D3.1: Technical issues and risks associated with general challenges of provisioning research infrastructures to deliver capabilities for EBV processing

Project acronym: *GLOBIS-B*

Project full title: “GLOBal Infrastructures for Supporting Biodiversity research”

Grant agreement no.: 654003

Due-Date:	29th April 2016
Actual Delivery:	*30th September 2016*
Lead Partner:	Cardiff University
Dissemination Level:	PU
Status:	Public version
Version:	1.0

DOCUMENT INFO

Date and version no.	Author	Comments/Changes
15th June 2016 16th June 2016, v0.1	Alex Hardisty, CU	Initial draft, structuring and outlining of content
16th June 2016, v0.2 17th – 21st June 2016, v0.3 23rd June 2016, v0.4	Alex Hardisty, CU	Continued writing while waiting for early review comment and input from others
30th June 2016, v0.5 1st July 2016, v0.6	Alex Hardisty, CU	Improved the section on translational risks; added more content to technical issues sub-sections
12th July 2016, v0.7	Alex Hardisty, CU	Filling outstanding gaps, ready for review
14th July 2016, v0.8	Alex Hardisty, CU	Near final draft for review by project staff; (still missing sections from David).
27th July 2016, v0.9	David Manset, Gnubila Alex Hardisty, CU	Inserted section on business model canvas as outcome of workshop 2. Other minor tidy-ups.
15th August - 3rd September 2016, v0.10	Lucy Bastin, Joint Research Centre of the European Commission Lee Belbin, Atlas of Living Australia	Independent critical review and editing.
30th September 2016, v1.0	Alex Hardisty, CU	Final editing and publication.

TABLE OF CONTENTS

LIST OF TABLES AND FIGURES

1 Executive summary

This document identifies and describes technical issues and risks associated with the general challenges of provisioning research infrastructures to deliver capabilities for processing Essential Biodiversity Variables (EBV) [Pereira 2013].

This document is the result of preparatory work for workshop 1 (Leipzig, 29th February – 2nd March 2016), the workshop itself and follow-up activities after the workshop and leading up to workshop 2 (Sevilla, 14 – 15th June 2016). It is a record of information available at a moment in time, together with some analysis as a result of activity carried out in tasks 3.1, 3.2 and 3.3 of the Description of Work. The document is concerned mainly with assisting to complete the work package objectives to:

- Map the user requirements with existing infrastructure data and processing capabilities, identifying gaps where they exist; and,
- Design a methodological framework to translate candidate EBVs into transnational and cross-infrastructure scientific workflows.

The document provides insights to the problem of how cooperating biodiversity research infrastructures can contribute to frontier research into the harmonised implementation of Essential Biodiversity Variables, by focusing on offering data, workflows and computational services.

This document summarises technical issues and risks that have to be addressed to enable practical development and delivery of EBVs data products. There are several unresolved general challenges and considerations associated with harmonised EBV implementation, and it places the technical issues and risks in the context of those. These general challenges are about assigning responsibility for EBV production, understanding the EBV production cycle and its needs in terms of data model structure and applicable standards. It is likely that new infrastructure will be needed to provide the tools and workflows for EBV production. A single and shared understanding of strategy toward the technical aspects of EBV data products production is needed among all stakeholders.

There is a significant unresolved problem in terms of the nature of the translation process that must be followed in order to move from frontier research that proves the principles of EBVs to a state of regular mass production of EBV data products, akin to climate variable data products.

We recommend the development of a technical roadmap for the next 3 – 5 years, in conjunction with identifying first steps to gain practical experience of the issues associated with producing EBV data products.

2 Contributors

The main contributors to the present document are Alex Hardisty (Cardiff University, UK), WP3 leader and David Manset (Gnubila, France). Other members of the project team and the participants of workshop 1 have provided information that supports and forms the contributions of Hardisty and Manset. In particular, we acknowledge contributions and inspirations from Daniel Amariles, Lee Belbin, Matthias Obst, Hannu Saarenmaa, Brian Wee, and Kristen Williams. Critical proof-reading and editing has been carried out by Lucy Bastin (who also provided suggestions for Table 6) and Lee Belbin.

Much of the work reported in the present document is based on the evolution of ideas, the conduct of workshops and analysis of the outcomes. In particular, the case studies mentioned in section 5 are works in progress that will yield important lessons regarding the practicalities (proofs-of-principle) of the informatics capabilities and capacities needed for producing EBV data products.

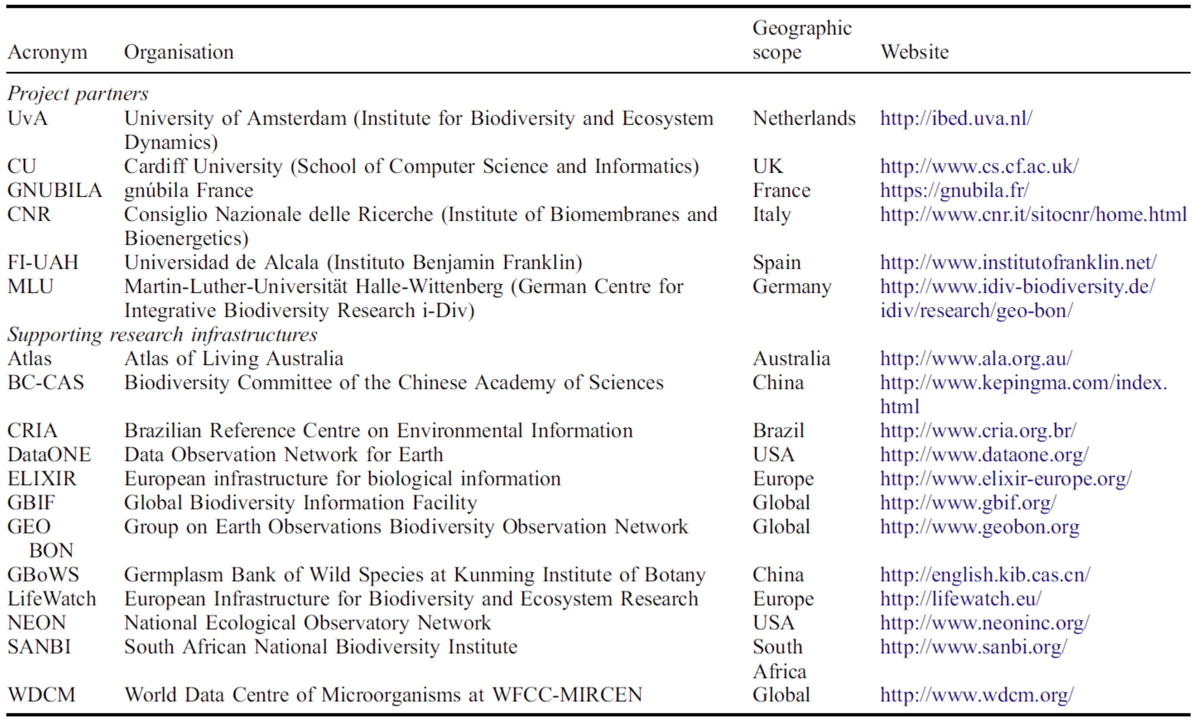
3 Background

A key question for global biodiversity monitoring is how the multi-lateral cooperation of data collectors, data providers, monitoring schemes, and biodiversity research infrastructures can be achieved at the global level to support the harmonised implementation of EBVs. As yet, there has been little applied evaluation of EBVs for their significance in constructing biodiversity indicators and their applicability at different spatiotemporal scales. Frontier research in this area requires the availability and accessibility of substantial data sets with sufficient spatiotemporal coverage. GLOBIS-B aims at elucidating how the cooperating research infrastructures (Table 1) may contribute to such an objective by focusing on offering data, workflows and computational services for supporting the calculation of EBVs…

- …for any geographic area, small or large, fine-grained or coarse;
- …at a temporal scale determined by need and/or the frequency of available observations;
- …at a point in time in the past, present day or in the future;
- …as appropriate, for any species, assemblage, ecosystem, biome, etc.
- …using data for that area / topic that may be held by any and across multiple research infrastructures;
- …using a harmonized, widely accepted protocol (workflow) capable of being executed in any research infrastructure;
- …by any (appropriate) person anywhere.

