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

D8.3

Report on workshop and surveys on current collection status of data management systems, metadata sources, semantic systems, and molecular, taxonomy, nomenclature, bibliography and environmental databases to be considered by the platform

Deliverable Date: October 2014
Actual Submission Date: November 2014
Lead Beneficiary: Partner 10 - JacobsUni
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 a result of the activity of work package 8, task 8.3, and it is tightly linked to the previous reports from the same work package, namely the report on minimum standards for data acquisition and data management and mechanisms to incentivize the deposit of quality data (D8.1) and the report on users' requests, desired features, and meta-analyses of the integrated platform (D8.5). These reports were taken into account while preparing this deliverable, in order to avoid needless repetitions and to better specify correspondences and synergies.

In the first report, deliverable D8.1, the MIRRI Information System (MIRRI-IS) vision and strategy was introduced. It describes how "*MIRRI will distinguish itself*" on the basis of the following features (which are here rephrased):

i) ensure high data quality by an intensive data curation activity,

ii) guarantee data integration across mBRCs/CCs and interoperability of related systems,

iii) provide an open platform for downstream data analysis and product development,

iv) establish complementarity with other disciplines, applying appropriate data structure and ontologies.

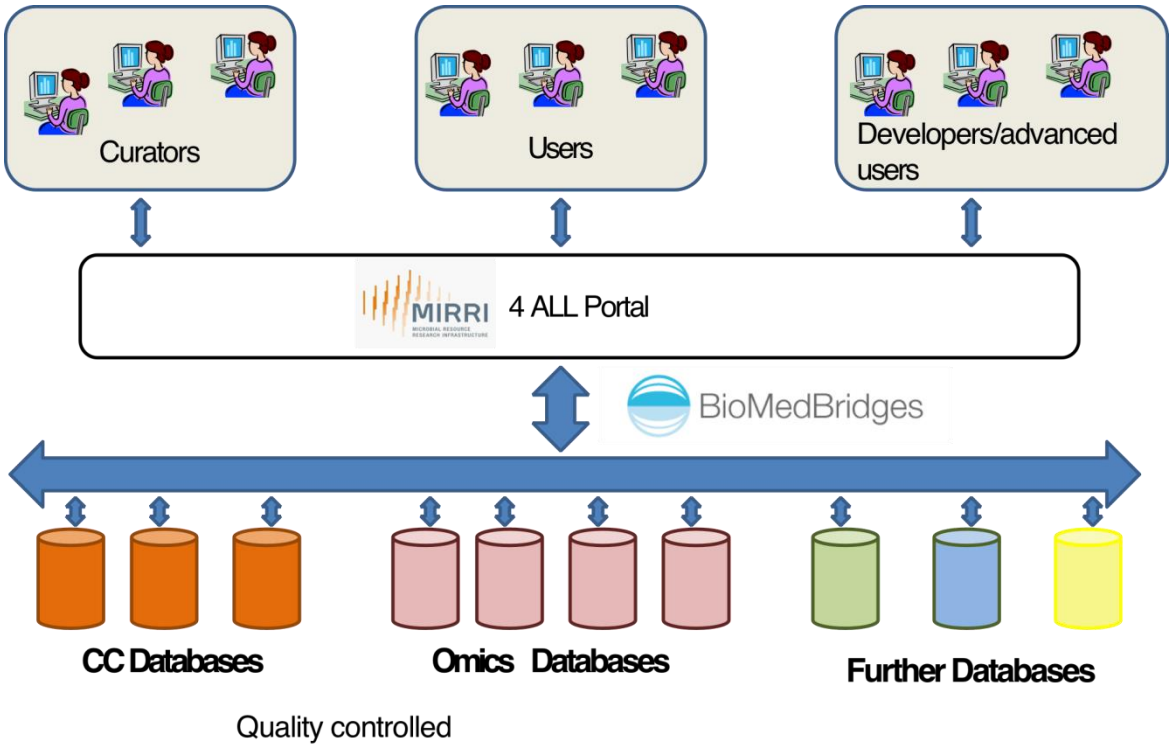


Figure 1: The MIRRI Information System (MIRRI-IS) providing access to quality controlled information from culture collections databases and integrating this data with "-omics" databases, as well as databases of other related domains, to the benefit of a varied community of users with different needs and skills.

Conclusions of deliverable D8.5 pointed out how the main obstacle for the building of the MIRRI-IS refers to "*defining a core set of data fields across BRCs invoking ontology/controlled vocabularies*

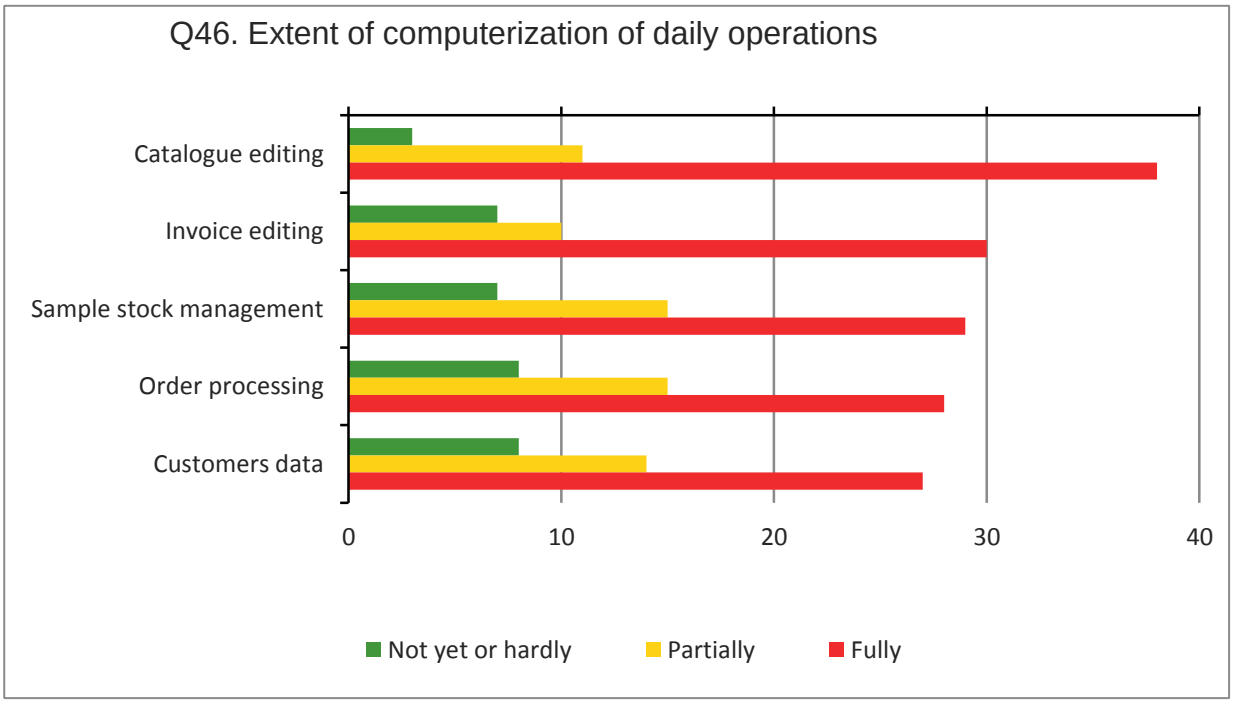
where necessary and feasible". This, essentially, is the core of this report. In this deliverable, we have mainly faced the following issues:

- the status of current culture collection data management systems,
- the present guidelines for culture collection data,
- the semantic resources that can support both an improved management of and an advanced access to culture collection data,
- the information systems that relate to the microbiological domain and should interoperate with the MIRRI-IS.

The status of current culture collection data management systems

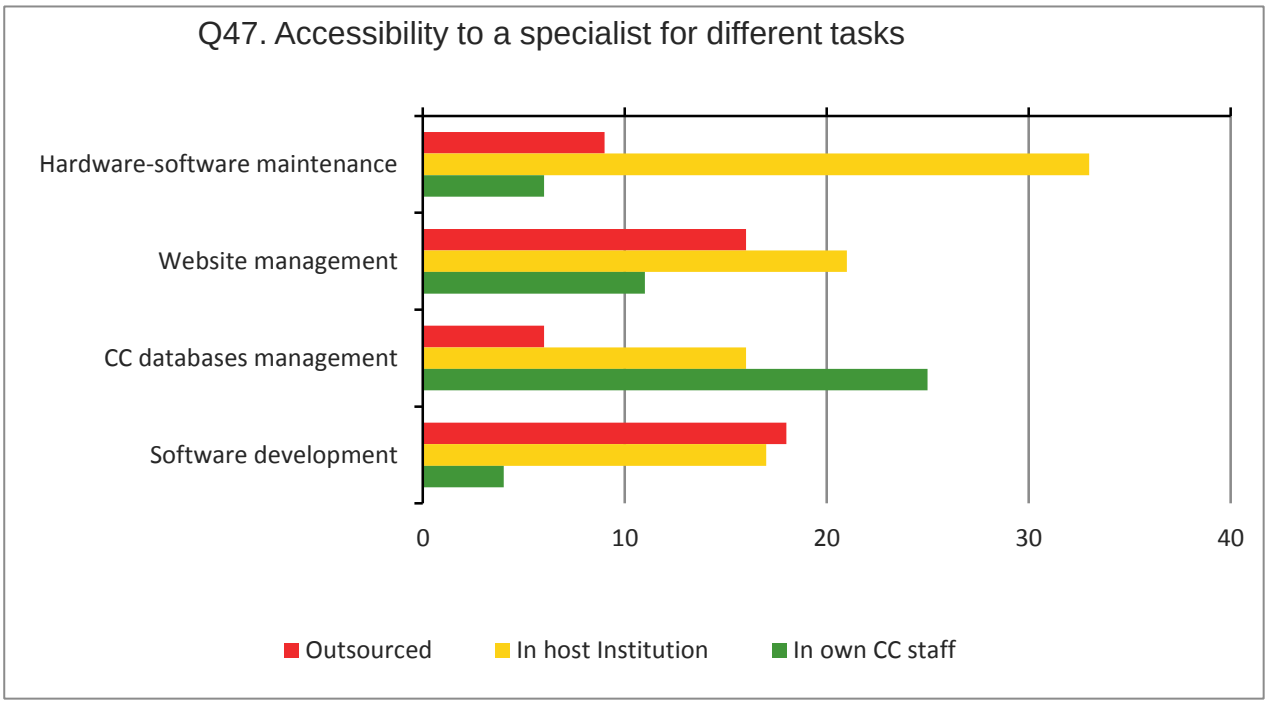
In the questionnaire for Culture Collections prepared by MIRRI WP2, which was presented on June 12th 2013 at the MIRRI meeting in Athens, a special section was devoted to data management by culture collections (section 6, questions 46 – 55). The analysis of results of this section was a required starting point for the assessment of the status of data management systems by culture collections.

As to the extent of computerization of daily operations, it turned out, as expected, that the activities related to catalogue editing was fully accomplished by software tools for 38 collections out of 52, with another 11 collections accomplishing it partially by computer tools (see Q46).

