# Carpe Diem. Retooling the "Publish or Perish" Model into the "Share and Survive" Model[1]

**Seung Yon Rhee***

Carnegie Institution, Department of Plant Biology, 260 Panama Street, Stanford, California 94305

Recent advances in high-throughput technologies such as genome sequencing and genome-wide gene expression profiling have energized and created a groundswell of interest among academic scientists, funding agencies, biotech companies, and even the general public. The development of these technologies in the last 20 years has culminated in drastic changes in how we conduct research and how we share the outcomes, as poignantly described in articles of this Editor's Choice series. The exponential growth of data that can be used to accelerate our understanding of biology provides a tantalizing hint that we may be on the verge of a scientific revolution. Individual researchers are digesting more information and expanding from their own "domain of expertise," fueled by access to the large body of information generated by other researchers and facilitated by advances in communications technologies such as the Internet, e-mail, and open community databases. Until recently, it was not unusual for an entire research group to focus on the characterization of just one gene, and many successful careers were built on that model. Now, it is more typical for graduate students and postdoctoral researchers to characterize a family of genes simultaneously. Even molecular genetics has expanded scope from the study of monogenic traits to the characterization of complex, quantitative traits and the eventual cloning of the genes responsible for these traits.

There is a dark side to this picture. Recent trends in public funding and in the biotech private sector in the last few years suggest that the enthusiasm for large-scale projects to make genomic resources available to the community may be winding down (http://www.lifesciencesnetwork.com/news-detail.asp?newsID=2156; Duyk, 2002; Reid, 2002; Brower, 2003; Lahteenmaki and DeFrancesco, 2003). This is possibly because of the lack of an immediate connection between generation of data and its transformation into new biological concepts and paradigms by the whole research community. We must come together as a community to leverage this opportunity and enable maximal use of the wealth of data by all to advance our understanding of plant biology. So, how can the plant research community carpe diem?

To transform the vast amount of data into knowledge efficiently, we need to connect several tasks seamlessly: (a) data generation, (b) data annotation, (c) information integration into existing databases, and (d) data presentation in an intuitive, organized layout. This organized information should be easily accessible and available without restriction. Only then can we utilize the information to make informed decisions about research directions, hypothesis building, and testing. There are currently three widely used models of information dissemination that will play key roles in this process: (a) On-line community database systems capture and present annotated information (Baxevanis, 2003). (b) Public repositories archive raw data permanently and make them accessible to the public (e.g. GenBank, http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=Nucleotide). 3) Scientific journals have been the traditional medium of knowledge dissemination. Recently, publishers have moved to electronic format, and are increasingly under pressure to consider free access (Public Library of Science, http://www.plosbiology.org/plosonline/?request=index-html; and BioMed Central, http://www.biomedcentral.com).

What is needed is a seamless connection of community databases, public repositories, and journals. This should facilitate free access to raw data (public repositories), aggregations of interconnected and annotated information (community databases), and thorough analyses and interpretation of experimental data (journal articles). Some problems need to be resolved before we can achieve this ideal state of information flow. For example, public repositories do not yet exist for all types of raw data. As a consequence, community databases often take on this role in addition to their primary role of providing value-added annotation and display of the information. This effectively diverts the efforts of community databases because limited resources are directed toward managing the avalanche of raw data. Also, journals are now facing the fact that results of microarray and proteomics experiments do not fit into publishable article pages, similar to the situation regarding publishing of sequence in articles some 15 years ago. Often, these results are archived on individual journals' Web sites and are not well connected to community resources. Community databases may

[1] This work was supported in part by the National Science Foundation (grant no. DBI–9978564). This is Carnegie Publication no. 1634.

* E-mail rhee@acoma.stanford.edu; fax 650–325–6857.

http://www.plantphysiol.org/cgi/doi/10.1104/pp.103.035907.

be an ideal place to contain these supplemental results. However, researchers are not accustomed to contributing their data and expertise to community databases. Here lies a conflict: Although there is a well-established reward structure for publishing in scientific journals and public repositories (largely through enforcement either as a condition for publication or for receiving grants), a similar reward system does not yet exist for contributions to community databases. Here, I describe several ways to resolve this conundrum of the need for high-quality, integrated information and the gaps in the existing infrastructure.