The scientific questions and methods related to the above aspects will assist in defining the user requirements for extracting and handling data to support measurement of biodiversity change. To this end, the project brings together key scientists with global research infrastructure operators, technical experts and legal interoperability experts to address research needs and infrastructure services underpinning the concept of EBVs. With this focus on research needs for calculating and testing EBVs, the short-term attention is on ad-hoc on-demand services (and related workflows) in the cooperating research infrastructures. The experiences obtained may later lead to a more systematic, periodic production cycle where EBV data products are produced, updated and extended annually, quarterly or monthly.

The listed supporting research infrastructures represent those that have agreed to contribute to the GLOBIS-B project. From Kissling *et al.* (2015)



Acronym	Organisation	Geographic scope	Website
Project partners			
UvA	University of Amsterdam (Institute for Biodiversity and Ecosystem Dynamics)	Netherlands	http://ibed.uva.nl/
CU	Cardiff University (School of Computer Science and Informatics)	UK	http://www.cs.cf.ac.uk/
GNUBILA	gnúbila France	France	https://gnubila.fr/
CNR	Consiglio Nazionale delle Ricerche (Institute of Biomembranes and Bioenergetics)	Italy	http://www.cnr.it/sitocnr/home.html
FI-UAH	Universidad de Alcala (Instituto Benjamin Franklin)	Spain	http://www.institutofranklin.net/
MLU	Martin-Luther-Universität Halle-Wittenberg (German Centre for Integrative Biodiversity Research i-Div)	Germany	http://www.idiv-biodiversity.de/idiv/research/geo-bon/
Supporting research infrastructures			
Atlas	Atlas of Living Australia	Australia	http://www.ala.org.au/
BC-CAS	Biodiversity Committee of the Chinese Academy of Sciences	China	http://www.kepingma.com/index.html
CRIA	Brazilian Reference Centre on Environmental Information	Brazil	http://www.cria.org.br/
DataONE	Data Observation Network for Earth	USA	http://www.dataone.org/
ELIXIR	European infrastructure for biological information	Europe	http://www.elixir-europe.org/
GBIF	Global Biodiversity Information Facility	Global	http://www.gbif.org/
GEO BON	Group on Earth Observations Biodiversity Observation Network	Global	http://www.geobon.org
GBoWS	Germplasm Bank of Wild Species at Kunming Institute of Botany	China	http://english.kib.cas.cn/
LifeWatch	European Infrastructure for Biodiversity and Ecosystem Research	Europe	http://lifewatch.eu/
NEON	National Ecological Observatory Network	USA	http://www.neoninc.org/
SANBI	South African National Biodiversity Institute	South Africa	http://www.sanbi.org/
WDCM	World Data Centre of Microorganisms at WFCC-MIRCEN	Global	http://www.wdcm.org/

Table 1: Project partners and supporting research infrastructures of the GLOBIS-B project.

The GLOBIS-B project is unique in bringing together biodiversity scientists and biodiversity informatics staff from research infrastructures to discuss and develop a framework for implementing EBVs. Further background is available in Kissling et al. 2015. The proposed 4 workshops are meant as *experiments,* where different scenarios are considered on how scientists may want to test the relevance of EBVs for building indicators, and which data, workflows and computational capacity each scenario will require. In turn, the cooperating research infrastructures will consider the challenges and potential solutions of providing the required data and workflow services to achieve global interoperability. This requires detailed discussion of the necessary steps and tools needed to move from data collection, integration, filtering, through modelling, testing and validation to the final presentation of an EBV (Figure 2, green).

