Biodiversity informatics, a basic ingredient of natural science museums

FRANCESC URIBE, Museu de Ciències Naturals de Barcelona and JOHN WIECZOREK, Museum of Vertebrate Zoology, University of California, Berkeley.

NATURAL SCIENCE MUSEUMS AS BIODIVERSITY INFORMATION PROVIDERS

Since ancient times and throughout the history of scientific knowledge, physical natural science collections have been the most prestigious asset of natural history museums. Samples in collections are a tangible testimony of several aspects of biological diversity: taxonomic, geographic, seasonal, life cycle, age or sex, etc. Collections illustrate current and past diversity, the latter being irreplaceable by new studies, but essential for understanding evolution. The importance of collections does not decrease, but rather increases with scientific progress. Molecular studies applied to the analysis of fundamentals of detected biochemical variability are currently so powerful that samples can now be used that were previously discarded due to their age and state of preservation.

Museum samples hold clues to the effects of current environmental conditions as well as those from the past and are available for research to discover pollutants, biological traces, etc. Material elements of collections can be permanently used by scientists, as they offer the possibility of replication, a fundamental condition in scientific research, since the sample remains for further testing, analysis and experiments.

Since the end of the 20th century, the traditional concept of heritage in natural history museums, based on material assets, has evolved to a new dimension: the digital highway, a metaphor in which management and dissemination of information are dominant aspects of the desire for dynamic and high speed access to knowledge. The power of collections to provide information (which is essentially immaterial but based on materials and not on simple observations) has become a requirement for new social and scientific roles in addition to the traditional uses of collections.

Technology tools applied to data, in combination with techniques for preservation and conservation of materials, constitute the new skills required to curate collections. In practice, technology has led to increased investment of time in digital projects and the related training for curators of collections. Natural science museums are formidable containers, with huge amounts of high quality sample-centered data information, accumulated over many years, (Ariño, 2010). It seems logical, objective, necessary, and even strategic to capitalize on this intangible resource that museums have to offer. Nevertheless, who could be interested in this?

INTENSIVE USE OF SCIENTIFIC DATA

Ecological and biogeographical research and environmental studies for diagnosing and predicting conditions of a territory or to design conservation actions, are, among others, examples of research that needs increasingly larger datasets. Substantial datasets can be analyzed with statistical and cartographic tools and many variables, either

continuous or discontinuous in space or time, contribute to the process (Frew & Dozier, 2012). Terms such as date-centered science, big data, data intensive science, etc., combine to reflect the need for particularly large amounts of data to detect trends in systems as complex as natural systems. Such complexity grows exponentially with the increase of spatial and temporal dimensions that must also be taken into account.

Biological datasets can be clustered: for instance, layers of vegetation with zoological studies or distribution maps of pollinators with vegetation atlases. Biological data can be linked to other sources of information: soils, hydrology, climatic charts, paleoclimatology, human activities, chemical pollution, or noise, and so on. Different combinations of content layers may disclose associations, inhibitions, and "hidden structures" (Heckerman *in* Service, 2013) that could help us to understand natural mechanisms. Biodiversity informatics is a very important discipline in scientific procedures. Computer tools have been devised for managing and analyzing huge data sets, as researchers are unwilling to overcome bureaucratic obstacles or to rely on phone calls to obtain information. Today, they expect to download data from recognized websites (warranted by relevant research authorities) directly to their desktop easily and quickly. Can data in natural science museums compete in the biodiversity information market?

MUSEUMS, FULL OF REASONS

The material and digital heritage of natural history museums is of paramount importance, especially when museums build alliances between each other to produce aggregate volumes of data. Such an asset, often publicly owned, obliges those who manage it to implement actions in order to increase its social value.

Until the mid-twentieth century, museums developed procedures to create data files that systematically captured descriptive data of each sample in their collections. The main objective was to offer querying capability for searching and locating samples. Catalogs or paper card inventories enabled linear classification by a single criterion (often taxonomy), but did not facilitate queries for other general categories (e.g., animals of one sex or of certain age) or for a combination of criteria (e.g. a particular animal species occupying a geographical area). It was often necessary to first choose one criterion and then choose cards according another one.

