

# On the Integration of Microbial Information

Peter Dawyndt<sup>1</sup>, Marc Vancanneyt<sup>2</sup> and Jean Swings<sup>1,2</sup>

<sup>1</sup>*Laboratory of Microbiology, Faculty of Sciences, Ghent University,  
K.L. Ledeganckstraat 35, B-9000 Ghent, Belgium.*

<sup>2</sup>*BCCM<sup>TM</sup>/LMG Bacteria Collection, Ghent University,  
K.L. Ledeganckstraat 35, B-9000 Ghent, Belgium.*

---

## 1 Introduction

Microbiologists routinely need to find their way through massive amounts of data, which excell in their variety and are physically distributed over a complex network of information sources. Replication of relevant information is the major mechanism that binds these data sources together, rather than the availability of solid and uniform cross-reference links. As a consequence, retrieval and interpretation of this information can become highly compromised by all kinds of inconsistencies within the data gathered from different information sources.

In order to cope with the fragmentary nature of microbial information sources, we are working on the development of an integrated microbial information gateway, which includes the complete integration of both data generated in-house and information retrieved from external data sources. At present, the information covered within this gateway is mainly biased towards bacteria, as this is the group of organisms worked with by the staff of the BCCM<sup>TM</sup>/LMG culture collection and its associated research laboratory.

The implementation of this microbial information gateway is built on top of a data warehouse [7] that seamlessly integrates the data gathered from different data sources, according to the following integration steps

- **data collection:** to get a *complete* picture of the available data, it is evident that we need to enclose all known information sources. The wide range of exchange formats used by the different data sources forms a major obstacle for the computerization of the data collection process. Consequently, separate parsers need to be developed for almost every source of information. Adherence to worldwide accepted and implemented data exchange standards would dramatically simplify this processing step.
- **standardization and normalization:** the collected data is transformed into a sound data model (standardization), where duplication of the information is replaced by cross-reference links (normalization) between the different objects in the data model. This keeps the data warehouse *manageable*.
- **integrity check:** to assure high quality of the information provided by the data warehouse, it is required to check the *correctness* of the collected data. After all, the reliability of the primary data sources might be compromised in that the data they provide might be both incorrect and obsolete.

---

*Email addresses:*

Peter.Dawyndt@UGent.be (Peter Dawyndt),

Marc.Vancanneyt@UGent.be (Marc Vancanneyt),

Jean.Swings@UGent.be (Jean Swings).

In the following sections of this paper we illustrate the application of these integration principles onto a number of real-world examples, that gradually build on top of each other. These examples all originate from the domain of bacterial taxonomy, but the results are as well applicable for other kinds of microorganisms. In section 2 we discuss the need to set up a complete and correct strain number equivalence relation and demonstrate how this was implemented in our integrated microbial information gateway. Section 3 demonstrates how the linear strain history described in different culture collection catalogs can be integrated into a complete strain history tree. Finally, section 4 tackles some of the problems encountered while linking experimental data onto the integrated strain database. As a result, it becomes much easier and more efficient to run advanced queries over data that originally came from different sources. This is illustrated in section 5.

## 2 Strain number equivalence

When a sample of biological material is deposited into a culture collection (either from another collection, a biological laboratory or a private person), a unique identifier (strain number) is assigned by that culture collection to label their proper culture of the biological material. In most cases, this identifier is constructed by appending the culture collection acronym with a unique serial number, internally managed by the collection. Other researchers may attach a personal reference label to the biological material they work with.

As the biological material gets distributed over multiple culture collections and other parties, the number of different labels referring to the same biological material grows accordingly. Moreover, data generated and information derived from

that biological material by different people at different locations gets completely defragmented, because in most cases reference is made to the microbial strain using only one of the possible strain labels. To avoid a *tower of Babel* situation, it is therefore essential to establish the equivalence relation of strain identifiers, which defines the collection of strain numbers that refer to the same biological material.

Currently, this equivalence relation is jointly set up in the catalogs of culture collections that harbour duplicate samples of the biological material. The equivalence information is recorded in a field called **other culture collection numbers** according to the CABRI standard [2]. When a biological sample is deposited from one collection into another, its equivalence information gets manually duplicated and completed based on the information taken from other catalogs. Therefore, if one needs to find all known labels used for a given strain, it is necessary to scan the catalogs of all culture collections that potentially harbour a culture of this strain. As an example, we have shown in the upper part of Table 1 the relevant information retrieved from a strain number equivalence search for the *Bacillus cereus* type strain. This table clearly demonstrates that the equivalence information gathered from different data sources contains a large amount of duplication. After normalization of the search results, all known strain numbers of the *Bacillus cereus* equivalence class are as shown in the last row of Table 1.

The cross-reference links between synonymous *Bacillus cereus* type strain labels, as they are recorded in the culture collection catalogs we have consulted to perform the equivalence search, have been graphically represented in Figure 1. We call this kind of graphical representation the *catalog cross-reference table* (CCRT) of the *Bacillus cereus* type strain. In such a CCRT, each row represents a single cat-

Source	Catalog entry	Species name	Other culture collection numbers
ATCC	ATCC 14579 <sup>T</sup>	<i>Bacillus cereus</i>	971; 13; NCIB 9373; NCTC 2599
CABRI	CIP 66.24 <sup>T</sup>	<i>Bacillus cereus</i>	ATCC 14579; CCM 2010; DSM 31; IAM 12605; JCM 2152; LMG 6923; NCIB 9373; NCTC 2599
CABRI	DSM 31 <sup>T</sup>	<i>Bacillus cereus</i>	ATCC 14579; CCM 2010; LMG 6923; NCIB 9373; NCTC 2599
CABRI	LMD 75.8 <sup>T</sup>	<i>Bacillus cereus</i>	ATCC 14579; NCIB 9373; NCTC 2599; CCM 2010; DSM 31; Gibson
CABRI	LMG 6923 <sup>T</sup>	<i>Bacillus cereus</i>	ATCC 14579; CCEB 625; CCM 2010; CCUG 7414; CECT 148; CIP 66.24; DSM 31; FIRDI 603; Ford 13; Gibson 971; IAM 12605; JCM 2152; Logan B0002; NCFB 1771; NCIB 9373; NCTC 2599; NRRL B-3711; OUT 8406
CABRI	NCIMB 9373 <sup>T</sup>	<i>Bacillus cereus</i>	ATCC14579; CCM2010; CECT148; CIP66.24; DSM31; IAM12605; IFO15305; JCM2152
CCM	CCM 2010 <sup>T</sup>	<i>Bacillus cereus</i>	ATCC 14579
CCUG	CCUG 7414 <sup>T</sup>	<i>Bacillus cereus</i>	CCM 2010; NCIB 9373; ATCC 14579; NCTC 2599; Ford 13; DSM 31
CECT	CECT 148 <sup>T</sup>	<i>Bacillus cereus</i>	ATCC 14579; CCEB 625; CCM 2010; CCRC 10603; CCRC 11026; CCTM La 3674; CCUG 7414; CIP 66.24; DSM 31; FIRDI 603; Ford 13; Gibson 971; IAM 12605; JCM 2152; LMD 75.8; LMG 6923; NCFB 1771; NCIMB 9373; NCTC 2599; OUT 8406; VTT E-93143
CECT	CECT 5050 <sup>T</sup>	<i>Bacillus cereus</i>	ATCC 14579; CCM 2010; CECT 148; DSM 31; Ford 13; Gibson 971; LMG 6923; NCIMB 9373; NCTC 2599
CIP	CIP 66.24 <sup>T</sup>	<i>Bacillus cereus</i>	ATCC 14579; CCM 2010; DSM 31; IAM 12605; JCM 2152; LMG 6923; NCTC 2599; NCIMB 9373
DSM	DSM 31 <sup>T</sup>	<i>Bacillus cereus</i>	ATCC 14579; CCM 2010; LMG 6923; NCIB 9373; NCTC 2599
IFO	IFO 15305 <sup>T</sup>	<i>Bacillus cereus</i>	ATCC 14579; CCM 2010; DSM 31; IAM 12605; JCM 2152; LMG 6923; NCIB 9373; NCIMB 9373; NCTC 2599
JCM	JCM 2152 <sup>T</sup>	<i>Bacillus cereus</i>	ATCC 14579; CCM 2010; CCRC 10603; CCUG 7414; CECT 148; CIP 66.24; DSM 31; IAM 12605; IFO 15305; KCTC 3624; LMG 6923; NBRC 15305; NCFB 1771; NCIMB 9373; NCTC 2599; NRRL B-3711; VKM B-504
KCTC	KCTC 3624 <sup>T</sup>	<i>Bacillus cereus</i>	ATCC 14579; CCM 2010; CCRC 10603; CIP 66.24; DSM 31; IAM 12605; IFO 15305; JCM 2152; LMG 6923; NCFB 1771; NCIMB 9373; NCTC 2599; VKM B-504
LMG	LMG 6923 <sup>T</sup>	<i>Bacillus cereus</i>	ATCC 14579; CCEB 625; CCM 2010; CCUG 7414; CECT 148; CIP 66.24; DSM 31; FIRDI 603; Ford 13; Gibson 971; IAM 12605; JCM 2152; Logan B0002; NCFB 1771; NCIB 9373; NCTC 2599; NRRL B-3711; OUT 8406
UKNCC	NCTC 2599 <sup>T</sup>	<i>Bacillus cereus</i>	Ford 13; ATCC 14579; NCIB 9373; DSM 31
UKNCC	NCIMB 9373 <sup>T</sup>	<i>Bacillus cereus</i>	Gibson971; Ford13; ATCC14579; CCM2010; CECT148; CIP66.24; DSM31; IAM12605; IFO15305; JCM2152; NCTC2599; NCDO1771
Integrated strain database	type strain	<i>Bacillus cereus</i>	ATCC 14579; CCEB 625; CCM 2010; CCRC 10603; CCRC 11026; CCTM La 3674; CCUG 7414; CECT 148; CECT 5050; CIP 66.24; DSM 31; FIRDI 603; Ford 13; Gibson 971; Gibson; IAM 12605; IFO 15305; JCM 2152; KCTC 3624; LMD 75.8; LMG 6923; Logan B0002; NBRC 15305; NCDO 1771; NCFB 1771; NCIB 9373; NCIMB 9373; NCTC 2599; NRRL B-3711; OUT 8406; VKM B-504; VTT E-93143; 13; 971