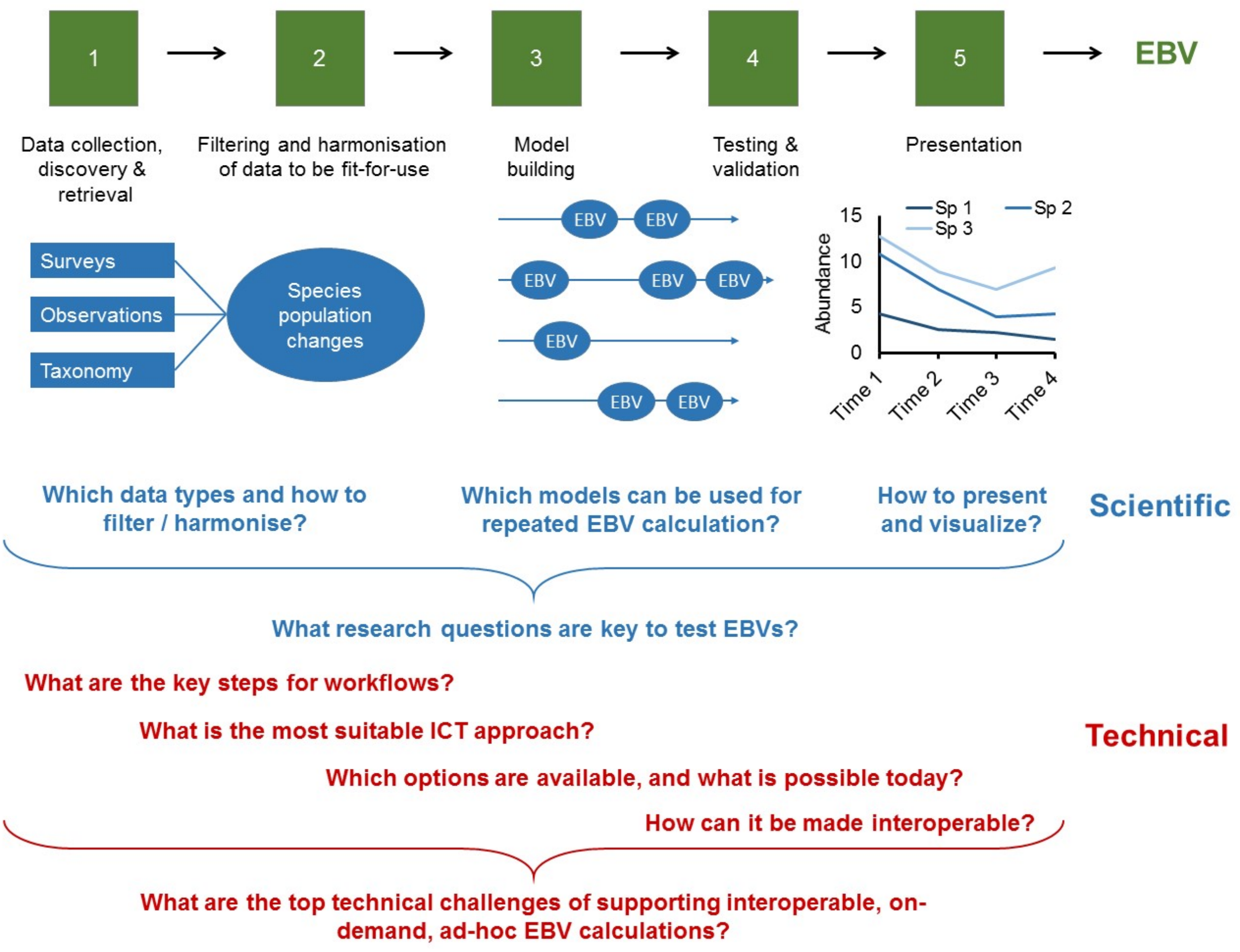

1
2
3
4
5
EBV
Data collection, discovery & retrieval
Filtering and harmonisation of data to be fit-for-use
Model building
Testing & validation
Presentation
Surveys
Observations
Taxonomy
Species population changes
EBV
Abundance
15
10
5
0
Sp 1
Sp 2
Sp 3
Time 1
Time 2
Time 3
Time 4
Which data types and how to filter / harmonise?
Which models can be used for repeated EBV calculation?
How to present and visualize?
Scientific
What research questions are key to test EBVs?
What are the key steps for workflows?
What is the most suitable ICT approach?
Which options are available, and what is possible today?
How can it be made interoperable?
Technical
What are the top technical challenges of supporting interoperable, on-demand, ad-hoc EBV calculations?

Figure 1: Potential steps for calculation of Essential Biodiversity Variables (EBVs, green); and related scientific (blue) and technical (red) questions and challenges.

Important scientific discussion points (Figure 1, blue) will be:

- Which data are needed and how they have to be integrated, filtered and harmonised;
- Which analytical tools and models need to be implemented and tested; and,
- How to present and visualize EBVs.

Related technical discussion points (Figure 1, red) are:

- How workflows have to be designed;
- Which Information and Communication Technology (ICT) approaches and options are available; and,
- How the approaches can be made interoperable.

The aim is to formulate key research questions for testing EBVs and to help identify which technical and legal challenges the research infrastructures are facing to support the interoperable, on-demand, ad-hoc calculation of EBVs.

4 Pre-workshop questions and responses

In preparation for the first workshop and having in mind the different perspectives (scientific, technical, legal/policy) of the experts, participants were asked to answer questions on scientific aspects of EBVs and the technical aspects of research infrastructures. Approximately 25 - 30 answers were received relating to the technical aspects (Annex 1) and these have been analysed thematically. The questions, with a summary of the emergent themes in the answers for each one are given in the following sub-sections.

4.1 Key steps of workflows for EBVs species distribution/abundance

Q5. What are the key steps of a workflow(s) for calculating species distribution and/or abundance EBVs, starting from accessing the raw data to presenting a visual result? What are the complexities involved? What data preparation is needed?

Many possible approaches and sequences of steps were mentioned by the respondents, varying according to their experience and backgrounds. These ranged, for example, from direct inference using remote sensing imagery through to derivative approaches based on some form of modelling. In all cases, data availability and quality is mentioned as being of particular concern. These concerns are described often in terms of understanding fitness for purpose of the available data, gaps in the data and accessibility of data. Respondents also supplied details of specific steps to be taken to prepare data for use e.g., by eliminating non-relevant data, by performing consistency checks and correcting biases, by standardising what is needed, metadata to assess precision and accuracy, etc. The tendency of both taxonomy and methods to evolve was highlighted as a significant challenge.

Many responses recognised the need to achieve a balance between having an automated approach based on common tools in a sequence of steps (a workflow) and the application of expert human input at multiple points to verify and assure the correct progress of the procedure. This is particularly important during the research phase while different approaches to producing EBV data products are being tested and assessed. Access to intermediate outputs for this purpose is essential, with documentation (i.e., traceability) being identified as the key. Several variations in the possible sequence of steps were suggested but generally these can be consolidated together in sequence to give a generalised workflow for EBV production that has approximately 12 - 15 steps. (Note: See also section 9 below.)

4.2 Suitable technical approach to perform workflow

Q6. What is a suitable technical (ICT) approach to perform this workflow(s) for calculating EBVs (any place, any time, using data anywhere, by anyone)? What special considerations have to be taken into account?

Many answers suggest using a standardised workflow approach, although there is variability in the details. Within these responses, adaptations of existing tools as well as several alternative approaches based on Species Distribution Modelling (SDM), or Online Analytical Processing (OLAP), etc. are suggested. These need to be investigated further. Understanding the consequences of demand on computational capability and capacity are important.

A common theme is wrapping of data resources and tools with standard APIs; and/or the use of standardized vocabularies for meta-description of those APIs. This is consistent with the general trend towards greater machine-to-machine (i.e., software-to-software) interactions. Work on reference architecture and profiles of recommended standards is also mentioned and needs to be pursued.

The level of practitioner skills needed to set up and operate any pipeline for producing EBV data products is often mentioned. There is a shared concern that it is not something that an untrained person can just turn the handle for. There is a set of opinions ranging from "I doubt that canned approaches are optimal" to "the analyses have to be adapted according to the research questions". These reflect a lot of uncertainty among the participants on the precise nature of EBV data products (section 11.6), what they will be used for and how they should be produced. This reflects the diversity of thought in the community on EBV research, development and production. This latter aspect is a theme taken up later in section 12 on translational aspects of EBVs methods research.

4.3 Technical options available

Q7. What are the technical options available and what is possible to achieve today or within the next 12 months (i.e., by mid-2017)? What data and/or workflows, software etc. are available today? Where is it and how can it be used?

Several respondents suggested that it should be possible to make substantial progress within 12 months by, for example working with different user groups and experts, conducting a survey and evaluation of the available tools and resources, adapting from existing tools s and moving towards a framework for data that avoids duplication in data preparation. There are individual tools available that are good for particular tasks. However, the overriding impression coming from the answers is that the landscape is fragmented. There are many tools and technologies that are used by different groups within the community so there is a priority to better understand the needs / requirements to converge on global interoperability.

4.4 Top 3-5 technical challenges

Q8. What are the top 3-5 technical challenges of supporting interoperable EBV calculations on a global basis? How can these be addressed and in what time period? Who has to do something?

Respondents mentioned challenges arising from:

- Data sparsity and variability, with most biodiversity being either completely unrepresented or very under-represented by the available data. In the specific context of abundance, the lack of accurate and widespread recording of abundance data with confirmed absences, (rather than just presence-only data) was mentioned several times as being a substantial barrier to producing useful data products.
- Metadata and quality assurance of data were mentioned again as technical challenges, as was the taxonomic resolution issue.
- Automation of modelling using consistent and rigorous approaches, designing for efficient interaction of computing and human steps, computational capacity, etc.