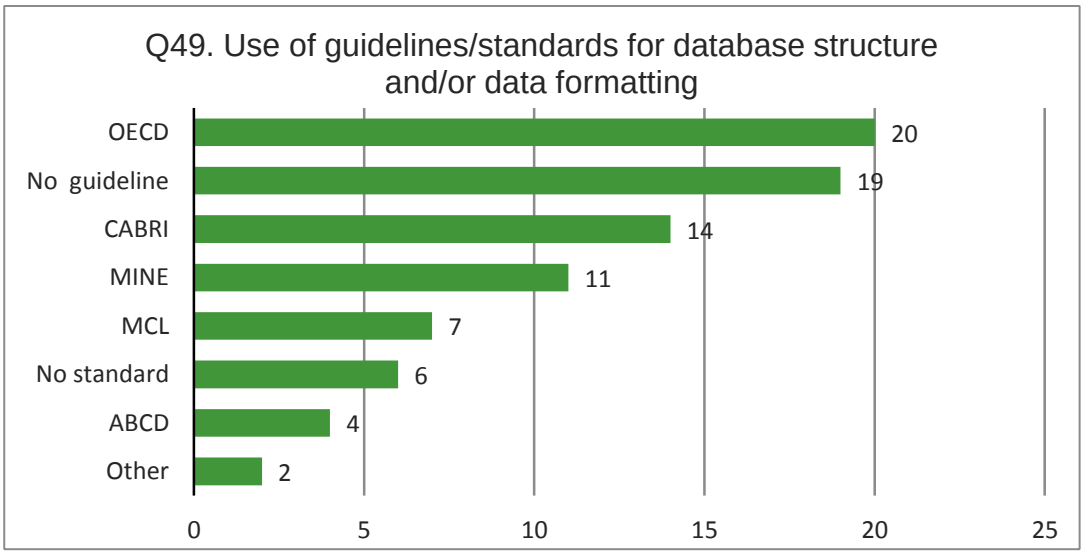


Access to specialists external to the collection is mainly performed for hardware-software maintenance, for which support is mainly sought by the host institution services, for website management, which is delegated both to host institute services and to external services, and for software development, which is required by a lesser number of collections (see Q47). In this

context, the only activity which is usually and predominately carried out by the culture collection staff is the management of the database.

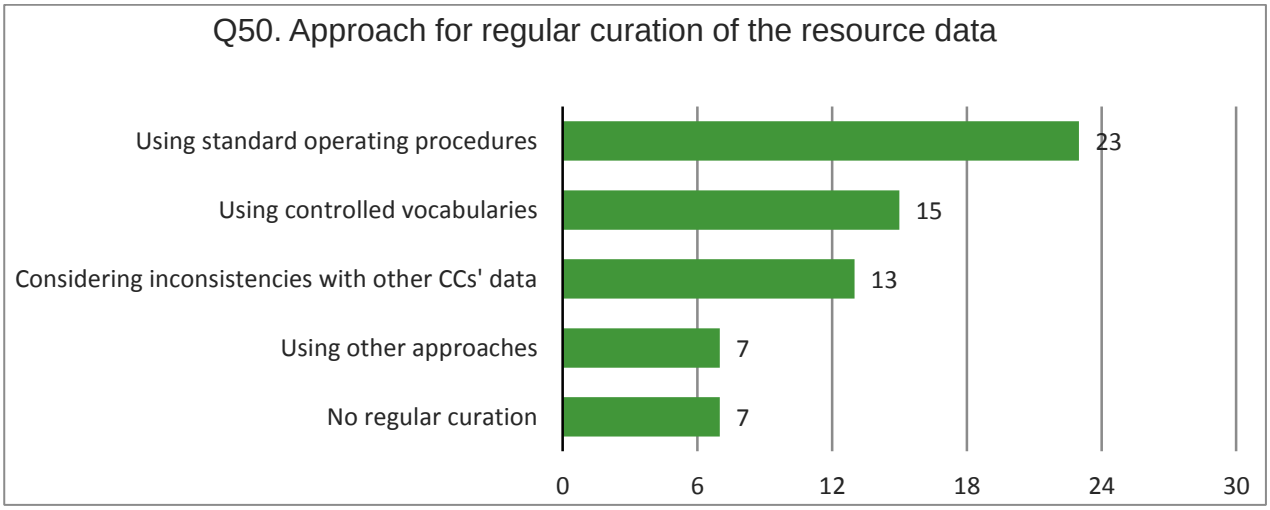


With reference to the adoption of guidelines or standards for database structure and/or data formatting (see Q49), a great heterogeneity among collections was shown. It is noteworthy that four out of 50 collections have declared that they do not use any standard for data management, and that 19 others have declared that they do not apply any specific guideline of the microbiological domain. This is even more surprising if we consider that the large majority of collections are aware of different international operational guidelines (e.g., 51 out of 52 collections are aware of OECD general best practice guidelines for mBRCs, see Q67-69-71) and all collections are involved in international networks, and are members of ECCO and/or WFCC (see Q66).



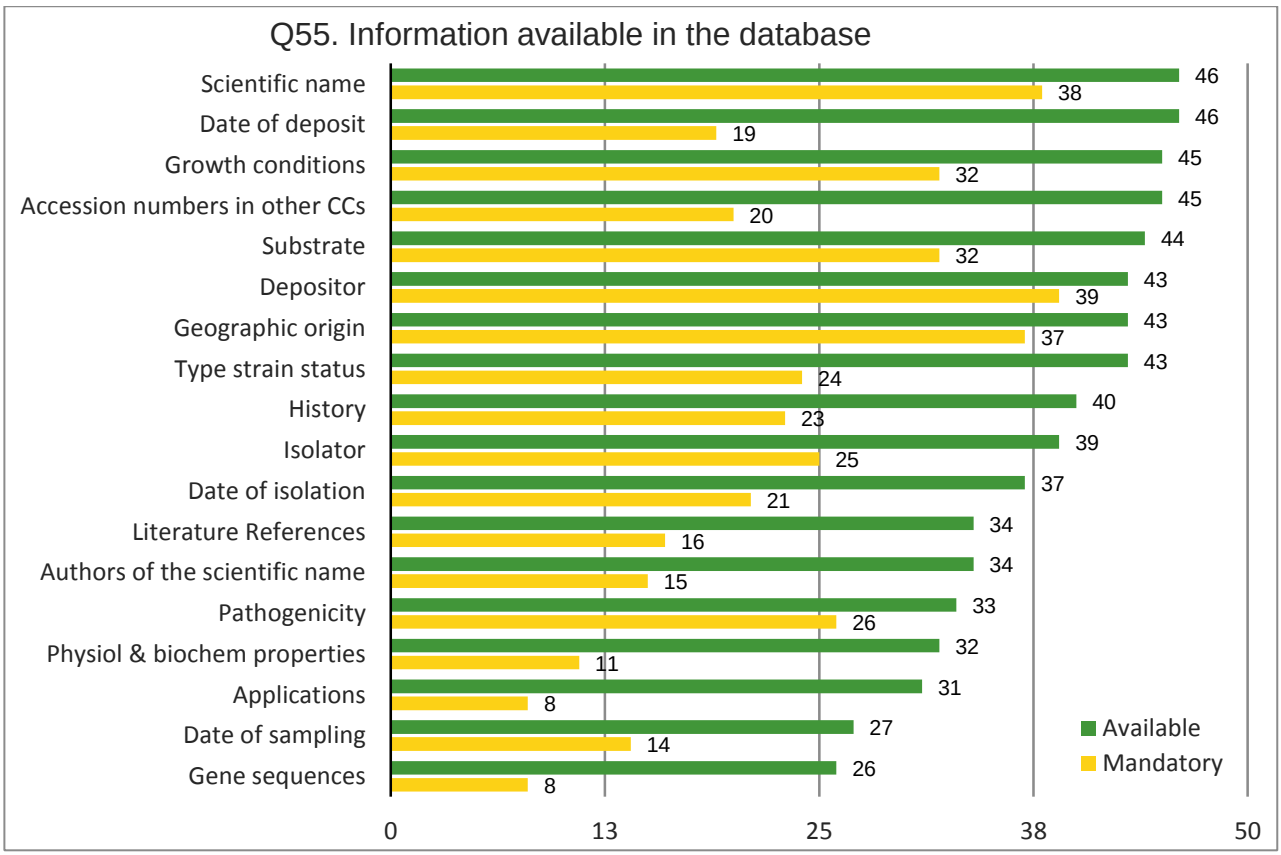
However, the most cited standards, i.e. “OECD Best Practice Guidelines” [1], “CABRI guidelines for catalogue production” [2,3], and “Microbial Information Network Europe (MINE) [4,5]”, have a strong overlap, because the OECD guidelines are based on CABRI guidelines, which in turn are derived from the MINE standard. Other standards, such as the “Access to Biological Collections Data (ABCD)” [6], have only been rarely cited. It therefore seems that the adoption of one of the MINE derived standards by MIRRI would only imply a limited impact on collection management procedures.

However, both ABCD and “Microbiological Common Language (MCL)” [7] may be of interest, since they have a better defined data model, ready for XML. MCL has a different approach, since it is not meant to express the contents of the catalogues, but to represent the origin, correspondences and equivalences of strains. Nonetheless, ABCD and MCL appear to be of interest because of their ability to cope with the most innovative software tools. A revised version of the MCL, able to express contents of CC catalogues by using a data definition derived from MINE, like the CABRI standard, could be a valuable hypothesis for the MIRRI standard.

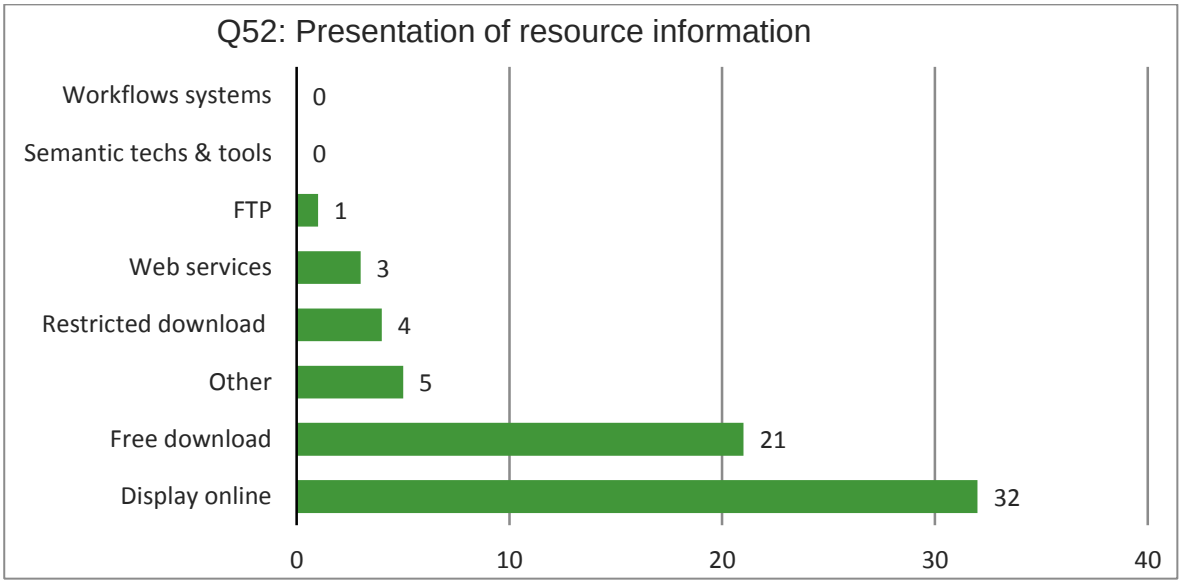


In order to properly curate data, controlled vocabularies are only used by 15 out of 49 collections, while some unspecified “standard operating procedures” are used by 23 out of 49 collections (see Q50). Regarding curation, it is surprising to note that there is no single information that is deemed mandatory for acceptance of a deposit by all collections (see Q55). Even the scientific name is considered mandatory only by 38 out of 52 collections.

