Nimis P. L., Vignes Lebbe R. (eds.)
Tools for Identifying Biodiversity: Progress and Problems – pp. 1-6.
ISBN 978-88-8303-295-0. EUT, 2010.

Devising the EDIT Platform for Cybertaxonomy

Walter G. Berendsohn

Abstract — This contribution describes the original ideas and preparatory work that led to the implementation of the EDIT Platform for Cybertaxonomy, a computing environment supporting the entire taxonomic workflow. It also briefly describes the current state of development of the project, which will end its EU-funded period in February, 2011.

Index Terms — biodiversity informatics, cybertaxonomy, taxonomic computing, taxonomy.

1 Introduction

Taxonomic research is traditionally a highly collaborative endeavour. The EU project EDIT (European Distributed Institute of Taxonomy) brings together a consortium including most of the largest European natural history museums. EDIT aims at integrating taxonomic research at multiple levels: research policies, collection management, training, outreach and public relations, and research infrastructure.

Natural history museums are information and knowledge institutions. In EDIT, the relatively new area of information technologies was seen as a major chance to integrate the activities of the project partners. Therefore, a third of the project funding was dedicated to create the “Internet Platform for Cybertaxonomy”.

W. G. Berendsohn is with the Botanic Garden and Botanical Museum Berlin-Dahlem, Freie Universität Berlin.
Instead of filling this page with co-authors, the author here lists the collaborators in the EDIT work-package 5 “Internet Platform for Cybertaxonomy” in a comprehensive form:
The task leaders in the work package were Wieslaw Bogdanowicz, Museum of Invertebrate Zoology, Polish Academy of Sciences (MIZPAN); Andras Gubanyi, Hungarian Natural History Museum, Budapest (HNHM); Anton Güntsch, BGBM; Christoph Häuser, State Museum for Natural History, Stuttgart (SMNS) (now at MfN); Mark Jackson, Royal Botanic Gardens, Kew (RBGK); Jorge Lobo, Museo Nacional de Ciencias Naturales, Madrid (CSIC); Karol Marhold, Institute of Botany, Slovak Academy of Sciences (IBSAS); Patricia Mergen, Royal Museum for Central Afrika, Tervuren (RMCA); Martin Pullan, Royal Botanic Garden, Edinburgh (RBGE); Henning Scholz, Museum für Naturkunde, Berlin (MfN); Jane Smith, Natural History Museum, London (NHML); Eduard Stloukal (CUB) and Régine Vignes, Université Pierre et Marie Curie Paris 6 (UPMC). The EDIT development team was first led by Markus Döring and from the second year on jointly by Andreas Kohlbecker and Andreas Müller (BGBM Berlin). Team members were (alphabetically; independent of the time span they worked for EDIT): Anahit Badadshanjan (BGBM), Elek Bozóky-Szeszich (HNHM), Garin Cael (RMCA),Pepe Ciardelli (BGBM), Ben Clark (RBGK), Nils Clark-Bernhard (independent), James Davy (RMCA), Marco Figuidero (NHML), Helene Fradin (UPMC), Giovanni

Thirteen institutions from 8 countries directly participated in the workpackage elaborating the Platform, 7 institutions were involved in software development (programming), with a total of 25 developers (12 concurrent) busy forging the code.

2 Arriving at the Specification

At the starting point of the project in March 2005 we knew that we had a well-resourced project, but we also knew that software development may exceed all cost expectations, especially when carried out in large cooperative projects and (worse) when relatively new technologies were to be used. We were determined not to re-invent the wheel and reuse existing software as much as possible.

The first project year was spent analysing existing software, standards, infrastructures at partner institutions, and requirements from taxonomists, especially from those involved in the EDIT "exemplar group" treatments.

We had realised that we would have to cope with a heterogeneous *institutional landscape,* with widely differing levels of IT capacity. Nevertheless, results from the analysis of institutional IST infrastructures were somewhat frustrating. In spite of the institutions' central role in taxonomic information provision, appropriate infrastructures were often completely lacking, and where in existence, they were rarely outward-looking. Intra-institutional coordination mechanisms were complex, if existing at all. IT developments were mostly depending on soft money funding, with all the consequences for personnel, scope and sustainability. In consequence, we knew that instigating institutional collaboration (the overarching aim of a Network of Excellence project) would present a long-term challenge. It also became clear that we needed a solution independent of database management and operation systems.

