Current Plant Biology xxx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Current Plant Biology

journal homepage: www.elsevier.com/locate/cpb

# The art of curation at a biological database: Principles and application☆

Sarah G. Odell[a,b], Gerard R. Lazo[a], Margaret R. Woodhouse[a], David L. Hane[a], Taner Z. Sen[a,c,*]

[a] U.S. Department of Agriculture – Agricultural Research Service, Crop Improvement and Genetics Research Unit, Albany, CA 94710, United States
[b] University of California, Department of Plant Sciences, Davis, CA 95616, United States
[c] Iowa State University, Department of Genetics, Development, and Cell Biology, Ames, IA 50011, United States

ARTICLE INFO

*Keywords:*
Biological databases
Curation
Genetic markers
Genetic maps
Genomic data
Genome browsers

ABSTRACT

The variety and quantity of data being produced by biological research has grown dramatically in recent years, resulting in an expansion of our understanding of biological systems. However, this abundance of data has brought new challenges, especially in curation. The role of biocurators is in part to filter research outcomes as they are generated, not only so that information is formatted and consolidated into locations that can provide long-term data sustainability, but also to ensure that the relevant data that was captured is reliable, reusable, and accessible. In many ways, biocuration lies somewhere between an art and a science. At GrainGenes (https://wheat.pw.usda.gov;https://graingenes.org), a long-time, stably-funded centralized repository for data about wheat, barley, rye, oat, and other small grains, curators have implemented a workflow for locating, parsing, and uploading new data so that the most important, peer-reviewed, high-quality research is available to users as quickly as possible with rich links to past research outcomes. In this report, we illustrate the principles and practical considerations of curation that we follow at GrainGenes with three case studies for curating a gene, a quantitative trait locus (QTL), and genomic elements. These examples demonstrate how our work allows users, i.e., small grains geneticists and breeders, to harness high-quality small grains data at GrainGenes to help them develop plants with enhanced agronomic traits.

## 1. Introduction

The value of a biological database is largely defined by the breadth and accuracy of its content. If the content is becoming limited and inaccurate, a database would steadily lose its value for its users, and will eventually become obsolete. The data coverage and accuracy of a database need to be therefore continuously enhanced, and a primary way of accomplishing this goal is through a critical process called biological curation, i.e., extracting biological data from scientific literature and integrating it into a biological database. Curators, usually combining computational skills with PhD-level biological expertise, peruse peer-reviewed scientific articles, extract data sets that they judge to be the most useful for their user base, and integrate them into a back-end database, so that these data sets can be displayed through a web interface. Because curators apply a set of subjective criteria and the extracted data sets need to be integrated into specific databases with different contexts and focus, the curated content from the same journal article can sometimes be curated slightly differently at different biological repositories (even for plant databases with similar user bases, such as GrainGenes (https://wheat.pw.usda.gov; https://graingenes.org) [1], TAIR [2], MaizeGDB [3], Gramene [4], Sol Genomics Network [5], and Soybase [6]. Yet, curators follow similar routes, workflows, and principles in curating biological data. Here, we provide general curatorial principles followed at GrainGenes, along with two examples of how curation is performed in practice.

GrainGenes [1] has a long history of serving the small grains communities via curation and many other activities. The repository was established in 1992 as a central data repository focused on Triticeae and Avena species, and has been continuously supported by the U.S. Department of Agriculture, Agricultural Research Service since then as a service to geneticists and breeders of wheat, barley, rye and oat worldwide. At times, the database resource has also facilitated research progress by hosting emerging projects such as EST sequencing, mapping, genome sequencing, and tools such as scripts for generating wheat genome-specific SNPs [7]. The database contains a wide variety of data types, including genome sequences, genetic maps, genes, alleles, molecular markers, phenotypes, QTLs, experimental protocols, and publications. In addition, GrainGenes serves the small grains communities by hosting small grains community newsletters such as the Annual Wheat Newsletter and Barley Genetics Newsletter, and community

☆ This article is part of a special issue entitled "Genomic resources".

* Corresponding author at: 800 Buchanan St. Albany, CA 94710, United States.
*E-mail address:* taner.sen@ars.usda.gov (T.Z. Sen).

https://doi.org/10.1016/j.cpb.2017.11.001
Received 26 September 2017; Received in revised form 28 November 2017; Accepted 29 November 2017

sites, such as the Triticeae Toolbox (T3) [8] repository. In addition, job openings, news updates, and links to other sites of interest are provided. A wide range of tools can be accessed by small grains researchers who use the website, including Generic Model Organism Database (GMOD) data visualization tools such as the CMap genetic map viewer [9] and the JBrowse genome browser [10] that visualizes genetic features along a reference sequence.

