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"

D20 – Version 2 of the EDIT Platform for Cybertaxonomy

Due date of deliverable: Month 42
Actual submission date: Month 42

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

Organisation name of lead contractor for this deliverable: 9 FUB-BGBM

Revision: Final

Project co-funded by the European Commission within the Sixth Framework Programme (2002-2006)		
Dissemination Level ("X" in the relevant box)		
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)	

1 Introduction

This deliverable presents a summary of the developments and results achieved with regard to the development of the „Internet Platform for Cybertaxonomy", a distributed computing platform assisting taxonomists in their daily work producing taxonomic revisions, checklists, inventories, etc. and making the results of the work accessible to others. The basic design was described in Deliverable 7 (month 18, September 2007), the first results of the implementation were presented in Deliverable 11 (D11, month 32, November 2008).

The document presented here will defer most technical details to the appended project component documents (underlined text). The present text attempts to summarise as far as possible. However, on the other hand the text should refrain from purely listing the components, it should provide an overview. The complexity of the endeavour and the amount of resources directly related to D20 once more justifies a deliverable document of some length: 34 project "components" and milestones delivered by 10 EDIT partners are included in D20.

Section 2 of the D11 document outlined the phases that led to Platform Version 1 (with months 0-11 occupied by preparation and definition activities; m12-17 providing the basic technical design, and m15-32 the first phase of the implementation). It stated that the "main purpose of Platform Version 1 [was] to demonstrate capabilities, not full usage". Platform Version 2 now represents the outcome of the development activities in the past 9 months, with key applications and developments now reaching production state.

For the continued development, testing and assessment of the EDIT Common Data Model (CDM) an the related software the availability of taxonomic data sets was essential (see C5.076 Taxonomic sample data available), as was the direct communication with the taxonomists responsible for such resources (see M5.32 Three significant taxonomic initiatives setup as CDM projects). Eight taxonomic datasets were used to test the real-world compliance of the CDM.

2 Software components and development

The software products and developments are presented here under headings representing the major components developed in WP5. Sections are related, but not restricted, to actual activities as defined in the WP5 workplan. Within the sections, a brief introduction (often with reference to more in-depth explanations given in D11) is followed by a description of developments and decision making processes since D11, a description of the current state, highlighting existing and/or imminent products, and a concluding outlook with respect to the planned development until the next deliverable ("Platform version 3").

The basic software equipment needed to carry out taxonomic work with the EDIT platform is a Common Data Model data store together with the desktop taxonomic editor and basic import and export functionalities, which can be used to set up the system with existing standardized data. We have bundled these functionalities within the CDM desktop taxonomic editor (M5.33 CDM Toolkit Bundle (data store, editor, conversion tools) available and documented to be used by EDIT partner institutions and beyond). The editor contains already a functional CDM store so that taxonomic work can be started immediately after installation.

2.1 Platform access: the EDIT Cybergate

The EDIT Cybergate provides access to the entire EDIT Internet Platform for Cybertaxonomy, as well as its underlying technologies, with a single graphical user access interface. Initially delivered as an experimental application following the need to have an overview of Platform developments and products, the Cybergate has proven to be an effective and stable tool.

The Cybergate is accessible under http://dev.e-taxonomy.eu/platform and will be further improved with respect to its information content and number of linked applications..

2.2 Basic technological architecture: CDM and Java Library

The EDIT Common Datamodel (CDM) and the CDM Library provide programmers with a set of functionalities that can be used in diverse applications related to taxonomic data. (see section 3.2 of D11 for details). A revised CDM was delivered with M5.22 Revised CDM frozen and documented for Platform release 2 in March 2009.

With most of the functionality laid out and prototyped in Platform v.1, implementation work focussed on 3 areas: (i) Bugfixing and adaptation of library functionalities to serve the emerging applications such as the taxonomic editor and the power user interface. (ii) The adaptation and further implementation of the EDIT web services. The services, built upon the REST (Representational state transfer) architectural design, provide a simple yet powerful way to collect data from a community store. The data is exposed to the web by a set of RESTful web services which expose each entity of information through a stable and unique URI (C5.059 CDM2 Community store interfaces documented). (iii) The implementation of data interfaces to and from the CDM-store. The Common Data Model (CDM) acts as the central "information broker" within EDIT's platform for cybertaxonomy architecture. Apart from its role as an advanced storage facility for individual researcher and taxonomic communities it has to communicate with decoupled EDIT platform components such as data portals, (print) publication tools, fieldwork and mapping software, as well as collection databases and specimen information systems. Consequently, the implementation of CDM interfaces providing and accepting data according to established international biodiversity informatics data standards and quasi-standards is considered essential for the entire EDIT-based taxonomic information workflow and a number of data interfaces are available (C5.077 CDM data interface components available).

