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)**

D2.2 Report on user feedback

Deliverable Status:	**Final**
File Name:	**pro-iBiosphere_D2.2_RBGK_VFFa_31082013.pdf**
Due Date:	**31 August 2013 (M12)**
Submission Date:	**31 August 2013 (M12)**
Dissemination Level:	**Public**
Task Leader:	**Don Kirkup (RBGK)**
Authors:	**D. Kirkup, A. Paton; E. Kralt, J. Miller, S. Sierra**

Consisting of:

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

Revision Control

Version	Author	Date	Status
1.0	Don Kirkup (RBGK)	14 August 2013	First Draft - Distribution: Donat Agosti, Henk Beentje, Laurence Bénichou, Quentin Groom, Peter Hovenkamp, Bente Klitgaard, Eva Kralt, Jeremy Miller, Alan Paton, Soraya Sierra, Jonathan Timberlake.
2.0	Don Kirkup, Alan Paton (RBGK)	27 August 2013	Introductory text added and expanded. First Revision (incorporating corrections received from Laurence Bénichou).
3.0	Daniel Mietchen (MfN), Soraya Sierra (Naturalis), Don Kirkup (RBGK)	28 August 2013	Revision
4.0	Laurence Bénichou (MNHN), Eva Kralt (Naturalis), Quentin Groom (NBGB), Anton Güntsch (FUB-BGBM), Daniel Mietchen (MfN), Educardo, Alan Paton (RBGK), Don Kirkup (RBGK)	29 August 2013	Draft
5.0	Don Kirkup (RBGK), Eva Kralt (Naturalis)	30 August 2013	Draft formatted
6.0	Don Kirkup (RBGK)		Final Draft
FF	Don Kirkup (RBGK)		Final Draft converted to PDF

Table of Contents

Executive summary

The present document is a deliverable of the pro-iBiosphere project, funded by the European Commission's Directorate-General Information Society and Media (DG INFSO), under its 7th EU Framework Programme for Research and Technological Development (FP7).

This document reviews how information held in Biotas (i.e. publications such as Faunas, Floras and Mycotas) is used by a variety of audiences. The methods used to gather evidence included interactive workshops, pre- workshop questionnaires, follow-up interviews and desk based research. The main audiences surveyed were taxonomists, informaticians, conservationists, ecologists, publishers and IT developers who routinely handle Biota information. Distribution, morphology, habitat and taxonomy are the most commonly used information types.

The major constraints to users of Biotas and the information they hold were:

- The time needed for information to be retrieved from a Biota and presented in a more appropriate format for a particular reuse, and locating the relevant source work in the first place;
- The lack of easy to use technical solutions for mining or presenting data currently in Biotas for reuse in other products or activities.
- Extraction of atomised information by markup is time consuming, technically difficult and potentially very costly. However, such data underpins analysis of information and may facilitate further synthesis, provided that data standardisation is adopted.
- Interfaces to data marked-up and atomised from Biotas need to be easy to use and follow data standards where appropriate.
- Limited access to information due to:
    - restrictive access conditions
    - difficulties in interpreting the data
    - inefficient discovery mechanisms
- Access to information varies, depending on where you live and work, the resources at your disposal and on your available finance. IPR restrictions are only one of the factors limiting access.
- There is a strong requirement from users of human expertise in addition to access to Biota information for activities such as identification or validation of data.
- People (and their expertise) are a valuable resource but as there is no standard index to them, significant time is spent locating them. User feedback channels are poor and this hampers understanding of user requirements. The anonymity of users of some online systems contributes to this.
- Users of Biotas have requirements for information traditionally not published in Biotas, but closely related and potentially available to Biota producers. Significant data gaps (such as species abundance) and monitoring could be resolved and undertaken by local people, provided they have easy access to existing information and means to add that information to the online Biota.
- Identification of species is a major activity both in the field and lab. The kinds of tools needed to help with this varies from low-tech to high tech depending on the particular situation. Where appropriate, these tools should be designed to integrate with local expertise.

Introduction

The importance of understanding users

The vision of pro-iBiosphere is to prepare the ground for the creation of a system for intelligent management of biodiversity knowledge by addressing technical and semantic interoperability of the challenges to improve the present system of taxonomic literature. Workpackage 6 investigates alternative business requirements and scenarios for a sustainable Open Biodiversity Knowledge System and aims to provide recommendations with regard to achieving sustainable delivery of core biodiversity data and information. This report provides information on how information held in Floras, Faunas, Mycotas and other related taxonomic literature (referred to here as Biotas) are used by a variety of audiences. Understanding how this information is used provides the basis of how this information can be better disseminated to these audiences. This is key to sustainability. A sustainable system must address the needs of the user community and the producers of Biotas need to understand more clearly what the demands of the user base are and the constraints to serving these users. This understanding, along with identifying what the benefits to the users are of this information and how these benefits can be maximised (the subject of the pro-iBiosphere workshop on "user engagement and benefits" to be held on 9 October 2013, in Berlin) provide vital information on how a sustainable system for Biota information needs to relate to the users of that system.

The challenges faced in trying to understand user needs

Traditionally Biotas were produced in hard copy. They sought to provide baseline biological data on a particular taxonomic group or the organisms from a particular area. As such there is an extremely broad range of potential users to this information and the content of Biotas was often repurposed to address the needs of particular audiences. For example, a detailed inventory of a national park compiled for management purposes may use much of the taxonomic information held in the Flora of that particular country. In the digital age the producers of Biotas want to make the their information to an extremely broad variety of users, but they need to do this in a flexible way, catering for as broad a range of users as possible and to try and ensure that they do not inadvertently constrain activity by making the information available in inappropriate data structures or format. It is not possible within the time constraints of this project to survey all potential users of Biota information. We concentrated on gathering information from the following categories of user that we know from previous experience and other collaborations are heavy users of Biota information: taxonomists, informaticians, conservationists, ecologists, publishers and IT developers who routinely handle Biota information.

Methods

A range of individuals from these sectors were invited to a workshop which aimed to understand the information requirements of our users in order to be able to specify well designed information services **(Annex 1).** A pre-workshop questionnaire was sent to the participants to capture information on their backgrounds and interests (**Annex 2).**The workshops took the form of small breakout groups who explored particular use-cases relevant to their actual use of Biota information. The use-cases were written up after the workshop and some participants were further interviewed to clarify particular points. Some desk top research also helped to clarify issues. Thirteen separate use-cases were investigated. These are listed below and detailed in **Annex 3**. Discussions were summarised on flip charts and summary presentations were recorded.

Figure 1 shows the use-cases located within the "information space" together with the participants' main disciplines. Distribution, morphology, habitat and taxonomy are common information requirements. The close proximity of taxonomists and conservationists to each other possibly reflects the dual role of major collections based biodiversity institutions. Publishers dealing with general outputs too occupy a central position. The informaticians and IT specialists are particularly concerned with nomenclature and validated content, whereas the ecologists' more specialised information requirements are outliers.

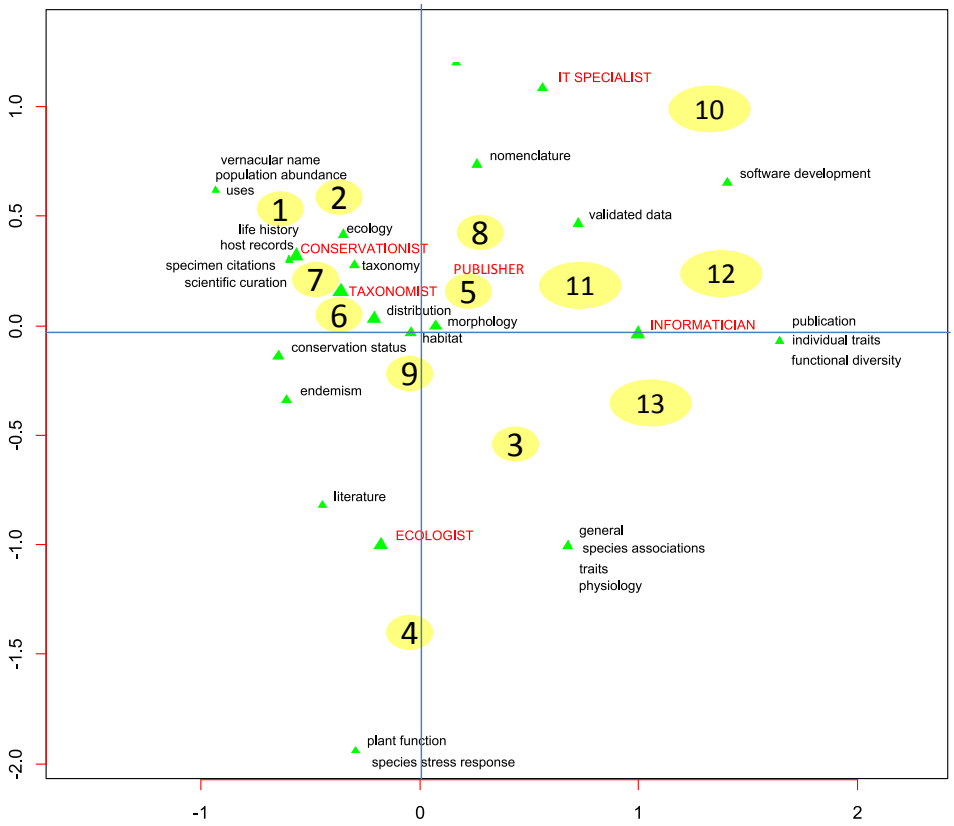

Figure 1. Use-cases and participants' roles plotted in "information space". A 2D plot based on a Joint Multiple Correspondence Analysis in principle components space. The use-cases are numbered within the yellow circles, the participants' roles are in red and the information types are in black. The areas of the green triangles are scaled to the relative frequency (mass) of the points at that location.

Two groups (use-cases 1 and 2) examined preparation of a species conservation assessment based on the IUCN procedures. This activity is a cornerstone of practical conservation work and the published assessments feed into national and international policy. Use-cases 3 and 4 are concerned with the compilation, databasing (3) and analysis (4) of trait information for use in vegetation modelling, and for the refinement of functional classifications. This is an important link with Earth Systems Science community modelling global change.

Use-case 6 looked at practical plant identification in the tropics, both in the field and with access to an herbarium. It reflects the typical experience of ecologists and conservationists working in the tropical countries. The key elements are the general lack of resources and the reliance on non-technical field guides, local knowledge and expertise.

Use-cases 5 and 7 are taxonomic studies dealing with the description and publication of a new species (5) and the preparation of a Flora for a poorly known region (7). They differ in the amount of time that can be spent resolving problems such as species delimitation and can be thought of as representing opposite ends of the spectrum of taxonomic studies.

Use-case 8 is from the publishers' perspective, dealing with the dissemination of high quality taxonomy in both printed and digital format, and is a central activity for dissemination of biodiversity information. Use-case 11 is also publishing, but is based on a proposal for re-publishing existing works digitally, with semantic enhancements.

Use-case 9 is a survey of a national park for specific management needs. Like conservation assessments (use-cases 1 and 2), the outputs feed directly into policy making.

Use-case 10, the production of a digital flora, use-case 12, the production of a field identification tool and use-case 13, ecological niche modelling based on specimens and observation from Floras, each explore the practical problems of producing digital taxonomy.

Further information on how users inter-react with Biota information was presented as a series of lightning talks. The titles and presenters of the lightning talks are presented in **Annex 7.**

Results

Summary of the pre-workshop questionnaire (from Annex 2)

Summary of responses

Tables 1 to 5 summarise the 35 responses to the pre-workshop questionnaire and are based on re-coded categories. In each case, the counts should be interpreted in the context of the sampling of participants.

Table 1. Major foci of the Participating institutions (from Annex 2).

major focus of institution	number of institutions
systematics	13
conservation	13
ecology	8
collections	8
data aggregation	3
sustainable use	2
education	2
research	1
publishing	1
horticulture	1
biogeochemistry	1

Note: an institution may have more than one focus.

Table 2. Major roles of the Participating individuals (from Annex 2).

participants' roles	number of participants
taxonomists	13
informaticians	12
conservationists	10
ecologists	6
publishers	3
IT specialists	3

Note: a participant may have more than one role.

Table 3. Major uses of Biota information sources (from Annex 2).

Major uses	number of participants
information extraction	30
species identification	18
dissemination	4
training	1
curation	1

Note: Information extraction includes all print based, manual and digital, programmatic approaches. A further breakdown of the kinds of information extracted is given below.

Table 4. The kinds of information extracted from Biotas (from Annex 2).

Kinds of information	Number of participants
distribution	20
morphology	7
habitat	5
validated data	5
taxonomy	5
conservation status	5
ecology	4
nomenclature	4
endemism	3
"content"	3
vernacular names	3
literature	2
general	2
"soft" traits	2
life history	1
host records	1
publication	1
specimen citations	1
uses	1

Note: "Content" is general output required by informaticians and IT specialists for the development of portals, services, tools and other software. "Publication" in this context refers to outputs for print and electronic dissemination. "Soft" traits are primarily morphological, easily observable features, which correlate with the underlying functional, physiological ("hard") traits.

Table 5. Other kinds of information sought but hardly ever found in Biotas (from Annex 2).

Other information sought	Number of participants
individual traits	1
life history	1
population abundance	1
functional diversity	1
plant function	1
species associations	1
species stress response	1
"hard" species traits	1
physiology	1

Summary of Use-case activities (from Annex 4)

Use-case activities are coded and summarised as follows:

- Processing includes activities such as specimen preparation, databasing, markup.
- Research activities include locating information, fieldwork, identifying species, data gathering and analysis, identifying gaps in data and gathering user feedback.
- Synthesis relates to the largely intellectual and manual process of summarising of data (merging of digital data is included under processing).
- Validation concerns quality control, editorial functions, peer and other forms of review, expert assessment.
- Capacity Building is mostly training.
- Scope includes defining projects, inclusion and exclusion criteria.

The difficulty levels are coded High/Medium/Low based on the participants rankings of the activities. The type of difficulty is coded as follows based on the participants' description of their activities:

- Time constraints - things that fundamentally take a long time (because of their scale), or take up too much of the available time within a project. This is a common element with most the difficulties recorded.
- Technical constraints - tasks that are hard to do maybe due to lack of expert knowledge or technical knowhow.
- Access constraints - things which make difficult or prevent the use of existing resources, for example due to the difficulty in finding them (e.g. a lack of indexes or unpublished information), their dispersed or remote location (so it's impractical to visit), or through a lack of infrastructure (e.g. library, internet services), or a lack of funds with which to buy access (e.g. books, internet, travel), or IPR restrictions.
- Missing information - resources that are required, but which do not yet exist.

	ACCESS PROBLEMS			IPR RESTRICTIONS			MISSING DATA			TIME CONSUMING			TECHNICAL DIFFICULTIES			
	H	M	L	H	M	L	H	M	L	H	M	L	H	M	L	
capacity building	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	*2*
dissemination	0	0	0	0	0	0	0	0	0	1	4	1	0	1	0	*7*
implementation	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	*1*
processing	3	0	0	1	0	0	0	1	1	25	19	1	14	3	2	*70*
research	11	1	4	0	0	0	11	3	4	32	14	7	21	3	1	*112*
scope	0	0	0	0	0	0	0	0	0	0	2	3	0	0	0	*5*
sustainability	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	*1*
synthesis	0	0	0	0	0	0	0	0	0	2	4	0	0	0	0	*6*
validation	0	1	0	0	0	0	0	1	0	5	7	0	3	2	0	*19*
	14	*2*	*4*	*1*	*0*	*0*	*11*	*5*	*5*	*67*	*51*	*13*	*38*	*9*	*3*	*223*

Things that participants find difficult with "processing" activities

(Virtually) all highly time consuming and technically difficult tasks, but access and missing information is not a major problem. Key digitisation tasks (from Annex 4) are the markup of and extraction of specific information from a broad range of information types (e.g. names occurrence, description). Less technically demanding but still time consuming is the handling of materials - processing of specimens, loans of collections material.

Things that participants find difficult with "research" activities

Most research based activities are time consuming in the gathering of data and require technical expertise to evaluate and analyse.
Access and missing information too, are significant components of the difficulties. Access problems relate to dispersed location of resources, the need for fieldwork, the lack of literature or of reference material in local collections, or the lack of access to people with needed expertise. Access to media such as images can be hampered by IPR restrictions.
Missing ecological information includes a range of important data on populations (e.g. species abundance, reproduction) which is vital for proper conservation assessment and ecological research. Missing operational information includes indexes: for resources such as e-Floras (there is no registry for these) and for people names (e.g. lists of experts), lists of institutions, the locations of specimens, and for controlled vocabularies needed for development of digital services.

Things that participants find difficult with "synthesis" activities

Synthesising large amounts of information from different sources, in different formats and with no, or differing standards is costly of time.

Things that participants find difficult with "validation" activities

Most of the use-cases include the activity of validation of information which requires significant time and expertise. Identifying qualified reviewers can be an access problem within publishing activities.

Summary of the importance of information types (from Annex 5)

The table below summarises the (coded) information types by importance, as ranked by the participants of the workshop. The numbers are the counts of the activities in which the information class is used.