A more specific survey should be requested to reach a deeper understanding of the curation activity that is carried out by collections. It seems, anyway, that semantic tools are not usually adopted, while taxonomic sources remain the fundamental reference for the definition of some data, like names. Other information is mainly introduced without any reference to shared syntax and controlled terminologies, e.g., for geographic information, or as textual comment only, e.g., for special features of the microorganisms, its activities, etc….

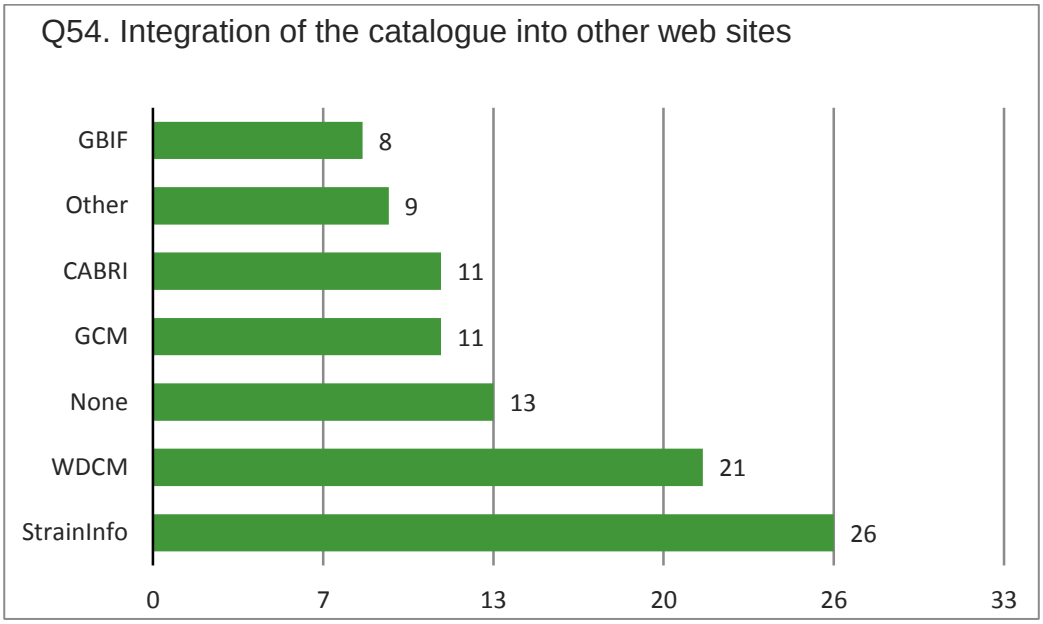


All culture collections present their catalogue information on-line and some of them allow end users to download it (see Q52). Advanced online tools, such as Web Services and other APIs (Application Programming Interfaces), automated workflows, or Semantic technologies, like those related to the Semantic Web, are almost unused (see Q52 again), although they have been demonstrated to be able to assist collections in validating and updating the catalogue data [8].



This may be a great disadvantage for the dynamic integration of systems (interoperability). The majority of catalogues (38 out of 51) are however included into other public web sites, including StrainInfo [9,10], the World Data Centre for Microorganisms (WDCM) [11,12], the Global

Catalogue of Microorganisms (GCM) [13,14], the Common Access to Biological Resources and Information (CABRI) web site [15], and the Global Biodiversity Information Facility (GBIF) [16] (see Q54). This is, however, a static integration, with catalogues been periodically transferred (i.e., uploaded to) these public sites.



Information available in the database is of course related to the guideline/standard that is followed by the single collection. Only a few data refer to information systems that are external to the catalogue, or to catalogues of other collections. Namely, only bibliographic references, gene sequences, and patent references are mentioned in the survey (see Q55 again).

Workshop on "Strain data and metadata, on semantic sources and on external databases to be considered for the integrated system"

The workshop on "Strain data and metadata, on semantic sources and on external databases to be considered for the integrated system" was considered as a fundamental step for MIRRI work package 8 "Data resources management" (WP8). It was included in the proposal as one of the project milestones (MS 8, M8.3.1). Initially set to be held within month 18, it was held on February 25-26, 2014, in Rome, at the University of Rome "La Sapienza".

Participants were selected in order to represent a wide and multidisciplinary team. Apart from work package- and task leaders, collections' curators, research scientists and computer scientists were invited to attend and bring their special expertise and points of view. This allowed the gathering of many useful presentations, as well as to have live, interactive and productive discussions.

The workshop was organized in three sessions, each of which was devoted to a single main topic. Each session was introduced by one or more presentations, which were then followed by open discussions. A final discussion allowed the identification of several key issues, for which an agreement was reached, and to identify the next steps towards the preparation of this deliverable.

In the first session, A. Vasilenko from VKM introduced the discussion by presenting the results of the analysis of existing standards for information on microorganisms. This analysis took into account the results of three relevant and significant past activities, i.e. Micro-organisms Information Network Europe (MINE) and Common Access to Biological Resources and Information (CABRI) projects and the OECD "Best Practice Guidelines for Biological Resource Centres", along with the initial results of the European – Russian initiative Banking Rhizosphere Micro-Organisms (BRIO).

In the second session, the first presentation was given by V. Robert from CBS. His talk had a natural connection to the first session. He pointed out two distinct, yet complementary, issues which are essential for the implementation of MIRRI Information System: the user interface and the underlying data model. Two short presentations were then given by P. Romano from USMI who presented an overview of the current status of biomedical ontologies and introduced Semantic Web technologies. The third session was introduced by interesting presentations given by F.O. Glöckner from JacobsUni and by B. Bunk and C. Söhngen from DSMZ, who respectively introduced databases on sequences, bacterial diversity, and biodiversity, along with possible interactions with CC catalogues. In the final discussion, three main issues were presented. For each of them, an agreement was achieved among participants resulting in the next steps for the MIRRI project being defined.

The concept of a Minimum Data Set (MDS) for MIRRI was discussed in detail. The general feeling is that such a dataset could be misleading, because it could drive the general attention towards a very limited set of information, while hiding at the same time the crucial aspects of both data formats, that must be especially stressed in order to achieve real interoperability, and of applications' oriented data, that are essential in order to implement innovative downstream analyses in the MIRRI-IS platform, although may not be commonly available at all collections.

As it was already pointed out in the deliverable 8.1, submission of strain data to mBRCs does not always follow standardized protocols or procedures. Also "*the quantity and quality of datasets available in different MRCs is very heterogeneous and far from being comprehensive*". This can be true even if the majority of mBRCs declare to follow one of the main standards. Finally, it was pointed out that "*mBRCs need to follow standard operation procedures (SOPs) to guarantee evaluation and consistent electronic storage of metadata*". As we have seen from the questionnaire on data management systems by mBRCs, the curation activity most often follows *ad hoc* approaches which are different from one mBRC to the others. Besides requirements and standards for data acquisition which are being defined in the task 8.1, the use of shared data formats, controlled vocabularies, terminologies, and ontologies to meet agreed standards are strictly required.

Having this target, participants agreed to concentrate the next activities along the following lines:

i) redefine information that is usually taken into account by micro-organisms collections on the basis of their proper application domains, by creating subsets of information that are of special interest for each domain,

ii) examine all data that is currently included in existing standards, i.e. MINE, CABRI, and OECD, in order to identify a proper data structure for each of them, including data type, values, syntax, best reference lists, terminologies, and ontologies when appropriate, with the aim of defining in a very precise way how the data must be encoded by collections.

It is expected that from such analyses new clear and effective definitions for involved data can be identified and agreed, so that a shared, improved standard can be established, towards the definition of an exhaustive data set, which could be the basis for a standard Minimum Information about Biological Resources (MIaBRe).

With reference to the biomedical databases which, being "external" to the datasets provided by microorganism collections, need to be interconnected with collections' data, the presentations given at the workshop demonstrated, on the one hand, that there is a great chance for integration that may indeed allow an innovative downstream analysis, and, on the other hand, that a careful analysis of the numerous existing databases of possible interest is needed.

With these objectives, the participants decided to proceed with a deep analysis of databases presented at the workshop, with the aim of identifying both the information that should be linked and the best way for establishing links. Taking into account that further databases could be of interest for the MIRRI-IS, the existing lists of biomedical databases, such as the NAR online Molecular Biology Database Collection that is provided by Oxford Journals through its web site [17], will also be carefully examined in order to identify more databases of possible interest.

Semantic sources have proven to be an essential tool for data integration in life sciences and medicine. Their main role may include the definition of a common terminology for all involved information sources, which is the best way to establish a real interconnection among them, and to enable interoperability of related information systems.

For this reason, they must be exploited to their best within the MIRRI-IS. However, the objective of this work package does not consist in the creation of new, or assessment of existing, semantic sources for the microbial knowledge domain, but on the use of the existing semantic sources that are the most useful for the interconnection / interoperability issues of MIRRI-IS.

Participants have therefore agreed to take into account all semantic sources that:

i) could prove useful for the definition of proper data structure/contents of data fields in existing standards for microorganism databases/catalogues,

ii) could support interconnection with external databases.

These sources will be identified during the analysis steps that have been defined for collections' catalogue data fields and for external databases.

Data sets for culture collections databases

As pointed out in deliverable D8.1, the minimum data sets for culture collections have been the topic of previous European Commission funded projects, including the Common Access to Biological Resources and Information (CABRI) [15], which were adopted by the OECD Best

Practice Guidelines for Biological Resource Centres [1]. These are the basis for the development of the MIRRI IS.

The minimum data sets (MDS), recommended data set (RDS) and full data set (FDS) defined by these best practices are the starting point for the CC database component of the MIRRI 4 ALL portal, a good basis for now while MIRRI focusses on how to add extra value to its data.

What is needed, first of all, is to:

- improve the definition of the contents of the information included in the data sets,
- define a proper data model for each piece of information,
- define a proper curation strategy and tools for ensuring high quality data in mBRC databases.

Already in deliverable 8.1, the core set of code data fields has been defined as: Strain Number, Other Strain Numbers, Name, Organism Type, Restrictions, Status, History of deposit, Growth conditions, Form of supply, Geographic Origin, and INSDC accession number(s) for reference sequences, namely 16S and 18S ribosomal RNA (rRNA) gene sequence where they exist. As specified, this is only a starting list for the better definition of contents: all data that is available in culture collections, especially mBRCs handling resources other than microbial strains, for example plasmids, will be taken into account and progressively analysed and included in MIRRI.

An initial analysis was carried out on data sets for bacteria and it is included as an annex to this report. The following examples are only meant to be representative of the work that is being carried out.

The Strain number is the collections' own accession number of the strain. It usually consists of two components: the collection acronym which is followed by a number or alphanumeric identifier, that must be unique in the collection for the given strain, separated with a blank. Well known examples are “LMG 19”, “DSM 1046”, “CBS 354.79”. In this case, the information can be considered as a single data item, even if it is composed of two terms. The data model can include just one field, which is simple to check for inconsistencies, at least apparent ones. Also for such simple information, errors can arise, e.g. when the mandatory space between acronym and identifier is omitted. A reference list of acronyms can be easily created by making reference to international organizations, like WDCM CCINFO [12], and indeed lists of acronyms are already available for checking the strain number correctness.

The information related to accession numbers of the strain in other collections is also apparently easy to define and check for curation. However, this may be much more complex if a complete analysis of “equivalent” strains in collections is required. To this aim, StrainInfo can provide extracts of the curated information. So, while the definition of the data field can just consist in a list of strain numbers, each of which must follow the related data model, the curation of this information requires much more attention.