We knew that the *taxonomic data domain* was well analysed and to a large extend covered by existing information models and data standards. What was missing was an information model incorporating all these existing, partially overlapping schemes.

It also became clear that good *software applications* were available for taxonomists, for example for descriptive data and the generation of identification keys. Good technologies and an open source environment were available for geographical applications. Handling of bibliographic references was well covered by existing scientific software, including access to citation databases.

Gaias (independent), Marcin Gąsior (MIZPAN), Marc Geoffroy (BGBM), Niels Hoffmann (BGBM), Patricia Kelbert (BGBM), Alexander Kroupa (SMNS, MfN), Dilan Latif (SMNS), Eun-Mok Lee (BGBM), Katja Luther (BGBM), Ôna Maiocco (UPMC), Bart Meganck (RMCA), Dominik Mikiewicz (MIZPAN), Maciej Posluszny (MIZPAN), Francisco Revilla (BGBM), Pere Roca Ristol (CSIC), Pablo Sastre Olmos (CSIC), Bernard Scaife (NHML), Dusan Senko (IBSAS), Lutz Suhrbier (Freie Universität Berlin), Franck Theeten (RMCA), Maxime Venin (UPMC), and Julius Welby (NHML). Exploratory work, information gathering, modelling, and software testing was carried out by Lisa Banfield (RBGE), Franck Dorkelt (INRA), Charles Hussey (NHML), Imre Kilian (HNHM), Boris Jacob (BGBM), Lellani Farina-Crespo (RMCA), Naomi Korn (NHML), Wolf-Henning Kusber (BGBM), Elise Kuntzelmann (UPMC), Barbora Šingliarová (SAVBA), Stanislav Spaniel (SAVBA), David Taylor (RBGK), Dorottya Varsányi (HNHM), Elke Zippel (BGBM), and Magda Zytomska (MIZ PAN). Workpackage coordination was effected by Malte Ebach, Anke Hoffmann, and Agnes Kirchhoff (in that sequence) at the BGBM.

The two most important *data access needs* of taxonomists were tackled by large-scale international initiatives: access to specimen information by the Global Biodiversity Information Facility (GBIF) and access to digitised taxonomic literature by the Biodiversity Heritage Library initiative (BHL). However, an overall integration to cover the needs of taxonomists was lacking. Some of the existing solutions would be difficult to fully integrate, because they depended on specific database or operating systems. Very few solutions existed that supported the full complexity of nomenclatural rules and taxonomic data relations. None was encompassing the full range of data.

Being faced with the unique chance the EDIT project offered, we took the decision to devise, implement, provide and propagate a comprehensive solution for taxonomic computing, the *EDIT Platform for Cybertaxonomy* [1]. The primary objective was to support, enhance and increase the efficiency of the taxonomic work process, for individuals and teams of taxonomists. An explicit aim was to hide the complexity of taxonomic information processing as far as possible, so that it was not inhibiting the workflow, as traditional software applications often did. We knew that new software technologies now offered solutions for some of the problems that had been in the way of creating user-friendly software earlier on. At the same time, the underlying framework had to ensure reusability of the data, seen as the key to future acceleration of taxonomic work processes. On the technical side, hard- and software platform independence had to be ensured to guarantee broad acceptance; at least the newly developed solutions had to be freely available and open source; and for developers wanting to use it for their software projects the solution should provide an API (Application Programming Interface) as well as web services.

In order to achieve these aims, we had to strive to professionalise taxonomic software development. Such a comprehensive solution needed adherence to a strict technological framework. Searching for this framework for development, we looked at content management systems, particularly because using this was a decision taken early-on by another EDIT workpackage -- the “Scratchpads” approach [2]. We saw and see the virtue of this approach for group communication, information dissemination, web publication and aggregation, but we continue to posit that this is not a viable solution for the kind of in-depth treatment of complex data that taxonomists require in their work process. Our aim was principally to support the actual generation of taxonomic data. After weighing several options, Java software development was accepted to provide the most acceptable general framework for Platform application development. Web publication for the Platform can still be realised using content management systems, taking advantage of the Platform’s web services (as demonstrated by the EDIT Data Portal implementations).

3 The Results

Space restrictions allow for only a brief summary of the results achieved so-far. Ciardelli & al. [3] provide a more extensive overview; for full information please refer to the Platform website [4].

