CANTO: A CONFIGURABLE TOOL FOR COMMUNITY LITERATURE CURATION

Antonia Lock[1], Kim Rutherford[2], Midori Harris[2], Steve Oliver[2], Jürg Bähler[1], Valerie Wood[2]

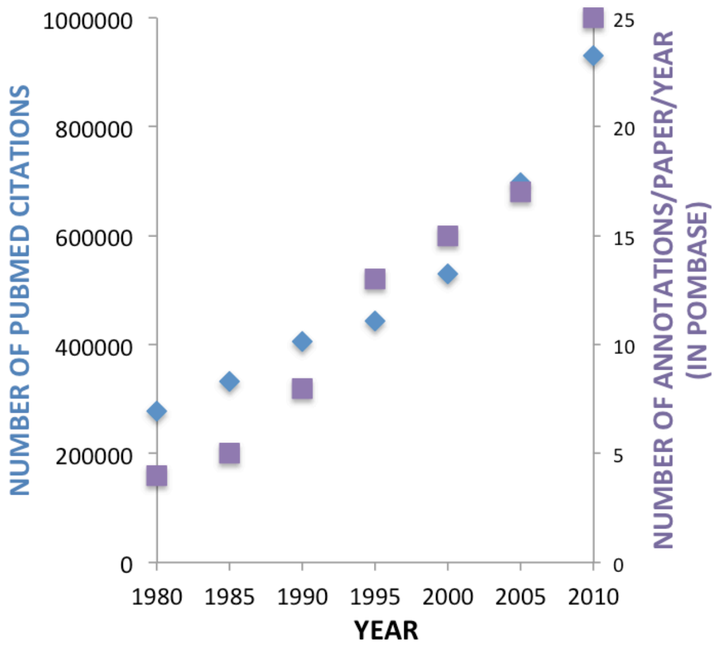
[1]Department of Genetics, Evolution and Environment, and UCL Cancer Institute, University College London, London WC1E 6BT, UK; [2]Department of Biochemistry, University of Cambridge, Sanger Building, 80 Tennis Court Road, Cambridge CB2 1GA, UK

We have developed a web-based annotation tool, Canto, to support community curation on a large scale. Canto is highly configurable, and can be used with minimal or extensive support from professional curators. It is therefore suitable for use by most research communities, including those not supported by a manual curation team, who want to contribute gene-specific experimental information from their organism of interest to public biological databases. Canto supports literature-based curation of a wide, and configurable, set of data types, including Gene Ontology (GO) annotations, phenotypes, interactions, and protein modifications. The tool is fully accessible online, requiring no software download or setup by the end user. Initial feedback from early community users indicates that Canto is easy to use, with an intuitive workflow and integrated help documentation. Canto was originally developed for community curation by the *S. pombe* database (PomBase) and its research community, who curate the most extensive set of data types. To date, Canto has also been adopted by the *K. pastoris* (Pichia) community and for GO annotation workshops at University College London in which researchers and post-graduates are invited to curate their own papers of interest. Ongoing Canto development ensures that feedback from users guides efforts to improve existing features or to implement new ones.

WHY COMMUNITY CURATION?

Research community benefits

- Information embedded in papers becomes easier to find in databases
- Formal syntax and defined language ensures that data from different sources is comparable
- Data integration and organization supports computational analysis of large gene sets to identify patterns



Many papers lack curation

- The number of papers published increases every year
- The amount of curatable information per paper is also increasing
- Database curators cannot keep pace with the resulting data deluge

Author benefits

- Authors learn about formal data representation, enabling them to use curated data more productively
- Participants' data are propagated to other resources, making it more accessible to other researchers which could lead to a higher citation index

CURATING IN CANTO

Getting started

Canto PMID Search

Curation is linked to a publication

Welcome to the PomBase Canto community curation environment. Researchers are welcome to evaluate existing annotations and create new annotations from their group's past publications.

If you wish to curate a paper from your laboratory, please enter the PubMed ID and proceed as directed.

Find ...

ID	PMID:6581157
Title	A meiotic mutant of the fission yeast Schizosaccharomyces pombe that produces mature asci containing two diploid spores.
Authors	Nakaseko Y, Niwa O, Yanagida M

Please annotate only the data that you have determined directly in the experiments described in this paper. If you have any questions, help is available on many pages, or you can contact the PomBase staff for assistance at any time.

Please enter your name and email address:

Name Antonia Lock
Email a.lock@ucl.ac.uk

Curation is linked to a community curator

Continue

Create gene list for PMID:6581157

Please enter the systematic identifier (eg. SPCC1739.10) or the primary identifier (eg. cdc11) of the genes referred to in **PMID:6581157**

The list of gene identifiers can be separated by commas, spaces, tabs or there can be one per line.

The tool is customizable and can accept different stable identifiers (UniProt, NCBI…)

Or: no for annotation in this paper

Continue

Curation workflow

- Customizable curation 'types'
- Lucene searching for terms
- Term definitions provided
- Child terms suggested
- Evidence code selector
- Annotation extensions supported
- Alleles and expression levels specified for phenotypes
- Comments can be added
- New terms can be suggested

Choose curation type for cdc2:

GO molecular function
GO biological process
GO cellular component
Single gene phenotype
Protein modification
Genetic interaction
Physical interaction

Search for GO biological process term

A biological process is series of events accomplished by one or more ordered assemblies of molecular functions. more ...

You may find it helpful to search for a broad term (e.g. cell cycle, transport), especially if you have trouble finding a specific term. more ...

protein phosphorylation

- **protein phosphorylation (GO:0006468)**
- regulation of **protein phosphorylation (GO:0001932)**
- serine **phosphorylation** of STAT **protein (GO:0042501)**
- serine **phosphorylation** of STAT3 **protein (GO:0033136)**

Term name
protein phosphorylation
Definition
The process of introducing a phosphate group on to a protein.

Table of new annotations

New GO molecular function annotations

Systematic identifier	Gene name	Term ID	Term name	Evidence code	With	Comment	Annotation extension
SPBC11B10.09	cdc2	GO:0004674	protein serine/threonine kinase activity	IDA		Add ...	has_substrate(PomBase:SPBC32F12.09)

New GO biological process annotations

Systematic identifier	Gene name	Term ID	Term name	Evidence code	With	Comment	Annotation extension	
SPBC11B10.09	cdc2	GO:0006468	protein phosphorylation	IDA		Add ...	Add extension ...	✗
SPBC11B10.09	cdc2	GO:0008361	regulation of cell size	IMP		Add ...	Add extension ...	✗

New phenotype annotations

Systematic identifier	Gene name	Allele	Term ID	Term name	Evidence code	Expression	Conditions	Comment
SPBC11B10.09	cdc2	cdc2-59(P248L)	FYPO:0000049	inviable	Cell growth assay	Endogenous	glucose minimal medium (PECO:0000126)	Add ...

New protein modification annotations

Systematic identifier	Gene name	Term ID	Term name	Evidence code	Comment	Annotation extension	
SPBC11B10.09	cdc2	MOD:00047	O-phospho-L-threonine	IDA	Add ...	residue=T167	✗

New genetic interaction annotations

Interactor A	Interactor B	Interactor A taxon	Interactor B taxon	Evidence code	
cdc2	pop1	4896	4896	Synthetic Lethality	✗

Customizable export e.g. GAF…

Supported by wellcome trust

CONTACT: HELPDESK@POMBASE.ORG

UNIVERSITY OF CAMBRIDGE
UCL