PRO-iBIOSPHERE
WWW.PRO-iBIOSPHERE.EU

Coordination & policy development in preparation for a European Open Biodiversity Knowledge Management System, addressing Acquisition, Curation, Synthesis, Interoperability & Dissemination

SEVENTH FRAMEWORK PROGRAMME

Project Acronym:	pro-iBiosphere
Project Full Title:	Coordination & policy development in preparation for a European Open Biodiversity Knowledge Management System, addressing Acquisition, Curation, Synthesis, Interoperability & Dissemination
Grant Agreement:	312848
Project Duration:	24 months (Sep. 2012 - Aug. 2014)

D.3.1 – Towards a set of Best Practices on Editorial Policies for the curation and publication of fundamental biodiversity data and information in an e-environment

Deliverable Status:	Final
File Name:	pro-iBiosphere_WP3_Naturalis_D3.1_VFF_31052013.pdf
Due Date:	31 May 2013 (M9)
Submission Date:	31 May 2013 (M9)
Dissemination Level:	Public
Task Leader:	Soraya Sierra, Naturalis
Authors:	S.Sierra, D.Agosti, Q.Groom, A.Güntsch, G.Hagedorn, P.Hoverkamp, L.Bénichou, R.Morris, S.Mota de Oliveria, L.Penev

European Commission

Consisting of:

Naturalis	Naturalis Biodiversity Center	Netherlands
NBGB	Nationale Plantentuin van België	Belgium
FUB-BGBM	Botanischer Garten und Botanisches Museum	Germany
Pensoft	Pensoft Publishers Ltd	Bulgaria
Sigma	Sigma Orionis	France
RBGK	The Royal Botanic Gardens Kew	United Kingdom
Plazi	Plazi	Switzerland
Museum für Naturkunde	Museum für Naturkunde Berlin	Germany

Disclaimer

All intellectual property rights are owned by the pro-iBiosphere consortium members and are protected by the applicable laws. Except where otherwise specified, all document contents are: "© pro-iBiosphere project".

All pro-iBiosphere consortium members have agreed to full publication of this document. The commercial use of any information contained in this document may require a license from the owner of that information.

All pro-iBiosphere consortium members are also committed to publish accurate and up-to-date information and take the greatest care to do so. However, the pro-iBiosphere consortium members cannot accept liability for any inaccuracies or omissions nor do they accept liability for any direct, indirect, special, consequential or other losses or damages of any kind arising out of the use of this information.

Revision Control

Version	Author	Date	Status
1.0	Sierra,S. (Naturalis)	16.05.2013	Initial draft
2.0	Sierra, S. (Naturalis) Agosti, D. (Plazi), Bénichou, L. (Muséum national d'Histoire naturelle), Groom, Q. (NBGB), Güntsch, A (FUB-BGBM), Hagedorn, G. (JKI, Plazi), Penev, L. (Pensoft), Mota de Oliveria, S., P. Hoverkamp (Naturalis), R. Morris (University Massachusetts)	28.05.2013	Comments
3.0		31.05.2013	Final version

Table of Contents

EXECUTIVE SUMMARY 6

INTRODUCTION 7

PART I. REVIEW AND ANALYSIS OF THE EXISTING METHODS FOR DATA ACQUISITION AND CURATION, BASED ON THE OUTCOMES OF THE QUESTIONNAIRE. 9

1.1 – OVERVIEW AND OUTCOMES OF QUESTIONNAIRE 9

WHAT KIND OF SOFTWARE/TOOLS/PLATFORMS IS MOST COMMONLY USED FOR ACQUISITION, CURATION AND PUBLICATION OF TAXONOMIC DATA AND INFORMATION?........ 9

A. ONLINE COLLABORATIVE MANUSCRIPT WRITING TOOL (PENSOFT WRITING TOOL, PWT)..... 17

B. ONLINE EDITORIAL MANAGEMENT: SUBMISSION, PEER-REVIEW AND EDITORIAL PROCESS 18

C. PRODUCTION: LAYOUT, PROOFREADING, ONLINE OPEN ACCESS PUBLICATION AND HOSTING.. 18

D. DISSEMINATION OF THE ONLINE CONTENT 19

PART II. TOWARDS BEST PRACTICES ON EDITORIAL POLICIES FOR ENTERING NEW DATA AND COLLABORATIVELY WRITING TAXONOMIC TREATMENTS........ 21

THE PUBLISHING LANDSCAPE IN TAXONOMY - WHERE ARE WE COLLECTIVELY? 21

DEFINITION OF BEST PRACTICES – WHAT WE AIM TO DO 22

TOWARDS A SET OF BEST PRACTICES TO IMPROVE THE XML WORKFLOW BETWEEN NOMENCLATORS AND QUERIES........ 24

TOWARDS A SET OF BEST PRACTICES TO FACILITATE PERMANENT DIGITAL IDENTIFICATION FOR SPECIMENS........ 26

TOWARDS A SET OF BEST PRACTICES TO OPEN ACCESS AND RE-USABILITY OF TAXONOMIC INFORMATION........ 27

CONCLUSIONS........ 28

PART I. REVIEW AND ANALYSIS OF THE EXISTING METHODS FOR DATA ACQUISITION AND CURATION 28

PART II. TOWARDS A SET OF BEST PRACTICES ON EDITORIAL POLICIES FOR ENTERING NEW DATA AND COLLABORATIVELY WRITING TAXONOMIC TREATMENTS........ 28

ACKNOWLEDGMENTS........ 29

REFERENCES........ 30

Executive summary

pro-iBiosphere WP3 “Scientific content and workflow coordination” aims at contributing to the establishment of new and standardised editorial policies for the curation and publication of fundamental biodiversity data and information in an e-environment. This includes editorial policies, Intellectual Property Rights (IPR), management, involvement of citizen scientists in data enhancement, use and re-use of data and information. In particular, pro-iBiosphere Task 3.1 - “Data acquisition and curation” focuses on (i) reviewing and analysing the existing methods for data acquisition and curation; and (ii) identifying and promoting good practices for entering new data and collaboratively writing taxonomic treatments.

In order to poll taxonomists and related professions on their use of digital tools for data acquisition and curation, an online questionnaire was designed and distributed among various stakeholders. A total of 220 responses were received.

In order to promote good practices for entering new data and collaboratively writing taxonomic treatments, a workshop on “Prospective Literature – Towards Best Practices for data acquisition and curation using e-tools for taxonomy” was organised on the 12th of February 2013 in Leiden, the Netherlands. The workshop brought together a group of diverse professionals (about 80 people), representing a wide range of experience and expertise (i.e. IT staff, online curators, publishers and editors of taxonomic literature).

The taxonomic community has shown its ability to digitalise legacy literature in biodiversity. Several challenges exist, and, this paper focuses on determining best practices on editorial policies for the curation and publication of fundamental biodiversity data and information in an e-environment.

Part I of this report presents a review and analysis of the existing methods for data acquisition and curation, based on the outcomes of the questionnaire. Part II focuses on best practices for entering new data and collaboratively writing taxonomic treatments, based on the results of the workshop and questionnaire.

Introduction

The core of biodiversity knowledge is composed of information on species and specimens. A wealth of such knowledge has been assembled in taxonomic literature over hundreds of years. Many European Natural History Institutions (NHIs), such as herbaria, botanic gardens, but also biodiversity specific university departments, publish taxonomic research journals, monographic series or dedicated biota, often in-house and run by their own staff. Some of these publications go back centuries to the founding dates of the NHIs. These publications are used to validate the in-house collections, document national collections, and communicate scientific and technical knowledge in the field of natural history.

A Biota is a published account of all plants, animals, fungi, protozoa or bacteria in a given geographic region. It typically includes descriptions of the species, identification tools ("keys"), synonymy, economic uses, geographical distribution and ecology. Each Biota project is generally coordinated by a natural history research institution, typically with a team of international specialists contributing. Accelerated degradation of natural resources and a growing need for informed conservation decisions have increased the recognition of Biotas as the key to study, understand and preserve biodiversity (see e.g. the Global Strategy for Plant Conservation,where Target 1 is an online world Flora. This recognises the need of all other Targets of Strategy for data held within a Flora).

In general, taxonomic literature typically provides new original information on taxa as well as an authoritative review and synthesis of the information that is already available elsewhere. This information is usually organi sed by a particular group of organisms (= higher taxa like genus, family, order) in a particular region and is compiled and evaluated by specialist taxonomists. It thus includes access to data of the highest quality available, including links to external resources upon which analyses are based. However, external and internal factors call for an urgent modernisation of the production and accessibility of these data, information and knowledge. External factors include the need for biodiversity data to support conservation decisions like the conservation of a particular region or a particular taxon, and for mitigation of climate change impact and adaptation studies. Internal factors are a consequence of the new opportunities due to the digital revolution, and the need to reconcile the increas ing amount of data that needs to be collected and curated by the decreasing numbers of taxonomists available for this type of work. Managing the quality of the data is already a major challenge. A strategy to adapt methods of acquisition, curation, synthesising and dissemination of biodiversity data to the digital era is paramount.

