

Metadata for Plant Seeds: Taxonomy, Standards, Issues, and Impact

Erin Wilson
Berkeley, CA

Introduction

It is an interesting time to be looking at metadata about plant seeds: a growing awareness of the importance of biodiversity has brought international attention to seed collecting, saving, and information sharing. At the same time, scientists are continually learning more about seeds at the most basic biological level, and this new knowledge is raising questions about the taxonomic systems scientists have used since the 1700s. Contemporary seed metadata is a complex and diverse area for study.

This paper is divided into four sections. First, it will begin by looking at the taxonomic systems employed by botanists and discussing current trends and issues in that area of seed metadata. Second, the most prominent seed metadata standards will be examined. Third, the emerging field of bioinformatics will be addressed, and finally the impact of biodiversity impact on seed metadata will be explored by looking at the major seed collecting and saving organizations and projects. Throughout this paper the focus is on seed metadata, but that cannot be separated from issues in the world of botanical metadata in general.

Taxonomy

Biological life, including seeds, has been classified using the Linnaean system since the mid-1700s. This classic taxonomic system forms our fundamental understanding of how the biological world is organized and is a model for other forms of taxonomy. To be clear, taxonomic data is a form of metadata. Botanists debate taxonomic classifications—whether a specific species is appropriately identified and classified—and much of the taxonomic literature is about these issues (Stevens, 2003), but it is the larger taxonomic system and how it relates to metadata that concerns this paper.

Linnaean taxonomy is in some way a model metadata standard: it is extensible, is the universal biological standard, and it has an agreed upon processes for proposing, revising, citing, and formalizing data changes (Godfray, Clark, Kitching, Mayo, & Scoble, 2007). Begun in the mid-1700s by Carolus Linnaeus (1707–1778), it has been revised and updated many times. It divides biological world into the familiar system of kingdoms and ranks, and includes the familiar binomial nomenclature, which uses the genus name and the specific name to form the species name (Minelli, 2005).

Botanical nomenclature is governed by International Code of Botanical Nomenclature (ICBN) (International Association for Plant Taxonomy, 2000). The ICBN is maintained and published by the International Association for Plant Taxonomy, an international association of scientists (International Association for Plant Taxonomy, 2007).

The acceptance of Linnaean taxonomy means that botanists have a good start toward interoperability, but Linnaean taxonomy does not provide one key to interoperability: unique identifiers (Godfray et al., 2007). Using the Linnaean taxonomic system, different species can have identical names,

which means that it is impossible to identify data in a way that is explicit and machine-readable. One prominent remedy is the proposed Life Sciences Identifiers (LSID).

LSIDs are being championed across the biological sciences. The system uses the construction of a Universal Resource Name (URN) using an authority name space ID, object ID, and authority ID (Smith & Szekely, 2005). LSIDs have their own resolution protocol and they “are persistent, global and location-independent object names” (Clark, Martin, & Liefeld, 2003). There has been some criticism of LSIDs based on their potential conflicts with other forms of internet protocol (NeuroCommons, 2008), but they are widely used. It is important to note that LSIDs are not a replacement for biological names since LSIDs are only for electronic resources (Huber & Klump, 2008), but that does not detract for their importance to metadata interoperability.

Metadata Standards

In addition to taxonomic metadata standards, there are a variety of other metadata projects and standards relevant to seeds.

The Biodiversity Information Standards group (known as TDWG) was formed to establish international collaboration among biological database projects and it focuses on developing standards for exchanging biological data. It is affiliated with the International Union of Biological Sciences and, in addition to being strong supporter of LSIDs, it leads the development of key metadata standards (Biodiversity Information Standards, 2007).

The SDD (Structure of Descriptive Data) standard from TDWG is designed to “allow capture, transport, caching and archiving” of descriptive, natural language. It is a platform- and application-independent standard that uses couplets of descriptive data fragments to make branching trees of data (Hagedorn, Thiele, Morris, & Heidorn, 2006).