Table 1

Strain number equivalence information for the *Bacillus cereus* type strain, collected from different culture collection catalogs. The last row of the table shows a normalized representation of the search results, which shows no duplication of the equivalent strain labels assigned to the *Bacillus cereus* type strain.

alog entry found in a culture collection catalog. Rows are annotated by the strain number and species name as assigned in their originating culture collection catalog. Each column of the catalog cross-reference table represents a strain identifier and is annotated by that identifier. If a given catalog entry references a given strain identifier as a synonymous strain number, the corresponding row-column intersection is marked by a colored box. Self-references are represented by means of a green box and cross-references by means of a black box. As such, CCRTs describe the presence/absence of the cross-references that constitute the strain number equivalence relation in an easy-to-interpret manner. By inspection of the catalog cross-reference tables for a number of strains, it can be easily derived that the list of equivalent strain numbers is far from complete in most collection catalogs.

In order to establish an up-to-date normalized view of the strain number equivalence relation, containing information on all known bacterial strains, we have developed a database and accompanying software tools, which automatically retrieve and process the equivalence information from the online data sources shown in Table 2. The type strain information extracted from the Bacterial Nomenclature Up-to-Date (DSMZ) list is included in the catalog cross-reference tables as rows marked by the term **type strain (DSM)**, followed by the species name of the type strain. Table 2 also shows that multiple instances of the same catalog may be available online, such as the NCIMB catalog for example, that is incorporated into both the CABRI and UKNCC websites. As nothing prevents that the equivalence information in different copies of the same catalog is out of synchronization, each instance is represented in the catalog cross-reference tables as a separate row. This is clearly illustrated in Figure 1, where the two rows that represent different copies of

the NCIMB 9373<sup>T</sup> catalog entry do not contain the same information.

From the first versions of this integrated strain database, it was immediately clear that quite a large number of incorrect cross-references were present in the information retrieved from the originating data sources. Some of these errors forced the integration software to illegitimately merge strain numbers of different strains into the same equivalence class. This highly compromised the applicability of the strain number equivalence database as a building block for an integrated microbial information gateway. To resolve this problem, an error detection/correction algorithm was developed. This software package represents each equivalence class as the catalog cross-reference table having rows constructed from all catalog entries that contributed information to the construction of the equivalence class. From a transversal grouping [1] of both the rows and columns of the catalog cross-reference table, the algorithm can then easily spot possible incorrect cross-references present in a catalog entry, as these incorrect references occur as outlying boxes in the CCRT representation. Figure 2 shows the catalog cross-reference representation of a case where three different type strains have been illegitimately merged into a single equivalence class, due to errors in two catalog entries. Horizontal and vertical classification of the CCRT was calculated using the UPGMA (unweighted pair-group method using arithmetic averages) hierarchical clustering method [8] working upon intermediate pairwise Dice similarity matrices [4] for both row and column vectors. From Figure 2 one can easily derive that the catalog entry of culture DSM 40066<sup>T</sup> makes reference to label CBS 498.68 as an equivalent strain number, while the correctly referenced label should be CBS 500.68. Similarly, the catalog entry NCIMB 8233<sup>T</sup> should refer to RIA 1040 as a synonymous strain number, instead of making refer-

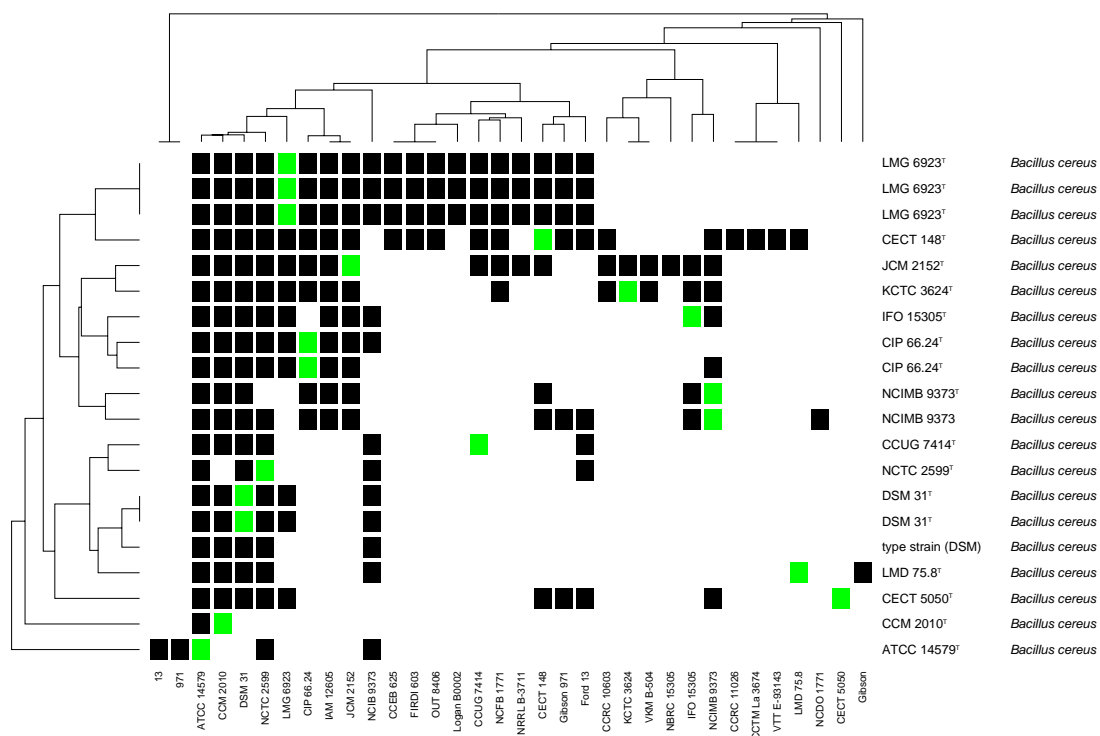


Fig. 1. Catalog cross-reference table for the *Bacillus cereus* type strain.