		IMPORTANCE			
		HIGH	MEDIUM	LOW	
INFORMATION TYPE	bibliography	3	1	0	*4*
	classification	2	0	0	*2*
	collection	2	0	1	*3*
	conservation	4	2	4	*10*
	description	9	0	4	*13*
	ecology	3	0	9	*12*
	function	0	1	5	*6*
	genomics	0	0	2	*2*
	habitat	3	10	7	*20*
	media	3	0	7	*10*
	name	12	0	0	*12*
	nomenclature	2	1	1	*4*
	observation	5	2	1	*8*
	geography	14	5	0	*19*
	operational data	10	0	6	*16*
		72	*22*	*47*	*141*

Geography (distribution and occurrence) and taxon names are the most highly rated information types, closely followed by taxon descriptions and operational data (the latter includes controlled vocabularies as well as standard lists of place-names, people and institutions). Habitat information is widely used but is deemed less important (possibly because it can be extracted from geographically referenced specimens). Conservation information (e.g. threat status, protected areas), functional and ecological information is also used but is of lower priority, perhaps reflecting the likelihood of finding the information. Classifications, nomenclature and bibliographies are important but to relatively few activities. Collection information is not frequently used but is of importance in some activities.

Summary of information sources (from Annex 5)

The sources of the information types are summarised in the table below. The source "literature" is printed, published information. It excludes online resources such as BHL which are included under "internet", together with GBIF and others and general search engines. The term "database" is intended for local, offline data-sets. "Lists" include gazetteers, standard lists and controlled vocabularies. "Grey" includes personal communication and other unpublished sources. "People" includes experts as well as those with local knowledge.

		INFORMATION SOURCE									
		people	field/lab	collections	maps	literature	lists	database	internet	grey	
INFORMATION TYPE	bibliography	0	0	0	0	1	3	1	6	1	*12*
	classification	0	1	0	0	1	1	1	0	0	*4*
	collection	0	1	3	0	1	0	2	0	0	*7*
	conservation	2	4	3	1	10	0	4	1	9	*34*
	description	2	2	3	0	16	2	7	3	3	*38*
	ecology	3	10	9	1	12	0	10	2	16	*63*
	function	3	2	2	0	6	0	4	2	5	*24*
	genomics	0	0	0	0	0	0	1	2	0	*3*
	habitat	1	4	4	6	9	0	8	3	4	*39*
	literature	0	0	0	0	6	0	1	6	0	*13*
	media	3	1	2	0	4	0	1	15	0	*26*
	name	2	0	2	0	15	1	6	11	1	*38*
	nomenclature	0	0	0	0	4	0	1	9	1	*15*
	observation	0	7	0	0	5	0	2	0	2	*16*
	occurrence	1	3	7	1	8	5	7	12	1	*45*
	operational data	1	2	1	0	1	8	0	8	1	*22*
		18	*37*	*36*	*9*	*99*	*20*	*56*	*80*	*44*	*399*

The printed published literature is the major resource for most types of information although the use of internet resources (such as IPNI, GBIF) is particularly important for names and occurrence data, and the nomenclatural information and literature used in taxonomic research (BHL). The grey literature and people are particularly important sources for conservation and ecological information, and often includes the "raw" data of measurements on individuals - however, they are very difficult to find.

Conclusion

Distribution, morphology, habitat and taxonomy are the most commonly used information types. These elements have also been recorded as important in taxonomic needs assessments carried out undertaken as part of the Global Taxonomic Initiative of the Convention on Biological Diversity (e.g. Taylor, A. 2006).

Users face time and technical constraints in trying to access information. Time is taken by having to locate the data, as often there is no way to quickly determine where that information might be found. When it is located, the data may not be easily reusable. Often legacy literature has not been made available online or it is not adequately marked-up and in a form which can readily be reused. These factors have technical solutions, but the implementation of these solutions needs to follow agreed standards and interfaces need to be easy to use. There are also constraints on reuse of data created by traditional processes of Biota production. There are some instances where Biota producers provide summary data, for example a description of organism morphology and distribution, but not the raw data such as the measurements of a morphological part or specimen information supporting the distribution. Providing access to this raw data in standard easily findable ways would support a variety of users in their activities. Biota producers could also consider capturing additional information at the same time as gathering information for taxonomically oriented works. Data on population size, life cycle or other factors facilitating conservation assessment could be gathered at the same time as taxonomic work during field work and presented, where available, in the Biota. This point was also made by Lowry & Smith (2003). Broader access to basic Biota information might stimulate and enable local users to add information such as field observations on populations to the account.

In addition to data provision there are instances where taxonomic assistance is required to interpret the Biota information. For example, the identification of plants can be done via traditional dichotomous keys or multi-access online resources, but non-taxonomists also value information on who can name organisms from a particular place or taxonomic group. Publishing and other synthetic activities also require information of who could review or validate information.

References

1. Harper, J. L. (1977). Population biology of plants. Academic Press.
2. Lowry, P. P., & Smith, P. P. (2003). Closing the gulf between botanists and conservationists. *Conservation Biology*, *17*(4), 1175-1176.
3. Taylor, A. (2006). United Kingdom Taxonomic Needs Assessment. http://www.cbd.int/doc/programmes/cro-cut/gti/gti-needs-uk.pdf

Annex 1. Notes for facilitators for the workshop "The users and uses of Biota publications and services".

Overall goal and workshop goals

To understand the information requirements of our users in order to be able to specify well designed information services. The workshop specific goals are outlined below.

Workshop session 1: use-case activities

- **The task**
    - for individual use-cases, identify the activities in which Biota info is used and establish which activities are the most time consuming and most difficult, i.e. potential "pain" factors for users
- **The outcome**
    - an activity map for each use-case (Post-it notes on flipchart paper)
    - a list of activities sorted by time/cost for each use-case (Post-it notes on flipchart paper)
    - photographic back-up of all of the above
    - notes made by facilitators during discussion
- **The process**
    - individuals or small groups first Post-up a list of activities the need to carry out for their use-case
    - same groups using Post-it action-map tool to make a map of the activities; each in turn give a short talk through for their map to the rest of the group, points noted by facilitators
    - same groups using Post-it swap-sort tool to prioritise the activities by cost/time; each give short talk through, points noted by facilitators
- **The rationale**
    - each individual or small group maps the activities for a specific use-case to show where Biota info. fits into their particular workflow and give their view of which activities are the most time consuming and costly to them; the size of the group depends on how similar each use-case is – may have to work as individuals to avoid over generalising
- **The expected time** – 90 minutes

Workshop session 2: use-case information

- **The task**
    - to list the types of information that are used in each particular use-case and to assess relative importance of each type to the user
    - to record the sources, interactions and the destinations of information used within workflows, to list the data standards in use and discuss any barriers to exchange/interoperability
- **The outcome**
    - a list of information sources used, sorted into categories (information type)
    - a list of categories ordered by importance to the user (Post-it notes on flipchart)
    - an information map for each use-case showing (in)compatibilities (Post-it notes on flipchart)
    - back-up photographs of the above
    - a list of standards used & info where standards not used (inc. notes from discussion)
- **The process**
    - Individuals or small groups use Post-it post-up tool to list the information and sources they use in carrying out their use-case
    - same groups Post-it Bottom-up-tree to sort their lists by information type (into columns). Add column heads for the categories and use Post-it swap-sort tool to order by value to the user.
    - add lines between sources to indicate compatibility e.g. (black), incompatibility (red)
    - list any standards used for each data type category (i.e., using the same column heads)
- **The rationale**
    - using the activities mapped in the previous exercise as a prompt, the groups think about the types of information and sources, how comparable different sources are and what are the standards (formal and informal) that they use; discussion could bring out reasons for adopting or not adopting standards
- **The expected time** – 90 minutes

Workshop session 3: What should a Biota of the future be able to do for you?

- **The task**
  - listen to and record users views on how Biota (fauna, flora, mycota) publications and services can be improved
- **The outcome**
  - a list of ideas from users
  - notes on discussion points
- **The process**
  - individual presentations to the whole group (lightning talks on "What should a Biota of the future be able to do for me?")
  - whole group brainstorming, open discussion notes by facilitators
  - a short presentation on 3 propositions (e.g. basic markup with links, semi-structured, fully atomised) and gradient of agreement taken for each proposition. Results lead into further discussion
- **The rationale**
  - the lightning talks lead in to the whole group discussion and brainstorming
- **The expected time** – 180 minutes

Annex 2. Participants' responses to pre-workshop questionnaire

Your name	Primary focus of your institution	Your role	Your use of flora/fauna/mycota in the above role	Geographic focus	Taxonomic focus
Henk Beentje	Taxonomy + conservation	Herbarium taxonomist, Red List assessor	Identification - extracting species distribution - extracting habitat information - extracting endemism information - extracting morphological information	Africa (Malesia)	Compositae Pandanaceae Palmae trees mangroves aquatic plants
Laurence Bénichou	Research, Knowledge dissemination, Conservation Teaching Expertise	Publications manager	I publish and disseminate them	World wide	All
Walter Berendsohn	Taxonomy, Biogeography, Collection (herbarium)	In this context: Checklist compiler (Dendroflora of El Salvador)	Identification of specimens. - extracting information about general distribution (countries) - specimen citations for target area - common names in target area - occurrence status in target area (native, possibly cultivated etc.) - nomenclatural detail (checking if in consensus with other sources) - taxonomic status (accepted, synonyms) - further literature and other notes that may be of use for determining the taxonomic concept to be used in the checklist.	El Salvador, C.A.	Large woody vascular plants (mainly trees)
Melanie Bilz	Conservation	IUCN Red List compiler	Extracting information on; - species distribution - ecology - habitat preferences - population abundance	Europe	Vascular plants
Christopher Chapano	Taxonomy, ecology and	Ecologist / Herbarium	Identification -species distribution,	Distribution of species and	Simple keys that can be used by amateur

Your name	Primary focus of your institution	Your role	Your use of flora/fauna/mycota in the above role	Geographic focus	Taxonomic focus
	conservation	taxonomist	- endemism - threat status - vegetation types	endemism	botanist
Thionois Charlotte	MNHN: research, conservation, teaching, knowledge dissemination and expertise	Scientific publishers	Publication of descriptive taxonomy	International valorisation of collections conserved in European Natural History Institutions	Publication of new data such as nomenclatural acts (zoology, botany, palaeontology)
Viola Clausnitzer	Conservation	Chair IUCN Dragonfly Specialist Group - member IUCN Red List Committee - field ecologist (Africa, Odonata)	Identification - Red List assessments	Africa	Dragonflies (Odonata)
Joe Cora	All (university)	Database manager, server administrator, software developer	Biodiversity information acquisition from primary sources, analysis and dissemination	worldwide	No focus per-se, but our resources are hymenopteran dominant
Eduardo Dalcin	Taxonomy and Conservation	IUCN Red List compiler, National plants checklist compiler, primary data acquire, validate and host	Offer validated data to the other systems and processes	Brazil (National)	Plants
Pablo Demaio	Conservation	IUCN specialist group chair	Identification - species distribution - taxonomy and nomenclatural features - ecological features	South America	Plants
Sonia Dias	Conservation and use	Trait, accesison, characterization and evaluation database; checklists, National Inventories and Conservation strategies	Documentation - data quality - support - publication -training	mostly European region (42 countries), and a global scope for some actions	Plant genetic resources withe xception of forestry (in my case)
Henry Ford	ecology	Trait database compiler, ecologist	Morphological and physiological features linked to associations of species.	UK	Vascular Plants
Quentin Groom	Taxonomy and ecology	various, biodiversity	Publishing information the internet and	North-western Europe and	Vascular plants

Your name	Primary focus of your institution	Your role	Your use of flora/fauna/mycota in the above role	Geographic focus	Taxonomic focus
		informatics roles	analysis of distributions data.	tropical Africa	
Jana Hoffmann	Taxonomy, Phylogenetics, Preservation, Ecology (Modelling)	Taxonomist (Marine Animals)	Identification of specimens - extracting morphological and distributional data - reference work	Caribbean Westpacific	Brachiopoda
Vololoniaina JEANNODA	Conservation (Madagascar Plant Specialist Group)	IUCN Survival Species Committee: Madagascar plants conservation status assessment and validation	Identification - vernacular name, - morphological features (description), - species distribution and ecology - uses - threats and pressure on habitat - presence in protected areas - endemism	Mainly Madagascar, but also Mascarene islands	Angiosperms and Pteridophytes mainly
Jens Kattge	Biogeochemistry	Trait database compiler	Information from Floras adds to the characterization of functional diversity in terms of trait values for individual plants and/or species. We use information from Floras on species name, trait values and species distribution.	Global	Global
Robert Kenward	Conservation through Sustainable Use of Biodiversity	Vice-chair of IUCN Sustainable Use and Livelihoods Specialist Group	Encouraging communities to map Biota for distribution and density estimations	Europe	Terrestrial and freshwater
Bente Klitgaard	RBG, Kew: conservation, taxonomy, ecology, systematics, horticulture	Managing and development of Neotropikey (identification tool for Angiosperm families of Latin America)	Extracting information for our tools	Latin America	All Angiosperm 318 Angiosperm families present in the Neotropics
Bente Klitgaard	RBG, Kew: conservation, taxonomy, horticulture, systematics, etc.	Herbarium taxonomist	Identification, extracting species distribution, morphological features etc.	Latin America	Leguminosae/Fabaceae
Bente Klitgaard	RBG, Kew: ecology, taxonomy,	Conducting biodiversity surveys	Identification	Latin America	All Angiosperm families present in Latin America

Your name	Primary focus of your institution	Your role	Your use of flora/fauna/mycota in the above role	Geographic focus	Taxonomic focus
	conservation etc.				
Patricia Mergen	All	Biodiversity information project management	All	worldwide some focus on Africa, Central Africa	All
Jeremy Miller	Taxonomy and biodiversity	taxonomist, developer of online fauna for megadiverse taxon	Investigating point species richness and rates of community change across landscape	Southeast Asia	Spiders
Chuck Miller	Systematic Botany - taxonomy, floras, monographs	Biodiversity Informatics	Building information systems for botanical taxonomic workers.	Global	Vascular plants and bryophytes
Andreas Müller	Taxonomy	database developer	requirements engineering to make data electronically available	None	Botany
Luciana Musetti	Taxonomy	Systematic entomology	Identification (keys and descriptions), species distribution, morphological characters, life history, host records, etc.	Worldwide	Insecta: Hymenoptera
Deborah Paul	iDigBio is an aggregator and data portal for vouchered specimens, so we are interested in all of this data.	User Services Helping providers get their data to iDigBio. Facilitating digitisation, mapping, data export,...	Getting useful vouchered specimen data into the iDigBio portal for anyone to use for their purpose / goals.	North American vouchered specimen data	everything
Lyubomir Penev	Publishing and dissemination of scientific knowledge in taxonomy, ecology and conservation	Publisher and technology developer	Publishing and dissemination of synthesized knowledge of a flora, fauna or mycota of a region.	Worldwide	Biota
Johannes Penner	Taxonomy (partly ecology, biogeography, macroecology)	Macroecologist (analyses, trait database compiler, field ecologist, biogeographer) + IUCN Red List (contribution to West African amphibians & reptiles; + Red	Gathering information on species (see above) + searching primary literature + keeping updated on current taxonomic knowledge	Depending on work (see role): World, Africa, West Africa, Liberia	Amphibians & reptiles

Your name	Primary focus of your institution	Your role	Your use of flora/fauna/mycota in the above role	Geographic focus	Taxonomic focus
		List Authority Coordinator Viper Specialist Group)			
Eckhard von Raab-Straube	Taxonomy, collections (herbarium and living), phylogeny	Taxonomic data and workflow curator, Checklist manager and editor	Extracting, databasing, comparing and critically evaluating taxonomic, nomenclatural, distributional and additional (e.g. common names) information from Floras	Europe, the Mediterranean, Atlantic Islands and Caucasus	Vascular Plants
Marianne le Roux	Biodiversity, taxonomy/systematics	e-Flora Coordinator and taxonomic researcher	Compilation of an e-Flora, i.e. the following is necessary: - populating the database with relevant information (use-cases) - extracting data such as descriptions, distributions, geographical checklists, morphological, habitat and ecological data	South Africa	I am studying provincial Floras to compile the e-Flora for South Africa. My own taxonomic research is focused on Crotalaria (Fabaceae) and Pelargonium (Geraniaceae)
Yashica Singh	SANBI focus is biodiversity: systematics, conservation and dissemination SANBI KwaZulu-Natal Herbarium (Durban) focus is plant taxonomy	Herbarium taxonomist, collections management	Identification, morphology, distribution, habitat, scientific curation Other: number of species, valid names, common names, flowering times	The KwaZulu-Natal Herbarium is regional, covers eastern seaboard of South Africa (KwaZulu-Natal, Eastern Cape Provinces) Taxonomic research is in the Flora of southern Africa region (SA, Namibia, Botswana, Lesotho, Swaziland)	Families: Araceae and Hypoxidaceae in southern Africa
Daniel C. Thomas	Taxonomy, evolution	Researcher in Systematic Botany (molecular phylogenetics, historical biogeography, alpha-taxonomy, implementation of the e-Flora Malesiana)	Identification, extracting species distributions, extracting habitat and morphological features	Southeast Asia	Flowering plants

Your name	Primary focus of your institution	Your role	Your use of flora/fauna/mycota in the above role	Geographic focus	Taxonomic focus
Jonathan Timberlake	Plant taxonomy and plant sciences research	Flora Editor. Conservation projects. Liaison for southern African region	Editing of flora treatments. Also using when it comes to secondary information (e.g. endemism, distribution, conservation status) of plant identification.	Flora Zambesiaca area in particular (Botswana, Caprivi, Malawi, Mozambique, Zambia, Zimbabwe), but generally southern and eastern Africa	For the Flora, all outstanding families. For myself, legumes, Acacia, Brachystegia
William Ulate	Literature / Floras / Botany	Provide basic information to users, develop systems, design solutions, facilitate knowledge extraction	Provide tools to mark up, mine information from text, discover structured knowledge	Worldwide	Legacy biodiversity literature in general
Holly Vincent	Conservation	PhD student, focusing on conservation strategies for priority global CWR.	Species distribution modelling, mapping of species occurrences	Global, with a focus on Middle East	Working on a global list of important crops such as legumes, cereals, fruits and vegetables
Mark Watson	Plant systematics	Herbarium taxonomist, floristic researcher	Identification Reuse of data - distribution, morphological description, nomenclature, further references, illustrations, taxon concepts, classification, voucher specimens, keys for characteristic characters Teaching - good and bad practice Research project development - critical comments flag up potential areas of taxonomic research (e.g. Flora Europaea). Contacts - authors of accounts as taxonomic experts to ask for further information/help Literature investigation - major cited references/monographic treatments Curation of herbaria -	China, Himalaya (especially Nepal)	Apiaceae, but these days more of a generalist

Your name	Primary focus of your institution	Your role	Your use of flora/fauna/mycota in the above role	Geographic focus	Taxonomic focus
			reclassifying to a new system and also finding duplicates of cited specimens		
Zerihun Woldu	Ecology	Field Ecologist	identification, extraction of species distribution	Ethiopia, east Africa	All
Shuangxi Zhou	Ecology	PhD student with field transect work	Identification, extracting species distribution, leaf traits, plant function, information on species response to environmental stressors	South Australia, Southwest China, Mediterranean area in Spain	Major tree and shrub species

Annex 3. Workshop outputs - Description of the use-cases

Workshop outputs

The following section documents the "use-cases" developed in the small workshop groups. The text enclosed in {} are part paraphrased, based on transcripts of recordings of the participants describing their use-case. Where referred to in the text, the ID numbers of the activities are enclosed in []. In the workflow diagrams the activities are colour coded according to their relative difficulty: red, hard; orange, medium; green, easiest.

