Linking CABRI catalogues to the EMBL Data Library: lessons learned and future perspectives

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

Outline of this presentation

- CABRI network services
- EBRCN project
- Methods for linking CABRI and EMBL
- Results
- Future perspectives

CABRI: the EU project

Common Access to Biological Resources and Information

Project funded by the EU from 1996 to 1999

Partners

- INSERM (coordinator L. Réchaussat)
- BCCM, CBS, DSMZ, ECACC, HGMP-RC, ICLC, NCCB (resources)
- HGMP-RC, IST, CERDIC (ICT)

Resources

- Microorganisms (bacteria, yeasts, fungi strains)
- Cells (animal and human cell lines, hybridomas, HLA typed B lines)
- Plasmids, phages, plant cells, plant cell viruses, DNA probes
- Overall, more than 90.000 items in catalogues

CABRI: the goals

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

CABRI: the SRS search engine

SRS was choosen because:

- it has a simple and effective interface
- it manages heterogeneous databases and it uses flat file format
- it offers both internal and external links and has a link operator
- it is easily expandible (new databases)
- it is widely used for many molecular biology database

CABRI: data structure

For each material, three data sets identified:

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

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

- Detailed textual description of the information
- In-house reference lists of terms and controlled vocabularies
- 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: DSMZ list of bacterial names**
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 Knösel 1962) Knösel 1984 VL
Infrasubspecific_names -
Status Type strain
History <- 1973, D.Knösel
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 Knösel 1962) Knösel 1984 VL
History <- D.Knösel (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: Extra features

Simple Search:

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

Shopping cart:

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

Links to reference databases:

- SRS and HTML links to reference databases (media, synonyms, hazard, etc…) for all catalogues

CABRI: Synonyms' support

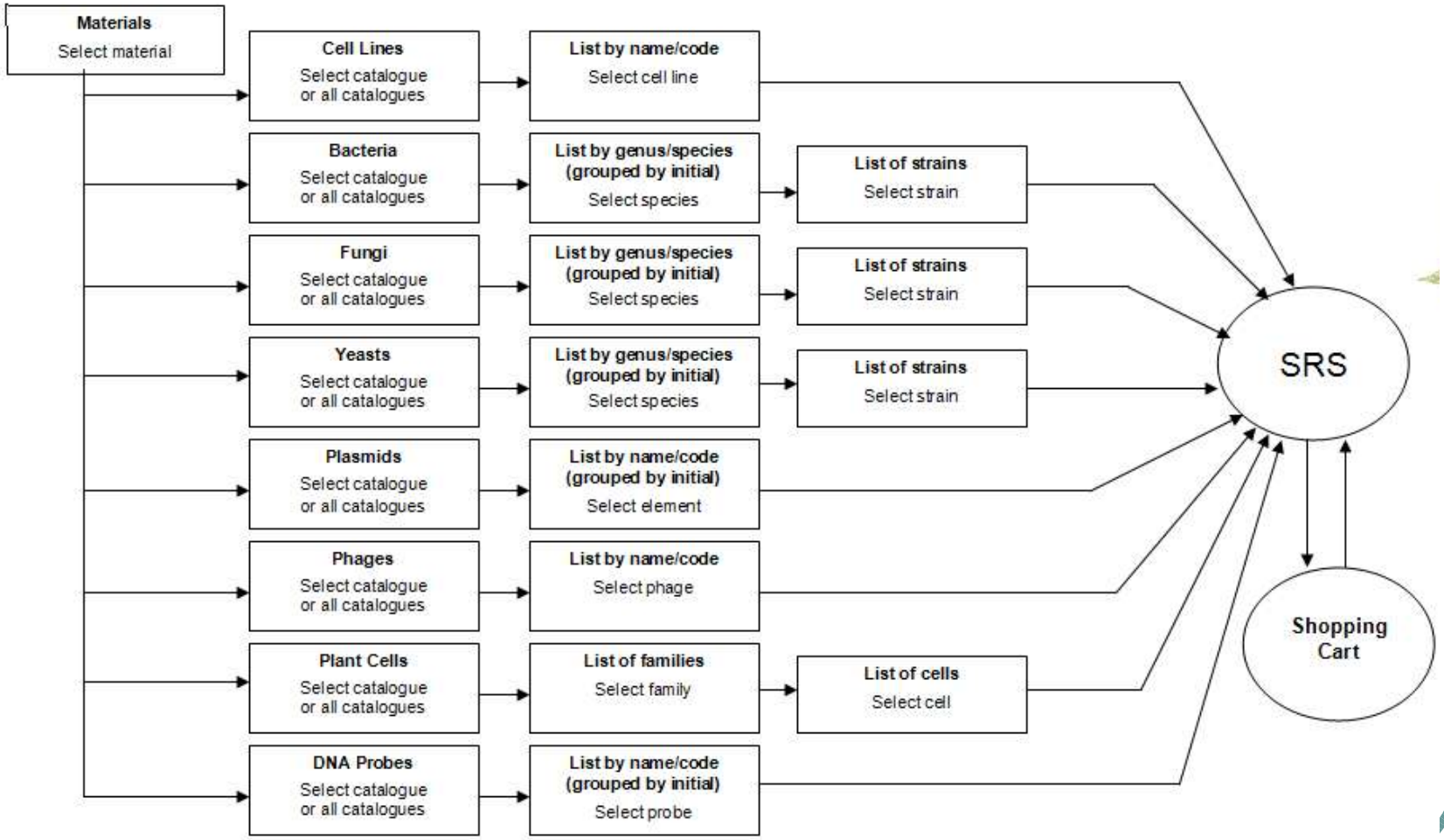
The synonyms' support allows to retrieve strains by using synonyms or alternate names

