Laying the Foundation for a Genomic Rosetta Stone:Creating Information Hubs through the User of Consensus Idenifiers

Bart Van Brabant[1], Nikos Kyrpides[2], Frank Oliver Glöckner[3,4], Tanya Gray[5], Dawn Field[5], Paul De Vos[1,6], Bernard De Baets[7], Peter Dawyndt[8]

[1]Laboratory of Microbiology, Ghent University,
K.L. Ledeganckstraat 35, B-9000 Ghent, Belgium.

[2]Department of Energy Joint Genome Institute (DOE-JGI),
2800 Mitchell Drive, Walnut Creek, California 94598, USA.

[3]Microbial Genomics Group, Max Planck Institute for Marine Microbiology,
Celsiusstrasse 1, D-28359 Bremen, Germany.

[4]Jacobs University Bremen gGmbH, D-28759 Bremen, Germany.

[5]Molecular Evolution and Bioinformatics Section, Oxford Centre for Ecology and Hydrology,
Oxford, OX1 3SR, UK

[6]BCCM^{TM}/LMG Bacteria Collection, Ghent University,
K.L. Ledeganckstraat 35, B-9000 Ghent, Belgium.

[7]Department of Applied Mathematics, Biometrics and Process Control, Ghent University,
Coupure links 653, B-9000 Ghent, Belgium.

[8]Department of Applied Mathematics and Computer Science, Ghent University,
Krijgslaan 281 S9, B-9000 Ghent, Belgium.

Summary

This paper presents a holistic approach that illustrates how the semantic hurdle for integration of biological databases might be overcome when mapping sources that provide information on individual genes and complete genomes to sources that provide information on the biological resources from which these sequences where derived, and vice versa. In particular we will explain how each of the completed and ongoing whole-genome sequencing projects in the Genomes OnLine Database and each of the ribosomal RNA sequences in the SILVA ribosomal RNA database have been persistently cross-referenced with the StrainInfo.net bioportal, serving both a genome centric and an organism centric view to the life on our blue planet as one more stepping stone towards the establishment of fully integrated and flexible biological information networks.

1 Introduction

As put forward by the 20^{th} century Austrian philosopher Ludwig Wittgenstein in his Philosophical Investigations [13], the meaning of a word is its usage in the language. Consider for example the term `F1`. When it is encoutered in a popular sports magazine, this term would most probably be used as an abbreviation for Formula One, the highest class of auto racing defined by the Fédération Internationale de l'Automobile (FIA). Users of programs running under Microsoft Windows from their part tend to rely on the `F1` function key to ask for assistance to be shown in the Help menu.

Some microbiologists, on the other hand, might immediately associate the strain `F1` to one of the most well-studied versatile environmental isolates, identified as *Pseudomonas putida*, that is capable to grow on several aromatic hydrocarbons, including benzene, toluene, ethylbenzene and *p*-cymene [6]. Its broad substrate toluene dioxygenase has been widely utilized in biocatalytic syntheses of chiral chemicals, as well as in the metabolism and detoxification of trichloroethylene, and in the production of indigo from indole. *P. putida* `F1` is known to be chemotactic to aromatic hydrocarbons and chlorinated aliphatic compounds. Well over 200 articles have been written about various aspects of *P. putida* `F1` physiology, enzymology, and genetics by microbiologists and biochemists, in addition to more applied studies by chemists and environmental engineers utilizing *P. putida* `F1` and its enzymes for green chemistry applications and bioremediation. Most recently, the completed whole-genome sequence of *P. putida* `F1` has been released to the community prior to publication [2].

To further complicate matter, microbial cultures of at least 23 strains other than *P. putida* `F1` that were labelled by the same term `F1` (or some syntactic variation thereof) are available from a global network of biological resource centers (BRC). An overview of the results retrieved from the StrainInfo.net bioportal after a search for strains that were labelled `F1` is given in Table 1. As a consequence of the ambiguity induced by a local labelling procedure used for tagging biological resources, the term `F1` hence also refers to at least 24 different microbial strains in research publications and public or proprietary databases.

The example above clearly illustrates how much the semantics of a given word or term might depend on the context wherein it is used. Whereas the associative behaviour of the human brain seems to cope well with extracting the right meaning based on contextual information, ambiguous terminology tends to cause trouble for software agents that are designed to incorporate autonomous and heterogeneous information sources into largely integrated information networks.

