DATABASE
The Journal of Biological Databases and Curation

Database, Vol. 2013, Article ID bas046, doi:10.1093/database/bas046

Original article

The Xenbase literature curation process

Jeff B. Bowes[1],*, Kevin A. Snyder[1], Christina James-Zorn[2], Virgilio G. Ponferrada[2], Chris J. Jarabek[1], Kevin A. Burns[2], Bishnu Bhattacharyya[1], Aaron M. Zorn[2] and Peter D. Vize[1]

[1]Department of Biological Sciences, University of Calgary, 2500 University Drive NW, Calgary, Alberta, Canada T2N1N4 and [2]Division of Developmental Biology, Cincinnati Children's Research Foundation and University of Cincinnati Department of Pediatrics, College of Medicine, Cincinnati, OH 45229, USA

***Corresponding author:** Tel: +403 220 2824; Fax: +403 282 9154; Email: bowes@ucalgary.ca

Submitted 18 June 2012; Revised 13 September 2012; Accepted 6 October 2012

Xenbase (www.xenbase.org) is the model organism database for *Xenopus tropicalis* and *Xenopus laevis*, two frog species used as model systems for developmental and cell biology. Xenbase curation processes centre on associating papers with genes and extracting gene expression patterns. Papers from PubMed with the keyword '*Xenopus*' are imported into Xenbase and split into two curation tracks. In the first track, papers are automatically associated with genes and anatomy terms, images and captions are semi-automatically imported and gene expression patterns found in those images are manually annotated using controlled vocabularies. In the second track, full text of the same papers are downloaded and indexed by a number of controlled vocabularies and made available to users via the Textpresso search engine and text mining tool.

Introduction

This article describes the curation workflow for Xenbase (www.xenbase.org) (1,2) and was presented as part of the BioCreative workshop 2012 track II. Xenbase is a large-scale model organism database (MOD) and community resource for researchers working with the frog species, *Xenopus tropicalis* and *Xenopus laevis.* It is a central source for genomic, gene expression, literature and other experimental data on *Xenopus* and also acts as a clearinghouse for *Xenopus* gene nomenclature.

Xenopus is a powerful model system for both developmental and cell biology. Developmental biologists use *Xenopus* embryos to study gene function during development and to model gene action in human congenital diseases, while cell biologists use *Xenopus* eggs and oocytes for exploring the mechanistic basis of central processes such as cell division. Xenbase also makes *Xenopus* data accessible to researchers using other model organisms by using gene symbols based on the symbols for human orthologues and by providing links to orthologous genes in humans and similar major vertebrate model organisms (e.g. mouse and zebrafish).

Our curation workflow focuses on associating papers with authors, genes and anatomy terms and extracting gene expression patterns. The *Xenopus* literature corpus contains >42 000 papers with ~1500 new papers added per year. Gene expression data are often presented in papers in the form of image evidence—typically images of embryos stained to display the distribution of an mRNA in the various tissues of the organism. Therefore, the principal Xenbase curation workflow centres on images and image captions.

The Xenbase literature curation workflow is illustrated in Figure 1 and discussed below. (The following 17 steps correspond to the numbered items in Figure 1.)

1. Every week an automatic process searches for new or updated papers in PubMed whose abstract, title or metadata contains the keyword '*Xenopus*'.

