The Diversity of Biodiversity Information: Is it Inevitable?

Ray Lester
The Natural History Museum
London

As so frequently these days, we will be concerned in this Paper with the effect of developments on the Internet and the World Wide Web on the public availability of information. The information arena which we will focus on is that of information relating to 'biodiversity' or, more generally, to 'the natural world'. But the generic points I shall make apply to any arena where there are:

- First, a diversity of organisationally independent suppliers of data and information for the public domain.
- Second, a diversity of potential users - or; as I shall call them, customers - for subsets of such data and information: each of these customers, usually, affiliated to one or more organisations who, as need be, can vouch for their authenticity as users of certain defined types of data and information.
- Third, a diversity of intermediating organisations between the various data and information supplier organisations, and the various data and information customer organisations: each intermediary intent upon adding value to the data and/or information as it passes along the Internet from supplier to customer.

The overall global information system I have just sketched out is not only diverse, it is also complex. I should therefore explain the simple lens I will be using to help us try to make sense of the complexity, and to answer the question whether the diversity of information available - at least of 'biodiversity' information - is inevitable; or there is instead some way or there are some ways by which that particular information arena can be made less diverse.

Customers, by visiting natural history museums and other places, need or wish to garner information about the natural world. Advances in computing and telecommunications are encouraging us digitally to capture representations of the objects and processes within that natural world, and to make the digital representations available to the customers via computer-based networks. Such digital resources can be representative:

- First, of objects and processes within the natural world itself.
- Second, of objects and processes within what we might term the natural world study system: the people and organisations charged with studying that natural world - including, naturally, the organisations we call natural history museums, and the people who work therein.
- Third, the digital resources can be representative of objects and processes within what we might term the natural world communication system: the panoply of information artefacts which increasingly, these days, are formally made publicly available as digital publications in addition to, or instead of print publications.

Thus, for instance, my own organisation might publish on its Web Site a digital version of an article by the palaeontologist Richard Fortey in our serial publication *The Bulletin of The Natural History Museum* - this, then an element of the natural world communication system. We might then arrange to provide hypermedial links from that article to information about the work in general of Richard Fortey - perhaps how he came to write his best-selling book, *Life: An unauthorised biography* - this then an element of the natural world study system. From the digital record of that interview, or from the Bulletin paper, there might then be links to digital images representative of some of the fossils stored in the museum's collections, and whose scientific examination helped generate the theories expounded in the interview, or in the formal publications: these fossils then being elements of the natural world itself.

A key underlying notion in what I have just described is that of granularity. As well as enabling us to hop, as it were, sideways from digital data about, say, my own museum, to data about that other natural history museum; or from data about this scientist to data about that fellow scientist; or from this digitally published scientific paper to that other cited paper: the Web enables us also to drill down to increasing levels of detail: to become more granular. Instead of linking directly to the resource he or she requires, the customer might wish or need first to link to something that describes the required resource: a resource description. Such a description can be anything from, say, a minimal catalogue record, to - at the other extreme - a full-blown Web Site with its own rich and diverse content. The catalogue record might describe a digital image, say, of a natural history museum fossil specimen; the Web Site might describe the full extent of the 'resource' which comprises the museum as a whole. I am trying to keep this simple: but we can already see some of the diversity that Web technology has visited upon us.

Such diversity applies - potentially - to all semantic domains. But if now we take this definition of 'the natural world' which appears in an exhibition policy document prepared by my own museum:

'The natural world is defined as comprising the Earth and its planetary environment; constituent minerals and organisms (including humans), living and past; and human perspectives of, and interactions with, nature both living and material'.

it soon becomes clear just how diverse the panoply of biodiversity data and information accessible via the Web might become.

