# Information system for culture collections: what is desirable

Vasilenko Alexander, Ozerskaya Svetlana, Stupar Oleg

VKM

- Information systems – GBRCN, WDCM, MIRRI, CABRI, BRIO
- Desirable components – a look from inside
- Desirable quality level

if we try to maximize the system efficiency

# Interdependence of system 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

1. Unification of Standard Operational Procedures (SOP), data structures, software tools. Help for orphan collections
2. Content search of microorganisms based on their properties
3. Navigation in information space of microbiology, bioinformatics, biotechnology, agriculture, medicine
4. Research tasks with the use of the human language

# 1. A way toward unification

- Stable terminology shared by microbiological and IT communities.
- Common formats of data processing operations and network access tools suitable for most of culture collections (CC). Shared information standards. Unification of informational SOPs. Example of CABRI.
- Help for the orphan collections. Example WDCM.
- Common software tools.
- Search for microorganisms according their names and CC numbers inside joint catalogue. CABRI MDS, OECD MDS, WDCM MDS standards.
- Ontology of microorganism names.
- Examples of advanced solutions in Straininfo, WDCM, Mycobank.

# 2. Content search based on properties

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.

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

# Fields description schema

1. Field's name
2. Short description of the content
3. Specification of the content

3.1. Detailed description of the content, difference from the other fields that may look similar

3.2. List of subfields and their descriptions

3.3. Specification of the coding for the field and subfields

3.4. The list of possible values (if short) or a reference to file with long list, reference to thesaurus, or reference to ontology

3.5. Reference to a manual which describes this field content

3.6. Reference to external standard used for this field

3.7. Samples of correct coding for this field.

# Organism name

1. "Organism_Name"
2. Text that defines the actual name of the strain in your collection.
3. Specification schema, proposed by prof. Alexander Kovalenko (BIN RAS, Saint Petersburg, Russia)

3.1. It is supposed that in the field "Organism Name" you try to give correct nomenclatural name of the strain

3.2. "Organism_Name" consists of subfields:

- "Genus" - Actual name of genus in your collection, for example "Fusarium"
- "Species" - Actual name of species in your collection, for example "bulbigenum"
- "Author_first" - Author(s) and year for first description of this species epitet. If more than one author they are separated by symbol "et", example "Cooke et Massee 1886"
- "Author" - Author(s) and year for description of this combination genus+species. If more than one author they are separated by symbol "et" or "and", or "et al." - according the naming rules presented in section 3.5. For example "Cooke et Massee 1887"
- "varietet" - type of intraspecies taxon, possible values: "var.", "subsp.", "f.", "f.sp."
- "Subspecies" - actual name of subspecies, for example "aechmeae"
- "Sub_Author" - Author and year for description of this combination "Genus"+"Species"+"varietet"+"Subspecies",
  Example of "Sub_Author" for combination "f.sp."+"aechmeae" is "Sauthoff et Gerlach 1958"
- "Validation_status" - refer to how the name became valid (the same system is used in Bergey"s Manual of Systematic Bacteriology):
  "Validation_status"=VP - stands for Validily Published (in IJSB/IJSEM)
  "Validation_status"=VL - stands for Validation Lists (if the name was published elsewhere and later was validated by inclusion on a validation list in IJSB/IJSEM)
  "Validation_status"=AL - stands for Approved Lists of Bacterial Names (if a name was included in the Approved Lists published in 1980 in IJSB)
  Examples: if "Validation_status"=VP, "Genus"=Ensifer, "Species"=fredii, "Author_first"=Scholla and Elkan 1984, "Author"=Young 2003
  then "Organism_Name"=Ensifer fredii (Scholla and Elkan 1984) Young 2003 VP

3.3. Organism Name is a string consisting of the subfields, separated by space symbol (" "). Possible combinations of subfields are valid:

- "Genus" "Species" "Author", for example "Fusarium bulbigenum Cooke et Massee 1887"
- "Genus" "Species" "Author" "varietet" "Sub_Author", for example "Fusarium bulbigenum Cooke et Massee 1887 f.sp. aechmeae Sauthoff et Gerlach 1958"

All the subfields are strings with no symbols "-_=+~!@#$%^*()|/<>|;:," inside.