Collection or catalog name	Acronym	URL
American Type Culture Collection	ATCC	<a href="http://www.atcc.org">www.atcc.org</a>
Czech Collection of Microorganisms	CCM	<a href="http://www.sci.muni.cz/ccm">www.sci.muni.cz/ccm</a>
Culture Collection, University of Göteborg, Sweden	CCUG	<a href="http://www.ccug.gu.se">www.ccug.gu.se</a>
Colección Española de Cultivos Tipo	CECT	<a href="http://www.cept.org/english/index.htm">www.cept.org/english/index.htm</a>
Collection de l'Institut Pasteur	CIP	<a href="http://www.pasteur.fr/externe">www.pasteur.fr/externe</a>
Deutsche Sammlung von Mikroorganismen und Zellkulturen	DSMZ	<a href="http://www.dsmz.de">www.dsmz.de</a>
Institute for Fermentation, Osaka	IFO	<a href="http://www.ifo.or.jp/index_e.html">www.ifo.or.jp/index_e.html</a>
Japan Collection of Microorganisms	JCM	<a href="http://www.jcm.riken.go.jp">www.jcm.riken.go.jp</a>
Korean Collection for Type Cultures	KCTC	<a href="http://kctc.kribb.re.kr/english">kctc.kribb.re.kr/english</a>
Laboratory of Microbiology, Ghent	LMG	<a href="http://www.belspo.be/bccm/lmg.htm">www.belspo.be/bccm/lmg.htm</a>
Pasteur Culture Collection of Cyanobacteria	PCC	<a href="http://www.pasteur.fr/recherche/banques/PCC">www.pasteur.fr/recherche/banques/PCC</a>
All-Russian Collection of Microorganisms	VKM	<a href="http://www.vkm.ru">www.vkm.ru</a>
Common Access to Biological Resources and Information	CABRI	<a href="http://www.cabri.org">www.cabri.org</a>
CABRI collects several culture collection catalogs into a uniform online catalog. Subcatalogs currently covered within integrated strain database: CBS, CIP, DSMZ, IMI, LMD, LMG, MUCL, NCIMB		
United Kingdom National Culture Collection	UKNCC	<a href="http://www.ukncc.co.uk/index.htm">www.ukncc.co.uk/index.htm</a>
UKNCC collects several culture collection catalogs into a uniform online catalog. Subcatalogs currently covered within the integrated strain database: NCIMB, NCPPB, NCTC		
Bacterial Nomenclature Up-to-Date (DSMZ)		<a href="http://www.dsmz.de/bactnom/bactname.htm">www.dsmz.de/bactnom/bactname.htm</a>

Table 2

Online culture collection catalogs currently covered within the integrated strain database.

ence to the strain number RIA 1038. The putative incorrect cross-reference links, as they are determined by our error detection/correction algorithm, are represented in the catalog cross-reference tables as red boxes.

The current version of the strain number equivalence database is heavily biased towards bacterial strain information, due to our initial selection of data sources. However, our experience has learned that the integration software can equally help unveil incorrect cross-references for other types of microorganisms. Even examples of clearly incorrect cross-references between bacterial and fungal strains have been found within the current version of the database. At the moment, the database contains information on 117227 microbial strains (equivalence classes), discriminating between a total of 303294 strain numbers. Based on our error detection/correction software package, a list of more than one thousand potential errors have been detected and/or corrected within the originating data sources. During the automated update process of the equivalence database, the integration software avoids illegitimate strain merges that were previously resolved, by consulting this list of errors. It is nevertheless highly recommended that these errors are corrected in the originating collection catalogs, as these are the primary sources of information, so that errors made during the manual collection of cross-referential information have a tendency to be unwillingly copied from one catalog into another. Therefore, the list of errors can be retrieved from the authors on request.

### 3 Strain history

Catalog cross-reference tables alone may not be sufficient to completely unravel the ins and outs of the more complex incon-

sistencies within the strain number equivalence database. However, integration of the fragmented strain history information into a complete history tree has proven to be a very helpful tool to help resolve some of these difficult cases. In this section we will explain the construction of the complete strain history trees and illustrate their error detection capabilities, by means of some real-world examples.

Culture collection catalogs record the strain history information within a field called **History of Deposit** according to the CABRI standard [2], which should describe the history of the cultured sample from its deposit into the collection up to its isolation. As an example, Table 3 shows the history information for the *Bacillus cereus* type strain as it can be found in the catalogs of a number of culture collections. This table clearly indicates that the different data sources have adopted different formats of coding the strain history information, due to the fact that the CABRI standard does not give a formal prescription of formatting the content of the **History of Deposit** field. As a consequence, it is a daunting task for software agents to process the history information in a fully automated way. Nevertheless, the history information of the *Bacillus cereus* type strain in Table 3 contains quite some duplication, which makes that after standardization and normalization of the data, the complete strain history tree of the *Bacillus cereus* type strain can be represented in a more informative way, as is shown in Figure 3. In this graphical representation of the complete history tree, the orange boxes represent cultures of the *Bacillus cereus* type strain as they are stored in different culture collections or private research collections. These boxes are labeled with the strain number assigned by the collection holder. If the collection holder additionally provides some identification of the cultured sample, the identification is shown in the bottom half of the orange box. De-



posit of a culture from one collection into another is represented by an arrow linking the corresponding boxes, and annotated with the information known about this deposition (date of deposit, depositor name, depositor institute, . . .). Strain labels from the *Bacillus cereus* type strain equivalence class that have an unknown position within the complete history tree, are lumped into a single blue box in the upper left corner of Figure 3.

According to the strain number equivalences found for the *Enterococcus gallinarum* type strain in different online culture collection catalogs, all strain numbers mentioned in the graphical representation of the complete strain history in Figure 4 constitute a single equivalence class. However, several catalog entries (NCTC 11428, ATCC 35038, LMG 11207) contain some direct or indirect evidence that the corresponding cultures should possibly be identified as *Enterococcus faecalis*. Within the history tree, one could interpret this as a putative contamination of the complete branch rooted at the node labeled NCTC 11428 (or any higher node). The nodes of this affected branch are coloured light gray in the complete tree representation of Figure 4.

#### 4 Link sequence and strain information

A major effort in the construction of our integrated microbial information gateway has been put into cross-referencing the integrated strain database we have discussed in section 2 with periferal data such as literature information and experimental information. This periferal data is typically generated by different instances on different locations, is stored over several distributed databases using a whole bunch of alternative storage formats, and refers to the biological material using only

a limited number of the known equivalent strain labels from the corresponding equivalence class. As such, the information gateway establishes a logical defragmentation of the information that is physically fragmented over different data sources. In this section we illustrate some of the problems encountered during linking the bacterial sequences included in the International Nucleotide Sequence Database (DDBJ/EMBL/GenBank) [5] with our integrated strain database.

A first deficiency of the public sequence database is that there is no consistent recording of the strain number of the individual culture from which the sequence was obtained. According to the specifications of the public sequence database, this information should be stored in the qualifiers `isolate` or `strain` of the `source` feature, but the sequence deposit procedures do not prevent that depositors provide the strain information within another field or – more critical – do not provide this information at all. Therefore, we have developed a software tool that automatically parses complete EMBL formatted sequence entries for extraction of the associated strain label information.

At first sight, these strain labels may seem good candidates for linking the sequence data with the integrated strain database, but unfortunately the labels associated with biological samples show some form of ambiguity. This is illustrated in Table 4, which shows all equivalence classes from the integrated strain database that harbour the strain number B2 or some alternative spelling. To resolve this problem, strain labels referring to samples of different microbial strains receive a different numerical identifier within the integrated strain database, which is shown in the first column of the example in Table 4. Fixed links between the sequence and strain database can be established using these unique identifiers. The integrated strain

Bacillus cereus  
type strain

CCEB 625  
CCTM La 3674  
FIRDI 603  
Logan B0002  
NBRC 15305  
NCDO 1771  
NCFB 1771  
OUT 8406  
VKM B-504  
VTT E-93143