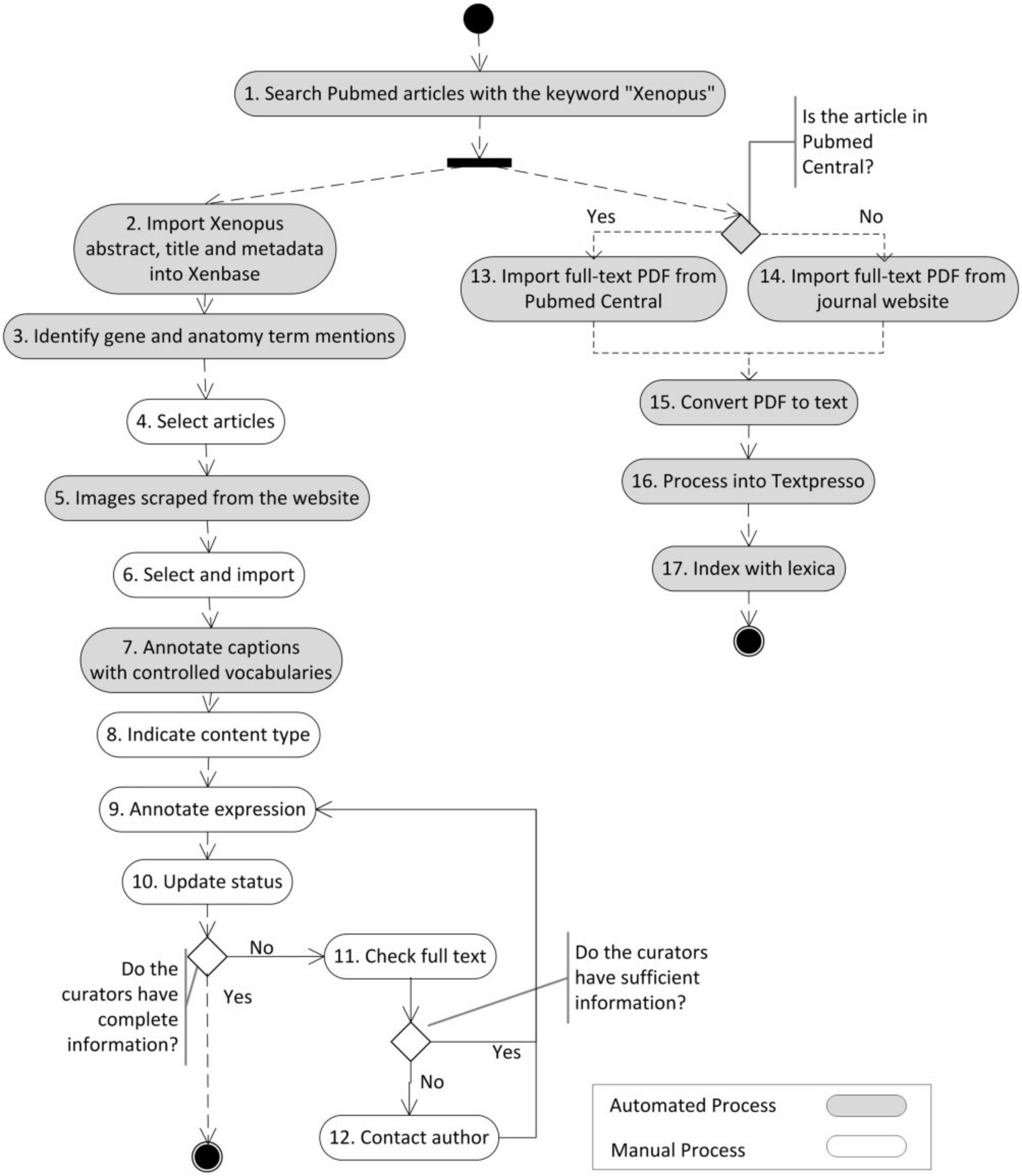
1. Search Pubmed articles with the keyword "Xenopus"
Is the article in Pubmed Central?
Yes
No
2. Import Xenopus abstract, title and metadata into Xenbase
13. Import full-text PDF from Pubmed Central
14. Import full-text PDF from journal website
3. Identify gene and anatomy term mentions
15. Convert PDF to text
4. Select articles
16. Process into Textpresso
5. Images scraped from the website
17. Index with lexica
6. Select and import
7. Annotate captions with controlled vocabularies
8. Indicate content type
9. Annotate expression
10. Update status
No
11. Check full text
Do the curators have sufficient information?
Do the curators have complete information?
Yes
Yes
No
12. Contact author
Automated Process
Manual Process

Figure 1. The Xenbase literature curation process splits into two streams. The first stream consists of a series of manual and automated steps that are used to annotate gene expression patterns in papers. The second stream describes the automated processes of capturing, transforming, inputting, and indexing papers into the Textpresso search engine.

The paper is then imported into two parallel curation processes.