The *EDIT Common Data Model (CDM)* now fully covers the data that are

used for systematic treatments resulting from the taxonomic work process (monographs, flora and fauna treatments, and taxonomic checklists). This includes the full complexity of nomenclatural information (botany and zoology), the entire range of taxonomic relationships (including multiple taxonomic hierarchies, synonymies, concept relationships etc.), structured and unstructured descriptive data, geographic information, literature, and specimen data. The CDM is based on existing information models (e.g. the Berlin Model for taxonomic information [5] or the BioCISE model for natural history collections [6]) as well as the standardisation efforts of "Biodiversity Information Standards (TDWG)" -- formerly known as Taxonomic Databases Working Group. Important TDWG standards in this context were the Taxonomic Concept Schema [7], SDD (Structured Descriptive Data) [8], and Access to Biological Collection data [9]. The CDM forms the base for the programming code implemented and made available as the *CDM Programming Library.* The application programming interface or the web services based on the CDM library can be used by programmers to create applications for taxonomists. New functionality created becomes part of the CDM Library after in-depth testing.

As a first step in a user project, a *Community Data Store* is created, i.e. a database that offers the entire scope of information that is covered by the CDM. This can be installed on an individual's computer, on a server in an institutional network, or on servers accessible through the Internet.

Three years of development within EDIT has resulted in a number of CDM-based applications, the two most important of which are the EDITor and the CDM Data Portal.

For data input, the *EDIT Taxonomic Editor (or EDITor)* was developed [10]. It combines an innovative user interface (e.g. allowing full text entry in place of the traditional form-based approach) with the possibility to edit every detail of the database content. The project database can be configured, e.g. by determining which kind of factual data is going to be available for data input (e.g. distributions, threat category, etc.) and which standard terms (if any) are allowed (e.g. TDWG area codes, IUCN threat categories). The taxonomic tree can be displayed and used for navigation and for restructuring by drag and drop. Apart from the taxon-centric standard interface, a "power user interface" presents the data in spreadsheet-like fashion and allows bulk editing and data cleaning. Import and export functionality with several pre-defined formats and standards is implemented. Users can install the EDITor locally on their computer for individual work or access to an institutional Community Data Store, or use it remotely.

The CDM Data Portal is a Drupal-based website used to publish the data in the Community Data Store. It is highly configurable as to displayed content and design. It also offers a taxonomic tree for navigation as well as simple and advanced search functions. The displayed taxon is linked to external resources such as GBIF, BioCASE, BHL, Tropicos, NCBI, Google Images etc. to offer integration with the existing biodiversity information infrastructure. The individual taxon page shows the standard taxonomic data (if the user has configured it that way), i.e. description and factual information. The distribution is visualised through the integrated map viewer (an application of the EDIT Geo-Platform).

All content can be bibliographically referenced. Synonyms can be displayed as homotypic groups, followed by the respective type information. Nomenclatural references are linked to the protologue record (scanned file or web link, where available). An unlimited number of images can be linked and the image gallery offers display in different resolutions and features the image metadata (artist, copyright etc.). CDM Data Portals are in productive use, examples include the EDIT exemplar group sites, for example that for the International Cichorieae Network [11].

Software bundles with the EDITor and Data Portal can be downloaded from the CDM Setup site at http://wp5.e-taxonomy.eu/cdm-setups [12].

Apart from on-line output, functions for pre-formatted print output are being implemented. Out of the (EDIT) box there will be ready made stylesheets for a botanical monograph, a zoological monograph, botanical and zoological checklist, and for the publication of new names in specific journals. Institutional developers will be able to create custom stylesheets conforming to the editorial rules of their in-house publication series.

EDIT has also developed a number of software applications that are not directly CDM based, of which three should at least be mentioned here: (i) The EDIT Geo-Platform [13], [14]; (ii) ViTaL, the Virtual Taxonomic Library, which (in close collaboration with the Biodiversity Heritage Library Europe project) provides an integrated index to taxonomic literature, and (iii) the observation databases and data input tools for the All Species Inventories and Monitoring sites of EDIT workpackage 7.