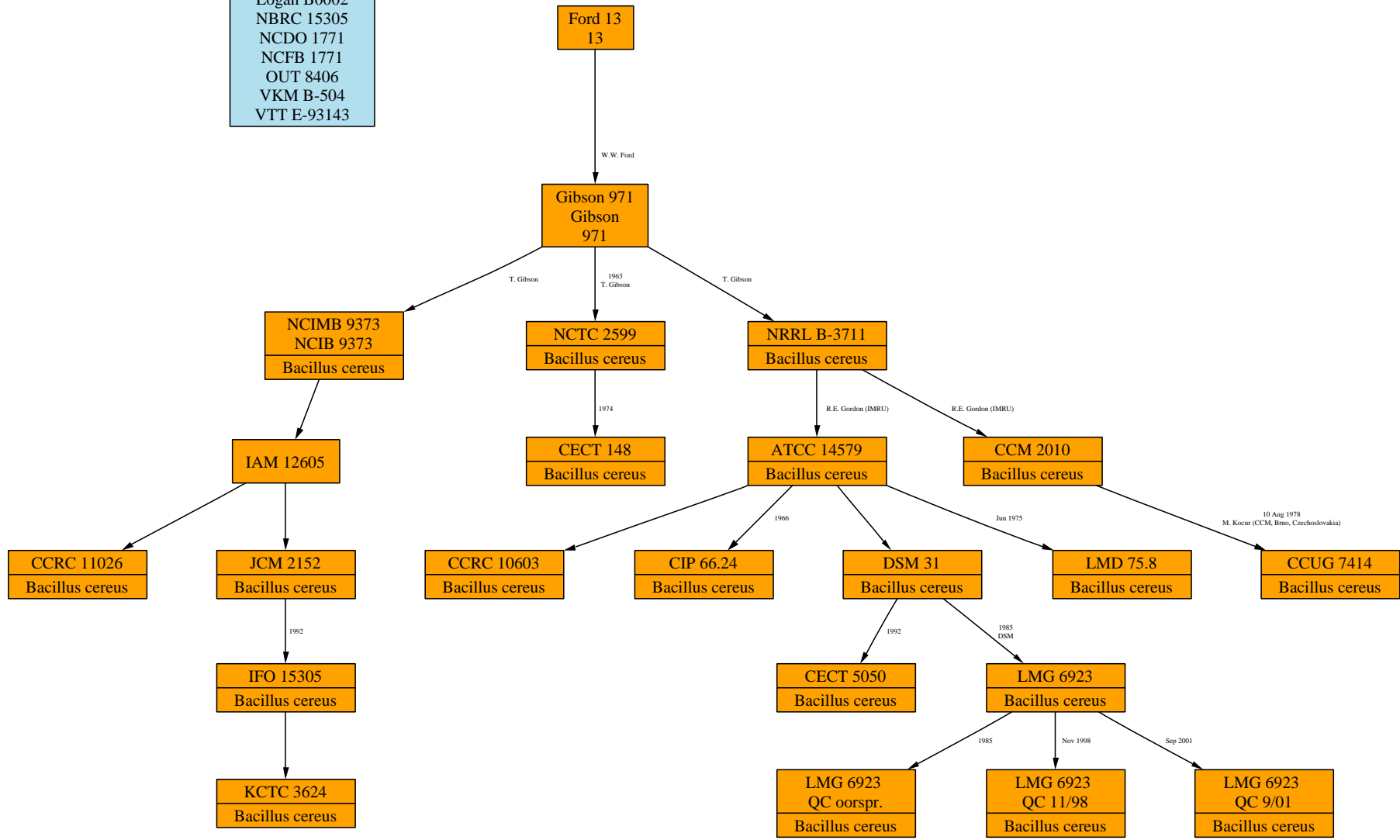


Fig. 3. Complete strain history tree of the *Bacillus cereus* type strain.

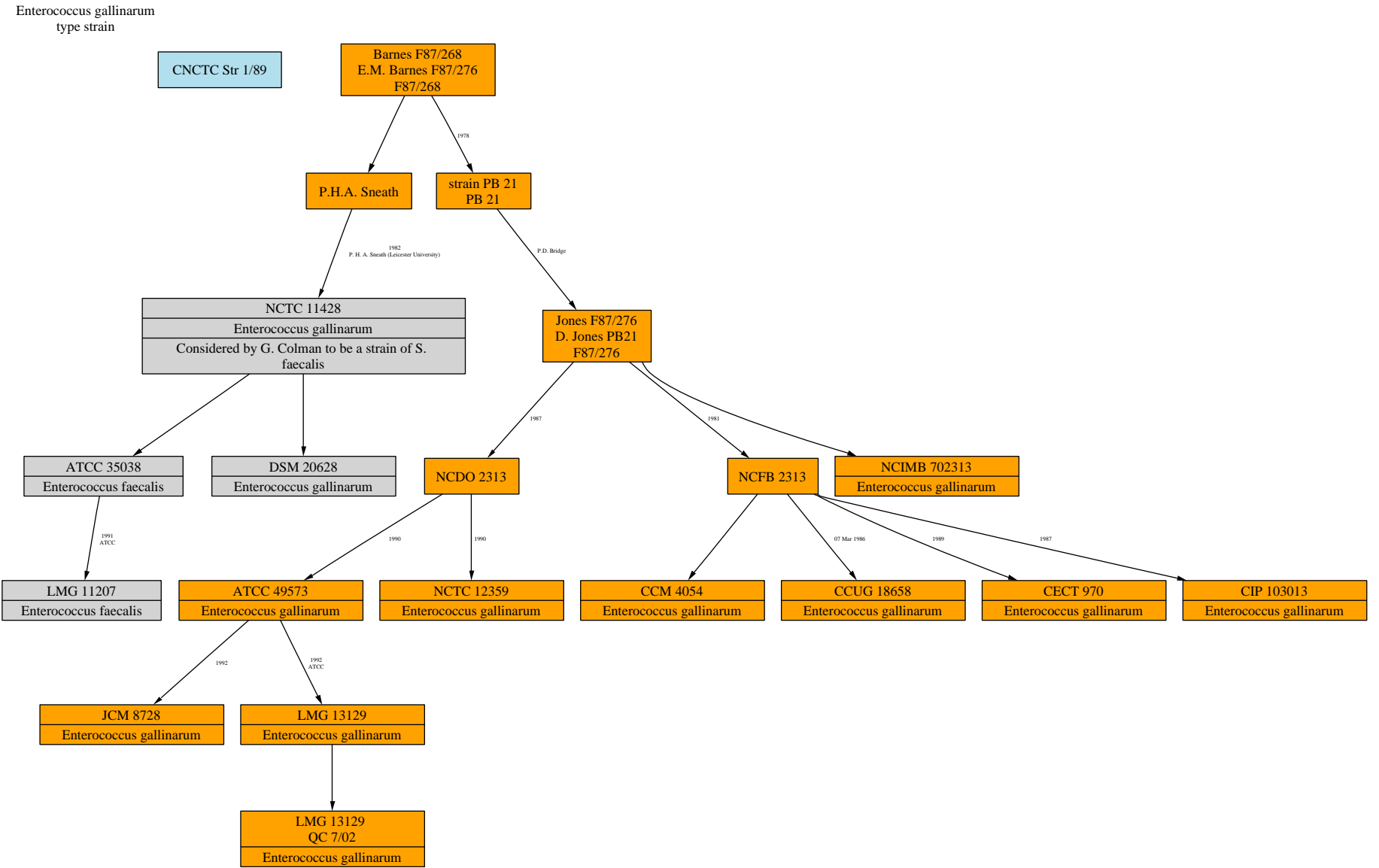


Fig. 4. Complete strain history tree of the *Enterococcus gallinarum* type strain, showing a putative contamination within the light gray coloured branch. Several sources indicate that the cultures of the affected branch should be identified as *Enterococcus faecalis*.

Source	Catalog entry	Species name	History
ATCC	ATCC 14579 <sup>T</sup>	<i>Bacillus cereus</i>	ATCC<<-RE Gordon <<-T. Gibson 971 <<- W. Ford 13
CABRI	CIP 66.24 <sup>T</sup>	<i>Bacillus cereus</i>	ATCC 1966 < R.E. Gordon: strain NRRL B-3711 < T. Gibson: strain 971 < W.W. Ford: strain 13
CABRI	DSM 31 <sup>T</sup>	<i>Bacillus cereus</i>	<- ATCC <- R.E. Gordon <- T. Gibson, 971 <- W.W. Ford, 13
CABRI	LMD 75.8 <sup>T</sup>	<i>Bacillus cereus</i>	LMD < Jun 1975, ATCC < R.E. Gordon < T. Gibson < W. Ford
CABRI	LMG 6923 <sup>T</sup>	<i>Bacillus cereus</i>	<- 1985, DSM <- ATCC <- R.Gordon <- T.Gibson <- W.Ford
CABRI	NCIMB 9373 <sup>T</sup>	<i>Bacillus cereus</i>	T.Gibson - W.W.Ford
CCM	CCM 2010 <sup>T</sup>	<i>Bacillus cereus</i>	R.E. Gordon <- T. Gibson <- W.W. Ford
CCUG	CCUG 7414 <sup>T</sup>	<i>Bacillus cereus</i>	M.Kocur,CCM,Brno,Czechoslovakia 10 Aug 1978 <R.E.Gordon,IMRU<T.Gibson<W.W.Ford
CECT	CECT 148 <sup>T</sup>	<i>Bacillus cereus</i>	CECT, 1974 < NCTC, 1963 < T. Gibson < W.W. Ford.
CECT	CECT 5050 <sup>T</sup>	<i>Bacillus cereus</i>	CECT, 1992 < DSMZ < ATCC < R.E. Gordon < T. Gibson < W.W. Ford.
CIP	CIP 66.24 <sup>T</sup>	<i>Bacillus cereus</i>	1966, ATCC <- R.E. Gordon: strain NRRL B-3711 <- T. Gibson: strain 971 <- W.W. Ford: strain 13
DSM	DSM 31 <sup>T</sup>	<i>Bacillus cereus</i>	<- ATCC <- R.E. Gordon <- T. Gibson, 971 <- W.W. Ford, 13
IFO	IFO 15305 <sup>T</sup>	<i>Bacillus cereus</i>	1992. JCM 2152 <== IAM 12605 <== NCIB 9373 <== R.E. Gordon
JCM	JCM 2152 <sup>T</sup>	<i>Bacillus cereus</i>	<- IAM 12605 <- NCIB 9373 <- R. E. Gordon
KCTC	KCTC 3624 <sup>T</sup>	<i>Bacillus cereus</i>	<- IFO <- JCM <- IAM <- NCIB <- R.E. Gordon
LMG	LMG 6923 <sup>T</sup>	<i>Bacillus cereus</i>	<- 1985, DSM <- ATCC <- R.Gordon <- T.Gibson <- W.Ford
UKNCC	NCIMB 9373 <sup>T</sup>	<i>Bacillus cereus</i>	T.Gibson - W.W.Ford