- Only allowed for micro-organisms
- Managed through a perl script
- Searched terms are first matched against synonyms' reference databases with getz
- Results are added to the initial search and a new search is carried out
- New results are then displayed
- Links to synonyms' databases are provided

CABRI: Contents updating

- New collections added after the end of the project:
  - CABI Biosciences (fungi, yeasts and bacteria)
  - CIP (bacteria)
  - NCIMB (bacteria and phages)
- No new catalogues since 2002
- Catalogues updated 1 to 2 times a year
- Now 28 catalogues for more than 110.000 items
- HyperCatalogue: hypertext for accessing records through a hyerarchical set of indexes

CABRI: HyperCatalogue



CABRI: collections and catalogues

	DNA probes	Bacteria strains	Fungi & yeasts strains	Plasmids	Phages	Cell lines	Hybrid omas	HLA typed B lines	Plant cells	Plant cell viruses
BCCM		11,912	15,865	681						
CABI		243	11,677							
CBS		1,038	31,746							
CIP		7,038								
DSMZ		8,366	2,296	284	89	526			476	426
ECACC	261					956	397	238		
ICLC						237				
NCCB		4,609		396	69					
NCIMB		6,986			69					
Total	261	40,192	61,584	1,361	227	1,719	397	238	476	426

CABRI: some limitations

- Closed system
- Lacking of external links
- Lacking of interoperability

Goals of the integration

- In biology, 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
- For biological resources, integration is useful for:
  - Improving catalogues descriptions of the resources by retrieving new data from well curated databases
  - Retrieving detailed information on resources while searching molecular biology databases
  - Improving visibility of high quality collections of biological resources
  - Extending use of biological materials of certified quality

Integration methods

From syntactical to semantical links:

- Explicit (reciprocal) links (xrefs)
- Implicit links (e.g., names)
- Common contents (vocabularies)
- Shared data models and schemas
- Ontologies

EBRCN: A focal point for BRCs

European Biological Resource Centres Network

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: Extending contents

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 external databases

Improve accessibility of catalogues

Create an interconnected biological resource database

EBRCN: links to external databases

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 (nomenclature, acronyms, genes) under evaluation

EBRCN: links to Medline

Syntax change:

add [PMID: <number>] after bibliographic reference

Links in place (> 7000):

Plasmids: LMBP (375), NCCB (30)

Cell lines: ICLC (294), DSMZ (905)

Fungi: CBS (454)

Yeasts: CBS (1132)

Phages: NCCB (30)

Literature reference file: DSMZ (3818)

EBRCN: revised 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

EBRCN: links to other databases

Plasmids' maps:

- Syntax:
  - New FDS field: 'External_links map <name>'
- Links in place:
  - Plasmids: LMBP (777)

Images of micro-organisms:

- Syntax:
  - New FDS field: 'External_links image <name>'
- Links in place:
  - None (waiting for next catalogues' update)

EBRCN: EMBL Data Library

- EMBL Data Library is the European database for DNA sequences
- It is updated daily and a coordination with NCBI and DDBJ ensures its completeness
- Static releases are issued every three months
- It is offered at EBI and other centers by means of SRS

EBRCN: EMBL Data Library size

- EMBL Data Library 74 (Mar 2003):
  - Sequences: 23,234,788, Bases: 30,356,786,718
- EMBL Data Library 81 (Dec 2004):
  - Sequences: 40,696,839, Bases: 44,285,259,441
  - WGS sequences: 5,408,558, Bases: 34,986,041,399
- EMBL Data Library 82 (Mar 2005):
  - Sequences: 43,246,005, Bases: 46,927,070,905
  - WGS sequences: 6,228,397, Bases: 38,207,643,477
- Size: 7,3% more vs 81 (3 months), 112,9% vs 74 (24 months)

EBRCN: EMBL links

Test have been conducted to identify how to link “on-the-fly” to EMBL Data Library through SRS

Tests performed on:

- Bacteria and Archaea
- Animal and Human Cell Lines
- Fungi and Yeasts
- Plasmids
- Viruses

EBRCN: EMBL links variability

Links vary according to the materials

Links can be created on the basis of various EMBL fields:

- All-text (not very useful)
- Organism (for micro-organisms)
- Division (useful for viruses and plasmids)
- Feature Table data (it allows a correct definition of a source through Key, Qualifier, Description)

Features

Key	Location	Qualifier	Value
source	1..513	db_xref	taxon:56450
		mol_type	genomic DNA
		organism	Xanthomonas cassavae
		strain	LMG673
misc_rna	1..513	note	16S-23S ribosomal RNA intergenic spacer

EBRCN: EMBL links variability

Example:

set up a link from the CBS fungi catalogue to EMBL for CBS 100.20 strain

Fields and values:

Organism: fungi

Ft-Key: source

Ft-Qualifier: strain

Ft-Description: "cbs 100.20"

EBRCN: EMBL links variability

Annotation problems in EMBL:

- CBS 100.20 can be annotated as:
  - CBS 100.20 or CBS100.20 or CBS 10020 or CBS10020
- CBS 112345 can be annotated as CBS12345

Indexing problems at EBI:

- 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 on-the-fly

Example of search “on-the-fly”:

