Work and Information Practices in the Sciences of Biodiversity

Geoffrey C. Bowker

Department of Communication
University of California, San Diego
La Jolla, CA 92093
USA
bowker@ucsd.edu

Abstract

This paper provides an introduction to data practices in biodiversity science. This is an area where multiple scientific domains are in constant interaction, and use data from multiple sources in that process. There is a consequent huge proliferation of technical standards in the field. Further, datasets used in the biodiversity sciences often extend over several decades — and thus attention must be paid to changing standards and the development of new storage media. Finally, the classification systems used in many of the contributing sciences are in a constant state of flux as more information is gathered. I describe the main categories of data development and use in the field of biodiversity, paying particular attention to work processes both in the generation and in the analysis of data.

1. The Age of Biodiversity Information

Diana Crane [2] claims that in scientific literature "The 'life' of a paper is very short, with the exception of a few classics. Papers published five years ago are 'old'. Papers published more than fifteen years ago are almost useless in many scientific fields". In this paper, I will examine a field of science in which this is emphatically not the case — the field of biodiversity science. Crane's model works best in physics, where there is no assumption that information collected in the early nineteenth century will still be of interest to the current generation of field theorists. There is the assumption [6, for example] that new theories will reorder knowledge in the domain effectively and efficiently; and since Kuhn [5] most would accept that a major paradigm change in, say, the understanding of 'gravity' renders previous work on incline planes literally incommensurable — not to mention technical improvements making the older work too imprecise. Astronomers trawl back further in time, seeking traces of supernovae in ancient manuscripts — but sporadically; they are just as likely to look at monastery records as at Tycho Brahe's original data.

Biodiversity information is fundamentally historical in three different ways. First, the sciences of biodiversity are trying to build up a picture of life on the planet since its inception up to the present. In the absence of the time and analytic ability to carry out experiments on complex ecological systems over tens to thousands of generations, the only real information about biodiversity comes from the history of life on this planet. Second, key work practices, especially in the field of systematics, require the recording of accurate information about publication place and date for publications extending back to the mid eighteenth century — a scope of interest unimaginable for any other scientific endeavour (and for few other academic endeavours). Third, the sciences of biodiversity more than other sciences depend on being able to triangulate between multiple disparate datasets produced by different agencies for different motives. The history of the dataset itself is a key to understanding the contribution it can make to the history of life.

Proceedings of the 26th International Conference on Very Large Databases, Cairo, Egypt, 2000

In this paper I shall briefly explore each of these three temporal aspects of biodiversity information with respect to work practices in biodiversity science. Be it noted that I am not thereby underplaying the vital role of issues of spatial scope and scale in biodiversity science — indeed I shall return to this issue in the conclusion.

2. Biodiversity Research and the History of Life

Doing good biodiversity research entails the collection and manipulation of massive datasets. All information about life on earth is potentially useful: from satellite photos through aerial photos and biodiversity inventories of forests down to minute descriptions of a square foot of soil and advances in molecular biology. No-one claims that it is easy to work these all into a single vast dataset: however to truly understand life on this planet we need to produce viable means for sharing information between these varied information sources.

There are two main models for understanding biodiversity, each with their own sets of information needs (and each with their own imperative to articulate with the other). The first is the view that life on earth can basically be understood informationally — there is a given amount of information in a gene pool, for example, and as the gene pool shrinks information is lost until a species becomes non-viable because it cannot react to changing conditions. By this view, the kind of data that we need in order to design good biodiversity policy is a fully ramified 'tree of life' which enables us to recognize key branch points and to identify specific species which carry a maximal amount of genetic information. The second is the view that biodiversity can best be understood ecologically: species develop in interaction with other species, and the basic unit of concern is the deme (an 'economically' active population group) rather than the gene. Although these views place their emphases differently, and sometimes lead to conflicting policy advice, they are not in principle irreconcilable.

Complicating the equation, is the fact that in the world of biodiversity research, difficult decisions need to be taken immediately on imperfect data. For example, at the current rate of completion, we could be well into the second half of this millennium before the various national floras are complete. Taxa are disappearing at a much faster rate; and decisions that in principle require the entire flora need to be taken now.