Xenbase curation process

The primary process, focussing on annotation of gene expressions and image data, is as follows:

2. Each paper's title, abstract and metadata are imported into Xenbase's DB2 relational database through which they are made available on Xenbase (www.xenbase.org/literature/literature.do) and to the literature curation pipeline.
3. An automatic process identifies mentions of genes and anatomy terms in the abstract and title based on gene symbols and anatomy terms as well as synonyms (see 'Data and controlled vocabularies' section).
4. Curators select papers for curation based on the presence of gene expression evidence and whether Xenbase may reproduce images either through open-access licensing or special permissions. At this stage, papers irrelevant to current curation processes, such as those using *Xenopus* oocytes for cellular biology experiments, are excluded.
5. Curators initiate a process that automatically scrapes images from journal websites.
6. Curators choose which images have relevant information (e.g. gene expression) and import those images and their corresponding captions.
7. Captions for imported images are automatically annotated with controlled vocabularies to identify genes and anatomy terms.
8. Curators assign content types to both papers and images from the paper. This may be types of content we currently curate (e.g. gene expression) or content types we plan to curate in the future once support is added to Xenbase (e.g. phenotypes, antibodies and morpholinos).
9. The curator annotates image captions for gene expression patterns via a custom web-based curation system (Figure 2). In the future, this curation is likely to be expanded to include additional data types such as phenotypes, gene regulation, transgenics and protein localization.
10. The curator updates the status to indicate whether or not curation is complete.
11. If there are insufficient data in the image captions to complete the annotation, the curator will proceed to the full paper text to obtain the required information (see 'Curation bottlenecks' section).
12. If the curator is still missing information, they may contact the author (see 'Curation bottlenecks' section). For example, curators often need to contact authors to determine which cDNA clone and, in the case of *X. laevis*, which alloallele was used in the experiment.

Textpresso process

In the second process, the full texts of papers are imported into Xenbase's Textpresso (3) text mining and search tool (see 'Use of text mining tools' section). Textpresso is currently independent of our curation process but we hope to integrate it into a semi-automated full-text curation process in the future (see 'Use of text mining tools' section).

13. If the paper is in PubMed Central, its full text is downloaded from PubMed central.
14. If the paper is not in PubMed Central, the PDF for the journal is imported from the journal website.
15. PDF documents are converted to xml using the open-source pdftohtml utility (pdftohtml.sourceforge.net).
16. Documents are processed into Textpresso (see 'Use of text mining tools' section).
17. Documents are indexed using Textpresso-controlled vocabularies.

To our knowledge, the Xenbase curation process, among MODs, is novel in a number of respects. First, it uses automatic processes to associate literature to genes and anatomy terms. Second, using html scrapers, we have semi-automated the task of extracting images from papers and uploading them to Xenbase. Third, the Xenbase curation process is centred around images.

These characteristics are deliberately designed to maximize the efficiency for human curators. Xenbase started with very limited resources for human curators. While more established MODs had teams of curators read papers and annotate them with genes and anatomy items described therein, Xenbase made use automated text mining methods for matching gene and anatomy term mentions in paper.

Evidence of gene expression data from *in situ* hybridization and immunohistochemistry experiments is most often presented as images. Hence, it was desirable to include these images drawn from literature into Xenbase. To eliminate the repetitive task of having curators extract images from papers and import them into Xenbase, we developed tools to scrape journal websites for images and upload those chosen by the curators into Xenbase. This resulted in significant time savings for curators.

The more difficult work of extracting gene expression patterns from texts has been left to human curators. However, even this process was designed to maximize curation efficiency. Standard MOD operating procedure for curators to extract facts, such as gene expression patterns, from papers would be to have a curator read the paper and extract those facts. Realizing the evidence for gene expression is almost always illustrated in images and described in