Searching for fil. fungi strain CBS 100.20
It involves: 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: Alternative approach

Test for linking “on-the-fly” to EMBL Data Library through SRS, without explicit IDs, gave negative results

An alternative approach:

- Identify “a priori” all existing crossreferences between EMBL and CABRI catalogues, based on CABRI IDs
- Add properly formatted links to CABRI in EMBL and use these links when linking from CABRI

CABRI IDs included in EMBL data library and distributed with it

EBRCN: procedure

Work carried out with EMBL 80 and CABRI catalogues version 2004.1

- Common SRS site for CABRI and EMBL established
- Modified indexing -> common keys format
- SRS internal links established for micro-organisms
- Automatically identified links downloaded and analysed
- List of dubious references sent to collections
- Some comments returned
- Data submitted to EBI and included in EMBL

EBCRN: Common site established

EBCRN: Common format index keys created

CABRI indexing: by whole ID

CBS 100.20 -> ‘CBS 100.20’

EMBL indexing: by single words

CBS 100.20 -> ‘CBS’ + ‘100’ + ’20’

CBS100.20 -> ‘CBS100’ + ’20’

Common indexing: name (only letters), possibly followed by space, followed by string (including letters, numbers, dot, dash), punctuation removed

CBS 100.20 -> ‘CBS10020’

CBS100.20 -> ‘CBS10020’

Special case (not currently managed):

NCCB LMD and Phabagen bacteria catalogues

EMBL (before)

```
ftqualx: ~ {$In:[ftvals c:qual] $Out}
            '/'
            ( /translation="[^"]*/
             | /[^= \n]+/ {$Uniq:name}) ('=' (/[a-zA-Z0-9]+/ {$Uniq:word} | /./)*)? ~
```

EMBL (after)

```
ftqualx: ~ {$In:[ftvals c:qual] $Out}
            '/'
            ( ( /strain/ {$Uniq:name} '=' (/[A-Za-z]+/ {$Wrt:[c:word]} ' '?
                                          /[A-Z0-9a-z.\n-]*/ {$App:[s:$StrRep:[s:$Ct
                                                                          r:"[ \n.-]" to:'']]}
    | /./ )*)
                | /translation="[^"]*/
                | /[^= \n]+/ {$Uniq:name}('=' (/[a-zA-Z0-9]+/ {$Uniq:word} | /./)*)?) ~
```

CABRI (before)

```
strain_no: ~ {$In:[fields c:strain_no] $Out}
               word ' '
               (/[ 0-9a-zA-Z_\\/\\.\\+\\*-]+/ {$Wrt})* ~
```

CABRI (after)

```
strain_no: ~ {$In:[fields c:strain_no] $Out}
               word ' '
               (/[ 0-9a-zA-Z_\\/\\.\\+\\*-]+/  {$Wrt:[s:$StrRep:[s:$Ct r:"[ ]" to:'']]})* ~
```

EBCRN: SRS links EMBL - CABRI

```
$Link: [from:$EMBLRELEASE_DB
        to:$BCCM_LMG_DB
        fromField:$DF_FtDescription
        toField:$DF_CABRI_Strain_number]

$Link: [from:$EMBLRELEASE_DB
        to:$CBS_BACT_DB
        fromField:$DF_FtDescription
        toField:$DF_CABRI_Strain_number]
```

