Meta-analysis of the possible MIRRI information system characteristics

A. Vasilenko, S. Ozerskaya, B. Bunk, A. Klindworth,
D. Smith, P. Romano, F. O. Glöckner

ICCC13, 26.09.2013, Beijing

MIRRI – Microbial Resource Research Infrastructure

- - Preparatory phase – Project development - 3 years,
- - Implementation phase – Project execution - 3 years,
- - Operation of real system – time frames not specified

- WP8 Data resources management
- Task 8.4 Provide strategies for access
- D8.4.1 Report on users requests, desired features, and meta-analyses of the integrated platform

The main recommendations

MIRRI
MICROBIAL RESOURCE RESEARCH INFRASTRUCTURE

MIRRI: ...to meet the needs of innovation in biotechnology.. fragmented resource distributed across Europe needs to be coordinated to common standards ... facilitating focus on the big challenges in healthcare, food security, poverty alleviation and climate change

WP8 Strategies will be defined to bridge the gaps of current storage and retrieval systems, to resolve the obstacles of data interoperability ...to add value for the users… . An assessment on existing tools, data-based platforms, standards and projects

Task 8.4. .. to assess how MIRRI can provide an optimal information system to access the integrated data of BRC collections ... ... A user survey on content … with the monitoring of user requests ... ... define the required features

Interdependence of desirable components

1.	**Information problems to be solved**
2.	Data contents, data structures
3.	Data standards, vocabularies
4.	Ontologies
5.	User interfaces, algorithms, interchange protocols
6.	Software
7.	Information system as a final product

Information problems

Information services for BRC and CC personal:

1. **Unification of Standard Operational Procedures (SOP), data structures, software tools.**
2. **Construction of integrated BRC knowledge system**

Information services for external users:

3. **Presentation of and search for BRC services**
4. **Content search of microorganisms based on their properties.**
5. **Navigation in information space of microbiology, bioinformatics, biotechnology, agriculture, medicine.**
6. **Content search of microorganisms experts based on user problems (problem discussed in WP6).**

A way toward unification

- Stable terminology shared by microbiological and IT communities. Dictionary.
- Shared information standards.
- Common formats of data processing operations and network access tools suitable for most of culture collections (CC). Unification of informational SOPs. Example of CABRI (http://www.cabri.org/guidelines/gl-framed.html).
- Common software tools.
- Shared formats of catalogue and services presentation on CC WEB site.
- Help with these tools for other collections.
- Ontology of microorganism names.
- Examples of advanced solutions in Strainfo, WDCM, Mycobank.

Construction of BRC knowledge system

1. Knowledge on key operations for the culture collections:
   - OECD Guidelines, WFCC Guidelines
   - SOPs in detail
   - IPR, biosecurity, biosafety, legal issues
2. Possible software tools survey
3. Links to the microbial research manuals
4. Links to microbial tutorials - morphology, cell structure, biochemistry, physiology, genetics, etc.
5. Links to educational programs
   - MIRRI internal programs
   - external programs
6. Links to experts

Presentation of BRC services

- Forms of supply of biological material.
- Forms of deposit, Deposit for patent purposes
- Expertise service: Isolation of pure cultures, Microbial count, Identification, Phylogenetic studies, Phenotypic characterization, DNA or RNA sequencing, Whole genome analysis, Gene sequence data analysis, Plasmid profile analysis, DNA-DNA hybridization, Gel electrophores, RAPD, AFLP, Ribotyping, Real-time PCR, Serotyping, FAME-MIDI, MALDI-TOF, Polar lipids determination, Metabolite production, Enzyme production, Pathogenicity tests, Antibiotic sensitivity tests, -omics, Screening for specific properties, Training
- Application fields:
  Agronomy, Bioremediation, Bioindustry, Environmental, Food

Content search based on properties

MIRRI
MICROBIAL RESOURCE RESEARCH INFRASTRUCTURE

What do we mean by the "Content search"? And why do we propose it?

On WDCM symposium in 2011 we demonstrated differences in typing.

We collected data on one strain of *Aspergillus brasiliensis* Varga et al. 2007 in four catalogues, strain numbers: ATCC 9642 = CBS 246.65 = DSM 63263 = VKM F-1119.

24 fields presented one strain in four collections, and only two fields in two collections were exactly the same. Potentially, it could be 144 coincidences, but formally the most of string values in four collections are different, and consistency level in this example is 1-2%.

But in fact there is no big difference in the content, and if we could compare the content of the fields the presentations are almost the same.

Examples

- The search for type strains for some taxa in some culture collections (CC) or BRCs.
- The search for microbial cultures of some kind in the list of CC or BRC, the answer in a table format. To help the customers with preferable purchase place.
- The search for rare and unique names of microorganisms (MO). The tools for identification and fixing misspellings in the names.
- The list of new MO taxa that were identified during some period of time and deposited into the Culture collections.
- Sublist of new taxa that have got descriptions according the actual Nomenclature Code.
- The search for strains, isolated by some authors, their strain numbers in Culture collections.
- The help in identification of new MO, if GBRCN has appropriate connection to databases with detailed description of MO (DSMZ/J.P. Euzeby, Mycobank, and/or other).