With the advent of digital structured and relational databases, easily accessible information was available, and natural science museums began a long process toward computerizing data from paper records and labels. Long before completing this goal, a new objective was envisaged in the last quarter of the 20th century: to make databases accessible on the Internet so that their potential use could expand without limits. One limitation would be, however, caused by the gigantic growth of the Internet, which created a risk of relative invisibility of the data in the new digital universe. At the beginning of the 21st century, in addition to the previous objectives, a new challenge has arisen, which is to support an efficient dissemination of information enabling potential consumers to locate and exploit museum data.

Many museums are still engaged in the first phase (computerization of collections: remember that they are extensive!), so Internet publication is still weak and it is not easy to design positioning strategy for the Internet. How to face three consecutive objectives, all of them highly time consuming for computerizing tasks is not a trivial question.

For reasons of coherence, one might conclude that it is necessary to concentrate resources on phases following a historical pattern; however, if we consider the current logic it may well be necessary to do otherwise. It could be better to collaborate with projects led by renowned institutions that could provide high visibility to museum data. A desired by-product of this participation could be a renewed interest in museums, which could channel resources to increase the information

digitization project. Each museum should have sufficient self-knowledge and independence to assess priorities. Fortunately, there are sufficient stories of success and failure to help determinations. Nevertheless, we must not lose sight of what is important.

ESSENTIAL CONDITIONS

Natural science museums have always classified samples and data; from the traditional display cabinets to the current containers and cloud services. Current fashionable and glamorous information and communication technologies do not preclude taking into account the strict criteria that should pave the way for their use. The key consists in controlling the structure and content of information. This may seem deceptively easy to do, but reality tends to be more demanding. The information must be organized into fields of identifiable information, contents must comply with controlled vocabularies (preferably shared by the expert user community), and descriptions in metadata should not be neglected. Museology has already taken these criteria into account, and standards and schemes have been developed and applied by communities of experts. The aim is to achieve robust, searchable and downloadable databases with the minimum possible ambiguity.

Natural science museums share biodiversity data with other agents such as powerful scientific and academic research centers. Museums deal with management parameters of specific interest for these kinds of centers, but content standardization must respond to intrinsic and shared values of the information on biodiversity, and to the requirements of users. The Biodiversity Information Standards organization, TDWG[1] provides leadership in the development of standards and communication protocols for biodiversity data. Its aim is the creation, development, use and promotion of standards, including the most influential and most commonly used Darwin Core (Wieczorek et al., 2012). Discussion forums and annual meetings are open to individual or institutional members of TDWG. The world's great natural history museums, among other actors interested in biodiversity informatics, participate in discussions. Beyond the difficulties of reaching agreements, or the technical subtleties that prolong debates, we are fortunate that such a promoter of standardization exists.

To manage and publish structured and controlled content and to connect it to collective repositories using established protocols is essential for achieving institutional trust. If authoritative sources of information are to be distinguished above those who cannot show quality references, natural history museums should exploit the resource of the inherent quality of the data referring to collections, by means of the precision of powerful and ambitious data-providing services. However, it is easy to be overwhelmed by the immensity of pending work for managing databases of collections with criteria of quality and efficiency. Certain strategies can help to discriminate the best way to move forward and to identify opportunities for involvement.

PATHWAYS FOR PROGRESS

It is easy to expect technological shortfalls in medium-sized or small museums. Managers and boards of directors are not happy with this situation, but the truth is that in a hypothetical technological race, museums are likely to lose. The progress of biodiversity informatics is constant and strives to operate in a distributed way, with various sources of information made available and stored in large-scale databases that are suitable for complicated analysis in order to interpret natural systems. Natural science museums find it difficult to maintain the pace of development of new computing biodiversity tools.

Natural science museums are in an unequal obstacle race in the pursuit of their goals. First of all, museums must computerize their legacy. Even before achieving this objective, museums have to spend most of their information management time

1. http://www.tdwg.org/

on Internet publishing projects. At present, museums are increasingly facing the challenge of becoming connected sources of information, as rigorous and accessible by analysis tools as possible. New objectives are piled onto old ones, and at the same time, analytical capabilities of communities of experts tend to be ahead of the capacity of museums to provide information in useful ways. How can we overcome this deficit? By:

- Professionalizing computer system management in museums.
- Prioritizing computerization objectives.
- Promoting automated data quality control and linked data.
- Participating in data aggregation platforms.
- Citizen collaboration: in situ and remotely.