EBRCN: Automatic identification of links

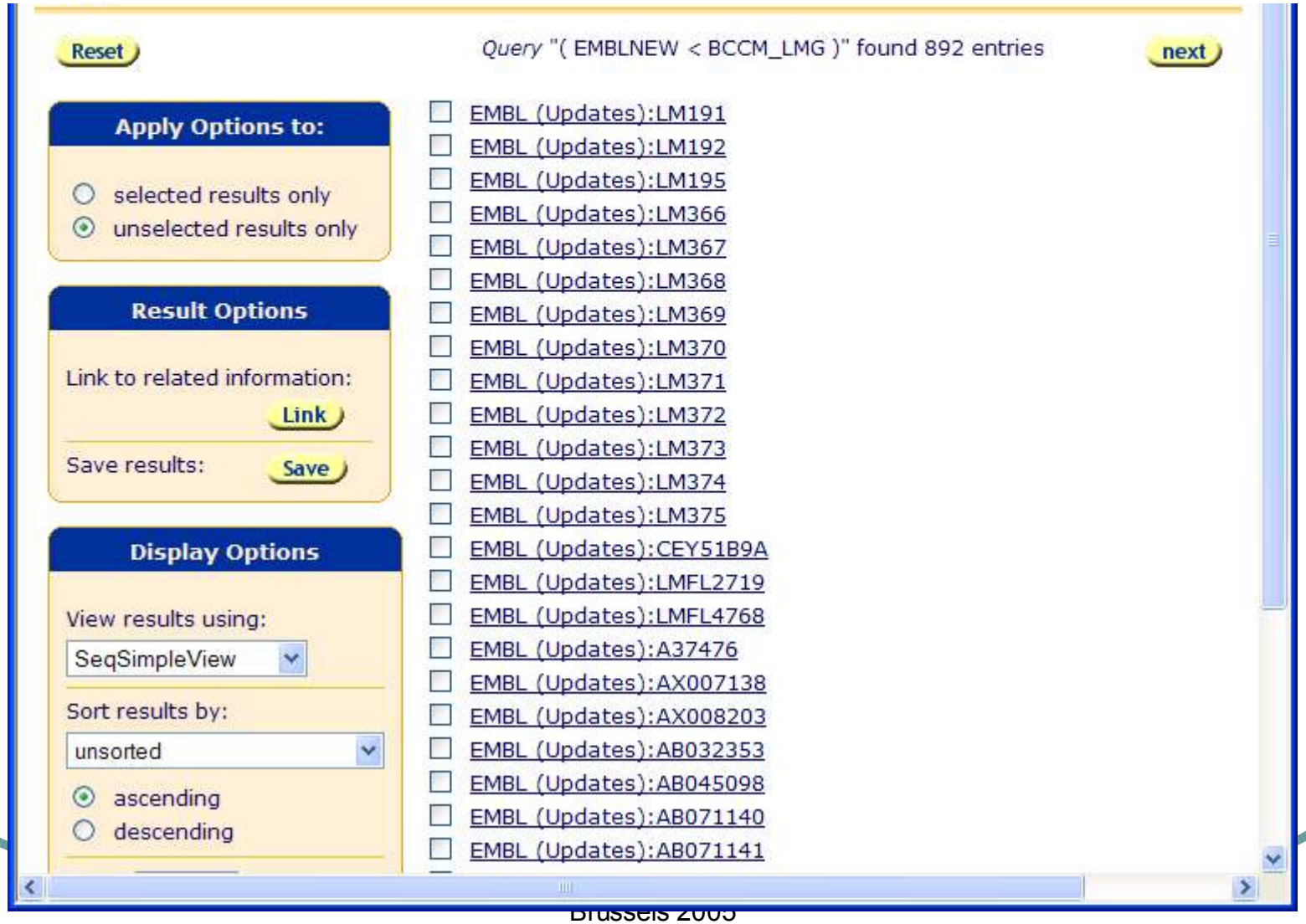
Links

From Databank	Entries Linked	To Databank	Entries Linked	Links Total	Indexing Date
EMBLRELEASE	0	EMBLRELEASE_features	0	0	
EMBLRELEASE	0	EMBLRELEASE_reference	0	0	
EMBLRELEASE	0	EMBLRELEASE_counter	0	0	
EMBLNEW	249034	EMBLRELEASE	249034	249034	06-Sep-2004
EMBLRELEASE	858	BCCM_LMG	659	864	06-Sep-2004
EMBLRELEASE	0	CABI_BACT	0	0	06-Sep-2004
EMBLRELEASE	0	CBS_BACT	0	0	
EMBLRELEASE	3910	DSMZ_BACT	1713	3910	06-Sep-2004
EMBLRELEASE	335	NCIMB_BACT	204	335	06-Sep-2004
EMBLRELEASE	1	NCCB_BACT	1	1	06-Sep-2004
EMBLRELEASE	717	CIP_BACT	388	717	06-Sep-2004
EMBLRELEASE	26	BCCM_IHEM	22	26	06-Sep-2004
EMBLRELEASE	162	BCCM_MUCL	133	162	06-Sep-2004
EMBLRELEASE	3148	CBS_FIL	2043	3155	06-Sep-2004
EMBLRELEASE	55781	CBS_YEAST	953	55789	06-Sep-2004
EMBLRELEASE	233	CABI_FIL	187	233	06-Sep-2004
EMBLRELEASE	0	CABI_YEAST	0	0	06-Sep-2004
EMBLRELEASE	47	DSMZ_FUNGI	34	47	06-Sep-2004

