ESF Workshop, Ginevra, Septembe 22nd, 2003

European Biological Resources Centers Network (EBRCN)

and metabolic pathways

Paolo Romano
National Cancer Research Institute, Genova
(paolo.romano@istge.it)

Summary

- Some ideas on data integration in biology
- CABRI: a “one stop shop” for biological resources
- EBRCN: interconnected biological resources database

Degrees of information integration

- Tightly integrated systems
  - Data: local warehouse
  - Applications: centralized or CORBA
  - Processes: static, repetitive services
  - Integration: early or predefined
  - Transparency: high
- Dynamicly (loosely) integrated systems
  - Data: decentrated, dynamic integration
  - Applications: Web Services
  - Processes: dynamic, based on users' requirements
  - Integration: on demand or data mining
  - Transparency: medium to low (interaction)

Integration longevity

- Integration needs stability
  - Standardization……
  - Good domain knowledge
  - Well defined data
  - Well defined goals

- Integration fears
  - Heterogeneicity of data and systems
  - Uncertain domain knowledge
  - Fast evolution of data
  - Highly specialized data
  - Lacking of predefined, clear goals
  - Originality, experimentalism (“*let me see if this works*”)

Biology data banks are distributed

- Distributed data banks means:
  - Different DBMS
  - Different data structures
  - Different information
  - Different meanings
  - Different data distribution methods

Goals of the integration

- Integration is needed in order to:
  - Achieve a better and wider view of all available information
  - Carry out analysis and/or searches involving more databases and softwares in one step only
  - Carry out a real data mining

Integration of databanks

- Integration of databanks implies:
  - Accurate analysis and definition of involved "biological objects"
  - Analysis of available information / data
  - Identification of logical links between objects and and definition of related data links between dbs
  - Definition and implementation of common data interchange formats, methods, tools

Integration of biological information

In biology:

- Goals and needs of researchers evolve very quickly according to new theories and discoveries
- A pre-analysis and reorganization of the data is very difficult, because data and related knowledge vary continuosly
- Complexity of information makes it difficult to design data models which can be valid for different domains and over time

Integration methods

- Explicit (reciprocal) links (xrefs)
- Implicit links (e.g., names)
- Common contents (vocabularies)
- Object oriented models
- Relational schemas
- Ontologies

CABRI: Objectives

Common Access to Biological Resources and Information (www.cabri.org)

- Setting Quality Management Guidelines
- Distributing biological resources of the highest quality
- Integrating searches and access to catalogues
- One-stop-shop for quality resources
- Ad hoc search (CABRI Simple Search)
- Shopping cart (pre-ordering facility)

CABRI: Partners and resources

Partners:

- INSERM (coordination)
- BCCM, CBS, DSMZ, ECACC, HGMP-RC, ICLC, NCCB (resources)
- HGMP-RC, IST, CERDIC (ICT)

Resources:

- Microorganisms (bacteria, yeasts, fungi)
- Cells (animal and human cell lines, hybridomas, HLA typed B lines)
- Plasmids, phages, viruses, DNA probes
- Overall, more than 100.000 items in catalogues

CABRI: Resources

	DP	B/A	F/Y	PL	PH	PC	PV	AC	HYB	BC
BCCM		X	X	X						
CABI		X	X							
CBS		X	X							
CIP		X								
DSMZ		X	X	X	X	X	X	X		
ECACC	X							X	X	X
ICLC								X		
NCCB		X		X	X					
NCIMB		X			X					

CABRI: why SRS

- Yes because
  - o Manages heterogeneous databases
  - o Flat file format
  - o Simple and effective interface
  - o Internal and external links
  - o Link operator
  - o Easily expandible (new databases)
  - o Flexibility in creation of indexes

CABRI: why SRS

- No because
  - Local databases, not remote (updates)
  - “Difficult” language (Icarus)
  - Commercial software (not free)

CABRI: data structure

For each material, three data sets identified:

- Minimum Data Set (MDS): essential data, needed to identify individual resources
- Recommeded Data Set (RDS): all data that are useful to describe individual resources
- Full Data Set (FDS): all data available on the resources

CABRI: data structure

For each information, data input and authentication guidelines, including:

- Detailed textual description of the information
- In-house reference lists of terms and controlled voca bularies
- Predefined syntaxes (e.g., Literature, scientific names)

CABRI: Data sets

Data set	Field label	Catalogues
MDS	Strain_number	All
MDS	Other_collection_numbers	All
MDS	Name	All
RDS	Race	All
MDS	Organism_type	All
MDS	Restrictons	All
MDS	Status	All
MDS	History	All
RDS	Misapplied_names	All
RDS	Substrate	All
RDS	Geographic_origin	All
RDS	Sexual_state	All
RDS	Mutant	All
FDS	Genotype	DSMZ
………	……….	……….

CABRI: Name field

Field	Name
Description	Full scientific and most recent name of the strain. It includes: ▪Genus name and species epithet ▪Subspecies ▪Pathovar ▪Authors of the name ▪Year of valid publication or validation ▪Approbation of the name
Input process	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: <u>DSMZ list of bacterial names</u>**
Required for	MDS