There are perhaps 1.8 million different types of organism or species that have so far been characterised by taxonomists. We, in this lecture theatre are all members of just one species out of the 1.8 million: each of us is a specimen of that one species. There are thought to be several million more species extant on the planet still to be characterised. However, not only do those concerned with 'biodiversity' need or wish to explore the behaviour of groups of specimens representative of particular species; also, they need or wish to study how each of those specimen populations interoperate with the populations of other species to form a biological community; and then how each particular community interacts with other communities and with their respective physical non-biological environments as a set of ecosystems within the world's overall biosphere.

In addition, scientists and others - frequently using the contents of the specimen collections found in natural history museums - need and wish also to capture data and information about the internal make-up of specific specimens. It is by now generally known how analysis of DNA and other related metabolites is potentially revolutionising our knowledge of which biological species are closely related in an evolutionary sense to which others, when compared to the knowledge that might be acquired from the specimens' visual morphological analysis.

The diversity of data and information and knowledge which is then potentially of relevance to a given biodiversity assignment is thus very large.

Now: I will be surprised if you have not noticed that I have been trying to be rather careful in my use within this Paper of the terms 'data', 'information', and just now also 'knowledge'. For the purposes here, I would like you to conceptualise 'knowledge' as being something that is internal to our beings: we increase our knowledge by acquiring 'information'. 'Information', I am roughly defining as 'data' which has been given context. Museum people spend a lot of time taking the data inherent in the objects in their collections and explaining the data's significance so that the resulting information will, they trust, increase the knowledge of those who visit their particular museums: to peruse exhibitions, attend educational sessions, use information services, and so on.

That is generally what happens in one real specific institution. In the environment of the Internet and the World Wide Web where any number of institutions can be linked virtually, things potentially become much more diverse. For, from what we might call the 'inside world' of organisations, such as museums, 'data' and/or 'information' and/or 'knowledge' and/or even 'wisdom' might be made available via the organisations' Web Sites. The 'knowledge' of the museum's curators - including their tacit knowledge - might for instance be tapped into by customers via the museum offering an enquiry answering service accessible via its Web Site. At specified times the 'wisdom' of individual curators might be displayed live online via some sort of video-conferencing facility.

I am focussing here, however, and for the remainder of this Paper, on just the constructs 'data' and 'information'. What the Internet and the Web have done is to encourage all organisations to mount on their Web Sites relatively raw unadorned 'data': numeric measurements, textual compilations, graphic images, dynamic representations, sonic sequences: with the result that those who interact with the totality of such Web Sites now potentially have access to a diversity of data far, far wider than is available within any one museum: even within a relatively large museum such as my own.

But for that data cornucopia to be converted into information which can increase knowledge, someone, somewhere, has to 'add value'. Such value-adding, clearly, can take place in one or a combination of three places. The data providers themselves can decide to try to offer information instead of, or in addition to, offering data. The 'outside world' customers might alternatively feel competent enough to add the value themselves: "Just give me the data: I will do the rest". Third, one or more intermediating bodies in 'the outside world' can take it upon themselves to convert the data into information.

It is not then difficult to conceive that, just as we have had appear on the Net a diversity of - in this context - biodiversity data collections, so we might start to have a diversity of biodiversity information collections. I do not feel the need here to try to define exactly what we might take to be the difference between biodiversity 'data' collections, and biodiversity 'information' collections: I have already suggested one possible parameter we might use to differentiate the two: the degree of 'context' provided to the 'data' for its customers. As a fan of the precepts of the philosopher Immanuel Kant who I believe stressed not only that human beings can never know exactly what the real world comprises, but also that each of our individual perceptions of that world are, indeed, 'individual', I am sure that I would find it very easy to find examples of one person's 'data' being another person's 'information'. The two words are simply used to capture the notion familiar to all who work in museums: that we will almost always need to contextualise unadorned 'data' or, more generally, to add value to it, if that data is to increase the 'knowledge' of the people - the customers - for whom it should ultimately be destined.