Apart from these areas of activities, the implementation of a versioning mechanism, support for LSIDs (Life Science Identifiers), user management and setting up a test frameworks are major new features of version 2.2 of the CDM library (C5.058 CDM2 Framework implementation available).

2.3 Output: Data Portal and print publication services

The **CDM Data Portal** is the central web output mechanism provided as a Platform component. It demonstrates the data output capabilities of the EDIT Platform, as well as the integration of CDM functionality into the Drupal Content Management System. However, the Platform design allows developers to use the web services and data interfaces to construct their own specialised web interfaces independent of Drupal and the data portal software.

The EDIT CDM Data Portal has been further developed into a mature product following user requirements from the EDIT exemplar groups (C5.126 Data portal update for Platform version 2 ready). Enhancements include: improved stability and documentation; increased functionality of the web services used to connect Data Portals to the CDM Community Store; implementation of multiple taxonomic views resulting in improved search functionality and taxonomic tree browsing; advanced search functions for taxa and synonyms as well as for geographic areas, with thumbnail image display in the results; dynamic distribution maps, which allow zooming and panning and optional display of legends; significant increase in the amount and granularity of data displayed; and an improved render mechanism for scientific names and references, which now is configurable to meet many different tastes.

Currently, the data portal is in use by the 3 EDIT Exemplar Groups of Work Package 6. To extend the user base, the software will next be used to implement the web interfaces of several checklist projects, most notably the PESI Euro+Med Plantbase and Fauna Europaea databases.

Printed publication output is still the most important result of the work of a taxonomist with respect to public recognition, professional credit, and official acceptance of newly published taxa. The Platform will therefore provide of print publishing support. The concept developed for this has been described in C5.088 Functional description for service to generate printed publication from the CDM. In essence, the CDM library will provide functions to output the data into an OpenDocument format (ODF) document. ODF is an ISO-certified open standard that has been adopted worldwide. Modern word processors or desktop publishing systems such as those included in OpenOffice or MS Office can be used to import and further edit the document, either directly or in an automated form by means of their macro facilities. Alternatively, standard output document formats conforming to the editorial rules of specific publication series can be programmed, which allows direct output as a print-ready formatted document.

A prototype of the service will be set up by the end of 2009, with full documentation becoming available by June and a number of pre-defined output formats by July 2010.

2.4 Input: The EDIT Taxonomic Editor

The EDIT Taxonomic Editor is an innovative data input and editing tool for the taxonomic data held in CDM data stores. Conceptual details are described in section 3.4 of D11. Version 1.0 already covered the basic functionality necessary for entering and editing taxonomic treatments, but it was still somewhat wanting in performance and in functionality. Version 2 (C5.078 Fully functional version of the CDM editor available) now offers all necessary functions for creating taxonomic trees, organizing synonymies and name relations, and maintaining all kinds off factual data, as required by the user; e.g. (textual) taxon descriptions, distribution information, common names, conservation status, economic use, etc. Apart from the necessary performance improvements and technical issues, Version 2 includes the following new functionalities:

- Extensive setting of user preferences (e.g. nomenclatural code, feature categories, name relationships and ranks to be used).
- Selecting and connecting data sources (CDM stores)
- Automated software updates
- Entering and organizing concept relations
- Search functions (taxa, synonyms, names, common names)
- Displaying distribution maps
- Importing and exporting data in various formats
- Bulk editing of references and names (as part of the Power User Interface, see below)
- De-duplication of references and names (ditto)
- User management (create, edit, and delete user accounts).

The latest version of the Editor runs on Microsoft Windows XP and Vista, as well as Mac OS 10. The software comes with a pre-installed H2 Java Database Engine for local use but can also be configured to communicate with a server database.

