**BioLexicon: Towards a reference terminological resource in the biomedical domain**

D. Rebholz-Schuhmann[1], P. Pezik[1], V. Lee[1], R. del Gratta[2], J.J. Kim[1], Y. Sasaki[3], J. McNaught[3], S. Montagni[2], M. Monachini[2], N. Calzolari[2], S. Ananiadou[3]

[1]European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SD, U.K.
[2]ILC-CNR, Area della Ricerca del CNR, Via Giuseppe Moruzzi N° 1, 56124 Pisa, Italy
[3]NaCTeM, The University of Manchester, Oxford Road, Manchester, M13 9PL, UK

BioLexicon is a large-scale terminological resource which has been developed to address the needs emerging in text mining efforts in the biomedical domain. In the first stage of the construction of the BioLexicon, potential terms are pooled together from several resources representing selected semantic types of entities, such as genes and proteins, chemical compounds, species, enzymes, as well as various entities found in biological ontologies [1]. Terms contained in this initial term repository are organized into sets of synonymous variants and annotated with a number of static features which improve the resolution of term ambiguity [2].

Once populated with terms from existing repositories, the BioLexicon is augmented with term variants extracted from the scientific literature and complemented with manually selected lexical items, such as biologically relevant verbs and multi word token expressions. Last but not least, a subset of terms in the BioLexicon is linked to Gene Regulation Ontology concepts to support the identification of gene regulatory events (http://www.ebi.ac.uk/Rebholz-srv/GRO/GRO.html) [3].

The schema of the BioLexicon preserves term annotations and metadata derived from the original data resources. At the same time, it provides consistent lexical representation for terms of different semantic types. BioLexicon thus offers the clear advantage of a uniform lexical format for a wide coverage of biological terminology.

The BioLexicon is publicly available both as an XML-formatted term repository and as a relational database (MySQL) and it adheres to the Eagle ISO standards for lexical resources (www.boostrep.org, http://www.ebi.ac.uk/Rebholz-srv/BootStrep/bootstrep.html) [4].

[1] Liu,H., Hu,Z., Zhang,J. and Wu,C., (2006). BioThesaurus: a web-based thesaurus of protein and gene names. Bioinformatics 2006 22(1):103-105.

[2] Pezik,P., Jimeno,A., Lee,V., and Rebholz-Schuhmann, D. (2008) Static Dictionary Features for Term Polysemy Identification. Proceedings of the Language Resources and Evaluation Conference (LREC-2008), workshop on "Building and evaluating resources for biomedical text mining", Marrakech (Morocco), 28-30 May 2008.

[3] Beisswanger,E., Lee,V., Kim,J.J., Rebholz-Schuhmann,D., Splendiani,A., Dameron,O., Schulz,S., Hahn,U. Gene Regulation Ontology (GRO): Design Principles and Use Cases. Medical Informatics Europe 2008, Göteborg, Sweden May 25-28, 2008.

[4] ISO FDIS 24613: 2008 Language Resource Management - Lexical Markup Framework, ISO/TC37/SC4 Geneva.