StrainInfo.net: Breaking down information barriers into holistic data integration scenarios using globally unique persistent identifiers

Peter Dawyndt[a,b]*, Bernard De Baets[b], Xianhua Zhou[c], Juncai Ma[c], Jean Swings[a,d]

[a] Laboratory of Microbiology, Ghent University, Belgium
[b] Department of Applied Mathematics, Biometrics and Process Control, Ghent University, Belgium
[c] Information Network Center, Institute of Microbiology, Chinese Academy of Sciences, PR China
[d] BCCM™/LMG Bacteria Collection, Ghent University, Belgium
*corresponding author: Peter.Daywndt@UGent.be

With the advent and rapid emergence of the Internet, world wide access to multiple public microbial information sources has given a strong impetus to research in the field of microbiology, by instantly disseminating the latest breakthroughs and insights within the problem domain and establishing in the long term a pool of the microbiologist's collective knowledge. The online catalogues of biological resource centres (BRCs) provide basic information on the isolation, identification and availability of many important and well-characterized micro-organisms. Empirical databases contain information on many of the genotypic and phenotypic traits of these microbial strains for a broad range of experimental techniques, which seriously vary in their inter-laboratory standardization and reproducibility. As such, the International Nucleotide Sequence Database[3] has emerged as one of the greatest successes in the accumulation of reproducible experimental information, providing parts or the whole genetic map of many of the life forms on earth. Completely new branches of research, such as computational genomics, have been established on the foundations of these sequence databases. Finally, probably the largest contributions to microbial research are at present only published in the scientific literature, which in itself forms a heterogeneous knowledge base that is progressively accessible in electronic form[6].

The bewildering proliferation of these massive amounts of information urges the establishment of solid cross-references between different autonomous and heterogeneous data sources, in order to reduce the amount of data duplication between the information providers, assist the researchers in the navigating this data-rich environment by merging all relevant information into a uniform view, discover new insights through knowledge discovery in databases and monitor the overall data quality provided by different web services through continuous quality control of the integrated information. Primordial to the establishment of durable cross-reference scenarios is the assignment of globally unique and persistent object identifiers for unequivocal discrimination, persistent localisation and autonomous integration of the different entities in the problem domain. As an example, the assignment of accession numbers as the unique object identifiers for genetic sequence data and the assignment of digital object identifiers (DOI[4]) that act as proxies to scientific publications, have enabled a cross-referencing scheme that maintains mutual links between the International Nucleotide Sequence Database and the large PubMed literature repository collecting scientific publications from the life sciences. However, as a consequence of the lack of globally unique identifiers for micro-organisms kept in BRCs, these living biological resources have only started to become involved in similar information cross-reference scenarios[5]. Instead, the necessary information about microbial resources is partially copied into the peripheral data sources, which perturbs the management of this information that is subjected to dynamic changes.

The StrainInfo.net portal (www.straininfo.net) envisions to overcome some of the problems related to the integration of basic information on biological resources as disseminated by hundreds of BRCs worldwide with the dynamically growing amount of downstream information that is generated on these organisms. A key issue in the philosophy of the portal is the compilation of the Integrated Strain Database[1,2], a central knowledge base that accumulatively learns about the equivalence relation that

exists amongst the strain numbers assigned to biological resources in a global research context, by means of the calculation of the transitive closure. Currently, information is gathered from 42 microbial culture collections that cover all earth's continents and range from small niche specific research collections to large general-purpose service collections. In addition, the information extracted from two lists of bacterial type strains is equally incorporated. This integration process has currently lumped over 600.000 strain numbers into some 250.000 equivalence classes that represent different strains of bacteria, archaea, filamentous fungi and yeasts. Special attention has been paid to error detection and correction within the equivalence classes due to irregularities in the data provided by the underlying information sources, through the design of novel intelligent tools that enable the automatic discovery of intrusions in the consistency of the integrated information. Without profound quality control of the integrated information, at least 719 (11.89%) of the bacterial type strains would have been affected by illegitimate merges into single equivalence classes[2].

While incrementally calculating the strain equivalence classes, new unique identifiers are assigned to strain numbers that were not previously encountered during the integration procedure. This helps to resolve some of the ambiguities that are a logical consequence of the local nature of the strain number assignment process and enables to set down context-dependant resolution of ambiguous strain numbers that often require some form of human-intervention. The latter is important to secure the tedious disambiguation procedure of existing cross-references for correct machine interpretation in the future. Moreover, it turns out that the information content of the Integrated Strain Database offers the perfect semantic context to guide the disambiguation process in a number of ways. To demonstrate the potential of the StrainInfo.net portal to fill the gap where there is no universally adopted system for assigning and recognizing persistent and unique identifiers for biological resources, we have consolidated the strain information captured within the Integrated Strain Database with relevant sequences and literature references assembled within public repositories. Not only does this offer a de-duplicated view on the downstream information that is available on the micro-organisms worldwide, but also allows for the execution of all sorts of dynamic queries that can automatically bridge over multiple web services that were physically separated before the integration process. The presented cross-reference model will however only show its full dynamic strength when the reverse references to the Integrated Strain Database are included in third party databases, thus establishing a true divide and conquer strategy for tracking related information within autonomously operating biological information sources.

[1] **Dawyndt P., Vancanneyt M. & Swings J. (2004).** On the integration of microbial information. *WFCC Newsletter* **38**, 19-34.

[2] **Dawyndt P., Vancanneyt M., De Meyer H. & Swings J. (2005).** 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.

[3] **Kanz C. *et al.* (2005).** The EMBL Nucleotide Sequence Database. *Nucleic Acids Research* **33**, D29-D33.

[4] **Paskin N. (2005).** The DOI handbook. Edition 4.2.0, International DOI Foundation, Inc.http://dx.doi.org/10.1000/182.

[5] **Romano P., Dawyndt P., Piersigilli F. & Swings J. (submitted).** Improving interoperability between microbial information and sequence databases. *BMC Bioinformatics*.

[6] **Wheeler D. L., Church D. M., Edgar R., Federhen S., Helmberg W., Madden T. L., Pontius J. U., Schuler G. D., Schriml L. M., Sequeira,E. *et al.* (2004).** Database resources of the National Center for Biotechnology Information: update. *Nucleic Acids Research* **32**, D35-D40.