## COMMUNITY DATABASE SYSTEMS

A community database system is defined as "an information resource that is created, maintained, or improved by a geographically distributed community" (http://www.scienceofcollaboratories.org/Resources/Community_Data_Systems.pdf). In plant biology, prototypical examples of community database systems are model organism databases such as The Arabidopsis Information Resource (TAIR; http://arabidopsis.org; Rhee et al., 2003) and MaizeGDB (http://www.maizegdb.org/). These databases contain information about the research community, bibliographies, genetic and genomic resources such as structural and functional annotations of genes, and genetic and physical maps of the genome. The resources typically have an in-house team of biologists (curators) who gather the information. This is similar to the initial business model of GenBank, where in-house curators copied sequences from published articles and entered them into the database (Burks et al., 1987). The community databases provide additional value by synthesizing and extracting annotations derived from experimental data in published articles. They also systematically analyze data; for example, by applying a series of quality control filters, normalizing, and clustering all available microarray data using a standard method to enable subsequent data mining by members of the community who otherwise might not have the tools to perform such analysis (http://arabidopsis.org/tools/bulk/microarray/analysis/index.jsp). In cases where there is not yet an obvious general public repository for raw data, community databases also fulfill that role.

In a sense, these community database systems are a prototype of the next generation of long-lived databases currently typified by GenBank, upon which future researchers will depend. However, as the amount of information increases, these resources may not be sustainable in the long term with the current model of mostly in-house curation. There are several ways to solve this issue. The most appealing is to engage the research community as active participants in development and maintenance of the databases. For example, TAIR currently has over 12,500 registered users in 4,500 laboratories around the world. If each laboratory was responsible for five to 10 genes, we would be ensured that the information about all 30,000 genes in the genome is kept up-to-date. If this concept were to work, it would be a monumental achievement that would serve as a model for other research communities. There have been some attempts to achieve these goals already (e.g. Genome Database, *Saccharomyces* Genome Database, TAIR, FlyBase, and MaizeGDB) but with limited success. Systematic analysis of what factors limited the success have not been conducted. However, a recent survey suggests that a lack of incentives for the user community to contribute to their databases might be the main cause of the lack of input (http://www.scienceofcollaboratories.org/Resources/Community_Data_Systems.pdf).

## THE RESEARCH COMMUNITY'S ROLES

Despite their definition, many community database systems primarily rely on "in-house" curation to capture information and keep it up-to-date (http://www.scienceofcollaboratories.org/Resources/Community_Data_Systems.pdf). For long-term success of community databases, contributions from the community are essential. However, this concept is new to most researchers and is not well embraced. For example, TAIR is one of the most widely used community databases, with a monthly average of 340,000 page views accessed by 16,000 unique IP addresses (http://www.scienceofcollaboratories.org/Resources/Community_Data_Systems.pdf; TAIR usage, http://arabidopsis.org/usage/). Since November 2002, TAIR has made available a function that allows any registered researchers to add their comments to any data detail page. In the 1st year since this "Add My Comments" function has been available, only 50 comments were added to 46 data pages (from more than 1.5 million available data pages generated from the database). The rate of increase in user-submitted comments would need to be enormous for their input to make a significant contribution. To improve the situation, it will require changes in attitude from a variety of people ranging from administrators at universities to research scientists and scientists developing the databases.

Members of individual research laboratories have useful information that may never be accessible to others. This includes protocols, genetic markers, sequences, microarray data, genetic mapping, and phenotypic characterization of mutants that do not end up in publications, either because of the page limitations or the tendency of journals for not publishing negative results. Because the current academic reward system emphasizes publication in journals with high citation indices, many scientists and students do not consider active participation in databases as part of their responsibility as a scientist. As a conse-

quence, information that does not end up in publications is not available to other researchers. Ironically, the public repositories and databases are far more widely accessible than any journal publications today. Therefore, contributing useful data and information to the databases and their subsequent use by others should have at least as much impact toward advancing science as publishing in journals. Administrators of academic institutions should recognize this activity as a crucial part of being a scientist and as an attribute of leadership.

The most effective way to increase contributions to community databases is for scientists to submit data and information that would benefit other researchers. For example, all sequence data, including transcript sequence from PCR reactions or EST clones that can be used to validate a gene structure, should be submitted to GenBank, where it can be accessed by and integrated into community databases. All other useful data and information particular to an organism should be sent to the specific organism database. The demand to make data accessible to the public will drive the development and enhancement of the databases to accommodate this need.