PROFESSIONALIZING COMPUTER SYSTEM MANAGEMENT IN MUSEUMS.

Those responsible for the care and use of public collections are usually specialists in the characteristics of the preserved heritage. Natural science museums often recruit people with biological or geological backgrounds. In practice, however, these people spend a large part of their time working on matters related to information technology and possibly also communication technologies.

As a rule, natural science museums that have a strong position in the information market possess their own teams, either external or mixed, to handle technological programs and discharge curators of responsibilities outside those related to their expertise. The pace of change and complexity inherent to biodiversity informatics has a direct impact on the technical skills of museum professionals. In the same way as a nobody doubts the unquestionable influence on the heritage of the museum of the department of preventive conservation, professional support in technologies for data dissemination should be another technical area able to increase the value of natural science collections. New professional profiles would include computer knowledge and information management.

Professional computer and information systems for museums should not necessarily be a headache for each center if participation structures, i.e., aggregation platforms, shared services resources exist. In other words, and taking advantage of the typical outsourcing of these technologies: networks of interoperability of information of collections (or other museums departments) may favor the scalability of expenses, if you create a shared technology service supported with proportional contributions of each center. From the perspective of the community of museums and museum professionals, this cooperation is the best foundation for a computing infrastructure that is able to transform information into knowledge (Krishtalka & Humphrey, 2000).

PRIORITIZING COMPUTERIZATION OBJECTIVES

Many forecasts have been made about the time required to computerize all the information contents of natural history collections into databases (ranging from several decades to centuries!). It is not fair to just follow simple cumulative procedures and wait for the end of a task that some of us might never see accomplished. A quick shortcut could consist in generating metadata descriptions of collections that could play the role of resource locators[2], especially taking into consideration that taxonomic and geographic information are the main searching factors (Berendson & Seltmann, 2010). In natural history museums, there is little tradition of describing collections, sets of samples, by means of metadata. The task of discriminating what records can be delimited to create a metadata file is not always easy. Nevertheless, the benefits outweigh the effort spent, due to the increased percentage of cases in which these resources are

2. Per exemple. http://www.bioexplora.cat/ncd/inici/lang-es

located by potential users. Note that the use of information standards is also welcome in this task.

Item by item computerization should not be substituted by metadata statements, but there are choices and decisions to make about which items to digitize first. It seems reasonable to establish priorities according to the needs of current or potential users of collections (Berents et al, 2010). A computerization work plan that only takes into account the interests of those in charge of the collections could miss opportunities to use them. According to Berents et al.(op.cit.) there are objective priorities such as type specimens (closely associated with the description of new species or subspecies), samples associated with future or current projects ("live" samples), records with historical significance (invoked by some publication project), or samples of species that have some value for which the center can earn uniqueness.

PROMOTING AUTOMATED DATA QUALITY CONTROL AND LINKED DATA

There is no need to reinvent all thesauri or all definitions that control database contents. It is preferable to take advantage of properly vetted thesauri and definitions that adequately cover each field of information in records of collections. Nevertheless, these consultations can be onerous if they lead to data cleaning procedures by means of printed resources, mainly if controlled vocabularies are very large. Participation in interoperability projects is an opportunity to make links to dictionaries available on the Internet and agreed upon by large communities of experts.

As a result of work carried out by such communities, we witness a new perspective on terminological control: the development of ontologies that define relationships between concepts that can be represented by metadata terms (Thessen & Patterson, 2011). This is the open door led by the semantic web for "intelligent" data search and browsing. Standard controlled vocabularies, metadata, and ontologies are all associated with communication technologies.

It is highly productive to have a plan on how to connect data in collections to authoritative web services from which to verify and expand the meaning of terms used. For example, the simple scientific name in our database can be used to get much more information from other sources (linked data); a whole package of content can be outside our system of information but linked to it. So, you do not have to worry about the names of the authors of the description of the mentioned taxon, the scientific publication reference, synonyms, or the validity of the name, etc. The names of species can also be used in pathways that connect to descriptions of their geographical distribution, vulnerability, etc.