A key aspect of NHIs' publishing over the past 250 years has been the descriptions of species and their relationships to other species, i.e. taxonomy and classification. An unusual aspect of taxonomy is that the description of new taxa such as species must follow specific rules under the relevant *Code of Nomenclature* (e.g. zoological, botanical, bacteriological, and geological) which describe the procedure and rules for creating or changing scientific organism names. Following the rules established in the appropriate code, taxonomic papers are legal documents that determine the legitimacy of a name of a new species. It is thus essential that these publications become globally and easily accessible. Being able to access species description publications, scattered across hundreds of journals, which may have been published any time between 1750 and today, quickly is critical to improve cost-efficiency in environmental management.

In today's digital era, the workflows for doing revisionary taxonomy and producing Biotas are rapidly changing. These new workflows require that aspects of the life cycle of data and information take place online. During the last decade, e-platforms have been developed to facilitate acquisition and curation of taxonomic data and information. Examples of these e-platforms are: Scratchpads, EDIT Platform for Cybertaxonomy , and Biowikifarm. Such e-platforms facilitate data and information acquisition, curation and exchange, and collaborative writing. They allow publication of information in human-readable form as well as atomised and structured content in various machine-readable formats (e.g. xml, json, rdf). All of these lead to higher quality data produced more efficiently.

Technological changes and the flexibility in use of these e-platforms raise various issues such as: how to increase awareness of the availability and use of e-platforms; how to address editorial policies, IPR and management; how to promote involvement of citizen scientists in data enhancement, use and re-use; and how to ensure that these data will remain available and accessible to future generations.

These issues are being addressed by the pro-iBiosphere project through several Tasks. On the one hand, Task 3.1 - “Data acquisition and curation” focuses on (i) reviewing and analysing the existing methods for data acquisition and curation; and (ii) promoting good practices for entering new field data and collaboratively writing taxonomic treatments. On the other hand, Tasks 2.3, 2.4 and 3.2 focus on creating awareness on these e -tools & platforms for taxonomy; the legal issues of data acquisition, curation and dissemination; and the semantic markup generation, data quality, and user-participation infrastructure, respectively.

Part I. Review and analysis of the existing methods for data acquisition and curation, based on the outcomes of the questionnaire.

The most common classes of software used by taxonomist for their research include:

1. Software for acquisition, curation and publication of data and information (e.g., Microsoft Office, Biowikifarm, Scratchpads; Taxonomic Editor, Pensoft Writing Tool)
2. Identification key building and maintenance systems (e.g., DELTA, LUCID, XPER2, SPECIFY)
3. Specimen management databases (e.g., Brahms, BgBase, various customised collection management tools used by the different NHIs, etc.)
4. Bibliography management tools (e.g., Endnote, Mendeley, Zotero)
5. Phylogenetic analysis tools (e.g., Mesquite, Winclada, R-phylo)
6. Online databases/aggregators of taxonomic information (e.g., Biodiversity Heritage Library (BHL), International Plant Name Index (IPNI), Worms, GBIF, etc.)
7. Other software (e.g., for semantic markup of documents (such as Golden Gate, Pensoft Markup Tool), creation of ontologies & geographic information systems).

The focus of the present report is class number one, and the links between this class and the others.

1.1 – Overview and outcomes of questionnaire

In order to poll taxonomists and related professions on their use of software for data acquisition and curation, an online questionnaire (see Annex 1) was designed by the institutions involved in Tasks 3.1 “Data acquisition and curation” and 2.2 “e-platforms and e-tools for taxonomy”.

To facilitate reaching a wide demographic, the questionnaire was distributed among the 100 participants of the pro - iBiosphere workshops that took place on the 11th -14th of February 2013 in Leiden, the Netherlands; three European taxonomic institutions (i.e., Botanischer Garten und Botanisches Museum Berlin, Naturalis, National Botanic Garden of Belgium); the distribution list “Taxacom” (i.e. a list that facilitates discussion on taxonomic issues); and the newsletter of Botanic Gardens Conservation International. The questionnaire consisted of 28 questions, combining a mixture of multiple choice and free text responses. 220 responses were received. The respondents were from 33 countries (from six continents, but predominantly Europe and N. America). 53% of the respondents described themselves as alpha-taxonomists, while the remainder is fairly evenly spread among ecologists, conservationists, citizen scientists, parataxonomists, editors, software developers and database managers. Respondents were given considerable liberty to skip questions they felt were inapplicable and 71% of them completed the whole questionnaire. For the purpose of writing this report, the answers of six questions formulated in the questionnaire were used. The remainder of the answers is being used for writing Tasks 2.2 and 2.3 reports on the feedback and use of e-tools & e-platforms for taxonomy (due August 2013).

What kind of software/tools/platforms is most commonly used for acquisition, curation and publication of taxonomic data and information?

For acquisition and curation considerable usage of desktop software like text processors or spreadsheets is apparent. 78% of the respondents use non-specialised desktop software; particularly Microsoft Office programs (see Figure 1). The use of web oriented software is yet very low. For publication the use of paper printing is not visible, but this is most likely due to a misinterpretation of the questionnaire. Similarly, we realise that known digital publication methods like web or PDF publications seem to be underrepresented in the responses.

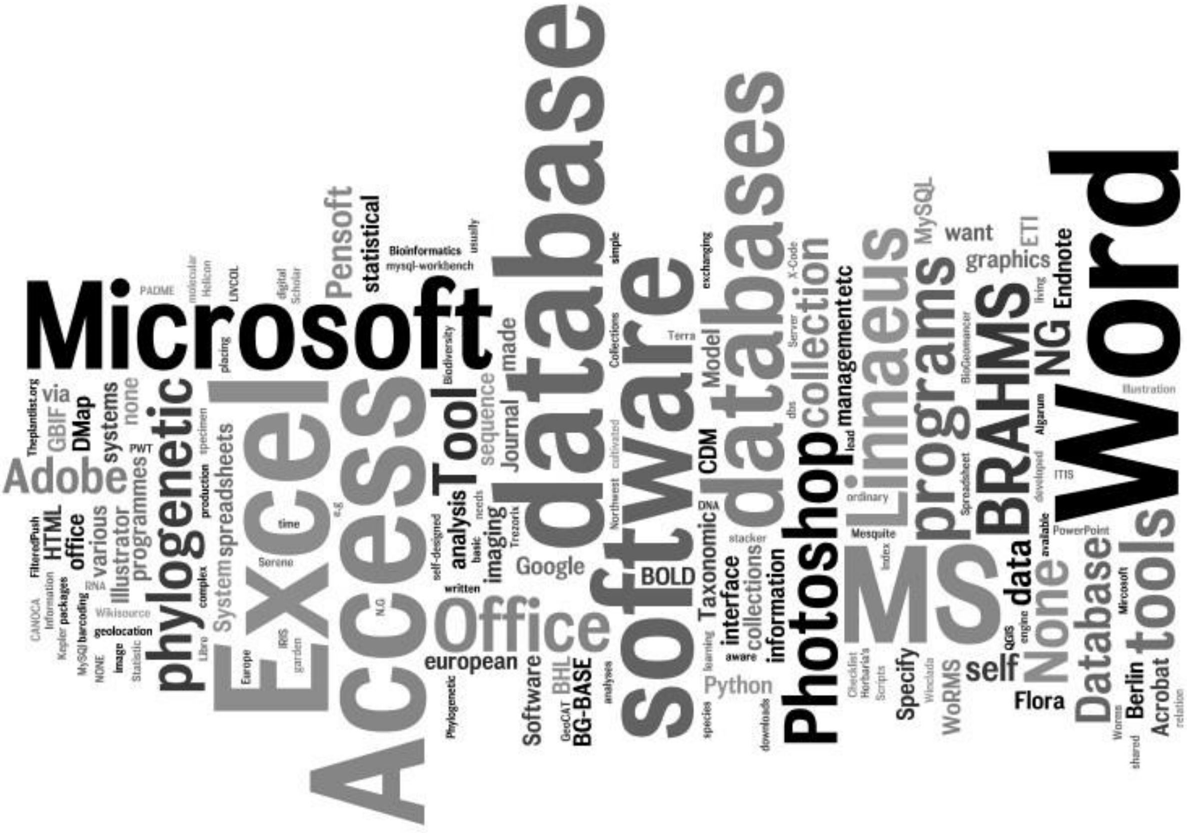

Figure 1. A word map showing the software used by respondents for their taxonomic work.

22% of the respondents use at least one of the following specialised software platforms for (i) acquisition, curation and publication of fundamental biodiversity data and information and (ii) creation of keys (see Figure 2). Yet it is clear from this graph that there are cohorts of early-adopters using several of these systems, these users tend to skew some statistics to give the impression that adoption of these technologies is higher than it is.

- Scratchpads, Taxonomic Editor, Biowikifarm, XPER2
- DELTA; LUCID

While the use of these particular custom software/platforms/tools is rather poor, about 70% of respondents do keep their primary observation data within some form of database.