Use-case 1: Making an IUCN Red list assessment (1)

Volololiaina Jeannoda, Yong-Shik Kim, Johannes Penner, Yashica Singh (facilitators Don Kirkup, Soraya Sierra).

Volololiaina :{The first activity is setting the priority for red listing [0], the priorities can be on species, regions or threatened species. For species it can for example take a <?> species or commercially traded species. Then you make a species list [1], with names that are taxonomically accepted. Then you have to compile the information from different sources [3], these can be published (including databases or web articles, journals) or unpublished "grey" literature. The specific information required for red listing s extracted [4] and the file is verified [5] by experts (who may be part of the group). Capacity building and training of the groups [6] also takes place within RLA process. With specialist groups, as in this case, the assessment can be carried out by the group itself within a group workshop [7], the participants will include experts on the species, on ecology, habitat assessment and so on. Alternatively, the assessment can be done by experts [8] previously trained and the assessment is then reviewed by the group [3] during the workshop. Publication of results [10] includes recommendations, conservation measures to be taken. Funding is a problem but in the case of Madagascar we work with what we have: once we identify data deficient taxa, we then try to get funding to work on them.}

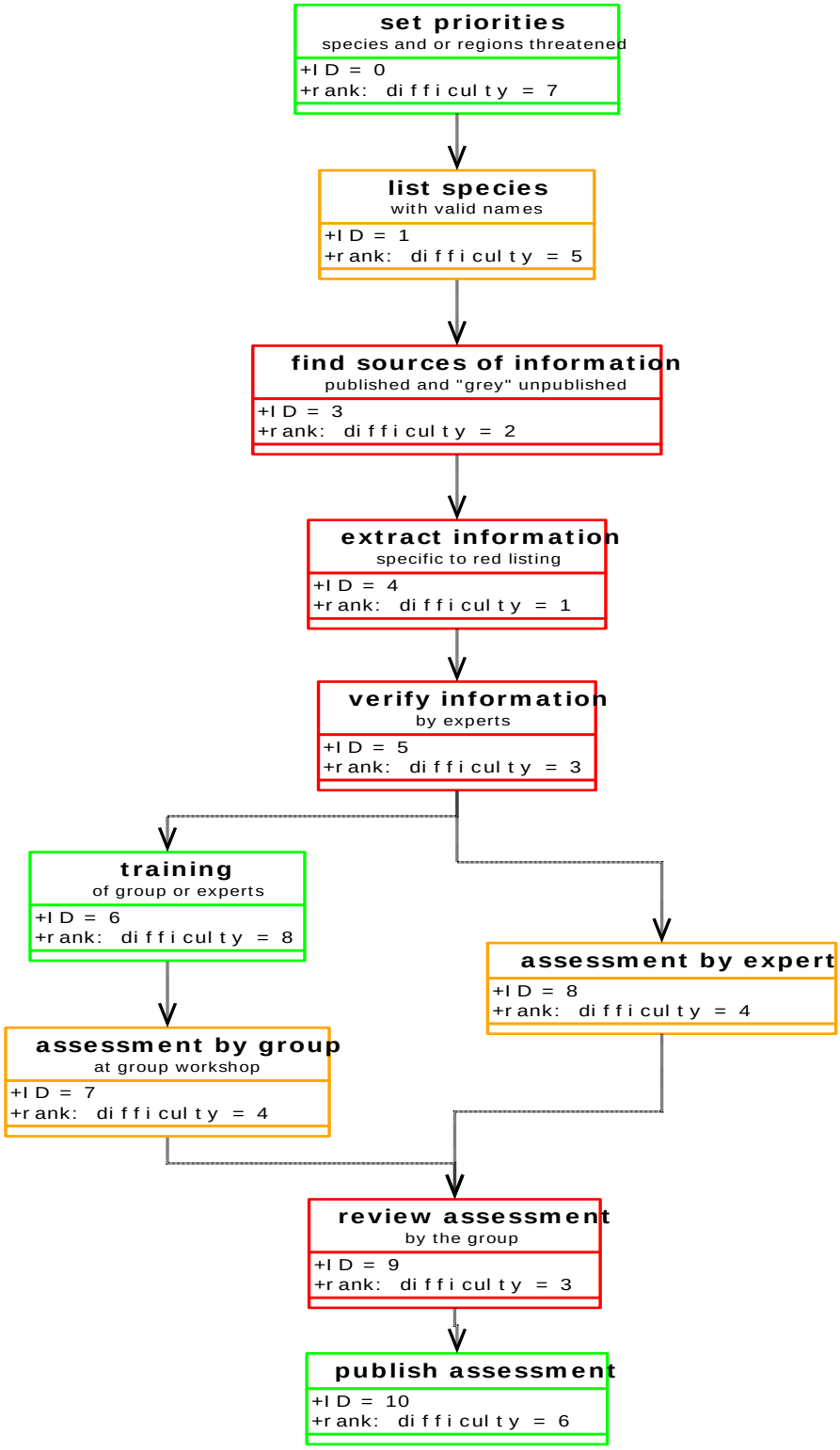

Figure 2. Making an IUCN Red list assessment (1).

Use-case 2: Making an IUCN Red list assessment (2)

Henk Bentje, Melanie Bilz, Viola Clausnitzer, Pablo Demaio (facilitators Don Kirkup, Soraya Sierra).

Henk: {The first thing is to assemble the team [4], botanists, technicians, who know the local circumstances and threats to populations, desk scientists and people who have done assessments before. Then you have to do the preparation [5], prepopulate the SIS forms, georeference specimens' records to maps ready for the Area of Occupancy (AOO) and Extent of Occurrence (EOO) [estimates].

In the actual assessment, problems emerge with gaps in the taxonomic data [6] and gaps in the distribution data [5], real problems which will either lead to data deficient taxa, or else "slightly less well proper" assessments. Population data [8] are often missing because they are not mentioned in Floras, they are in the brains of the people who do a lot of fieldwork, which is why you have to get these people to the actual workshops.

Data that are also missing as well are data on ecological vulnerability [9]: habitat data but also the specific reactions of these taxa to eco-threats, or impacts of various other kinds. Information on threat [10] is vital and is also very difficult to get a hold of - you have to get those local people with the threat information, and this all feeds into the selection of participants. These problems all feed into the assessment process itself, which means that some of the time, although not desirable, data deficient taxa or data deficient assessments are inevitable.

The final aspect of the workshops is capacity building [12]. Many field botanists have not done assessments before. At the end of a good workshop they will have done a couple of assessments under the guidance of other people and they feel comfortable with it, which means next time in the field they will come up with slightly different field notes: they'll make more notes on the threats and the populations.

Once the workshop assessments have been done the processing for publication of the official assessment [14] happens at the IUCN, Cambridge and this is a major bottleneck, as currently, owing to largely financial constraints, only two people are employed on the task of processing all Red List assessments. The result is that although a lot of assessments have been done, only a few appear on the official list.

Although attendance at the workshops is funded, -many people-invest a considerable amount of their own time into the assessments for which they are not paid. It would be really useful if each individual assessment was recognised as a publication which would help with finance and with the assessment of people as well as that of species.

Another financial bottleneck is with re-assessing. As well as researching the data deficient taxa [16], we also follow-up taxa that are either highly endangered or exceeding vulnerable: these taxa need to be monitored [17] on a regular basis and there is very little money for this, even less money than there is for funding the workshops themselves. This places a severe constraint on the proper

assessment process. Data deficient taxa really need fieldwork to gather more data (e.g. What is the distribution?. What are the threats in areas that may not have been visited since 1924 or even earlier?). Field work to monitor species that are at very high risk is also essential for a proper Red List process.

Finally, after the assessment has been officially published, we get the implementation phase at various levels, from national to global. In part this comes down the areas with the most threatened species which are usually the tropical regions, and these are often the regions with the least money to follow-up the recommendations or (act on) the threat levels. The rich nations, we feel, have a responsibility to help the areas where the high biodiversity is and where the problems are, to address those problems there is an international responsibility. The red list data feed into the prioritisation of conservation actions and into conservation planning. We know from practical examples that big environmental impact assessments can only use things that have been properly published (i.e. once they have cut through the IUCN bottleneck) and there are many cases where we know that assessments have been made, haven't been published yet and they are unable to be used in conservation planning, so this bottleneck is a vital one.}

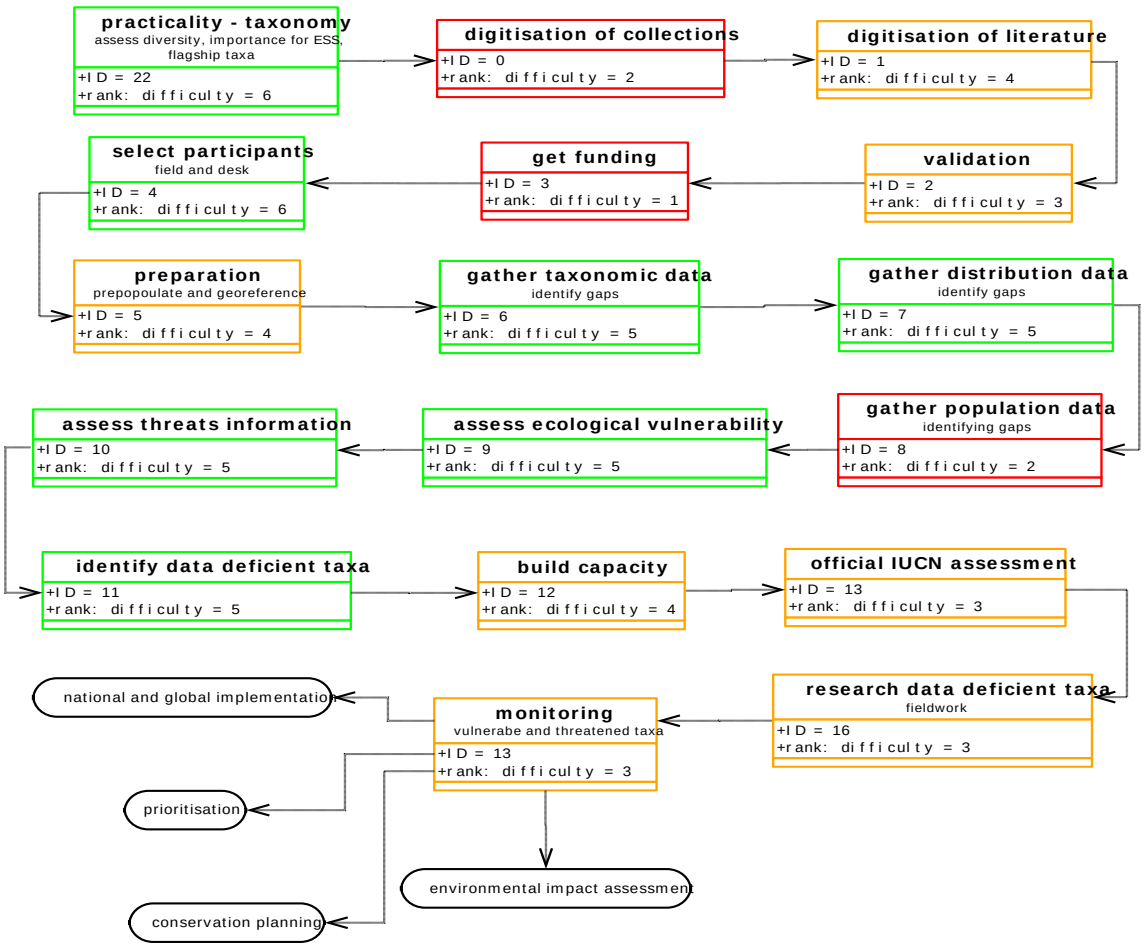

Figure 3 Making an IUCN Red list assessment (2).

Viola: {Digitisation of collections [0] takes a lot of real scientific work, and may take weeks or months. Population data might be available for some groups but mostly is not, and for some groups, such as invertebrates, it is impossible to gather within a realistic time-frame and we have virtually no data. Obtaining funding [3] takes a long time and results in bottlenecks, such as the official publication of assessments on the IUCN website [13]. As well as requiring funding, researching data deficient taxa [16] often involves one or more expeditions into remote areas and thorough searches requires a long time-scale. Validation of data [2] is also time consuming, for example all the synonymies cannot be true if there is a misidentification made even more complicated if you don't have the specimens but only the literature. The technical process of the digitisation of literature [1] is also time consuming, for example we sent pdf publications to India for digitisation and we get the information back. Capacity building, preparation and georeferencing [5] are also time consuming.

Assessing ecological vulnerability [9], gaps in the taxonomic [6] and distribution [7] data is straightforward to extract once the harder to obtain information is available}

Use-case 3: Plant-trait database compilation

Henry Ford, Jens Kattge (facilitators Don Kirkup, Soraya Sierra).