Tools for the traceability of the entire process, based on an effective system of unique identifiers for biodiversity occurrences / objects are considered important. However, more important is the reproducibility and robustness of the EBV calculations, ensuring that they scale correctly, that results are consistent and that they are trusted by decision makers.

One of the outstanding conceptual challenges for the development of EBVs is agreement on common scales/units/indices of measurement to allow data on, for example, species distributions to be integrated sensibly with data on habitat extent, or allelic diversity with habitat extent. It was suggested that combining different EBVs in a standardised way is the real power of the conceptual approach.

Distributed versus centralised responsibility and procedures for producing EBVs is an important issue. Trade-offs here are to ensure consistency of calculation and adequate resourcing at, for example national level versus the need for resources centrally to carry out this task, and errors arising from limited understanding of the data in a centralised operation.

Many respondents recognised the need for a global, cooperative effort for 3-5 years on developing and documenting a system that is truly accessible to a range of stakeholders that need access to robust and reliable EBV data products. That system is envisaged to comprise a set of specifications, processes and procedures and their technical implementation at national and international level. The role of national and international infrastructure with guaranteed long-term funding is critical in this value chain from raw data (GBIF and BONs) to creating EBV data products through to indicators demanded by IPBES, CBD and others. It is critical and has to be clarified.

This effort has to include agreement on topics such as standard data representations, standard data content, standard methods for defining workflows for manipulating the data into an EBV data product. There are requirements for continued data mobilisation, e-Services as standard building blocks and encouraging more open data access and sharing.

Having the right people with the skills and knowledge to deal with the technical implementation issues (e.g., statistics, workflows, databases) and the associated interoperability and legal issues at a global level is seen as critical to success and as an issue that needs attention.

5 Outcomes from workshop 1

In 5 parallel groups, participants to workshop 1 discussed the above 4 questions and reported back to plenary session. Important points relating to likely workflow steps are summarised in Figure 2 below. Technical issues arising from these discussions, report backs and further plenary discussion are listed thereafter.

Major workflow step	Group 1	Group 2	Group 3	Group 4	Group 5
Data preparation, quality assurance	1. Require automated means for: - Assessing data quality (e.g., ALA's data quality flags) - Assessing restrictions on use (e.g. licensing restrictions). 2. Remove duplicates (same record from multiple institutes: an issue of designing persistent unique identifiers) 3. Encourage use of taxonomic name resolution service	1. Get good quality species distributions data - What species locations available now. Need for good metadata, and to understand the data (presence records). Deal with bias. Characterise where people may have sampled; alongside what? 2. Get good quality co-variance data - Do we have those data available as GIS covariates 3. Expert input on needed co-variance data; possible use of traits information and databases; how variables affect distrib'n.	Workflow to mobilize data more easily e.g., - Automatic extract and convert data and publish through IPT, BDJ (needs to include eDWC) - Data cleaning / check the quality - Create eDWC - Provide polygons in a defined format - GBIF window to extract EBV data DWC as basis; extended DWC including the needs of EBV	Example: Detection of corrected occupancy based on camera trap data 1. Identification of potential data sources - Images along with metadata 2. Observations - Image tagging along with commonly agreed upon metadata standard (Camera Trap federated minimum data standard) 3. Data quality - Based on the metadata proceed to quality control 4. 'Validated' observations - Species pres/abs matrix	1. Define a dataset of species. 2. Integrate and put in relation with other data and metadata: location information, habitat, human density, land cover, genetic information, etc. 3. You don't want to homogenize heterogenic data and loose information.
Modelling	Large matrix inversion 1: Use biodiversity models identified by EU-BON as relevant = {models} 2: For each M in {Model}: Select N dimensions of the hypercube. Run model M against dimensions N. Add model output as another dimension into hypercube	4. Importance of experts' role interacting with the workflows: to know likely distribution, to understand observations that have been made, internally in the model does it look like its predicting well		5. Model (selection of suitable model) 6. Detection corrected occupancy (EBV)	4. Build model keeping track of all parameters (and quality); version it.
Validation		5. Expert validation of EBV / model output / results - Look at the model outputs to see if the expected distribution matches reality - Discuss with experts and adjust, according to understanding - Make a model of records against sampled places.			1. Expert / peer-review validation of models

Table 2: Main workflow steps suggested by group discussions.

The questions about suitable ICT, what can be achieved today and the top 3-5 challenges of supporting interoperable EBV calculations prompted many discussion responses. Participants described these in their own terms, according to their own expertise, backgrounds and experiences. The list below groups similar responses in a logical sequence to provide emerging issues:

1. A range of ICT approaches is likely due to the particular EBV and the organisation(s) responsible for producing the data products associated with that EBV. For example, a remote sensing based EBV could be quicker to operationalise than one that is based on distribution modelling.
2. Different ICT systems may be needed for prototyping and production.
3. Reproducibility.
4. Repeatable 'computer actionable' workflows / processes / procedures that can integrate data from multiple sources. Workflows are a means of documenting and implementing steps in sequence (expressed also as 'a platform to merge the different analysis methods') capturing expert knowledge needed to repeat the procedure.
5. The outputs of the calculations / models have to be captured for re-use e.g., as reference dataset. This is a provenance issue.
6. Workflows should be open to peer review and thus transparent in their description and operation.
7. Execution services are needed to run the workflows and while some are available, they need to be more robust.
8. One set of data services / products is needed for scientists and another peer-reviewed, carefully documented set is needed for policy makers.
9. New products should be easily produced as new data becomes available.
10. Should EBV data products be published via some kind of federated database with links to all the underlying data providers? This is what GEO BON has been considering but practicalities and details are not clear yet. There are similarities with the hypercube approach.
11. A hypercube approach provides a conceptual model that aids community engagement. Everyone can project his/her own contribution to the whole. The hypercube approach allows us to switch to a new paradigm of thinking; thinking in terms of a continuum "hypervolume" (which is an n-dimensional representation of an ecological niche delimited by the values of n variables. Hutchinson 1957, Blonder et al 2014). Delta between two points in the hypercube is a measure of change.
12. Data in hypercubes, of which there may be many has to be discoverable and accessible, so metadata standardization would be helpful. Metadata improves discoverability and navigation around and through datasets.
13. Agreements about data format and safe data transfer protocols are needed for each type of information.
14. No data standards are available yet. Detailed specification and documentation is needed for each EBV dataset/product.
15. For presentation / visualization of information we should use maps showing data in time slices in layers.
16. Substantial and easily accessible (cloud) computing resources are needed for running models that may depend on large datasets of environmental variables and primary biodiversity data. There may be a substantial cost associated with analyses.
17. Some tools are available today that can be applied to producing EBV data products but agreement on applicability is needed.
18. The systems should be usable non-expert users (user-friendly interface).
19. It is likely that new platforms for producing EBV data products will need to be developed.
20. Managing the survey data (plot field data, remote sensing, etc.) is important for any analyses. Maintaining the data structure, information relating to the data quality, completeness and quality of the metadata is fundamental.