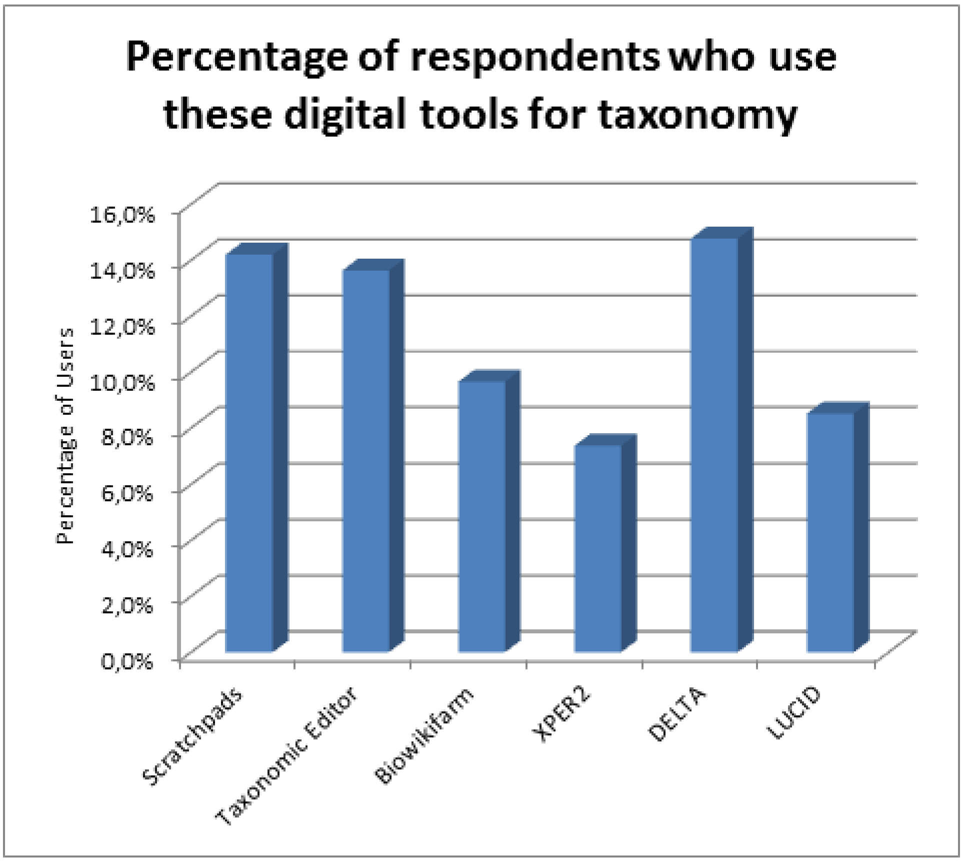

Figure 2. The percentages of respondents that actually use Scratchpads; Taxonomic Editor; Biowikifarm; XPER2; DELTA; LUCID. Note: some respondents use multiple systems so the total percentage of people using digital tools for taxonomy cannot be estimated by adding these columns.

Review and analysis of Scratchpads, Taxonomic Editor, Biowikifarm

We selected Scratchpads, Taxonomic Editor and Biowikifarm for review because (i) all three platforms share a common vision of transforming the traditional working environment of biodiversity research into a modern e -Science environment supporting all aspects of data curation and (ii) they provide for efficient use through portals and services. Interoperability among them is at present being addressed by the FUB-BGBM and other partners in the framework of ViBRANT WP4 (Standardisation). A list of trade-offs between the platforms is presented on Table 1.

Scratchpads

Overview: Scratchpads provides content management functionalities to communities who want to share and publish biodiversity-related data. Content management systems allow the arrangement and publication of material on the internet without having to resort to writing HTML pages from scratch. The website content is stored in a database and presented in an uniform format controlled by the developer. This allows content creators to concentrate on their content, while the HTML coding and storage is handled by the software. Users interact with the program through menus and add new content through a graphical user interface. The system can present data and text in various different ways to create blogs, message boards and forums, individual species pages, etc.

Registered users can be given a variety of different access permissions to enable them to contribute collaboratively at different levels. Master users have full access to configure the website, while contributing users can add text and comments to pages.

Scratchpads has proven useful in bringing people together to collaborate on building content. According to the Scratchpads website, at present (13/05/2013) their user base consists of "*585 Scratchpads by 9,883 active users covering 29,870 taxa in 550,607 pages*". Just their top 10 site have had over 10 million visits and contain over 135,000 pages of content.

Scratchpads is good at accommodating heterogeneous and semi-structured data of all kinds and can be seen as a general and effective tool for data integration. The unstructured format of content allows users the flexibility to create the sort of content they want. Yet the downside of this tolerant approach is that Scratchpads' data often lack granularity and standards-compliance which hinders its use for specialised scientific purposes. However, in the recent change from Scratchpads v. 1.0 to v.2.0 the ability to structure data has improved. Within the ViBRANT FP7 project, Scratchpads developed a publication module that generates and exports highly structured manuscripts in XML format to the Pensoft Writing Tool and from there to the desired journal (e.g., Biodiversity Data Journal, Zookeys, PhytoKeys, MycoKeys, etc.).

Where semantic information is missing this impedes the re-use of the data, but one exception to this is the highly structured way that Scratchpads deals with bibliographical information. This has facilitated the construction of Refbank from Scratchpads data, which aims to be a universal, open, bibliography of taxonomy. This simple use case illustrates the greater utility of structured data if users can be persuaded to do the extra work required to create it.

Software programming. Scratchpads is based upon the Drupal Content Management System. One of the features of new versions of Drupal is that back compatibility with older versions is not maintained. This allows for rapid and innovative development, but it does require reprogramming by the biodiversity community to take advantage of new functionality. The latest upgrade, from Drupal 6 to 7, took place from 2011-2012 (performed by the Natural History Museum of London – NHML, with ViBRANT funding). The new version, Scratchpads 2.0, was released in March 2012. Since then, several tools have been released, including a publication module, and a "bibliography of life" which will accumulate bibliographic references in the biodiversity domain, among others.

Funding. Scratchpads was developed during the FP6 NoE EDIT project. At present, they are part of the EU funded ViBRANT FP7 project and the NERC funded e-Monocot project , both with external funding available until November 2013. Dave Roberts (ViBRANT project manager) announced during the pro-iBiosphere workshops that the NHML will commit dedicated resources for Scratchpads after the EU funding period.

Examples of best use of Scratchpads.
As Scratchpads is based upon a generic content management system it is quite versatile in potential uses. It is not restricted to purely taxonomic uses, but it can be used for any topic or activity that needs a website for collaboration, communication or promotion. Particular examples of the varied uses are below:

- Project promotion OpenUp!
- Community engagement The Phasmid Study Group
- Taxon identification and information Manual of the Alien Plants of Belgium
- Fund raising Wallace Memorial Fund
- Conservation information Cites Bulbs

Biowikifarm

Overview: Biowikifarm is an open content and social networking platform, hosted by the ViBRANT FP7 project, JKI, and BGBM. It allows the maintenance of published data over a long time period, allows more efficient work, and distributes administrative and maintenance work among several partners. Biowikifarm and Scratchpads are fairly related; they use a similar base of mature technologies, namely MySQL and PHP. However, the import complexity level, MediaWiki versus Drupal, has some significant differences.

The Biowikifarm architecture is based on the "MediaWiki" open source authoring system that is also used by pro jects of the Wikimedia Foundation (e.g., the Wikipedia, Wikispecies, Wikibooks, Wikiversity, Wikisource, and the Wikimedia Commons Media Repository). Much of the strength of the Biowikifarm approach derives from this long-term publishing and collaboration platform.

Essential features are:

- the support of the requirements of creative commons licences (perpetuating licences, tracking contributions and attributing all authors of text and media),
- a version management and comparison system making changes in a large community transparent to the end-user,
- a layered development system consisting of core-mediawiki-developers (PHP), mediawiki extension developers (PHP), mediawiki template and form-definition language (over the web, community driven) and javascript development (over the web, community driven).

This approach uniquely empowers the community to participate in the functional development of the system, and provides a flexible and agile environment in which users themselves supply functionality that they find missing in the base system.
Beyond the Wikipedia functionality, the Biowikifarm provides several specialised extensions to support single-(dicho/polytomous) and multi-access (matrix) keys, data harvesting, and the Semantic Media Wiki functionality. The latter is a functionality developed by several Universities to make Wikipedia a part of the Semantic Web (RDF/OWL support); the Wikimedia foundation plans to roll this out to Wikipedia in 2013.
Flexibility. Biowikifarm offers an environment that is very flexible and that can be quickly updated by users. It is safely programmable (with limited programming abilities).

The Biowikifarm is not a complex database model, but a general purpose tool, which can be adapted to a broad variety of projects without recourse to a redevelopment and without compromising the long-term maintainability. It is related to general purpose online-office tools, but with significantly enhanced collaboration, community and data-markup and harvesting functionality. The MediaWiki-based platform is suitable for the development of collaboratively edited flora and fauna projects. The wiki relies to a large part on human error checking, and hence, is less suitable for data analyses when compared to the Common Data Model used by the Taxonomic Editor.

Long-term cost. The underlying software (MediaWiki) is widely distributed, open source and very likely to have the long outside support since the existence of all Wikipedias depend on this system. Once deployed in a well-supported network environment, MediaWiki installations rarely any need IT expertise to manage and support them. In wiki style a large number of talented people can simply add their own functionality on top of an existing one. Compared to Scratchpads and the Platform for Cybertaxonomy, the Biowikifarm platform has the lowest long-term cost.