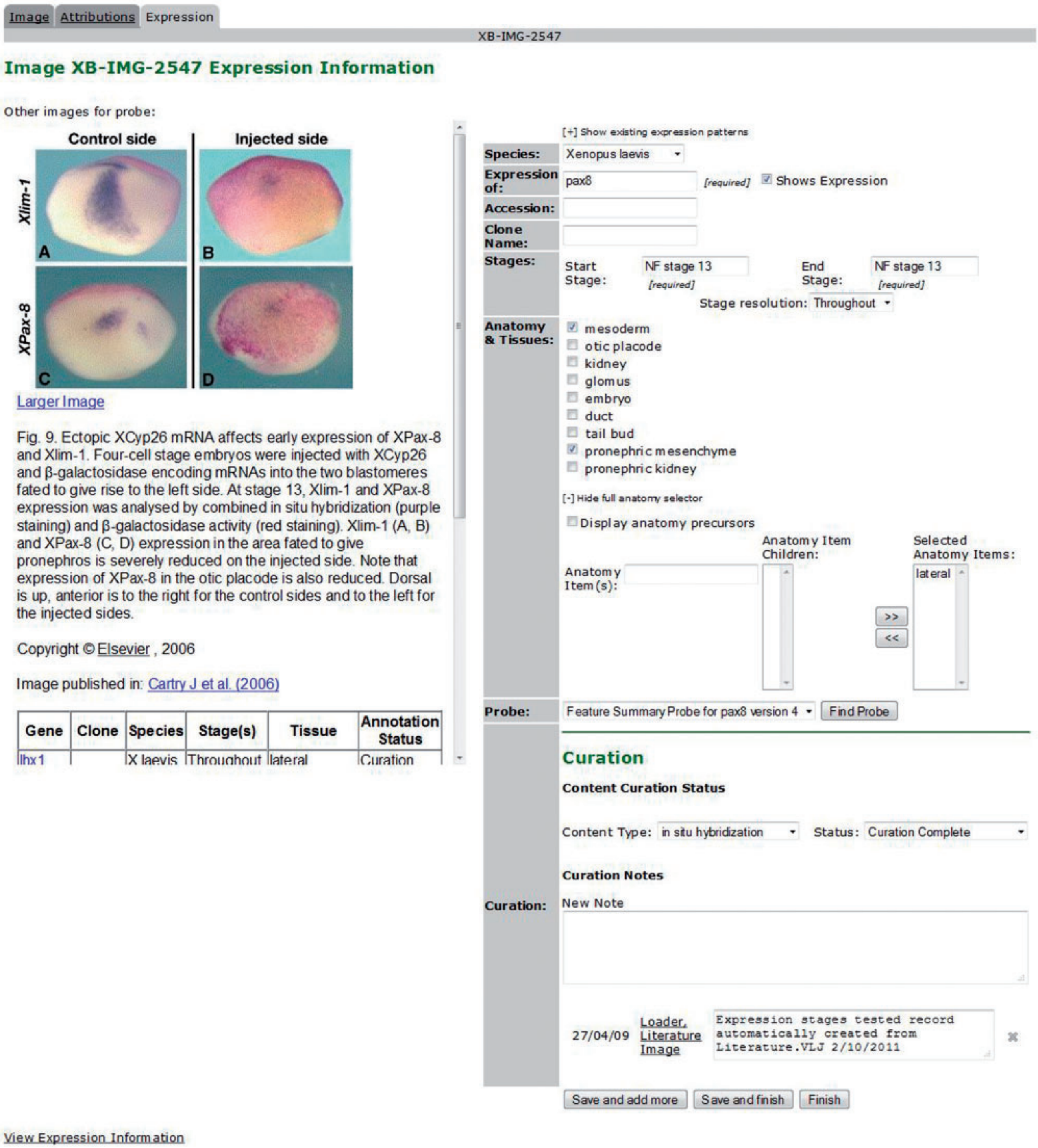

Image XB-IMG-2547 Expression Information

Other images for probe:

Larger Image

Fig. 9. Ectopic XCyp26 mRNA affects early expression of XPax-8 and Xlim-1. Four-cell stage embryos were injected with XCyp26 and β-galactosidase encoding mRNAs into the two blastomeres fated to give rise to the left side. At stage 13, Xlim-1 and XPax-8 expression was analysed by combined in situ hybridization (purple staining) and β-galactosidase activity (red staining). Xlim-1 (A, B) and XPax-8 (C, D) expression in the area fated to give pronephros is severely reduced on the injected side. Note that expression of XPax-8 in the otic placode is also reduced. Dorsal is up, anterior is to the right for the control sides and to the left for the injected sides.