This paper presents a holistic approach that illustrates how the semantic hurdle for integration might be overcome when mapping sources that provide information on genes and complete genomes to sources that provide information on the biological resources from which these sequences where derived, and vice versa. In particular we will explain how each of the completed and ongoing whole-genome sequences in the Genomes OnLine Database (GOLD) [7] and each of the ribosomal RNA sequences in the SILVA ribosomal RNA database [10] have been cross-referenced with the StrainInfo.net bioportal [4], enabling direct navigation to additional information on the biological resources from which these sequences have been derived. These use

CID	Label	Taxon	Other known strain numbers
608639	F-1	*Cladosporium cladosporioides*	ATCC 58227
347348	F1	*Clostridium bifermentans*	CECT 550, CN4781, NCDO 1711, NCFB 1711, NCIB 506, NCIMB 506, NCTC 506, strain F1
596704	F-1	*Culcitalna achraspora*	ATCC 34328, CBS 143.63, Ko-45
302418	F1^{T}	*Desulfovibrio furfuralis*	DSM 2590
268505	F1	*Desulfovibrio vulgaris* subsp. *vulgaris*	DSM 6618
603298	F1	*Epidermophyton floccosum*	ATCC 44685
603162	F-1	*Fusarium oxysporum*	ATCC 46074, SUF 1318
266953	F1	*Klebsiella pneumoniae* subsp. *pneumoniae*	ATCC 13886, CCM 5789, CCTM 2047, CCUG 31616, CIP 52.144, CUETM 78-58, CUETM 78-64, E.P. Snijders F 1, F. Kauffmann A 5054, IPL 1823, Kauffmann A 5054, LMG 3086, NCDC 420-68, NCTC 5054, Snijders F1
347838	F1	*Leuconostoc mesenteroides* subsp. *cremoris*	NCDO 1085, NCIMB 701085
267281	F-1	*Leucothrix mucor*	ATCC 25907
267439	F1	*Mannheimia haemolytica*	ATCC 33374
696388	F-001	*Metarhizium anisopliae*	ATCC 200614, DAT F-001
268338	F1	*Methanobrevibacter smithii*	DSM 2374
603974	F-1	*Monascus kaoliang*	ATCC 46594
327828	F1	*Mycobacterium heidelbergense*	CCUG 42424 B
597846	F1	*Pleurotus ostreatus*	ATCC 38539
349438	F-1	*Pseudoalteromonas piscicida*	NCIMB 850
268202	F1	*Pseudomonas putida*	ATCC 700007, BCRC 17059, CCRC 17059, DSM 6899, LMG 24210, PpF1, R-36290
595499	F1	*Scopulariopsis brevicaulis*	ATCC 36009
290548	F1^{T}	*Staphylothermus marinus*	ATCC 43588, DSM 3639, JCM 9404
345570	F1	*Streptomyces* sp.	NCIMB 11785
328128	F1	*Vagococcus fluvialis*	CCUG 42871
268203	F-1	*Vibrio ichthyoenteri*	ATCC 700024, CECT 7095, IFO 15846, NBRC 15846, strain F-1
267940	F1	*Xanthomonas campestris*	ATCC 43177, P843048

Table 1: Strains having a strain number that matches the term F1, as found by searching the Integrated Strain Database underlying the StrainInfo.net bioportal. Table columns represent the globally unique culture identifier (CID) assigned during compilation of the Integrated Strain Database [4], the original strain number (type strains are indicated by ^{T}), species name and other known strain numbers for the given strain. To ease further navigation, the above table is decorated in the StrainInfo.net bioportal by mapping species names to relevant taxonomic information sources and mapping strain numbers to their online catalogue records using persistent hyperlinks.

cases demonstrate the knuckles-and-nodes strategy for integrating distributed biogical databases as followed by the StrainInfo.net bioportal in order to gather and deduplicate all downstream information known on a given microbial strain [11]. At the same time they underline the ultimate goal of the Genomic Standards Consortium to come up with ways of integrating all information related to genome or metagenome projects [5].

A short description of the three information sources that take part in the use cases is given first. This is followed by a technical description of the URI mappings and the different scenarios that were implemented and consumed to establish and manage bidirectional link integration between the different information sources. Finally, the outcomes of these initial integration efforts are discussed and held against the light of future work that needs to be done to further improve global integration of biological databases.

2 Information sources

Genomes OnLine Database – Whole-genome sequencing efforts are currently being performed by different sequencing centres, through a variety of funding sources, whereas analysis results tend to end up in a multitude of different databases. The Genomes OnLine Database [7], or GOLD for short, was established in 1997 with the overall goal to monitor all completed and ongoing whole-genome and metagenome sequencing projects worldwide from instigation to completion and to provide the community with a largely integrated database derived from diverse sources related to those projects. This comprehensive web resource currently provides information on over 2500 whole-genome sequencing projects, of which 578 have been completed and have their data deposited into the public sequence databases (statistics as of June 2007). Apart from project information and basic sequence statistics, GOLD also provides information related to organism properties such as phenotype, ecotype and disease. GOLD is available at `www.genomesonline.org`.