3. Biodiversity Research and Publications

In the field of biodiversity research, there are multiple needs for extensive manipulable online versions of publications dating back two hundred and fifty years. This is partly a technical issue of nomenclature: the rules of zoological and botanical nomenclature require that naming priority be given to the first published instance of a name. There are innumerable cases in the literature of plants and animals receiving multiple names (and there is natural reluctance on the part of a given community to change the names that they are used to), so this adjudicatory mechanism holds quite an important place. However, this need goes further, in that we need very long datasets in general in the field of biodiversity research: historical observations of a given community can have immediate relevance for undertanding their current structure and nature. Much of this information might be hidden in obscure journals in different fields and a variety of languages, each using their own naming conventions. Some of the earliest incunabula are botanical and zoological field guides; and there is an enormous wealth of locally-generated material since about the flora and fauna of most countries.

4. Datasets in Biodiversity Research

4.1 Nomenclature and Systematics

The current dogma, accepted in most parts of the world except Kansas, is that there is a single origin for life as we know it; and the work of systematists involves producing an hierarchical classification system that more or less accurately reflects the history of life. In the best of all possible worlds, one would imagine assigning a single identifier to flora and fauna found over the world using a facetted classification system that permitted easy 'on the fly' ordering of classifications when new data (e.g. the discovery of new species) demands. This is far from being workable, for a number of reasons. The first is that not all users have the same need or desire to keep up with the latest classification — a change in the genus of the tomato (recently approved) would cost nurserymen millions of dollars if they instituted it; a change in the name of a wild orchid might well remove it from legal protected status in a given country [4]. Second is that some agencies will follow their own name list, which often conflicts with other local, state, national and international lists — GAP analysis is a good example here [3].

One could multiply the reasons but the effect is clear. Although in principle one wants to have single identifiers for all taxa, in fact this is not achievable. Naming conventions will vary by group and discipline, and it will always be difficult to reconcile the different datasets. I should note that there are a number of national and international initiatives (some conflicting with each other) to bring some order into the field; however is is clear from the history of such attempts at order that they will generate their own local differences: this is not a fault of the field but a fact about the maintenance of global classification systems [1].

4.2 Regional and Temporal Variation

Many individual scientists or teams of scientists specialize both temporally (studying the fauna of the Eocene period for example) and spatially (studying the flora or fauna of a particular region). In principle this division of labor is efficient and in practice it is frequently effective. However, when one tries to aggregate data from a set of regions or from a series of tranches of the fossil record, one has difficulties Naming conventions differ from region to region, and from tranche to tranche. Thus simply fitting the pieces together to make a global overview does not work: there needs to be a process of negotiation at the same time between local specialists about their processes of identification and naming in order to prevent an artificial reification of contingent modern differences — Koch points to the possibility of tracing the Austro-Hungarian empire on the floral map of Europe, because of the differing naming conventions between the Austro-Hungarian and British Empires!

4.3 Range of Metadata Needs

Information about species tends to rely on a very small sample. In the case of plants, for example, a 'type specimen' is held by a herbarium in folders on library shelving. When a researcher wants to verify that such and such a species described in the literature is the same as another one described elsewhere, then he or she has to either visit the herbaria where the two desiccated type specimens are held or request that it be sent out through the mail. In some cases, even this impoverished material basis is not present, and the 'type specimen' is taken to be the description of the specimen in a published article. The type specimens carry their own context with them; in the form of annotations, lists of previous consultants, date and time of collection and so forth. The dispersed nature of the collections of type specimens means that reconciliation of the holdings of different herbaria is extremely difficult — thus when conjuring these data into electronic form there is a need to maintain sufficient flexibility in order to merge and disaggregate certain species.

Further, there is a need to preserve information about the precise details of a given measurment technique. Thus if an ecological measurement entails collecting lake water samples which are then measured for carbon dioxide content, it makes a difference to the measurement whether or not it is done immediately at lakeside or on return to the laboratory — perhaps later that day. In the short term, local research communities know and understand their own practices. However, over time this information is lost — in general when a scientific paper was published in the past, the accompanying dataset gradually decayed, and people forgot the full details of data collection (no scientific paper is long enough to be complete on these details). In practical terms, this means that there needs to be good and easy provision within biodiversity databases for the recording of as much contextual information as possible — both immediately and retrospectively. Research in the field has repeatedly shown that scientists will not have the time to fill in complex forms which record data that is not immediately relevant to their purposes. However, in the case of biodiversity science, later interest focuses often less on the paper and its conclusion than on the dataset and its construction. There needs to be a dual effort to on the one hand educate domain scientists about the nature of this shift and its implications for their work practices and on the other to design systems which make it truly easy to enter the maximum amount of contextual data.