Table 3

Strain history information of the *Bacillus cereus* type strain, as it was found in different catalogs of culture collections that are available online.

database is constructed in such a way that strain labels starting with an acronym that appears in the WDCM directory of culture collections [10] occur only once in the database, such that these labels unambiguously refer to a single culture within a single equivalence class. This kind of strain number is called a *unique strain number*, as opposed to *ambiguous strain numbers* that may appear more than once in the strain database. As a consequence, sequence entries that refer to a unique strain number can be automatically linked with the strain database, while linkage of entries with ambiguous strain references cannot be performed without user intervention. This is illustrated by the sequence entries with accession numbers AF509820 and AJ309324 that both refer to their sequenced bacterial strain using the label B2. However, the former sequence is identified as *Acinetobacter baylyi* and should be linked to the integrated database entry with identifier 368362 according to the search results in Table 4. This is confirmed by looking up the equivalent strain numbers in the publication by Carr *et al.* [3] linked to sequence entry AF509820. Similarly, the latter se-

quence is identified as *Chryseobacterium defluvi*, indicating that it should be linked to the integrated database entry with identifier 65830. Again, this is confirmed by the equivalent strain numbers mentioned in the paper by Kämpfer *et al.* [6] that is linked to this sequence entry. From the link with the integrated strain database it can be easily derived that both sequences represent type strain sequences, although this information was not directly provided in the sequence database entry AJ309324.

We have put quite some effort in linking a selection of 93680 bacterial sequence entries that are potentially related to the 16S rRNA gene with the integrated strain database. At present, only 11895 (13%) of these entries have been successfully linked in the way described above. Although a vast number of the currently unlinked entries are sequences related to uncultured or uncultureable bacterial strains, our experience from working with this database is that still a significant number of the unlinked entries can be manually linked, at the cost of a time-consuming lookup process. This may include looking up informa-

ID	Label	Species name	Equivalent strain labels
368362	B2 <sup>T</sup>	<i>Acinetobacter baylyi</i>	CIP 107474, DSM 14961
268815	B-2	<i>Actinoadura madurae</i>	A 124, DSM 43381, IMET 7144, IMRU 1136
347460	B2	<i>Aeromonas hydrophila</i> subsp. <i>hydrophila</i>	NCIMB 640
345118	B2	<i>Bacillus</i> sp.	NCIMB 10936
65830	B2 <sup>T</sup>	<i>Chryseobacterium defluvii</i>	CCUG 47675, CIP 107207, DSM 14219
350023	B2	<i>Clostridium butyricum</i>	KCTC 1902, NCIMB 9575
267132	B-2	<i>Corynebacterium glutamicum</i>	ATCC 21269, KCTC 9853
267459	B2	<i>Curtobacterium flaccumfaciens</i>	ATCC 33802
303025	B2 <sup>T</sup>	<i>Methylocapsa acidiphila</i>	DSM 13967, NCIMB 13765
267727	B2	<i>Morganella morganii</i> subsp. <i>sibonii</i>	ATCC 51596
269342	B-2	<i>Neisseria gonorrhoeae</i>	CCUG 13573
346764	B-2	"Other unnamed bacteria"	NCIMB 17
268485	B2	<i>Pseudomonas putida</i>	DSM 6376
266888	B2	<i>Streptococcus salivarius</i>	ATCC 9759
267562	B2	<i>Tenacibaculum maritimum</i>	ATCC 43397, CCM 3965, CECT 4276, CIP 103529, IAM 14118, IFO 16015, JCM 8137, LMG 10398, LMG 11611, NCIMB 2153, NCMB 2153, strain B2, Wakabayashi B-2

Table 4

Equivalence classes containing strain number B2, found within the current version of the integrated strain database. Variations in spelling were taken into account during the search.

tion within the integrated strain database or within external data sources (mainly publications) linked to the public sequence database. As a result, the sequence database is continuously enriched and advanced searches based on the cross-reference links with the biological material become more reliable.

Based on the cross-reference links between the integrated strain database and the public sequence database, it is again possible to perform some integrity checks on the information provided by the sequence database. Table 5 enumerates a small excerpt of the incorrect strain references that have been discovered within the International Nucleotide Sequence Database. The first part of this table shows some inconsistencies that were detected by comparison of the strain identification information taken from both the sequence database and the integrated strain database. The first column of this table gives the accession number of the EMBL sequence entry together with the associated strain identification stored in the `organism species` field. The second column shows the strain number extracted from the sequence entry with the corresponding strain identification retrieved via the link with

the integrated strain database. Clearly, for these examples there is a discrepancy between this identification information that is taken from the two different data sources. By consulting the literature references linked to the EMBL entries and searching the integrated strain database, the cross-reference links have been corrected as indicated in the last column of the table. The last two rows of Table 5 show examples of entries in the public sequence database that refer to strain labels that do not exist within the catalog of the BCCM<sup>TM</sup>/LMG culture collection. These incorrect references have been resolved in a similar way as described before.

## 5 Advanced dynamic queries

The implementation of a microbial information gateway by means of a data warehouse that contains an integrated strain database (constructed as described in sections 2 and 3) that is cross-referenced with a significant amount of literature and experimental data (as explained in section 4), enables us to perform all sorts of advanced queries on the fly. Such queries

accession number	incorrect strain reference	correct strain reference
M58730 ( <i>Bifidobacterium asteroides</i> )	ATCC 29510 ( <i>Stenotrophomonas maltophilia</i> )	ATCC 25910 <sup>T</sup> ( <i>B. asteroides</i> )
X71855 ( <i>Clostridium xylanolyticum</i> )	ATCC 4963 ( <i>Lactobacillus gasseri</i> )	ATCC 49623 <sup>T</sup> ( <i>C. xylanolyticum</i> )
AB089482 ( <i>Derxia gummosa</i> )	ATCC 15594 ( <i>Arthrobotrys conoides</i> )	ATCC 15994 <sup>T</sup> ( <i>D. gummosa</i> )
Y17361 ( <i>Lactobacillus amylolyticus</i> )	DSM 1664 ( <i>Clostridium sporogenes</i> )	DSM 11664 <sup>T</sup> ( <i>L. amylolyticus</i> )
AJ224308 ( <i>Aeromonas popoffii</i> )	LMG 317541 (does not exist)	LMG 17541 <sup>T</sup> ( <i>A. popoffii</i> )
X81623 ( <i>Shewanella putrefaciens</i> )	LMG 26268 (does not exist)	LMG 2268 <sup>T</sup> ( <i>S. putrefaciens</i> )

Table 5  
Examples of incorrect strain references in the International Nucleotide Sequence Database.

can automatically bridge over multiple data sources that were physically separated before the integration process. In the past, answering these advanced queries could be quite a labour-intensive and time-consuming task. Consequently, the search results did not always give a complete and up-to-date picture of all the information available in-house or within the public domain.

Throughout this paper we already illustrated the applicability of advanced queries for the implementation of some extensive integrity checks on data that is distributed over several microbial information sources. In this section we give some examples of advanced queries that are more directly useful for microbiologists.

Knowledge of all the experimental data for a given microbial strain is primordial for setting up polyphasic taxonomy studies [9]. As an example, Table 6 shows the results of a query that searches within the microbial information gateway for all known experimental data generated for the *Enterococcus faecium* type strain. To answer this question, it was required to initially lookup all strain numbers that were used to label the *E. faecium* type strain and subsequently gather all experimental data linked to each of these strain labels. Detailed information is available for each experiment in the search results, on request of the user of the microbial information gateway, as well as statistical tools for the appropriate numerical analysis of the data.

The International Nucleotide Sequence

Database lacks some information needed to perform a direct search of all 16S rRNA gene sequences of all *Enterococcus* spp. type strains that have been deposited in the public domain. The only appropriate search option within the public sequence database is to retrieve all 16S rRNA sequences associated with *Enterococcus* strains, followed by manually filtering out the entries that are associated with a type strain. This search strategy might be hampered by outdated identification information and missing, ambiguous or incorrect strain information in the public sequence database. However, based on the cross-reference links between the sequence entries and our integrated strain database, this query can be resolved more accurately as is shown in Table 7.

## 6 Conclusion

The importance of data integration for the retrieval and interpretation of massive amounts of microbial data, has stimulated us into the development of an integrated microbial information gateway. In this paper we have highlighted some of the challenges and benefits of data integration that occur behind the scenes of the gateway portal interface, such that actions may be taken to cope with the lessons learned from this project.

