



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

Joint and collaborative work of the following teams

**D. Rebholz-Schuhmann, P. Pezik, V. Lee, J.J. Kim**

European Bioinformatics Institute, Wellcome Trust Genome Campus, Cambridge, CB10 1SD, U.K.

**N. Calzolari, M. Monachini, S. Montemagni, R. del Gratta, S. Marchi, V. Quochi**

ILC-CNR, Area della Ricerca del CNR, Via Giuseppe Moruzzi N° 1, 56124 Pisa, Italy

**S. Ananiadou, J. McNaught, Y. Sasaki**

School of Computer Science, The University of Manchester, 131 Princess Street, M1 7DN, U.K.

## Generation of the BioLexicon

The BioLexicon is a large-scale terminological resource developed to address text mining requirements in the biomedical domain. In the first stage of the construction of the BioLexicon, potential terms are pooled from several resources representing selected semantic types of entities, such as genes and proteins, chemical compounds, species, enzymes, as well as various entities and concepts found in biological ontologies [1]. The content has been optimized to serve the needs of the BOOTStrep project to extract gene regulatory events from the scientific literature and to serve as a large-scale reference terminological resource. 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 [2,3].

Currently, the BioLexicon contains transcription factors (160 entries), operons (2,672), sequence ontology concepts (1,431), enzymes (4,016), protein domains (16,940), protein complexes (2,104), genes and proteins (358,335), chemicals (CHEBI, 19,637), diseases (19,457), gene ontology concepts (25,219), Taxa (NCBI, 482,992), molecular role concepts (8,850), CELLO (842), and verb forms for the support of information extraction. The verb forms are subdivided into sets of domain specific verbs (658), of derived verb forms (2,764) and inflected verb forms (8,356). The number of term variants is several times bigger than the number of term entries.

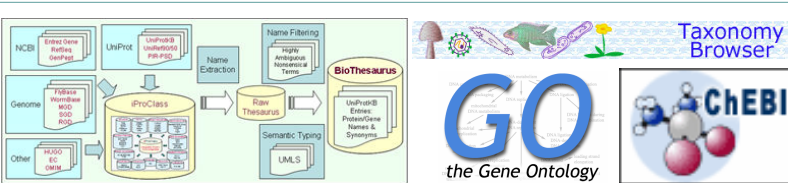
Once populated with terms from existing repositories, the BioLexicon is augmented with term variants extracted from the scientific literature [3,4] and complemented with manually selected lexical items, such as biologically relevant verbs and multi word token expressions. For each biologically relevant verb, information about its subcategorization behaviour [5] and associated semantic frames is acquired from domain corpora and recorded in the BioLexicon. Lastly, 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.nml>) [6].

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

## Availability of the BioLexicon

The BioLexicon is publicly available both as an XML-formatted term repository and as a relational database (MySQL) and it adheres to the LMF ISO standards for lexical resources ([www.boostrep.org](http://www.boostrep.org)) [7,8]. It can be downloaded from all three sites which developed it.

- [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] Tsuruoka, Y., McNaught, J. and Ananiadou, S. (2008) Normalizing biomedical terms by minimizing ambiguity and variability. *BMC Bioinformatics*, 9(Suppl 3):S2.
- [4] Sasaki, Y., Tsuruoka, Y., McNaught, J., and Ananiadou, S. (2008) How to Make the Most of Named Entity Dictionaries in Statistical Named Entity Recognition. *ACL BioNLP*, pp. 63-70.
- [5] Lenci, A., McGillivray, B., Montemagni, S., Pirrelli, V., Unsupervised acquisition of verb subcategorization frames from shallow-parsed corpora. In *Proceedings of the 6th International Conference on Language Resources and Evaluation, Marrakech (Morocco)*, 28-30 May 2008
- [6] 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.
- [7] ISO FDIS 24613: 2008 Language Resource Management - Lexical Markup Framework, ISO/TC37/SC4 Geneva.
- [8] Quochi, V., Monachini, M., Del Gratta, R., and Calzolari, N. (2008) "A lexicon for biology and bioinformatics: the BOOTStrep experience" *Proceedings of the LREC'08*. 28-30 May 2008, Marrakech, Morocco.