CABRI: Reference paper field

Field	Reference paper
Description	Original paper [if available]
Input process	New entries: JournalTitle Year; Volume(issue): beginning page#-ending page# The title is abbreviated following international standard rules (ISSN). Abbreviations 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
Required for	MDS

```
Strain_number LMG 1(t1)
Other_collection_numbers CCUG 34964;NCIB 12128
Restrictions Biohazard group 1
Organism_type Bacteria
Name Phyllobacterium rubiacearum, (ex Knsel 1962) Knsel 1984 VL
Infrasubspecific_names -
Status Type strain
History <- 1973, D.Knsel
Conditions_for_growth Medium 1, 25C
Form_of_supply Dried
Isolated_from Pavetta zimmermannia
Geographic_origin Germany, Stuttgart-Hohenheim
Remarks Stable colony type isolated from LMG 1. See also Agrobacterium sp.
    LMG 1(t2)

Strain_number LMG 1(t2)
Other_collection_numbers -
Restrictions Either Biohazard group 1 or Biohazard group 2
Organism_type Bacteria
Name Agrobacterium sp.
Infrasubspecific_names -
Status -
Other_names Phyllobacterium rubiacearum, (ex Knsel 1962) Knsel 1984 VL
History <- D.Knsel (Phyllobacterium rubiacearum)
Conditions_for_growth Medium 16, 28C
Form_of_supply Dried
Isolated_from Pavetta zimmermannia
Geographic_origin Germany, Stuttgart-Hohenheim
Remarks One (t2) out of two stable colony types isolated from the original
    culture LMG 1.
```

CABRI: integration

For each catalogue:

- SRS and HTML links to reference dbs (media, synonyms, hazard, etc…)

For each material:

- Common data structure and syntax
- Integrated searches/results through SRS

CABRI: Extra features

CABRI Simple Search:

- Search by ID(s), name(s), all other fields
- Search by name(s) with synonyms support

CABRI Shopping cart:

- Set of mixed javascripts and perl scripts
- Pre-order facility (email or fax)

CABRI: Simple Search

Synonyms' support

- Only allowed for micro-organisms
- Managed through a perl script
- First searched terms are matched against synonyms' reference dbs with getz
- When available, names are added to the initial search and a new search is carried out
- Results are then displayed and a link to synonyms' dbs is added

EBRCN: Extending integration

European Biological Resource Centres Network

(www.ebrcn.org)

Wp1 Co-ordinate European BRC policies, prepare a co-ordinated European response to international initiatives on biodiversity and become the European focal point for BRCs

Wp2 Develop new and maintain existing quality standards for European BRCs

Wp3 Establish a framework to maximise complementarity and minimise duplication among European BRCs

Wp4 Introduce new techniques in Information Technology to the EBRCN to add value to current catalogue information and enhance accessibility

Wp5 Collate and disseminate relevant information to the BRCs

EBRCN: Workpackage 4

Workpackage 4

“Introduce new techniques in information technology to the EBRCN to add value to current catalogue information and enhance accessibility”

Objective

Link catalogue data to literature, to nucleotide and to related genetic databases

EBRCN: new links

For all catalogues:

- Links to Medline through Pubmed ID
- Links to representative EMBL records

For selected catalogues:

- Links to plasmids' maps (plasmids)
- Links to microscope images (microorganisms)
- Links to other dbs under evaluation

Interconnected Biological Resources Database

EBRCN: Linking to EMBL

- Test for linking to EMBL Data Library through SRS, without explicit IDs, gave negative results:
  - Links are different for different materials and can use various EMBL fields:
    - Organism (micro-organisms), Division (viruses and plasmids), Feature Table (definition of the source through Key, Qualifier, Description)
  - Annotation and indexing problems

EBRCN: EMBL links variability

- Annotation problems:
  - CBS 100.20 can be annotated as CBS 100.20 or CBS100.20
  - CBS 12345 can be annotated as CBS12345

- Indexing problems:
  - CBS 100.20 is indexed as CBS, 100 and 20
  - The dot is not included and is used as a separator
  - CABRI unique index key is “CBS 100.20”

EBRCN: Linking to EMBL (ii)

Examples of search:

- Query: Fungi & source & cbs 100.20

```
( ( ([emblrelease-FtKey:source] &
    [emblrelease-FtQualifier:strain] &
   ( ( [emblrelease-FtDescription:cbs] &
       [emblrelease-FtDescription:100] ) |
       [emblrelease-FtDescription:cbs100] ) &
       [emblrelease-FtDescription:20]) )
  < [emblrelease-Organism:fungi*] )
```

EBRCN: Linking to EMBL (iii)

A possible approach:

- Identify xrefs for linking from EMBL to CABRI catalogues, based on CABRI IDs
- A huge number of EMBL records could be linked to a single CABRI item
- Add links in EMBL and use these links when linking from CABRI (search by means of SRS)
- CABRI Ids included in EMBL data library and distributed with it

EBRCN: Extracted databases

- Extracted databases made available for SRS based sites in academic/no-profit Institutes
- Selected meaningful subset of information: MDS+link to main CABRI site
- FTP site with data and SRS syntax/structure files

CABRI & EBRCN: what next?

Following SRS and ITC developments:

- SRS 5.1 -> SRS 7.1 -> SRS 8
- Flat file -> XML -> Web Services

Adding contents:

- New catalogues
- New materials
- Links to further external dbs
- Extended catalogue contents (further characterization or improved data structure)

CABRI & pathways

Quality materials are essential for research:

- Extracted databases can be made available to the pathways community
- Information in catalogues could be enhanced by adding links to pathways dbs

Suggestions are welcome, esp. on:

- Links to further external dbs
- Extended catalogue contents (further characterization of materials OR improved data structure)

Some acknoledgements…..

A. Doyle (ECACC)
B. Dutertre (CERDIC)
J. Franklin (ASFRA)
D. Fritze (DSMZ)
F. Guissart (BCCM)
M. Kracht (DSMZ)
F. Malusa (IST)
D. Marra (IST)
L. Réchaussat (INSERM)
D. Smith (CABI)
E. Stackebrandt (DSMZ)
J. Stalpers (CBS)
G. Stegehuis (CBS)
M. Vanhoucke (BCCM)
B. Vaughan (HGMP-RC)