Ed: {This use-case is based on the participants' experiences with two different trait databases: The Ecoflora of the British Isles, and the TRY database. Whereas the databases have a lot in common, there are some differences. The Ecoflora has an open access web portal whereas TRY protects the IPR of its data donors. This difference would appear pertinent to the comments below on difficulties experienced with data release and in monitoring user feedback.}

Henry: {First, what are traits [0,1] and what are they used for? This is initially a committee decision. Historically, the traits that could be found from the accessible Floras (e.g. Flora of the British Isles, Clapham Tutin and Warburg) - were actually quite useful; one of the people came up with life history strategies based on trait selection e.g. guerilla and phalanx strategies for plants (see Harper 1977). The initial committee decision deciding on what traits you think that you ought to have, in fact what happens is that you generally take anything that you can get your hands on, and this involves finding data sources [2], and then you have to put it up on a database. This is a fairly critical item: you do not want to spend your time re-designing your database. Modern systems can accommodate large flat files so it is not necessary to have a normalised data structure to save space and you can use a program structure to get the data out of it. Having found the data sources and decided what a trait is, prioritising your data selection [17], is again, generally you take anything you can get your hands on. It is a continuous process, trait selection, database structure [4], getting a little bit of user feedback [12] if you possibly can, prioritising data on the basis of which traits are most useful, and going back and re-programming. From here people need to extract the data so you need a process of data release [8], programming is part of this, quality assessment [7] - this is part of the data curation process [6]. Designing the web or personal download access [10], possible going back into scientific publication sometimes [9]. Also from the data extraction and user feedback you need to find out what the traits are being used for: generally we have no idea, people don't tell us what things are being used for. As an ecologist, I have personally used for community ecology [13] and vegetation modelling [14], invasion ecology [15]. With the Ecoflora there is also a lot of non-specialist use [16] - with people just looking up particular plants and seeing what information we have on them. Data curation [6] comes back into long term sustainability: because we have already lost one or two databases, the funding has gone, they are not being maintained and you can't get hold of the data. For example, the cost to keep the Ecoflora database is £70 pa, but in order to put more data up and ensure sustainability requires more funding. }

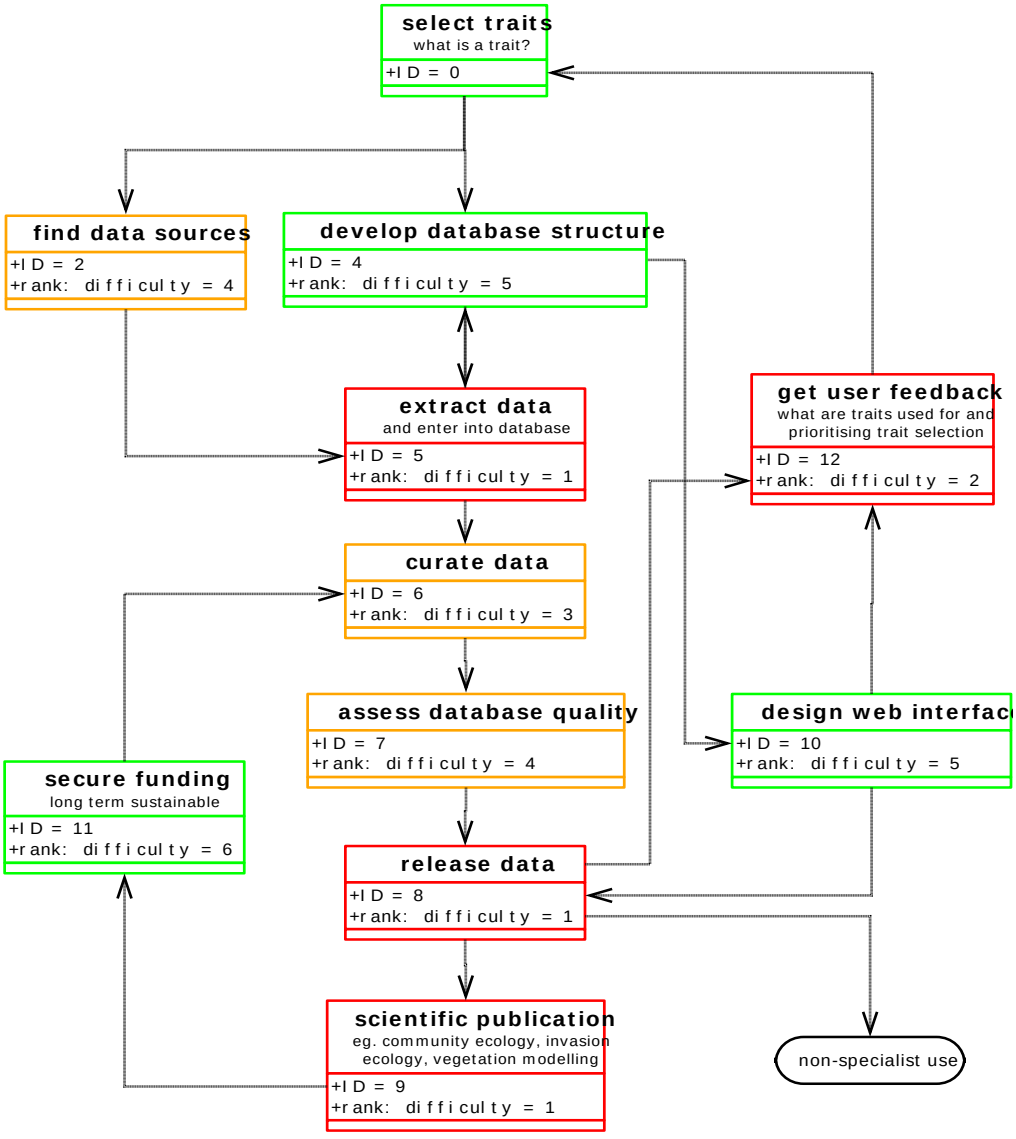

Figure 4. Plant-trait database compilation.

Use-case 4: Linking ecophysiology to vegetation modelling

Shuangxi Zhou (facilitators Don Kirkup, Soraya Sierra).

Whereas use-case #3 dealt with compilation of trait databases, this example demonstrates a downstream use of databases such as TRY.

Zhou : {My research project aims to establish a synthesis of field experimental data on the response of different plant functions to environmental changes, systematically study the relationship between plant traits and processes and key environmental factors. By incorporating recent advances in plant ecophysiology and biophysics and rapid accumulation of quantitative data on plant functional traits into current vegetation dynamics models, the project can contribute greatly to the evaluation and improvement of dynamic global vegetation models. The first paper, "How should we model plant responses to drought? An analysis of stomatal and non-stomatal responses to water stress", is in press now by the *Agricultural and Forest Meteorology*. Two glasshouse drought experiments in Sydney and Barcelona, and two transect fieldwork in south Australia and southwest China will be completed this year. (Details: *https://sites.google.com/site/shuangxizhou2014/publications*)}

The workflow is quite simple. "Establishing the transect network"[6] entails a lot of field work and is the most difficult and time consuming part of the process. "Identification and mapping of species" [1] is relatively easy (since there are only a few target species to deal with). The ease of "Collecting global trait data" [2] depends on the availability and the format of data available for the species of interest in trait databases such as TRY.}

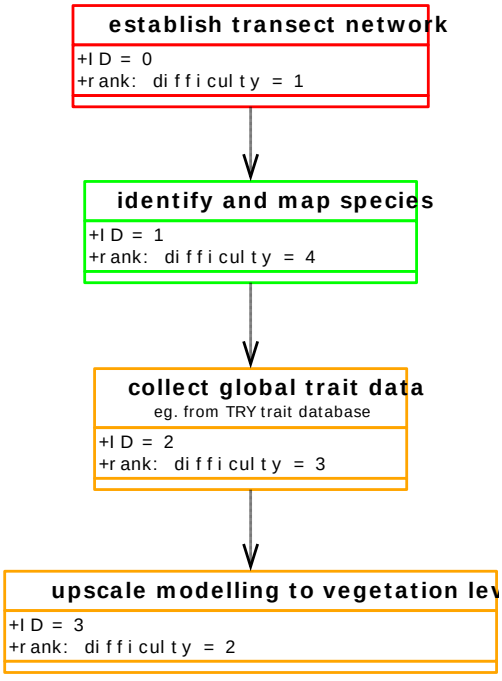

Figure 5. Linking ecophysiology to vegetation modelling.

Use-case 5: I want to describe a new species

Norman Johnson, Luciana Musetti (facilitators Jeremy Miller, Alan Paton).

Norman: {First we need to acquire specimens (e.g. field work or loans from other institutions) then capture the specimen label data, adding georeferences to the localities. After capturing the label data we go into the process of developing a character set which can come from the literature or from examination of the specimens, and we develop taxonomic concepts on that basis. We revise our characters on the basis of that sorting, then in turn we revise our taxonomic concepts (- we can go on to repeat the process forever in a circle). We can also get either of those from the literature. We capture images for the specimens, locate the primary types and compare the primary types to the taxonomic concepts, then summarise the geographic distribution and characters for the taxonomic concept. We then coalesce all of that and submit it somewhere for publication. Finally we return the specimens.}

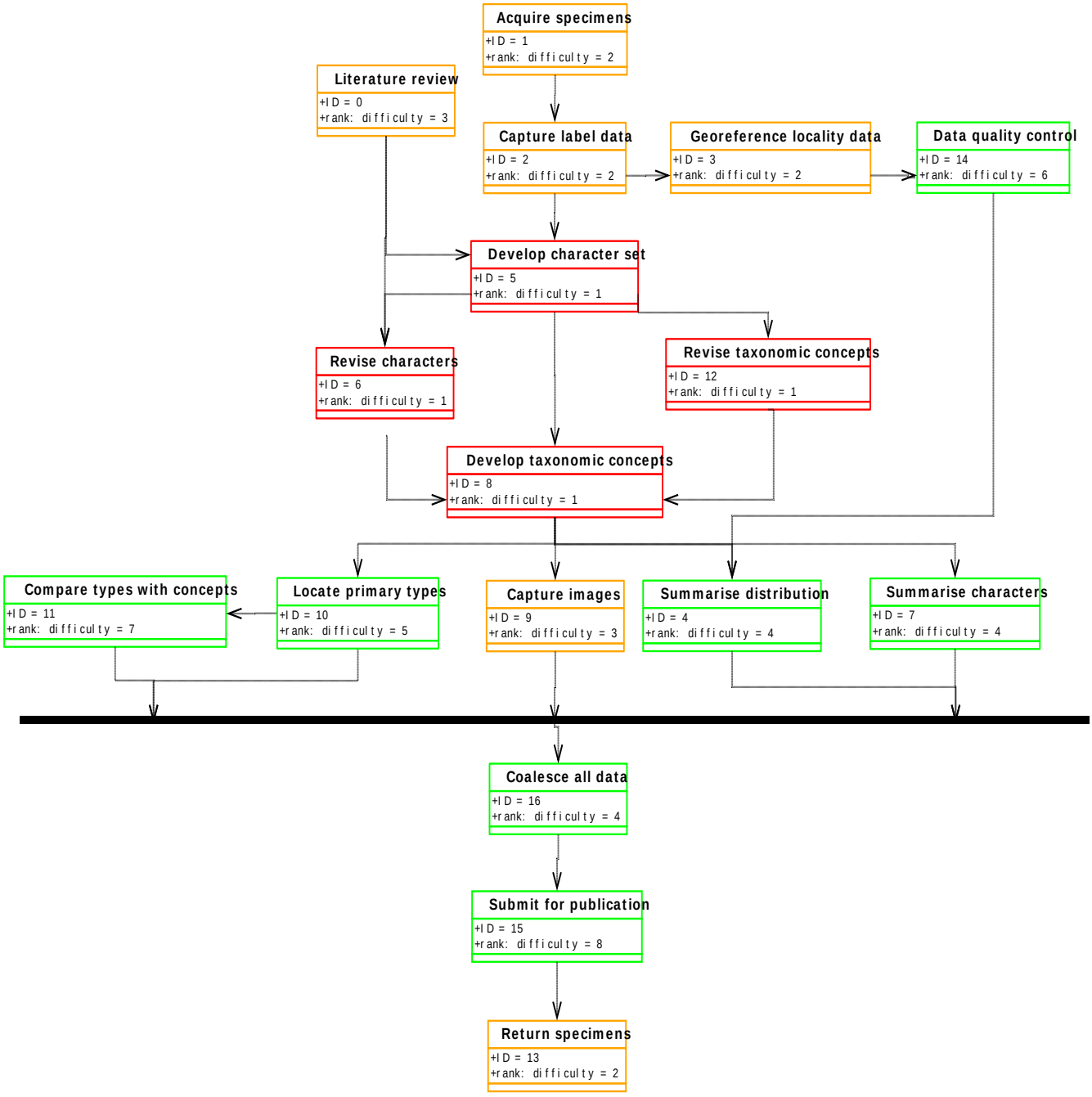

Figure 6. I want to describe a new species.

Norman: {For the most difficult activities what we have at the top is the feedback loop of testing our taxonomic hypotheses [5, 6, 8, 12], going back and forth, and we don't know where it ends, so we have a kind of quantum gap here between these and the next activities: acquisition of specimens [1], either finding out where they exist in collections and getting them, or doing field work - it turns out returning specimens to the right place takes a lot of time.}

Use-case 6: How do I identify a plant?

Eduardo Dalcin*, Bente Klitgaard*, Marriane le Roux, Daniel Thomas, Zerihun Waldu (facilitators Jeremy Miller, Alan Paton).

6a* (local botanists in the field)

Bente: {Once finding a plant in the field we observe the plant and mentally collect and process the data. We then make a decision based on our experience, such as plant smell, or whatever things we observed in the field [5], or based on photographs from photo-picture books or e-guides which are actually quite useful [4], we make an initial field identification [1]. If we can't identify the specimen, we always take the material back: process it (usually as herbarium specimens) and (if they are available) use more sophisticated field guides, floras or e-identification tools, which then will allow us to match it in the herbarium.

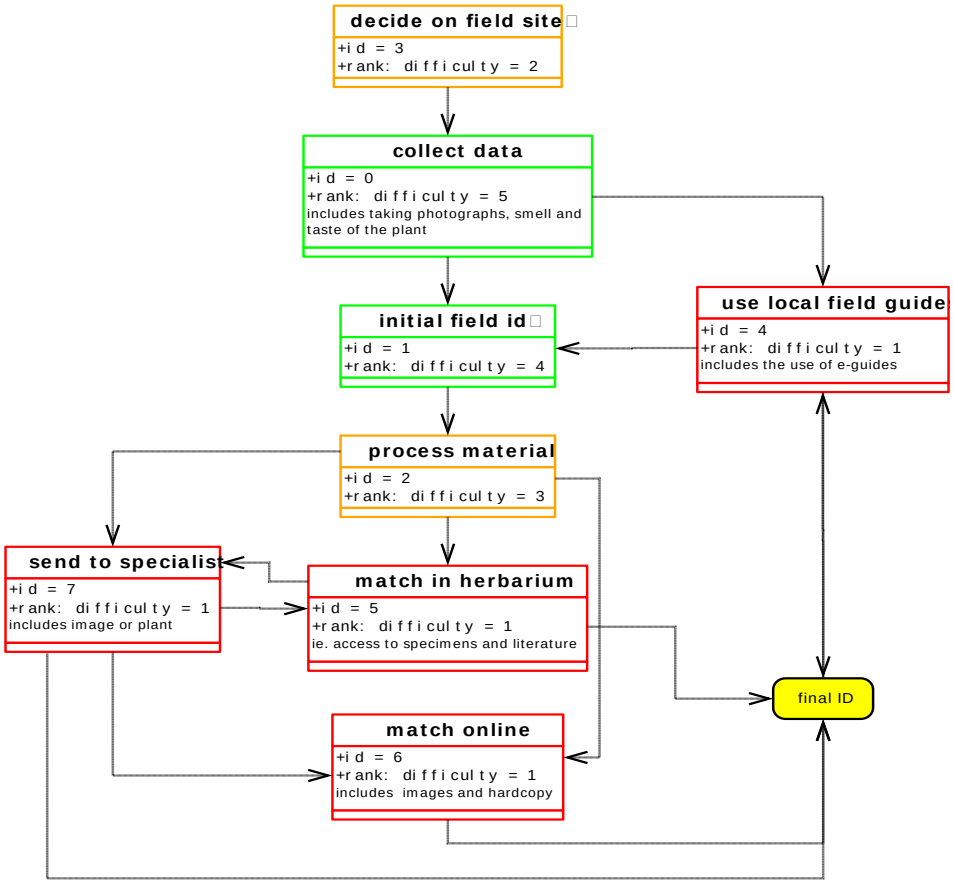

Figure 7. How do I identify a plant? (in the field).

This process may be iterated several times until we get a final ID. If we cannot identify the specimen ourselves, then we may have to send it to a specialist [7], either as an image first, or the plant as such. Our inability to identify may be because we don't have

enough reference material to match or not enough literature (access to field guides or e-material) ourselves. Once the final ID is made it can then be fed into new field guides or identification tools.}

6b (general identification in the herbarium)

Zerihun: {In identifying a plant the first thing to do is to collect a voucher specimen [6] and simultaneously collect field information (which consist of GPS coordinates, altitude and all other site information). For the pre-identification [7] we use "expert knowledge". After arriving at a [provisional] identification we process the specimen and send or take it to the herbarium. Once in the herbarium we can match with specimens in the collection or use quick guides or Floras including species keys. Duplicate specimens can be sent to other experts for their advice for identification. }

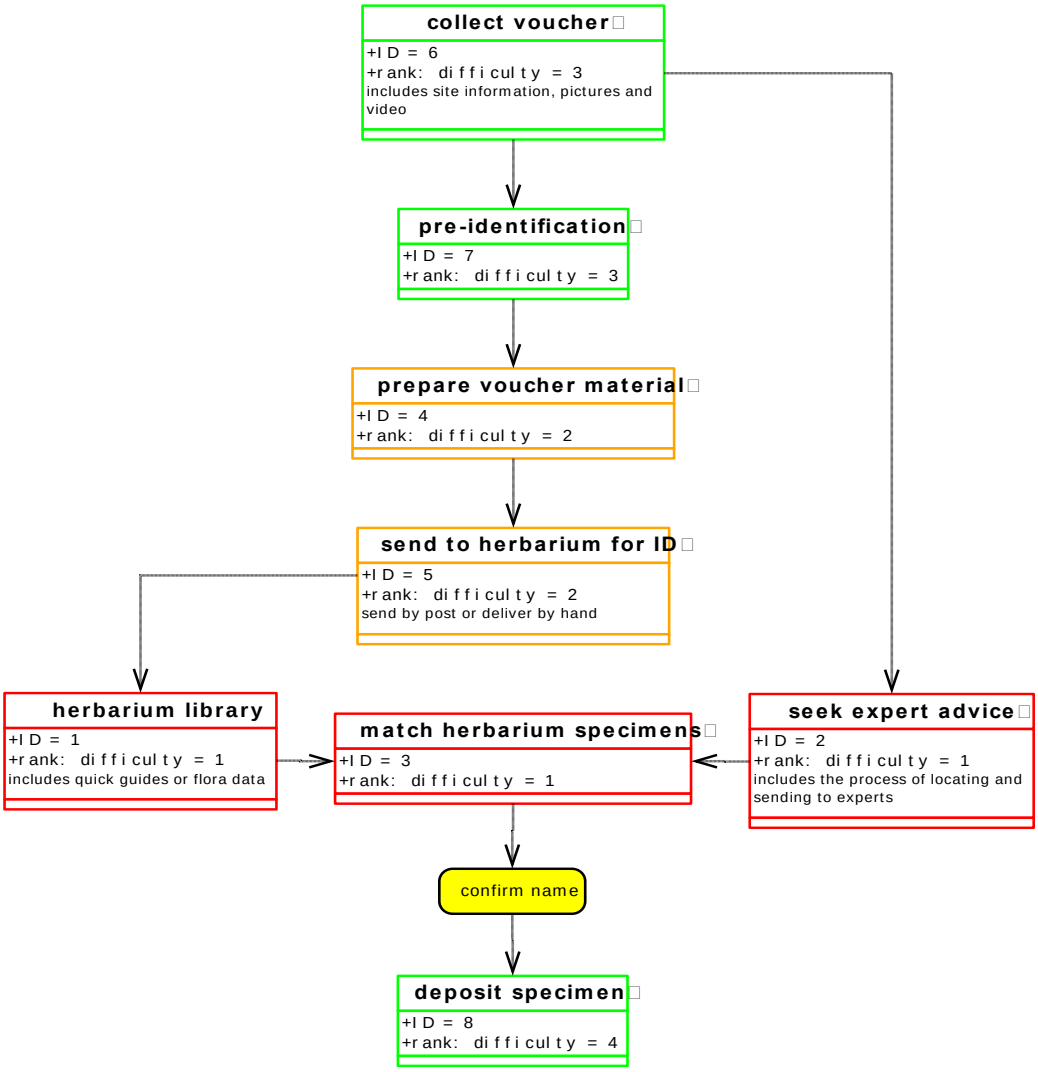

Figure 8. How do I identify a plant? (general).