Expert users vs dedicated developers. To use MediaWiki, it is required to have competent staff (i.e. expert users with an understanding of biology and affinity for computers) available to do most of the work. These expert users should like experimenting, learning and dealing with wiki syntax – it is just more decentralised. A dedicated developer is not required. For the wiki pilot using the static content of Flora Malesiana (e.g. Cyclopedia of Collectors, FM Bulletin:

newsletters, expedition reports, bibliography) new functionality was added by an "expert user" and did not require involving a dedicated developer. This activity turned out to be very easy and time effective.

Potential for community development. Biowikifarm scales to many users. It can however be used without compromise for a group of 1 to 20. If needed, users can make parts of the information (whole namespaces) invisible. None of the wikis on Biowikifarm works in "Wikipedia" mode (i.e. allowing all users to edit). Editing is restricted always to a relatively small group. In fact, the one thing MediaWiki does not do very well is to have many users and many pages and try to assign individually who can write and read where. It is more about trust building, supported by g roup-rights management, than it is about individual rights management. One can, however, easily assign write rights to a limited number of groups. Most wikis choose to allow all contributors from all other bio -wikis to have editing rights as well, but that is a choice which simplifies user management in cases where there is overlap in contributions.

Workflow management. Mediawiki has good tools for collaborative management, however, it has poor tools for single document multi-level supervisor style workflow. Mediawiki does not support older database standards for export, while it does support modern Semantic Web RDF. Mediawiki offers high transparency (users can see what the previous worker has done), however, it allows only one person to make changes at a time.
Funding. The IT-Center of the Natural Science Collection of Bavaria (serving all state natural history, geology and science museums of Bavaria) will guarantee for long-term online availability should dedicated project funds run out.

Examples of best use of Biowikifarm.
Wiki style websites are very versatile, they can be used for practically any type of communication and presentation, but are particularly suited to collaborative works.
A few of the varied uses include

- Identification Guides Species-ID and Key to Nature
- Information on agricultural pests Pest Info Wiki
- Collaborative project organisation pro-iBiosphere
- Regional Checklists of species Vice County Census of South Northumberland

Platform for Cybertaxonomy, CDM library and Taxonomic Editor

Overview. The Platform for Cybertaxonomy offers an object-oriented data management system allowing one to (i) re-use existing data easily, (ii) publish up-to-date information online directly from an underlying database and, (iii) foster collaboration among remote experts to create missing treatments of taxa.

The architecture of the Platform for Cybertaxonomy consists of several layers of specifications, information models, web service layers and APIs (application programming interface) and software applications. At the basis is the "Common Data Model" (CDM), which specifies the content and structure of the data handled in the Platform. It covers the entire scope of data needed in the taxonomic workflow, aiming at the production of monographs, Faunas, Floras, and checklists, both on the web and as print-output. The CDM provides the specification for a "CDM Data Store", which is the repository of data used by a specific group of users (e.g. the authors of a web revision or contributors to a checklist).
To access, edit and process data in the CDM Data Store, functions and procedures have been programmed, which are united in the "CDM library". In addition to general programming code needed for the handling and processing of data, the CDM library contains a comprehensive set of routines specific to the taxonomic domain, for example those based on the rules laid down in the international codes of nomenclature. Database access is isolated from the functional code in the CDM library, which has been programmed using the Java programming language. This allows institutions to freely choose their database management system and operating system.

The CDM library provides programmers with an API, as well as with web services, both defining functions which programmers can use to develop software applications. Within the EDIT project, several applications have been

produced in order to put these developments to use. The most important examples are the " EDITor" (Taxonomic Editor), a desktop application to edit CDM data as the tool for data entry and manipulation, a nd the "Data Portal", which is used to publish the content on the web. The Platform has also been integrated with pre -existing tools of partner institutions, for example the CATE software, which is fully based on the CDM library, and the Xper2 software, which handles the descriptive data for CDM stores while maintaining its stand-alone capabilities. Further development and integration into the Platform as well as updating of the documentation is ensured by permanent staff of the LIS in collaboration with staff at the BGBM, but also by means of collaboration within the ViBRANT project.
In addition to this, there are a number of corollary services, mainly in the geographic and bibliographic domain, which are used by the Platform but can also be used independently

The CDM offers a highly granular object-oriented data model and can be seen as both a working platform for scientists and an information broker supporting all major data exchange standards for biodiversity research. It is therefore very well equipped for the implementation of robust services of high quality required in plant and animal sciences.

Software development. One of the requirements defined by EDIT institutions was scalability of the Platform, ranging from single-user application to institutional networks and internet-based collaboration. In order to meet the needs of the EDIT institutions (in terms of taxonomic workflow, administration and public relations, etc.) new software architecture was developed.

Flexibility. Compared to the other platforms, the CDM is less flexible, and, hence, users can only do what has been anticipated a long time in advance. Despite this, with joined efforts it may be in the long term the most feature -rich model. The Platform is based on very mature Java foundations. The modern technological basis of the Platform architecture forms a base for long term joint development and support, as well as for versatility and fancy software features.

Addition of new functionality. The Platform allows the addition of new features (types of descriptive data) or extensions (additional information that is added to any entity such as a taxon or specimen). For instance, the EU FP7 Palms project uses the Platform. For this project, an external developer wrote code to extend the existing functionality for structured usage data. The CDM is a highly atomised model that is designed to find or prevent inconsistencies, and hence, very suitable to perform data analyses. Compared to the other platforms, the CDM better supports data analysis questions. In order to do this, functionality needs to exist. In case the functionality has not yet been developed, users will need to ask an expert from their local institution (i.e., someone who has understanding of the CDM data structure and basic IT knowledge on data manipulation, e.g. via Excel) to extract the data from the underlying database. Often such questions do not take more than 5-10 minutes to retrieve the underlying data. Most of the CDM 'data analysis' functions are not yet implemented with a user-friendly interface. Often, questions are very specific, and hence, difficult to implement in a generic way. In case questions are frequently addressed by institutions/projects, the FUB-BGBM team can easily implement them either in the EDITor or in the CDM Data Portal.

Potential for development by the biodiversity community. The Platform for Cybertaxonomy and the CDM library have become central elements of the FUB-BGBM's biodiversity informatics strategy. Compared to Scratchpads (software development based at an individual institution), the CDM depends slightly more on the collaboration of the biodiversity community. The EDIT's Information Science and Technology Committee – ISTC (a committee under the Consortium of European Taxonomic Facilities - CETAF) formed a subgroup to ensure coordinated and joint technical development and to provide adequate representation of the stakeholder institutions and projects. Successes in project applications and the increasing number of uses of the Platform is an encouraging sign that EDIT's comprehensive and collaborative approach is starting to provide the promised synergies. The FUB-BGBM can currently offer to drive the development in collaboration with other interested parties. At present there are no formal policy rules for how the development is done. But this is more because the Platfor m developer community is still small enough to follow an informal demand driven approach. Whenever a more formal approach is required, the FUB - BGBM is prepared for it.

Funding. Like Scratchpads, the Platform for Cybertaxonomy, CDM library and EDITor were also developed during the FP6 NoE EDIT project. At present, larger externally funded projects that support parts of the Platform functionality include: SYNTHESYS II, ViBRANT, i4Life (all of these until 2013); BioVel and the German Campanula and Red List 2020 projects. CDM Platform development has been assured beyond any project contract period (since mid -2012 onward the FUB-BGBM has 1 permanent full time position for CDM Platform development). Institutional commitments, funded projects, initiatives and proposals provide a sound base for further steady development of the EDIT Platform. As there are several internal projects directly connected to this task, resources are available to continue the process, especially with respect to hosting the CDM development environment and tools, training of new software developers in using the CDM library, documentation of the CDM library, and incorporation of the Platform development into new funding applications.

Examples of best practice in the use of the Platform for Cybertaxonomy
The Platform for Cybertaxonomy, in comparison to Scratchpad and Biowikifarm, is much more focused on a single use. That is the writing and management of taxon treatments. Specific examples include...