When using electronic databases fed into by several people, it is common to have a person in an editorial role with extended right to “massage” the data in the database. In short, this “power user” is a non-technical user responsible for the integrity and cleaning of the data. A functional prototype of the “Power User Interface” was integrated with the Editor for users with such a role (C5.113 CDM power user interface design study and pilot implementation). Its functionality currently includes spreadsheet-like display of data with sorting options and duplicate removal for scientific names and references (this can be extended to other data types). Forthcoming is the integration of an import functionality by parsing text pasted directly into the tabular interface.

For the Editor and Power User Interface, the emphasis in development efforts towards Platform Version 3 will be on implementing interface and functionality improvements demanded by users as a result of thee software testing and promotion activities.

2.5 Tools: The Geo-Platform

Progress in the development of Platform components related to geographic information processing in taxonomy mainly consisted of functional improvements of the basic tools and services already described in section 3.5 of D11. The development can be summarised as follows:

The **EDIT Mapviewer** is an online toolbox for taxonomists, offering data visualisation and analysis. The version available at the time of D11 had to be ported to a new software framework (OpenLayers), because the development of the one previously used (Mapbuilder) was abandoned. This also allowed to unify the development cycles for the mapViewer online tool and the REST services (see below). A number of new functions were developed for version 2 (C5.108 The EDIT MapViewer with enhanced functionality available), including:

- additional map layers
- inclusion of environmental variables
- linkage to GBIF data (C5.072 Dynamically link GBIF data to core geoplatform allowing GBIF data to be visualised in addition to custom user data)
- an upload facility for point data
- downloads of maps for printing and publication
- drawing of polygons
- spatial analysis using statistics about the data points (C5.070-2 Practical implementation of a converter webservice that transforms point occurrence data to distribution data).

The other important component of the Geoplatform are the **EDIT mapViewer REST services.** They make a selection of the mapViewer functionalities accessible as a web service for machine-to-machine communications (C5.109: The EDIT Map Web Services for occurrence and distribution data with enhanced functionality available).

Availability of the webservices made it possible to include mapViewer functionality in the existing websites, namely the EDIT Exemplar Group data portals and the websites of the ATBI/M sites (C5.060 ATBI site data linked to geoplatform). While the usability of the geo-platform for taxonomists was confirmed in principle (C5.107 Specification for the application of EDIT Geo-Software for taxonomic inventory sites), WP7 has identified issues that need to be solved in preparation of version 3 of the Geoplatform (C7.3.7 Report on the usability of the ATBI map visualization tools for ATBI+M scientists and stakeholders), among them the current restriction to less than about 300 data points. There are also a number of items on their wish list, e.g. refining the map visualization to allow users to change the area displayed on the map and zoom in and out, display all georeferenced inventory sites, and others.

Currently the Geoplatform is undergoing testing with input from several EDIT partners. It is already clear that the hardware platform for the services needs to be improved, e.g. by mirroring the server. It is also necessary to further test the application under a variety of hard- and software environments and to provide proper documentation. In the context of user testing (see section 2.11 below) the development efforts until version 3 need to focus on resolving these issues.

2.6 Tools: Descriptive components

This topic mainly includes the data structures, functionalities and tools needed for structured descriptive data, i.e. the atomised content of taxon descriptions and identification keys (see section 3.6 of D11 for more details). For D11 it had already been decided not to include input and editing, mainly because existing tool sufficiently cover the needs of taxonomist, and standard output in the Structured Descriptive Data format (SDD) allows to import data into the CDM.

An analysis of the subject area (C5.081 Use case model and functional description of CDM descriptive data editor) provided an overview of the functionalities that could possibly be reused from existing tools: output of printed keys and natural language descriptions, interactive identification, building diagnosis, building forms synthesizing descriptive data on a taxon, automated comparison of taxa, and validation of the coherence of descriptions. The report also identified a number of issues not yet covered: an inventory of descriptive resources and associated search functionalities, a tool for syntactic reading of descriptive text in natural language, the generalisation of descriptions (e.g. building taxon descriptions based on specimens descriptions), complete handling of languages, statistical analysis of descriptive data (data summary, ANOVA), and collaborative editing of descriptive data. These functionalities have been further analysed and documented in the document accompanying M5.19b Functional description for descriptive user interface.

Task 5.6 has analysed possibilities for the creation of an inventory of descriptive resources; in essence, the use of a specialised wiki site in collaboration with the Key2Nature project is suggested (C5.116 Inventory of descriptive knowledge bases). This will be followed up in collaboration with that project.