Copyright © Elsevier , 2006

Image published in: Cartry J et al. (2006)

Gene	Clone	Species	Stage(s)	Tissue	Annotation Status
lhx1		X laevis	Throughout	lateral	Curation

[+] Show existing expression patterns

Species:	Xenopus laevis
Expression of:	pax8 *[required]* ☑ Shows Expression
Accession:	
Clone Name:	
Stages:	Start Stage: NF stage 13 *[required]* End Stage: NF stage 13 *[required]* Stage resolution: Throughout
Anatomy & Tissues:	☑ mesoderm □ otic placode □ kidney □ glomus □ embryo □ duct □ tail bud ☑ pronephric mesenchyme □ pronephric kidney [-] Hide full anatomy selector □ Display anatomy precursors Anatomy Item(s): Anatomy Item Children: Selected Anatomy Items: lateral
Probe:	Feature Summary Probe for pax8 version 4
Curation:	Curation Content Curation Status Content Type: in situ hybridization Status: Curation Complete Curation Notes New Note 27/04/09 Loader, Literature Image Expression stages tested record automatically created from Literature.VLJ 2/10/2011

View Expression Information

Figure 2. The Xenbase image curation interface consists of two panels. The left panel shows the image, caption and a tables of existing annotations made to that image. The right side of the panel contains a form used for editing existing or new annotations. It starts with a series of fields to specify species, gene name, clone names or accessions. Next the curator can specify a range of development stages when a particular expression pattern occurs. Finally, anatomy terms describing where expression occurs can be chosen from checkbox lists of commonly uses anatomy terms or a suggestion box. Anatomy terms are restricted to those that exist during the stage range previously entered by the curator. Finally, there is a section for updating the curation status of the item and entering curation notes.

those image captions, we created a curation process centred on images and their captions. Xenbase, curators first examine images for gene expression information. They only proceed to read the full paper if important information is missing from the captions. Even when information is missing from captions, it can often be searched from the paper without reading the entire paper.

In conclusion, this process has been very successful at leveraging the use of human curator time. The association of papers with genes has been acceptably accurate. Finally, the image-centric curation process can be expanded to other forms of data that commonly used image-based evidence such as phenotypes.

While it has been largely successful, there are a number of problems. Early attempts to make a 'first pass' extraction of gene expression facts based on cooccurrences of anatomy terms and genes in captions proved too inaccurate. Image scrapers save curator time but frequently need to be fixed when publishers change the structure of their journal websites. The image-centric curation process is not really applicable to problems such as identifying antibody, morpholinos or gene ontology terms. Thus, Xenbase is exploring the idea of creating semi-automated curation processes of the full text of papers, using Textpresso.

Data and controlled vocabularies

Xenbase uses various different controlled vocabularies to capture the 'what', 'where' and 'when' of gene expression. Further metadata on these experiments such as specific clone or antibody used are also captured.

A controlled vocabulary of gene symbols and synonyms drawn from the Xenbase gene catalogue describe what is expressed. Xenbase gene symbols are used by NCBI Gene for *Xenopus* gene naming and are based on the names of human orthologues. Synonyms represent historical or alternative identifiers for genes. This is especially important as the same gene is often referred to by more than one name in archival literature. Xenbase allows any user to enter synonyms for genes, helping ensure synonym lists are comprehensive and up to date. Because *X. laevis* is pseudotetraploid, having two versions of each gene (alloallele-a and alloallele-b), we also must consider which version of each gene was expressed.

Anatomy items (i.e. tissues, organs and cell types), 'where' genes are expressed, are represented using the Xenopus Anatomy Ontology (XAO) (4). The XAO is structured as a directed acyclic graph {DAG [A DAG is a directional data structure similar to a tree or hierarchy except that nodes may have more than one parent and cycles are not allowed. For example, in the XAO, a DAG represents the part-of relationship between anatomy items. For example, the brain is part of the nervous system and the head (two) parents. The brain has the parts (children) hindbrain and forebrain.]} and the XAO is constructed using best practices outlined by the OBO foundry (The OBO foundry is a repository of biomedical ontologies and a source of best practices in constructing those ontologies.) (5) allowing it to be interrelated to anatomy ontologies from other model organisms. Nieuwkoop and Faber (NF) development stages, defined in (6), have long been the accepted standard for delineating periods of development in *Xenopus* and are used to describe when expression occurs. NF stages are also included in the XAO.