Some data fields may be simply defined by an enumeration of allowed values. In this case, a reference list must be defined and used by all collections. Examples are the form of supply of a

strain to customers, whose values maybe enumerated as ‘Active', ‘Freeze-dried', 'Dry ice', and the organism type, which would be enumerated with values "B" for bacteria and "A" for archaea.

The majority of other data fields require a deeper analysis. The scientific name is certainly among the most complex pieces of information and it requires careful evaluation. The CABRI guidelines define a unique data field, manually curated, that is meant to include many single data items: the genus name, the species epithet, subspecies (if applicable), pathovar (if applicable), author(s) of the name (which indeed is not a single value because many authors may be cited), year of valid publication or validation, and approbation of the name.

In this context, each single data has its defined input process, while some simple syntactic rules make the separation of single values possible. E.g., the scientific name must be provided as given by depositor and confirmed (or changed) by collection, the names of the authors of the name, the year of valid publication or validation and the approbation must be included after a comma. Possible values for approbation are listed as: AL = approved list, cf. International Journal of Systematic Bacteriology (IJSB) 1980, VL = validation list, in IJSB after 1980, VP = validly published, paper in IJSB after 1980. The DSMZ list of bacterial names [18] is cited as a reference list for names. It is even more complex for the fungi but where MycoBank [29] and Index Fungorum [30] can offer guidance.

For a proper curation and validation of information, this kind of data field should probably be split into single data, each of which can be automatically curated and validated both as a single piece of data, and in conjunction with one or more of the others in an effective manner.

The same issue may arise, and the same approach can be applied, to many further data fields, including of course infrasubspecific names, which is meant to contain variety, designation, epithet, authors and reference, but also the status of the bacteria strain, which should contain information on type and scientific name of the organism from which the strain was isolated as type material. In the latter case, e.g., the type should be defined on the basis of an enumeration (‘T’, ‘NT’, ‘PVRS’) and the name should follow the same format as the scientific name, as described above.

A different issue is represented by those data fields that report some of the strain properties, usually in a narrative way, using free text. These data should obviously be enhanced with the aim of exploiting them for integration/interoperation purposes. The first obvious choice consists in the separation of data, i.e. in the creation of a greater number of fields each of which have a semantically consistent meaning and a precise syntactic definition, and the creation, or adoption when available, of reference terminologies.

For bacteria and archaea, the data fields on Pathogenicity, Enzyme production, Production of metabolites, and Applications are all of this kind. It is noteworthy that some of these data can also be used for interoperation with external databases, such as those related to disease symptoms, chemical compounds, proteins, and with ontologies, including well known Gene Ontology [19,20], largely applied for annotation of sequence, interaction networks, and pathways databases.

Public databases and information systems related to microbial resources

Many public information systems and databases exist that may support, in tight connection with mBRC databases, the proposed downstream analysis tools. These systems relate to different knowledge domains and research environments, including sequence data, ecology, geography, climate, literature. During the Rome workshops, some of these were presented, but a complete list of all systems and tools of possible interest is very difficult to achieve. Moreover, new information sources and databases are being created continuously and this calls for a recursive procedure, able to assess new systems periodically. It is likely that this procedure can best be implemented when a new system, targeting a given application and analysis, is being developed.

This procedure should be based on a careful analysis of information included in the databases of interest, with the aim of identifying possible links between mBRC databases and external resources and the best way for establishing these links. Such analysis should be applied to databases included in public lists, such as the NAR online Molecular Biology Database Collection [17]. A list of information systems, databases and tools of interest for the microbial resources could be compiled and made available on-line along with annotations of possible links.

Molecular and associated data resources

Sequence databases

Sequence data is archived in one of the databases participating the International Nucleotide Sequence Database Collaboration (INSDC) [21], namely GenBank [22], ENA [23], and DDBJ [24]. In these databases, a huge number of sequences determined from microbial strains are available. A few values may demonstrate this: in ENA release 120, 8,923,904 sequences from filamentous fungi strains are included, as well as 5,585,707 sequences from prokaryotes, and 43,049,094 sequences from environmental samples; there presently are 1,850,526 links from ENA sequences to StrainInfo, while 683,444 links from ENA to CABRI will be established with next release.

In this context, while it is clear that adding so many links to mBRC databases would be impossible, it is essential that on one side links from ENA are established towards mBRC databases and at least some essential links to main sequences, e.g. 16S rRNA, are included in mBRC catalogues for each strain.

The Barcoding of Life Database (BOLD) [25,26] platform comprises a set of integrated databases among which the Public Data Portal and the Barcode Index Number (BIN) database may be of interest for MIRRI. BINs are sequence clusters that closely approximate species and may allow for rapid validation and use of barcode data where taxonomic data are lacking or unverified.

The SILVA rRNA database project [27] is a comprehensive resource for ribosomal rRNA sequence data providing quality checked and regularly updated datasets of aligned small (16S/18S, SSU) and large subunit (23S/28S, LSU) rRNA sequences for all domains of life. It has established links to StrainInfo and to ENA and it may then represent a valid source of information for a quick

incorporation of sequence accession numbers in mBRCs catalogues and for an effective interoperation with ENA.

The NCBI manage a set of coherent databases, some of which are of interest to MIRRI. The NCBI taxonomy identifier (taxonID), in conjunction with a specific microbial resource strain number, may be a valid data combination that can be used to establish links to many information resources.

The Braunschweig Enzyme Database (BRENDA) [28], that contains detailed information on enzymes and their ligands, is an example of such databases. By searching BRENDA with a given taxonID, it is then possible to retrieve information on all molecules related to a given strain number. The EC number (Enzyme Commission number) is retrieved and it can then be used to gather further additional information from other databases, like UniProt, KEGG, etc…

This is a clear example of how the contents of mBRC catalogues can be validated and extended. While some mBRC catalogues have information on enzyme production, this rarely includes the EC number or a similar identifier and free text is usually preferred. However, this limits the possibility of integrating the catalogue with other databases. Through some automatic procedure, it would be possible to retrieve EC numbers for a given strain, compare this information with the contents in the catalogue, and possibly correct or update it, by also including the EC number that may later enable a more effective connection of the catalogue with other databases, including protein, protein interactions and networks, pathways.

Semantic resources for microbial databases

The StrainInfo (SI) portal [9], which collects information from many CC catalogues, has been able to reconstruct connections among strains from different collections, thus representing a fundamental curation tool for information on other collection numbers for each strain. Also, it includes an impressive number of established links to sequence, taxonomy and literature databases that can be exploited by the MIRRI-IS.

The Microbiological Common Language (MCL) [7] is also a useful resource, although it presently does not allow the inclusion of all the contents of a CC database that follows one of the main standard/guidelines (OECD [1], CABRI [2], MINE [4,5]).

Another downside to StrainInfo is that it was designed specifically for bacterial, yeast and fungal strains. The portal needs to be adapted to include other biological resources, such as plasmids and phages. To this end, an extension must be developed and implemented. See Annex B on CABRI and MCL for a first analysis of equivalences between these standards.

In the following excerpt, e.g., a simple example of a strain represented in MCL is reported. All tags enclosed between “<!—“ and “ -->” (also highlighted in red) are not presently defined in MCL. Some information is also truncated for readability.

```
<mcl:Culture>
        <mcl:strainNumber>DSM 5079</mcl:strainNumber>
        <mcl:otherStrainNumber>LMG 7912</mcl:otherStrainNumber>
        <!-- <mcl:restrictions>Risk group 2 (A)</mcl:resrictions> -->
        <!-- <mcl:organismType>Bacteria</mcl:organismType> -->
```

```
        <mcl:speciesName>Yokenella regensburgei</mcl:speciesName>
        <mcl:qualifiedSpeciesName>Yokenella regensburgei, Kosako et al. 1985 VP</mcl:qualifiedSpeciesName>
        <!-- <mcl:otherName type="Subjective">Koserella trabulsii, Hickman-Brenner et …</mcl:otherName> -->
        <mcl:typeStrainOf>Type strain</mcl:typeStrainOf>
        <mcl:Deposit>
                <mcl:resultingStrainNumber>DSM 5079</mcl:resultingStrainNumber>
                <mcl:depositDate>-</mcl:depositDate>
                <mcl:depositorInstitute>Georgia State Hlth. Dept</mcl:depositorInstitute>
        </mcl:Deposit>
        <mcl:history><- ATCC <- G. Carter, CDC 3349-72 <- Georgia State Hlth. Dept</mcl:history>
        <mcl:Medium>
                <mcl:mediumName>Medium 1</mcl:mediumName>
                <mcl:growthTemperature>37</mcl:growthTemperature>
        </mcl:Medium>
        <mcl:Sample>
                <mcl:sampleLocationCountry>-</mcl:sampleLocationCountry>
                <mcl:sampleLocationPlace>-</mcl:sampleLocationPlace>
                <mcl:sampleHabitat>human wrist wound</mcl:sampleHabitat>
        </mcl:Sample>
        <!-- <mcl:formOfSupply>Dried</mcl:formOfSupply> -->
        <mcl:Publication>
                <dc:title>Validation of the publication of new names and new …</dc:title>
                <prism:publicationName>Journal Int. J. Syst. Bacteriol.</prism:publicationName>
                <prism:volume>35</prism:volume>
                <prism:pageRange>223-225</prism:pageRang>
                <dcterms:issued>1985</dcterms:issued>
                <dc:creator>Hickman-Brenner, F. W., Huntley-Carter, G. P., …</dc:creator>
                <dc:title>Koserella trabulsii, a new genus and species ….</dc:title>
                <prism:publicationName>Journal J. Clin. Microbiol.</prism:publicationName>
                <prism:volume>21</prism:volume>
                <prism:pageRange>39-42</prism:pageRang>
                <dcterms:issued>1985</dcterms:issued>
                <!-- <mcl:pmid>85105491</mcl:pmid> -->
        </mcl:Publication>
</mcl:Culture>
```

The World Data Centre for Microorganisms (WDCM) maintains CCINFO, the catalogue of CCs in the world [12]. It provides a unique system of identifiers for strains. For this reason, it is a reference for description of collections and a link to/from MIRRI-IS should be provided.

In the following excerpt, e.g., a simple example of a collection represented in MCL is reported. The red tag, which is also enclosed between "<!—" and " -->", is not presently defined in MCL.

```
<mcl:Catalog>
  <mcl:CatalogDescription>
    <dc:creator>Paolo Romano</dc:creator>
    <mcl:catalogVersion>2004.1</mcl:catalogVersion>
    <mcl:catalogLastUpdateDate>27-03-2007</mcl:catalogLastUpdateDate>
      <!-- <mcl:CabriName>CBS_FIL</mcl:CabriName> -->
    <mcl:BRC>
      <mcl:WDCMNumber>133</mcl:WDCMNumber>
      <mcl:fullName>Centraalbureau voor Schimmelcultures, Filamentous Fungi … </mcl:fullName>
      <mcl:acronym>CBS</mcl:acronym>
    </mcl:BRC>
  </mcl:CatalogDescription>
</mcl:Catalog>
```

The WDCM also hosts the Global Catalogue of Microorganisms (GCM) [13], which already includes various types of microbial resources: Fungi, Yeast, Bacteria, Archaea, Microalgae, Cyanobacteria, Protozoa, Plasmid, Phage, and Virus. As such, the GCM may serve as a basis for

the integration of the different types of data from different types of resources into one common information system. However, currently this too does not include all data fields and is restricted to a defined MDS.