With respect to the implementation of Platform tools, the major aim is to develop CDM library functionalities that allow to the output of structured descriptive data to CDM-based applications such as the EDIT Data Portal. To that end, the C5.117 Specification of a CDM Library functionality for the output of structured descriptive content in structured and textual form, provided the base for a pilot implementation, demonstrating two core functionalities: user upload of an SDD file into a CDM store, and display of selected (or all) descriptive content contained in a CDM store. The tool consists of an HTML page including a form and a program in PHP processing the input information. The PHP code calls JAVA functions that interact with some CDM library classes (for details see C5.118 Pilot implementation of a CDM Library functionality for the output of structured descriptive content in structured and textual form).

The aim is to have a fully functional output mechanism for structured descriptive data by the release of platform v. 3. The next steps planned are to finalise the SDD-CDM2 import and export module, to select representative taxonomic groups for testing, to continue to develop the prototype towards the full implementation and to test the outcomes with the selected group.

2.7 Tools: Bibliography (ViTaL)

From the start, this activity has suffered setbacks leading to re-definition of tasks and deadlines, including severe software and personnel problems. After D11 the long standing issue concerning the commercial MetaLib application was finally solved, but then the responsible developer in the task left EDIT. In parallel, the global situation changed, not only with the international BHL project progressing in their provision of functionality, but also through the successful funding application for BHL-Europe.

The current situation calls for an intense review with respect to the complementarities of activities in these projects and with respect to the possibilities to finalise the tasks in EDIT that can contribute to the overall literature access system for taxonomists. It is clear that there are features in ViTaL which could be highly useful, in particular the (human and machine) access to all kinds of digitised resources (not only BHL documents) through a distributed catalogue search fed by the specialised taxonomic libraries of EDIT partners. This has at least partly been implemented, we are currently re-assessing the resources needed to finalise this service. The involved partners have agreed to jointly find a solution which may include a proposal to reallocate funding to EDIT members that can provide the necessary expertise and/or staff.

Although it was attempted to stick to the workplan and indeed the milestone M5.26 Soft Launch of ViTaL was delivered, we have now decided to postpone the component “C5.030 Launch of ViTaL as a functioning component of the EDIT Platform for Cybertaxonomy including system and user documentation to enable effective access” as well as the components depending on this (C5.102-C5.106) until the situation has been clarified.

2.8 Tools: Specimen access

The specimen access portal has been operational since D11. Further development including the incorporation of direct access to the search results via the taxonomic editor are features of Platform version 3 and will be incorporated in due course. We have received the formal consent from RBG Kew to incorporate the World Checklist of Plants into the query expansion features of the tool, and from Missouri Botanical Garden the consent in principle to include the data from Tropicos for the same purpose. Zoological data from the MNHN can be negotiated once the databases at that institution have been unified in a common data centre.

2.9 Technology: Security and Single Sign-on (SSO)

This activity covers the design of a security infrastructure for the platform and the introduction of a secure Community Single Sign-On (CSSO) for the Platform. The CSSO enables the various EDIT service providers to protect their services and resources defining individual access control policies, while users can access different services using only one identity. The security infrastructure is based on the international standard SAML v2 protocol family and provides a federation concept to realise the community aspect.

A number of problems have been solved that were defined while extending the scope of applications that could be incorporated (C5.085 Integrate further platform components with the CSSO security infrastructure). As a test scenario, the EDIT applications used by developers are already accessed via an SSO mechanism. Problems that were incurred with the use of the Spring Framework were solved. A solution for problems connected to the usage of SSL server certificates issued by certification authorities and recognised by popular web browser software will now be proposed, following the discussion at the ISTC meeting in September. In essence, the FUB-BGBM offers to act as a registration agency for security certificates. It is qualified to do so because it's IT infrastructure conforms to the rules laid out by the German Research Network DFI. The FUB will coordinate this with the partners over the next months. In a next step, the single sign on can be implemented for the existing EDIT applications.

2.10 Tools and developments for ATBI and Monitoring activities

In close collaboration of WP7 and WP5, a technical workflow for the handling of data generated by the "All Taxa Biodiversity Inventories + Monitoring" (ATBI+M) sites in protected areas was developed (C5.096 Report on WP7 data management).