SILVA ribosomal RNA database – Sequencing of ribosomal RNA (rRNA) genes is currently the method of choice for phylogenetic reconstruction, nucleic acid based detection and quantification of microbial diversity, resulting in a rapid increase of publicly available rRNA sequence data over the last few years. The SILVA system (from Latin *silva*, forest), was implemented to support the users with a central comprehensive web resource for up to date, quality controlled and aligned rRNA databases comprising *Bacteria*, *Archaea* and *Eukarya* [10]. All sequences are checked for anomalies, carry a rich set of sequence associated contextual information to assist the further curation process, have multiple taxonomic classifications and the latest validly described nomenclature. Furthermore, two types of precompiled sequence datasets compatible with ARB [8] are offered for downloading on the SILVA website: (1) Reference (Ref) datasets, comprising only high quality, nearly full length sequences suitable for in-depth phylogenetic analysis and probe design and (2) the comprehensive Parc datasets with all publicly available rRNA sequences longer than 300 nucleotides suitable for biodiversity analyses. The latest publicly available database release 90 (May 2007) hosts a total of 496,842 sequences split into 422,987 small subunit and 73,855 large subunit rRNAs. SILVA is available at `www.arb-silva.de`.

StrainInfo.net bioportal – Information from 44 BRCs that cover all earth's continents and range from small niche specific research collections to large general-purpose service collections is processed into an Integrated Strain Database that underpins the StrainInfo.net bioportal [4]. In addition, information extracted from three taxonomic reference databases together with their type strains is equally incorporated. This integration process has currently lumped over 630.000 strain numbers into some 275.000 equivalence classes that represent different strains of *Bacteria*, *Archaea*, filamentous fungi and yeasts. Organisms can be searched for either by a given strain number, or by a given taxon name. A series of predefined workflows allows the resolution of some specific queries. For convenience of the users, all searches are made insensitive to many forms of orthographic variations. Query results redirect the user on-the-fly to the online catalogues of the different culture collections that have a given strain in their holdings. Additional pointers allow to navigate the downstream information available on these organisms, thereby rendering superfluous the need to browse a multitude of autonomous and heterogeneous information sources. For each organism, a list of all sequences deposited in the International Nucleotide Sequence Database Collaboration (INSDC) is given as a separate view. This overview is dynamically compiled irrespective of the strain numbers assigned to different cultures of the same organism during the sequence deposition process, solving possible Babel-like confusions. Sequence records are linked out to their corresponding records in the INSDC databases. Similarly, another view assembles all scientific publications on a given organism, with links to the PubMed repository wherever available. Finally, the integrated history and geographic views allow simple inspection of strain authenticity and availability. The StrainInfo.net bioportal is available at `www.StrainInfo.net`.

3 URI Mappings

The StrainInfo.net bioportal supports a series of template Uniform Resource Identifiers (URI) that allow third party information providers to embellish their resources by hooking up with information on micro-organisms provided by the bioportal. While designing these URIs, special attention has been paid to maximize the manageability of bidirectional cross-references to and from the bioportal, keeping in mind that both resources at either end of the reference channel must be able co-evolve independently. These template URIs turn out to become extremely helpful tools for retrieving additional information on biological resources through a simple point and click strategy, and may be integrated into online applications and web pages to enhance navigation.

Other information sources such as GOLD and SILVA may link out to information on biological resources provided by the StrainInfo.net bioportal, either by organismal references or by references to taxonomic names. In addition, the bioportal supports itself some mappings of its own unique biological resource identifiers onto a series of external identifiers, such as goldstamp identifiers attached to completed or ongoing whole-genome sequencing projects by GOLD and accession numbers assigned to sequence records by the INSDC. This avoids that external information sources need to maintain these mappings themselves.

All URI mappings return relevant information formatted as HTML documents that are retrieved through the http protocol. The URIs supported by the StrainInfo.net bioportal use the following basename:

```
http://www.StrainInfo.UGent.be/
```

All URIs mentioned further in this document use this basename as a prefix, which is silently removed for the sake of brevity.

What follows is a detailed discussion of the different mappings, including some enveloping examples and supplementary comments.

3.1 Direct mappings

Most information sources refer to a particular organism using a strain number attached to that organism. This is an alphanumeric label either assigned by an individual researcher, a research institution or a biological resources center in order to be able to discriminate between different organisms in a local context. This kind of identifier is particularly made for interpretation by human beings. However, as illustrated by the *P. putida* `F1` example in the introduction, orthographic variations and various kinds of ambiguities lead to the fact that these strain numbers are less adequate for setting up solid global cross-reference scenarios, wherein software agents can take over part of the reasoning. It is therefore strongly recommended that more and more resources make reference to a particular organism using the globally unique culture identifiers assigned during compilation of the Integrated Strain Database that underpins the StrainInfo.net bioportal, in order to get rid of problems raised by ambiguity in a global context.