Additional metadata on experiments such as clones or antibodies used to test for gene expression and the species used in the experiment are also captured. cDNA clones are identified via references to NCBI accession numbers. Various data on antibodies used in immunohistochemistry experiments are captured from the literature by curators. Finally, the species, *laevis* or *tropicalis*, used in the experiment must be determined.

Additionally, Xenbase uses a small ontology structured as a DAG to describe different content types of papers. These include antibodies, *in situ* hybridizations and phenotypes. Finally, Textpresso uses numerous controlled vocabularies. Some examples are gene regulation relationship, and morpholino and antibody terms. These are all structured as single-level-controlled vocabularies with synonyms, described in (3).

To get a better feel for the data elements captured in Xenbase, we recommend the reader to view some sample pages in Xenbase: www.xenbase.org/literature/article.do?method=display&articleId=39749 provides an example of a curated article. Note that names of genes and anatomy items are hyperlinked in the abstract and image captions. (To view captions click 'show captions'.) Clicking on an image will open a box showing the image, caption and a table of curator-entered annotations. www.xenbase.org/gene/expression.do?method=displayGenePageExpression&geneId=484814&tabId=1 is the expression tab of the *sox3* gene page. Click an image under the 'Summary' or 'Literature Images' section to see sample annotations. Finally, the XAO can be downloaded from obofoundry.org/cgi-bin/detail.cgi?id=xenopus_anatomy.

Curation bottlenecks

Xenbase curators experience a number of obstacles such as missing information and anatomy terms that are not in the XAO.

Curators often find pertinent information on the experiment missing. For example, the authors may not have specified which species of *Xenopus* was used, which *X. laevis* gene alloallele was tested or the Genbank accession number of the clone used in the experiment. In the case of antibodies, there may be information missing on its construction. This requires first searching the entire text of the paper, literature search of secondary references describing the missing data and then, if necessary, contacting the author.

Another problem is that the controlled vocabularies Xenbase uses are not yet comprehensive enough to capture all the terminology used in literature. This is especially the case with the XAO. There are cases where the author has described expression in a tissue that is not currently represented in the XAO. This requires the curator to research the tissue, determine whether the term is a synonym of an existing XAO term or whether it is a valid term missing from the XAO. Xenbase curators are continually expanding the XAO and keeping it synchronized with other model organism ontologies. Curators will define new terms and integrate them into the XAO in relation to existing terms. This may involve examining anatomy ontologies for other vertebrates such as mouse or zebrafish or through consultation with other ontology development teams such as National Center for Biomedical Ontology (www.bioontology.org), Uberon (www.uberon.org), the common amphibian anatomical ontology (www.amphibanat.org) or Phenoscape (www.phenoscape.org).

Use of text mining tools

Xenbase currently successfully makes use of its built-in link matching tool for text mining. We are also starting to make used of Textpresso for text mining. However, there are many more potential text mining applications in Xenbase.

Our link matching tool identifies gene and anatomy term mentions in titles, abstracts and captions. This uses a combination of inverted indices and regular expressions to match gene symbols, anatomy terms and their synonyms. In the case of genes, synonyms may be added by any user. Terms with common homonyms can be entered into a table of exclusions that are ignored by the matching process, to reduce false positives. Identified genes or anatomy terms are hyperlinked to gene and anatomy pages, respectively. This tool comes into play at both Steps 3 and 7 in our curation workflow diagram. Although this is a fairly basic approach, the identification of synonyms is a particularly important step allowing this tool to associate genes and anatomy terms from different papers, despite the plethora of alternative gene names and variant anatomical descriptions used in the scientific literature.

We have implemented Textpresso, a biological text search and mining tool. Textpresso is used to index the full text of documents, and in particular, index the corpus by controlled vocabularies. Textpresso also segments the paper into sections such as abstract, body, discussion, materials, results and citations. Currently, Textpresso is used as an advanced query tool (www.xenbase.org/cgi-bin/textpresso/xenopus/home) that allows users to return documents or sentences matching particular criteria. Users can pose questions such as 'return sentences with two genes and a regulation term' [by selecting three categories: 'gene (Xenopus)' twice and 'regulation' or 'return sentences containing a gene mention and a morpholino from the materials section' [by turning advanced search on, unselecting all fields but materials and entering the categories: 'gene (Xenopus)' and 'morpholino'.