If now I list just some of the organisations who are intent upon adding value to the specimen and other data held within my own natural history museum so as to generate information for their customers, the idea that the Internet has begun to spawn a diversity of biodiversity 'information' providers, alongside the already well-established diversity of biodiversity 'data' providers begins to look only too true. These are all networks with whom the museum is doing business; or is seriously talking about doing business:

- The National Biodiversity Network aims to gather biodiversity data from all manner of stakeholders within the political jurisdiction which is the United Kingdom. The Natural History Museum will be providing the master species list for that network.
- The 24 Hour Museum aims to be a portal to data about and contained within all those types of institutions which are museums. Our museum is of course referenced on that portal and some of its members helped to set it up.
- GLOBIS is projected to be a global electronic catalogue of all the world's butterflies. The museum is one of the network's lead partners.
- CETAF is the Consortium of European Taxonomic Facilities and was formed especially to be a body which would represent large museums, botanical gardens, and similar bodies with reference to the European Commission and its funding streams. CETAF is slated shortly to receive a grant under the EU Framework V programme with my museum as the lead partner.
- The UK Public Record Office - the depository for national official archives - has a collaborative Project 2000 which is aiming to provide unified access to archival material and in which The Natural History Museum is participating.
- AMICO you are now familiar with: the focus here is a particular format of resource - images - and we would like to see the museum's images represented there.
- Finally, potentially dwarfing all those is an initiative of the Organisation for Economic Cooperation and Development (OECD). This is GBIF, the Global Biodiversity Information Facility, which in June was given the green light to receive several million dollars. The museum is closely involved with a number of aspects of the planning for GBIF.

And as if there is not enough going on in the outside world to provide customers with the biodiversity information they need, The Natural History Museum itself is working to become a major player in the biodiversity information market. Some of the processes of value-adding to data with which we are involved are:

- First, the choice of literature to be indexed in our bibliographic databases inherent in the Museum library's collection development policy.
- Second, the summaries tailored to customer-needs which will be created from our collections of digital scientific data.
- Third, the use of Dublin Core records to provide a cross-database searchable system: a system searchable both at the individual item level, and at the less granular collection of items level.
- Fourth, the use of authority lists of terms and consistent subject classifications across the different types of digital artefact: a feature absolutely critical for inter-operability.
- Fifth, the creation of a 'natural world' Web portal we have provisionally called RING: 'Resources in Nature Gateway.

In the terminology I touched on at the outset of this Paper, virtually the whole of this architecture comprises resource descriptions at various levels of granularity: many of the descriptions are, of course, descriptions of descriptions.

Faced with these developments, the rather fundamental questions which then seem to need asking are:

- Why The Natural History Museum? Surely it would be better if we left the value-adding to others, such as the various networks I mentioned a moment ago, and concentrated our limited internal resources on capturing and making available our frequently unique raw data for others then to add value to.

- If others are to add value, with what lens would their customers wish to approach the resulting sets of information. By planning to provide a 'natural world' portal RING, we presumably believe in the Museum that there is something special or appealing about such a concept. But is that true? Do customers go around wanting 'natural world' information in the way that they go around wanting 'art' information? Why the natural world?

- Even if we can convince ourselves that there is a market for information about biodiversity or, more generally, the natural world: Why not the private sector? The natural world data providers would then concentrate on just that - the provision of data, leaving commercial operations to add value, and provide information.

- Maybe, however, there are genuine customer needs for biodiversity information which would not in the event be satisfied by the private sector. There would then indeed be a role for non-profit making institutions - such as my own Museum - to fill the gaps in the marketplace. But given that competitive market forces - or at least the need primarily to survive and prosper by earning income from the sale of products and services - is by definition not the dominant driver of each of those non-profit making institutions' existence, who decides which of the institutions should fulfil such a value-adding information generating role, and which institutions not?

- Faced with that question, perhaps we should ask - rather late in the piece: What do customers need? The overriding answer to that question I would give is that customers need systems which are **trustworthy**. And by that I do not just mean trustworthy at the 'micro' level of the individual biodiversity information provider: important and difficult as that is to be able to achieve. I also mean at the 'macro' level of the system - if we can call it that - which decides who does what in the non-profit making part of the economy.