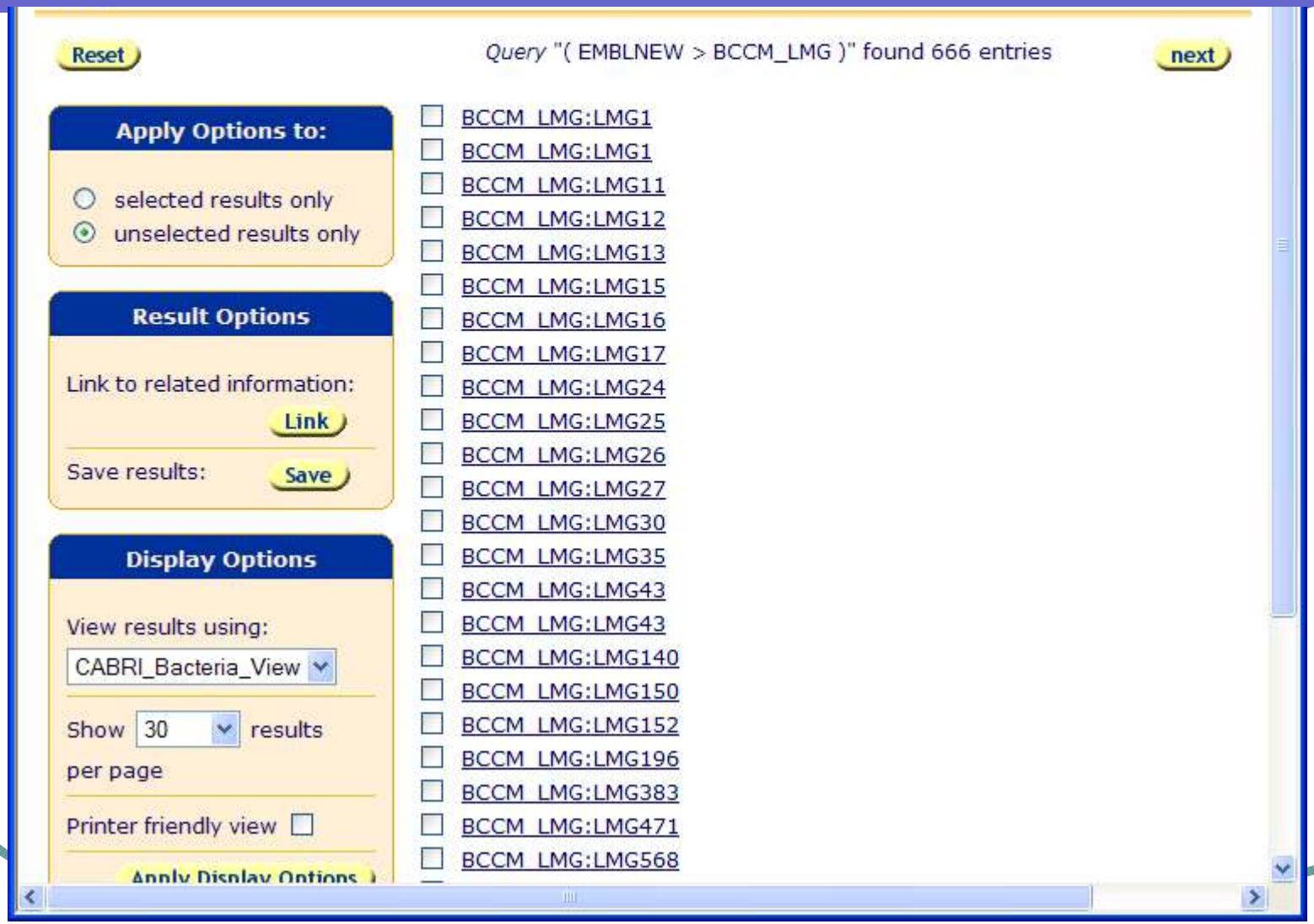
SRS Description

Structure (srsgen.i , srsdb.i , href.i)
Syntax
Information

EBRCN: EMBL records citing CABRI strains



EBRCN: CABRI records cited in EMBL



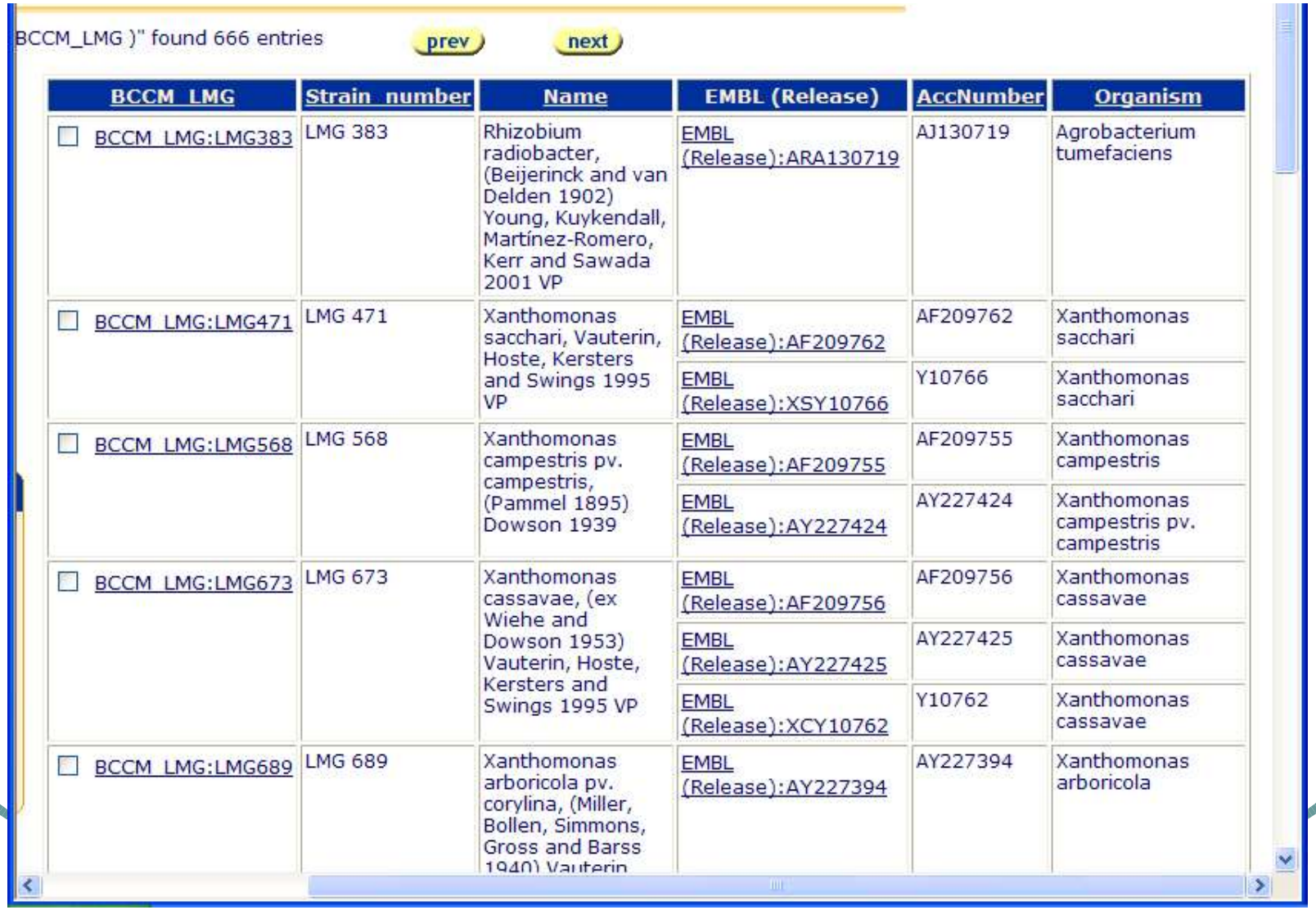
EBRCN: EMBL custom views

EMBL (Release)	Organism	CIP_BACT	Other_collection_numbers	Name
EMBL (Release):CMDNAADE1	Candida maltosa	CIP_BACT:CIPA10	-	Pseudomonas aeruginosa
EMBL (Release):CMADNAPSS	Candida maltosa	CIP_BACT:CIPA101	-	Corynebacterium diphtheriae
EMBL (Release):AB008684	Acinetobacter baumannii	CIP_BACT:CIP7034	ATCC 19606; CCUG 19096; DSM 30007; IAM 12088; LMG 1041; NCTC 12156; NCIMB 12457	Acinetobacter baumannii
EMBL (Release):AB008686	Acinetobacter sp.	CIP_BACT:CIP7029	ATCC 19004	Acinetobacter genomospecies 3
EMBL (Release):AB008687	Acinetobacter sp.	CIP_BACT:CIPA165	ATCC 17979	Acinetobacter genomospecies 6
EMBL (Release):AB008688	Acinetobacter sp.	CIP_BACT:CIP7031	ATCC 9957	Acinetobacter genomospecies 9
EMBL (Release):AB008689	Acinetobacter sp.	CIP_BACT:CIP7012	ATCC 17924; LMG 1003; NCIMB 9019	Acinetobacter genomospecies 10
EMBL (Release):AB008690	Acinetobacter sp.	CIP_BACT:CIP6346	ATCC 11171; DSM 590; NCIMB 8250	Acinetobacter genomospecies 11
EMBL (Release):AB008691	Acinetobacter haemolyticus	CIP_BACT:CIP643	ATCC 17906; CCM 2358; CCUG 888; LMG 996; NCTC 10305; NCIMB 12458	Acinetobacter haemolyticus
EMBL (Release):AB008692	Acinetobacter	CIP_BACT:CIP646	ATCC 17909; CCUG 19095;	Acinetobacter

EBRCN: EMBL custom views (ii)

	/pseudo	
EMBL (Release):AB008684	/db_xref="taxon:470" /mol_type="genomic DNA" /organism="Acinetobacter baumannii" /strain="CIP 70.34; ATCC 19606"	CIP_BACT:CIP7034