In the near future, we plan to expand Textpresso's application to identify papers that contain information on antibodies and/or morpholinos for particular genes. This information will be presented to users on Xenbase gene pages.

We have had success with text mining; we believe its potential to improve curation has barely been tapped. There are many places in the curation process where text mining could be applied to further improve our curation workflow.

At Steps 3 and 7, our current technology does an effective job of finding gene and anatomy term mentions, especially by taking advantage of synonyms. However, we are aware that more effective text mining methods for capturing this information exist. Furthermore, because of numerous ways used to describe a range of development stages (e.g. stages 1, 2, 3 and 4; 1–5; St. 1 to 5), our current methods do not capture this information. Furthermore, it would be useful to capture other entities such as Gene Ontology terms or NCBI accession numbers.

At Step 4 of the process, classifiers that could identify the content of papers using our content type ontology could help in triaging papers for curation.

Much more ambitiously, at Step 7, actual gene expression relationships could potentially be captured from captions. If false positives were kept to a reasonable level, extracted relationships could be approved or edited by curators, increasing the efficiency of the process. This would be a difficult problem to solve. If this was made possible, extracting other relations such as gene regulations (i.e. gene *a* regulates gene *b)* or phenotypes (e.g. knocking out gene *a* results in phenotypes 1, 2 and 3) would also be valuable.

At Step 16, Textpresso attempts to segment papers into sections. However, Textpresso does this poorly because it is difficult to automatically recognize the many different ways to title and delineate sections of a paper. It would be very useful to segment papers well. For example, we have found that associating papers with gene mentions produces many false positives as paper titles in the paper's references mention genes that are unrelated to the work described in the paper. Being able to properly distinguish between the references section and other sections of a paper would allow us to more accurately associate papers with genes by excluding genes referenced in the citations section. Extending the markup of papers to identify paragraphs and part-of-speech tagging would also be extremely useful.

As current curation in Xenbase is very image and caption centric, it may miss information found only in the full body text of papers. Beyond gene expression, other types of curatable information may not be presented via images (e.g. gene regulation). While limited curation resources still preclude examining every paper, we have considered developing a pipeline that would use Textpresso to extract sentences from papers that may contain useful biological information (gene expression, antibodies, phenotypes, morpholinos, etc.). This could also be implemented with other text mining tools. The interface would be designed to allow curators to approve, reject or edit extracted facts and zoom out from sentences to surrounding text to further assess facts in the data.

Software and tools

Java libraries for constructing image scrapers are available upon request. Most other code involved in this application is heavily embedded in the Xenbase application.

Funding

The Eunice Kennedy Shriver National Institute of Child Health and Human Development (NICHD) at the National Institutes of Health (5P41HD064556) and an NSERC Discovery Grant. Open access publishing is funded by the NICHD grant (5P41HD064556).

Conflict of interest. None declared.

References

1. Bowes,J.B., Snyder,K.A., Segerdell,E. *et al*. (2008) Xenbase: a Xenopus biology and genomics resource. Nucleic Acids Res., **36**(Database issue), D761–D767.
2. Bowes,J.B., Snyder,K.A., Segerdell,E. *et al*. (2010) Xenbase: gene expression and improved integration. *Nucleic Acids Res.*, **38**(Database issue), D607–D612.
3. Muller,H.M., Kenny,E.E. and Sternberg,P.W. (2004) Textpresso: an ontology-based information retrieval and extraction system for biological literature. *PLoS Biol.*, **2**, e309.
4. Segerdell,E., Bowes,J.B., Pollet,N. *et al*. (2008) An ontology for Xenopus anatomy and development. *BMC Dev. Biol.*, **8**, 92.
5. Smith,B., Ceusters,W., Klagges,B. *et al*. (2005) Relations in biomedical ontologies. *Genome Biol.*, **6**, R46.
6. Nieuwkoop,P.D. and Faber,J. (1994) *Normal Table of Xenopus laevis (Daudin)*. Garland, New York.