Daniel: { Getting input from the experts [2] at the top here as it is the most time consuming step because you may have a whole batch of specimens from a general collection and you need species-level ID for example for a survey. So you have a lot of collections, some of which may be easy to identify but many may need expert knowledge for identification: you need [to reach all] the people who know where the necessary information is located. The preparation of the specimen [4] including the physical step of handling the specimens and the databasing and collecting all the information [6], although vital for identification, is not so time consuming as it usually involves a much smaller group of people.}

Use-case 7: I want to prepare a quick and dirty flora account for a taxon

Mark Watson (facilitators Jeremy Miller, Alan Paton).

Mark: {This use-case describes a quick and dirty Flora account for a country which doesn't have a Flora yet (which is what we are doing for example for the Flora of Bhutan). We are not looking at a deep monographic type flora but are looking at what we can do quickly to get some information out there which can then be the first cut of a more sophisticated, deeper Flora. We are looking at how much data we can reuse and the degree to which we can reuse that data.}

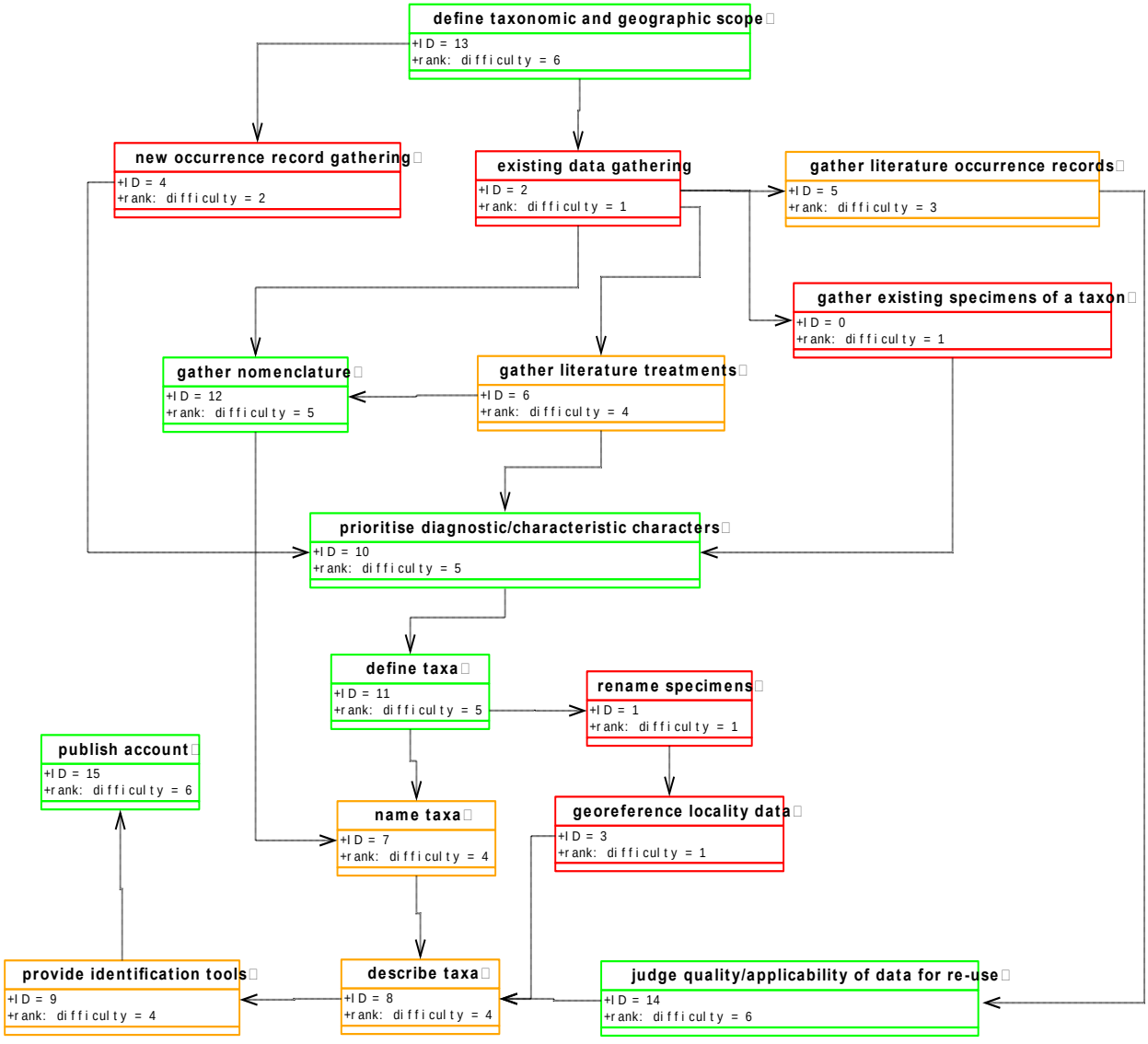

Figure 9. I want to prepare a quick and dirty flora account for a taxon.

Mark: {First we have to decide which area and which taxonomic groups [13]. Then we go into the data gathering phase - both existing [2] and new data [4]. For example you may need to do fieldwork where you feel there aren't enough specimens from a particular area. Existing data gathering we can split into getting occurrence records from the herbarium specimens in the collections [0] or there might be literature references of occurrences [5] as well, not just to herbarium specimens, but also from secondary sources. There also could be literature treatments [6] - in Floras and monographs - thus drawing on existing information to see if we can reuse all of that. Once we have a handle on the names of the taxa we are going to be dealing with, then we can start looking for nomenclatural information [12]; the place of publication, their types and so forth. We need to judge the quality of the data, for example a statement in the literature that a plant occurs in this particular area [5], we have to decide whether we will use it or not, and how we would reuse it [14]. Specimens can be re-identified [1] and georeferenced [3]. Once you have all the existing data together plus your new specimens, you often use the past literature to prioritise the taxonomic characters [10] you are going to be homing in on, because you haven't got time to do a full monographic treatment: so you may have a group of five species and you look at what are the features that other people have found to be important for defining the taxa. You have a quick definition of the taxa [11]. If there any taxonomic problems, usually you will be time limited as to whether or not you can investigate them. You may just have to put in what you think and put in some taxonomic records to identify the problem. Once you have defined the taxa (the classification side) you can go on to name the taxa [7] (the nomenclature side), and once you have names for your taxa you can go on to rename the herbarium specimens, and you can see what you can do about renaming existing literature records, sometimes you can rename and reuse them, but if you split a taxon you can't reuse the existing literature. Once you have defined and named your taxon you can describe it [8]: morphology, distribution, habitat (you might be lucky enough to have your own data collected in the field but often gained from herbarium specimen labels and the literature, again reusing data). You can then look at providing identification tools [9], whether it's keys or illustrations. You can then publish your paper [15], whether that's electronic format field guides or in print.

Gathering existing specimens [0] usually involves visiting herbaria and takes a long time. Renaming specimens [1] and georeferencing [3] takes a long time. Doing new fieldwork takes a long time [4]. Gathering literature records takes a long time because they are sprinkled across difficult to access publications, local journals but their value may be limited so we don't spend a lot of time doing it - it would be worth doing well but it would take a long time. At an intermediate level: gathering literature treatments [6] is so much easier to find with BHL; naming taxa [7] is pretty easy for experienced nomenclaturalists, but it could take other people a lot longer; describing taxa [8] doesn't take too long once you've done all the groundwork; providing identification tools [9] is a bit open-ended, if you write a key it can be quite quick, but if you try to produce more sophisticated ID tools then it can take a lot longer. Getting down to the relatively quick things: prioritising characters [10] is relatively quick once you've got the information in place; defining taxa [11] for a quick and dirty flora we can't afford to spend much time doing that (c.f. describing taxon); gathering nomenclatural information [12] is pretty quick, most of it is in BHL. And then really quick; defining the area [13]; judging the quality of data [14]; publishing accounts [15]. }

Use-case 8: I want to publish and disseminate high quality taxonomy

Laurence Bénichou, Patricia Mergen, Charlotte Thionois (facilitators Jeremy Miller, Alan Paton).

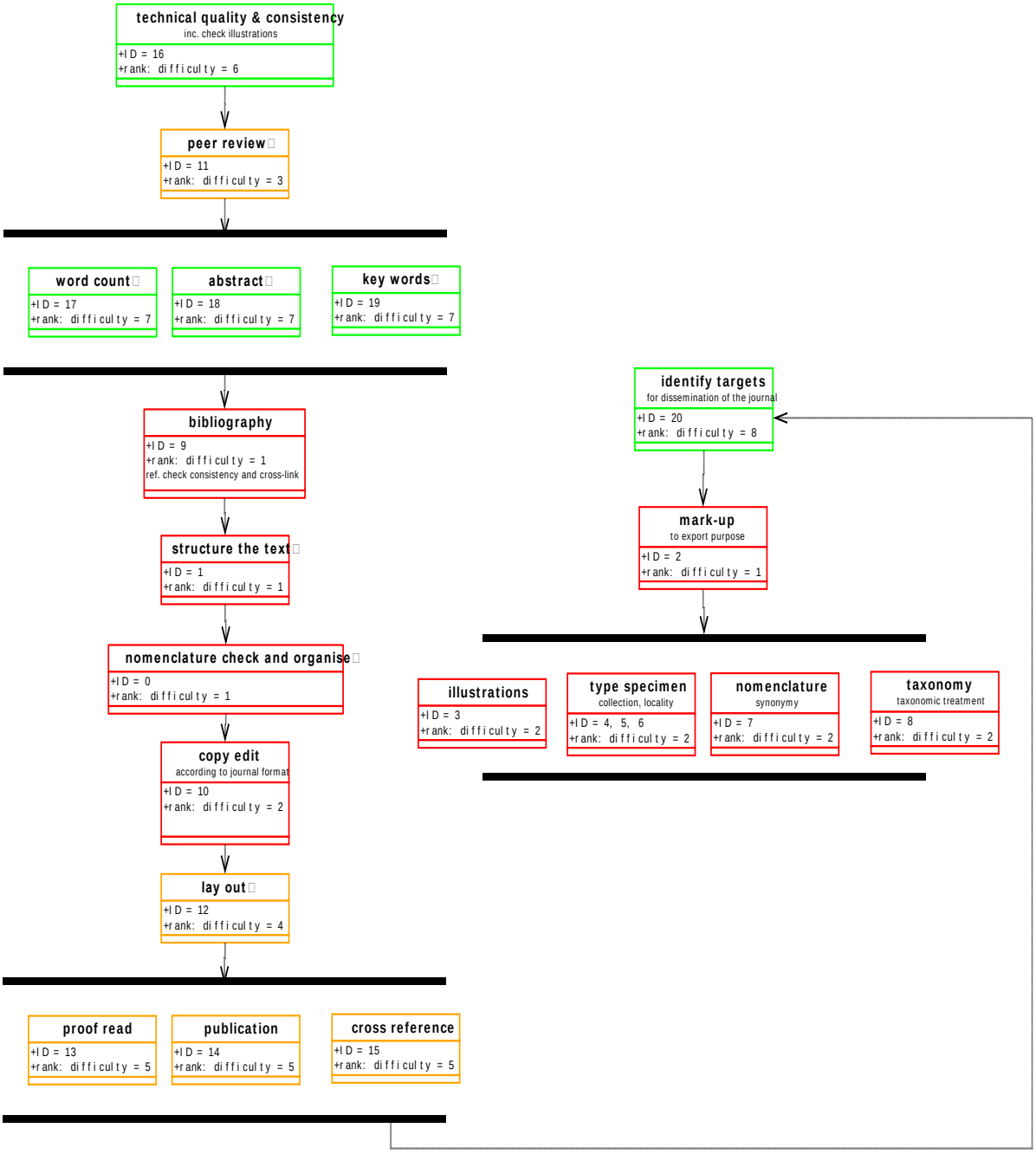

Figure 10. I want to publish and disseminate high quality taxonomy.

Laurence: { From the desk editing point of view, when we have an article submitted to a journal, first of all we check [16] that everything is in order – e.g. that the illustrations are consistent with the text, that their captions are clear, that nothing is missing. Then we send it for peer review [11]. Our work truly begins only once, the paper is accepted. We check that the numbers of words [17] are what we expect, that the keywords [19] are appropriate for the article, the consistency and citation of all the references within the paper [9], the consistency of text and cross-link all the references, we check the spelling and typesetting, we standardise the text, footnotes if any, the bibliography according to the journals' instructions and use in the field. We check and organise the text in order to structure it [1] for an appropriate journal, but we often struggle with the organisation of the nomenclatural section and the materials examined [8]. There are two reasons for this; the more structured is your text the better the dissemination will be as the result markup of the text will enable the extraction of data, and of course we need to structure each article the same way all the time. This is very difficult sometimes because each author is particular. Then we copy edit the full text and illustrations [10] according to the journal format, layout [12], proof read [13]. We then send it to publication [14] and then send the numbers to cross-reference [15].

The following activities are carried out in parallel to the above: we have to identify all the targets [20] in which we want to disseminate the journal, this can change as some of the databases die; we have to mark up [2] all our text (illustrations [3], type specimens [4,5,6], nomenclature [7], taxonomy [8.]) in order to export it to databases; we have to send all the information to the databases, the dissemination is almost as important as the production of course as the access to the information is critical to our field.

Probably the most time consuming activity is the nomenclature check [8] because we need to go back to the scientist, expert or editor for that particular purpose. Structuring of the text [1] and markup [2] are similar and depend on the number of species in your article and the degree of markup we want to perform; the bibliography [9] is the same, if it is very long then it can take a long time to check. The copy editing [10] and the proof reading [13] could be one way or the other. I have separated peer review [11] from the others because it's not something that we do ourselves but have to find others to do this. However, we have to manage the peer review process, find referees (which is sometimes very difficult), then send the referees' reports back to the authors and check that the revised manuscript is revised accordingly with their remarks before it is accepted (or rejected) by the editor in chief.}

Use-case 9: I want to carry out a plant survey of a small national park for management's decision making

Christopher Chapano, Chuck Miller, Eckhard Von Raub-Straub, Jonathan Timberlake (facilitators Jeremy Miller, Alan Paton).

