Information infrastructure for global biological networks

Meredith A Lane

Global Biodiversity Information Facility Secretariat
Tel: (45) 3532 1484
Fax: (45) 3532 1480
E-mail: mlane@gbif.org

Any sort of global network of research facilities, in order to maintain a truly 'network' status, requires a mechanism for sharing information among the participating centres. The more immediate, transparent and effective the mechanism is, the more useful the research network is, not only to its participants but to the community at large.

In this digital age, information sharing is of course via the Internet. However, for data and information as complex as that found in the biological research domain, the Internet alone (e-mail, file transfers) is not sufficient. What is needed is interoperability among the databases maintained at each centre, and transparent means of querying multiple databases and representing query results to users anywhere in the world. This requires an information infrastructure that has several component layers.

Important among these components are community-agreed standards for data and metadata, data schemas, transfer protocols, networked collection(s) of controlled vocabulary terms (thesauri), and single (or few) access points from which users can retrieve data from a wide variety of sources (portals). If these are utilised together in an information architecture based on 'web services', or extensible markup language (XML), great benefit is perceived by both internal (researchers at the centres) and external (other researchers, the public) users.

An additional benefit of such an information architecture, constructed for use by one sort of network, is that the same construct can be accessed and used by other networks as well. Thus there is a great saving by avoiding duplication of effort. Such information architectures are also modular, so that each network, while utilising the main components in common, may add modules that suit its own specific needs. Likewise, this type of architecture is expandable, so that as new data and metadata standards are developed for various data types, the network(s) can present ever-richer arrays of data and information.

The Global Biodiversity Information Facility (GBIF) has built a web services based information architecture for the initial purpose of sharing specimen and observational primary biodiversity data [1]. Important components of the architecture include the Darwin Core (DwC) data schema, together with the Distributed Generic Information Retrieval (DiGIR) data sharing protocol, or the Access to Biological Collection (ABCD) data schema, together with the Biological Collection Access Service (BioCASe) protocol. These have been in use for several years, and soon it will be possible to access either DwC or ABCD data using the 'next generation' Taxonomic Databases Working Group (TDWG) Access Protocol for Information Retrieval (TAPIR).

In 2003, a workshop entitled *"Towards a global infrastructure for microbial information"* was cosponsored by GBIF, the World Federation for Culture Collections (WFCC) and three Belgian agencies. At the time, the minimal data set for BRCs was not entirely reflected in either the DwC or ABCD schemas. Since then, with the help of the microbial community, microbial extensions have been developed for schemas [2] (Figure 1).

In the meantime, the World Data Center for Microorganisms (WDCM), CABI Bioscience, Japan's National Institute of Genetics, the Belgian Node of GBIF and others began sharing microbial culture data using the GBIF information architecture. The contributions of several of these have added not only (culture) specimen data to the information available to the world, but have also helped to build the microbial portion of the *Electronic Catalogue of Names of Known Organisms (ECAT)*. This catalogue provides a taxonomic names service to the GBIF information architecture that is essential to searching the Internet for data on biological organisms and materials.

Figure 2 provides a simplified graphical overview of the GBIF information architecture. The pyramidal shape should not be interpreted as exclusionary in any way. Any number of networks, for example the Global Biological Resource Centres Network (GBRCN), can tap into its functionality at several levels. It is possible to use the whole architecture by building a portal that is designed for a specific purpose which would tap into the data index and the UDDI registry. It is also possible to build a different registry and index against lower levels of the architecture, and then specialised portal(s) against those. The system is very flexible and scalable.

GBIF is working toward adding more functionalities to the information architecture in the near future. Of particular importance will be globally unique identifiers (GUIDs), because these will improve the capability of the system to:

- Maintain and track connections between related data items even when they are served by different data

providers.

- Assist in the detection of duplicate data items in the system.
- Help in making sure that data providers are given appropriate attribution by users of those data.

Additional data types are also of importance, and some of the most in-demand of these are images and tools for the identification of organisms. There are TDWG working groups for both data types. Progress has also been made in developing best-practices for digital imaging [3], and there are examples of web-based identification tools for microorganisms already functioning on the web [4].

Also, in the next few years, GBIF will be working with the molecular and ecological informatics communities to make searching across all levels of biological organisation possible. It will facilitate linking into digital libraries not only of literature but also of various media types (Figure 3). In undertaking these activities, it will work with TDWG and other partners, as well as the various research and user communities. The vision is to facilitate an information space that allows the generation of 'species pages' from any and all data that are available on the web about a particular organism, and do it 'on the fly'.