- To collect the lists of differences in description of cultures in one CC with cultures of the same strains in other CC. To do this search on all the fields of the data base, or on some list of fields, with all the other CC or for some of them. To help in fixing of mistakes in own CC Catalogue.
- The search for bibliographical references on some cultures of one CC and on the cultures of other CCs that are in the same strain. To sort the table produced by the names of cultures, or by authors, by journal name, or by year of publication.
- The search for bibliographical references on the cultures with some properties, with some nomenclature or taxonomical problems.
- The search for MO, typical for some ecological conditions.
- The search for microorganisms-identifiers of some plants (poppy, hemp, etc.) or of some ecological conditions.
- To find a list of strains isolated from some substrates (oil pollution, the vegetative rests, alkaline soils, food stuff, etc.). The answer sorted by MO names and by CC acronyms, in table format.

- To find a list of strains with some parameters of growth (thermophiles, psychrophilic, anaerobic, etc.). The answer sorted by MO names and by CC acronyms, in table format.
- To find a list of strains with some activities (ferments, organic acid production, transformation of steroids, etc.). The answer sorted by MO names and by CC acronyms, in table format.
- The search for microbial associations.
- Information on taxa diversity for microorganisms that have the sequence data in Genbank for type strains. The list of MO that has no sequence data for type strains. Calculate dynamics of growth for sequence data information on type strains.
- Cross statement of most questions from the list above. For example, the list of new taxa with some parameters of growth, or isolated from some substratum.

Crucial components

MIRRI
MICROBIAL RESOURCE RESEARCH INFRASTRUCTURE

According to experience of TDWG, GBIF, BioMedBriges the crucial components inside should be:

- **Community supported vocabularies and ontologies expressing shared semantics of data**
- **Common exchange protocols**
- **Persistent identifiers**