Both types of organismal identifiers described above (eg. strain numbers and culture identifiers) can be used to make mappings to additional information on a particular organism as provided by the StrainInfo.net bioportal. Using either one of the organismal identifiers as a parameter to the `strainGet.jsp` URI template, direct access and integrated views are provided to a wide range of downstream information that is known about a particular organism. The following mandatory and optional parameters are used with the `strainGet.jsp` URI template:

- **`snr`** (mandatory when no value is supplied for the **`cid`** parameter) provides the strain number on which information needs to be searched for. Only in particular cases (strain numbers assigned by a selection of BRCs), the StrainInfo.net bioportal regards strain numbers as unique identifiers in a global context. Searches for organisms based upon a given strain number might thus not be specific enough to unambiguously point to a single organism. For non-unique labels, the **`view`** parameter therefore loses its meaning. Searches based upon a given strain number may use the **`queryOption`** parameter to alter search semantics, as discussed below.
- **`cid`** (mandatory when no value is supplied for the **`snr`** parameter) provides the globally unique culture identifier assigned to each labelled culture during the accumulative learning proces of compiling the Integrated Strain Database [4]. In contrast to strain numbers, this kind of identifier can always be considered unique within a global context. When this parameter is used, the **`queryOption`** parameter loses its meaning. The **`cid`** parameter takes precedence over the **`snr`** parameter if both parameters are supplied.
- **`view`** (optional, default value **`syn`**) provides the type of information (the so-called information view) that needs to be returned for the referenced organism. The following types of information are supported by the bioportal:
  - **`syn`** (default) shows a collapsible panel containing a complete list of strain numbers that are known to be assigned to the referenced organism. If available, also redirects the user to the online catalog record of the BRC that has the referenced culture of the organism in its holdings.
  - **`lit`** shows a list of literature references that provide additional information on the referenced organism, irrespective of the strain number used for that organism in the publication.

  - **`seq`** shows a list of records in the INSDC databases that provide partial or complete genome sequences derived from the referenced organism, irrespective of the strain number used for that organism in the sequence record.
  - **`his`** shows the completely integrated strain history of the referenced organism (if available).
  - **`map`** shows the geographic distribution of all BRCs that have a culture of the referenced organism in their holdings.

- **`queryOption`** (optional, default value **`1`**) influences semantics of the searches based upon a given strain number (**`snr`** parameter) using the following options:
  - **`1`** (default): for convenience of the user, the given strain number is transformed into normalized syntactical form to deal with orthographic variations, using a parsing procedure described by Dawyndt *et al.* [4]. Based upon this normalized format, a perfect search is successively performed against the database. Unique strain numbers (e.g. `BCRC 17059`) are directly resolved and automatically redirected to the online record of the corresponding BRC catalogue, while ambiguous strain numbers (e.g. `F1`) result in a list of possible matches. This parameter option corresponds to the **`normal`** search option in the web query interface of the bioportal.
  - **`2`**: only the number component is extracted from the alphanumeric strain number, after which all strain numbers that have the same number component are matched. For example, when searching for the strain number `BCRC 17059` using **`2`** as value for the **`queryOption`** parameter, all strain numbers that have `17059` as their number component will be matched. This parameter option corresponds to the **`number`** search option in the web query interface of the bioportal.
  - **`3`**: all strain numbers in the database that contain the given search term as a substring are matched. This parameter option corresponds to the **`contains`** search option in the web query interface of the bioportal.

We demonstrate how to make use of the `strainGet.jsp` URI template, by referring to the *P. putida* `F1` example that was discussed in the introduction. It can be derived from Table 1 that when a culture of this strain was deposited in the American Type Culture Collection (ATCC), it was assigned the strain number `ATCC 700007`. As this strain number is regarded unique by the StrainInfo.net bioportal, the reference

```
strainGet.jsp?snr=ATCC 70007
```

will result in an automatic redirection to the corresponding record in the online catalogue of the ATCC. In addition, a complete list of hyperlinked strain numbers known for this strain is made available through the bioportal in a collapsible panel at the left side of the page. This allows users to easily navigate the different collections that provide additional information on this strain. Remark that the value of the **`snr`** parameter is insensitive for some syntactic variations that are commonly found where strain numbers are used. As a result, the same mapping as above is returned when passing the following values for the **`snr`** parameter: `ATCC700007`, `ATCC-700007` or `atcc700007T`. By encapsulating redirection to the corresponding online catalogue records behind the URI template, location transparency and long-term persistence of organism references is achieved. These properties are of utmost importance in order to set up solid cross-reference scenarios.