Jonathan: {This use-case was something [that turned out] much broader than we had actually chosen. "I want to carry out a plant survey of a small national park for management's decision making" So it's very clear objectives, a specific area for a specific purpose. It's for management so it's not to do with the creators of the information but with the users. There actually should be a couple more sheets [of activities] inserted before the bottom line but we ran out of time. The first things that we do is to gather a map of the area [20], define the area [13], understand from management what it is that they are after [24], which is a pretty critical point. We then gather existing information, studying Floras, checklists [5]; collating existing information on plants [12,13], this is going to be a critical one if your survey is going to be looking at individual species as opposed to habitats, identify some of your threatened species [11], at least ones that have already been documented as threatened.}

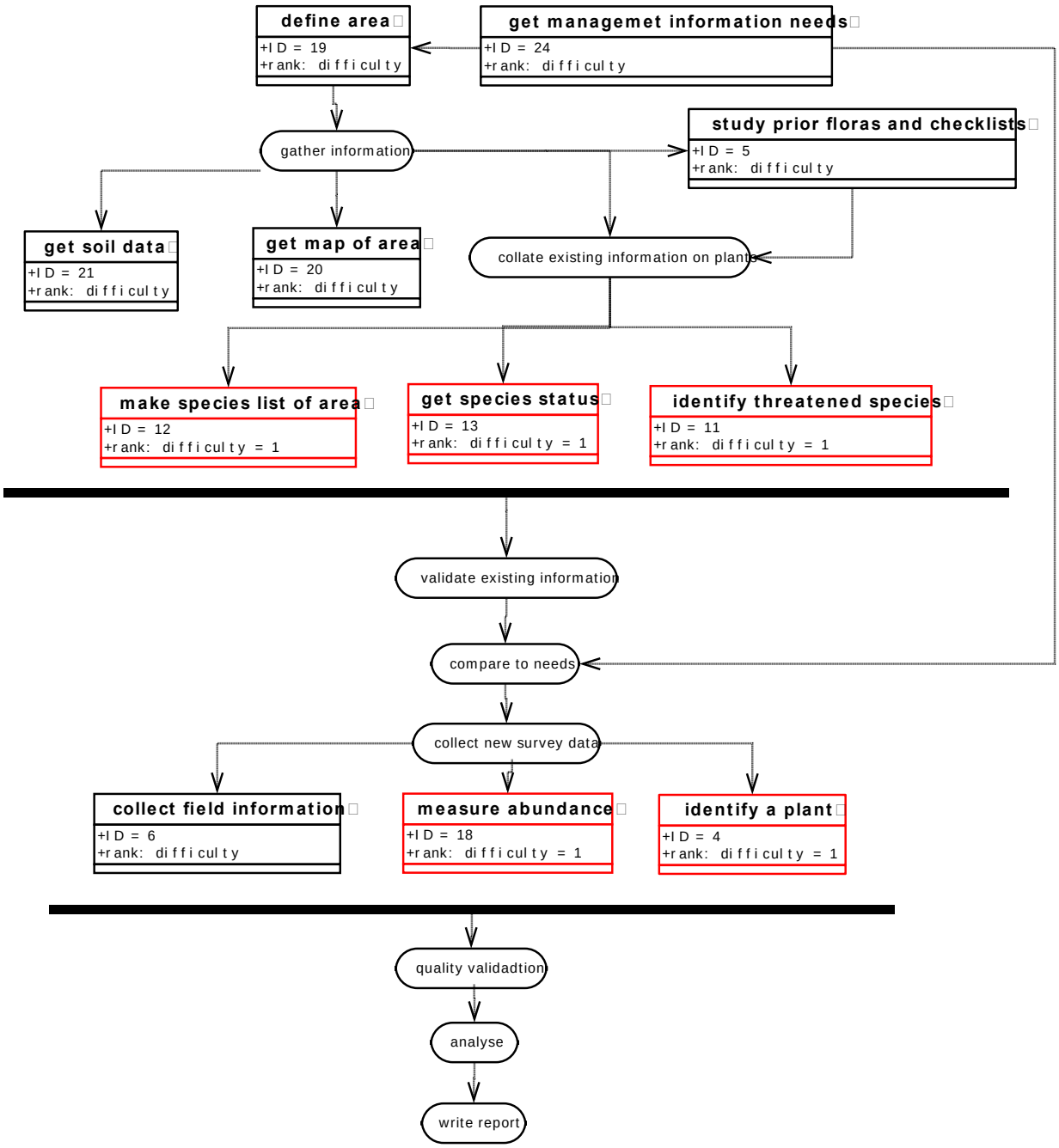

Figure 11. I want to carry out a plant survey of a small national park for management's decision making.

Jonanthan: {So it's collating existing information, trying to validate it, and then to identify what you are going to do in the field. So you go out into the field and record information [6] (it's this decision making network for recording field data which lacks detail in the diagram). The sort of information you are going to collect depends on the sort of information that management needs: they don't necessarily need full checklists, they certainly don't need information on subspecies - perhaps, unless they are threatened

species. They might need very coarse level information in which case you can run through quite quickly and come to the point where you guys... <recording ends> }

Chuck: { Synthesising the data and analysing it takes about 40% of the time, 30% on collecting and 30% on processing. So taking samples, identifying them, identifying a plant in the field [4], studying the prior Floras and checklists [5], gathering all your information, collecting your field information [6] - those are the things that are time consuming.

Getting down the gathering [8], validating [9], collating [10] identifying threatened species [11], making lists for the area, which is where a flora or checklist would come in. Then getting the status of the species that you put together by doing all this work.

Using identification guides [14], comparing how we are doing versus what we started out being required to do [15], validating the quality of what we gathered [16] then sorting and sifting out [17] - do we have too much detail or not enough detail? All that kind of decision making as you do that kind of project.

Finally, doing something like measuring abundance is down after analysis, then all the steps related to the beginning of the project like getting the soil data, putting together the needs, these are all things that don't take a lot of time.

In other words, easy to say, hard to do. You could say "Just go write that report" but there's a huge amount of work for that simple request.}

Use-case : 10 Producing a Digital Flora

Andreas Muller, Quentin Groom (facilitators Donat Agosti, Eva Kralt).

Andreas: {We wanted to show the workflow for prospective, not legacy data, these are all the steps you have to go through. First there are all the different types of information that you need to gather [11, 12, 13, 15], then the process of cleaning [9], putting it all in a database or some other digital system[8], then the editorial process [7] to improve the data and bring it into a form that becomes a flora. Then we have the applications; we want to publish things online [13], or on paper [14], or we want to create keys [3,4] which is an important part of the flora. Another important part is to have the links to the sources.}

Quentin: {When thinking in terms of costs, the original research, the editorial side of things and the putting it all together at the end are actually the most expensive parts since it's that which takes the most time and involves the most expensive people, the technologists. We thought the digitisation of data comes around the middle.}

Andreas: {It's difficult to say where the high and low costs are since it depends where you get your data from. If you reuse a lot of existing data then maybe getting the data is not so expensive. If you have to write your own database then that is expensive whereas if you can adapt an existing database to your purpose that will not take much time. If you have to do original research that is expensive, editorial work is definitely expensive. Creating multi-access keys is not that easy because you need highly structured data and with dichotomous keys you still have to think a lot about how to do it. Gathering data can be easy if you have a good library and the internet, but gathering specimens requires visits to herbaria (for our example to Africa) and is cost intensive}

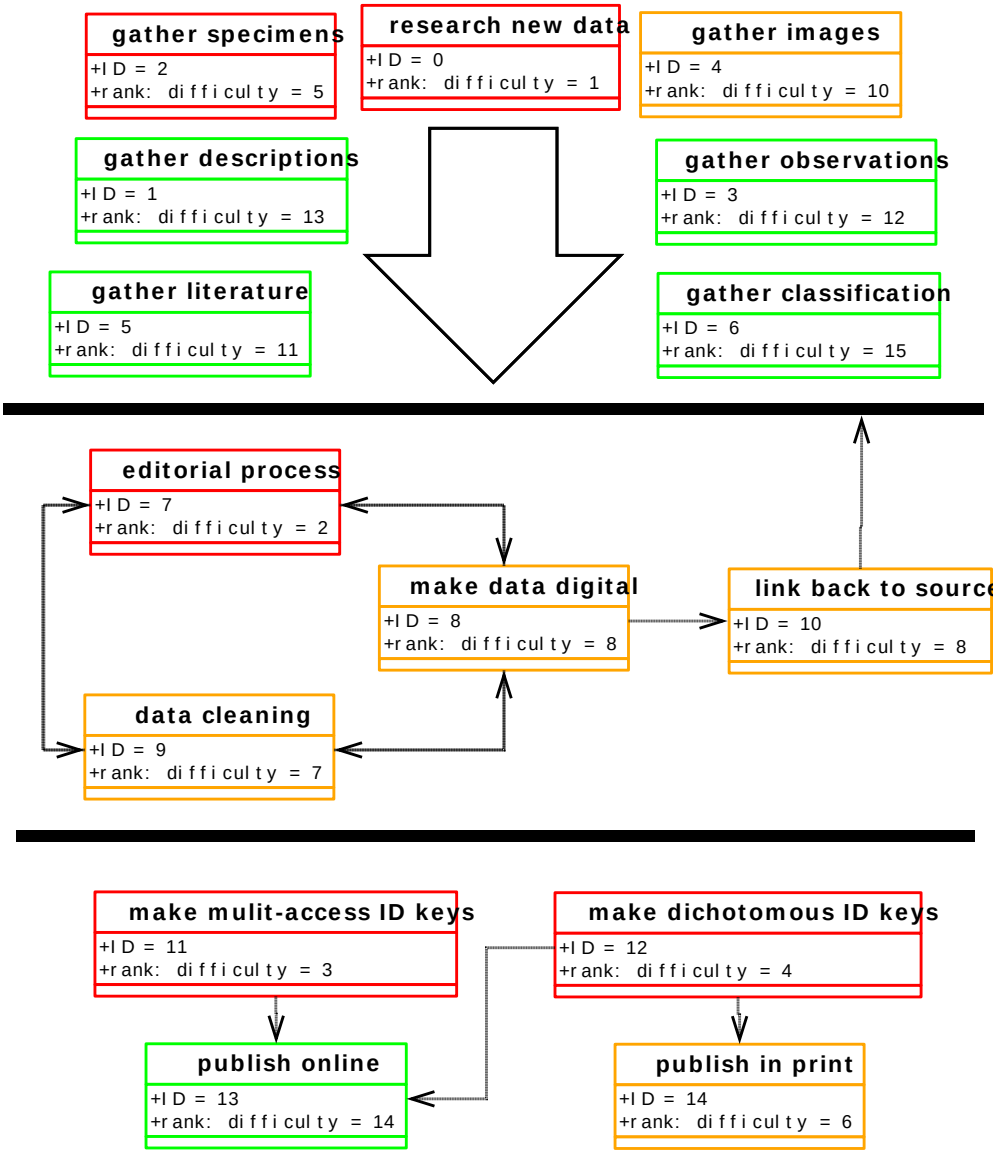

Figure 12. Producing a Digital Flora.

Andreas: {Looking at the kind of data we need and the sources; the main sources are definitely literature and online databases, maybe also offline databases. We also identified research that you may do yourself in a lab, where the results are images, descriptions of taxa, or classification information such as phylogeny, trees and so on. Herbaria provide physical information on what the specimens look like; you can do fieldwork which provides the same. Other sources that we didn't think of initially include social media which might be a good source of images and literature, and point to new information. Conferences might be interesting sources as they are good for networking and exchange of information.}

Andreas: {Looking at the standards that are in use, or could be in use; for images we have the Audubon Core standard for media data, for specimen observation data we have ABCD, Darwin and Dublin core; for descriptive and maybe key information there is SDD and DELTA. From our experience there are a lot of standards out there for literature data and we commonly use them for importing literature into our databases. A lot of people use Endnote which is not really a standard; others commonly used include MARC 21, MODS and such. For classifications there is the official standard TCS but as far as I know almost no one is using it, so we would use Darwin Core Archive, but usually classification data doesn't come with any standard at all. For phylogeny we have Nexus. Of course we also have many general standards such as XML, HTML and so on. General standards for geographic data such as the TDWG areas and ISO list of countries and language standards ISO-639 which is very helpful for common names and so on. These are the most important for us. There are standards missing in the general part that we didn't mention but these are not so important for us.}

Quentin; {We haven't mentioned standards for how you write a taxonomic name down. One of the problems we have with putting all this data together is because species names are actually written in many different ways and when trying to link databases on names, you encounter different versions of the same name, particularly with hybrids or cultivars. If there was a standard for such a thing then that would be very convenient.}

Sonia Dias: {In our community we have passport data which tells you exactly how you should describe such things.}

David Patterson: {I think the thing with that is that you have one standard, someone else has another standard and someone else another. The Global Names Architecture is building a reconciliation tool that will take all those variants, map them against each other, then allow you to select whichever format is suitable for you.}

Use-case 11: Re-Publishing Biotas

Lyubo Penev, Joe Cora, William Ulate (facilitators Donat Agosti, Eva Kralt).

Lyubo: {We've talked a lot about markup and we all know how costly and time consuming it is. If you were to do the markup of a historical Floras, why not go one step further and publish these volumes as second digital semantically enhanced editions, which could be linked to external data sources? They can be linked to a new edition of the same flora for that region, or even to a database like e-Floras CDM, so people can create new editions of that published flora. This is the basic idea behind this use-case. If we have the OCR and markup why not republish this volume or series of volumes as open access, online, digitally improved, semantically enhanced flora, fauna or mycota so that it cannot only be used in databases, but it can be reused, downloaded by anyone in the world for free. Doubling the effort and gaining one more significant benefit from the marked-up text.}

Jordan: {The next chart looks at time versus money. The most time consuming tasks are the top, the least on the bottom. The costliest are on the left and the cheapest are on the right. The most expensive tasks is to OCR the text, [2] extracting the images [3]and the tables [4] from the text (hopefully these are a small percentage). Also time consuming are the markup [6] and the scanning [0]. Copyright issues [1] can be expensive involving a lot of research and lawyers. Moving down, next we have to get IDs [8] for the markup from external sources so that we can link [7] to those external sources. Assembling the manuscript [5] is one important task because when you scanned in the text, cropped images and retyped the tables you have to reassemble in the same order to preserve the actual intention of the author, then when you enrich semantically [9], you are able to export it atomised to several databases [12]. Then you actually publish the whole thing [10,11]. }

William: {First we defined the data source types; nomenclatural lists, publication lists, external images, sequences, geographical names, people names, collection lists, multimedia and taxonomic treatments. Those are things that we could access to do the semantic enhancement to the original scanned documents. Some of the sources that we defined for scanned documents are BHL, paper (the original thing) or digital born and published online. For the list of taxa there are plenty of nomenclators. For publication lists (?). For sequences; Genbank, BoL}

William: {For the standards we've been using; TaxonX TaxonPUB, ABCD, GCS, GFF for genetic information, Audubon Core, Dublin Core. Then for literature; BibTex, MARC, MODS. For unique identifiers DOI, LSIDs. Then the standards that we all know; XML, jpeg2. Then the specialised one like OGC and wc3 standards which are already being used along with the country names and so on. We find that there are a lot of things that are built on top of the others; Dublin Core for example is used in several of the ones that we use in different environments, but it's there at the base of the definition so hopefully will allow for interconnection later when we go through the process of enriching the content. For standards that are missing; there is no standard ID for people names, collection lists, organisations. For multimedia presentations you could say that Audubon Core is it, but you could probably conceive of things that it doesn't cover. We don't have a standard way to get funding}

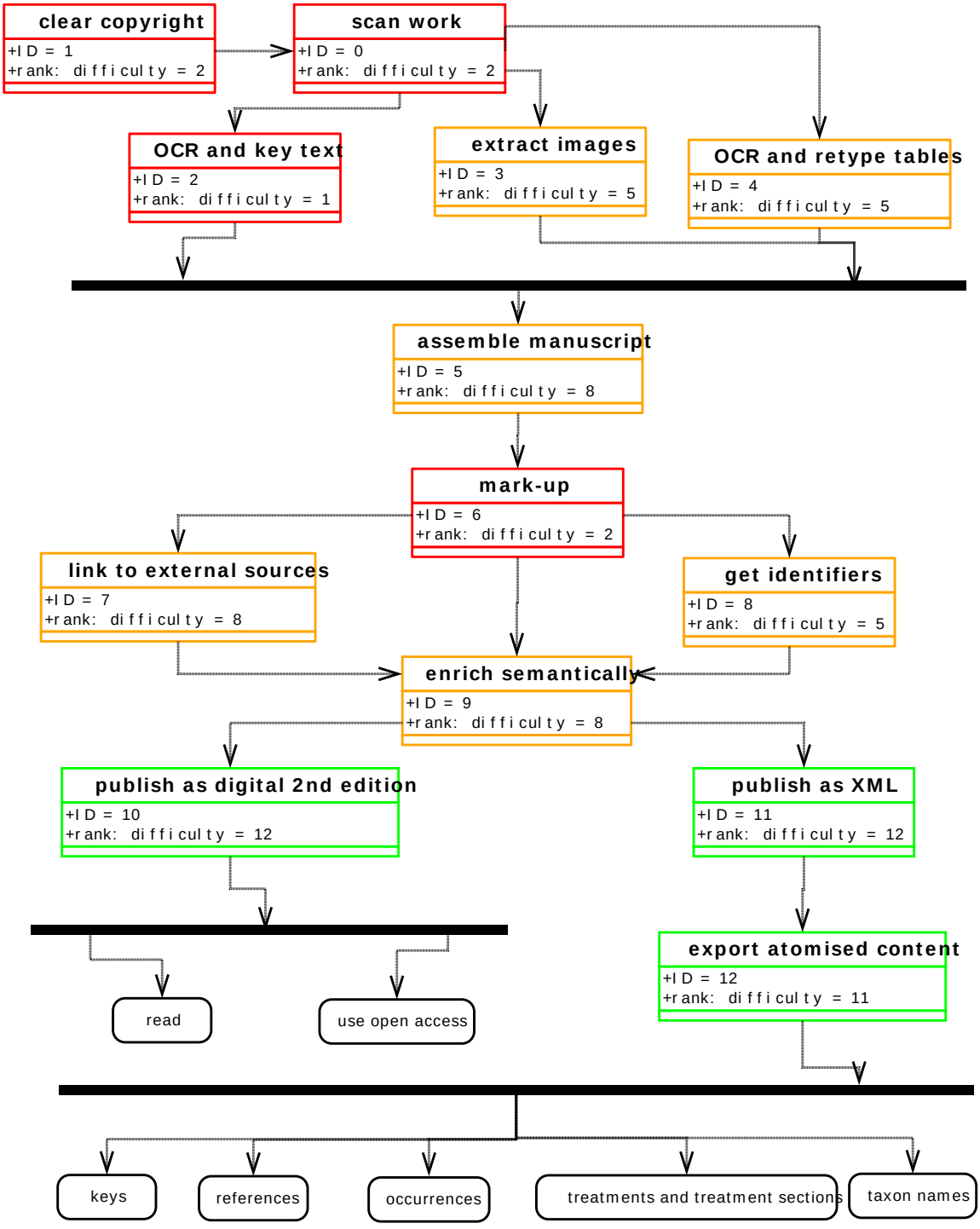

Figure 13. Re-Publishing Biotas.

