# Text mining for characterizing cells and tissues

Mariana Neves[1,2], Alexander Damaschun[2], Nancy Mah[3], Fritz Lekschas[2], Stefanie Seltmann[2], Harald Stachelscheid[2], Jean-Fred Fontaine[3], Andreas Kurtz[2], and Ulf Leser[1]

[1]Humboldt-Universität zu Berlin, WBI, Berlin, Germany
[2]Berlin Brandenburg Center for Regenerative Therapies, Charité, Berlin, Germany
[3]Max Delbrück Center for Molecular Medicine, Berlin, Germany

Regenerative medicine is an important field for translational medical research provided its potential for repair, restoration and replacement of tissues [1]. One of the requirement of regenerative approaches is the well characterization of therapeutic cell populations based on reliable measurement and analysis techniques. The biological scientific literature contains a huge number of citations on the expression of genes/proteins in a variety of cell lines, cell types or tissues, for instance (PMID 20085633): *Pros is never expressed in glial cells.* In this work, we show that text mining can help characterizing cell and tissues and we describe the results we have achieved so far in the scope of the CellFinder database.

The CellFinder database[1] is a repository of cell research which aims to integrate data derived from many sources, such as literature curation and microarray data. As part of the text mining development, a set of 20 full text documents on human embryonic stem cell [2] and on kidney research have been manually annotated with almost 3,000 gene expression events on cells and tissues. Later, they have been utilized for evaluation and training of supervised learning methods as part of a text mining pipeline [3], which include document triage, named-entity recognition for a variety of types and event extraction. A first evaluation of the pipeline resulted on the curation of more than 1,800 gene expression events.

CellFinder currently includes more than 4,500 facts on the expression of particular genes in particular cells, derived from more than 800 full text publications. This literature-derived data have been automatically normalized to a variety of ontologies (CL, CLO, FMA, EHDAA2, UBERON) and to the EntrezGene database to allow integration into the database, after previous manual validation of the identifiers. Currently, validated data corresponds to more than 150 and 800 distinct anatomical terms and genes/proteins, respectively.

**Funding**: This work was supported by the DFG [LE 1428/3-1 and KU 851/3-1].

[1] `http://www.cellfinder.org/`

## References

[1] F.-M. Chen, Y.-M. Zhao, Y. Jin, and S. Shi. Prospects for translational regenerative medicine. *Biotechnology Advances*, 30(3):658 – 672, 2012.

[2] M. Neves, A. Damaschun, A. Kurtz, and U. Leser. Annotating and evaluating text for stem cell research. In *Proceedings of the Third Workshop on Building and Evaluation Resources for Biomedical Text Mining (BioTxtM 2012) at Language Resources and Evaluation (LREC) 2012*, pages 16–23, Istanbul, Turkey, 2012.

[3] M. Neves, A. Damaschun, N. Mah, F. Lekschas, S. Seltmann, H. Stachelscheid, J.-F. Fontaine, A. Kurtz, and U. Leser. Preliminary evaluation of the cellfinder literature curation pipeline for gene expression in kidney cells and anatomical parts. *Database*, 2013, 2013.