Data integration in the MicrobeDB.jp using Semantic Web technology

Hiroshi Mori, Ikuo Uchiyama, Yasukazu Nakamura, Hideaki Sugawara, Ken Kurokawa, and MicrobeDB.jp Project Team

1) Tokyo Institute of Technology, 2) National Institute for Basic Biology, 3) National Institute of Genetics, Japan
Many microbial databases (DBs) exist ...

Which DBs should we use?

- Ortholog
- Taxonomy
- Culture Collection
- Genome
- Metagenome
- Gene Function
- Pathogen
- Collection
- Culture
Microbes inhabit almost everywhere on Earth and interact with their environments.

Knowledge of microbes will have high potential scientific and commercial applications.
Promoting the Integrated Use of Life Science Databases in Japan

- FY 2007-2010 “Integrated Database Project”
  → Database Center for Life Science (DBCLS)

- FY 2011-
  → National Bioscience Database Center (NBDC)

About NBDC
- Established in April 2011
- As part of the Japan Science and Technology Agency (JST), a funding agency supported by MEXT

URL: http://biosciencedbc.jp/?lng=en
Activities by NBDC

1. Formulation of strategies related to coordination and integration of DBs, and international cooperation

2. Creation and management of a portal website from existing life science DBs [http://biosciencedbc.jp/?lng=en](http://biosciencedbc.jp/?lng=en)

3. Funding of R&D of new technology necessary for organizing and linking life science DBs

4. Funding of R&D that coordinate existing and emerging DBs in specific research fields

   Includes microbes (PI: Ken KUROKAWA)

Aim of [Microbe DB JP](http://biosciencedbc.jp/?lng=en)

- to integrate several microbial data (include omics, taxonomy/cultures, habitats)
- using semantic web technology
Microbe DB integrates lots of data related to microbes. Especially, we integrate the microbial data that can be linked to genomes.

How to simplify the process of integration of other domain’s data?

Ortholog: MBGD
Genome: GTPS/RefSeq
Annotation: TogoAnnotation

Taxonomy: NCBI Taxonomy
Culture Collection: NBRC/JCM

Metadata: INSDC SRA
Metagenome: INSDC SRA

Red color indicates our collaborators.
RDF is a standard data model of Semantic Web technology

RDF (Resource Description Framework)
Data model which uses Triples (Subject – Predicate – Object)

**gtps:**Gene1  **rdfs:**label  “16S rRNA gene”

URI node can be linked to other nodes

To prepare data in RDF, the database management system automatically recognize same resources (same URI).
How to integrate the data from two different DBs?

1. When two DBs use same URI, already two DB’s data are integrated.
2. If not, you can integrate two DB’s data by adding one Triple (db1:A owl:sameAs db2:B).

You don’t need to place all of these data in one DB management system.

How can we discriminate whether two DB’s resources are same or not?
You should describe your resource by using some **Ontologies**

**Ontology** is a structured controlled vocabulary to describe properties and types of resources.

For example, to answer: What is soil? What is a relationship between soil and sand?

**MEO (Microbes Environmental Ontology)**

**PDO (Pathogenic Disease Ontology)**

**MCCV (Microbial Culture Collection Vocabulary)**

**MSV (Metagenome Sample Vocabulary)**

**MPO (Microbial Phenotype Ontology)**

**MBGD Ortholog Ontology**

Most of them can be obtained from BioPortal
We have converted most of our data to RDF, developed many ontologies, and developed a RDFized microbial DB.

More than 1 billion Triples!

---

**Ortholog:** MBGD

**Genome:** GTPS/RefSeq

**Annotation:** TogoAnnotation

**Taxonomy:** NCBI Taxonomy

**Culture Collection:** NBRC/JCM

**Metadata:** INSDC SRA

**Metagenome:** INSDC SRA

Red color indicates our collaborators.
<table>
<thead>
<tr>
<th></th>
<th>Field</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Strain_Number</td>
</tr>
<tr>
<td>2</td>
<td>Other_Collection_Numbers</td>
</tr>
<tr>
<td>3</td>
<td>Name</td>
</tr>
<tr>
<td>4</td>
<td>Organism_Type</td>
</tr>
<tr>
<td>5</td>
<td>History_of_Deposit</td>
</tr>
<tr>
<td>6</td>
<td>Date_of_Isolation</td>
</tr>
<tr>
<td>7</td>
<td>Isolated_from</td>
</tr>
<tr>
<td>8</td>
<td>Geographic_Origin</td>
</tr>
<tr>
<td>9</td>
<td>Status</td>
</tr>
<tr>
<td>10</td>
<td>Optimum_Temperature_for_Growth</td>
</tr>
<tr>
<td>11</td>
<td>Maximum_Temperature_for_Growth</td>
</tr>
<tr>
<td>12</td>
<td>Minimum_Temperature_for_Growth</td>
</tr>
<tr>
<td>13</td>
<td>Medium</td>
</tr>
<tr>
<td>14</td>
<td>Application</td>
</tr>
<tr>
<td>15</td>
<td>Literature</td>
</tr>
</tbody>
</table>
Overall data structure of MicrobeDB.jp
http://microbedb.jp/
Keyword example: lake

- Lake meo:pond is_a meo:lake
- Strain_A mccv:isolation_source meo:pond Strain_A

Genome sequenced strains isolated from lake

Abundant Orthologs in metagenome samples obtained from lake

JCM/NBRC Strains isolated from lake

Metagenome samples obtained from lake

Taxonomic composition of 16S amplicon sequencing which sampled from lake

MicrobeDB.jp will facilitate the exploration of the existing scattered information of microbes.
Acknowledgements

· Ken Kurokawa (Tokyo Institute of Technology)
  Junichi Takehara, Koji Yoshino, Nozomi Yamamoto, Takuji Yamada, Fumikazu Konishi

· Yasukazu Nakamura (National Institute of Genetics, DDBJ)
  Takatomo Fujisawa, Eri Kaminuma, Hideaki Sugawara

· Ikuo Uchiyama (National Institute for Basic Biology)
  Hirokazu Chiba, Hiroyo Nishide

Advisor (DataBase Center for Life Science)
  Shinobu Okamoto, Shuichi Kawashima, Toshiaki Katayama, Yasunori Yamamoto, Shoko Kawamoto

NBRC Culture Collection data
  Ken’ ichiro Suzuki, Masami Ichihara, Natsuko Ichikawa

JCM Culture Collection data
  Moriya Ohkuma, Takuji Kudo

Funding