The Access to Biological Collections Data (ABCD) Schema, also from TDWG, is comprehensive and highly structured. It is attempting to set the standard for the access to, and exchange of, information about specimens and observations. One of its strengths is that proposes parallel structures so that it can handle free-text and atomized data (Biodiversity Information Standards, 2009a). ABCD includes a great deal of metadata on intellectual property rights which means that organizations that use ABCD can make claims about the copyright, accreditation and use of their data (BioCASE, 2009). ABCD is very widely used.

Another TDWG standard, Taxonomic Concept Transfer Schema (TCS), is an attempt to navigate differences in taxonomic concept models. As discussed above in the section on taxonomy, the same name can be applied to different organisms—but in addition the same name can also be applied to different taxonomic concepts. Although individual databases or collections rely on a consistent set of names and concepts, the differences between datasets clearly makes interoperability impossible. TCS is an XML schema document that proposes a standard to allow exchange of data between different data models and “it aims to capture data as understood by the data owners without distortion, and facilitate the query of different data resources according to the common schema model” (Biodiversity Information Standards, 2009c).

Darwin Core, also from TDWG, is a standard developed to be a stable reference for standard terms about biodiversity. Its form is primarily based on Dublin Core Metadata Initiative (and it cleverly borrows its name from it). It consists of:

a vocabulary of terms (properties, elements, fields, concepts), the policy governing the maintenance of these terms, decisions resulting in changes to terms, the complete history of terms including detailed attributes, a generic application schema for use in the construction of new application

schemas based on Darwin Core, a simple (flat) application schema for the use of these terms and a metafile schema to allow for the description of Darwin Core fielded text files (Biodiversity Information Standards, 2009b)

The Biological Collection Access Services (BioCASE) provides access to European biological collections through a central web portal. BioCASE works by supplying data providers (the original collections) with software that uses a XML based query language to abstract the original collection's schemas and map it onto BioCASE's database (BioCASE, 2009). To do this it relies on agreed upon metadata schemas, including Darwin Core, ABCD, and even Dublin Core.

Global Biodiversity Information Facility (GBIF) is an independent, international scientific organization committed to the digitization and global dissemination of primary biodiversity data. Its web portal provides access to 7788 datasets from 283 data providers. To make these disparate datasets interoperable it uses many of the standards discussed in this paper, including ABCD, LSID, BioCASE, and Darwin Core (GBIF, n.d.). As such a mammoth project, its adoption of a standard (for example Darwin Core or LSID) has a profound impact on that standard's widespread use.

Bioinformatics

Bioinformatics is the “study of biological information using concepts and methods in computer science, statistics and engineering” (Rhee, Dickerson & Xu, 2006). It is a relatively new field and bioinformatics professionals combine knowledge of molecular biology with the ability to create and manipulate databases, algorithms, computational and statistical techniques to solve biological problems.

Bioinformatics arose from advances in gene sequencing that provided scientists with unparalleled amounts of information about individual species at the same time that computers also provided new ways to manipulate and explore data. Thus far most gene sequencing projects have focused on humans and animals, but there is an increasing focus on plants—including several important crops like barley and peas (Kuenne et al., 2007)

As plants, and seeds, become more of a focus of gene research, bioinformatics will have an increasing effect on seed metadata. One of the key features of bioinformatics is the manipulation of very large datasets, often culled from multiple sources (Rhee, Dickerson & Xu, 2006), which means seamless access to multiple databases. The taxonomic and metadata standards discussed above are, of course, crucial for facilitating this kind of exchange.