- Detailed information on a tribe of the Asteraceae — The Cichorieae Portal
- Worldwide information on Palms — Palmweb
- Regional checklists — Checklist of the Flora of Central Africa

Table 1. Trade-offs among the three Platforms for e-taxonomy

	Scratchpads	Platform for Cybertaxonomy	Biowikifarm
Base technology	MySQL, PHP	Java, MySQL or other databases, PHP	MySQL, PHP
Software	Drupal	Custom (EDIT CDM)	Mediawiki
Re-use of non-biodiversity IT resources	High	Low	High
Required maintenance through biodiversity IT resources	Medium	High	Low
Software mediated quality control	Medium	High	Medium
User flexibility	Medium	Low	High
External Software updates	New Drupal versions are incompatible with earlier version, time consuming. Reprogramming by biodiversity	Not applicable (custom software)	Fully compatible, tested on very broad and variable

	community required		dataset (Wikipedia)
Type of tool & services	General to specialised	Specialised	General to specialised
Type of data	Unstructured or structured	Structured and highly atomised; unstructured only localised as text fields	Unstructured to semi-structured (Semantic Web markup) or structured
Potential for addition of new functionality	Dedicated developers from biodiversity organisations: High; Expert users: Medium	Dedicated developers from biodiversity organisations: Very High; Expert users: Very low	Dedicated developers from biodiversity organisations: Medium; Expert users: High
Potential for development by biodiversity community	Moderate to High (depending on fragmentation)	High	High
Funds supporting further development	Developed during the EU FP6 EDIT project. The ViBRANT EU FP7 project and the NERC funded e-Monocot project are supporting further development (funding available until Nov 2013). The NHML will commit dedicated resources for Scratchpads after the external funding period.	Developed during the EU FP6 EDIT project. The following projects are supporting further development: SYNTHESYS II (until 2013), ViBRANT and i4Life (until 2013), BioVel and German Red List 2020 (until 2014). From mid-2012 the FUB-BGBM has 1 fte for CDM Platform development	ViBRANT EU FP7 project, JKI and FUB-BGBM host it. The SNSB is giving a guarantee for long-term online availability should dedicated project funds run out

Besides the e-platforms mentioned in the analysis of the questionnaire, there is a new XML-based collaborative online platform for authoring, peer-review, editing, publishing and dissemination: Pensoft Journal System (PJS 1.0 and 2.0). The main features of PJS 2.0 platform are:

A. Online collaborative manuscript writing tool (Pensoft Writing Tool, PWT)

- A collaborative environment for authors and additional contributors (e.g., mentors, potential reviewers, linguistic and copy editors, etc.) to create and work on an online document (manuscript)
- Email and chat communication tools within the group of co-authors and contributors associated with a manuscript
- Automated import of data-structured manuscripts generated in various platforms (e.g., Scratchpads, authors' databases)

- Track change and comments tools
- Revision history, version control and version comparison
- Pre-defined templates for different types of articles, for instance, research article, review paper, data paper, taxonomic treatments, and others
- Different templates for taxonomic and nomenclatural acts compliant to the international codes of biological nomenclature (International Code of Nomenclature for algae, fungi and plants, ICNafp; International Code for Zoological Nomenclature, ICZN)
- Various models of inclusion of data into manuscript (supplementary files, multimedia, import of data tables, linking to external data repositories, etc.)
- Markup of text and data during the writing process, with no additional effort for the authors
- Acquiring and inserting data from external sources into a manuscript in accordance with internationally accepted standards (e.g., species occurrence data in Darwin Core, and others)
- Pre-submission validation of the manuscript
- Automated submission from the PWT to the desired Pensoft's journal

B. Online editorial management: Submission, peer-review and editorial process

- Online submission of manuscripts
- Possibility to opt for a conventional, open, public and post-publication peer-review
- Information on funding agencies and projects included in the submission metadata
- Information on conflicts of interest included in the submission metadata
- Customisation of the financial conditions of publishing services in the manuscript submission form
- Customisation of waiver and discount options in the manuscript submission form
- Customisation of taxonomic, subject, geographical and other classifications used in a journal
- Online manuscript flow control and email alerting network between editors, authors, reviewers and publisher
- Straightforward peer-review process: maintenance of domain-specific reviewers' databases, joint presentation of all reviews and quick integration of referees and editors' comments into the manuscript; email review requests, online review forms, reminders, etc.

C. Production: Layout, proofreading, online open access publication and hosting

- Online layout-to-publication management (formatting, proofreading, copy-editing, etc.)
- Assigning CrossRef DOI (Digital Object Identifier) to the article
- Assigning DOIs to supplementary files, figures, data sets (optional)
- Semantic tagging of the edited text through the Pensoft Markup Tool (PMT)
- Markup of funding agencies, grants and grant numbers acknowledged in the paper
- Markup of taxon names and treatments based on the TaxPub XML schema
- Registration of new plant, fungi and animal taxa in IPNI, MycoBank, Index Fungorum, or ZooBank, respectively
- Dynamic creation of an up-to-date Web taxon profile (through the Pensoft Taxon Profile, PTP) for each taxon name mentioned and automated linking to various biodiversity resources (GBIF, Encyclopedia of Life, BHL, the National Center for Biodiversity Information (NCBI), Genbank and Barcode of Life, PubMed, PubMedCentral, Google Scholar, IPNI, MycoBank, Index Fungorum, ZooBank, PLANTS, Tropicos, Wikispecies, Wikipedia, Species-ID, and others)
- Cross-linking of in-text citations of references and figures with their online visualisations (provided in the HTML version of the paper)
- Obtaining and inserting DOIs through CrossRef for the literature references (when DOIs are available)

- Multiple choice data publishing model that enables the publication of data of different types and complexity as follows:
    - Supplementary data files published alongside with the perspective papers;
    - Support and infrastructure for open data publishing through internationally recognised data repositories, such as Genbank, GBIF, Barcode of Life, Dryad, TreeBASE, Pangaea and others;
    - Specific data types indexed by large data aggregators (e.g., Genbank or GBIF);
    - Data can also be published in the form of marked-up, structured and machine-readable texts;
    - Integration of the editorial workflow with the Dryad Data Repository (useful for a wide array of biodiversity-related data) (optional per journal);
    - Integration of the editorial workflow with the GBIF Integrated Publishing Toolkit (IPT) (useful for any kind of taxon occurrence data and checklists/inventories in Darwin Core Archive format) (optional per journal);
    - Assigning DOI numbers to the published datasets;
    - Possibility to describe and publish data in the form of stand-alone, peer-reviewed "data papers";
    - Publication of multimedia images/files associated with an article;
- Online publication in PDF, semantically enhanced HTML and XML formats
- XML version compliant to PubMedCentral's archiving requirements, based on the TaxPub XML schema
- Option to publish either (1) separate articles, when ready, and/or (2) separate issues or yearly volumes when completed
- Unlimited number of articles or issues per journal per year
- Special issues under distinct title and editorship (monographs, conference proceedings, collections of papers, Festschrift volumes, etc.)
- Assigning ISBN numbers to special issues, in addition to the journal's ISSN, to provide dissemination through book industry networks
- No restrictions or additional charges on use of colour

D. Dissemination of the online content

- Indexing in: ISI Web of Science, Scopus, Zoological Record, Google Scholar, CAB Abstracts, DOAJ, Wikispecies, Vifabio, BHL Citebank, Global names, and others when appropriate
- Automated export of article metadata and linking to social networks (Twitter, Mendeley, Facebook, Google+, and others when appropriate)
- Providing information on funding agencies, grants and grant numbers acknowledged in the paper
- Download citation tool
- Sharing and bookmarking tools
- Post-publication comments and annotations
- Archiving content in one or more repositories, e.g. CLOCKSS or PubMedCentral or others when appropriate
- Export of XML-based metadata and XML versions of the papers to PubMedCentral (in case the journal is accepted for coverage by the latter)
- Updating publication details in specialised registries (e.g., ZooBank for systematic zoology, IPNI for botany, MycoBank and Index Fungorum for mycology)
- Export of all new taxa descriptions, including images, to Encyclopedia of Life (for example: http://eol.org/pages/21232877/overview)
- Export of all new taxa descriptions, including images, to the Wiki environment (Species-ID, example: http://species-id.net/wiki/Spigelia_genuflexa)
- Export of metadata descriptions of identification keys to KeyCentral
- Export of images to Wikimedia Commons (examples: http://species.wikimedia.org/wiki/Wikispecies:Collaboration_with_ZooKeys_and_PhytoKeys, http://commons.wikimedia.org/wiki/Category:Images_from_ZooKeys, and http://commons.wikimedia.org/wiki/File:Acontia_albida.JPG)

- Indexing of new taxa in Wikispecies (example: http://species.wikimedia.org/wiki/Spigelia_genuflexa)
- Automated email acknowledgments to editors and reviewers upon publication of an article
- RSS feeds and article email alerts on newly published articles/issues
- "Saved searches" options in the platform search module
- Public relations (PR) and communication workflow and support, through press releases, social networks and blog postings, Wikipedia articles, etc.

Note: All services described above are provided by the publisher.

Part II. Towards Best practices on editorial policies for entering new data and collaboratively writing taxonomic treatments.

The publishing landscape in taxonomy - where are we collectively?

Natural History Institutions, such as research institutions, herbariums, botanical gardens, and museums, were created with the purpose to contribute to the understanding of the natural world and to the dissemination of this knowledge. Their core mission can be divided into three main objectives, to:

- establish and maintain biological collections (carried out by herbaria, zoological archives, etc.)
- conduct scientific research associated with the collections
- disseminate scientific knowledge within the scientific community and to the general public.

NHIs generate taxonomic publications in the form of checklists, monographs, Floras, Faunas, Mycotas, etc. Within the domain of taxonomy, these publications are equivalent to legal documents validating the names of organisms, which makes the link between publication and taxonomy particularly intertwined. Publishing taxonomic liter ature is part of the mission of most NHIs, promoting dissemination of scientific information in natural history sciences.