The *P. putida* `F1` strain can also be referenced using `F1` as one of its strain numbers. However, as this strain number is not considered unique by the bioportal, the reference

```
strainGet.jsp?snr=F1
```

will result in a list of possible matches that resembles Table 1. To ease navigation, the search results are further decorated in the StrainInfo.net bioportal by mapping species names to relevant taxonomic information sources and mapping strain numbers to their online catalogue records using the same persistent hyperlinks as discussed above.

Unambiguous organism references are only guaranteed when globally unique culture identifiers are passed to the **`cid`** parameter of the `strainGet.jsp` URI template. It can easily be derived from Table 1 that the following reference

```
strainGet.jsp?cid=268202
```

establishes a direct mapping to the *P. putida* `F1` strain.

All examples given so far return a reference to the default synonyms view whenever unique organism identifiers were used. Mappings can be made to several other information views provided by the StrainInfo.net bioportal. The following reference

```
strainGet.jsp?cid=268202&view=his
```

for example, will return the completely integrated strain history of the *P. putida* `F1` strain. The resulting view at the time of writing is depicted in Figure 1.

3.2 Mappings based on taxonomic names

Searching the StrainInfo.net bioportal using a given taxonomic name (e.g. *Pseudomonas putida* or *Saccharomyces cerevisiae*) results in a list of organisms that were identified as the given taxon by at least one of the BRCs that have a culture of the organism in their holdings. It should be noted here that it is well possible that different BRCs might identify the same organism differently. A programming interface to this type of searches is provided through the `taxonGet.jsp` URI template. For convenience of the user, searches are made insensitive to common orthographic variations (including case insensitity and removal of tokens that refer to taxonomic rank). The following mandatory and optional parameters are used with the `taxonGet.jsp` URI template:

- **`taxon`** (mandatory) provides the taxonomic name for which a list of organisms that were identified as the given taxon by at least one of the BRCs that have a culture of the organism in their holdings needs to be searched for.

- **`subtax`** (optional, binary, default value **`1`**) allows to widen searches by including all subtaxa of the given taxon in the request. This parameter option corresponds to the **`include subtaxa`** search option in the web query interface of the bioportal.

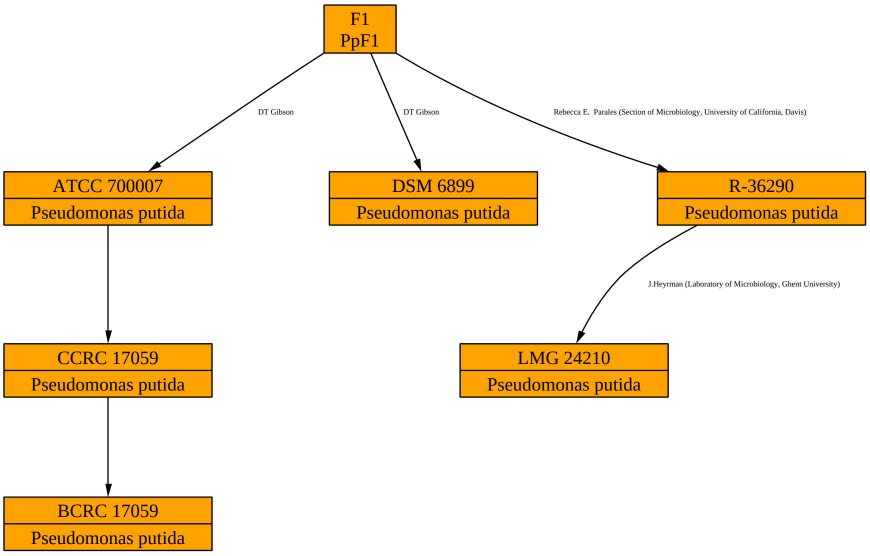

Figure 1: Completely integrated strain history of *Pseudomonas putida* `F1` as provided by the StrainInfo.net bioportal. The graphical representation shows how this strain, originally isolated from a polluted creek in Urbana (Illinois) by D.T. Gibson [6], got distributed over several BRCs worldwide that make it available for further investigation to the research community. Acronyms of BRCs: ATCC (American Type Culture Collection, United States); BCRC (Bioresource Collection and Research Center, Taiwan; formerly known as CCRC); DSM (Deutsche Sammlung von Mikroorganismen und Zellkulturen, Germany); LMG (BCCM$^{\mathrm{TM}}$/LMG Bacteria Collection, Belgium).

- **`restrict`** (optional, binary, default value **`0`**) allows to narrow searches by restricting the search results to type strains only. Type strain information of bacteria, archaea and fungi is kept up-to-date by automated orchestration with the following external information sources: i) List of Prokaryotic Names with Standing in Nomenclature (J.P. Euzéby), ii) Bacterial Nomenclature Up-to-Date (DSMZ) and iii) Mycobank (CBS) [3]. This parameter option corresponds to the **`restrict to type strains`** search option in the web query interface of the bioportal.