PARTICIPATING IN DATA AGGREGATION PLATFORMS

An interest in participation in proposals of aggregating biodiversity data sources has already been mentioned. Users of these platforms have the enormous advantage of obtaining classified information from many sources with a single query, and probably will not give too much importance to whether the source is a museum or not. However, this possible lack of interest should not hide the truth: it is likely that the scientist does not see any other data source than aggregation platforms, and for this reason, museums should be present in federated data portals.

The most popular and comprehensive biodiversity aggregation platform is the Global Biodiversity Information Facility, GBIF[3], with over 530 million records now accessible from its aggregate portal. This initiative, born in the 1992 Summit in Rio de Janeiro, has a variety of global members. Biodiversity data are scattered throughout the world, although its representation in shared data sources seems to reflect more the socio-economic contribution than the natural wealth[4].

3. http://www.gbif.org/
4. http://iphylo.blogspot.com.es/2013_09_01_archive.html

Projects aggregating biodiversity data are diverse, both geographically and thematically. From a global scale to a local level, the Biodiversity Data Bank of Catalonia[5] is a platform suitable for harvesting data from museums. There are also data federations restricted to an area and devoted to a biological group. Among them, we can highlight VertNet[6], a global project focused on vertebrates, originating under the National Science Foundation in the United States. One feature of this network of collections of vertebrates is the assumption of a collaboration style that facilitates the inclusion of centers with limited technological capability. There is data exchange technology, in which leading centers in the network make available information management tools, which are necessary to accommodate smaller institutions that wish to include their contents, no matter where they might be.

CITIZEN COLLABORATION: *IN SITU* AND REMOTELY

Almost inseparable from the life of natural science museums, one can see that their work can be complemented by external, voluntary, and practically free collaborations. This has been so since the beginning of natural science museums. People who collaborate with museums may furnish scientific knowledge about a specific biological group, contribute with communication skills deployed in public activities, or assist in projects dealing with some very specific action.

The supporting community of volunteers can go beyond the physical dimensions of the museum and not just in the sense of participating in field research protocols. Pending operations in museums regarding computerization, debugging, or validating collection data now have a new resource; remote collaboration (Hill et al., 2012).

The difficulties of pattern recognition of samples, reading difficult handwriting on tags correctly, controlling the variable terminology of names of people or places, among other things, are easily achievable tasks in collaborative spaces arranged on a museum website. Controls for verifying data are a keystone of these projects and should be conscientiously planned and the time spent by volunteers would only require a minimal supervision by museum technicians. The fundamental condition is to adapt the objectives to the procedures of volunteer participation.

THERE IS NO SINGLE RECIPE

There is a new, complex and splendid scenario in natural science museums. Museums that would have been lost are now in a position to regain high scientific value if they are firmly inserted in the flow of scientific information. Nevertheless, these museums should be aware of the demise of the old centralist mentality, when information on biodiversity was practically reserved to museums and a few naturalist entities. This privileged position has changed (although it is not threatened) by the presence of active research centers, environmental companies and organizations, etc. It will not be easy to regain the central position of museums, nor may it be always logical to do so.

Regarding affordable knowledge, natural history museums must clearly be useful for society and for their peers. To play a renewed leading role, museums must show their capability of channeling and making their own information on biodiversity strongly profitable. At the same time, museums must also be useful for other centers that act as information providers, helping them to make their contents highly visible and attractive for potential users. The sum of relevant heritage and the coordination of shared services strengthens natural history museums, obliging them to have a permanent double perspective on information management: bot internal and external.

How to gain a new perspective for attaining the new objectives? Each museum can progress in it own particular way according to size and placement, and to the alliances, initiatives, and friendships it can promote. In spite of the common tendency to go on their own, it is equally true that

5. http://biodiver.bio.ub.es/biocat/index.jsp
6. http://vertnet.org/index.php

there is a way forward with clear advantages for all natural history museums: networking data sources, knowledge and services.

ACKNOWLEDGEMENTS

We are indebted to Marc Folia, Martí Pericay, Agustí Escobar, and Jordi Agulló for the talks and discussions held around the recent seminar on biodiversity informatics organized at the Museu de Ciències Naturals de Barcelona, 15/10/2013.