3.4. If "Organism_Type" is "Fungi" or "Yeast", the list of possible values for "Genus" is given in Dictionary of the Fungi, Ed. P.M.Kirk, CABI, UK, P.F.Cannon, CABI, UK, J.A.Stalpers, CBS, The Netherlands, D.W.Minter, CABI, UK. 10th Edition, 2008. - 784 p. The list values for all the other subfields of "Organism_Name" is given in http://www.indexfungorum.org/ and http://www.mycobank.org/ for each value of "Genus" separately.
If "Organism_Type" is "Archaea" or "Bacterium" the list of possible values for "Organism Name" is given in document "LPSN. List of Prokaryotic names with Standing in Nomenclature" see http://www.bacterio.cict.fr/.

3.5. Reference to manuals which specify Organism Name coding rules:
For "Archaea" and "Bacterium": International Code of Nomenclature of Bacteria - International Committee on Systematics of Prokaryotes (ICSP) (www.the-icsp.org/),
For "Fungi" and "Yeast": International Code of Nomenclature for algae, fungi, and plants - Nomenclature Committee for Fungi (http://www.ima-mycology.org/CFF/)

# RDS fields with descriptions

- **Organism_Name, 7 subfields:**
  - Genus
  - Species
  - AuthoritySp
  - Variant
  - Name_variant
  - AuthoritySubSp
  - Validation_status
- **Strain number**
- **Organism type**
- **Status**
- **ReceivedFrom**
- **History_of_Deposit**
- **Collect_Country**
- **Collect_GeoLocation**
- **Collect_Habitat_Locality**
- **Collect_Latitude**
- **Collect_Longitude**
- **Collect_Altitude**
- **Collect_Depth**
- **Collect_Habitat_Locality**
- **Collect_Source**
- **Other_Collection_Numbers**
- **GrowthMedium**
- **IncubationTemp**
- **Preservation**
- **Restrictions**
- **Accession_Date**
- **Reference**
- **DNASeq, 16 subfields:**
  - Genome sequence
  - Transcriptome
  - 16S rRNA
  - cpn60
  - gyrB
  - Mitochondrial Genome
  - 18S rRNA
  - 28S rRNA
  - 5,8S rRNA
  - ITS1
  - ITS2
  - RPB1
  - RPB2
  - ATP6
  - EF1A
  - mitSSU

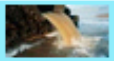
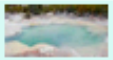
# 3. Navigation in information space

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.