21. The challenge of accurate taxonomic identification raises associated resolution and reconciliation issues, and questions about tools to support that. Synchronisation of taxonomies needs improvement.
22. Not only the difficulty of consistently producing any single variable but also the difficulty of consistently producing across several variables so that when they are used together they are meaningful.
23. A suite of different activities is needed to deliver the evidence level (See, for example Hobern et al. 2013). One suggestion is that data is presented as a linked open graph with portals as integrated views of that graph. Data portals of aggregated data are views of the evidence we have. EBV data products will be based on access to available data.
24. What is the minimum metadata informative for EBVs? What is the minimum data to be mobilised for relevance? What kind of metadata do we need for the different kinds of data we are using?

6 Case studies

Workshop participants focused on the EBV class 'Species populations' as the data, models and understanding are the best developed to gain more in-depth understanding.

Participants were encouraged to think about practical implementation cases that could be used to uncover the issues affecting EBV dataset production. These cases were built on current activities and capabilities already in place today. Participants were encouraged to continue discussions after the workshop.

Each case study summary below explains the key activity, the purpose and what is original/novel about it.

e-Bird and atlases: Reviewing all biodiversity monitoring projects but focusses especially on eBird and other projects that employ volunteers to collect observations. The goal is to provide a global monitoring system for biodiversity in general and birds specifically. Participants in such projects can submit observations, photos, sounds, and video via the internet and mobile applications. All data is openly available. For eBird the software stack and functionality is original, as are the data analyses and visualizations of models created from eBird data.

Preparing invasive species data towards EBV species distribution: Developing a workflow to identify key issues towards automation of the use of species distribution data for EBVs across two Data Publishers (GBIF and the ALA). The purpose is to evaluate the viability of Data Publishers' support for species distribution EBVs, document the likely workflow and the issues arising. The originality lies in aligning Data Publishers' support for species distribution EBVs, identification of issues of data integration and evaluation (fitness for use) of GBIF and ALA data using invasive species as the exemplar.

Publishing LPI data through GBIF: The Living Planet Database (LPD) is working with GBIF on the publication of the Living Planet Index (LPI) data through GBIF. This involves converting the data from the LPD into Darwin Core Format and installing the GBIF Integrated Publishing Toolkit (IPT) to provide regular updates from the LPD to GBIF.

Marine EBV pilot: The activity is to identify representative problems in the structure and processing of marine data when calculating Essential Biodiversity Variables for Species populations. The purpose is to gain empirical insight into the validity of existing marine data for EBV calculations and the identification of obstacles for their calculation.

Metagenomics-Metabarcoding-ELIXIR: The goal is to include the entire biodiversity of microbiomes from any habitats into an EBV, shedding new light on their increasingly evident role as promoters or indicators of ecosystem changes. EBVs can be computed simultaneously for a number of species, in the order of hundreds or thousands of microbial taxa from a given sample. Several large metagenomic efforts are already generating datasets that can potentially be used to model microbial EBVs. These include the Earth Microbiome Project and the MetaSUB project. The first aims at generating a comprehensive atlas of the microbial diversity on the planet, with hundreds of thousands of samples collected worldwide. MetaSUB focuses on the microbial diversity of cities and subways and it is attractive as most of these environments types are consistent even in different continents. EBVs based on this data can thus be directly compared. Although the temporal dimension is missing at the moment, association between species presence/absence and geography or environment types will be possible at unprecedented scale. Finally, the relevant complementary activities of ELIXIR, the European infrastructure for biological data, concerning data resources and analytical tools designed specifically for marine metagenomics will also be considered in the composition of the workflow for modelling microbial EBV.

Implementation of Wildlife Picture Index (WPI) as an EBV: The aim is to create an EBV from primary, standardized, high quality species data (~250 species and ~500 populations of tropical forest ground-dwelling mammals and birds) from camera trap data coming from 17 sites in 15 countries. The secondary aim is to expose these data to the GEO BON community of practitioners to improve data use and analytic development of these key datasets. The originality is in the use of standardized sensor (camera trap image) data, easily replicable and scalable.

Table 3 below brings together any immediate information from each of the above mentioned case studies that might be helpful to support identification of ICT-related technical issues and risks.

Case study name	Challenges and risks	Specific missing services	Essential resources
eBird and atlases	Volunteer participation with little training requires specific methods of data analyses to control for observation bias and data quality challenges. These are being overcome and eBird data is now being used in many peer-reviewed scientific publications.	Most monitoring projects need to connect with individuals within country or regions. This is often the most difficult aspect of monitoring projects. Additionally, languages could be a barrier and all aspects of the application(s) need to be translated into multiple languages.	The most essential resource to operate a project like eBird is open access to the web by all people in every country. This is not always the case. For example, mainland China does not provide access to Google maps, which is an important part of eBird for finding the location you were observing birds. Even still, more than 100,000 observations have been collected in China and the rate of growth is @ 40% annually.
Preparing invasive species data towards EBV species distribution	1. Having consistent web services across Data Publishers. 2. Data integration and specifically taxonomic name integration. 3. Criteria for record acceptance (fitness for use). 4. A standard for records testing and assertions across Data Publishers.	1. Consistent web services across Data Publishers. 2. Tools for taxonomic names integration. 3. Criteria for record acceptance (fitness for use). 4. A standard for records testing and assertions across Data Publishers.	1. Infrastructure for executing workflows. 2. Workflows development capability / capacity.
Publishing LPI data though GBIF	1. Getting data into the right format. 2. Gaining agreement from all data providers to LPD to permit publishing the data through GBIF.	Attribution of data with the involved multiple data sources.	Support from GBIF for converting the data to Darwin Core format and for setting up the IPT to continually update the published data.
Marine EBV pilot	- -	Improvement of the data processing pipeline.	- -

Case study name	Challenges and risks	Specific missing services	Essential resources
Metagenomics-Metabarcoding-ELIXIR	1. Spatial dimension is a particularly challenging aspect of a genetic-related EBV. There is an extremely large heterogeneity in any macro-environment and habitats that are directly in contact show only a small number of common species. 2. There is still a major disagreement as to define the "species concept" for microbial organism, even if some operational procedures are commonly adopted. 3. The relative abundance of a species can change as a consequence of abundance shifts of other species and it is difficult to interpret them in isolation. 4. Reproducibility of the sequence data quality and quantity in long term assessments and of the relevant analyses. 5. Internationally supported gold standards for reference molecular databases and their relevant taxonomic classification	Further advances in high-throughput molecular and bioinformatic resources will significantly improve future microbial classifications and identifications.	- Rich and properly annotated molecular reference standard resources, internationally supported, including consistent metadata. - Powerful computational infrastructures and pipelines to manage and analyse in a reproducible and accurate way huge amount of genomic data.
Implementation of WPI as an EBV	Data scarcity, unwillingness of people to share. Knowing when to make data openly available and how to control that, taking into account any legal ramifications of the information contained in the data (e.g., revealing location of protected species).	Rather than exposing data directly to GBIF TEAM are considering sharing data via Vertnet. Ready to produce and share an EBV from millions of camera trap records under TEAM and Wildlife Insight; working to register the EBV with GEO BON and publish the metadata.	To build better API and webs services for TEAM as well as all the Wildlife Monitoring Analytics software (including Wildlife Insights which has WCS data in it).

Table 3: ICT related-challenges arising from case studies.

7 Role of established and emerging infrastructures

7.1 Current and emerging infrastructures

Research infrastructures (RIs) are facilities, resources and services designed to support efficient research. Most usually, research infrastructures are deployments of technologies (scientific instrumentation, computational capabilities, databases, software, networking, etc.) integrated for data management, for experimentation, analysis pipelines, model building and testing, simulation, sharing and collaboration.