### 1.1. The age of big data

The effectiveness of data curation depends on the initial triage of papers, choosing the ones with the most impactful research outcomes. The amount and the heterogeneity of data that are included in the papers influence triage decisions. A curator needs to consider how data sets will be entered into a back-end database and displayed through the web interface. So-called “Big Data” has made these triage decisions more important than ever [11]. What do we mean by big data? Big data is hard to define, but for most in biological fields, it means data sets from megabytes to terabytes in size with a wide variety of data types. Big data is a direct result of technological innovations. Within the last few decades, rapid advancements in high-throughput technology and high-performance computing have resulted in an explosion of biological data production, both experimental and predictive. As a result, we now have access to multiple high-quality genome assemblies, transcriptomes, proteomes, and genome-wide association studies (GWAS). The resolution and accuracy of these data sets are usually high and they have opened up new possibilities for research and scientific discovery. However, our current data infrastructure, analysis methods, and visualization capabilities are being continuously challenged. Against this data deluge, indexing and standardization are becoming more crucial for ensuring that data are available for knowledge extraction, and research communities are getting together to create guiding principles, such as FAIR, i.e., findability, accessibility, interoperability, and reusability [12]. Grassroots organizations, such as AgBioData (https://www.agbiodata.org) have formed to help standardize data representation across groups and to make recommendations for responsible data sharing and management in developing data and metadata standards in the form of templates that would facilitate scaling of curation. At GrainGenes we developed data templates for researchers to help them upload their data into GrainGenes (some GrainGenes templates with metadata fields and example data entries can be found here: https://wheat.pw.usda.gov/GG3/submit), but better standards across communities are needed. The age of big data is only starting, and big data will definitely present more opportunities and challenges for biocuration in the future.

There has been an increasing effort to use standardized ontologies for labeling of genetic data. Standardization of data labels, in an object-oriented sense, has helped to build data-connections within and between databases as resources have grown over time. The aim of the Gene Ontology (GO) Consortium is to create unity in the description of gene terminology. Out of this community, the Trait Ontology (TO) and Plant Ontology (PO) have also blossomed. The evolution of terminology into ontology terms has enabled some types of classification and curation to be automated, easing the workload of curators. However, by no means does this automation make manual curation obsolete. Rather, high-throughput automation complements manual curation by allowing curators to focus more on the tasks of curation that require a human mind – those that call for critical thinking, investigation, and creativity.

### 1.2. Why curate?

To researchers who would like to have access to most recent, high-quality data in their field, the importance of curation is obvious. But, unfortunately, curation is not always seen as a critical part of scientific work. Here we want to emphasize the importance of curation for the advancement of science.

If new data sets are not curated into databases for long-term sustainability and integrated with pre-existing data, they may lose their accessibility and utility over time. If new, important data sets are not used, knowledge production and discovery rates will lag behind data production rates. In other words, data must be captured, standardized, organized, and made accessible to the scientific community if it is going to have a significant and lasting impact. In addition, a database is only as good as its data. If members of the scientific community do not find the data in their popular databases up-to-date, accurate, or transferable, then the database is of little use and will be obsolete soon. Likewise, if an online database’s interface is not intuitive, few researchers will utilize the database. The role of a biocurator is therefore to provide up-to-date, accurate, and accessible information, and, through this critical activity, facilitate scientific discovery.

## 2. Curation workflow at GrainGenes

Although each publication that is curated into GrainGenes might use distinct data types, the general protocol for manual curation is the same (Fig. 1). By following an established procedure, data can be formatted in a manner that is compatible with work done by past and different curators, assuring that as much meaningful information as possible is stored in the database.

What differentiates community databases like GrainGenes from primary data repositories such as NCBI and EMBL is that the content of community databases is geared toward a particular organism or a set of closely-related organisms to cater to the needs of researchers in that particular community. The manual curation required to maintain a community database is time-intensive, and making incremental updates are an ongoing challenge, but it ultimately results in an indispensable resource.

### 2.1. Identification of peer-reviewed journal articles for curation

Every curator has limited time for curation, and therefore the first and most important step in the curation workflow at GrainGenes is to identify the peer-reviewed, high-impact journal articles that would most enrich the database and make the database most useful to small grains researchers. The identification step is primarily done by monitoring the release of publications in scientific journals relevant to small grains research. PubMed, Google Scholar and Scopus are sites where a wide range of journal articles is regularly updated, but these listings are not all-inclusive or strictly plant-focused. Our experienced curators are familiar with the journals most likely to contain articles of interest to our users and devote special attention to each new issue of them. Citation indices also help identify research with impact. We do not however use a specific list of journals or quality-metrics to identify articles.

Although web search tools are very useful, our experience shows that personal interactions are the best way of being informed of high-impact articles (Fig. 1). There are two beneficial ways of interacting with researchers to learn of new advances and therefore new publications coming our way. First, by attending conferences and listening to presentations, our curators are apprised of cutting-edge research that has been published or is about to be published. Second, when curators establish personal relationships with small grains researchers, then the researchers are more likely to contact the curators when they have new data sets. Some journals have actually made data submission mandatory for article authors in an effort to promote open access to research data, and GrainGenes is among the biological databases that greatly benefit from this requirement.

### 2.2. Curation prioritization

Selected papers then go through the triage stage, where the priority