The overwhelming variety of data exchange formats used for the dissemination of information on microorganisms, makes the implementation and maintenance of

Experiment type	Experiment date	Strain number	Accession number
CHARACTER\API\RAPID ID 32 STREP (2.0)	1998-12-22	LMG 11423	
CHARACTER\API\RAPID ID 32 STREP (2.0)	2003-05-14 17:16:55	LMG 11423	
CHARACTER\FAME\TSBA40 (4.00)	1990-10-25 07:12:50	LMG 8149	
CHARACTER\FAME\TSBA40 (4.00)	1990-10-31 12:51:14	LMG 8149	
CHARACTER\FAME\TSBA40 (4.00)	1992-10-19 17:21:04	LMG 11423	
CHARACTER\FAME\TSBA40 (4.00)	1992-10-19 19:21:39	LMG 12692 t2	
CHARACTER\FAME\TSBA40 (4.00)	1992-10-20 07:13:47	LMG 12692 t1	
FINGERPRINT\REP-PCR (GTG5)	2003-11-19 17:11:53	LMG 11423	
FINGERPRINT\SDS-PAGE		LMG 8149	
FINGERPRINT\SDS-PAGE		LMG 8149 t1	
FINGERPRINT\SDS-PAGE		LMG 8149 t2	
FINGERPRINT\SDS-PAGE		LMG 11423	
FINGERPRINT\SDS-PAGE		LMG 12692 t1	
FINGERPRINT\SDS-PAGE		LMG 12692 t2	
FINGERPRINT\SDS-PAGE		LMG 12692 QC 10/92	
SEQUENCE\DNA\16S rRNA	1994-06-08	JCM 5804	D31676
SEQUENCE\DNA\16S rRNA	1998-01-08	CCUG 542	Y12906
SEQUENCE\DNA\16S rRNA	1998-03-24	JCM 5804	AB012213
SEQUENCE\DNA\16S rRNA	1999-07-22	NCFB 942	Y18294
SEQUENCE\DNA\16S rRNA	2000-07-08	DSM 20477	AJ276355
SEQUENCE\DNA\16S rRNA	2000-11-24	LMG 11423	AJ301830
SEQUENCE\DNA\16S rRNA	2001-12-21	CECT 410	AJ420800
SEQUENCE\DNA\16S-23S rRNA spacer	1997-02-21	ATCC 19434	X87180

Table 6

Polyphasic search results showing all experimental data generated for the *Enterococcus faecium* type strain, known within the integrated microbial information gateway.

separate parsers for all information sources very costly. Moreover, most exchange formats have been designed for human interpretation, which makes them unsuited for automated processing by computer agents. Adherence to worldwide accepted and implemented data exchange standards would dramatically simplify the case.

Error detection/correction algorithms can automatically discover inconsistencies by evaluation of the integrated information within the data warehouse. In most cases these inconsistencies could not have been resolved locally within the isolated data sources. This makes data integration the perfect vehicle for monitoring the data quality provided by different sources of microbial information. Performing these integrity checks improves both the information stored within the integrated data warehouse and the originating data sources.

## References

- [1] Anderberg, M.R. (1973). Cluster analysis for applications. *Academic Press*, New York and London.
- [2] CABRI (1998). Guideline for Catalogue Production, available at [www.cabri.org](http://www.cabri.org).
- [3] Carr, E.L., P. Kämpfer, B.K.C. Patel, V. Gürtler and R.J. Seviour (2003). Seven novel species of *Acinetobacter* isolated from activated sludge. *Int. J. Syst. Evol. Microbiol.*, 53, pp. 953–963.
- [4] Dice, L.R. (1945). Measures of the amount of ecological association between species. *J. Ecology*, 26, pp. 297–302.
- [5] International Nucleotide Sequence Database, publicly accessible through the DDBJ ([www.ddbj.nig.ac.jp/Welcome.html](http://www.ddbj.nig.ac.jp/Welcome.html)), EMBL ([www.ebi.ac.uk/embl/index.html](http://www.ebi.ac.uk/embl/index.html)) and GenBank ([www.ncbi.nlm.nih.gov](http://www.ncbi.nlm.nih.gov)) portals.
- [6] Kämpfer, P., U. Dreyer, A. Neef, W. Dott and H.-J. Busse (2003). *Chryseobacterium defluvii* sp. nov., isolated from

