

USING TEXTPRESSO FOR INFORMATION RETRIEVAL, FACT EXTRACTION AND DATABASE ENTRY

Karen Yook, Kimberly Van Auken, Paul Sternberg and the WormBase Consortium, Division of Biology, California Institute of Technology, Pasadena, CA, USA

Ten years ago WormBase¹ started as a repository for sequence data for the model organism *Caenorhabditis elegans* and has since striven to include the curation of all genetic and molecular data published for this nematode. With a publication rate in the *C. elegans* field of approximately 800 papers per year, WormBase (WB) has the opportunity to include information from every paper published. Currently there are ~11,000 full text research papers (mid-1970's to the present) downloaded into the WB curation database, from which over 27 data types (i.e. genetic interactions, transgene objects, gene expression patterns, mutant phenotypes etc.) are extracted by curators. Textpresso² is an open source text-mining tool capable of rapid searches for keywords, as well as concepts, from the full text of research papers. Curators at WB use Textpresso on a daily basis for many aspects of literature curation, from simple keyword searches to semi- or fully automated entity and fact extraction, which feed into curation pipelines or directly into the curation database itself. In addition, Textpresso greatly aids prioritization of literature curation by retrieving papers based on their full contents rather than solely on their abstracts. Such retrievable contents can range from the very particular (such as a gene simply being mentioned in the Materials and Methods section of a paper) to the complex (such as molecular functions that involve cellular components). As WB expands to incorporate the genomes of other nematodes, we will be working with Textpresso developers to set up a library of literature for related nematodes. We expect Textpresso to be crucial for most efficiently directing our efforts in literature curation, and for most quickly providing data to users searching the literature. In this workshop we will show how we use Textpresso in our curation pipeline to help with literature queries, to prioritize our workflow, and to automate data and fact extraction.

¹ <http://www.wormbase.org>

² <http://www.textpresso.org>