However, what if no format exists for submission of a specific type of data? In this case, researchers should contact the databases and request the feature, which most databases will be happy to provide. The development of new and improved functionalities should be inspired by the needs of the community as they are communicated to the databases. Finally, there are many ways for users to contribute to a community database beyond the crucial role of submitting high-quality data. For example, researchers can notify the database curators of data errors such as out-of-date information or incorrect annotations. In addition, if users do not find information they are looking for, they should notify the database, rather than assume that the information is not there. Similarly, users can report software problems, such as performance issues and bugs that database software developers need to resolve.

## THE ROLES OF COMMUNITY DATABASE SYSTEMS

If we can mobilize the research community to actively contribute data to community databases, we will have to resolve engineering and technical issues to ensure that the contributions are represented consistently and accurately. First, databases need to clearly communicate what data are suitable for contribution, and in what format they should be submitted. A few plant community databases actively encourage information submission and seek to provide guidance for users (TAIR, http://Arabidopsis.org/info/data_submission.jsp; Gramene, http://www.gramene.org/submission.html; and GrainGenes, http://wheat.pw.usda.gov/ggpages/forms/index.shtml). Also, attributions and acknowledgment to contributors should be made clear. If multiple users update information for the same data, explicit attribution to each user must be made. To assess the quality of data and information, it is important for submitters to include information about experimental design and methods of data collection and annotation (so called "meta data"). For example, if a researcher is submitting results from a protein localization experiment, they should provide information on methodology (e.g. green fluorescent protein translational fusion, immunolocalization, in vitro transport assays, etc.), how the data were collected, and methods for analysis of raw data. The same standards used by journals should be applied, not only for "wet lab" methods but also for computational analysis. The data should be made publicly available with minimal delay, and the submitter should have a chance to review the data in the database before the public release to ensure that he or she is satisfied with the display. In addition, the history of any significant changes made to a data set (history tracking and versioning) must be transparent. The ability to capture and present the dynamic nature of information as we gain more insight and better analysis methods is one of the advantages of databases over traditional publication in journals. Last but not least, there should be an easy way for users to report errors, provide updated information, and make suggestions for improvement. Most databases include contact e-mail links on every Web page to make it easy for users to send these comments and critiques immediately.

## COOPERATION AND COLLABORATION AMONG COMMUNITY DATABASE SYSTEMS

Researchers rarely encounter difficulties in extracting information from papers, even when the subject is unfamiliar. In part, this is because of standardized formats used by scientific publications and common standards for data reporting. Unlike the scientific publication in journals, community databases have developed more or less independently from each other because they cater to distinct audiences, and they have a much shorter history than journals. The need for common database data types and interfaces has increased in the last few years as more sequence and comparative information became available. As the boundaries between the communities served by each database become more diffuse, the need for standard ways of displaying and disseminating information will increase. In response to this need, the National Institutes of Health's National Human Genome Research Institute division initiated a program called the Generic Model Organism Database (http://www.gmod.org). Its goal is to enable established model organism databases to share their software, standard operating procedures, and expe-

  

riences. This will allow them to come up with a set of useful toolkits for new community database systems to be rapidly deployed. As a part of the initiative, software developers and curators of community database systems meet on a regular basis and share their experiences in developing open source software and in annotating data (e.g. Biocurator, http://biocurator.org).

The way that information is described is as important as the way it is displayed. The language of biology is often specific to each research system and comparison among different systems is impossible unless there is a common vocabulary. For example, comparison of gene functions will be difficult if the annotations in different database systems use different language. To address this problem, a few model organism databases initiated a project called the Gene Ontology (GO) Consortium to develop a common set of vocabularies to describe the biochemical functions and cellular locations of each gene product, along with biological roles in the organism (GO Consortium, 2001; http://geneontology.org). The vocabularies developed by the consortium are explicitly defined and structured in hierarchies such that broader concepts encompass narrower, more specific concepts. This consortium has increased in size since its inception in 1998 and now has 17 member databases that have collectively developed some 16,000 terms and used them to annotate about 1 million gene products from approximately 58,000 organisms. The implementation of the GO project has enabled researchers to query for a gene or biological topic of interest and retrieve all relevant gene products and get their detailed information from a large number of community databases. However, some aspects of biological processes are not yet handled by GO, such as temporal and spatial patterns of gene expression and mutant phenotypes. To address this issue, a few plant community databases have come together recently to coordinately develop plant anatomy and developmental stage vocabularies in a collaborative project called the Plant Ontology Consortium (http://plantontology.org).