Semantic type	Resources
Cell	Cell ontology
CellComponent	Gene Ontology GO:0005575 cellular component
Chemical	CHEBI, IMR:0000947 chemical
Disease	OMIM
Enzyme	Enzyme commission
Gene	BioThesaurus
Ligand	IMR - INOH Protein name/family name ontology
NuclearReceptor	GO:0004879 ligand-dependent nuclear receptor activity
NucleicAcidRegion	Sequence Ontology :Region
Operon	RegulonDB, ODB (Operon DataBase)
Organism	NCBI Species
TranscriptionFactor-BindingSite	Sequence Ontology
Protein	BioThesaurus
ProteinComplex	Corum database
ProteinDomain	InterPro
TranscriptionRegulator	RegulonDB, TransFac, Gene Ontology Annotation

## Term Repository (XML)

**Term features:** id, source, variants, preferred term, Acronym

**Statistics:** Frequencies in British National Corpus, Medline and MeSH;

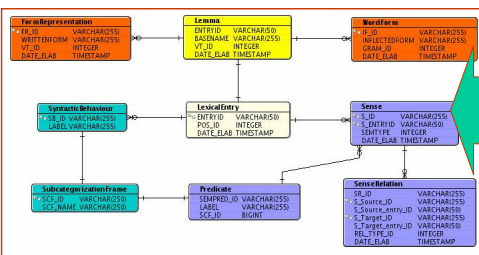
**Qualifiers:** isAcronym, isGenericEnzymeName

## BioLexicon

MySQL

Integration of terms and senses

Augmented by terms manually curated and mined from the literature  
Verb subcategorization frames



For technical support contact:

[textmining-support@ebi.ac.uk](mailto:textmining-support@ebi.ac.uk)

**BootStrep Web page:**

<http://www.boostrep.org>

**BootStrep resources at the EBI:**

<http://www.ebi.ac.uk/Rebholz-srv/BootStrep/bootstrep.html>



**Appreciation of collaborative work:** EMBL-EBI generated and reproduced the term repository and added statistical and feature information to the contained terms. UoM extracted relevant terms from the scientific literature and coordinated the population of the BioLexicon, to which also CNR-ILC contributed for what concerns the acquisition of verb subcategorization frames. CNR-ILC produced the BioLexicon database and integrated all data from the different sites. Friedrich-Schiller University (Prof. Udo Hahn, Julie) is the coordinator of the research project.

## Acknowledgements

This research was sponsored by the EC STREP project "BOOT-Strep" (FP6-028099, [www.bootstrep.org](http://www.bootstrep.org)) including the development of the Term Repository.

Dietrich Rebholz-Schuhmann  
Rebholz group, text mining  
European Bioinformatics Institute  
Wellcome Trust Genome Campus  
Hinxton, Cambridge, CB10 1SD  
U.K.  
[rebholz@ebi.ac.uk](mailto:rebholz@ebi.ac.uk)  
<http://www.ebi.ac.uk>

Sophia Ananiadou  
National Centre for Text Mining  
School of Computer Science  
University of Manchester  
Manchester Interdisciplinary Biocentre  
Manchester, M1 7DN, U.K.  
[Sophia.Ananiadou@manchester.ac.uk](mailto:Sophia.Ananiadou@manchester.ac.uk)  
<http://www.nactem.ac.uk>

Nicoletta Calzolari  
ILC-CNR  
Area della Ricerca del CNR  
Via Giuseppe Moruzzi N° 1  
56124 Pisa  
Italy  
[nicoletta.calzolari@ilc.cnr.it](mailto:nicoletta.calzolari@ilc.cnr.it)  
<http://www.ilc.cnr.it/>

Udo Hahn  
Julie Lab  
Friedrich Schiller University Jena  
Fürstengraben 30  
07743 Jena  
Germany  
[hahn@coling.uni-jena.de](mailto:hahn@coling.uni-jena.de)  
<http://www.coling.uni-jena.de/>