	/db_xref="GOA:Q43907" /db_xref="UniProt/Swiss-Prot:Q43907" /transl_table=11 /EC_number="5.99.1.3" /gene="gyrB" /product="DNA gyrase subunit B" /protein_id="BAA75401.1" /translation="DNSYKVSGGLHGVGVSVVNALSSKLHLTIYRAGQIHEQEYHHGDP QYPLRVIGETDNTGTTVRFWPSAETFSQTIFNVEILARRLRELSFLNAGVRIVLRDERI NLEHVYDYEGGLSEFVKYINEGKNHLNEIFHFTADADNGIAVEVALQWNDSYQENVRCF TNNIPQKDGGTHLAGFRAALTRGLNQYLENENILKKEKVNVTGDDAREGLTAIISVKVP DPKFSSQTKEKLVSSEVKPAVEQAMNKEFSAYLLENPQAAKSIAGKIIDAARARDAARK AREMTRRKSALDIAGLPGKLADCQEKDPALSELYLVEGDSAGGSAKQGRNRKMQAILPL KGKILNVERARFDKMISSQEVGTLITALGCGIGREEYNPDKLRYHKII"	
EMBL (Release):AB008686	/db_xref="taxon:472" /mol_type="genomic DNA" /organism="Acinetobacter sp." /strain="CIP 70.29; ATCC 19004"	CIP_BACT:CIP7029
	/db_xref="GOA:Q44270" /db_xref="UniProt/Swiss-Prot:Q44270" /transl_table=11 /EC_number="5.99.1.3" /gene="gyrB" /product="DNA gyrase subunit B" /protein_id="BAA75403.1" /translation="DNSYKVSGGLHGVGVSVVNALSSKLHLIINRAGQVHEQEYHHGDP QYPLRVIGETDSSGTTVRFWPSELTFSQTIFNVEILARRLRELSFLNAGVRIVLRDERI NLEHVYDYEGGLSEFVKYINEGKTHLNEIFHFTADTENGIGVEVALQWNESYQENVRCF TNNIPQKDGGTHLAGFRAALTRGLNQYLENENILKKEKVNVTGDDAREGLTAIISVKVP DPKFSSQTKEKLVSSEVKPAVEQAMNKEFSAYLLENPQAAKSIAGKIIDAARARDAARK AREMTRRKSALDIAGLPGKLADCQEKDPALSELYLVEGDSAGGSAKQGRNRKMQAILPL KGKILNVERARFDKMISSQEVGTLITALGCGIGREEYNPDKLRYHKII"	
EMBL (Release):AB008687	/db_xref="taxon:472" /mol_type="genomic DNA" /organism="Acinetobacter sp." /strain="CIP A165; ATCC 17979"	CIP_BACT:CIPA165
	/db_xref="GOA:Q44271" /db_xref="UniProt/Swiss-Prot:Q44271"	