From activities like this, which will draw research communities together, will grow a global network of networks that each can function well within their own domains, but that will also function synergistically, and thus be of great service to science and society.

Reference

1. www.gbif.net
2. http://digir.net/schema/conceptual/darwin/extension/microbial/0.1/microbial.xsd
3. Fritze D. Digital imaging of prokaryotes for taxonomic purposes. In: Häuser *et al.* (Eds). *Digital Imaging of Biological Type Specimens. A Manual of Best Practice.* Results from a study of the European Network for Biodiversity Information, Stuttgart 2005, p.153-171.
4. See, for just one example, http://www.cbs.knaw.nl/yeast/(arpfokyzpdoebb45kcqo0jav)/BioloMICSID.aspx

Figure 1. Representation of the microbial extension to the ABCD data schema.

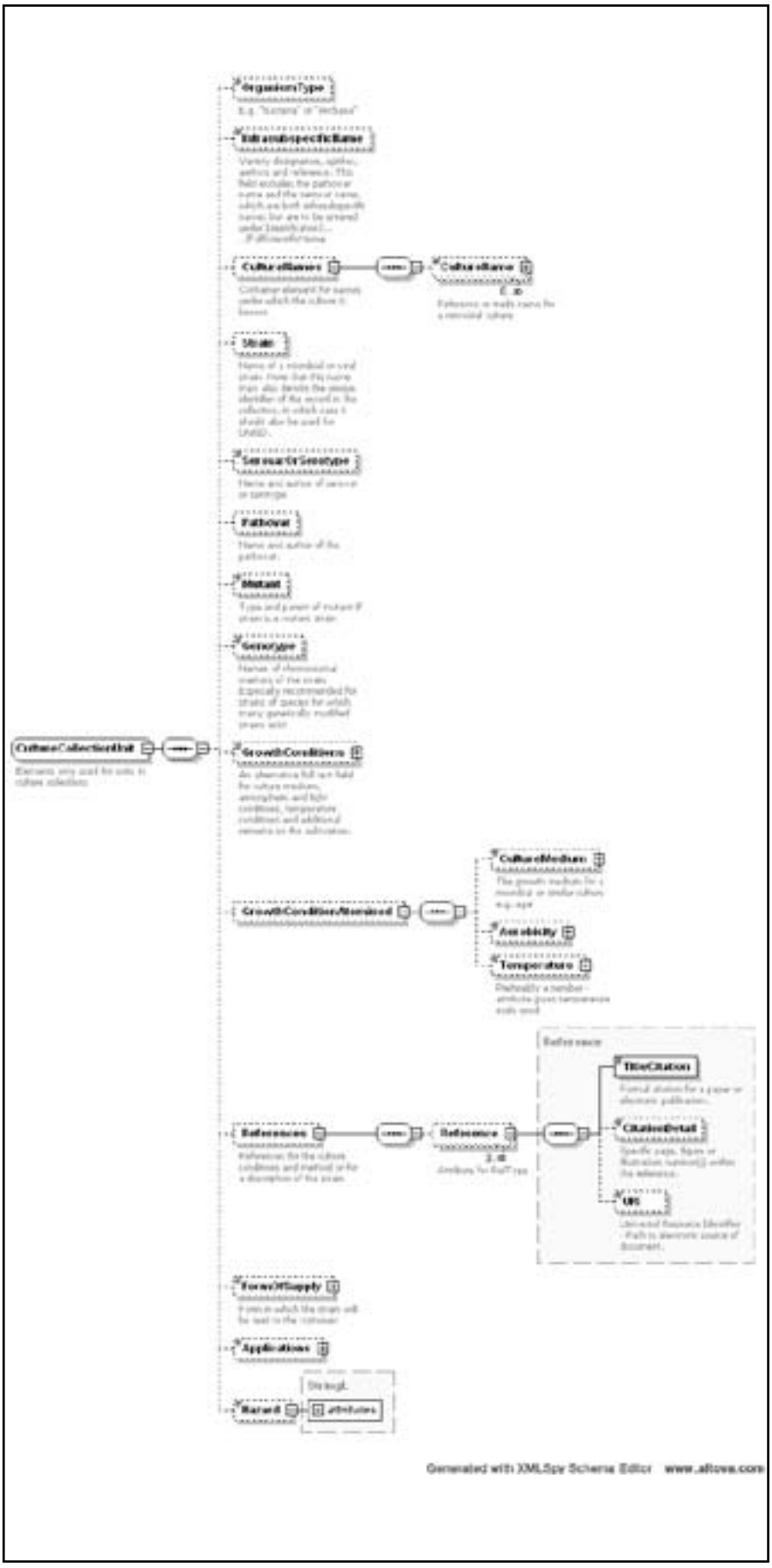

Figure 2. Overview of the GBIF information infrastructure. GBIF utilises protocols common to the entire Internet, such as HTTP, XML and web features service. It has also adopted standards and protocols for biodiversity data, including the ABCD and DwC schemas and the BioCASe and DiGIR protocols. In addition, GBIF is playing a role in the development of additional standards (indicated in blue), as well as leading the way toward adoption of GUIDs for biodiversity data objects. The GBIF UDDI service registry and data index are infrastructures that allow the GBIF data portal or any internet portal that uses these standards and protocols to access the data that GBIF makes available.