The scientific literature dealing with biodiversity has been estimated at approximately 5.4 million volumes (around 800,000 monographs and 40,000 journal titles) since 1469 (Gwinn & Rinaldo, 2009). A key aspect of taxonomic publications over the past 250 years has been descriptions of species and their comparison to other species. The access to publications with descriptions of species (like in Floras, Faunas, etc.) is critical to the field. Many of these publications are 50-years old (or more) and are still being used by various stakeholders outside taxonomy (e.g. climate researchers, ecologists, ethnobotanists/pharmacologists). Taxonomic publications have very long-standing titles with long shelf life fields.

The value of legacy literature was formulated by Page (2013): "many taxa are poorly studied, and hence the chances that someone will find data on a certain organism in the recent literature are likely to be low (unless one studies an economically or medically important taxon, or a model organism). This means that in the case of less well -studied taxa, the data (if it exists at all) will be found not in the modern literature (where the focus has long since moved on from the organism to genomics and system biology) but in the corpus of taxonomic and ecological literature that are being scanned and stored in digital archives" (blog post: Does the legacy literature matter? posted by Roderic Page, Glasgow University, as a result of discussions with participants during the pro-iBiopshere meeting that took place in February 2013).

The difficulty in accessing published literature on biodiversity has always been one of the major obstacles to efficient and productive research. This is particularly true in the field of taxonomy, since the descriptions are scattered across thousands of journals, many of which are difficult to find and access. Furthermore, while many recent data and information is "born digital", in the case of taxonomic works (like Faunas and Floras), most of these data and information were created (usually with government or institutional funding) before the start of the digital revolution. As a result of this, much of the data and information is now available as printed copy, or digitally in PDF - format. At present, despite the availability of state-of-the-art tools/platforms for taxonomy, these data and information are still being gathered in the traditional way (see section 1 of the report). Due to these challenges, the data and information needs transformation into a digital form, by scanning the images/text, doing OCR and conversion to XML through markup activities. Biodiversity literature is currently being digitised by many institutions around the world and the recent eContent plus project Biodiversity Heritage Library for Europe (BHL -Europe) has achieved substantial progress in coordinating and integrating these efforts in the EU. Since the resolution and quality of the BHL scans is not suitable for importing the data to e-platforms like the EDIT Platform for Cybertaxonomy, various institutions interested in creating eFloras/eFaunas have re-OCRd their legacy literature.

Several markup tools and workflows have been developed over the last years, each of them addressing specific use cases. The procedures available are time consuming and very expensive. At present no automated markup procedure is available for Floras and due to the highly heterogeneous structures of source documents, developing fully automated processes can be very challenging. In order to achieve a higher degree of automation in the markup procedures, various challenges need to be addressed. For instance, written language often contains inconsistencies in format and the OCR of the scanned documents is by no means perfect due to the original documents having bad print quality, typos, symbols and fonts (= lettertypes) that do not exist anymore, etc. This has two implications. Firstly, the automated markup cannot be performed on the OCR'ed documents because they first need to be cleaned up to remove everything that does not need to be included. Secondly, the automated markup process usi ng scripts will not add XML everywhere where it is needed, requiring a thorough check of the resulting file. pro -iBiosphere Task 3.3 “Semantic integration of Biodiversity literature” (led by the MfN) is in charge of aligning ongoing and forthcoming efforts to semantic mark up of biodiversity literature and will provide technical and social solutions for their use. To facilitate further markup, Plazi and the other partners will analyse the XML schemas currently implemented in their workflows. Three viable paths for future improvement of semantic markup are presently recognised: (1) fully automated natural language processing (NLP), (2) base markup complemented by automated processing and specialist correction, and (3) social crowd-sourcing models (citizen involvement).

In order to facilitate the open access and re-use of data and information of taxonomic literature by various stakeholders (i.e. non-taxonomists), a more efficient way of working is needed. New workflows will allow data to be structured in a coordinated and consistent way.

As the field of scholarly publishing is changing rapidly NHIs are facing an increasing abundance of information that requires consistent dissemination. They face complex, strategic and technical issues related to the access, format and financial structure of their existing titles. The challenge addressed here is to adopt modern electronic publishing practices so as to enable increased production and a wider circulation of the results of taxonomic research both for the legacy and the prospective literature.

Many NHIs still publish one or more research journals, often in-house. Most of the time, journals are run by isolated members of staff or owned by learned societies with institutional support. In addition, publishing institu tions are confronted with various technological revolutions in scientific publishing that pose complex and strategic questions. Very few institutions have a real publishing team able to address the technological revolution challenges.

This section of the report focuses on providing recommendations towards a set of best practices on editorial policies for entering new data and collaboratively writing taxonomic treatments. All this in order to facilitate structuring these documents, exporting, harvesting, and organising the data more efficiently, and thus increase its accessibility.

Definition of Best Practices – What we aim to do

According to Wikipedia on best practice, “A best practice is a method or technique that has consistently shown results superior to those achieved with other means, and that is used as a benchmark. In addition, a "best" practice can evolve to become better as improvements are discovered. Best practice is considered by some as a business buzzword, used to describe the process of developing and following a standard way of doing things that multiple organisations can use”.

Several recommendations were seen as an essential to ensure better coordination in the use of data, in a consistent way both for the legacy and the prospective literature.

The Best practices on editorial policies for entering new data and collaboratively writing taxonomic treatments will be available and updated on the pro-iBiosphere wiki. The pro-iBiosphere workshops organised in February 2013 (Leiden) and May 2013 (Berlin) were useful to establish where are the needs for best practices and to create discussion on the various topics. At this stage the consortium can only provide insights on various best practices that have been identified on the following topics:

- Best Practices for collaboratively writing taxonomic treatments
- Best practices to improve the XML workflow between nomenclators and queries
- Best practices to facilitate permanent digital identification for specimens

Towards a set of Best Practices for collaboratively writing taxonomic treatments
According to the answers received on the pro-iBiosphere questionnaire (see section 1 of the report), the main barriers that prevent people from using e-platforms and e-tools for taxonomy are the:

- Lack of training on the use of e-platforms/e-tools (54% of respondents)
- Time it takes to learn a new system (35% of respondents)
- Lack of application support (30% of respondents)

Respondents to this question had the opportunity to give free text responses. These covered a wide range of subjects and were often contradictory, reflecting the diversity of the community. Nevertheless, an implied theme that seems to permeate the responses is that the advantages do not justify the investment in time and effort. Some of these answers were:

- “Most taxonomic tools are too complicated for my purposes”.
- “There is currently no need to use digital taxonomic tools”.
- “There are too many different tools and if I need to make a whole survey of everything that exists in order to make a choice, it would take too much time”.
- “My examination of some digital tools showed them to be too flexible”.
- “The single biggest reason is that the tools I have are more than adequate to the task.”

Indeed, most respondents did not agree that their research was so unique that it was incompatible with current systems. Nor was language a barrier.

The emphasis in responses towards training suggests that developers should devote time to usability. Few modern software tools come with manuals and users are expected to be able to follow the logic of a software system, using reference material infrequently. Time spent on usability can help to lower the uptake barrier and increase the use of e-tools/e-platforms for taxonomy. Furthermore, software with a large user base is more likely to attract further development.

From the perspective of availability and adaptability, the use of generic software, such as Microsoft Word, can be advantageous. However, generic software has many disadvantages in terms of standardisation, collaboration, validation and re-use. The use of e-platforms/e-tools for taxonomy needs to be incentivised among the persons involved in writing the taxonomic treatments. This facilitates entering the data in a structured way and avoids the use of further markup (an expensive and time consuming process).

One very specific use of Microsoft Word has been the mark up of legacy texts. Microsoft Word has been extensively employed in this process at the Royal Botanic Gardens, Kew and at the National Botanic Garden of Belgium, Meise (Kirkup, D., Malcolm, P., Christian, G. & Paton, A. (2005). Towards a digital African flora. *Taxon* **54** (2) 457–466). Paper documents scanned and transformed to digital text with OCR are further marked up either partially or wholly within MS-Word using a combination of macros and manual editing. Extensive use of the styling and XML functionality with in MS-Word made this possible, as did the fact that this is a relatively stable, mature product that inexperienced users can work with easily.

The demand for accessibility to digital information from the end-users of taxonomic information needs to filter down to the taxonomists, and may also help to increase awareness on the use of e- tools/platforms.

1. Online curation of data should be seen as an important priority.

The digital curation of dynamic datasets, information and knowledge, as available in Floras and Faunas, involves continuous enrichment or updating. This continuous process is essential because it enhances the long-term value of existing data by making it available for further high quality research and reduces duplication of effort in research data creation.

The writing of new taxonomic treatments is seen as a main priority, whereas the online curation of fundamental biodiversity data and information is not. Some of the reasons for these are:

- Online curation is a time consuming process and there are no incentives for practising taxonomists to spend time on this. Fundamental biodiversity data and information is published online by aggregators (e.g. GBIF, EoL, etc.) and is not being curated. The data and information they publish comes from various sources, and hence can only be updated if the first aggregator that collected the data is approached (by the specialist taxonomist) and that aggregator regularly exchanges data with others.
- Present methods to maintain taxonomic accuracy of online biodiversity data are non-existent or inadequate. The available annotation systems do not support taxonomists in their basic activities, while still requiring them to spend additional time on making their knowledge public.
- Information on taxonomic identity is usually taken for granted by the various stakeholders and too little credited to the taxonomists who provide the basic information.