As it was already pointed out in deliverable 8.1, keeping pace with the taxonomy and name changes being continuously introduced for species is one of the fundamental unsolved problems for microorganism collections curators. Although many attempts to solve this problem have been made in last years, the problem is still evident and can be seen when making access to the WDCM and StrainInfo. In this context, the objective of MIRRI would be to highlight possible problems in its CCs and to offer exhaustive query tools, able to find all strains linked to the queried names.

Many taxonomies are of strict interest for MIRRI, as MycoBank (nomenclature and taxonomy) [29] and Index Fungorum (nomenclature) [30] for fungi, and the Prokaryotic Nomenclature Up-to-date provided by DSMZ (nomenclature and taxonomy) [18] and the List of Prokaryotic Names with Standing in Nomenclature (LPSN) [31] for bacteria.

All these systems have serious interoperability issues, because they offer a standard user-oriented web interface and no APIs. Moreover, data is not always stored in databases but in simple HTML pages. Therefore, for the MIRRI project, new tools are needed. A first search engine for bacterial names that can be queried by using Web Services has been created at CBS and it is available at http://www.mycobank.org/bacteria/. However, it is likely that more such APIs will be developed in the very near future and these should definitely be exploited within MIRRI-IS.

Conclusion

In this report, we have focused on some of the fundamental issues related to the development of the MIRRI-IS, the platform that will allow to exploit all available information from mBRCs catalogues, together with various databases and tools which are being maintained outside the MIRRI infrastructure, thus enabling an innovative downstream data analysis of microbiological information in various application domains.

The analysis of data management issues as reported by CCs in a recent survey has shown some interesting aspects. Although all the CCs are members of some International network in the field and the vast majority of them are aware of international standards for data structure and format, almost half of them do not use or apply any standard to its database. Anyway, the most cited standards are all derived from the MINE one, that do not offer interoperability features, contrary to the Microbiological Common Language (MCL) which, in turn, does not include all information that are usually included in catalogues. A revised version of the MCL, able to express contents of CC catalogues, may therefore be a valuable hypothesis for the MIRRI standard.

The use of taxonomic sources and of shared controlled vocabularies, which is fundamental both for data curation and exchange, is still limited, although it is much more frequent than that of semantic tools. From this point of view, much work is needed to enforce the adoption of shared terminologies, especially for information such as compound consumption and production, physico-

chemical properties, applications, which could better be used for linking to external databases. For this reason, it is unlikely that the MIRRI-IS will be able to exploit, in its initial phase, the wealth of information that is indeed contained in the CC catalogues. This will be achieved only after the wide adoption by CCs of a common data model.

With reference to interoperability of CC information systems, due to the lack of proper tools, it may be envisaged that the majority of catalogues will be included into the common MIRRI-IS by some kind of uploading (and converting) procedures, similar to what is presently done for, e.g., StrainInfo, GCM and CABRI. One of the objectives of the MIRRI infrastructure could then be the development of public APIs to enable interoperability of CC catalogues with the MIRRI-IS.

Previously defined standards for the description of data sets for CCs has been re-analyzed and compared with MCL in order to define a proper standard for interoperability, both among CCs and with external information resources. Annex A and Annex B of this document presents the preliminary results of this comparison.

The analysis of the CABRI data fields was carried out by taking into account all fields that were used by the majority of collections and that provided a good wealth of data, thus overcoming the limits which are implicit in the definition of minimum, recommended and full data sets: for the MIRRI-IS all information that can be given by collections is of the same relevance and should be provided to the central node. Only by this approach a real downstream data analysis environment may be realized. For all fields in the bacteria data sets, an analysis of their contents and of a possible improved definition, with the aim of describing each single data, was performed. Along with this, possible curation activities were outlined. This analysis is attached as Annex A.

A comparison of CABRI data fields with MCL tags was also carried out, in order to: i) identify equivalences and differences, ii) identify information for which the MCL language does not provide any tag, iii) suggest a list of new MCL tags that allow to incorporate the full contents of a CC catalogue described on the basis of CABRI data sets into an MCL file. This comparison is attached as Annex B.

References

1. OECD (2007). OECD Best Practice Guidelines for Biological Resource Centres http://www.oecd.org/health/biotech/oecdbestpracticeguidelinesforbiologicalresourcecentres.htm
2. CABRI Guidelines for catalogue production http://www.cabri.org/guidelines/catalogue/CPdata.html
3. 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
4. Stalpers JA, Kracht M, Janssens D, De Ley J, van den Toorn J, Smith J, Claus D, Hippe H. Systematic Applied Microbiol 1990;13:92-103
5. Gams W, Hennebert GL, Stalpers JA, Janssens D, Schipper MAA, Smith J, Yarrow D, Hacksworth DL. Structuring Strain Data for Storage and Retrieval of Information on Fungi and Yeasts in MINE, the Microbial Information Network Europe. Journal of General Microbiology (1988), 134, 1667-1689.
6. Access to Biological Collections Data (ABCD) http://www.tdwg.org/standards/115/

7. Microbiological Common Language (MCL) http://www.straininfo.net/projects/mcl
8. Romano P, Marra D, Milanesi L. Web services and workflow management for biological resources, BMC Bioinformatics 2005, 6(Suppl 4):S24
9. StrainInfo bioportal http://www.straininfo.net/
10. Van Brabant B, Dawyndt P, De Baets B, De Vos P. A Knuckles-and-nodes approach to the integration of microbiological resource data. Lecture Notes in Computer Science, 4277, 740-750, 2006
11. World Data Centre for Microorganisms (WDCM) http://www.wdcm.org/
12. Culture Collections Information Worldwide (CCINFO) http://www.wfcc.info/ccinfo/
13. Global Catalogue of Microorganisms (GCM) http://gcm.wfcc.info/
14. 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
15. Common Access to Biological Resources and Information (CABRI) http://www.cabri.org/
16. Global Biodiversity Information Facility (GBIF) http://www.gbif.org/
17. NAR online Molecular Biology Database http://www.oxfordjournals.org/our_journals/nar/database/c/
18. Prokaryotic Nomenclature Up-to-date http://www.dsmz.de/bacterial-diversity/prokaryotic-nomenclature-up-to-date.html
19. Gene Ontology http://geneontology.org/
20. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G. Gene Ontology: tool for the unification of biology. Nature Genetics 25, 25 - 29 (2000)
21. International Nucleotide Sequence Database Collaboration (INSDC) http://www.insdc.org/
22. GenBank http://www.ncbi.nlm.nih.gov/genbank/
23. European Nucleotide Archive (ENA) http://www.ebi.ac.uk/ena
24. DNA DataBank of Japan (DDBJ) http://www.ddbj.nig.ac.jp/
25. Barcoding of Life Database (BOLD) http://www.barcodinglife.com/
26. Ratnasingham S, Hebert PDN. BOLD: The Barcode of Life Data System. Molecular Ecology Notes (2007) 7, 355-364.
27. SILVA rRNA database project http://www.arb-silva.de/
28. BRENDA (BRaunschweig ENzyme Database) http://www.brenda-enzymes.org/
29. MycoBank http://www.mycobank.org/
30. Index Fungorum http://www.indexfungorum.org/
31. List of Prokaryotic Names with Standing in Nomenclature (LPSN) http://www.bacterio.net/

Annex A

From CABRI data fields to MIRRI data objects: the evolution of a standard

(version 1.0, October 29, 2014)

P. Romano, D.P. Colobraro

In the context of the Common Access to Biological Resources and Information (CABRI) project, participant collections and IT experts made an effort to compare information available in collections' catalogues and define a common format for the fundamental information included in the catalogues so that these could be included in a unique information system able to index all contents in a coherent way, thus offering a common access point for all catalogues of participating collections. The resulting guidelines for catalogue production have enabled the implementation of the CABRI Network services, which are offering a common access to more than 120,000 high-quality resources from ca. 30 collections in Europe since 2000 at www.cabri.org.

CABRI guidelines define three distinct data sets for each of the resource types which are taken into account. The Minimum Data Set includes all information that is necessary to identify the resource. The Recommended Data Set (RDS) includes all information that is usually made available by the majority of collections and may improve the characterization of resources. The Full Data Set (FDS) includes all further information that may be uniquely provided by each single collection. The guidelines are meant to define a shared format for all information that is usually included in mBRCs catalogues. However, they only define a tight syntax for included information only in few cases, e.g. the scientific name and the bibliographic reference. The vast majority of information is free text and there is no strict validation procedures for data included in catalogues.

This approach, which has demonstrated to have a good efficacy in achieving the objectives of the project, is based on the idea that all catalogues are stored in a single server. With the evolution of ICT tools, and especially of those tools that make the development of interoperable systems possible, a different approach, able to connect servers of participating collections and allow them to interoperate, is viable and offers some advantages, mainly related to availability of always up-to-date information for end users, a very limited work load for collections, and an improved interoperation with other biological databases. This can only be achieved if a better model for collections' data is designed and implemented, at least at an API level.

For this purpose, the following table shows how the CABRI data model could evolve to a data model allowing for interoperability, by taking as an example the bacteria and archaea CABRI data sets. The information on CABRI guidelines are taken from the CABRI web site (http://www.cabri.org/guidelines/catalogue/CPdata.html), while the information on the MINE standard are taken from Stalpers et al., Systematic Applied Microbiol 1990;13:92-103. Similar tables can be designed for fungi and yeast strains and for phages, plasmids and other biological resources.

CABRI Field	Contents	CABRI Input process	MIRRI data model	Curation tasks
Accession number	Collections' own accession number of the strain	Consists of collection acronym followed by a number or alphanumeric identifier separated with a blank. Collection acronyms: CABRI agreed list: http://cabri.org/guidelines/catalogue/CPocnbact.html One accession number refers to a unique deposit of the biological material .	Character string Reference list of acronyms	Check acronym and syntax Verify uniqueness
Other culture collection numbers	Accession numbers of the strain in the collection or in other collections. Cannot be given if the strain is unique in the collection.	New strain: as given by depositor Existing strain: if strain is sent to another collection, get and add accession number used in that collection. Collection acronyms: CABRI agreed list: http://cabri.org/guidelines/catalogue/CPocnbact.html Numbers are separated by semicolon	Character string Reference list of acronyms Multiple values allowed	Check acronym and syntax Verify reciprocity between collections and StrainInfo annotations

Restrictions	Contains: - Release conditions / restrictions - country-specific shipping restrictions - restrictions imposed by the depositor etc.	Value "no", or a text field describing the appropriate restrictions: hazard group, import/export regulations Enter 'No', or actual restrictions, which may be linked to a file for conditions of delivery The text may contain codes which refer to text files explaining the restrictions in more detail. Most frequent contents in CABRI catalogues: no (14,715), requires signed release (10,706), Biohazard group 1 (9,473), ACDP Category 1 (5,831), For research only (D) (2,839), Biohazard group 2 (2,669), Risk group 2 (A) (1,813), Not Yet assigned (1,327), ACDP Category 2 (1,151), Biohazard group 2; In addition to the general BCCM MTA, for this strain a depositor's MTA is applicable, which prohibits the commercial use of this strain (526)	It should be split. Enumerated hazard values Evaluate the creation of a list of allowed terms Make explicit country specific restrictions Add a free text comment field for remarks	Check enumeration values Check coherence among equivalent strains in catalogue and in other collections Guarantee updating
Organism type	Value "B" for bacteria or "A" for archaea.	Enter "B" or "A"	Enumeration: "A","B"	

