

Resource Summary Report

Generated by NIF on Aug 28, 2026

Textpresso

RRID:SCR_008737

Type: Tool

Proper Citation

Textpresso (RRID:SCR_008737)

Resource Information

URL: <http://www.textpresso.org/>

Proper Citation: Textpresso (RRID:SCR_008737)

Description: An information extracting and processing package for biological literature that can be used online or installed locally via a downloadable software package, <http://www.textpresso.org/downloads.html> Textpresso's two major elements are (1) access to full text, so that entire articles can be searched, and (2) introduction of categories of biological concepts and classes that relate two objects (e.g., association, regulation, etc.) or describe one (e.g., methods, etc). A search engine enables the user to search for one or a combination of these categories and/or keywords within an entire literature. The Textpresso project serves the biological and biomedical research community by providing:

- * Full text literature searches of model organism research and subject-specific articles at individual sites. Major elements of these search engines are (1) access to full text, so that the entire content of articles can be searched, and (2) search capabilities using categories of biological concepts and classes that relate two objects (e.g., association, regulation, etc.) or identify one (e.g., cell, gene, allele, etc). The search engines are flexible, enabling users to query the entire literature using keywords, one or more categories or a combination of keywords and categories.
- * Text classification and mining of biomedical literature for database curation. They help database curators to identify and extract biological entities and facts from the full text of research articles. Examples of entity identification and extraction include new allele and gene names and human disease gene orthologs; examples of fact identification and extraction include sentence retrieval for curating gene-gene regulation, Gene Ontology (GO) cellular components and GO molecular function annotations. In addition they classify papers according to curation needs. They employ a variety of methods such as hidden Markov models, support vector machines, conditional random fields and pattern matches. Our collaborators include WormBase, FlyBase, SGD, TAIR, dictyBase and the Neuroscience Information Framework. They are looking forward to collaborating with more model organism databases and projects.
- * Linking biological entities in PDF and online journal articles to online databases. They have established a journal article mark-up pipeline that links select content of Genetics journal articles to model organism databases such as WormBase and

SGD. The entity markup pipeline links over nine classes of objects including genes, proteins, alleles, phenotypes, and anatomical terms to the appropriate page at each database. The first article published with online and PDF-embedded hyperlinks to WormBase appeared in the September 2009 issue of Genetics. As of January 2011, we have processed around 70 articles, to be continued indefinitely. Extension of this pipeline to other journals and model organism databases is planned. Textpresso is useful as a search engine for researchers as well as a curation tool. It was developed as a part of WormBase and is used extensively by *C. elegans* curators. Textpresso has currently been implemented for 24 different literatures, among them Neuroscience, and can readily be extended to other corpora of text.

Abbreviations: Textpresso

Synonyms: Text presso, Textpresso - literature search engine

Resource Type: data or information resource, database, software application, software resource, text-mining software

Defining Citation: [PMID:18949581](#), [PMID:15383839](#)

Keywords: literature, extract, process, bibliographic resource, database application, linux, macos, pdf, perl, posix/unix-like, sh, bash, unix shell, web service, search engine, curation tool, dicty, neuroscience, regulon db, ecoliwiki, ecocyc, curation, text-mining

Funding: NHGRI HG004090

Availability: Textpresso License

Resource Name: Textpresso

Resource ID: SCR_008737

Alternate IDs: nlx_143812, OMICS_01199

Alternate URLs: <http://www.nitrc.org/projects/textpresso-2-0/>

Record Creation Time: 20220129T080249+0000

Record Last Update: 20260821T193842+0000

Ratings and Alerts

No rating or validation information has been found for Textpresso.

No alerts have been found for Textpresso.

Data and Source Information

Source: [SciCrunch Registry](#)

Usage and Citation Metrics

We found 10 mentions in open access literature.

Listed below are recent publications. The full list is available at [NIF](#) .

Nenni MJ, et al. (2019) Xenbase: Facilitating the Use of Xenopus to Model Human Disease. *Frontiers in physiology*, 10, 154.

Venkatesan A, et al. (2016) SciLite: a platform for displaying text-mined annotations as a means to link research articles with biological data. *Wellcome open research*, 1, 25.

Ozyurt IB, et al. (2016) Resource Disambiguator for the Web: Extracting Biomedical Resources and Their Citations from the Scientific Literature. *PloS one*, 11(1), e0146300.

Collier N, et al. (2013) Toward knowledge support for analysis and interpretation of complex traits. *Genome biology*, 14(9), 214.

Bandrowski AE, et al. (2012) A hybrid human and machine resource curation pipeline for the Neuroscience Information Framework. *Database : the journal of biological databases and curation*, 2012, bas005.

Schindelman G, et al. (2011) Worm Phenotype Ontology: integrating phenotype data within and beyond the *C. elegans* community. *BMC bioinformatics*, 12, 32.

Warden R, et al. (2010) The Internet and science communication: blurring the boundaries. *Ecancermedicalscience*, 4, 203.

Van Auken K, et al. (2009) Semi-automated curation of protein subcellular localization: a text mining-based approach to Gene Ontology (GO) Cellular Component curation. *BMC bioinformatics*, 10, 228.

Hunter L, et al. (2006) Biomedical language processing: what's beyond PubMed? *Molecular cell*, 21(5), 589.

Krallinger M, et al. (2005) Text-mining and information-retrieval services for molecular biology. *Genome biology*, 6(7), 224.