accession	species name	strain number	date	description	length
Y11621	<i>E. asini</i>	AS2 <sup>T</sup> (LMG 18727 <sup>T</sup> )	1998-06-02	E.asini 16S rRNA gene	1551
D31674	<i>E. avium</i>	NCDO 2369 <sup>T</sup> (LMG 10744 <sup>T</sup> )	1994-06-15	E.avium gene for 16S ribosomal RNA, partial sequence	166
Y18274	<i>E. avium</i>	NCFB 2369 <sup>T</sup> (LMG 10744 <sup>T</sup> )	1999-07-22	E.avium (strain NCFB 2369T) 16S rRNA gene	1429
AJ301825	<i>E. avium</i>	LMG 10744 <sup>T</sup>	2000-11-24	E.avium 16S rRNA gene, strain LMG 10744	1833
Y12907	<i>E. avium</i>	CCUG 7983 <sup>T</sup> (LMG 10744 <sup>T</sup> )	1998-01-08	E.avium 16S rRNA gene, partial (strain ...	366
X76177	<i>E. canis</i>	LMG 12316 <sup>T</sup>	1994-07-30	E. sp. (LMG12316) 16S rRNA gene	1440
Y18161	<i>E. casseliflavus</i>	NCIMB 11449 <sup>T</sup> (LMG 10745 <sup>T</sup> )	1999-07-22	E.casseliflavus 16S rRNA gene, strain NCIMB 11449	1421
AJ301826	<i>E. casseliflavus</i>	LMG 10745 <sup>T</sup> (LMG 10745 <sup>T</sup> )	2000-11-24	E.casseliflavus 16S rRNA gene, strain LMG 10745	1904
Y12908	<i>E. casseliflavus</i>	CCUG 18657 <sup>T</sup> (LMG 10745 <sup>T</sup> )	1998-01-08	E.casseliflavus 16S rRNA gene, partial (strain ...	366
AJ420804	<i>E. casseliflavus</i>	CECT 969 <sup>T</sup> (LMG 10745 <sup>T</sup> )	2001-12-21	E.casseliflavus 16S rRNA gene, strain CECT969T	1451
Y18355	<i>E. cecorum</i>	NCDO 2674 <sup>T</sup> (LMG 11741 <sup>T</sup> )	1999-07-22	E.cecorum 16S rRNA gene	1409
AJ301827	<i>E. cecorum</i>	LMG 12902 <sup>T</sup>	2000-11-24	E.cecorum 16S rRNA gene, strain LMG 12902	1667
AF061009	<i>E. cecorum</i>	ATCC 43198 <sup>T</sup> (LMG 11741 <sup>T</sup> )	1999-02-08	E.cecorum 16S ribosomal RNA gene, partial sequence	1509
Y12917	<i>E. cecorum</i>	CCUG 27299 <sup>T</sup> (LMG 11741 <sup>T</sup> )	1998-01-08	E.cecorum 16S rRNA gene, partial (strain CCUG 27299)	366
Y18275	<i>E. columbae</i>	NCIMB 13013 <sup>T</sup> (LMG 11740 <sup>T</sup> )	1999-07-22	E.columbae (strain NCIMB 13013T)16S rRNA gene	1443
X56422	<i>E. columbae</i>	NCIMB 13013 <sup>T</sup> (LMG 11740 <sup>T</sup> )	1992-03-12	E.columbae 16S rRNA gene	1493
AJ301828	<i>E. columbae</i>	LMG 11740 <sup>T</sup>	2000-11-24	E.columbae 16S rRNA gene, strain LMG 11740	1818
AF061006	<i>E. columbae</i>	ATCC 51263 <sup>T</sup> (LMG 11740 <sup>T</sup> )	1999-02-08	E.columbae 16S ribosomal RNA gene, partial sequence	1481
Y12918	<i>E. columbae</i>	CCUG 27894 <sup>T</sup> (LMG 11740 <sup>T</sup> )	1998-01-08	E.columbae 16S rRNA gene, partial (strain CCUG 27894)	366
Y18358	<i>E. dispar</i>	NCIMB 13000 <sup>T</sup> (LMG 13521 <sup>T</sup> )	1999-07-22	E.dispar 16S rRNA gene	1397
AJ301829	<i>E. dispar</i>	LMG 13521 <sup>T</sup>	2000-11-24	E.dispar 16S rRNA gene, strain LMG 13521	1875
AF061007	<i>E. dispar</i>	ATCC 51266 <sup>T</sup> (LMG 13521 <sup>T</sup> )	1999-02-08	E.dispar 16S ribosomal RNA gene, partial sequence	1514
Y12920	<i>E. dispar</i>	CCUG 33309 <sup>T</sup> (LMG 13521 <sup>T</sup> )	1998-01-08	E.dispar 16S rRNA gene, partial (strain CCUG 33309)	366
X87178	<i>E. durans</i>	ATCC 19432 <sup>T</sup> (LMG 10746 <sup>T</sup> )	1997-02-21	E.durans 16S-23S rRNA spacer DNA	277
Y18359	<i>E. durans</i>	NCFB 596 <sup>T</sup> (LMG 10746 <sup>T</sup> )	1999-07-22	E.durans 16S rRNA gene	1434
AJ276354	<i>E. durans</i>	DSM 20633 <sup>T</sup> (LMG 10746 <sup>T</sup> )	2000-07-08	E.durans 16S rRNA gene, strain DSM20633	1534
X87177	<i>E. durans</i>	ATCC 19432 <sup>T</sup> (LMG 10746 <sup>T</sup> )	1997-02-21	E.durans 16S-23S rRNA spacer DNA & tRNA-Ala gene	331
Y12909	<i>E. durans</i>	CCUG 7972 <sup>T</sup> (LMG 10746 <sup>T</sup> )	1998-01-08	E.durans 16S rRNA gene, partial (strain ...	366
AJ420801	<i>E. durans</i>	CECT 411 <sup>T</sup> (LMG 10746 <sup>T</sup> )	2001-12-21	E.durans 16S rRNA gene, strain CECT411T	1506
D31675	<i>E. faecalis</i>	NCDO 581 <sup>T</sup> (LMG 7937 <sup>T</sup> )	1994-06-08	E.faecalis gene for 16S ribosomal RNA, partial sequence	199
X87182	<i>E. faecalis</i>	ATCC 19433 <sup>T</sup> (LMG 7937 <sup>T</sup> )	1997-02-21	E.faecalis 16S-23S rRNA spacer DNA	226
Y18293	<i>E. faecalis</i>	NCIMB 775 <sup>T</sup> (LMG 7937 <sup>T</sup> )	1999-07-22	E.faecalis 16S rRNA gene	1449
AJ301831	<i>E. faecalis</i>	LMG 7937 <sup>T</sup>	2000-11-24	E.faecalis 16S rRNA gene, strain LMG 7937	1556
AB012212	<i>E. faecalis</i>	JCM 5803 <sup>T</sup> (LMG 7937 <sup>T</sup> )	1998-03-24	E.faecalis gene for 16S rRNA, partial sequence	1517
L16515	<i>E. faecalis</i>	NCTC 775 <sup>T</sup> (LMG 7937 <sup>T</sup> )	1993-05-20	E.faecalis (NCTC 775) 16S ribosomal RNA,...	418
Y12905	<i>E. faecalis</i>	ATCC 19433 <sup>T</sup> (LMG 7937 <sup>T</sup> )	1998-01-08	E.faecalis 16S rRNA gene, partial (strain ...	366
AJ420803	<i>E. faecalis</i>	CECT 481 <sup>T</sup> (LMG 7937 <sup>T</sup> )	2001-12-21	E.faecalis 16S rRNA gene, strain CECT481T	1477
D31676	<i>E. faecium</i>	JCM 5804 <sup>T</sup> (LMG 11423 <sup>T</sup> )	1994-06-08	E.faecium gene for 16S ribosomal RNA, partial sequence	1479
Y18294	<i>E. faecium</i>	NCFB 942 <sup>T</sup> (LMG 11423 <sup>T</sup> )	1999-07-22	E.faecium 16S rRNA gene	1459
AJ276355	<i>E. faecium</i>	DSM 20477 <sup>T</sup> (LMG 11423 <sup>T</sup> )	2000-07-08	E.faecium 16S rRNA gene, strain DSM20477	1533
AJ301830	<i>E. faecium</i>	LMG 11423 <sup>T</sup>	2000-11-24	E.faecium 16S rRNA gene, strain LMG 11423	1651
AB012213	<i>E. faecium</i>	JCM 5804 <sup>T</sup> (LMG 11423 <sup>T</sup> )	1998-03-24	E.faecium gene for 16S rRNA, partial sequence	1523
X87180	<i>E. faecium</i>	ATCC 19434 <sup>T</sup> (LMG 11423 <sup>T</sup> )	1997-02-21	E.faecium 16S-23S rRNA spacer DNA, strain ...	344
Y12906	<i>E. faecium</i>	CCUG 542 <sup>T</sup> (LMG 11423 <sup>T</sup> )	1998-01-08	E.faecium 16S rRNA gene, partial (strain ...	366
AJ420800	<i>E. faecium</i>	CECT 410 <sup>T</sup> (LMG 11423 <sup>T</sup> )	2001-12-21	E.faecium 16S rRNA gene, strain CECT410T	1489
Y18295	<i>E. flavescens</i>	NCIMB 13226 <sup>T</sup> (LMG 13518 <sup>T</sup> )	1999-07-22	E.flavescens 16S rRNA gene	1425
AJ301832	<i>E. flavescens</i>	LMG 13518 <sup>T</sup>	2000-11-24	E.flavescens 16S rRNA gene, strain LMG 13518	1847
Y12923	<i>E. flavescens</i>	CCUG 30567 <sup>T</sup> (LMG 13518 <sup>T</sup> )	1998-01-08	E.flavescens 16S rRNA gene, partial (strain CCUG 30567)	366
AJ420802	<i>E. flavescens</i>	CECT 4481 <sup>T</sup> (LMG 13518 <sup>T</sup> )	2001-12-21	E.flavescens 16S rRNA gene, strain CECT4481T	1514
AJ301833	<i>E. gallinarum</i>	LMG 13129 <sup>T</sup>	2000-11-24	E.gallinarum 16S rRNA gene, strain LMG 13129	1568
Y12910	<i>E. gallinarum</i>	CCUG 18658 <sup>T</sup> (LMG 11207 <sup>T</sup> )	1998-01-08	E.gallinarum 16S rRNA gene, partial (strain CCUG 18658)	366
AJ420805	<i>E. gallinarum</i>	CECT 970 <sup>T</sup> (LMG 11207 <sup>T</sup> )	2001-12-21	E.gallinarum 16S rRNA gene, strain CECT970T	1516
AY033814	<i>E. gilvus</i>	PQ1 <sup>T</sup> (CCUG 45553 <sup>T</sup> )	2002-04-04	E.gilvus 16S ribosomal RNA gene, partial sequence	1295