Biodiversity and Seed Metadata

There is rising international recognition that seed varieties are being lost through agri-business and climate change. Traditionally seeds have been saved from year-to-year by individual farmers or exchanged at the local level. That has changed so that now the vast majority of farmers buy their seeds from major seed producers. These purchased seeds have been bred for increased disease resistance as well as other types of plant characteristics—for example height, nutrient content, or crop shelf life. In many ways this has been a boon for farmers, but it has also contributed to diminished seed biodiversity. A stark, but unfortunately typical, example of this diminished biodiversity is the apple varieties in the United States: in the 1800s over 7100 named varieties of apples were grown while today there are only about 1000 (Global Crop Diversity Trust, 2006). Climate change is also having a profound impact on seed biodiversity. As weather and land conditions change, some varieties simply no longer grow and are lost that way.

The loss of biodiversity has several important consequences. First, having only a few varieties of a particular crop means that the entire production of that crop is vulnerable to infestation and disease. The fewer the varieties of crop that are grown, the most vulnerable the whole crop is to destruction.

Second, crop variety allows scientists and crop breeders to adjust to climate change by finding, or developing, new varieties to fit changed growing conditions. Biodiversity is the tool kit for evolution (Global Crop Diversity Trust, 2006).

How does this relate to metadata? Seed biodiversity is an international issue that must be addressed beyond the confines of nations or even continents. There are innumerable seed metadata projects but consistent concerns are access and interoperability. Under the pressure of rapid climate change, scientists are scrambling to collect and save seeds with accessible and relevant metadata that allows them to be both preserved for the future and used now. There is fierce pressure to collect all the information possible about all available seeds, but—as will be illustrated below—seed metadata standards vary widely depending on the ultimate goal of the project and audience.

The United Kingdom's Royal Botanic Gardens at Kew have long been at the forefront of seed metadata creation. One of its current major efforts is the Millennium Seed Bank Project (MSBP). MSBP currently has more than a billion seeds, it will have over 24,000 species by 2010 and by 2020 they will have the seeds to safeguard over 25% of the world's botanical biodiversity, and including the species and regions most at risk from climate change (Royal Botanical Gardens, n.d.).

The metadata infrastructure for a project of this scale is obviously crucial. MSBP is developing and populating a database, the Seed Information Database, during the course of building the seed collection. The database is has taxon-based information about seeds, both from MSBP and from other sources. The main purpose of the database is to analyze these data for predictive patterns that support seed conservation operations (Royal Botanical Gardens, n.d.). The database is tightly controlled by MSBP curators, but some information about its standards can be gleaned by examining it from the outside.

The database contains fields like storage behavior, seed dispersal, seed protein content, plant life form, and seed morphology (Liu, Eastwood, Flynn, Turner, & Stuppy, 2008). The database provides detailed descriptions, which are clearly part of the metadata standards, of each field. For example, “seed weight is actually composed of a number of different pieces of information: 1000 seed weight (g); cultivar name (where appropriate); source reference; type of material that was weighed (e.g. seed, achene); notes relevant to the interpretation of the seed weight data” (Royal Botanical Gardens, n.d.).

There are many independent seed banks and seed saving networks around the world. The Seed Savers Exchange (SSE), based in Decorah, Iowa, is one of the largest seed non-governmental organizations. It conserves plant genetic resources in North America (Lewis, 1997). SSE permanently maintains more than 25,000 endangered vegetable varieties, while its distributed network of gardeners offers 13,263 unique varieties of heirloom seeds. Maintaining biodiversity and having that biodiversity fully accessible is a primary focus for SSE.

The seeds collected and offered by SSE gardeners reveal an interesting facet of metadata. The information collected includes: plant type, variety, maturity, description, and source information. Almost no guidance is given as to the standards for these fields, so that there are wildly different concepts reflected in the material. For example, a search for celery revealed the “description” field to be completely empty on some seeds, while some included plant characteristics (“needs early start and plenty of water”), or flavor descriptions (“large ribs have nutty taste”), and still others had physical descriptions (“forms a dense clump of narrow thin green stalks”) (Seed Savers Exchange, 2009). The advantage to this kind of loose standard is that there is no barrier to participation by non-scientists and there is no learning curve to accessing the metadata. At the same time there is, of course, no way that the data could be shared with another database, and indeed within the database the lack of standards means that it is almost impossible to do a meaningful search.