Use-case 12: Producing a Field Identification Tool

Sonia Dias, Gregor Hagendorn, Richard Old, David Patterson, Holly Vincent (facilitators Donat Agosti, Eva Kralt).

Richard: {We are talking about a field ID tool with the assumptions that this would be a high tech, modern, computerised ID tool, not another book with a dichotomous key in it. Initially we wanted to be able to identify any unknown object but later restricted to just plants because we needed more specific information. We need a key which will handle parts, not just the whole organism, this is something I'm painfully familiar with, since I get specimens in the mail and you often don't get the whole specimen. Even in the field you need a key where you can deal with just portions of a specimen. We are also looking at general characteristics which do not require an absolute expert: things like size, shape, colour, habitat, those sorts of things that you can do without a high level of expertise. In any successful system you need to be able to identify your characteristics based on illustrations. The user needs to be able to say "My characteristic looks like THAT." They don't need to know the terminology of what that structure is, what it does, any of that, they just need to be able to say "Mine looks like this one over here." So illustrations are especially important in here. We didn't discuss this in the group, but one of the most powerful features of a key is the ability of a user of a key to say "Or" – i.e. "Mine looks like this one, or that one." You don't force the user to make up their mind. The other important characteristic is to provide guidance when they don't have it. Dichotomous keys have probably scared more people out of the sciences than other any single tool but they do provide guidance. So with a random access computerised key you need to have some way of providing guidance, which generally is just allowing the user to say "What do I do next?"

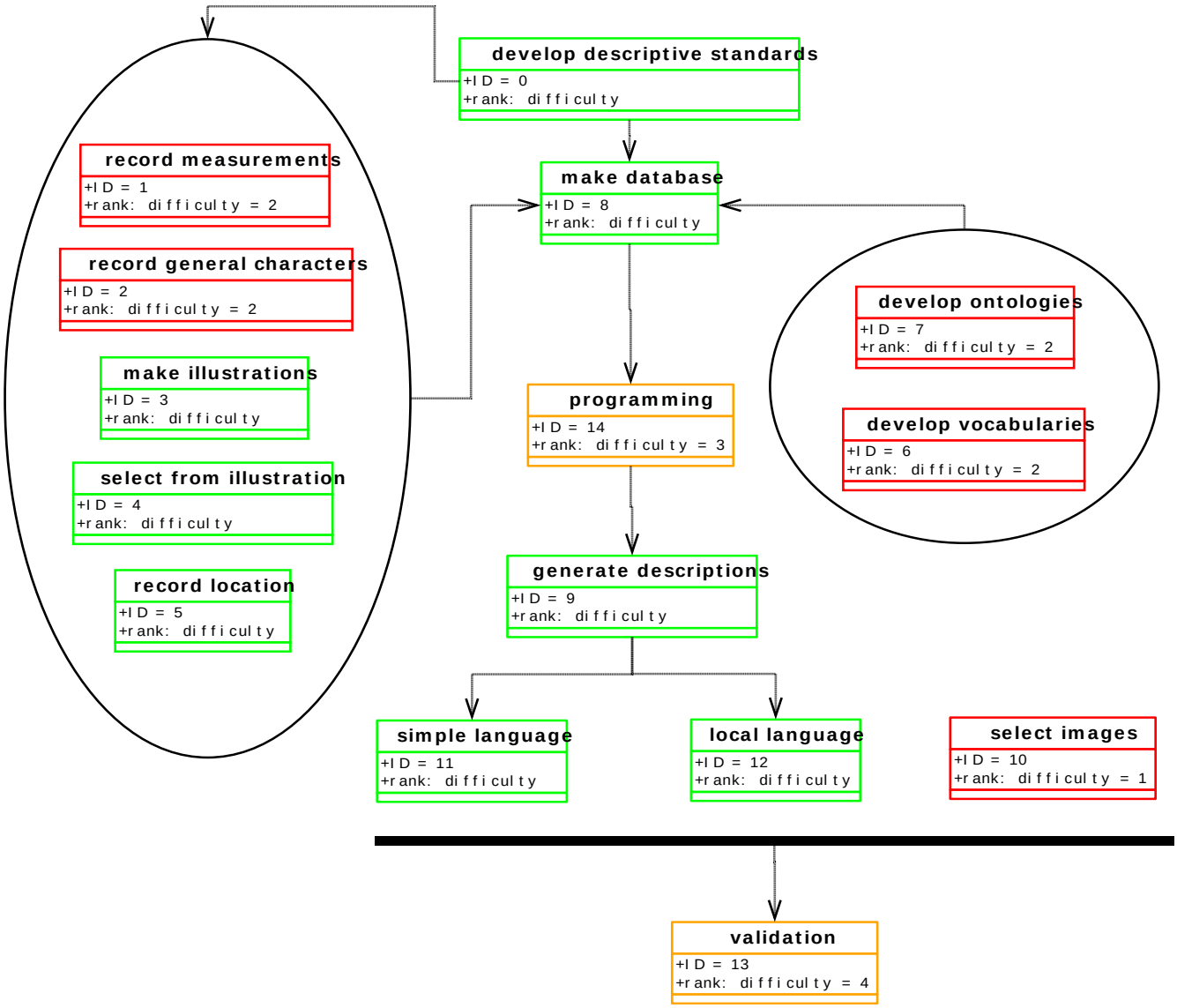

Figure 14. Producing a field identification tool.

Gregor: {One of our aims is to bring the tools to the non-expert, to the lay audiences be they farmers, high school students or otherwise. At the same time we need to relate to the experts because they can tell correct from incorrect. With a database like this there will be lots of bugs in the data you have collected. There are lots of bugs in the published flora. You can try to correct these bugs but it is a process that goes through centuries, so we can't just say that we have the data and the algorithms and just give it to the people. We need a curatorial process to continuously evaluate and correct the data, based on professional feedback and from the users of the key. That's a non-trivial thing because you need to speak two languages at once; that of the experts and that of the lay people. Experts distinguish between phylloclade and leaf whereas for normal people everything is leaf. So an important point of

such is system is an ontology system which can reason from an exact database to generalise to such a level that it can be used by lay people. Building the curation feedback platform will be an important task.}

Sonia: {We have prioritised five actions; 1 images, 2 data accumulation, 3 program, 4 verification of information and 5 infrastructure needs. At all these levels there is a need for curation, verification and validation. In terms of costs of these priorities we just start at the images because we had this discussion in terms of having high quality images for what Richard just explained, for a non-expert who need to look at images of things that are similar to what he's looking at the site and that is very costly. Then there is how things could be brought together in the way that Gregor has explained. We have what we call the community of parties, basically people with a common interest who could contribute to the different phases of gathering the information, curating and validating the data. On the bottom part you just have the different data categories as input which will be the aggregator of the different data types into the common "plant form" and that will provide the keys for the users. The middle layer became the "data in" and the "data out." So it's all these different phases, from the user to the plant form, from the plant form to the aggregator, and to the originator of the data, there needs to be a good annotation system, curation and good quality validation. All this is to make sure that we get it right; the risk is that if bad things go in then bad things come out. We mention IPR issues but didn't discuss them at all, but I think the objective is to have open access to this data as much as possible, so it can be used and disseminated widely, not only within the framework of a project but scaled out and up for continuous use. There is a <?> of wiki data within the cloud which could be used, but is more technology than needs driven, and within this framework needs thought. What drives all of this is are the four key aspects of sharing collaboration, facilitation, and a platform for input and output of information}

David Patterson: {I'd like to add a minority view; I don't agree with the view of "bad data in, bad products out" since that is a mindset that is going to delay progress. I think in here is incorporated this feedback loop which is the annotation systems that we have been hearing about - Annosys being built here in Berlin and Filtered Push being built in the US. That would allow free release of all content irrespective of quality. The end user or any player in the game can annotate any item to say that there is something wrong with this, it needs attention. That allows you to release data of any standard quickly and get the system moving quickly}
Holly: {Next we talked about the data sources that were needed for a field ID system and we came up with six data types; morphological data, geographical data, ecological data, taxonomic data, images and sociological, conservation and use. Floras came up as sources under each one of those heading this and stress the importance of Floras and what they should be doing. There are a lot of online sources as well; IPNI, focusing on plant lists, there are all sorts of different naming systems. Local expertise is highlighted as well, obviously that is very important. In terms of the field that I work in, Red Listing is important since if you are collecting in the field you do not want to collect species if they are seriously threatened, so you need that information to hand. We also discussed images and the issues surrounding those including copyright and cost.}

David: {There are certain areas of standards where we are almost ready, which is georeferencing; things that are emerging which are names and taxonomy - Global Names fits in there. It doesn't mean that because they are ready they are perfect yet, there is work to be done. Personally I think that we need to be charging particular organisations to act as the coordinating sites for all of this kind of stuff: GBIF declares itself as being in the space of georeferencing therefore they have a significant role to play. An

important point was made earlier about TCS - a lot of time and effort was put into developing the TCS system by TDGW - turned out nobody wanted to use it. So it was that it was essentially set to one side and a simpler system brought out. Now we are beginning to recognise that the simpler system does not have the capacity to resolve edge cases, there's ambiguity which gets fed into this because it's not rigid, so this is a problem for us. When we think about standards we've got to try and work out how we are going to marry things that are easy to implement and bring into use versus the things that will carry the level of discrimination that we ultimately want. Then the stuff that needs investment is going to be all the phenotypic, morphological stuff, for which there is an array of standards but for which we still find that there are different standards being used for different taxonomic areas i.e. different taxonomic areas have their own system so we need to find a way to universalise that. My own personal view is that it should be built around a phylogenetic framework.}

Use-case 13 Ecological niche modelling based on specimens and observation from Floras

Anton Guensch, Karol Marhold, Deborah Paul (facilitators Donat Agosti, Eva Kralt).

Anton: {Our use-case is to exploit occurrences, specimen data which is base data for databased Floras and e-Floras so that they can be used for scientific studies such as ecological niche modelling. This is a real world example that we are currently working on in BGBM. There are two branches in this workflow, the first one is the situation where you only have the paper-based Flora [0] available, or maybe not even that available. The second one is where we have an e-Flora [3], but making the point that we have a very low specification for an e-Flora, in our sense it's a database in electronic form; it doesn't have to be on the web, it could be a spreadsheet. }

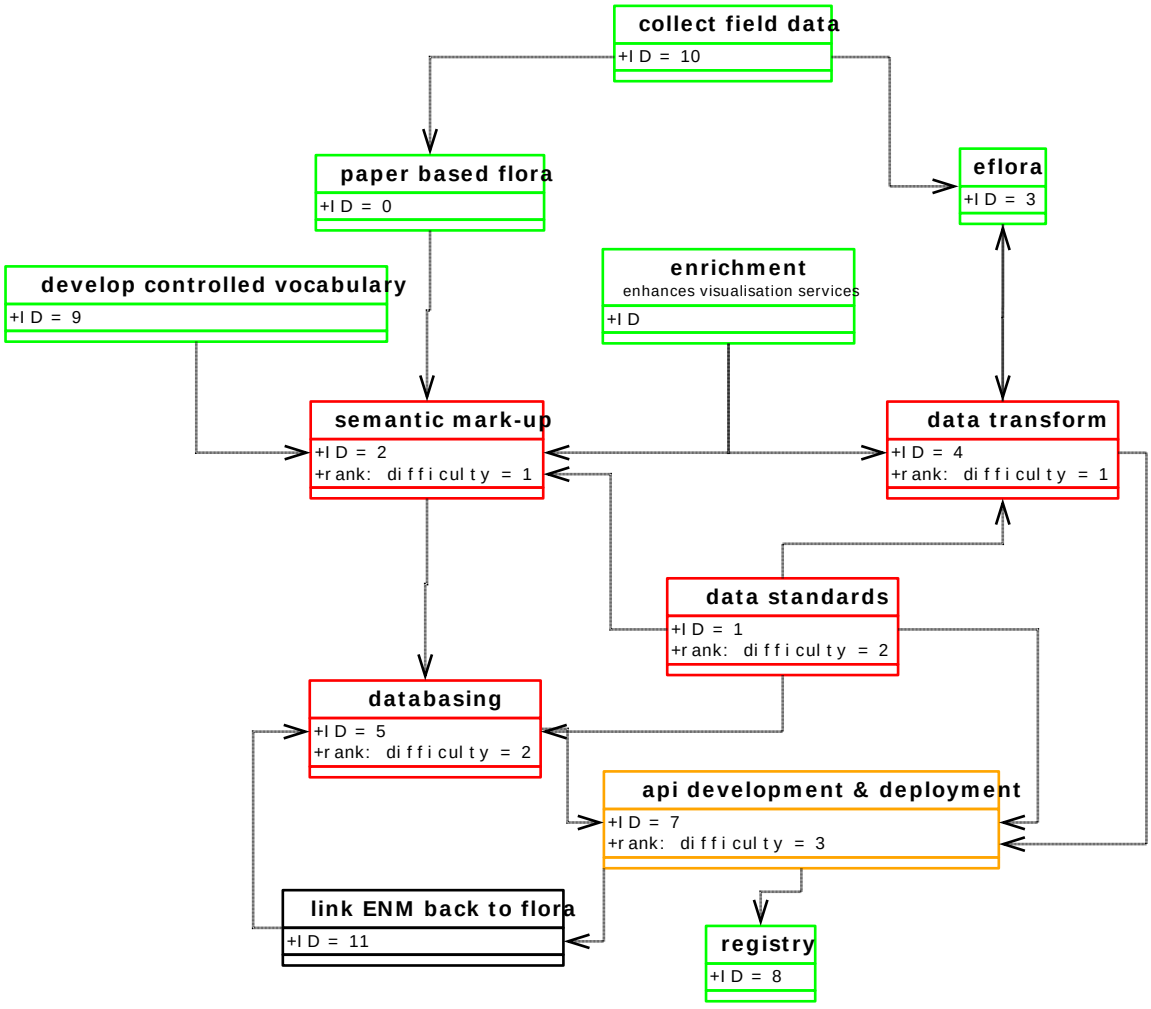

Figure 15. Ecological niche modelling based on specimens and observation from Floras.