In addition to the semantic (description or definition of information) challenges in sharing information across databases, syntactic (format in which the information is presented) challenges of exchanging data and information across different databases must be addressed. The challenge of sharing information across different databases is not limited to biology, and many fields dependent on the Internet for communication are developing and testing mechanisms of information sharing. A widely used method is to hyperlink to relevant Web pages by using a combination of a URL base rule and an external identifier for the data object of interest. All community databases should provide a simple URL base rule and external identifiers of data objects to allow others to hyperlink to their pages easily. Hyperlinking is fairly rudimentary, and more sophisticated ways of exchanging information are available. These include simple object access protocol (http://www.w3.org/TR/SOAP/) or common object request broker architecture (http://www.omg.org/gettingstarted/corbafaq.htm), which allow more powerful ways of connecting between databases so that one could form a query from one site and search databases at another site. In addition to these database exchange methods, applications have been developed that allow researchers to exchange and share genome annotations with each other and community databases. One example is the distributed annotation system (http://biodas.org/) used by WormBase (http://wormbase.org). Ultimately, successful implementation will follow only after a wide acceptance by community database systems.

## ROLE OF INTERNATIONAL FUNDING AGENCIES

An effective way to enforcing sharing of resources has been the requirement of funding agencies for grant recipients to promptly release genomic data to the public (Silverthorne, 2003). However, such policies are country specific, and there is not yet an international agreement among funding bodies that encourages public release of data. Despite the huge amounts of funding dedicated to generating genomic resources in many developed countries, much of the information is not made available to the research community. A successful international agreement among funding agencies for releasing sequence information to GenBank was established at the Bermuda International Meeting on Human Genome Sequencing (http://www.sanger.ac.uk/HGP/policyforum..shtml). Similar international agreements for releasing genomic resources to the public could stimulate contributions of extremely valuable data to public repositories and community databases.

## CONCLUDING REMARKS

New technologies drive new ways of thinking. With the explosion of biological data available today, researchers increasingly incorporate other people's data and results in their own research. Community database systems that collect and disseminate curated genomic resources allow researchers to gain access to more information faster and more easily than ever before and facilitate new ways of thinking about biological problems and conducting research. Community databases can also level the playing field for members of the community by democratizing access to information, not unlike the printing press. Just as importantly, these databases serve to create a "virtual community" of researchers who share their findings, thoughts, and questions with each other. There is no doubt that these changes will bring about

new paradigms and advances in biology in our lifetime. However, to maximize the return on our investment, it is essential that members of the community actively participate in the improvement of the community databases by contributing data, biological insight, and feedback. The research community, academic institutions, community databases, publishing groups, and funding agencies all have a role to play and a responsibility to make these changes.

## ACKNOWLEDGMENTS

I am grateful to Leonore Reiser, Eva Huala, Tanya Berardini, David Jackson, Suparna Mundodi, and Peifen Zhang for their critical reading of the manuscript.

Received November 6, 2003; returned for revision November 14, 2003; accepted November 14, 2003.

## LITERATURE CITED

**Baxevanis AD** (2003) The Molecular Biology Database Collection: 2003 update. Nucleic Acids Res **31:** 1–12

**Brower V** (2003) Biotech industry tries to recover from 2-year lull. J Natl Cancer Inst **95:** 348–349

**Burks C, Fickett J, Goad W** (1987) GenBank status report. Science **235:** 267–2688

**Duyk GM** (2002) Sharper tools and simpler methods. Nat Genet Suppl **32:** 465–468

**GO Consortium** (2001) creating the gene ontology resource: design and implementation. Genome Res **11:** 1425–1433

**Lahteenmaki R, DeFrancesco L** (2003) Public biotechnology 2002: the numbers. Nat Biotechnol **21:** 607–612

**Reid B** (2002) HGS drug flop latest genomics setback. Nat Biotechnol **20:** 533

**Rhee SY, Beavis W, Berardini TZ, Chen G, Dixon D, Doyle A, Garcia-Hernandez M, Huala E, Lander G, Montoya M et al.** (2003) The Arabidopsis Information Resource (TAIR): a model organism database providing a centralized, curated gateway to Arabidopsis biology, research materials and community. Nucleic Acids Res **31:** 224–228

**Silverthorne J** (2003) Ensuring access to the outcomes of community resource projects. Plant Physiol **132:** 1775–1778

  