EBRCN: CABRI custom views



BCCM_LMG)" found 666 entries prev next

BCCM_LMG	Strain_number	Name	EMBL (Release)	AccNumber	Organism
BCCM_LMG:LMG383	LMG 383	Rhizobium radiobacter, (Beijerinck and van Delden 1902) Young, Kuykendall, Martínez-Romero, Kerr and Sawada 2001 VP	EMBL (Release):ARA130719	AJ130719	Agrobacterium tumefaciens
BCCM_LMG:LMG471	LMG 471	Xanthomonas sacchari, Vauterin, Hoste, Kersters and Swings 1995 VP	EMBL (Release):AF209762	AF209762	Xanthomonas sacchari
			EMBL (Release):XSY10766	Y10766	Xanthomonas sacchari
BCCM_LMG:LMG568	LMG 568	Xanthomonas campestris pv. campestris, (Pammel 1895) Dowson 1939	EMBL (Release):AF209755	AF209755	Xanthomonas campestris
			EMBL (Release):AY227424	AY227424	Xanthomonas campestris pv. campestris
BCCM_LMG:LMG673	LMG 673	Xanthomonas cassavae, (ex Wiehe and Dowson 1953) Vauterin, Hoste, Kersters and Swings 1995 VP	EMBL (Release):AF209756	AF209756	Xanthomonas cassavae
			EMBL (Release):AY227425	AY227425	Xanthomonas cassavae
			EMBL (Release):XCY10762	Y10762	Xanthomonas cassavae
BCCM_LMG:LMG689	LMG 689	Xanthomonas arboricola pv. corylina, (Miller, Bollen, Simmons, Gross and Barss 1940) Vauterin	EMBL (Release):AY227394	AY227394	Xanthomonas arboricola

EBRCN: Download of data

	delicatum	/mol_type="genomic DNA" /organism="Aquaspirillum delicatum" /strain="LMG 4328" /product="16S ribosomal RNA"		delicatum, (Leifson 1962) Hylemon, Wells, Krieg and Jannasch 1973 AL
ates):AF085226	Wautersia paucula	/db_xref="taxon:82633" /mol_type="genomic DNA" /organism="Wautersia paucula" /strain="LMG 3413" /product="16S ribosomal RNA"	BCCM_LMG:LMG3413	Ralstonia paucula, Vandamme, Goris, Coenye, Hoste, Janssens, Kersters, De Vos and Falsen 1999 VP
ates):AF090993	Lactobacillus fermentum	/db_xref="taxon:1613" /mol_type="genomic DNA" /note="conserved RAPD fragment UF" /organism="Lactobacillus fermentum" /strain="LMG 6902" /codon_start=2 /db_xref="GOA:Q9XDB9" /db_xref="UniProt/TrEMBL:Q9XDB9" /transl_table=11 /gene="hda" /product="D-2-hydroxyacid dehydrogenase" /protein_id="AAD41656.1" /translation="RALNARVLGYDLKPREEMEGIVEYVSKEELLRQSDVVSLHVDLNP TSEGLLTKAEFDQMKPGASLVNASRGPVVNTADLINALKDGPLAFAALDTVEGEGPIFN ADHRQDGLKGEPLVEELHAMDNVILTPHIAFFTNIAVQNMVDISLDDVITILNGQPSEH EVAN"	BCCM_LMG:LMG6902	Lactobacillus fermentum, Beijerinck 1901 AL
ates):AF090994	Lactobacillus fermentum	/db_xref="taxon:1613" /mol_type="genomic DNA" /note="conserved RAPD fragment LF" /organism="Lactobacillus fermentum" /strain="LMG 6902" /codon_start=1 /db_xref="UniProt/TrEMBL:Q9XDB8" /note="putative" /transl_table=11 /gene="ap1" /product="unknown" /protein_id="AAD41657.1" /translation="TRPYIGSSSTNSVLNLAFGYNGTQRLFGQTTGTGGTFSGMGTKAS GKTKTGAPTGGTKQGGMKKPAGGAPTGQAKPTSGMKKGTTNGGKPSGKSLSSMPKTTKG RPTGMKGGAGQGGNGIFNVGTAGVTRLFQTALRRQISWFLPLALFGMVAAYLAAYDRKK KWWQTNRKQQDVLYWAGWLIPVGAFFSVANFFHPYYTIMLAPPLAVLAALAITSGLKRP	BCCM_LMG:LMG6902	Lactobacillus fermentum, Beijerinck 1901 AL