These challenges need to be addressed in order to facilitate efficient re-use of fundamental biodiversity data and knowledge, and increase its dissemination. Linking of these data and information, and annotation tools for curation will be addressed during the pro-iBiosphere workshops that will take place on the 23rd of May 2013.

An important aspect to be considered in this new way of working is involvement of citizen scientists in dat a enhancement, use and re-use. pro-iBiosphere D.3.2.1 “Concept paper for involvement of individual experts, commercial vendors, and citizen scientists” addresses questions such as how the different actors may be incentivised, recognised and rewarded for these activities.

2. Create engagement and commitment from the community

It is important to raise awareness and educate publishers and (isolated) editors to change editorial practices. This can be done by

- Providing guidelines for publishing.
- Providing pre-defined templates.
- Providing vocabularies in e-tools (authors are inconsistent & language complex).
- Defining and monitoring data quality.

Towards a set of Best practices to improve the XML workflow between nomenclators and queries

Nomenclators provide an exhaustive list of species and, hence, they facilitate knowing what names have been published and assist users to avoid and disambiguate homonyms (a homonym is a name for a taxon that is identical in spelling to another name that belongs to a different taxon).

Taxonomic names on plants, fungi and animals are registered in nomenclators such as:

- IPNI (Royal Botanic Gardens Kew - RBGK)

- Index Fungorum (RBGK – Index Fungorum Partnership)
- Mycobank (CBS-KNAW Fungal Biodiversity Center)
- Zoobank (Bishop Museum)

The data and information stored on these nomenclators is used by various aggregators/projects/initiatives such as Catalogue of Life, Genbank, etc.

At present, data is entered to these nomenclators following two approaches:

1. Scanning of literature upon receipt of hard copy journals by libraries. This results in a time lag between publication data and entry (in the case of IPNI, this may be as much as 2 years).
2. Users report missing nomenclatural acts (usually accompanied by a link to digitised literature page (BHL)).

During the pro-iBiosphere workshop, the operators of these nomenclators expressed their intention of moving away from the hard copy approach and instead, moving towards new workflows involving XML integration. pro -iBiosphere is conducting various pilots with the nomenclators on this topic. Information on these pilots is available on: http://wiki.pro-ibiosphere.eu/wiki/Pilots

At present, when online journals are used (instead of the hardcopies), two parallel workflows for register ing names are followed by nomenclators, consisting of: (i) a pre-submission registration (by the author or registry agent) and (ii) a pre-publication registration (by the publisher). This approach increases the possibility for errors and duplication, and is time consuming.

This can be avoided by adopting best practices involving a common XML query/response model for automated publication-to-registration pipeline, consisting of the following steps:

Step 1. XML query to the registry upon acceptance of the manuscript (containing the type of act, taxon names, and preliminary bibliographic metadata).
Step. 2a. XML query response containing the unique identifier (e.g., LSID, PURL, or other resolvable URLs) of the act and potential error messages.
Step. 2b. Correcting potential errors and duplicates: human intervention, at either registry's or publisher's side (or at both).
Step. 3. Inclusion of identifiers in the published treatments (protologues, nomenclatural acts).
Step. 4. Final XML report sent by publisher on the day of publication (exact bibliographic details of the published article: authors, title, journal, issue no, date of publication, pagination). Once the article is approved for publication, an XML query is sent to the appropriate registry. The registry responds with the identifiers for the taxonomic act.

Since the different registries support a different set of taxonomical or nomenclatural acts and save different pieces of information, a common XML registration model is hardly possible and three workflows will need to be created (and associated XML schemas).

This new common XML query/response model for automated publication-to-registration pipeline allows users to:

- resolve nomenclatural problems pre-publication
- switch the editorial role from keying to checking
- seed identifiers into literature
- ensure data consistency for curation
- increase the usability of published data

Towards a set of Best practices to facilitate permanent digital identification for specimens

When referring to a specimen (describing its properties, citing it as evidence) the conventional methods of material identification are, e. g.:

- name or code of institution owning or curating the specimen (e.g., in the case of plants, the code might be according to Index Herbariorum)
- institutional accession number
- a combination of collection data such as collector, collection date or number

The above-mentioned methods often lead to confusion, misinterpretation, or an inability to restudy the cited specimen. The community lacks a generally agreed simple and digitally enabled identifier mechanism to cite specimens in an unambiguous manner.

Historically, the biodiversity community decided to use Life Science identifiers (LSIDs) for all kind of resources, including taxa or specimens. However, while LSIDs where supported for taxonomic names, the acceptance for specimens was very low. The absence of support for LSIDs in general web browser and the difficulty to support them in the Semantic Web motivated only few institutions to support them.

Recently, the desire to participate in the Semantic Web and Linked Open Data has caused new interest in modern alternative identifiers for natural history collection specimens. The Royal Botanic Garden Edinburgh (RBGE) published a paper (Hyam et al. 2012; see also Stable Citations for Herbarium Specimens on the internet) on using the Linked Data principles (http://www.w3.org/TR/cooluris/) to issue HTTP URIs (URLs) for their specimens. The pro-iBiosphere project was instrumental in furthering this discussion by addressing the issue in depth during both the Leiden (2013-02) and Berlin (2013-05) workshops and developing a best practices document (http://wiki.pro-ibiosphere.eu/wiki/Best_practices_for_stable_URIs).

In the Semantic Web, each specimen, each specimen itself is referred to by a stable URI. When dereferencing this URI (e.g. using a browser), the system returns the information that a second URI contains a related information document. This information document might contain data about the specimen, images, etc. It can be returned either in a human-readable form or as Semantic Web readable data (RDF).

The Information Science and Technology Commission of the Consortium of European Taxonomic Facilities (ISTC-CETAF) decided to recommend further investigations and a workshop will be held in Edinburgh in June 2013. Several European collections already decided to implement their identifier system using this approach (Royal Botanical Garden Edinburgh, Museum für Naturkunde in Berlin, Botanical Garden and Botanical Museum Berlin) and Zoobank will offer this approach as an alternative to its present LSID-based system. We expect other institutions in the biodiversity domain to follow.

In some cases DOIs may have been issued for specimens. This is a permissible alternative that should not be prevented by editorial practices.

Concrete recommendations for editors of journals or monographs: Encourage authors to add a web address (a URI that starts with http://) for each specimen they cite. Not all collections provide this yet; it is therefore desirable to develop a list of all known collections that do. The URIs provided by the author for a specimen in the publication should be tested in a web browser. If it results in a web page, but is not shown as the recommended identifier specified on that web page, the author mistook the URI-identifier of the web page for the URI-identifier of the specimen itself. Editors should educate their authors about the difference.

Towards a set of best practices to open access and re-usability of taxonomic information

From a legal point of view, most elements of taxonomic information do not qualify as distinct creative works for which copyright is recognised. Content cannot be copyrighted, only the creative form of it. Well established scientific practices exist for the form in which taxonomic knowledge is communicated. Its form is thus copyrightable only in exceptional circumstances. This includes observation data, names, list of names, or treatments. Even where the entire publication may be copyrighted, these parts are not copyrighted (Agosti & Egloff, 2009). This includes compilations of lists of names, reporting knowledge about taxa or making the discovery and description of a new taxon.

In general, copyright law does not protect the effort invested into or the creative analysis necessary for the preparation or discovery of the content. The scientific merit is covered by moral and scientific practices requiring the citation of the data. This not only gives credit, but also has a function of building trust into a dataset by citing the source and the authority of the creator. Citing is standard practice in science. Disregarding it is considered plagiarism and can lead to harsh consequences like the loss of an academic title or occasionally even employment.

The goal of scientific publishing is to share data and knowledge and make these available as widely as possible for others to use. This is the intention of the taxpayer when funding such work and, in general, also the drive of the ones who fund scientific work. In this regards, scientific knowledge is contributed to the welfare of the society by allowing the re-use of information and applications of data innovation (see e.g. Whitehouse, 2013).

Even where it is legally possible to extract non-copyrightable content parts from publications that as a whole are copyrighted, it is recommended to clarify and simplify the situation by making the entire publication open access under an open access licence. The most common licence is CC BY (attribution required) which models the traditional best practice of scientific publishing. In some cases, the CC BY-SA (similar to OpenSource Software development) may be an alternative. The “non-commercial” licence (CC BY-NC) is not recommended since it excludes the use for non-profit (but commercial) use cases such as university-based education (see Hagedorn et al. 2011). For data the permissive CC0 licence is the most appropriate solution. Data are not copyrightable, including most free -form text fragments expressing facts. However, where data contains free-form text, a danger exists that in rare occasions some of these free-form texts may be copyrightable. The CC0 licence avoids this uncertainty and allows unhindered re-use of the data. This does not affect the necessity to cite such a source in scientific works.

To sum up, the proposed instrument for a future biological knowledge management system, such as the use of identifiers or data mining work, can only be used successfully in an open acc ess environment when information delivery and analysis requests are not prevented by paywalls.