In order to generate a list of all organisms available in public BRCs that were identified as *Saccharomyces cerevisiae*, the following URI may be used:

```
taxonGet.jsp?taxon=Saccharomyces+cerevisiae
```

Note that the correct URI encoding needs to be followed, wherein spaces are replaced by `%20` or plus signs. To get the type strain of *Pseudomonas putida* without making a reference based on specific strain numbers or culture identifiers, the following URI may be used:

```
taxonGet.jsp?taxon=Pseudomonas+putida&subtax=0&restrict=1
```

Finally, all type strains of the genus *Bacillus* can be listed using the following URI:

```
taxonGet.jsp?taxon=Bacillus&subtax=1&restrict=1
```

NAMESPACE	IDENTIFIER DESCRIPTION	EXAMPLE
INSDC	Accession number assigned by the International Nucleotide Sequence Database Collaboration (INSDC)	CP000712
GOLD	Goldstamp assigned by the Genomes OnLine Database (GOLD)	Gc00572
ENTREZ_PID	Genome Project Identifier assigned by Entrez, the life science search engine of NCBI	13909
IMG	Integrated Microbial Genomes (IMG) Identifier assigned by the DOE Joint Genome Institute	638341163
GCAT	Genome Catalogue Identifier assigned by the Genomic Standards Consortium	001392_GCAT

Table 2: Currently supported mappings from identifiers assigned by third party information providers to additional organism information provided by the StrainInfo.net bioportal through the strainMapping.jsp URI template. Columns contain i) the value that needs to be passed to the namespace parameter, ii) a description of the identifier assigned by the corresponding information provider, and iii) an example identifier that needs to be passed to the id parameter.

3.3 Mappings to external identifiers

The StrainInfo.net bioportal aims to provide its users with as much information on a given organism as possible. One way to achieve this goal is to map the globally unique identifiers assigned to each organism during the accumulative learning stage of compiling the Integrated Strain Database [4] to a series of identifiers assigned by third party information sources that provide additional information on that organism. This way, the information content of the bioportal itself remains fairly lightweight, whereas synchronisation requirements are reduced to a strict minimum.

As all mappings from internal to external identifiers are inherently bidirectional in nature, the StrainInfo.net bioportal opens up all its mappings for further use by third party information providers through the `strainMapping.jsp` URI template. The following mandatory and optional parameters are used with this URI template:

- **`namespace`** (mandatory) indicates the information provider authorized for the assignment of a series of external identifiers that are mapped onto the organismal identifiers supplied by the StrainInfo.net bioportal. A list of mappings that are currently supported by the bioportal, together with some example identifiers, is given in Table 2.
- **`id`** (mandatory) provides the alphanumeric identifier assigned by the information provider that is referenced by the **`namespace`** parameter.
- **`view`** (optional, default value **`syn`**) has exactly the same meaning as it has when used with the `strainGet.jsp` URI template.

The StrainInfo.net bioportal for example knows that the GOLD record with goldstamp `Gc00572` describes a publicly available whole-genome sequence of the *P. putida* `F1` strain (which is uniquely referenced by the culture identifier 268202). Third party information providers may then establish a reference to the organism that was used in this whole-genome sequencing project based on the following URI:

```
    strainMapping.jsp?namespace=GOLD&id=Gc00572
```

Similarly, the bioportal knows that the accession number assigned by the INSDC to the whole-genome sequence of the *P. putida* `F1` strain is `CP000712`. Consequently, the same organism reference as above would also result from the following URI:

```
    strainMapping.jsp?namespace=INSDC&id=CP000712
```

Given the examples above, it immediately becomes clear that following this strategy, third party information providers do not even need to know that the GOLD and INSDC records that correspond with the whole-genome sequence of the *P. putida* `F1` strain are mapped to an organism identified by the culture identifier `268202` in order to provide links to additional information on that organism supplied by the StrainInfo.net bioportal. The bioportal simply takes care of all required mappings and establishes redirection to other relevant primary data sources. Hence, this approach enables to avoid duplication of efforts across multiple information providers.

4 Discussion

A broad diversity of biological resources is deposited and made publicly available through a global network of public BRCs. Hence the long-term preservation of this valuable gene pool is safeguarded for further exploitation in clinical, industrial and research applications. In terms of the integration of biological databases, the *laissez faire* attitude towards using locally assigned strain numbers is seriously hampering the effort to set up solid cross-reference scenarios involving these biological resources. After all, strain numbers are semantically weak due to the orthographic variations and ambiguities that pop up when they are used in a global context. If two gene sequences are deposited in one of the INSDC databases using the `F1` label, are they then derived from the same organism? Is the `F1` label even referring to an organism at all? From a semantic web perspective, where machines are equally expected to understand the exact meaning of shared information, it is generally recommended that these strain numbers are replaced or augmented by globally unique identifiers. The culture identifiers assigned during the accumulative learning stage of building the Integrated Strain Database, a perpetually ongoing process that underpins the StrainInfo.net bioportal, are the first organismal identifiers put in place that may fulfill this role.