RI representatives have been encouraged to think about how RIs can practically support the needs of EBVs, from the perspective of supporting research into the EBVs themselves and for the production of EBV data for use by others.

The information below provided by RIs summarises how they will support EBV data production.

Distributed System of Scientific Collections (DiSSCo): Is developing a European Distributed System for Scientific Collections that relies on: i) generation of accessible and interoperable content (Digitisation); ii) development of e-Infrastructures that makes content openly interoperable and linked; and iii) harmonisation of policies, standardization of processes and expert training. The purpose of the initiative is to allow the scientific collections to respond to urgent societal challenges by embarking on an unprecedented digitisation and data mobilisation effort. Original and novel aspects of the initiative include: i) addressing digitization in a holistic way (taxonomy, imaging, trait (morphology, ecology) extraction, molecular and chemical information, biogeography); ii) scaling up, linking together the largest Natural History Collections in Europe; iii) creating a registry of specimens and experts across Europe; iv) providing a panEuropean access point for all collection data; and v) serving data to other e-infrastructures (e.g., GBIF, GenBank, etc.).

CRIA: Tools and services support data gathering, data transformation and data sharing to enable sharing of biological collection data (speciesLink) and providing mechanisms to generate ecological niche models (openModeller). The purpose of such tools is to support, for example initiatives like “Biogeography of Flora and Fungi from Brazil”[1], where scientists use both resources through a user-friendly interface to generate potential distribution models in the Brazilian territory for ~3,500 plants species until now.

LifeWatch: Is initiating a range of activities that includes working to improve species observations data for EBV use; new individual and community (terrestrial and marine) sensors for generating EBV usable data; developing a Virtual Laboratory offering Earth Observation information for EBVs; and development and use of species and individual trait databases with work on ontologies, semantics and (trait) identifiers for different groups to make systems interoperable across taxonomic and geographic boundaries. The purpose of these activities is to provide support for researchers on EBV developments, and subsequently support to governmental and private organisations interested in EBVs and Indicators. Dedicated virtual environments allowing researchers to focus on their science rather than on technical problems is the innovative and original aspect of this initiative. A number of past and current EU projects, including BioVeL are serving as pilot implementations. Moreover, the Biomolecular Thematic Centre (CTB) of the Italian node of LifeWatch, along with the related Molecular Biodiversity Laboratory (MoBiLab) is providing skills and advanced facilities for molecular and bioinformatics analyses of metagenomic data for supporting microbial EBVs modelling.

LTER Europe: For a network of long-term monitoring sites across Europe focussed on understanding ecosystem processes, LTER Europe is collecting, managing and providing data and information for biodiversity for the development and validation of EBVs. LTER Europe is enhancing the availability and accessibility of long term observation data. LTER Europe seeks to innovate on novel methodological aspects of species and community monitoring that could lead directly to EBV usable data.

CAS Biodiversity Committee: Activities are concerned with: i) building up species distribution database focused on mammals, birds and higher plants; ii) monitoring biodiversity changes via the Biodiversity Monitoring Network in China; and iii) analysing biodiversity status and trends. The activities are for studying rules for biodiversity origin, evolution and change and for supporting implementation activities across China for the Convention on Biological Diversity (CBD).

National Germplasm Bank of Wild Species in China (GBOWS): The key activity is banking of seeds, DNA, plants *in vitro*, microbial cultures, and animal cell lines with a focus on endangered, endemic species and those with potential economic or scientific value. This is for the purpose of collecting and safeguarding China’s wild species biological resources.

NEON: No information available at time of writing.

[1] http://biogeo.inct.florabrasil.net

SANBI: No information available at time of writing.

DataONE: No information available at time of writing.

Elixir: The CNR GLOBIS-B team is also part of the Italian node of Elixir and participates in the H2020 EXCELERATE work package focused on Marine Metagenomics. Its research is focused on the development of molecular and bioinformatic resources for microbiome analysis. Both the pipelines for data management and taxonomic analysis and the molecular data resources can support the modelling of microbiome EBVs.

Table 4 below brings together any immediate information from each of the above mentioned infrastructures initiatives that might be helpful to support identification of ICT-related technical issues and risks that are faced by the infrastructures as they prepare to support EBVs.

Infrastructure name	Challenges and risks	Specific missing services	Essential resources
Distributed System of Scientific Collections (DiSSCo)	Generating the content (new data) and linking them together is a herculean task at the scale envisaged for this Infrastructure, involving data standards and large-scale digitisation activity.	1. Complete taxonomic backbone (as a service). 2. Common URI framework. 3. Agreed upon ontologies and controlled vocabularies. 4. Robust data brokering mechanisms.	1. Cloud Services (Cloud computing and storage). 2. Access to web-services by international data aggregators. 3. Links to services provided by other RIs (e.g. LifeWatch, ELIXIR).
CRIA	- -	- -	- -
LifeWatch	1. Developing generic virtual environments that suit a variety of research interests. 2. Trusted data from multiple sources is needed.	Data services.	Accessing large computational capacity.
LTER Europe	1. Data availability, data harmonization, using a common species list. 2. Data has to be provided by the local LTER sites in a harmonized manner.	1. Standardised data services from data providers (currently under development). 2. Common discovery and access portal (currently under development).	Aspects of data processing and workflow engines are delegated to RIs (e.g., LifeWatch) and European scale projects (e.g., EUDAT2020, ENVRIplus) focusing on these aspects.
CAS Biodiversity Committee	- -	- -	- -
National Germplasm Bank of Wild Species in China	The technology of keeping live materials for as long as possible.	- -	- -
NEON	- -	- -	- -
SANBI	- -	- -	- -
DataONE	- -	- -	- -
Elixir	Standards and guidelines for molecular data interoperability with other EBVs.	1. Metagenomic and Meta-barcoding data/metadata retrieval system and analysis 2. Services for comparative and statistical analyses.	1. Powerful computational platforms 2. Standard reference taxonomies and ontologies

Table 4: ICT related-challenges arising from infrastructure initiatives.

7.2 Provisioning research infrastructures to deliver capabilities for EBV processing

Implementing EBVs at production-scale requires global cooperation among biodiversity research, observation and data infrastructures to serve comparable data sets and offer analytical processing capabilities. This cooperation implies interoperability between the RIs at the level of the service logic (García 2014, Kissling 2015), based on agreed data standards and widely accepted methods (workflows) capable of being executed in any research or observation infrastructure from anywhere in the world. This poses several challenges of provisioning infrastructures to deliver capabilities for EBV processing.

8 General challenges and considerations

8.1 Responsibility for EBV production

Which kinds of organisation(s) are going to produce EBV data products? It is probably not the role of data publishers to take on the full responsibility of building particular EBV workflows, so new infrastructure is likely to be is needed to provide the tools and workflows for EBV production.

There are few RIs today in the area of biodiversity science and ecology that are well-founded, sustained and in a position to take on the responsibility for producing EBV data products. The business case needs to be made for this kind of activity, and for associated sustained funding. The regional Biodiversity Observation Networks (BON) have a role to play. There are multiple kinds of data products related to EBVs. Some products are needed during the phase of research into EBVs themselves. Other products are needed for new science based on EBVs data. A third kind of product is needed to support policy-making and decisions. Different infrastructures can have different responsibilities for each component.