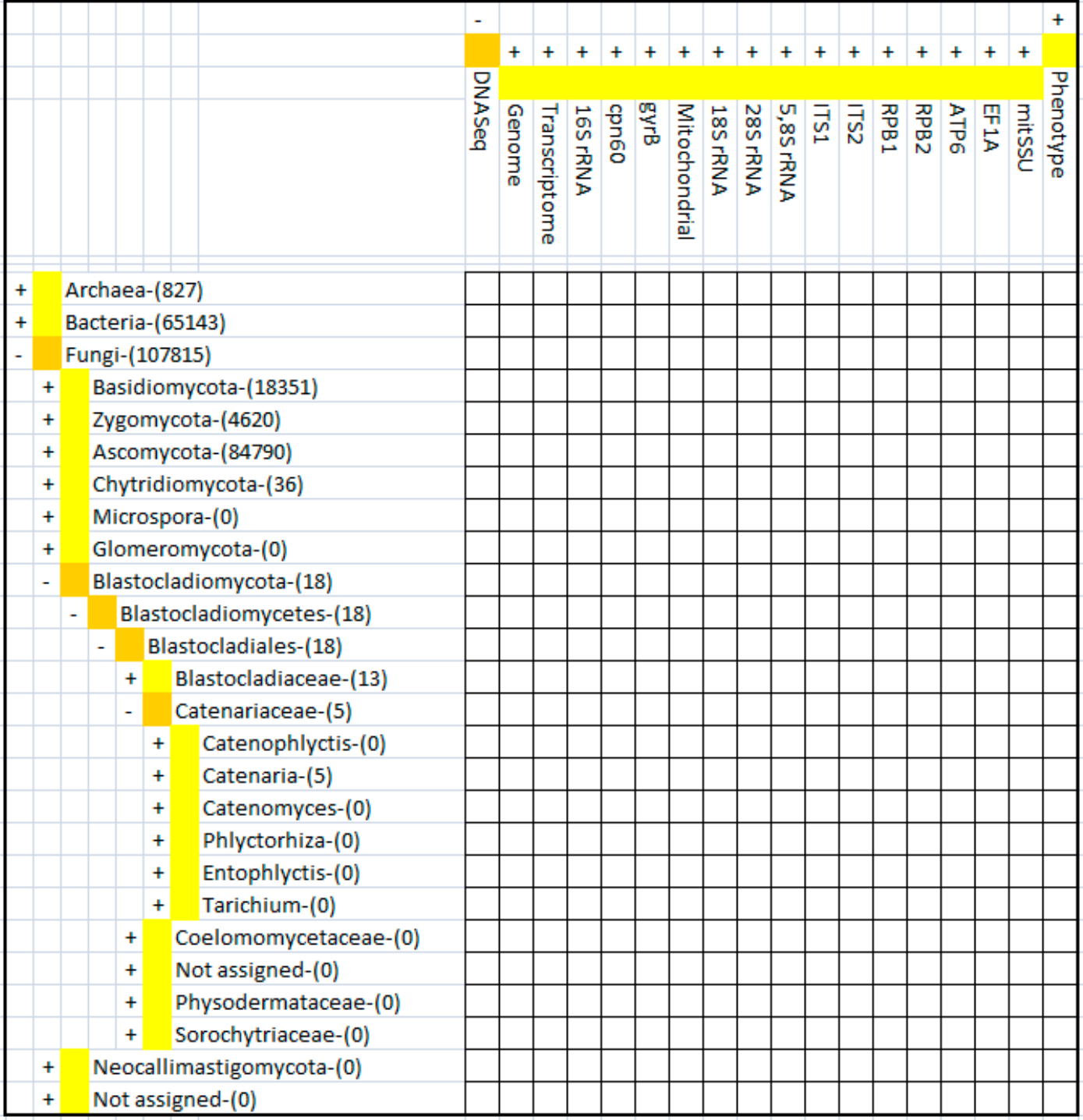
The data standard may have to keep all the types of data that partners are ready to share on-line, we could call it Extanded Datasets or Full Datasets, its size could be more that 300 fields. And we may need ontologies for all of them.

List by Strain Name | Strain Number | List by CCs | List by Isolation Sources | Species Info | Map View

| | 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 |

Browse taxonomic tree Search

- 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)


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

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 (See example *)
- One question should give one page or one file answer

Example:

- - if the WEB site shows information on some culture and gives links to bibliography and sequencing, the addressed pages should give bibliographical and sequencing information just on the same culture.

2. Data

- Database content must be curated

A total of 460 Species , 23 page

All

- *aspegillus awamori* (2),Species
- *aspergillus* (9),Species
- *aspergillus fumigatus* (15),Species
- *aspergillus niger* (5),Species
- *aspergillus calidoustous* (2),Species
- *aspergillus flavipes* (1),Species
- *aspergillus flavus* (4),Species
- *aspergillus fumigatiaffinis* (3),Species
- *aspergillus fumigatus* (1),Species
- *aspergillus insuetus* (1),Species
- *aspergillus japonicus* (1),Species
- *aspergillus niger* (2),Species
- *aspergillus ochraceus* (2),Species
- *aspergillus penicillioides* (1),Species
- *aspergillus sclerotiorum* (1),Species
- *aspergillus terreus* (15),Species
- *aspergillus tubingensis* (1),Species
- *aspergillus versicolor* (4),Species
- *aspergillus acanthosporus* (2),Species

# Conclusions

On each step of this brief analysis we see that merits of our proposals should depend on response of experts in microbiology.

To illustrate this, let us look at desirable ontology of Names. It would be natural if taxonomy system is inside its objects structure, and all the corrections of this system might be the activities of microbiological communities (Bergey Trust, Euzeby Data base, MycoBank, etc.). The job of WDCM is only to give convenient access for them, and WDCM taxonomy tree will be hopefully always correct!

# Thank you

Vasilenko Alexander
vanvkm@gmail.com