Verslyppe et al. BMC Bioinformatics 2010, 11(Suppl 5):P3
http://www.biomedcentral.com/1471-2105/11/S5/P3

BMC Bioinformatics

POSTER PRESENTATION Open Access

Semantic integration of isolation habitat and location in StrainInfo

Bert Verslyppe[1,2*], Wim De Smet[1], Paul De Vos[1,3], Bernard De Baets[4], Peter Dawyndt[2]

From Workshop on Advances in Bio Text Mining
Ghent, Belgium. 10-11 May 2010

StrainInfo (http://www.straininfo.net) is a global catalog of microbial material, building upon the catalogs of Biological Resource Centers (BRCs) by integrating catalog entries of equivalent microbial material. Currently, the integration algorithm resolves the equivalent cultures and links all downstream information [1]. However, in order to increase the information content of StrainInfo, it is necessary to add fine-grained semantic information. This information enters StrainInfo on the culture level (synchronization with BRC catalogs), but must be integrated to the strain level (i.e. the set of equivalent cultures) in order to be presented on so-called strain passports.

The adoption of Microbiological Common Language (MCL) XML synchronization quickly increased the volume of semantic data in StrainInfo [2]. However, the effective data values of the different semantic fields still are raw textual entries and therefore are of varying detail, can have different forms or languages, and sometimes contain inconsistencies or even true errors. By consequence, in order to generate a strain level consensus value for each field, a specialized semantic integration of this data needs to be developed. As a case study for semantic integration in StrainInfo, the focus was put on the isolation habitat and location information fields due to their importance from both biological and legal (IP rights) perspective. An example of such data can be found in Table 1.

To integrate geographical information, named entity recognition is performed by annotating all geographic names with features from the GeoNames ontology. This yields a multitude of annotations, each annotation matching a name with one or more geographical features. As a large number of geographic names is not unique (e.g. Cambridge becoming annotated with both the USA and the UK instance), irrelevant annotations are removed by using other higher order features such as countries or continents found in the strain. In addition, the most specific feature is selected by removing the higher order features as this is redundant information that can be inferred from the ontology. The remaining annotation is the integration result; multiple remaining annotations or features being too distant indicate inconsistent data.

Table 1 Example isolation habitat and location data of a *Pichia guilliermondii* strain, as listed by different BRCs. For each column, we want to calculate a consensus value for the complete strain

Strain number of culture	Isolation habitat	Isolation location
CECT 1456	insect frass on Ulmus americana (elm tree)	n/a
CLIB 515	Insect	USA
MUCL 49143	insect frass on Ulmus americana (elm tree)	America, United States, Illinois, Peoria
VKM Y-1018	frass of insects infesting Ulmus americana	USA
VKM Y-1256	frass of insects infesting Ulmus americana	Illinois, USA
STRAIN	**?**	**?**

* Correspondence: Bert.Verslyppe@UGent.be
[1]Laboratory of Microbiology, Department of Biochemistry and Microbiology, Ghent University, K.L. Ledeganckstraat 35, 9000 Ghent, Belgium
Full list of author information is available at the end of the article

The habitat fields can also be integrated using a similar algorithm. However, in order to have enough ontological coverage, a combination of the Environmental Ontology (EnvO), the NCBI Taxonomy and Foundational Model of Anatomy (FMA) ontology is used. This possibly yields multiple orthogonal annotations, but for this field, having multiple annotations increases the information content and therefore does not indicate inconsistencies.

Author details
[1]Laboratory of Microbiology, Department of Biochemistry and Microbiology, Ghent University, K.L. Ledeganckstraat 35, 9000 Ghent, Belgium. [2]Department of Applied Mathematics and Computer Science, Ghent University, Krijgslaan 281, 9000 Ghent, Belgium. [3]BCCM™/LMG Bacteria Collection, Ghent University, K.L. Ledeganckstraat 35, 9000 Ghent, Belgium. [4]Department of Applied Mathematics, Biometrics and Process Control, Ghent University, Coupure links 653, 9000 Ghent, Belgium.

Published: 6 October 2010

References

1. Dawyndt P, Vancanneyt M, De Meyer H, Swings J: **Knowledge accumulation and resolution of data inconsistencies during the integration of microbial information sources.** *IEEE Trans. Knowl. Data Eng* 2005, **17**:1111-1126.
2. Verslyppe B, Kottmann R, De Smet W, De Baets B, De Vos P, Dawyndt P: **Microbiological Common Language (MCL): a standard for electronic information exchange in the Microbial Commons.** *Res. Microbiol* 2010, **161(6)**:439-445, doi:10.1016/j.resmic.2010.02.005.

doi:10.1186/1471-2105-11-S5-P3
Cite this article as: Verslyppe *et al.*: Semantic integration of isolation habitat and location in StrainInfo. *BMC Bioinformatics* 2010 **11**(Suppl 5):P3.