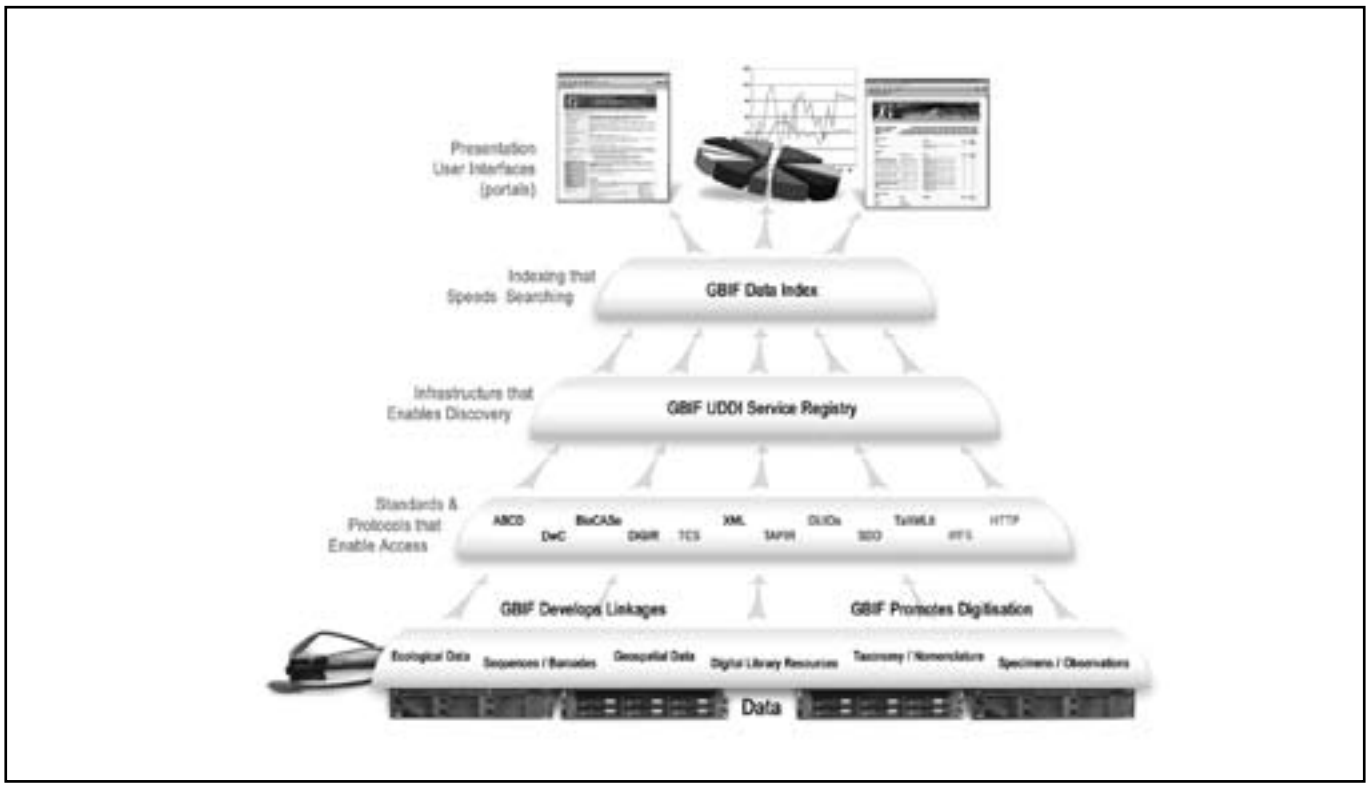

Figure 3. GBIF's information architecture will link together data from all levels of biodiversity and with other digital resources. The key to modern biological information is the scientific names of organisms, and the electronic catalogue of the names that GBIF is building is fundamental to searching within and among the types of data indicated. The linkages are completed by the registry and index that GBIF provides, because these are accessible to search engines from other domains.