Perhaps the most famous seed bank is the Svalbard Global Seed Vault in Svalbard, Norway. Opened for just over one year, the vault has been called the “Doomsday Vault” by the media. The Svalbard Global Seed Vault collects very little metadata about the seeds in its collection. Unlike BSBP or Seed Savers Exchange, Svalbard does not collect seeds for use. It is called the “Doomsday Vault” because the seeds are designed to stay safely in the seed vault in the case of an ultimate emergency. Svalbard’s seeds are our global insurance plan against total crop destruction (Svalbard Global Seed Vault, 2009).

Svalbard also provides a searchable database. Svalbard metadata includes: institute code, collection name, accession number, full scientific name, country of collection or source, number of seeds, and regeneration month and year. The metadata standards and submission forms are freely available, and are surprisingly vague. For example, the number of seeds is defined as “based on a full count or on an estimate from the weight of the sample”—but in fact no guidelines are provided for system of measurements should be used (Svalbard Global Seed Vault, 2009). The Svalbard metadata reflects the purpose of the seed vault: it is for preservation but not access and has no true interest in interoperability. The seeds are there, separate from the rest of humanity, for posterity. Everyone hopes that we will never have to actually access the Svalbard Global Seed Vault.

A different approach to addressing the role of seeds, and seed metadata, in biodiversity is the Generation Challenge Programme (GCP). It is a global crop research consortium dedicated to improving biodiversity through using comparative biology and genetic resources to improve breeding. GCP is a partnership between international agricultural research institutes and its goal is to “exploit advances in molecular biology to harness the rich global heritage of plant genetic resources and contribute to a new generation of stress-tolerant varieties” (Bruskiewich et al., 2008). A primary component of this vision is the development of what they call a “crop informatics platform.” This platform will be freely available to all crop researchers and breeders, and provide a system for agricultural researchers in developing countries to become aware of new developments in crop breeding.

GCP adopted existing standards for identifying and naming seeds and has made those standards available in a dedicated online database, the GCP Pantheon (Bruskiewich et al., 2008). Among the standards used by GCP are LSIDs for unique identifiers and Library of Congress Subject Headings for subject headings (Generation Challenge Programme, 2008). GCP is committed to access for scientists and crop breeders around the globe and the standards it uses are within that paradigm.

Conclusion

Any research into the metadata about seeds cannot be separated from the larger field of biological metadata. There is a diversity of concerns affecting seed metadata, many of them driven by the pressures of biodiversity and new advances in biology. Seed metadata has a good foundation in that Linnaean taxonomy is almost universally accepted, and there is wide spread commitment to the idea of data sharing. There are many projects committed to setting metadata standards, like TDWG, and setting up systems for information sharing, like GBIF. Change is on the horizon with the growth of bioinformatics and it will be important to track how bioinformaticists and more traditional botanists work together to resolve the pressing issues of seed metadata.

Works Cited

Biodiversity Information Standards. (2007). About. Retrieved March 5, 2009, from <http://www.tdwg.org/about-tdwg/>

Biodiversity Information Standards. (2009a). Access to Biological Collections Data. Retrieved March 12, 2009, from <http://www.tdwg.org/standards/115/>