The internal structure of the StrainInfo.net bioportal has entirely been built up around these unique culture identifiers. However, as strain numbers are still commonly used (and people are expected to continue to do so for the time being), the web interface of the bioportal allows using these strain numbers as an entry point to start navigating the information on biological resources provided through the bioportal. Queries based on strain numbers that are considered ambiguous identifiers by the bioportal, need some user intervention (taking into account contextual information) in order to resolve them to the correct culture identifier. Once this resolution has taken place, the bioportal for example knows exactly what INSDC sequence records are linked to each of the different strains labelled `F1` in Table 1. This relies on a mapping between the accession numbers assigned by the INSDC and the culture identifiers resolved from the strain numbers extracted from the corresponding sequence records, wherever this information was available [9].

The latter mapping underscores that the StrainInfo.net bioportal was never meant to be used solely as a standalone tool. As the relation expressed through the persistent mapping of unique identifiers assigned to biological objects by different information providers are always bidirectional in nature, these mappings might equally be consumed by the two information providers. Both providers at either end of the relationship may then make use of each others identifiers in order to establish linkouts as one way of implementing the relationship. However, this link integration strategy requires both providers to maintain a local copy of the bidirectional mapping. This brings up a synchronisation issue of the mapping between all its consuming parties. Even more so if the same mapping is consumed by several information providers. Good database design, even in a distributed environment, dictates that bidirectional mappings are maintained by one provider only, which makes it accessible for all other parties.

One way of disclosing the integrated information on biological resources supplied by the StrainInfo.net bioportal, is implemented through the template URIs presented in this paper. A first series of organism-based mappings, discussed in detail in Section 3.1, allows information providers that locally maintain a mapping of some biological object identifier onto a strain number or a culture identifiers to make linkouts to the bioportal. A second series of organism-based mappings, discussed in detail in Section 3.3, allows information providers to consume mappings that are maintained by the bioportal itself, using their local object identifiers instead of the culture identifiers at the other end of the mapping. Information providers that may better serve their users with easy navigation to additional information on biological resources, are free to make linkouts based on either of the provided template URI schemas presented in this paper.

Two frequently consulted information providers have already implemented such an organism-based linkout strategy. The effort to map culture identifiers onto goldstamp identifiers assigned by the Genomes OnLine Database (GOLD) for all completed and ongoing whole-genome sequencing projects has unveiled that only 58% of the Prokaryotes that were completely sequenced have been deposited in at least one public BRC. Only 37% of the sequenced Prokaryotes have been deposited in at least two culture collections. We are thus very far off from a situation where all sequenced strains are saved from extinction by depositing them in at least two different BRCs. A general recommendation that was previously made by several groups of authorities [1, 12]. The situation is somewhat more unclear for the Eukaryotes, as the StrainInfo.net bioportal currently only covers filamentous fungi and yeasts. From a first inspection, however, the situation there seems to be quite comparable to that of the Prokaryotes. Apart from linkouts from the GOLD database to the StrainInfo.net bioportal, the bioportal itself also offers predefined workflows that allows to dynamically generate lists of organisms used in completed and ongoing whole-genome sequencing projects.

An integration effort of somewhat larger dimensions is established by extracting organism information from the sequence records of the INSDC databases, followed by a resolution to culture identifiers that yields semantic disambiguation [9]. Apart from the sheer amount of manual interaction required to perform context-dependent disambiguation, this process is also seriously hampered by the fact that organism information is inconsistently coded in the sequence records, is subject to orthographic variations, is often missing or needs to be indirectly extracted from the publications linked to the sequence records, and may well contain several errors. While further efforts are needed to complete and correct this mapping, its current version is used to cross-reference the StrainInfo.net bioportal and SILVA databases, thus adding molecular and phylogenetic entry points to the organismal information since SILVA provides a taxonomic browser and additional search facilities that take into account different taxonomic outlines. A sequence based search engine for SILVA is currently under con-struction, opening up new opportunities for extending and performing quality control on the mapping.

Integration of the StrainInfo.net bioportal with the GOLD and SILVA data bases might be of special interest for users coming from the molecular world, where gold stamps for sequencing projects and accession numbers for sequence records seem to be used as primary access points to information, whereby the need to access organismal and ecological information only pops up further down the information extraction chain. Serving both this genome centric versus organism centric view to the life on our blue planet is one more stepping stone towards the establishment of fully integrated and flexible biological information networks.