As Lorcan Dempsey of UKOLN and others have argued persuasively in a recent important report to the European Commission, the virtual society we are moving into - potentially bringing seamlessly together for instance, data and information from museums, libraries and archives located worldwide - has thrown up the need for a new type of institutional landscape. Perhaps the key decision is what in the overall evolving global information architecture is perceived will in the future be public sector resource supported, and what left to the private sector. (Obviously the dividing line will shift from time to time; and clearly, as we have noted, there will be 'mixed-economy' operations.) The types of intervention possible and appropriate from national and cross-national bodies such as the OECD and the potential effectiveness of such intervention will all be strongly influenced by the public or private nature of the type of organisational entity one is dealing with.

I must say that personally - and this may be controversial - I start from the notion that the generation and distribution of publicly available information is wholly a private sector business unless there are good reasons for public subsidy. That is, I believe that one should not start from what will inevitably on examination turn out to be rather vague notions about access to all information in all locations being available to all free of charge: and then wonder how ever the information supply and demand functions are going to be managed. One starts by assuming a commercial information market-place: and then asks what intervention is needed from the public sector to achieve the information and communication goals of society. Such a stance seems to work well with print 'libraries' - whose artifacts for the most part all start life in the public arena as commercially financed 'publications'. Government then directly or indirectly finances public sector libraries so that they are able to buy and make accessible to their customers free of direct charge a proportion of these commercially produced publications.

However, such a stance clearly does not work so well with (public sector) museums and archives: who generally start from the premise of being public sector funded - and then try to make ends meet by charging for certain services for certain people. So there is already a potential conflict of perspective when one brings digital content from the three types of institution together even before one adds in the fact that the artifacts in museums and galleries that are 'digitised' will be unique; those from libraries will almost always be copies of each other. Thus libraries might compete with each other for attention to their items; museums and archives more often than not will not need to.

Irrespective of how we resolve those differences of legacy and philosophy amongst us all, some way has to be worked out which will avoid more and more biodiversity information providers being funded by the world's public sectors ultimately to fulfil identical or at least seriously overlapping customer needs for such information. The Natural History Museum has been commissioned by the UK Higher Education sector to contribute descriptions of Web-based resources to a comprehensive Resource Discovery Network. The specific service within that Network which we will be involved with - called a 'hub' - is named BIOME: our gateway part of BIOME will be called Natural Selection - a nice designation I think dreamt up by the Project Manager for BIOME within the Museum, Anne Freeman.

But of course there are already many other hubs, gateways, portals, etc. servicing one or another aspect of natural world information. A number of these services happened to meet recently under the auspices of an embryonic organisation IMesh: but many of us felt that we did not get as far as we would have liked in trying to figure out how - internationally - we were doing to avoid unnecessary diversity: principally I believe because we could not decide whether we were all intent on cooperating with each other; or competing with each other! For a number of organisations represented at the meeting, if they stopped being a subject gateway, they would stop existing as an organisation - with all that implies for job security.

A wider and more fundamental type of challenge faces the last initiative I will mention: the UK Interoperability Focus - whose Web Site can be accessed via the UKOLN site. I happen to chair the Focus's Advisory Committee: and the problem there, again, is the great number and diversity of overlapping initiatives: many of them funded by different agencies within UK Government: but with none of us - certainly none of us on the Advisory Committee - being in a position to say: 'Don't you think it would be better if your organisation stopped trying to do that, and left these other people to get on with it'.

However, it is early days: we will, for instance, soon have in being in the UK the newly formed Museums, Libraries and Archives Council which Matthew Evans will chair. There are now so many people talking about the proliferation and consequent increasing diversity of biodiversity and other information providers on the Net, all funded by one or another element of the public sector, that I am optimistic that - somehow - we will find a way other than leaving it all to the private sector to ensure that 'the diversity of biodiversity information is NOT inevitable'!