RIs are one option for implementing support for EBV development responsibility, while Biodiversity Observation Networks (BON) focus is more on production-quality. BONs are a better fit for producing EBVs data products as they have the necessary political and financial support. The RIs are more likely to be focussed on supporting research into the development of EBVs data products and EBV methods rather than in their actual production.

It is necessary to understand the scope and responsibility of each of the actors / stakeholders having an interest in EBVs. This includes the established / emerging research infrastructures data providers / publishers and monitoring infrastructures. Data publishers, for example must ensure data have the necessary ancillary information and consistent implementation of data quality assertions [Belbin 2013] that enable potential users to make appropriate decisions about fitness for use. Other actors have to provide the 'super-tools' and workflows for implementing different aspects of EBV production, for example imputing missing variables, generating covariates to facilitate bias correction, applying workflows for standardised modelling applications that allow a level of user intervention, and so forth. Knowing what the steps are in EBV processing is therefore important. This is addressed in section 9.2.

The 'innovation chain' by which the emerging EBV methods progress from being a scientific research output through several stages: from proof of principle, through method / product / service development and into science / business / policy production and use is addressed further in section 12. At all stages there are different ICT issues that have to be addressed.

8.2 The EBV production cycle

Once we understand who is responsible for what, do we aim for an organised, cyclic (periodic) and systematic approach to produce and publish EBV data products, e.g., annually? Or do we intend a much more ad-hoc, on-demand, on-the-fly kind of production process? Each approach places different requirements on the production process for EBV data products, the necessary infrastructure and the informatics.

An analogy: Cyclic versus on-demand printing of books: Consider a book printing. In a cyclic organised approach, the book content (the data) is held by the publisher and sent to be printed in e.g., 10,000 copies every few years, with some revisions each time. In an on-demand approach, the content is held by the publisher and made available for individual readers to print when they want it. The printing, binding and distribution process differs substantially between the two cases.

As with Essential Climate Variables [Bojinski 2014] production of EBV data products has to follow standard, repeatable, open and transparent procedures with clear and accessible documentation and scrutiny at each step.

Some of the technical implications of the two possible approaches to the EBV production cycle are considered in section 10.1 below.

8.3 EBV data model and structures

Hypercubes are multi-dimensional data arrays organised along lines of time, space, species and potentially other dimensions. These arrays can be “sliced, diced, rolled-up or drilled into” [Chaudhuri 1997]. Hypercubes are a logical structure that sits between applications (e.g., specific indicators, specific scientists, or other software examples) wanting to make use of EBV data products, and the underlying data on which the products are based. For each EBV, are we expecting to have a single hypercube that grows over time? Or are we expecting multiple-interlinked hypercubes delineated, for example by geography, by time and distributed by ownership?

Are there more effective conceptual and practical structures that are suitable for EBV production?

8.4 Data standards

When we have decided on macro-data structures, what is the set of core data classes that we consider important enough to standardise (specimen, collection, taxon name, taxon concept, sequence, gene, publication, taxon trait, etc.)? What metadata is required to make the EBV data products discoverable, accessible and usable?

How are data objects linked and how are these links maintained? For example, from specimen to taxon concept to taxon name to publication, or from sequence to associated sequences to taxon concepts to species occurrences?

How can Darwin Core and its extensions help in the production process of EBVs related to species distributions and abundances?

There is a need for controlled vocabularies and eventually for ontologies expressing relations between the concepts defined by the controlled vocabularies. Controlled vocabularies have to be kept to the minimum necessary. Multilingualism must also be supported to allow for mapping between concepts in different countries.

9 Outcomes from workshop 2

Participants to workshop 2 continued the discussions and provided further technical information by answering a survey from RIs and EBV groups standpoints. The latter provided initial material to brainstorming on a possible methodological framework to structure EBV developments based on existing strategic management tools.

9.1 Only open data for EBVs

There is support for the principle to only use open data in the production of EBV data products. This probably means (in the short-term at least) that only a few data sets / sources can be used. However, a requirement for input data to be open could be an incentive for data providers to publish and share under open licensing, since many data providers are under pressure to show value-added use and re-use of their products.

GBIF advocates free and open access to biodiversity data and recently passed a milestone [GBIF 2016] whereby “*all species occurrence datasets on GBIF.org now carry standardized licences, giving data users and* *publishers greater clarity and enabling them to proceed with confidence about the terms of use for data accessed through the GBIF network.*" These are open, Creative Commons licenses.

9.2 Generalised workflow for EBV processing

As illustrated in Figure 2 below, three major types of EBV datasets can be differentiated: EBV-usable datasets (orange workflow steps), EBV-ready datasets (green workflow steps), and derived & modelled EBV data (blue workflow steps). Each set of workflow steps can lead to published datasets of the type indicated.