Name	Full scientific and most recent name of the strain, including: - Genus name and species epithet - Subspecies * - Pathovar * - Authors of the name - Year of valid publication or validation - Approbation of the name *: only if applicable	Enter full scientific name as given by depositor and confirmed (or changed) by collection. Names of authors of the name, year of valid publication or validation and approbation are included after a comma. Values for approbation: AL = approved list, c.f.r. IJSB 1980, VL = validation list, in IJSB after 1980, VP = validly published, paper in IJSB after 1980 Reference list: DSMZ list of bacterial names: http://www.dsmz.de/bactnom/bactname.htm	Split components Manage multiple values for authors Add taxonID Enumeration for approbation: “AL”, “VL”, “VP”	Check against nomenclatures both single components and the full name Check for accented and language specific letters Check coherence among equivalent strains in catalogue and in other collections

Infrasubspecific names	Contains: - variety - designation - epithet - authors and reference Does not apply to all strains.	Enter type and epithet of the variety as given by depositor and confirmed (or changed) by collection. Names of authors and reference are included. This field excludes the pathovar name and the serovar name, which are both infrasubspecific names, but are to be entered in the Name field From MINE format: Infrasubspecific names Contains variety designation, epithet and author(s) and/or references. The authors and references are separated from the epithet or designation by a comma. Infrasubspecific subdivisions are designated by the term with the suffix -var (not -type) as biovar, chemovar, cultivar, morphovar, pathovar, phagovar, serovar or as forma specialis; phase. They have no nomenclatorial status. For practical reasons serovar and pathovar have been accomodated in separate fields. When the same strain belongs to more than one -var, these -vars are listed, separated by a semicolon. The "-var" indicator must therefore be entered individually. Example: in [Peudomonas fluorescens LMG 1244] biovar C, Stanier et l. 1966	Split components Manage multiple values for authors	Check against nomenclatures both single components and the full name Check for accented and language specific letters Check coherence among equivalent strains in catalogue and in other collections

Status	Nomenclature status of the strain (like "type", "neotype", etc.). Does not apply to all strains.	Enter information on type and scientific name of organism, for whose material from which strain was isolated is type material. From MINE format: Status T = type strain or type species, NT = neotype, PVRS = pathovar reference strain Since 1980 only type strains and neotype strains have status in nomenclature. Use of the rems pathotype (PTT), lectotype (LT), paratype (PT) and authentic strain (AUT) is not recommended. OR (original strain) and REFS (reference strain for infrasubspecific taxa) also have no status Status either refers to the name mentioned in the NAME fields, and then only one of the above abbreviations is given, or it refers to another taxon, then the name of that taxon with author(s) (no comma!) is given, preceded by "of". Examples: [in Acetobacter pasteurianus LMD 22.1] T [in Acetobacter pasteurianus LMD 51.1] T of Acetobacter pasteurianus ssp. Ascendens (Henneberg 1898) De Ley and Frateur 1974 AL	Separate type and name Enumeration for type: “T”, “NT”, “PVRS” Scientific name: as the Name field	As Name field for the name part Check allowed values for type

History of deposit	The history of the element from depositor to isolator, including: - Name of depositor - Institute of depositor - Name of isolator - Institute of isolator - Names and institutes in between	Enter information as given by depositor. From MINE format: History Fate of the isolate between isolation and deposit in the present collection. The backward sequence of deposits is used separated by "<" meaning "received from". Each entry may contain the name of the collection, (month and) year of the acquisition. Between parenthesis can be entered: strain designation or collection numbers (only when confusion is possible between two or more numbers from the same collection) and/or a name when a name change has occurred. Example: [in Bacillus sphaericus DSM 488] NCTC, Nov. 1973 (Bacillus loehnisii) < T. Gibson, 1935 < Kral Collection (Bacillus probatus)	Split components of the history Define special id for first isolator Keep trace of the sequence of deposits NB: in the example, other names are also included. This also appears in CABRI contents. Revise guidelines	Check for accented and language specific letters Build a reference list for researchers and institutes Check coherence among collections
Conditions for growth	Contains: - Culture medium - Atmospheric and light conditions - Temperature conditions	Each collection has to provide its own list of culture media and recipes, medium provided for strain is linked to medium in this file. List of recommended media at: http://www.dsmz.de/catalogues/catalogue-microorganisms/culture-technology/list-of-media-for-microorganisms.html Atmospheric and light conditions should only be given if they are special. Additional remarks on the cultivation like 'freshly prepared medium is necessary' or 'extended incubation time…' if necessary'	Split components of the medium Allow multiple values Add a free text comment field for remarks	Check terms for recipe ingredients

Form of supply	The form in which the strain will be sent to the customer.	Values: ‘Active', ‘Freeze-dried', 'Dry ice', ‘DNA’	Enumeration: “Active”, “Freeze-dried”, “Dry ice”, “DNA”. Multiple values allowed.	
Serovar	The serovar name and author. Required for some medically important species. Does not apply to all strains.	From MINE format: Serovar The word serovar is to be entered and (between parentheses) the code. If the serovar belongs to a serogroup, the serogroup may be entered after a semicolon. References are separated from the name by a "<". Example: [in Salmonella sp. LMD 73.6] Serovar dublin (1,9,12:g,p:-)		Check syntax and values of the code
Other names	Names used for the strain in the past (not synonyms) that may still be in use. Does not apply to all strains.	Enter previous names used for the strain if they are no synonyms of the current name. Same format as Name field.	Split names as for the Name field Different types: Objective name Subjective name Alternative state Add type value in a separate field	Same as for the Name field

Geographic origin	Name of country where strain originated from, followed by details on location.	Enter information as given by depositor or as retrieved from public information sources. From MINE format: Geographic origin (location of original material) Sequence: Country, state; local details A major delimiter ";" may separate subfields. Spelling of country names is English, other names according to national spelling ("Hannover, München, Lisboa, Warszawa, Moska"), Cyrillic names are transliterated according to ISO norm 833-1974. Examples: Canada, Quebec; Montreal, La Salle Woods, 400 m alt. Neherlands; Baarn, Estate Groenveld, near ditch CABRI agreed list for country names based on country code file retrieved from: ftp://ftp.ripe.net/iso3166-countrycodes.txt (updated by the RIPE Network Co-ordination Centre, in co-ordination with the ISO 3166 Maintenance Agency at DIN Berlin). Exception for UK, USA, USSR, Netherlands, North Korea	Separate components. Add standard geographic coordinate.	Check among components (e.g., town in proper country) Check country vs geographic coordinates Check for coherence among collections.
Mutant	Type and parent of mutant if strain is a mutant strain.	Free text: enter information as given by depositor or as retrieved from public information sources	Add separate value for mutation Add a free text comment field for remarks	Check the syntax of mutated strain number

Genotype	Names of chromosomal markers of the strain. Especially +recommended for strains of species for which many genetically modified strains exist. Does not apply to all strains.	From MINE format Genotype, chromosomal marker Examples: [in Escherichia coli PC 43] thyrA mal (lam)phx Most frequent values from CABRI contents: rif (63), thr; leu; thi; pyrF; codA; thyA; argG; ilvA; his; lacY; fhuA; tsx (59), lambda- (52), trp (40)	Separate markers Allow multiple values	Check syntactic correctness of information Use markers as link to external resources
Literature	Reference to publication in which the strain was originally described. Especially recommended for type strains.	Formatted reference as follows: Journal Title Year;Volume(issue):beginning page#-ending page# The journal title is abbreviated following ISSN abbreviations, which are without dot. Authors and title of the article are not mentioned. The reference can be followed by the Pubmed ID enclosed within square brackets as follows: [PMID: 1234567], where '1234567' is the Pubmed ID of the paper	Separate components Add PMID and DOI, when available	Check coherence between data and pubmed record Check coherence among collections
Sexual state	Undefined	From MINE standard: Sexual state Indicates the sexual condition of the strain. Examples: [in E. coli PC 1299] F- [in Pseudomonas aeruginosa PC 1412] FP+ Values from CABRI catalogues: F- (2,025), Hfr (289), F' (185), F+ (127), FP- (59), FP5+ (2), FP+ (2), FP2- (1)	Enumeration of possible values	Check for coherence among collections

Pathogenicity	Undefined	From MINE: Pathogenicity and virulence Name and/or symptoms of disease, susceptible group of organisms, separated by ":". If possible, use Latin names of organisms. Subfields for different diseases are separated by ";". Remarks can be entered after "<". The expression "causing dysentery in man" is entered as: dysentery: man. Example: crownrot: Rheum rhaponticum Values from CABRI catalogues (often in Remarks): group A (24), group E (19), group C (7), group B (4), group D (1), plant pathogenic (1), wooping cough symptoms (1), transmissible murine colonic hyperplasmia (1), strongly pathogenic on cotton (1)	Split components and add remarks (free text) Enumeration for "Pathogenicity group": "A", "B", "C", "D", "E", … Species name in "Pathogenetic for" Effect of pathogenicity Free text "Remarks"	Check syntax of species name Reference list for effects
Enzyme production	Undefined	From MINE standard: Name of enzyme(s). Subfields are separated by ";". Remarks after "<" may include optimal conditions for production and the information whether the enzyme is inducible. Prefixes (1, 2, D, L, alpha, beta) may be entered after the name, separated by "/". No abbreviation s in the database. Abbreviations or EC numbers for input will be organized through a thesaurus (based on Enzyme Nomenclature, 1984). Example: urease; xylosidase/beta; amylase More than 200 distinct values from CABRI catalogues. Most frequent values: restriction endonuclease (57), xylosidase/beta- (34), restriction endonucleases (9), hydrogenase (9), ferredoxin; formyltetrahydrofolate synthetase (9	Separation of single enzyme, multiple values allowed. Adoption of a reference vocabulary Use of EC number Reference link to enzyme databases (BRENDA) Separated free text remarks.	Check terms in reference vocabulary Check enzyme to validate EC number

Production of metabolites	Undefined	From MINE standard: Metabolites. Includes antibiotics, toxins, fermentation products, also “acid”, “gas”, etc… name of metabolite, possibly substrate (especially in case of biotransformation). No abbreviations in database. Abbreviations or CAS (Chemical Abstracts Service) numbers for input will be organized through a thesaurus. Subfields separated with “;” without spaces. Prefixes (1, 2, D, L, alpha, beta) to be entered after the name, separated by “/”. After “<” optimal conditions for production and origin of information may be entered in each subfield. Examples: lactic acid hydroxylation/12-beta from ergosterol streptomycin More than 940 distinct values from CABRI catalogues. Most frequent values: streptomycin (19), L-glutamic acid (15), inhibitors for glycoside hydrolases (14), steroids (12), oxytetracycline (12), tetracyclin (11), aureomycin (11), actinomycin (11), gentamycin (10), L(+)-lactic acid (9), lincomycin (9), chloramphenicol (9)	Separation of single metabolite, multiple values allowed. Adoption of a reference vocabulary Use of CAS number Reference link to compound database Separated free text remarks.	
Applications	Undefined	From MINE standard: General and industrial applications. Subfields separated by “;”. Keyword, description More than 1180 distinct values from CABRI catalogues. Most frequent values: mapping of mutants (199), Genome sequencing strain (82), Degradation of poly-3-hydroxyalkanoates; Degradation of poly-beta-hydroxybutyrate (70), Reference strain of Burkholderia cepacia complex strain panel (41), Quality control of media (35), Fixation of nitrogen (34)	Separate: type, domain, remarks Reference to external vocabularies (e.g., patents) for industrial categories?	