Validation of crossreferences

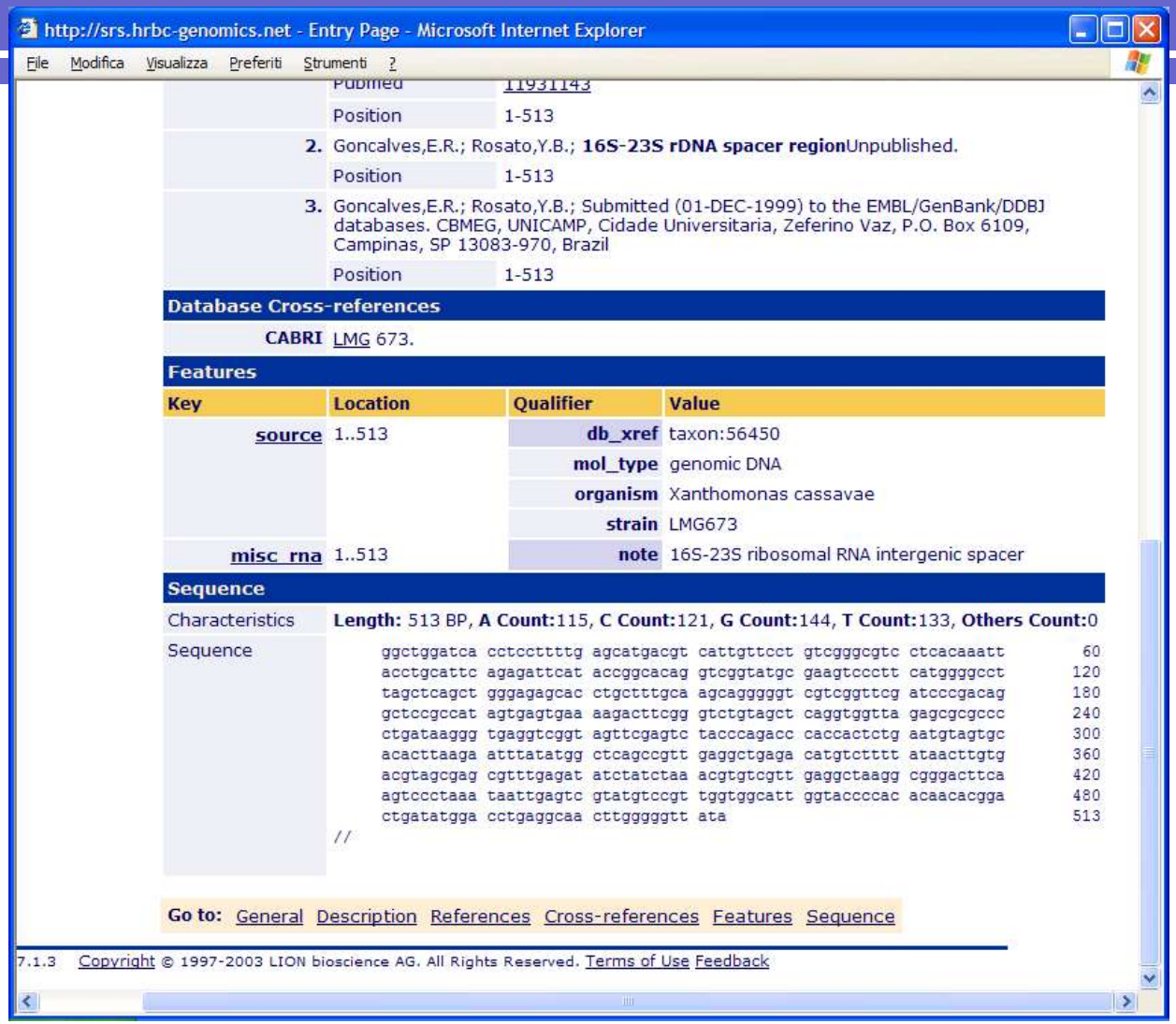
- A custom view including organism name, as specified by EMBL, and scientific name, as specified by CABRI, was created and data downloaded
- For each catalogue, three lists were created including:
  - **Valid links**: records having identical strain numbers and the same organisms/scientific name
  - **Dubious links**: records where organisms and name were different, but could be checked for synonyms of names, alternate names and previous names
  - **Invalid links**: records where names were different and clearly not coherent (e.g., bacteria versus human sequences)
- For bacteria strains validity was further validated with data provided by P. Dawyndt (Univ. Ghent)

Submission of crossreferences

- Validated links have been submitted to EBI and included in EMBL version 81
- Collections are invited to return validated data
- Analysis carried out every two EMBL versions

CBS Yeasts catalogue (huge records)	51103
CBS Filamentous fungi catalogue (huge records)	7116
DSMZ Bacteria catalogue	2618
CBS Filamentous fungi catalogue	2551
CBS Yeasts catalogue	1880
BCCM/LMG Bacteria catalogne	799
CIP Bacteria catalogne	546
NCIMB Bacteria catalogne	336
BCCM/MUCL Fungi and yeasts catalogue	119
DSMZ Fungi catalogne	31
BCCM/IHEM Fungi and yeasts catalogue	18
Total	67528

Links added to EMBL and set up at EBI



Some lessons learned

- Although SRS is a powerfull indexing and search engine, the EMBL size and complexity actually limits the possibility of performing complex searches on it
- Linking to EMBL is best done by using direct explicit links
- Since annotation of strain numbers is not coherent, an a priori identification of crossreferences is needed and it can be effectively carried out by using SRS
- A complete validation can only be achieved by manually analysing dubious data
- Comparison with crossreferences identified by using different methods is helpful for reducing manual intervention

Future perspectives

- The procedure should be further refined and made as automatic as possible
- Any effort should be done for requesting the insertion of strain numbers in a proper format and in a separate field
- The implementation of links from CABRI to EMBL should be done by adding a purpose information in those CABRI records that are actually cited in the EMBL Data Library
- This link can simply be implemented by searching the external references field of EMBL with the CABRI strain number through SRS

I wish to thank for their contributions…

National Cancer Research Institute (Genoa, IT)
Federico Malusa, Domenico Marra, Francesca Piersigilli

LMG, Gent, BE
Peter Dawyndt

BCCM, Brussels, BE
François Guissart

DSMZ, Braunschweig, DE
Manfred Kracht

CBS, Utrecht, NL
Gerrit Stegehuis

This work was partially supported by the EU within the projects Common Access to Biological Resources and Information (CABRI) and European Biological Resource Centers Network (EBCRN)