4.4 Interdisciplinary Communication

Biodiversity work in general entails communication between multiple overlapping scientific disciplines. In general, each discipline has grown up with its own information infrastructure and information standards. This is true at a very mundane level: some scientific communities use almost exclusively Macintoshes, others rely on Unix boxes. It is also true at many other levels. For example, there are a number of different geological timelines available for paleoecological work. Any one subcommunity might be using a different, slightly conflicting timeline. This will not make a difference when there is no need to integrate information across disciplines: however this form of integration is of the essence of biodiversity research.

More generally, a slight lack of fit between models in different disciplines only becomes apparent when one is trying to integrate information across them. Thus, in a study I am doing of work to map the environmental hydrology of the Mississippi River Basin (work that has direct implications for the preservation of biodiversity in that region) there are inconsistencies between the atmospherical, groundwater and river flow models. The negotiation of these inconsistencies is occurring precisely at the time when a team of computer scientists are trying to build up a general model of the whole water cycle in the region. My point here is that a major part of the task of building robust databases in biodiversity is facilitating interdisciplinary communication — this communication cannot just be a desired outcome, it must be designed into the data collection and representation work that is being done. In the case I have described, and in innumerable others, this communication work is not receiving the attention it deserves.

4.5 Computer Science and Biodiversity Science

It is unclear what the career trajectory is for someone in computer science who turns to biodiversity science; or vice versa of someone in biodiversity science who turns to computer design. It is certainly clear in this regard that the interests of the two communities taken separately are not the same. A computer scientist often wants to use the

latest database techniques in order to produce a maximally effective and flexible tool. Indeed she has to, since without this she will not get promotion within her university department. However, the biodiversity scientist is aware that much of the work in the field is being done by relatively untrained parataxonomists, say, who have little or no access to computing equipment. For them, the simplest and most basic database form is the only one that they can use. Perhaps, as suggested by Schnase in an accompanying paper in these proceedings, these two fields must develop together over time to become truly synergistic.

5. Conclusion

I started this paper by talking about the importance of history in the design of good databases for biodiversity research. I could equally have talked about issues of space; and then walked this theme through an analogous set of issues. In both cases, the central lessons would be the same:

- Biodiversity research increasingly requires the generation of very large scale easily manipulable datasets.
- The creation of suitable database structures necessarily entails attention being paid to local work practices — in both the developing and developed world — with respect to both access to computing and modes of data collection.
- The creation of such databases is a site at which genuine interdisciplinary communication occurs — and so the role of databases in making such communication possible should be recognized at the moment of design. This can be done both through sensitising designers to different scientific cultures and carrying out studies of work practice concurrent which can feed into the design process.

Biodiversity research relies fundamentally on database design; and the present is a great opportunity for designing database structures which can further research through facilitating interdisciplinarity and can thus make a major contribution to the development of workable biodiversity policy.

6. References

[1] Bowker, Geoffrey C., and Susan Leigh Star. *Sorting Things Out: Classification and its Consequences*. Cambridge, MA: MIT Press, 1999.

[2] Crane, Diana. *Invisible colleges; diffusion of knowledge in scientific communities*. Chicago: University of Chicago Press, 1972.

[3] Edwards, Thomas C., Collin G. Homer, Scott D. Bassett, Allan Falconer, R. Douglas Ramsey, and Doug W. Wight. *Utah GAP analysis: An Environmental Information System.* Logan, UT: National Biological Service, Utah Cooperative Fish and Wildlife Research Unit, Utah State University, 1995.

[4] Klemm, Cyrille de, International Union for Conservation of Nature and Natural Resources, and World Wide Fund for Nature. *Wild plant conservation and the law.* Gland, Switzerland: IUCN-The World Conservation Union, 1990.

[5] Kuhn, Thomas S. *The Structure of Scientific Revolutions*. Chicago: University of Chicago, 1970.

[6] Poincaré, Henri. *Science and Hypothesis*. New York: The Science Press, 1905.