Remarks	Undefined	In CABRI, unique field devoted to include varied comments. Some examples of recurrent contents: “Atypical in fatty acid analysis”, “Does not belong to Bacillus circulans according to rDNA restriction fragment analysis (ARDRA)”, “Contains plasmid pEMT1”, “3-Ketolactose positive”, “Genomovar III”, “Pathogenicity group A”, “Race 3. Biovar 2”, “Avirulent strain”, “Willems et al. group 3”, “Tetracycline resistant”, “Possibly Rhizobium vitis”, “Possibly Alcaligenes sp.”, “Serotype III”, “Possibly Erwinia herbicola”, “Probiotic strain”.	Provide distinct remark fields for various data (cf. other data fields)	
Plasmids	Undefined	From MINE standard: Contains plasmid name, recombinant or natural plasmid, cryptic or non-cryptic and accession number in plasmid database. RP: recombinant plasmid; NP: natural (original) plasmid; C+: cryptic; C-: non-cryptic. Information on various plasmids is separated by a semicolon. If plasmid is unnamed, enter “unnamed”. More than 380 distinct values in CABRI catalogues. Most frequent values: F+ (124), PO1 of HfrH (96), pTiC58 (51), PO100 of HfrR4 (40), pTiB6S3 (35), PO2A of HfrC (Hfr Cavalli) (26), PO3 of P4X (J2, Hfr type 2) (23)	Separate components and add remarks. Define reference plasmid database. Allow multiple values.	Forced encoding. Check plasmid name in database.

Annex B

A preliminary comparison of CABRI data fields and the Microbiological Common Language (MCL) with a proposal for an extension of MCL

(version 1.0, October 29, 2014)

D.P. Colobraro, P. Romano

The Microbiological Common Language (MCL) was introduced for interoperability within the StrainInfo system. It aims at simplifying data exchange on microorganisms. It provides data in an XML-based format with special tags to identify information about microorganisms. CABRI data fields, which are defined in CABRI guidelines for catalogue production, are meant to define a shared format for all information that is usually included in mBRCs catalogues. It is used by the CABRI Network services (http://www.cabri.org/).

CABRI data fields and MCL tags do not overlap, mainly because of the different aims for which they have been defined. A comparison of CABRI data fields and MCL tags can be a useful for an improved definition of information included in MRCs catalogues.

MCL has 10 domain tags, see table 1, that cluster the full set of 102 property tags in groups related to main information domains.

Domain	**Description**
mcl:Culture	An instance of a strain, held at a given place and time. In practice, a mcl:Culture is associated with a strain number. Each issue of a new strain number yields a new culture.
mcl:Strain	The result of the StrainInfo integration of cultures. Staley and Krieg define a strain as the descendants of a single isolation in pure culture.
mcl:Sample	Environmental sample from which a microorganism was isolated. Multiple cultures can be isolated from the same Sample.
mcl:Isolation	The process of isolating a pure culture from an environmental sample. If multiple cultures have been isolated from a sample, each isolate corresponds to one Isolation entity.
mcl:Medium	Culture medium used to grow cultures.

mcl:Publication	Describes a scientific publication. A human-readable citation of the work with enough information to enable the user to find the intended publication is obligatory when describing publications. If the components of the citation are available separately (or the citation can be easily split into components), it is recommended to include the separate components.
mcl:Deposit	The transition of one culture to another. When a BRC transfers a culture to another BRC, a new culture originates and a new strain number is assigned. The process of deposit and assignment of new strain numbers results in a strain with an exchange history. This tag is used to model one transfer of the exchange history.
mcl:CatalogDescription	Metadata on a BRC catalog.
mcl:BRC	Used to describe a BRC related to a catalog .
mcl:StrainInfo	The root element of a MCL file for exports.

Table 1. MCL domain tags from Microbiological Common Language (MCL) reference (http://www.straininfo.net/projects/mcl/reference) and from Staley and Krieg. Classification of prokaryotic organisms: Bergeys Manual of Systematic Bacteriology, pages 1-4, 1984.

MCL tags are able to convey many different information types on microorganisms. However, some information that is available in mBRC catalogues cannot be represented by MCL. MINE standards and CABRI guidelines provide the elements that are requested for the definition of new tags able to extend the MCL and its ability to represent the full content of a catalogue.

In the following, a comparison of existing MCL tags and CABRI fields and a proposal for new MCL tags able to represent all CABRI data are reported (for more information on CABRI data sets and guidelines, see http://www.cabri.org/guidelines/catalogue/CPcover.html).

The MCL version of a catalogue must begin with some metadata on the related mBRC. This is done by the <mcl:CatalogDescription> tag, which must include information on the creator of the catalogue (<dc:creator> tag), the catalogue version (<mcl:catalogVersion> tag), the date of last catalogue update (<mcl:catalogLastUpdatedate>) and the information on the mBRC, which is organized in the <mcl:BRC> tag that includes sub-tags for mBRC name and acronym (respectively, <mcl:fullName> and <mcl:acxronym>). This section of the catalogue is identical for all resource types.

In the following MCL tags, information on the BCCM/LMG catalogue, that is available in the CABRI web site, are shown for demonstration purposes. Here two new tags are also included (highlighted in red in the excerpt below) in order to take into account the catalogue name (<mcl:catalogName> tag) and the unique number of the mBRC at the WDCM registry (<mcl:WDCMNumber> tag).

```
<mcl:CatalogDescription>
        <dc:creator>CABRI staff</dc:creator>
        <mcl:catalogVersion>2012.1</mcl:catalogVersion>
        <mcl:catalogLastUpdateDate>17-06-2014</mcl:catalogLastUpdateDate>
        <mcl:catalogName>BCCM_LMG</mcl:catalogName>
        <mcl:BRC>
                <mcl:WDCMNumber>296</mcl:WDCMNumber>
                <mcl:fullName>Belgian Coordinated Collections of Microorganisms/ LMG Bacteria Collection</mcl:fullName>
                <mcl:acronym>LMG</mcl:acronym>
        </mcl:BRC>
</mcl:CatalogDescription>
```

Biological resources included in the catalogue are then described by using one <mcl:Culture> tag for each resource.

As anticipated, there isn't a complete correspondence between data sets included in standards, such as CABRI and OECD guidelines, and information that is accounted for by MCL. In the following tables, the data fields included in the CABRI Recommended Data Sets are compared to MCL tags. When no MCL tag is available for a given field, a proposal for a new tag is reported. For sake of clarity, a separate table is shown for each organism.

Table 2. A comparison between the CABRI Recommended Data Set for Bacteria and Archaea and MCL.

CABRI FIELD	MCL TAG (existing or proposed)	NOTE
Strain_number	<mcl:strainNumber>	Same format, one occurrence
Other_collection_numbers	<mcl:otherStrainNumber>	Same format, multiple occurrences. To identify origin of information, a comment could be

		inserted, e.g., <!-- Strain from CABRI -->
Organism_type	<mcl:organismType>	Same format, one occurrence
	<mcl:speciesName>	Not requested in CABRI data sets
Name (Infrasubspefic_name)	<mcl:qualifiedSpeciesName>	Almost same format, one occurrence. When available, it includes CABRI Infrasubspefic_name data. See examples in table 2a.
Other_names	<mcl:otherName type="TYPE">	Format: same as <mcl:qualifiedSpeciesName>. TYPE may be one of the following: Synonym, Objective, Subjective, AlternateState. E.g.: <mcl:otherName type="Synonym">
Restrictions	<mcl:restrictions>	
Status	<mcl:typeStrainOf>	
History	<mcl:history>	Same format, one occurrence
Conditions_for_growth	<mcl:Medium> including sub tags: <mcl:mediumNumber>, <mcl:mediumName> and <mcl:growthTemperature>	Multiple occurrences are allowed for <mcl:Medium>. For each <mcl:Medium>, only one sub tag can be used for each type
Form_of_supply	<mcl:formOfSupply>	
Serovar	<mcl:serovar>	
Mutant	<mcl:mutant>	
Genotype	<mcl:genotype>	Multiple occurrences are allowed
Isolated_from	<mcl:Isolation> including sub tags: <mcl:isolationDate>, <mcl:isolator>, <mcl:isolatorInstitute>, <mcl:isolationMethod>, <mcl:isolatedFromSample>, <mcl:comments>	Multiple occurrences are allowed for <mcl:Isolation>. For each <mcl:Isolation>, only one sub tag can be used for each type

Geographic_orign	<mcl:Sample> including sub tags: <mcl:sampleLocationCountry>, <mcl:sampleLocationPlace>, and <mcl:sampleHabitat> For geographic coordinates, use <wgs84_pos:lat> and <wgs84_pos:long>	
Literature	<mcl:Publication>, which includes one <dcterms:bibliographicCitation> for each reference. Within <mcl: bibliographicCitation>, the following shared tags from Darwin Core and PRISM (Publishing Requirements for Industry Standard Metadata) are used: <dc:title> (title), <dc:creato> (author, multiple), <prism:publicationName> (journal / book), <prism:volume> (volume number), <prism:number> (issue number), <prism:startingPage> (first page number), <prism:pageRange> (first and last page numbers), <dcterms:issued> (Year)	

Table 2a. Some examples of scientific name definitions in CABRI guidelines and MCL definition.

CABRI Name	CABRI Infrasubspecific_name	<mcl: speciesName >	<mcl: qualifiedSpeciesName >
Aeromonas veronii, Hickman-Brenner, MacDonald, Steigerwald, Fanning, Brenner and Farmer III 1988	biogroup sobria	Aeromonas veronii, biogroup sobria	Aeromonas veronii, Hickman-Brenner, MacDonald, Steigerwald, Fanning, Brenner and Farmer III 1988
Xanthomonas campestris pv. campestris, (Pammel 1895) Dowson 1939	-	Xanthomonas campestris pv. Campestris	Xanthomonas campestris pv. campestris, (Pammel 1895) Dowson 1939
Pectobacterium carotovorum subsp. carotovorum, (Jones 1901) Hauben, Moore, Vauterin, Steenackers, Mergaert, Verdonck and Swings 1999 VL	-	Pectobacterium carotovorum subsp. carotovorum	Pectobacterium carotovorum subsp. carotovorum, (Jones 1901) Hauben, Moore, Vauterin, Steenackers, Mergaert, Verdonck and Swings 1999 VL

Table 3. A comparison between the CABRI Recommended Data Set for Filamentous Fungi and for Yeasts and MCL.

CABRI FIELD	MCL TAG (existing or proposed)	NOTE
Strain_number	<mcl:strainNumber>	Same format, one occurrence
Other_collection_numbers	<mcl:otherStrainNumber>	Same format, multiple occurrences. To identify origin of information, a comment could be inserted: <!-- Strain from CABRI --> or <!-- Strain from Straininfo -->
Organysm_type	<mcl:organismType>	Same format, one occurrence
	<mcl:speciesName>	Not requested in CABRI data sets
Name	<mcl:qualifiedSpeciesName>	Same format, one occurrence.
Misapplied_names	<mcl:otherName type="TYPE">	For Filamentous fungi only Format: same as <mcl:qualifiedSpeciesName>. TYPE may be one of the following: Synonym, Objective, Subjective, AlternateState. E.g.: <mcl:otherName type="Objective">
Restrictions	<mcl:restrictions>	
Status	<mcl:typeStrainOf>	
History	<mcl:history>	Same format, one occurrence
Conditions_for_growth	<mcl:Medium> including sub tags: <mcl:mediumNumber>, <mcl:mediumName> and <mcl:growthTemperature>	Multiple occurrences allowed For each <mcl:Medium>, only one sub tag can be used for each type
Form_of_supply	<mcl:formOfSupply>	Multiple occurrences allowed
Sexual_state	<mcl:sexualState>	Multiple occurrences allowed
Mutant	<mcl:mutant>	Multiple occurrences allowed

Applications	<mcl:application> or <mcl:application type="TYPE">	Multiple occurrences allowed. The alternative tag, that includes the' type' attribute, would support an improved definition of applications. The attribute could include a keyword, or a value from a vocabulary, or an ontological term, and the content of the tag would include an extended free text description. This would correspond to the MINE format that foresees a keyword followed by a description.
Geographic_orign Substrate	<mcl:Sample> including sub tags: <mcl:sampleLocationCountry>, <mcl:sampleLocationPlace>, and <mcl:sampleHabitat> For geographic coordinates, use <wgs84_pos:lat> and <wgs84_pos:long>	CABRI field Substrate is equivalent to the MCL tag <mcl:sampleHabitat>, which is a sub tag of <mcl:Sample> (for Filamentous fungi only)
Race	<mcl:race>	For Filamentous fungi only.
Literature	<mcl:Publication>, which includes one <dcterms:bibliographicCitation> for each reference. Within <mcl: bibliographicCitation>, the following shared tags from Darwin Core and PRISM (Publishing Requirements for Industry Standard Metadata) are used: <dc:title> (title), <dc:creato> (author, multiple), <prism:publicationName> (journal / book), <prism:volume> (volume number), <prism:number> (issue number), <prism:startingPage> (first page number), <prism:pageRange> (first and last page numbers), <dcterms:issued> (Year)	

Table 4. A comparison between the CABRI Recommended Data Set for Phages and MCL.

