28.

6. Belleau F, Nolin M-A, Tourigny N, Rigault P, MorissetteJ: Bio2RDF: Towards a mashup to build bioinformatics knowledge systems. Journal of Biomedical Informatics 2008, 41:706–716.
7. Romano P, Kracht M, Manniello MA, Stegehuis G, Fritze D. The role of informatics in the coordinated management of biological resources collections, Applied Bioinformatics. 2005;4(3):175-186
8. Robert V, Szoke S, Jabas B, Vu D, Chouchen O, Blom E, Cardinali C. BioloMICS software: Biological data management, identification, classification and statistics. Open Applied Informatics Journal 2011 (5):87-98.
9. Wu L, Sun Q, Sugawara H, Yang S, Zhou Y, McCluskey K, Vasilenko A, Suzuki K-I, Ohkuma M, Lee Y, Robert V, Ingsriswang S, Guissart F, Desmeth P, Ma J. Global catalogue of microorganisms (gcm): a comprehensive database and information retrieval, analysis, and visualization system for microbial resources, BMC Genomics 2013, 14:933
10. Van Brabant B, Dawyndt P, De Baets B, De Vos P. A knuckels-and-nodes approach to the integration of Microbial Resource data. Lecture Notes in Computer Science. LNCS 4277. 2006. pp. 740-750.
11. Verslyppe B, Kottmann R, De Smet W, De Baets B, De Vos P, Dawyndt P: Microbiological Common Language (MCL): a standard for electronic information exchange in the Microbial Commons. Research in microbiology 2010, 161(6):439-445.
12. Verslyppe B, De Smet W, De Baets B, De Vos P, Dawyndt P: Make Histri: reconstructing the exchange history of bacterial and archaeal type strains. Systematic and applied microbiology 2011, 34(5):328-336.
13. De Smet W. Explicit sequence-culture-strain-taxon links in StrainInfo and their role in quality assessment and assurance. 2013. PhD dissertation. ISBN 9789461971623
14. Kim M, Oh H-S, Park S-C, Chun C. Towards a taxonomic coherence between average nucleotide identity and 16S rRNA gene sequence similarity for species demarcation of prokaryotes. Int. J. Syst. Evol. Microbil. 2014 64:346-351.
15. Smedley D, Haider S, Durinck S, et al. The BioMart community portal: an innovative alternative to large, centralized data repositories. Nucleic Acids Res 2015, 43(W1):W589-598. (PMID: 25897122; doi: 10.1093/nar/gkv350)
16. Kasprzyk A. BioMart: driving a paradigm change in biological data management. Database (Oxford). 2011 2011:bar049 (PMID: 22083790; doi: 10.1093/database/bar049)
17. Oinn T, Addis M, Ferris J et al. Taverna: a tool for the composition and enactment of bioinformatics workflows. Bioinformatics 2004;20:3045-54.
18. Bhagat J, Tanoh F, Nzuobontane E, Laurent T, Orlowski J, Roos M, Wolstencroft K, Aleksejevs S, Stevens R, Pettifer S, Lopez R, Goble CA. BioCatalogue: a universal catalogue of web services for the life sciences, Nucl. Acids Res. 2010, 38 (Suppl 2): W689-W694. doi:10.1093/nar/gkq394
19. Goble CA, Bhagat J, Aleksejevs S, Cruickshank D, Michaelides D, Newman D, Borkum M, Bechhofer S, Roos M, Li P, De Roure D. myExperiment: a repository and social network for the sharing of bioinformatics workflows. Nucl. Acids Res. 2010, 38 (Suppl 2): W677-W682. doi:10.1093/nar/gkq429]
20. Goecks J, Nekrutenko A, Taylor J, Galaxy Team. Galaxy: A comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. Genome Biology 2010, 11(8): R86 (PMID: 20738864; doi: 10.1186/gb-2010-11-8-r86)
21. Blankenberg D, Coraor N, Von Kuster G, Taylor J, Nekrutenko A, Galaxy Team. Integrating diverse databases into an unified analysis framework: A Galaxy approach. Database 2011, 2011:bar011 (PMID: 21531983; doi: 10.1093/database/bar011)

22. Söhngen C, Bunk B, Podstawka A, Gleim D, Overmann J. BacDive - the Bacterial Diversity Metadatabase. Nucleic Acids Res. 2014 42(1):D592-599 (PMID: 24214959; doi: 10.1093/nar/gkt1058)
23. Verslyppe B, De Smet W, De Baets B, De Vos P, Dawyndt P: StrainInfo introduces electronic passports for microorganisms. Systematic and applied microbiology 2014, 37(1):42-50.
24. De Smet W, De Loof K, De Vos P, Dawyndt P, De Baets B: Filtering and ranking techniques for automated selection of high-quality 16S rRNA gene sequences. Systematic and applied microbiology 2013, 36(8):549-559.
25. Colobraro DP, Romano P. A Galaxy approach to integrate microbial data: the USMI Galaxy demonstrator. Proceedings of BITS 2015, 11th Annual Meeting of the Italian Society of Bioinformatics, 3-5 June, 2015, Milano. Guffanti A, Masseroli M, Milanesi L (eds) (in press)
26. Schomburg et al. Enzyme data and metabolic information: BRENDA, a resource for research in biology, biochemistry, and medicine. Gene Funct Dis 3 (4): 109–118.
27. Dawyndt P, de Vos P, Bunk B, Söhngen C, Overmann J. Towards an integrated portal of biological resource centers. (position paper, personal communication)