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”

C5.096 Report on WP7 data management

Due date of deliverable: Month 40
Actual submission date: Month 41

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

Organisation name of lead contractor for this deliverable: **MfN with BGBM**

Revision [end version]

Project co-funded by the European Commission within the Sixth Framework Programme (2002-2006)		
Dissemination Level (“X” in the relevant box)		
PU	Public	
PP	Restricted to other programme participants (including the Commission Services)	X
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)	

1. Introduction

One task of work package 7 is the establishment of “All Taxa Biodiversity Inventories + Monitoring” (ATBI+M) sites in protected areas. ATBI+M are large-scale efforts to record, identify, and document the entire biodiversity of a given area. EDIT's ATBI+M sites differ from traditional approaches in their longer-term orientation: from an initial species inventory, they will form the basis for monitoring biodiversity changes over time.

Active European ATBI+M sites are located in France, Italy and in Slovakia. More than 4,500 species including 2,700 insects have been recorded by more than 100 scientists so far. The data generated by EDIT's ATBI+M’s are also accessible world-wide through the Global Biodiversity Information Facility (GBIF).

2. Workflow

All scientists collecting data in the field have to fill in special Excel sheets developed for the data recording in ATBI sites. The predefined Excel sheets consist of three spread sheets for data of collected specimen, observations and publications. The first spread sheet is for localities, the second for events and the third for taxonomic data. Each scientist has to complete all three spread sheets and no other format for the delivered data is accepted. The delivered data is checked for completeness by WP7 staff and each Excel file gets a unique package ID. The data in the Excel file are also checked for correct values of longitude and latitude and the fields for date and time are checked for correct format. Fields for number values (minimum and maximum altitude, radius, accuracy) are also verified for correct content.

After this first check the data of these three spread sheets are imported into the central Access database (Fig. 1). First the data of the locality spread sheet, then data from the event spread sheet and at last the taxonomy spread sheet data are imported. Error messages will advise of mistakes if field types are not consistent with the database. These errors have to be corrected in the Excel-file before importing the data another time. Importing wrong or missing keys between the tables ‘locality, event and taxonomy’ into the relational database will result again in an error message. In the majority of cases these errors have to be corrected by the data deliverer.

After the clean data is imported into the Access database some additional fields have to be completed: in the locality table one field for a standardized name of each national park and in the event table one field for a standardized name of each collecting method.

The last step is to assign the species to an authorized species list. This is necessary to eliminate especially typing errors of species names and to assure that one species occurs only once in the species list.

3. Images

Each scientist can provide images for the delivered records. Images are also used for the species sheets on the ATBI website to illustrate the species information. The data for all images are stored in one central table, without any relational connection to any other table.

4. Export

The data are exported periodical to
BioCASE (http://ww3.bgbm.org/biocase/querytool/main.cgi?dsa=EDIT_ATBI),
GBIF (http://data.gbif.org/datasets/provider/236) and
Cardobs (http://inpn.mnhn.fr/isb/index.jsp).

For BioCASE and GBIF the data of the tables’ locality, event and taxonomy are combined to one flat Excel table. Only the taxonomic data – higher taxonomy and species data – is taken from the table ref_species to ensure that the correct higher taxonomy and the correct species names

will be shown on the external webpages. The data of the pictures are exported to a separate Excel file.
For the atbi-websites (www.atbi.eu/gemer and www.atbi.eu/mercantour-marittime) all Access tables are converted to MySQL files in the same structure as in the central Access database.
The export for Cardobs combines the two tables locality and event to one flat Excel spread sheet and the taxonomy, with species name and higher taxonomy from ref_species, is exported to another Excel spread sheet.
To export the data to EDITs CDM (Common Data Model) a transformation to an xml file is necessary. The format of the xml file is based on the ABCD Schema (http://www.bgbm.org/tdwg/codata/schema/ABCD_2.06/HTML/ABCD_2.06.html).