CABRI FIELD	MCL TAG (existing or proposed)	NOTE

Collection_number	<mcl:accessionNumber>	
Other_collection_numbers	<mcl:otherCollectionNumber>	Same format, multiple occurrences. To identify origin of information, a comment could be inserted: <!-- Strain from CABRI --> or <!-- Strain from Straininfo -->
Organism_type	<mcl:organismType>	Only one occurrence. Allowed value : phage
Name	<mcl:resourceName>	
Restricted_distributions	<mcl:restrictions>	
History_of_deposit	<mcl:history>	
Host_for_propagation	<mcl:hostForPropagation>	
Host_used_for_propagation	<mcl:hostUsedForPropagation>	
Lisogenity	<mcl:lisogenity>	
Virus_used_for	<mcl:application>	
Cell surface receptor	<mcl:cellSurfaceReceptor>	
Literature	<mcl:Publication>, which includes one <dcterms:bibliographicCitation> for each reference. Within <mcl: bibliographicCitation>, the following shared tags from Darwin Core and PRISM (Publishing Requirements for Industry Standard Metadata) are used: <dc:title> (title), <dc:creato> (author, multiple), <prism:publicationName> (journal / book), <prism:volume> (volume number), <prism:number> (issue number), <prism:startingPage> (first page number), <prism:pageRange> (first and last page numbers), <dcterms:issued> (Year)	

Table 4. A comparison between the CABRI Recommended Data Set for Plasmids and MCL.

CABRI FIELD	MCL TAG (existing or proposed)	NOTE
Collection_number	<mcl:accessionNumber>	
Other_collection_numbers	<mcl:otherCollectionNumber>	Same format, multiple occurrences. To identify origin of information, a comment could be inserted: <!-- Strain from CABRI --> or <!-- Strain from Straininfo -->
Type	<mcl:type>	Only one occurrence. Allowed value from the following reference list: plasmid, phasmid, cosmid, shuttle vector, transposon, minitransposon, IS element
Name	<mcl:resourceName>	
Restricted_distributions	<mcl:restrictions>	
Class	<mcl:class>	Only one occurrence. Allowed value from the following reference list: non-recombinant, recombinant
History_of_deposit	<mcl:history>	
Host_for_distribution	<mcl:hostForDistribution>	
Host_range	<mcl:hostRange>	
Medium	<mcl:Medium> including sub tags: <mcl:mediumNumber>, <mcl:mediumName> and <mcl:growthTemperature>	Multiple occurrences allowed For each <mcl:Medium>, only one sub tag can be used for each type
Lisogenity	<mcl:lisogenity>	
Selectable_phenotype	<mcl:selectablePhenotype>	Multiple occurrences allowed

Properties_and_Applications	<mcl:application> or <mcl:application type="TYPE"> <mcl:property> or <mcl:property type="TYPE">	Two new mcl tags are proposed, having the same format. Multiple occurrences allowed for both tags. The alternative tags, that include the' type' attribute, would support an improved definition of applications and of properties. The attribute could include a keyword, or a value from a vocabulary, or an ontological term, and the content of the tag would include an extended free text description. This would correspond to the MINE format for applications that foresees a keyword followed by a description.
Replicon	<mcl:replicon>	
Literature	<mcl:Publication>, which includes one <dcterms:bibliographicCitation> for each reference. Within <mcl: bibliographicCitation>, the following shared tags from Darwin Core and PRISM (Publishing Requirements for Industry Standard Metadata) are used: <dc:title> (title), <dc:creato> (author, multiple), <prism:publicationName> (journal / book), <prism:volume> (volume number), <prism:number> (issue number), <prism:startingPage> (first page number), <prism:pageRange> (first and last page numbers), <dcterms:issued> (Year). Suggested extension: <mcl:pmid> and <mcl:doi>	CABRI guidelines already foresees the use of the Pubmed ID, when available.

Table 5: MCL tags (existing in black and proposed in red) and their corresponding CABRI field(s) and the resource types where it can be used. When a CABRI field cell is empty, the corresponding MCL tag does not have a direct equivalent in CABRI data sets.
Legend for resource types: A: archaea, B: bacteria, F: filamentous fungi, Y: yeasts, Pl: plasmids, Ph: phages

MCL tag (existing or proposed)	**CABRI field(s)**	**Resource type(s)**
<mcl:strainNumber>	Strain_number	A, B, F, Y
<mcl:collectionNumber>	Collection_number	Pl, Ph

<mcl:otherStrainNumber>	Other_collection_numbers	A, B, F, Y
<mcl:otherCollectionNumber>	Other_culture_collection_numbers	Pl, Ph
<mcl:speciesName>		
<mcl:qualifiedSpeciesName>	Name Name + Infrasubspecific_name	A, B, F, Y A, B
<mcl:resourceName>	Name	Pl, Ph
<mcl:otherName type="TYPE">	Other_names Misapplied_names	A, B F, Y
<mcl:typeStrainOf>	Status	A, B, F, Y
<mcl:history>	History	A, B, F, Y
<mcl:history>	Hystory_of_deposit	Pl, Ph
<mcl:Medium>	Condition_for_growth	A, B, F, Y
<mcl:Medium>	Medium	Pl
<mcl:restrictions>	Restrictions	A, B, F, Y
<mcl:restrictions>	Restricted_distrinution	Pl, Ph
<mcl:organismType>	Organism_type	A, B, F, Y
<mcl:type>	Type	Pl, Ph
<mcl:formOfSupply>	Form_of_supply	A, B, F, Y
<mcl:Isolation> <mcl:isolationDate> <mcl:isolator> <mcl:isolatorInstitute> <mcl:isolationMethod>	Isolated_from	A, B

<mcl:isolatedFromSample> <mcl:comments>		
<mcl:Sample> <mcl:sampleLocationCountry> <mcl:sampleLocationPlace>	Geographic_origin	A, B, F, Y
<mcl:Sample> <mcl:sampleHabitat>	Substrate	A, B, F, Y
<mcl:genotype>	Genotype	A, B
<mcl:mutant>	Mutant	A, B, F, Y
<mcl:sexualState>	Sexual_state	A, B, F, Y
<mcl:race>	Race	F
<mcl:applications>	Applications Virus_used_for Properties_and_Applications	F, Y Ph Pl
<mcl:proprierties>	Properties_and_Applications	Pl
<mcl:hostForDistribution>	Host_for_distribution	Pl
<mcl:hostForPropagation>	Host_for_propagation	Ph
<mcl:selectablePhenotype>	Selectable_phenoype	Pl
<mcl:replicon>	Replicon	Pl
<mcl:hostRange>	Host_range	Pl
<mcl:hostUsedForPropagation>	Host_used_for_propagation	Ph
<mcl:lysogenicity>	Lysogenicity	Ph
<mcl:cellSurfaceReceptor>	Cell_surface_receptor	Ph

<mcl:Publication>	Literature	A, B, F, Y, Pl, Ph

External links

Presently, no sequence information is explicitly included in CABRI data sets, apart from the plasmid one. It is clear, however, that at least the information related to some characterizing sequences should be included. INSDC accession number(s) for reference sequences, namely 16S and 18S rRNA, were identified as the best option. In the new MIRRI data model, this data can easily be added, as a single value with a well-defined format. Also for MCL it would be straightforward to define one or more tags for this information in the description of a strain, i.e. in the context of the <mcl:Culture> tag. Two options may be identified: i) a single tag devoted exclusively to the accession number, ii) a set of tags in a new domain devoted to sequence information. In the first option, a tag <mcl:sequenceAccessionNumber> could be defined. In the second, the following tags could be instead defined.

```
<mcl:sequenceInfo>
    <mcl:sequenceOrigin>ENA-EMBL</mcl:sequenceOrigin>
    <mcl:sequenceDescription>Lactobacillus zymae 16S rRNA gene, type strain LMG 22198</mcl:sequenceDescription>
    <mcl:sequenceAccessionNumber>AJ632157</mcl:sequenceAccessionNumber>
    <mcl:sequenceLink>http://www.ebi.ac.uk/ena/data/view/AJ632157</mcl:sequenceLink>
</mcl:sequenceInfo>
```

In the latter case, the information stored in the MCL file would be redundant, since all data can be easily derived from ENA given the accession number, but it would be at the same time self-standing.

A similar approach may be applied to taxonomic information, where the taxonID could play the same role of the sequence accession number. The alternative options here would relate to: i) a single tag devoted exclusively to the taxon identifier, ii) a set of tags in a new domain devoted to taxonomic information. The second option would produce a set of tags such as the following:

```
<mcl:taxonomyInfo>
    <mcl:taxonomyOrigin>NCBI</mcl:taxonomyOrigin>
    <mcl:taxonomyID>1081613</mcl:taxonomyID>
    <mcl:taxonomyLink>http://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?id=1081613</mcl:taxonomyLink>
</mcl:taxonomyInfo>
```

International characters encoding

Due to its international nature, every effort for sharing data must take into account the need for appropriate management of international characters.

In CABRI, which is focused on European BRCs, accented letters (à, ü), other national letters (ñ, ß) and symbols (©, etc...) should be limited. The language to be used is English and the following rules were defined to adopt a standardised approach to some scientific symbols. To avoid any errors due to incorrect reading of a character set, standard ASCII alternatives to symbols must be used. These alterations must be made before submitting catalogues to CABRI for indexing. They include:

- Greek letters cannot be used, they must be fully spelled (e.g., write “alpha” instead of “α”, beta” instead of “β”)
- The “°” symbol (degree) for temperature is to be omitted entirely (e.g., 37C replaces 37° C).
- No subscripts or superscripts are allowed (e.g., cm3 replaces cm^3 and CO2 replaces CO_2)

Due to XML requirements, MCL adopts UTF8 encoding in order to properly manage international characters and words (for more information, see http://www.w3schools.com/xml/xml_encoding.asp). In this context, so called “special characters”, that is characters having a special function in XML (see a few examples in table 1), as well as special visualization issues, e.g. subscripts and superscripts, must be represented by their HTML entity/tag (for more information, see http://www.w3schools.com/html/html_entities.asp).

Character	HTML encoding
<	&lt;
>	&gt;
&	&amp;
3 (as superscript>	³
2 (as subscript>	₂

Table N. Special characters in XML