4 Conclusion

For more than 2 decades there are efforts in joint modelling, standard-building and application development that provide us with excellent knowledge of the taxonomic domain's information structures and business rules. The EDIT Platform is the attempt by European institutions to create a sustainable, collaborative, and comprehensive software solution to increase the efficiency of the taxonomic work process

Acknowledgements

Apart from the EDIT collaborators mentioned in the title page footnote, we would also like to thank numerous taxonomists for their input, in particular those involved with the EDIT exemplar groups: Irina Brake (NHML), Bill Baker, Simon Mayo and Soraya Villalba (RBGK), and Norbert Kilian, Ralf Hand and Eckhard von Raab Straube (BGBM). Gregor Hagedorn gave most valuable advice especially with regard to descriptive data modelling. This work was supported by the European Commission's 6th Framework Programme (Contract No.: 018340).

References

[1] M. Döring and W. G. Berendsohn, "A general concept for the design of the EDIT Platform for Cybertaxonomy", *EDIT newsletter,* vol. 3, pp. 13-15, 2007.

[2] V. S. Smith, S. D. Rycroft, K. T. Harman, B. Scott and D. Roberts. "Scratchpads: a data-publishing framework to build, share and manage information on the diversity of life", *BMC*

Bioinformatics, vol. 10 (Suppl 14): S6doi:10.1186/1471-2105-10-S14-S6, 2009.
[3] P. Ciardelli, P. Kelbert, A. Kohlbecker, N. Hoffmann, A. Güntsch and W. G. Berendsohn, "The EDIT Platform for Cybertaxonomy and the taxonomic workflow: selected Components", *Lecture Notes in Informatics (LNI),* vol. 154, pp. 625-638, 2009.
[4] Anonymous, "EDIT Platform for Cybertaxonomy, "http://wp5.e-taxonomy.eu, 2010.
[5] W. G. Berendsohn, M. Döring, M. Geoffroy, K. Glück, A. Güntsch, A. Hahn, W.-H. Kusber, J. -J. Li, D. Röpert and F. Specht, "The Berlin Taxonomic Information Model", *Schriftenreihe Vegetationsk.,* vol. 39, pp. 15-42, 2003.
[6] W. G. Berendsohn, A. Anagnostopoulos, G. Hagedorn, J. Jakupovic, P. L. Nimis, B. Valdés, A. Güntsch, R. Pankhurst and R. J. White, "A comprehensive reference model for biological collections and surveys", *Taxon,* vol. 48, pp. 511-562, 1999. (Preprint: http://www.bgbm.org/biodivinf/docs/CollectionModel/, accessed 2010).
[7] R. Hyam (Ed.), "Taxonomic Concept Schema – User Guide", Biodiversity Information Standards (TDWG), http://www.tdwg.org/fileadmin/subgroups/tnc/User_Guide.pdf, 2008.
[8] G. Hagedorn, K. Thiele, R. Morris and P. B. Heidorn, "The Structured Descriptive Data (SDD) w3c-xml-schema, version 1.0", Biodiversity Information Standards (TDWG), http://www.tdwg.org/standards/116/, 2005 (accessed 2010).
[9] W. G. Berendsohn (ed.), "Access to Biological Collection Data", Biodiversity Information Standards (TDWG), http://wiki.tdwg.org/ABCD/, 2010.
[10] P. Ciardelli, A. Müller, A. Güntsch and W. G. Berendsohn, "Introducing the EDIT Desktop Taxonomic Editor". In: A. L. Weitzman and L. Belbin (eds.), P*roceedings of TDWG 2008*, Fremantle, Australia, http://www.tdwg.org/proceedings/article/view/325, 2008.
[11] R. Hand, N. Kilian and E. von Raab-Straube (eds.), *International Cichorieae Network: Cichorieae Portal*, http://wp6-cichorieae.e-taxonomy.eu/portal/, 2009+ (continuously updated).
[12] A. Kirchhoff, A. Kohlbecker, N. Hoffmann and A. Güntsch, "CDM setups site - How to install the software modules of the EDIT Platform for Cybertaxonomy", *EDIT Newsletter,* vol. 21, pp. 6-7, 2010.
[13] P. Sastre, P. Roca, J. M. Lobo and EDIT co-workers: "A Geoplatform for improving accessibility to environmental cartography", *J. Biogeogr.*, vol. 36, p. 568, 2009.
[14] P. Mergen and B. Meganck, "Geospatial components for EDIT", *EDIT Newsletter,* vol. 5, pp. 14-17, 2007.