- Biodiversity Information Standards. (2009b). Darwin Core. Retrieved March 12, 2009, from <http://www.tdwg.org/standards/450/>
- Biodiversity Information Standards. (2009c). Taxonomic Concept Transfer Schema. Retrieved March 12, 2009, from <http://www.tdwg.org/standards/117/>
- Biological Collection Access Services. (2009). BioCASE provider software. Retrieved March 12, 2009, from http://www.biocase.org/products/provider_software/
- Bruskiewich, R., Senger, M., Davenport, G., Ruiz, M., Rouard, M., Hazenkamp, T., et al. (2008). Generation Challenge Programme platform: Semantic standards and workbench for crop science. *International Journal of Plant Genomics* . doi:10.1155/2008/369601
- Clark, T., Martin, S., & Liefeld, T. (2004). Globally distributed object identification for biological knowledgebases. *Briefings in Bioinformatics* 5 (1), 59-70.
- Generation Challenge Programme (2008). GCP Pantheon. Retrieved March 12, 2009, from <http://pantheon.generationcp.org/index.php>
- Global Biodiversity Information Facility. (n.d.). Standards. Retrieved March 12, 2009, from <http://www.gbif.org/links/standards>
- Global Crop Diversity Trust. (2006). Main. Retrieved March 10, 2009, from <http://www.croptrust.org/main/>
- Godfray, H., Clark, B., Kitching, I., Mayo, S. & Scoble, M. (2007). The web and the structure of taxonomy. *Systematic Biology* 56 (6), 943-955.
- Hagedorn, G., Thiele, K., Morris, R. & Heidorn, P. B. (2006). The Structured Descriptive Data (SDD) w3c-xml-schema, version 1.1. Retrieved March 11, 2009, from <http://rs.tdwg.org/UBIF/2006/rddl.html>
- Huber, J. & Klump, J. (2008). What LSIDs are good for. Stratigraphy.net intervals. Retrieved March 14, 2009, from <http://stratigraphy.net.blogspot.com/2008/06/what-lsids-are-good-for.html>
- International Association for Plant Taxonomy. (2000). International Code of Botanical Nomenclature. Retrieved February 27, 2009, from <http://www.bgbm.org/iapt/nomenclature/code/SaintLouis/0000St.Luistitle.htm>
- International Association for Plant Taxonomy. (2007). About. Retrieved February 27, 2009, from http://www.botanik.univie.ac.at/iapt/index_layer.php
- Kuenne, C., Grosse, I., Matthies, I., Scholz, U., Sretenovic-Rajcic, T., Stein, N., Stephanik, A., Steuernagel, B., & Weise, S. (2007). Using data warehouse technology in crop plant bioinformatics. *Journal of Integrative Bioinformatics* 4 (1). doi:10.2390/biecoll-jib-2007-88
- Lewis, V. and Mulvany P.M. (1997). *A typology of community seed banks* . Natural Resources Institute, University of Greenwich.
- Liu, K., Eastwood, R.J., Flynn, S., Turner, R.M. & Stuppy, W.H. (2008). Seed Information Database (release 7.1, May 2008). Retrieved March 8, 2009, from <http://www.kew.org/data/sid>
- Minelli, A. (2005). Classification. In *Encyclopedia of Life Sciences* . DOI: 10.1038/npg.els.0004121

NeuroCommons. (2008). Life sciences identifier. Retrieved March 8, 2009, from http://neurocommons.org/page/Life_sciences_identifier

Rhee, S. Y., Dickerson, J., & Xu, D. (2006). Bioinformatics and its applications in plant biology. *Annual Review of Plant Biology* (57). 333-60.

Royal Botanical Gardens. (n.d.). Millennium Seed Bank Project Royal Botanical Gardens, Kew. Retrieved March 6, 2009, from <http://www.kew.org/msbp/index.htm>

Seed Savers Exchange (2009). Celery. Seed Savers Exchange Yearbook. Retrieved March 4, 2009, from http://seedsaversyb.dreamhosters.com/pt_search.php?PlantType=CELERY

Smith, D. & Szekely, B. (2005). A guide to deploying Life Science Identifiers. Retrieved March 13, 2009, from <http://www.ibm.com/developerworks/opensource/library/os-lsidbp>

Stevens, P. F. (2003). History of Taxonomy. In Encyclopedia of Life Sciences. DOI: 10.1038/npg.els.0003093

Svalbard Global Seed Vault. (2009). Svalbard Global Seed Vault. Retrieved March 10, 2009, from <http://www.regjeringen.no/en/dep/lmd/campain/svalbard-global-seed-vault.html?id=462220>