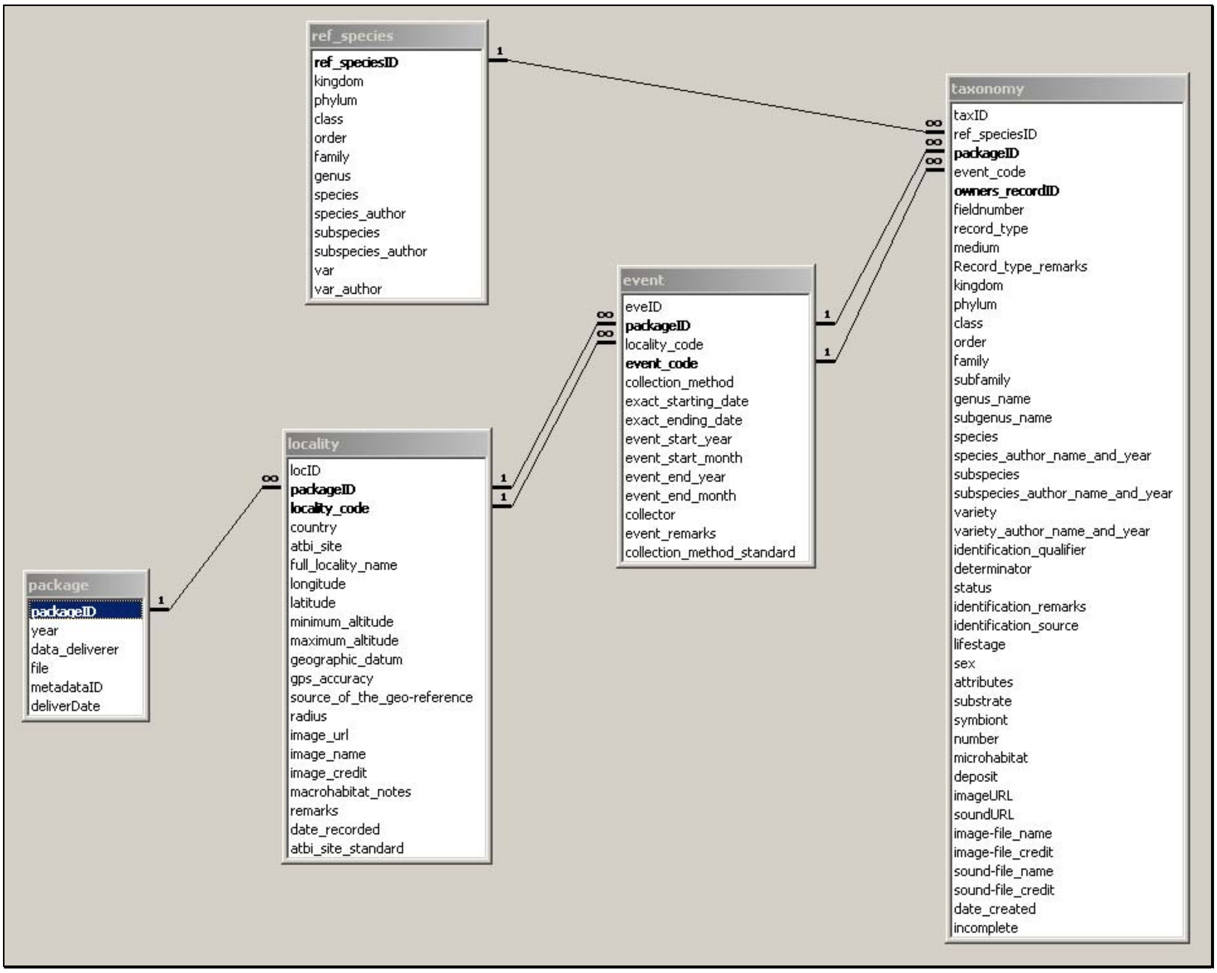

Figure 1: Entity-relationship diagram of the central Access database with five relevant tables (package, locality, event, taxonomy and ref_species)

5. Problems of delivered data

Delivered data has to be mostly corrected in regard to the georeferenced values and fields of the type number. The georeferenced values have to be converted into decimal values, in case this format has not been used. Many scientists enter for example in the field for the altitude not only the value, but also in addition "meter" or "m". The fields for date and time are also often not in the format that they should be delivered.
Another major problem for data deliverers seems to be the connections between the tables' locality/event and event/taxonomy. The connections are the fields "Locality code" and "Event code". Each "Locality code" in the table Event has to have one entry in the table Locality. If one

"Locality code" is misspelled or missing, it is not possible to import the table with the events. The same applies for the "Event code". In most cases these errors have to be corrected by the data deliverer.

6. Problems of a uniform species list

Most problems with the species names are connected to typing errors. Typing errors happens not only between different data deliverers but also within one data delivery from one scientist. Also different writing of the author name of one species is a source for duplicate species. Sometimes the author's name is abbreviated or the year of publication is different or the name and the year are separated with or without a comma. To get a consistent species list for presentations it is necessary to have a uniform species list and the species name of each new record has to be linked to one name of this uniform species list. Most of the correlation between the species name of the data deliverers and the uniform species list can be done automatically. However, new species for the ATBI sites and typing errors have to be checked manually and the link to a uniform species name has to be set by hand as well.