For data input, a format for portable devices was developed that can be used to input data at the site (C7.3.11 Report on field testing of customized data recording tools). The report documents the software application in detail, which represents a major update of a previous version that had been tested at the EDIT summer school 2008. Further testing is to be carried out by researchers at the ATBI+M sites, during the summer school 2010, and possibly within regional projects.

For data output, the pre-existing ATBI websites (and those newly developed) were moved to Drupal to align with the EDIT standard CMS (C7.4.7 Content management system for ATBI+M administration and outreach implemented for every active ATBI+M pilot site). Incorporating the functionality provided by the Geoplatform, the ATBI sites can now include "species sheets" for the individual species found at the specific site (C7.3.10 Prototype of web-based species information sheets with detailed information of the distribution within the ATBI+M pilot sites).

2.11 Usability testing

User input for software development was obtained throughout the project, either indirectly, by analysing the specifications of existing tools, or directly, by means of input received from exemplar groups, in demonstration sessions, or in workshops with users. However, a systematic approach was needed and finally became possible when RBGE joined the EDIT as a partner.

Active work started in May 2009. At the start of the testing process a "Test Plan Discussion Document" detailed the initial plans for testing of the Cyber-platform. This document was immediately circulated to the developers and other members of the EDIT community so as to obtain early and general agreement on the methods, focus and scope of the structured testing to be undertaken. The scope of the tests was agreed to include testing for module compatibility, regression stability, functional completeness, usability / consistency and security. Test results are directly fed back to developers using the Track system (the project management tool used to coordinate software development). The testing will include unstructured as well as structured procedures, for the latter it was decided to use the Testlink framework to document, implement and record the results of the test cases. TestLink will be used to log the results each test case, allowing statistics and metrics (number of failures, outstanding failures etc) to be produced for each application as a whole, or for specific test cases, test suites or test plans. Automated test will be used where possible using the MacroExpress software. Details are provided by C5.130 Reliability and usability testing summary report for Platform version 2.

The work on testing the Cyber-platform components is proceeding well and a good relationship between the testers and the developers has been established. Much useful and timely feedback has already been provided to the Taxonomic Editor development team which led to an enhanced Version 2 of the editor now being released. Several partners have been allocated a small budget to support the testing exercise. During the ISTC it was agreed that the introduction of the tools at member institutions should be paralleled by information events in order to reach out to the taxonomists at the institutions.

It has been agreed that the testers will produce a manual for the Taxonomic Editor based on the application as released for Platform version 2. Work on creating the manual has begun.

3 Integration and sustainability

D11 details the opportunities for synergetic collaboration among institutions and taxonomic researchers offered by modern data and information processing. With the software components finally becoming products, the decisive prerequisite for success is in place. Ultimate success will depend on broadening the support within the taxonomic and biodiversity informatics communities. Acceptance of the products and the necessary accompanying software developments in the sense of adapting the now available tools to user needs will be the focus of the activities during the remainder of the EDIT project. Acceptance of the products is also a prerequisite for the institutional support needed to make the products and the entire Platform sustainable beyond the EDIT project.

Apart from the testing procedures described in section 2.11, increasing the outreach activities undertaken by WP5 staff is planned, including the offer of hands-on training workshops at congresses and institutions.

For the acceptance of Platform products it is also important to offer interfaces to existing applications. Priorities among the many possibilities (see C5.098: List of additional software applications requiring interface definitions) will have to be set in accordance with user needs.

A solid coordination infrastructure will continue to be an important factor to hold together the diversity of platform components developed at different European institutions and to integrate these with intra-institutional developments. Measures in place and to be taken are detailed in C5.097 First report on integration of software developments in partner institutions.

The ISTC will continue to play an important role in efforts to coordinate activities (M5.34 4th ISTC Meeting held). It is proposed to widen the scope of the group to include the largely overlapping membership of taxonomic institutions in SYNTHESYS II and CETAF in the ISTC. Several of the IT departments represented in the ISTC stated that they will support the Platform if the institution's researchers request it.

The existence of projects collaborating with Platform developments (e.g. PESI, BHL-Europe, SYNTHESYS II) is encouraging. The open source software development pathway suggested in a draft of the report on sustainability due in February 2010 is generally supported in the membership, but again: this will only work if the tools developed are used by the institution's researchers. Creating and maintaining a momentum towards usage of tools will be the challenge for the remaining period of the EDIT project.