Navigation in information space

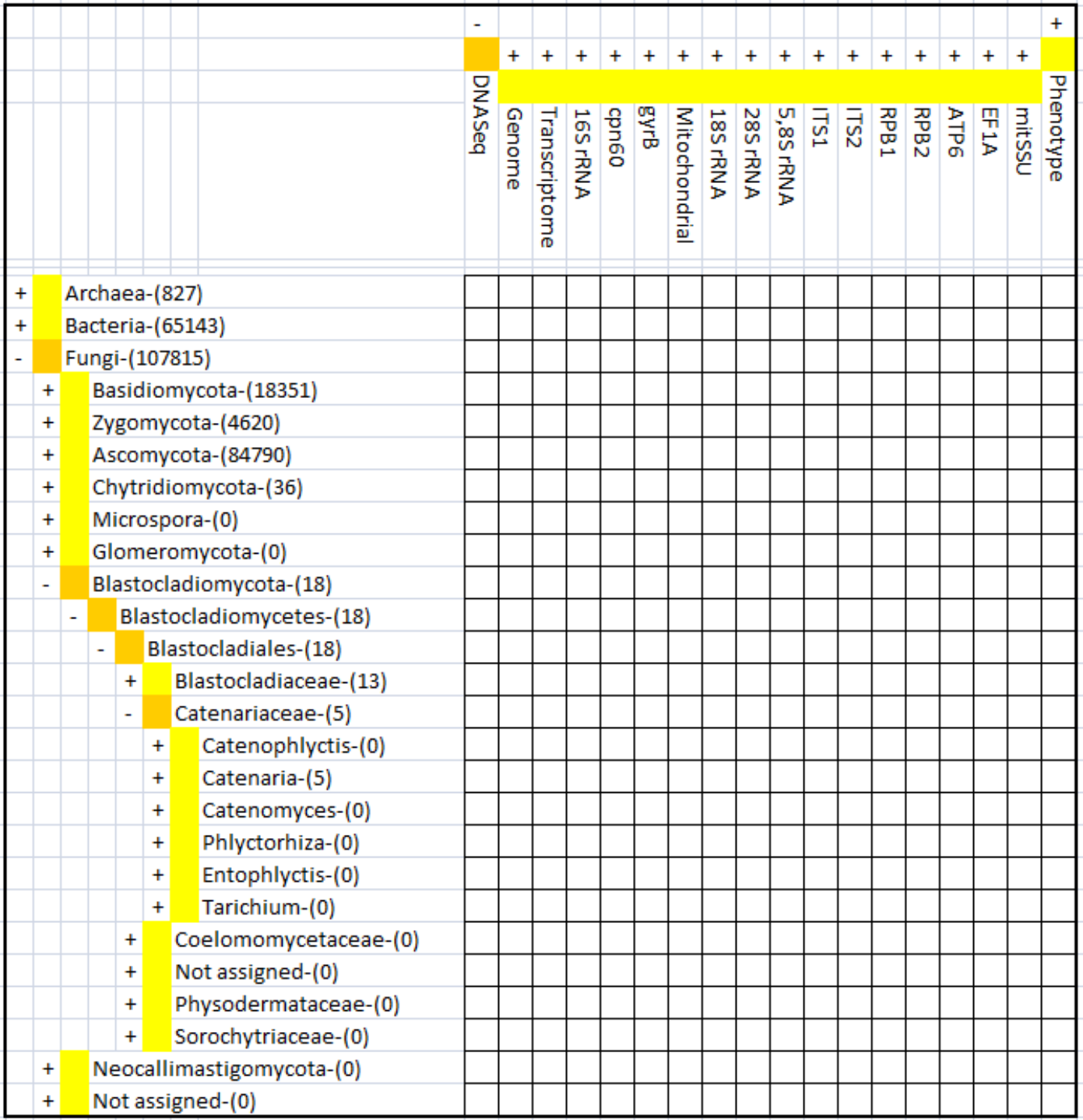
MIRRI
MICROBIAL RESOURCE RESEARCH INFRASTRUCTURE

To give the end-user convenient tools to see what knowledge on what strains he can get and what data can not be found in the particular culture collections.

To show all the information space of strain data kept in partners databases, to give the opportunity of comparison, cuts, to see what is done, the gaps, the contact data, etc. The main idea – to make research easier, to cut the time. One-stop access.

The data standard may have to keep all the types of data that partners are ready to share on-line, we could call it Extended Datasets or Full Datasets, its size could be more that 300 fields. And we may need ontologies for all of them.

	Fungi	Bacteria	Yeasts	Actinomycetes	Archaea	Phage	Microalgae	Virus	Antibody	Total
 Sludge/Wasterwater										NaN
 Soil										NaN
 Sediment										NaN
 Fermentation products										NaN
 Plant-associated										NaN
 Host-associated										NaN
 Human-associated										NaN
 Water										NaN
 Microbial-mat/Biofilm										NaN
 Air										NaN
 Genetic engineering strain										NaN
 Food										NaN
 Others										NaN
Total	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN


DNASeq
Genome
Transcriptome
16S rRNA
cpn60
gyrB
Mitochondrial
18S rRNA
28S rRNA
5,8S rRNA
ITS1
ITS2
RPB1
RPB2
ATP6
EF1A
mitSSU
Phenotype
+ Archaea-(827)
+ Bacteria-(65143)
- Fungi-(107815)
+ Basidiomycota-(18351)
+ Zygomycota-(4620)
+ Ascomycota-(84790)
+ Chytridiomycota-(36)
+ Microspora-(0)
+ Glomeromycota-(0)
- Blastocladiomycota-(18)
- Blastocladiomycetes-(18)
- Blastocladiales-(18)
+ Blastocladiaceae-(13)
- Catenariaceae-(5)
+ Catenophlyctis-(0)
+ Catenaria-(5)
+ Catenomyces-(0)
+ Phlyctorhiza-(0)
+ Entophlyctis-(0)
+ Tarichium-(0)
+ Coelomomycetaceae-(0)
+ Not assigned-(0)
+ Physodermataceae-(0)
+ Sorochytriaceae-(0)
+ Neocallimastigomycota-(0)
+ Not assigned-(0)

Desirable quality level

MIRRI
MICROBIAL RESOURCE RESEARCH INFRASTRUCTURE

The quality criteria:

1. Functions

- The most fast response of the system, maximum help with the request options and parameters, the maximum clarity of problem if data base response does not give the full answer, the easiest ways to the information required.

- To keep the subject of request: if the WEB page with answer have additional links, the pages addressed by them should give more information just on initial user request .

- **One question should give one page or one file answer.**
- **Attractive interface and intuitive navigation.**

2. Data

- Database content must be curated.

How it should be done

MIRRI
MICROBIAL RESOURCE RESEARCH INFRASTRUCTURE

We need to know what data is out there and how we link to them and make our data interoperable with them. We need to know what data we need to add and how we go about making them available. We need to work closely with ELIXIR and other research infrastructures such as EU-OPENSCREEN. We need to work with the other players STRAININFO.NET, WDCM - their Global strain database. We need to work with the user to see how best we deliver this information. We need strategy - all the rest is detail that we can prioritize and deal with once we have the concept of what the MIRRI system will look like.

MIRRI needs to identify the data sets maintained by others that we need to link to facilitate data mining. We need user input to help us design this aspect but we already know some of the data types that must be shared. This is an area we need more work; for example we cannot ask collections to screen and re-characterize all their organisms but by linking out to different datasets we can create data mining tools that can help predict properties or identify potential for the user. This would be utilizing a combination of literature databases, chemistry, habitat and ecosystem information, taxonomic hierarchy and relationships, etc; some of the databases to provide this exist whereas in other areas such as environmental or ecosystem data may not be so

Conclusions

MIRRI
MICROBIAL RESOURCE RESEARCH INFRASTRUCTURE

The goals that look real at the moment:

1. **Unification of Standard Operational Procedures (SOP), data structures, software tools.**
2. **Construction of BRC knowledge system**
3. **Presentation/search of BRC services**

Content search of microorganisms is an ambitious goal (problem 4). The first difficulty is ontologies. Problems 5 and 6 may be considered as well.

This is result of the first year of research. We have two more years for analysis in top-down schema.

Thank you

Vasilenko Alexander
vanvkm@gmail.com