Project no. 018340

Project acronym: EDIT

Project title: Toward the European Distributed Institute of Taxonomy

Instrument: Network of Excellence

Thematic Priority: Sub-Priority 1.1.6.3: "Global Change and Ecosystems"

C5.90 Inclusion of missing data items in GBIF index

Due date of component: Month 27
Actual submission date: Month 27

Start date of project: 01/03/2006 Duration: 5 years

Organisation name of lead contractor for this component: 9 BGBM

Revision: Draft 1

Project co-funded by the European Commission within the Sixth Framework Programme (2002-2006)		
Dissemination Level		
PU	Public	x
PP	Restricted to other programme participants (including the Commission Services)	
RE	Restricted to a group specified by the consortium (including the Commission Services)	
CO	Confidential, only for members of the consortium (including the Commission Services)	

C5.90 Inclusion of missing data items in GBIF index

The Global Biodiversity Information Facility (GBIF) was established to make biodiversity data openly and universally available. One of the products of GBIF is a central index database that stores a limited number of core data elements.

Because original biodiversity data remain with the data providers, the data not stored in the index database can be viewed but not be searched. Filtering and search functionality in user interfaces (GBIF portal, http://data.gbif.org) and other specialized portals set up on the GBIF infrastructure (e.g. BioCASE portal http://search.biocase.org, EDIT portal), rely on the data stored in the GBIF index database.

This component is to document which data items are found missing in the GBIF index (June 2008) from the user perspective of taxonomists.

This evaluation is based on the results of component C5.89 "List of data items and searchable data items in existing portals" (Activity 5.9.1: Define specimen access interface for taxonomists) and on interviews with taxonomists of different fields (botany, mycology, zoology, phycology) aimed at getting an overview of taxonomist's needs. The taxonomists weighted the categories listed in C5.89 as high priority (= data item is a key to taxonomic information and must be searchable) or priority (data item is identified as important and must be displayed independently of the provider accessibility). The resulting priority list contains data items which need to be stored in the GBIF index database. Other data items should be shown when the original data record is requested, for example by displaying the original data of the data provider (XML-file or text by means of style sheet transformation).

Items in the GBIF index database

In many cases the GBIF index database fits the taxonomist's demands (in descending order):
- high priority: Scientific name incl. infraspecific scientific names, country, nomenclatural type, collector, date collected, higher rank, accession number
- priority: region, elevation, image respectively multi media content, collector's number, identified by, field number, common or vernacular name, type status
- further: sex, life stage, preparation etc.

Items missing in the GBIF index database

In some cases the taxonomist's demands are currently not supported by the GBIF index database (descending order).
Institution (see below, high priority), collection (see below, high priority), barcode no. (high priority), Genbank (NCBI) no. (priority), Exsiccatae (priority), host of parasite (priority).

1. Institution and collection are two data types essential in the daily work of taxonomists which are not searchable yet. The name of the institution means the herbarium, natural history collection etc. were the physical specimen is housed. Bigger institutions hold mostly several collections (e.g. separate collections of various higher taxa). As long as the data resource is a collection database and the data provider is the institution holding the physical specimens, institution and collection are represented in the GBIF index database. However, in many cases specimen data are provided by institutions other than the institution/collection housing the specimens, e.g. B (abbreviations explained below) is providing specimen data of Desmids compiled by researchers from HBG but housed at M; M is providing lichen data of B; B is

providing diatom specimen data of BHUPM and BRM etc. (B = Botanischer Garten und Botanisches Museum Berlin-Dahlem, Zentraleinrichtung der Freien Universität Berlin, HBG = Biozentrum Klein-Flottbek, M = Botanische Staatssammlung München, BHUPM = Museum für Naturkunde Berlin, BRM = Alfred-Wegener-Institut für Polar- und Meeresforschung). The GBIF index is excellent in tracing back records to a single data resource (called collection) or data provider (called institution) but the GBIF index lacks data on place where the physical specimens can be found. Two additional fields for storing the Institution name and Collection name (independent of the data providers and data collection names) are recommended.

2. A further item is the so called barcode number, this is a current number given by a herbarium/nature history collection, often including a collection code and a number (this may or may actually be a barcode; the decisive feature is that it is physically connected to the specimen). In some institutions this modern number differs from accession numbers (which are in the GBIF index). Publications often refer to the barcode or accession numbers; hence this information is needed by taxonomists.

3. The Genbank (NCBI) number is an identifier which can be linked to a specimen in a collection. One to many Genbank numbers can refer to a single specimen. At least for some taxonomists this information has high priority, and the importance of this item is likely to rise in the future, so it is highly recommended to include Genbank (NCBI) numbers in the GBIF index.

4. "Exsiccatae" are sets of specimens from the same species and collection event which have been scrutinised by specialists before being distributed to several collections. These specimens are in many respects similar to duplicates, but they can be treated differently by different institutions (i.e. lectotypification of one single specimen etc.). The storage of Exsiccatae information in the GBIF index could facilitate the usability for taxonomists.

5. At least for some researchers working on parasites the knowledge of the host species is essential (this is implemented in single data bases). Similar specimen-to-species relationships represented in specimen data include pollination, predation, feeding, etc. This kind of information is of high relevance also for ecological research but not supported by the GBIF index.

The data items under point 2-5 are to some degree already supported by the ABCD standard data definition but would need to be defined in an extension to the Darwin Core. Items 2 and 3 require the possibility for multiple entries for a single specimen.