Anton: {In the e-Flora information flow we identified two steps. The first one being data transformation [4] to bring the data into some kind of standard-based representation [1], together with external services to enrich [6] the data so that we can map using latitude and longitude, ISO country, things like that. The second step would be to deploy some kind of API [7] for this data service which makes the data available in a standardised way, for example aligned to the existing protocols we have - BioCase, Digar, Tapir. Once you have this, then the data can be immediately applied to niche modelling workflows. The second branch is a little bit more complicated because we have the data on paper, we have to do the OCR first, or in some, or in many cases we already have representation in BHL. Second we have to do some kind of semantic markup [2] using one of the tools available, and we see some need for vocabulary controls [9] for those markup processes, which help us to make the data comparable so to speak. The outcome of this markup process is the data in some format such as TaxonX or TaxML, then we need to import to a database platform [5], at which point the two branches come together with the exposure of the data through the API [7].}

Anton: {Data standards are involved throughout the process. We have indicated in blue the ones that we are using already, like ABCD, Darwin Core, Darwin Core Archive. For the registration of such new data sources we can use the Biodiversity Catalogue which has recently been released. TaxonX, TaxML and ISO standards - that's what we are already using.}

Deborah: {We'd like to link the products of the ecological niche modelling back to a flora for example or to the specimen data it came from, and also the collection of the data in the field in the first place, as there may be some things that we want to change about what's in a flora, based on what's possible in the field now, and what's traditionally been done in the past, for example instead of just an area having physical data, [we have data] for each individual specimen.}

Anton: {Considering the costs, no surprise that the semantic markup [2] needs a substantial amount of human interaction and is therefore the most expensive thing in the whole process. From experiences with the Biovel project, data transformation [4] tasks are very expensive - bringing the data up to a level where it can be used in niche modelling workflows. Databasing [5] i.e. imports from markup is not so expensive, also data standardisation [1] effort as well as OCR is not so expensive. The cheapest is the API development or deployment [7], as we have the APIs already, it's the usual provider software packages, we just have to install and configure them and it's done.}

Sabrina Eckert: {The OCR can be time consuming as well depending on how good the image is, but still less expensive compared to the markup}

Deborah: {Several people have mentioned the notion of putting the data out there so people can help us get this done - for example getting the tools for sematic markup so that people can get involved}

Anton: {The sources were the most difficult part. We have 4 categories for sources for the information flow; the literature, the controlled vocabularies, the services for semantic enrichment and the e-Floras information. For the literature we have BHL and paper libraries, and very often text files in word processing format which people give us and say please markup (i.e. personal communication). Controlled vocabularies we've mentioned several times; TDWG standards, distribution codes, GBIF name bank for

example; ISO standards, country codes and so forth. For the enrichment services, there are plenty of services which can be used, we've named just a few of them; a service published by CRIA which does the mapping from coordinates to country fields; the IPNI name services and Catalogue of Life services; Google (now called open) refine is also one of the services we deploy. For e-Floras, there is no global registry for e-Floras - it would be nice to have one but as long as we don't, it's personal communication usually. You find only a small percentage of what is available as flora information in the pockets of scientists so to speak on the web, so you really need to talk to people to get the data framework.}

Annex 4. List of use-case activities and their relative difficulties

USE-CASE	ID	DIFFICULTY	CAT	ACTIVITY
1	0	7	L	set priorities - species/regions threatened
1	1	5	M	make a list of species with valid names
1	2	2	H	compile information
1	3	2	H	finding sources (published and unpublished info)
1	4	1	H	extract specific information for red listing
1	5	3	H	verification by experts
1	6	8	L	training of group or experts
1	7	4	M	assessment by the group - group workshop
1	8	4	M	assessment by experts
1	9	3	H	review of the assessment by the group
1	10	6	L	publication
2	22	6	L	practicality/taxonomy
2	0	2	H	digitisation of collections
2	23	6	L	diversity
2	1	4	M	digitisation of literature
2	24	6	L	importance for Economic & Ecological Services
2	2	3	M	validation
2	25	6	L	flagship taxa
2	3	1	H	get funding
2	4	6	L	select participants (field and desk)
2	5	4	M	preparation: prepopulate and georeference
2	6	5	L	gaps in taxonomic data
2	7	5	L	gaps in distribution data
2	8	2	H	population data (often missing)
2	9	5	H	ecological vulnerability
2	10	5	H	threats info
2	11	5	L	data deficient taxa
2	12	4	M	capacity building
2	13	3	M	RLA recognises as publication
2	16	3	M	researching DD taxa
2	17	2	H	monitoring
3	0	-	L	trait selection
3	1	-	L	what is a trait

USE-CASE	ID	DIFFICULTY	CAT	ACTIVITY
3	2	4	M	finding data sources
3	4	5	L	developing database structure
3	5	1	H	data extraction [+entering]
3	6	3	M	data curation
3	7	4	M	quality assessment
3	8	1	H	data release
3	9	1	H	scientific publication
3	10	5	L	web design
3	11	6	L	long term sustainable funding
3	12	2	H	getting user feedback
3	17	-		prioritising trait selection
4	0	1	H	establishing transect network
4	1	4	L	species identity and mapping
4	2	3	M	collect global data of same interest (e.g. TRY)
4	3	2	M	upscale modelling to vegetation level
5	0	3	M	literature review
5	1	2	M	acquire specimens
5	2	2	M	capture label data
5	3	2	M	georeference locality data
5	4	4	M	summarise distribution for taxon concepts
5	5	1	H	develop character set
5	6	1	H	revise characters
5	7	4	M	summarise characters for taxon concepts
5	8	1	H	develop taxonomic concepts
5	9	3	M	capture images (characters, taxon concepts)
5	10	5	M	locate primary types
5	11	7	L	compare types with taxon concepts
5	12	1	H	revise taxonomic concepts
5	13	2	H	return specimens
5	14	6	L	data quality control
5	15	8	L	submit for publication
5	16	4	M	coalesce all data
6a	0	5	L	collect data
6a	1	4	L	initial field id
6a	2	3	M	process material

USE-CASE	ID	DIFFICULTY	CAT	ACTIVITY
6a	3	2	M	decide on field site
6a	4	1	H	use local field guides (inc. e-guides)
6a	5	1	H	match in herbarium (inc. online)
6a	6	1	H	match online images and hardcopy
6a	7	1	H	send to specialist (image or plant)
6b	1	1	H	herbarium guides or flora data
6b	2	1	H	seek expert advice
6b	3	1	H	match herbarium specimens
6b	4	2	M	prepare voucher material
6b	5	2	M	send to herbarium for ID
6b	6	3	L	collect voucher
6b	7	3	L	pre-identification
6b	8	4	L	deposit specimen
7	0	1	H	gather existing specimens of a taxon
7	1	1	H	rename specimens
7	2	1	H	data capture
7	3	1	H	georeference locality data
7	4	2	H	new occurrence record gathering
7	5	3	H	gather literature occurrence records
7	6	4	M	gather literature treatments
7	7	4	M	name taxa
7	8	4	M	describe taxa
7	9	4	M	provide identification tools
7	10	5	L	prioritise diagnostic/characteristic characters
7	11	5	L	define taxa
7	12	5	L	gather nomenclature
7	13	6	L	define taxonomic and geographic scope
7	14	6	L	judge quality/applicability of data for reuse
7	15	6	L	publish account
8	0	1	H	ensure consistency of taxonomic treatment: collections numbers, locality-type, synonymy
8	1	1	H	structure the text
8	2	1	H	markup to export
8	3	2	H	illustrations editing
8	9	1	H	bibliography ref. check consistency and cross-link

USE-CASE	ID	DIFFICULTY	CAT	ACTIVITY
8	10	2	H	copy edit according to journal format / standard in use in the field
8	11	3	M	manage the peer review process
8	12	4	M	lay out
8	13	5	M	proof read
8	14	5	M	publish (print or online)
8	15	5	M	link the cross-reference
8	16	6	L	ensure technical quality & consistency, check illustrations quality
8	18	7	L	rewrite abstract and ensure it is not too long (risk of being cut in databases)
8	19	7	L	define key words
8	20	8	L	identify targets for dissemination of the journal
8	21		L	export standards for official databases
8	0	1	H	ensure consistency of taxonomic treatment: collections numbers, locality-type, synonymy
8	1	1	H	structure the text
8	2	1	H	markup to export
8	3	2	H	illustrations editing
8	9	1	H	bibliography ref. check consistency and cross-link
8	10	2	H	copy edit according to journal format / standard in use in the field
9	0	-	H	synthesise data
9	1	-	H	analyse
9	2	-	L	take samples of unknowns
9	3	-	L	identify samples
9	4	1	H	identify a plant
9	5	-	H	study prior Floras and checklists
9	6	-	H	collect field information
9	7	-	L	collect new survey data
9	8		M	gather information
9	9		M	validate existing information
9	10		M	collate existing information on plants
9	11	1	H	identify threatened species
9	12	1	H	make species list of area

USE-CASE	ID	DIFFICULTY	CAT	ACTIVITY
9	13	1	H	get species status
9	14		M	use identify guide
9	15		M	compare to needs
9	16		M	quality validation
9	17		M	sort and sift
9	18	1	H	measure abundance
9	19		L	define area
9	20		L	get map of area
9	21		L	get soil data
9	22		L	draw graphs of survey data
9	23		L	draw maps of survey
9	24		L	get management information needs
9	25		L	write report
9	26		L	explain report
10	0	1	H	research new data
10	1	13	L	gather descriptions
10	2	5	H	gather specimens
10	3	12	L	gather observations
10	4	10	L	gather images
10	5	11	L	gather literature
10	6	15	L	gather classification
10	7	2	H	editorial process
10	8	8	M	make data digital
10	9	7	M	data cleaning
10	10	8	M	link back to source
10	11	3	H	multi-access identification keys
10	12	4	H	dichotomous identification keys
10	13	14	L	publish internet
10	14	6	M	publish paper
11	0	2	H	scan
11	1	2	H	clear copyright
11	2	1	H	OCR and key text
11	3	3	M	extract images
11	4	3	M	OCR and retype tables
11	5	4	M	assemble manuscript

USE-CASE	ID	DIFFICULTY	CAT	ACTIVITY
11	6	2	H	markup
11	7	4	M	link to external sources
11	8	3	M	get identifiers
11	9	4	M	enrich semantically
11	10	6	L	publish as 2nd digital edition
11	11	6	L	publish as XML
11	12	5	L	export atomised content
12	0		L	develop descriptive standards
12	1	2	H	record measurements
12	2	2	H	record general characters
12	3		L	make illustrations
12	4		L	select from illustration
12	5		L	record location
12	6	2	H	develop vocabularies
12	7	2	H	develop ontologies
12	8		L	make database
12	9		L	generate descriptions
12	10	1	H	select images
12	11		L	simple language
12	12		L	local language
12	13	4	M	validation
12	14	3	M	programming
13	0		M	paper flora
13	1	2	H	data standards
13	2	1	H	semantic markup
13	3		M	e-Flora
13	4	1	H	data transform
13	5	2	H	Databasing
13	6	-	L	enrichment - enhanced viz services
13	7	3	L	API development or deployment
13	9	?	L	controlled vocabulary development
13	10	?	L	field data collection
13	11	?	L	linking ENM back to flora?

Annex 5. List of use-case information types, relative importance (and difficulty)

ns= not scoped (due to lack of time in the workshop to complete the exercise)

USE-CASE	IMPORTANCE	INFORMATION	DIFFICULTY
1	H	georeferenced localities	L
1	H	georeferenced localities	H
1	H	EOO &AOO	L
1	L	bioclimatic zone	H
1	L	geology	H
1	L	ecosystem	H
1	L	vegetation	H
1	H	habitat	L
1	L	altitude	H
1	H	protected area	L
1	H	taxonomy	L
1	L	vernacular names	L
1	L	description features	L
1	L	ecosystem services	H
1	L	uses	H
1	L	trade/CITES	H
1	L	climate change vulnerability	H
1	L	conservation measures	H
1	L	age of maturation	H
1	L	abundance	H
1	L	longevity and trends	H
1	L	reproduction	H
1	L	dispersal	H
1	L	pollination	H
1	L	migration	H
1	L	phenology	H
2	H	taxonomy	ns
2	H	distribution	ns
2	H	population size	ns
2	M	habitat	ns
2	M	threat status	ns
2	M	threats	ns
2	L	use and ecological services	ns
2	L	research & conservation action	ns
3	H	taxon	ns

USE-CASE	IMPORTANCE	INFORMATION	DIFFICULTY
3	H	trait	ns
3	H	source	ns
3	M	exposition	ns
3	M	measurement details	ns
3	M	lat/long	ns
3	M	region	ns
3	M	country	ns
3	L	vegetation classification	ns
3	L	biome	ns
3	L	soil	ns
4	H	major species	ns
4	H	species distribution	ns
4	H	gas exchange	ns
4	H	leaf traits	ns
4	H	hydraulic traits	ns
4	H	leaf nutrients and isotopes	ns
4	M	climate	ns
4	M	elevation	ns
4	M	soil	ns
4	L	genomics	ns
4	L	morphology	ns
5	ns	existing taxonomies	H
5	ns	existing taxonomic names	H
5	ns	geo loc	H
5	ns	location of specimens	H
5	ns	location of specimens	L
5	ns	acquisition of specimens	L
5	ns	anatomical terms	H
5	ns	anatomical terms	L
5	ns	literature	H
5	ns	literature	L
5	ns	images	L
6	L	list of experts	L
6	L	list of herbaria	L
6	L	collecting permits	L
6	M	locality information	H
6	M	occurrence data	H
6	M	vegetation maps	L

USE-CASE	IMPORTANCE	INFORMATION	DIFFICULTY
6	M	climate data	L
6	M	physical map	L
6	H	description	H
6	H	specimen	H
6	H	images	H
7	H	original publication references	L
7	H	original publication of names	L
7	H	bibliography standards	H
7	H	distribution - geo-range	H
7	H	distribution - geo-range	H
7	H	GIS shape files	H
7	H	geo-locality	H
7	H	collector information	H
7	H	herbarium specimens	H
7	H	digital illustrations and photographs	H
7	H	digital illustrations and photographs	H
7	H	descriptions and keys	H
7	H	flowering and fruiting time	H
7	M	habitat	H
7	M	habitat	H
7	M	altitudinal range	H
7	M	altitudinal range	H
7	L	plant uses	H
7	L	plant uses	H
8	H	key word check	L
8	H	references	L
8	H	description	L
8	H	collection numbers	L
8	H	locality	L
8	L	IMRAD	H
8	L	drawings	H
8	L	SIMILIS	H
8	L	identification key	H
9	ns	correct name	M
9	ns	plant ID visually	M
9	ns	plant ID visually	L
9	ns	endemic/non-endemic	M
9	ns	invasive/non-invasive	M

USE-CASE	IMPORTANCE	INFORMATION	DIFFICULTY
9	ns	presence/absence	M
9	ns	presence/absence	L
9	ns	threat status	L
9	ns	habitat plant needs	M
9	ns	habitat we have	M
10	ns	identification keys	ns
10	ns	descriptions	ns
10	ns	specimens	ns
10	ns	observations	ns
10	ns	images	ns
10	ns	literature citations	ns
10	ns	classification	ns
11	ns	scanned documents	ns
11	ns	list of taxa	ns
11	ns	nomenclatural list	ns
11	ns	publication list	ns
11	ns	external images	ns
11	ns	sequences	ns
11	ns	geographical names	ns
11	ns	people names	ns
11	ns	collections list	ns
11	ns	multimedia	ns
11	ns	taxon treatment	ns
12	ns	morphological data	ns
12	ns	geographical data	ns
12	ns	ecological data	ns
12	ns	taxonomic data	ns
12	ns	images	ns
12	ns	sociological data, uses and conservation	ns
13	ns	literature	ns
13	ns	controlled vocabulary	ns
13	ns	enrichment services	ns
13	ns	e-Floras	ns

Annex 6. List of data standards mentioned in the use-cases

ns= not scoped (due to lack of time in the workshop to complete the exercise)

USE-CASE	INFORMATION TYPE	STANDARD
11	semantic markup	TaxonX
11	semantic markup	TaxPub
11	ns	DwC
11	ns	ABCD
11	ns	TCS
11	ns	GFF
11	ns	Audubon Core
11	ns	Dublin Core
11	ns	BibTex
11	ns	MARC
11	ns	MODS
11	ns	DOI
11	ns	LSID
11	ns	XML
11	ns	JP2
11	ns	OGC
11	ns	ISO
11	ns	W3C
10	images	tiff
10	images	jpg
10	images	exif
10	specimen observations	ABCD
10	specimen observations	Darwin Core
10	Keys & descriptions	DELTA
10	Keys & descriptions	SDD
10	literature	BibTex
10	literature	endnote
10	classification	DWC-A
10	classification	TCS
10	classification	Nexus
10	general	UTF8
10	general	XML
10	geography	TDWG geography

USE-CASE	INFORMATION TYPE	STANDARD
10	geography	ISO-639 (lang)
13	semantic markup	TaxonX
13	semantic markup	TaxML
13	semantic markup	agreed format
13	databasing	agreed format
13	data transform	ISO
13	API development	DwC
13	API development	DwC-A
13	API development	ABCD
13	Registry	Biodiversity Catalogue
12	geography	georef coordinates
12	geography	polygons
12	geography	point range
12	geography	GBIF
12	geography	TDWG geography
12	names & taxonomy	multiple systems
12	names & taxonomy	vernacular
12	names & taxonomy	GNA
12	names & taxonomy	CoL
12	names & taxonomy	TCS
12	names & taxonomy	DWC-A
12	?	Bioversity Descriptors

Annex 7. Lightning talks

Key themes and discussion points

SCOPE
access to primary literature
amelioration, prediction
distribution
distribution
ecology
ecophysiological
functional types
habitat maps
identification aids
images
incorporate different opinions
IUCN conservation ratings
leafsnap but with more information
measurements of identifying features
niche models
prediction and modelling
presence/absence data
regional to global
species maps
traits

AUTHORITY
check taxonomic content
no competing databases
static citable version
validation of data

INTEROPERABILITY
combine existing but dispersed sources
data standards
links to all information
list important and useful info and agreement on structure
match with other data
official names database
standard export

INFRASTRUCTURE
"filtered push"
cloud-based data
distributed nodes holding data
information as an unrooted network
LSID

CHANNELS
app. on phone
database
derivatives
electronic database main product - many delivery channels
field guides
hardcopy

STAKEHOLDERS
access for users globally
accreditation explicit
attract interest groups
citizens as a resource
updates available to the community
user help
users "down to farmers"

SUSTAINABILITY
advertisement
don't plan just do
economic argument for open access
free access
keep raw data
literature online - free
no "walls" [barriers]
open access
orphaned data - what happens?
period of limited access
revenue models
secret sites vs. open access

List of speakers

Speaker	Organisation
Donat Agosti	Plazi
Laurence Bénichou	MNHN
Viola Clausnitzer	Senckenberg Museum of Natural History Görlitz
Sonia Dias	Bioversity International, Rome
Quentin Groom	NBGB
Vololoniaina Jeannoda	University of Antananarivo, Faculté des Sciences, Botany Department
Jens Kattge	Max Planck Institute for Biogeochemistry
Robert Kenward	IUCN
Bente Klitgaard	RBGK
Jeremy Miller	Naturalis
Richard Old	XID Services, Inc.
Deborah Paul	Florida State University
Lyubomir Penev	Pensoft
Johannes Penner	MfN
William Ulate	Missouri Botanical Garden
Mark Watson	RBGE
Shuangxi Zhou	Macquarie Univ., AU