Conclusions

Part I. Review and analysis of the existing methods for data acquisition and curation

When deciding what e-platform is the most suitable for an institution, the IT capacities (or capabilities) of the institution and the requirements of the projects that will use the platforms, need to be taken into account.

Scratchpads (Drupal), Biowikifarm (Wikimedia) and Platform for Cybertaxonomy (CDM) provide integrated platforms for compiling and sharing biodiversity data online. These platforms have proven very useful in bringing people together to collaborate in building content. They are suitable for the development of online and hardcopy Flora and Fauna versions. Interoperability among them is at present being addressed by the ViBRANT project.

- Scratchpads are very good in accommodating heterogeneous and semi-structured data of all kinds and can be seen as a very general and effective tool for data integration. Scratchpads use Drupal as software, and hence programming by the biodiversity community is required (old and new versions are not compatible with each other).
- Biowikifarm is not a complex database model, but a general purpose tool, which can be adapted to a broad variety of projects without recourse to a redevelopment and without compromising the long-term maintainability. It is related to general purpose online-office tools, but with significantly enhanced collaboration, community and data-markup and harvesting functionality. Its strongpoint is that due to the close link with Wikipedias software development, its longevity outlook as a collaboratively maintained open source tool is excellent, in effect as long as Wikipedia will exist.
- The CDM offers a highly granular object-oriented data model and can be seen as both a working platform for scientists and an information broker supporting all major data exchange standards for biodiversity research. It is therefore very well equipped for the implementation of robust services of high quality required in plant and animal sciences. The Platform is based on very long-term Java foundations. The modern technological basis of the Platform architecture forms a base for long term joint development and support, as well as for versatility and fancy software features.

Part II. Towards a set of Best Practices on editorial policies for entering new data and collaboratively writing taxonomic treatments.

We advocate the open access and free re-use of data publicly-funded and produced by NHIs (including publications). While we do not question the convenience and usability of paper/PDF publication for human reading, their applicability for the forthcoming e-environment and Semantic Web is approaching a "less than zero" value. PDFs can be produced from XML master documents. So changing the fundamental data archive from PDF or paper to XML will have little effect on the way most scientists actually consume publications, but adds the benefits of (potentially) semantically structured data and non-proprietary format. This is because the content from PDF or paper needs to undergo the huge effort of digitisation (OCR) and markup to enable text-mining, import into databases, etc. Being this is an expensive and time consuming process.

Despite the technological developments, most content of Biotas (e.g., Floras, Faunas, Mycotas, as of biodiversity information in general, is still being published in "closed', non-machine-readable formats, such as paper and PDF. Those closed formats are in most cases available through a pay-wall. Scientists continue to gather high quality, well-structured data, which are then being "closed" into non-machine-readable publication formats, which lead on its turn on doubling the effort to get back the published data into databases.

A standard that is widely available and specifies a minimum of fundamental biodiversity data, e.g.: references/citation, taxon treatments, collection specimen numbers, names, material citations, descriptions illustrations (line art to multimedia), etc., needs to be formulated. A best practice should specify that publishers of biodiversity data should export these data, irrespective of the form of publication, and allow its wide dissemination.

Acknowledgments

We would like to thank the following persons for reviewing the R. Morris (Harvard University), Jeremy Miller (Naturalis),Eva Krait (Naturalis), and Peter Hovenkamp (Naturalis).

References

Agosti, D., & Egloff, W., 2009. Taxonomic information exchange and copyright: the Plazi approach. BMC Research Notes 2009, 2:53 doi:10.1186/1756-0500-2-53.

Gwinn, Nancy E., & Rinaldo, Constance. 2009. The Biodiversity Heritage Library: Sharing biodiversity literature with the world. IFLA Journal 35(1), 25-34.

Hagedorn, Gregor; Mietchen, Daniel; Morris, Robert A.; Agosti, Donat; Penev, Lyubomir; Berendsohn Walter G. & Hobern, Donald 2011. Creative Commons licenses and the non-commercial condition: Implications for the re-use of biodiversity information. Zookeys 150: 127-149. – doi: 10.3897/zookeys.150.2189.

Hyam, R.D., Drinkwater, R.E. & Harris, D.J. 2012 Stable citations for herbarium specimens on the internet: an illustration from a taxonomic revision of Duboscia (Malvaceae) Phytotaxa 73: 17–30.

Kirkup, D., Malcolm, P., Christian, G. & Paton, A. 2005. Towards a digital african flora. Taxon 54 (2) 457–466.

Whitehouse, 2013. Expanding access to results of publicly funded research. http://www.whitehouse.gov/blog/2013/02/22/expanding-public-access-results-federally-funded-research.

http://wiki.pro-ibiosphere.eu/wiki/Workshops_Leiden_February_2013
http://scratchpads.eu/
http://wp5.e-taxonomy.eu/
http://biowikifarm.net/
http://wp5.e-taxonomy.eu/taxeditor/
http://pwt.pensoft.net
http://code.google.com/p/open-delta/
http://www.lucidcentral.com/
http://lis-upmc.snv.jussieu.fr/lis/?q=en/resources/software/xper2
http://specifysoftware.org/
http://herbaria.plants.ox.ac.uk/bol/
http://www.bg-base.com/
http://www.mesquiteproject.org/mesquite/mesquite.html
http://www.cladistics.com
http://www.r-phylo.org
http://www.biodiversitylibrary.org/
http://www.ipni.org/
http://www.gbif.org/
http://www.bgci.org/resources/article/0736/
http://www.wordle.net http://vbrant.eu/
http://vbrant.ipd.uka.de/RefBank/search
http://www.e-taxonomy.eu/
http://www.nerc.ac.uk/
http://e-monocot.org/
http://open-up.eu/
http://phasmid-study-group.org/
http://alienplantsbelgium.be/

http://wallacefund.info/ http://citesbulbs.myspecies.info/
http://biowikifarm.net/meta/Julius_K%C3%BChn_Institute_%28JKI%29
http://www.wikipedia.org/
http://species.wikimedia.org/wiki/Wikispecies:Collaboration_with_ZooKeys_and_PhytoKeys
http://www.openstarts.units.it/dspace/bitstream/10077/3739/1/Hagedorn%20et%20al,%20bioidentify.pdf
http://wiki.floramalesiana.org/wiki/Main_Page
http://species-id.net/wiki
http://www.keytonature.eu/wiki/
http://wiki.pestinfo.org/wiki/Main_Page
http://biowikifarm.net/v-ukcounty/web/Vice_County_Census_of_South_Northumberland
http://www.e-taxonomy.eu/files/C5.133_second_Report_on_sustainability_of_the_Internet_Platform_for_Cybertaxonomy_software_Final.pdf
http://wp6-cichorieae.e-taxonomy.eu/portal/
http://www.palmweb.org/
http://floreafriquecentrale.org/
http://www.pensoft.net/services-for-journals
http://ibot.sav.sk/icbn/main.htm
http://www.nhm.ac.uk/hosted-sites/iczn/code/
http://rs.tdwg.org/dwc/2009-02-20/terms/guides/text/index.htm
http://www.crossref.org/
http://www.sourceforge.net/projects/taxpub
http://www.mycobank.org/
http://www.indexfungorum.org/
http://zoobank.org/
http://www.ptp.pensoft.eu/
http://www.eol.org/
http://www.ncbi.nlm.nih.gov/
http://www.ncbi.nlm.nih.gov/genbank/
http://www.boldsystems.org/
http://www.ncbi.nlm.nih.gov/pubmed/
http://www.ncbi.nlm.nih.gov/pmc/
http://www.scholar.google.com/
http://plants.usda.gov/java/
http://www.tropicos.org
http://www.pensoft.net/page.php?P=23
http://www.genbank.org/
http://www.datadryad.org/
http://www.treebase.org/
http://www.pangaea.de/
http://www.pensoft.net/ipt.pensoft.net/ipt/
http://www.gbif.org/informatics/standards-and-tools/publishing-data/data-standards/darwin-core-archives/
http://science.thomsonreuters.com/ http://www.scopus.com/home.url
http://thomsonreuters.com/products_services/science/science_products/a-z/zoological_record
http://scholar.google.com/scholar?q=zookeys&hl=en&btnG=Search&as_sdt=2001&as_sdtp=on
http://www.cabi.org/default.aspx
http://www.doaj.org/doaj?func=openurl&issn=13132989&genre=journal&uiLanguage=en
http://www.vifabio.de/search/?q=source:zookeys&c=BASE&lang=en

http://citebank.org/search/apachesolr_search/zookeys?filters=ss_biblio_remote_db_provider%3A%22Pensoft%20Publishers%22
http://www.globalnames.org/
http://www.clockss.org/clockss
http://eol.org/pages/21232877/overview
http://species-id.net/wiki/Spigelia_genuflexa)
http://keycentral.identifylife.org/
http://commons.wikimedia.org/wiki/Category:Images_from_ZooKeys,
http://commons.wikimedia.org/wiki/File:Acontia_albida.JPG)
http://species.wikimedia.org/wiki/Spigelia_genuflexa
http://iphylo.blogspot.nl/search?q=legacy+literature
http://en.wikipedia.org/wiki/Best_practice
http://stories.rbge.org.uk/archives/1377
http://stories.rbge.org.uk/archives/1284