AF286832	<i>E. haemoperoxidus</i>	CCM 4851 <sup>T</sup> (LMG 19487 <sup>T</sup> )	2001-07-11	E.haemoperoxidus 16S ribosomal RNA gene, partial sequence	1512
X87184	<i>E. hirae</i>	ATCC 8043 <sup>T</sup> (LMG 6399 <sup>T</sup> )	1997-02-21	E.hirae 16S-23S rRNA spacer DNA	232
Y18354	<i>E. hirae</i>	NCFB 1258 <sup>T</sup> (LMG 6399 <sup>T</sup> )	1999-07-22	E.hirae 16S rRNA gene	1445
Y17302	<i>E. hirae</i>	DSM 20160 <sup>T</sup> (LMG 6399 <sup>T</sup> )	1999-02-24	E.hirae 16S rRNA gene	1535
AJ276356	<i>E. hirae</i>	DSM 20160 <sup>T</sup> (LMG 6399 <sup>T</sup> )	2000-07-08	E.hirae 16S rRNA gene, strain DSM20160	1535
AJ301834	<i>E. hirae</i>	LMG 6399 <sup>T</sup>	2000-11-24	E.hirae 16S rRNA gene, strain LMG 6399	1787
AF061011	<i>E. hirae</i>	ATCC 8043 <sup>T</sup> (LMG 6399 <sup>T</sup> )	1999-02-08	E.hirae 16S ribosomal RNA gene, partial sequence	1507
Y12912	<i>E. hirae</i>	CCUG 18659 <sup>T</sup> (LMG 6399 <sup>T</sup> )	1998-01-08	E.hirae 16S rRNA gene, partial (strain ATCC 8043...	366
AJ420799	<i>E. hirae</i>	CECT 279 <sup>T</sup> (LMG 6399 <sup>T</sup> )	2001-12-21	E.hirae 16S rRNA gene, strain CECT279T	1514
Y18339	<i>E. malodoratus</i>	NCFB 846 <sup>T</sup> (LMG 10747 <sup>T</sup> )	1999-07-22	E.malodoratus 16S rRNA gene	1461
AJ301835	<i>E. malodoratus</i>	LMG 10747 <sup>T</sup> (LMG 10747 <sup>T</sup> )	2000-11-24	E.malodoratus 16S rRNA gene, strain LMG 10747	1701
AF061012	<i>E. malodoratus</i>	ATCC 43197 <sup>T</sup> (LMG 10747 <sup>T</sup> )	1999-02-08	E.malodoratus 16S ribosomal RNA gene, partial sequence	1500
Y12911	<i>E. malodoratus</i>	CCUG 30572 <sup>T</sup> (LMG 10747 <sup>T</sup> )	1998-01-08	E.malodoratus 16S rRNA gene, partial (strain ...	366
AF286831	<i>E. moraviensis</i>	CCM 4856 <sup>T</sup> (LMG 19486 <sup>T</sup> )	2001-07-11	E.moraviensis 16S ribosomal RNA gene, partial sequence	1509
Y18340	<i>E. mundtii</i>	NCFB 2375 <sup>T</sup> (LMG 10748 <sup>T</sup> )	1999-07-22	E.mundtii 16S rRNA gene	1447
AJ301836	<i>E. mundtii</i>	LMG 10748 <sup>T</sup>	2000-11-24	E.mundtii 16S rRNA gene, strain LMG 10748	1864
AF061013	<i>E. mundtii</i>	ATCC 43186 <sup>T</sup> (LMG 10748 <sup>T</sup> )	1999-02-08	E.mundtii 16S ribosomal RNA gene, partial sequence	1529
Y12913	<i>E. mundtii</i>	CCUG 18656 <sup>T</sup> (LMG 10748 <sup>T</sup> )	1998-01-08	E.mundtii 16S rRNA gene, partial (strain CCUG 18656)	366
AJ420806	<i>E. mundtii</i>	CECT 972 <sup>T</sup> (LMG 10748 <sup>T</sup> )	2001-12-21	E.mundtii 16S rRNA gene, strain CECT972T	1521
AY033815	<i>E. pallens</i>	PQ2 <sup>T</sup> (CCUG 45554 <sup>T</sup> )	2002-04-04	E.pallens 16S ribosomal RNA gene, partial sequence	1294
AY028437	<i>E. phoeniculicola</i>	JLB-1 <sup>T</sup> (DSM 14726 <sup>T</sup> )	2001-07-02	E.phoeniculicola 16S ribosomal RNA gene,...	1479
AF335596	<i>E. porcinius</i>	ATCC 700913 <sup>T</sup> (CCUG 43229 <sup>T</sup> )	2001-01-29	E.villorum 16S ribosomal RNA gene, partial sequence	1536
Y18356	<i>E. pseudoavium</i>	NCFB 2138 <sup>T</sup> (LMG 11426 <sup>T</sup> )	1999-07-22	E.pseudoavium 16S rRNA gene	1424
AJ301837	<i>E. pseudoavium</i>	LMG 11426 <sup>T</sup>	2000-11-24	E.pseudoavium 16S rRNA gene, strain LMG 11426	1636
AF061002	<i>E. pseudoavium</i>	ATCC 49372 <sup>T</sup> (LMG 11426 <sup>T</sup> )	1999-02-08	E.pseudoavium 16S ribosomal RNA gene, partial sequence	1513
Y12916	<i>E. pseudoavium</i>	CCUG 33310 <sup>T</sup> (LMG 11426 <sup>T</sup> )	1998-01-08	E.pseudoavium 16S rRNA gene, partial (strain...	366
Y18296	<i>E. raffinosus</i>	NCIMB 12901 <sup>T</sup> (LMG 12888 <sup>T</sup> )	1999-07-22	E.raffinosus 16S rRNA gene	1425
Y12914	<i>E. raffinosus</i>	CCUG 29292 <sup>T</sup> (LMG 12888 <sup>T</sup> )	1998-01-08	E.raffinosus 16S rRNA gene, partial (strain ...	366
AF539705	<i>E. ratti</i>	ATCC 700914 <sup>T</sup> (NCIMB 13635 <sup>T</sup> )	2002-09-12	E.ratti 16S ribosomal RNA gene, partial sequence	1503
X55766	<i>E. saccharolyticus</i>	NCDO 2594 <sup>T</sup> (LMG 11427 <sup>T</sup> )	1992-03-12	S.saccharolyticus 16S rRNA gene (5')	144
Y18357	<i>E. saccharolyticus</i>	NCDO 2594 <sup>T</sup> (LMG 11427 <sup>T</sup> )	1999-07-22	E.saccharolyticus 16S rRNA gene	1456
AJ301839	<i>E. saccharolyticus</i>	LMG 11427 <sup>T</sup>	2000-11-24	E.saccharolyticus 16S rRNA gene, strain LMG 11427	1902
U30931	<i>E. saccharolyticus</i>	NCDO 2594 <sup>T</sup> (LMG 11427 <sup>T</sup> )	1996-07-31	E.saccharolyticus 16S ribosomal RNA partial sequence	1521
AF061004	<i>E. saccharolyticus</i>	ATCC 43076 <sup>T</sup> (LMG 11427 <sup>T</sup> )	1999-02-08	E.saccharolyticus 16S ribosomal RNA gene, partial sequence	1506
Y12919	<i>E. saccharolyticus</i>	CCUG 33311 <sup>T</sup> (LMG 11427 <sup>T</sup> )	1998-01-08	E.saccharolyticus 16S rRNA gene,...	366
X55767	<i>E. saccharolyticus</i>	NCDO 2594 <sup>T</sup> (LMG 11427 <sup>T</sup> )	1992-03-12	S.saccharolyticus 16S rRNA gene	1293
Y18338	<i>E. solitarius</i>	NCIMB 12902 <sup>T</sup> (LMG 12890 <sup>T</sup> )	1999-07-22	E.solitarius 16S rRNA gene	1411
AF061010	<i>E. solitarius</i>	ATCC 49428 <sup>T</sup> (LMG 12890 <sup>T</sup> )	1999-02-08	E.solitarius 16S ribosomal RNA gene,...	1341
AJ301840	<i>E. solitarius</i>	DSM 5634 <sup>T</sup> (LMG 12890 <sup>T</sup> )	2000-11-24	E.solitarius 16S rRNA gene, strain DSM 5634	1653
Y12915	<i>E. solitarius</i>	CCUG 29293 <sup>T</sup> (LMG 12890 <sup>T</sup> )	1998-01-08	E.solitarius 16S rRNA gene, partial (strain CCUG 29293)	367
Y18341	<i>E. sulfureus</i>	NCIMB 13117 <sup>T</sup> (LMG 13084 <sup>T</sup> )	1999-07-22	E.sulfureus 16S rRNA gene	1391
X55133	<i>E. sulfureus</i>	MUTK 31 <sup>T</sup> (LMG 13084 <sup>T</sup> )	1991-06-17	E.sulfureus 16S ribosomal RNA	1495
AJ301841	<i>E. sulfureus</i>	LMG 13084 <sup>T</sup>	2000-11-24	E.sulfureus 16S rRNA gene, strain LMG 13084	1902
AF061001	<i>E. sulfureus</i>	ATCC 49903 <sup>T</sup> (LMG 13084 <sup>T</sup> )	1999-02-08	E.sulfureus 16S ribosomal RNA gene, partial sequence	1498
Y12921	<i>E. sulfureus</i>	CCUG 33313 <sup>T</sup> (LMG 13084 <sup>T</sup> )	1998-01-08	E.sulfureus 16S rRNA gene, partial (strain CCUG 33313 )	366
AJ271329	<i>E. villorum</i>	LMG 12287 <sup>T</sup>	2001-06-13	E.villorum 16S rRNA gene, strain LMG 12287	1512

Table 7

Integrated microbial information gateway search results showing 16S rRNA gene sequences of *Enterococcus* spp. type strains, deposited within the International Nucleotide Sequence Database.

wastewater. *Int. J. Syst. Evol. Microbiol.*,  
53, pp. 93–97.

- [7] Lane P. and G. Lumpkin (1999). Oracle8i Data Warehousing Guide, Release 2 (8.1.6), Oracle Corporation, USA.
- [8] Sneath, P.H.A. and R.R. Sokal (1973). Numerical Taxonomy. The Principles and Practice of Numerical Classification, W.H. Freeman and Co., San Francisco.
- [9] Vandamme, P., B. Pot, M. Gillis, P. De Vos, K. Kersters and J. Swings, (1996). Polyphasic taxonomy, a consensus approach to bacterial systematics, *Microbiological Review*, 60, pp. 407–438.
- [10] WFCC-MIRCEN, World Data Centre for Microorganisms (WDCM), directory of culture collections, online available at [wdcm.nig.ac.jp](http://wdcm.nig.ac.jp).