5 Acknowledgements

The authors are indepted to all members of the Genomic Standards Consortium for their inspirational suggestions and valuable contributions to this work. The work of B. Van Brabant, P. Dawyndt, B. De Baets and P. De Vos was supported by grants C3/00/12 and IAPVI/06 from the Belgian Science Policy. P. Dawyndt wishes to thank Mark Segal (Environmental Protection Agency, US) for interesting discussions on the trace ability of biological resources and bringing the P. putida F1 strain under attention. This work was performed under the auspices of the US Department of Energy's Office of Science, Biological and Environmental Research Program, and by the University of California, Lawrence Berkeley National Laboratory under contract No. DE-AC02-05CH11231, Lawrence Livermore National Laboratory under Contract No. DE-AC52-07NA27344, and Los Alamos National Laboratory under contract No. DE-AC02-06NA25396.

References

[1] Coenye T., Vandamme P. Bacterial whole-genome sequences: minimal information and strain availability. *Microbiology* **150(7),** 2017–2018, 2004.

[2] Copeland A., Lucas S., Lapidus A., Barry K., Detter J. C., Glavina del Rio T., Hammon N., Israni S., Dalin E., Tice H., Pitluck S., Chain P., Malfatti S., Shin M., Vergez L., Schmutz J., Larimer F., Land M., Hauser L., Kyrpides N., Lykidis A., Parales R., Richardson P. (unpublished).

[3] Crous P. W., Gams W., Stalpers J. A., Robert V., Stegehuis G. MycoBank: an online initiative to launch mycology into the 21st century. *Studies in Mycology* **50**, 19–22, 2004.

[4] Dawyndt P., Vancanneyt M., De Meyer H., Swings J. Knowledge accumulation and resolution of data inconsistencies during the integration of microbial information sources. *IEEE Transactions on Knowledge and Data Engineering* **17(8),** 1111–1126, 2005.

[5] Field D., Garrity G., Gray T., Morrison N., Selengut J., Sterk P., Tatusova T., Thomson N., Allen M. J., Ashburner M., Baldauf S., Ballard S., Boore J., Cochrane G., Cole J., de Pamphilis C., Edwards R., Faruque N., Feldman R., Glockner F. O., Haft D., Hancock D., Hermjakob H., Hertz-Fowler C., Hugenholtz P., Joint I., Kane M., Kennedy J., Kowalchuk G., Kottmann R., Kolker E., Kyrpides N., Leebens-Mack J., Lewis S. E., Liste A., Lord P., Maltsev N., Markowitz V., Martiny J., Methe B., Moxon R., Nelson K., Parkhill J., Sansone S., Spiers A., Stevens R., Swift P., Tay-lor C., Tateno Y., Tett A., Turner S., Ussery D., Vaughan B., Ward N., Whetzel T., Wilson G., Wipat A. Towards a richer description of our complete collection of genomes and metagenomes: the ”Minimum Information about a Genome Se-quence” (MIGS) specification. *Nature Biotechnology* (submitted).

[6] Gibson D. T., Koch J. R., Kallio R. E. Oxidative degradation of aromatichydrocarbons by microorganisms. I. Enzymatic formation of catechol from benzene. *Biochemistry* **7**, 2 653–2661, 1968.

[7] Liolios K., Tavernarakis N., Hugenholtz P., Kyrpides N. C. The Genomes On Line Database (GOLD) v. 2: a monitor of genome projects worldwide. *Nucleic Acid Research* **34**, D332–334.

[8] Ludwig W., Strunk O., Westram R., Richter L., Meier H., Yadhukumar, Buchner A., Lai T., Steppi S., Jobb G. et al. ARB: a software environment for sequence data. *Nucleic Acid Research* **32**, 1363–1371, 2004.

[9] Romano P., Dawyndt P., Piersigilli F., Swings J. Improving interoperability between microbial information and sequence databases. *BMC Bioinformatics* **6(4),** S23, 2005.

[10] Prusse E., Quast C., Knittel K., Fuchs B. M., Ludwig W., Peplies J., Glockner F. O. SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. *Nucleic Acid Research* (submitted).

[11] Van Brabant B., Dawyndt P., De Baets B., De Vos P. A knuckles-and-nodes approach to the integration of microbiological resource data. *Lecture Notes in Computer Science* **4277**, 740–750, 2006.

[12] Ward N., Eisen J., Fraser C., Stackebrandt E. Sequenced strains must be saved from extinction. *Nature* **414(6860),** 148, 2001.

[13] Wittgenstein L. Philosophical Investigations (Philosophische Untersuchungen). Translated by G.E.M. Anscombe. NewYork: The MacMillan Company, 1953.