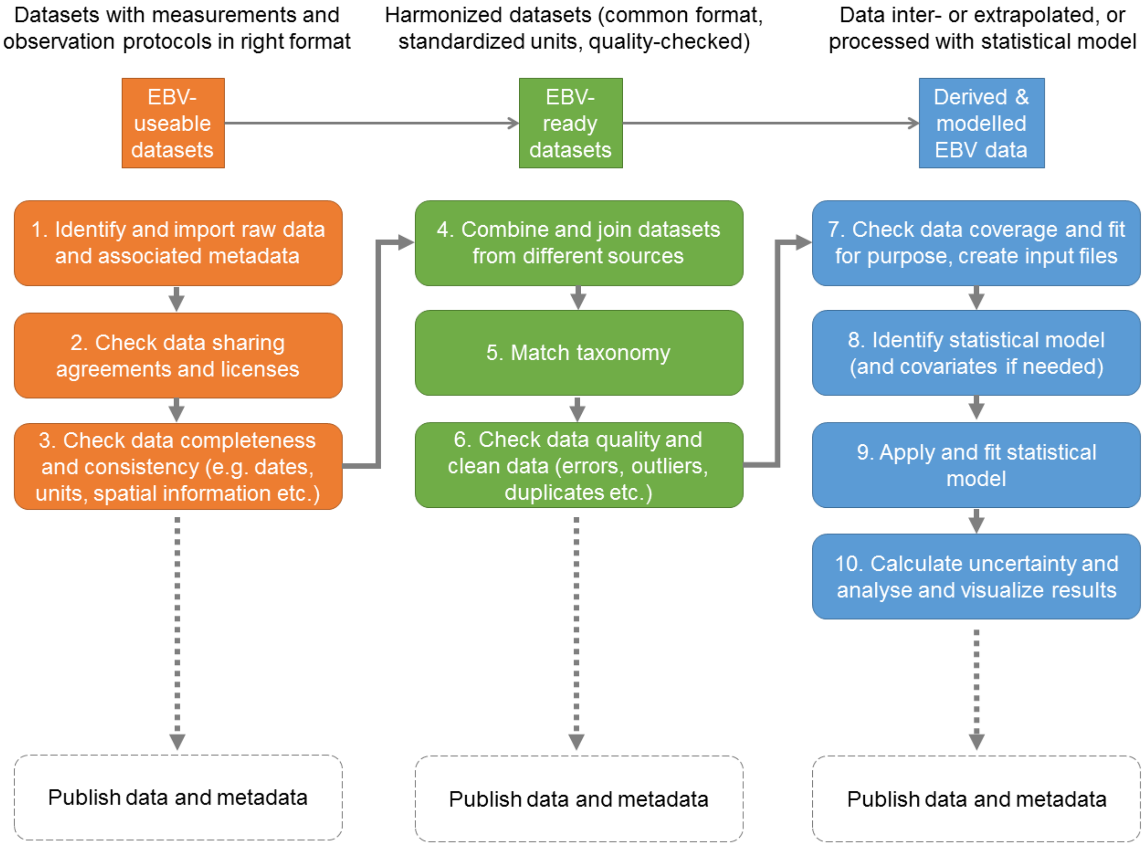

Datasets with measurements and observation protocols in right format
Harmonized datasets (common format, standardized units, quality-checked)
Data inter- or extrapolated, or processed with statistical model
EBV-useable datasets
EBV-ready datasets
Derived & modelled EBV data
1. Identify and import raw data and associated metadata
2. Check data sharing agreements and licenses
3. Check data completeness and consistency (e.g. dates, units, spatial information etc.)
4. Combine and join datasets from different sources
5. Match taxonomy
6. Check data quality and clean data (errors, outliers, duplicates etc.)
7. Check data coverage and fit for purpose, create input files
8. Identify statistical model (and covariates if needed)
9. Apply and fit statistical model
10. Calculate uncertainty and analyse and visualize results
Publish data and metadata
Publish data and metadata
Publish data and metadata

Figure 2: Examples of key workflow steps for building EBV datasets on species distributions and abundances.

It is possible to map each Research Infrastructure to the steps shown in Figure 2 for an indication their scope of responsibility, for their ability to support EBVs production and their gaps in capability and capacity.

9.3 Mapping RIs scope to generalised workflow steps

9.3.1 Dashboard of workflow steps, risks and needs

The technical information has been analysed and visualized in a dashboard, Figure 3 below (expanded for readability in annex 2). The analysis is unfortunately biased since not all questions were answered. Nevertheless, this first exploration highlights some interesting trends and gaps.

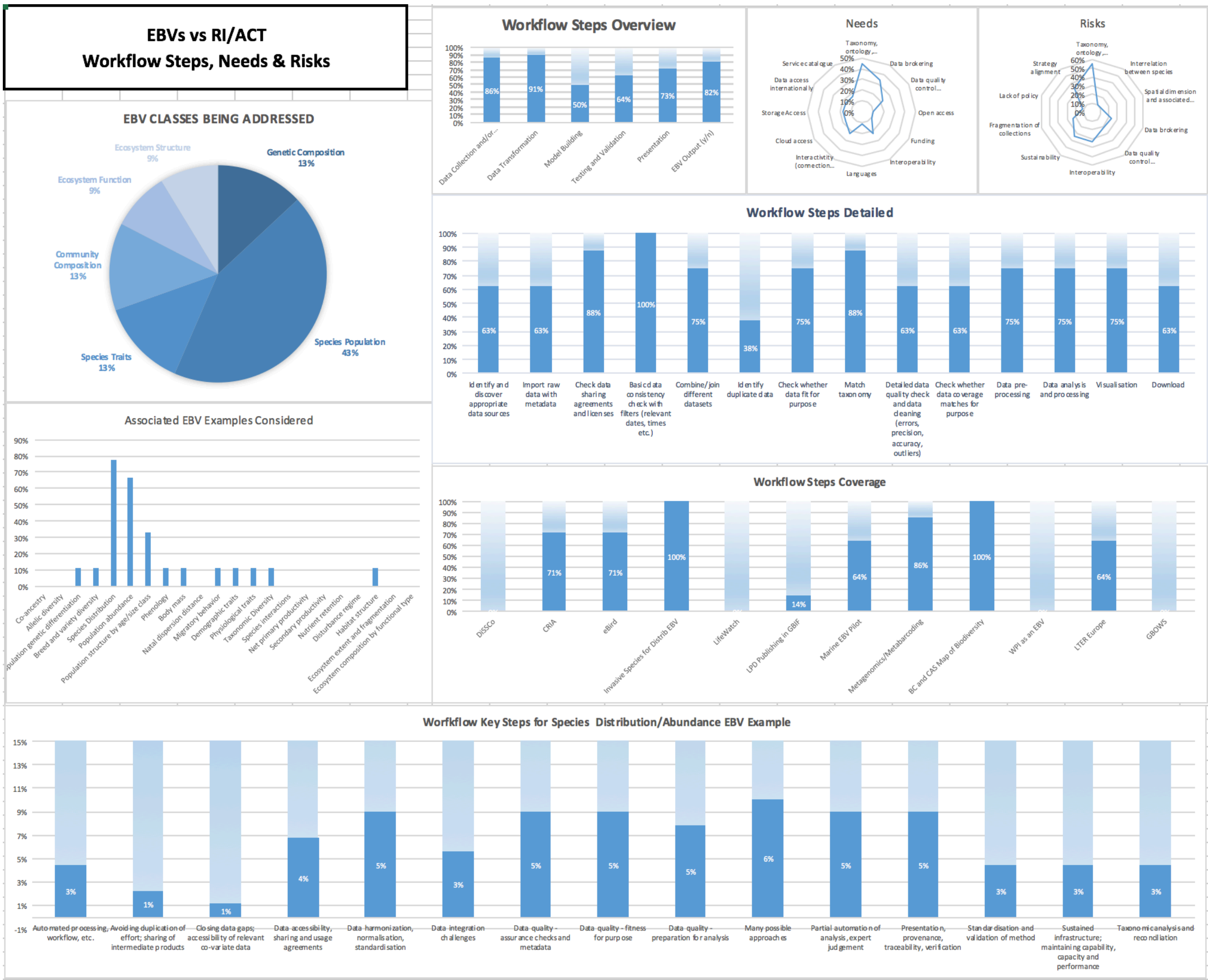


Figure 3: Dashboard: Workflow steps, risks and needs (expanded for readability in annex 2)

This visualization opened the discussion on a methodological framework to support EBV groups in organising their future developments and assessing their related key performance indicators. David Manset (Gnubila) suggested to use a business model canvas as a focal collaboration tool and to adapt it to the needs of EBV developments, i.e., to create a “strategy model canvas for EBVs”.

9.3.2 Strategy model canvas for EBVs

The Business Model Canvas[2] is a strategic management[3] template for developing new or documenting existing business models. It is a visual chart with elements describing an organisation's or product's value proposition, infrastructure, customers, and finances. It assists organisations in aligning their activities by illustrating potential trade-offs. The Business Model Canvas, illustrated in Figure 4 below, was initially proposed by Alexander Osterwalder[4]. It is based on his earlier work on Business Model Ontology (Osterwalder 2004).

[2] "Business Model Canvas." Wikipedia. Wikimedia Foundation, n.d. Web. 9 July 2016.

[3] "Strategic Management." Wikipedia. Wikimedia Foundation, n.d. Web. 9 July 2016.

[4] The Business Model Canvas nonlinearthinking.typepad.com, July 05, 2008. Accessed Feb 25, 2010.

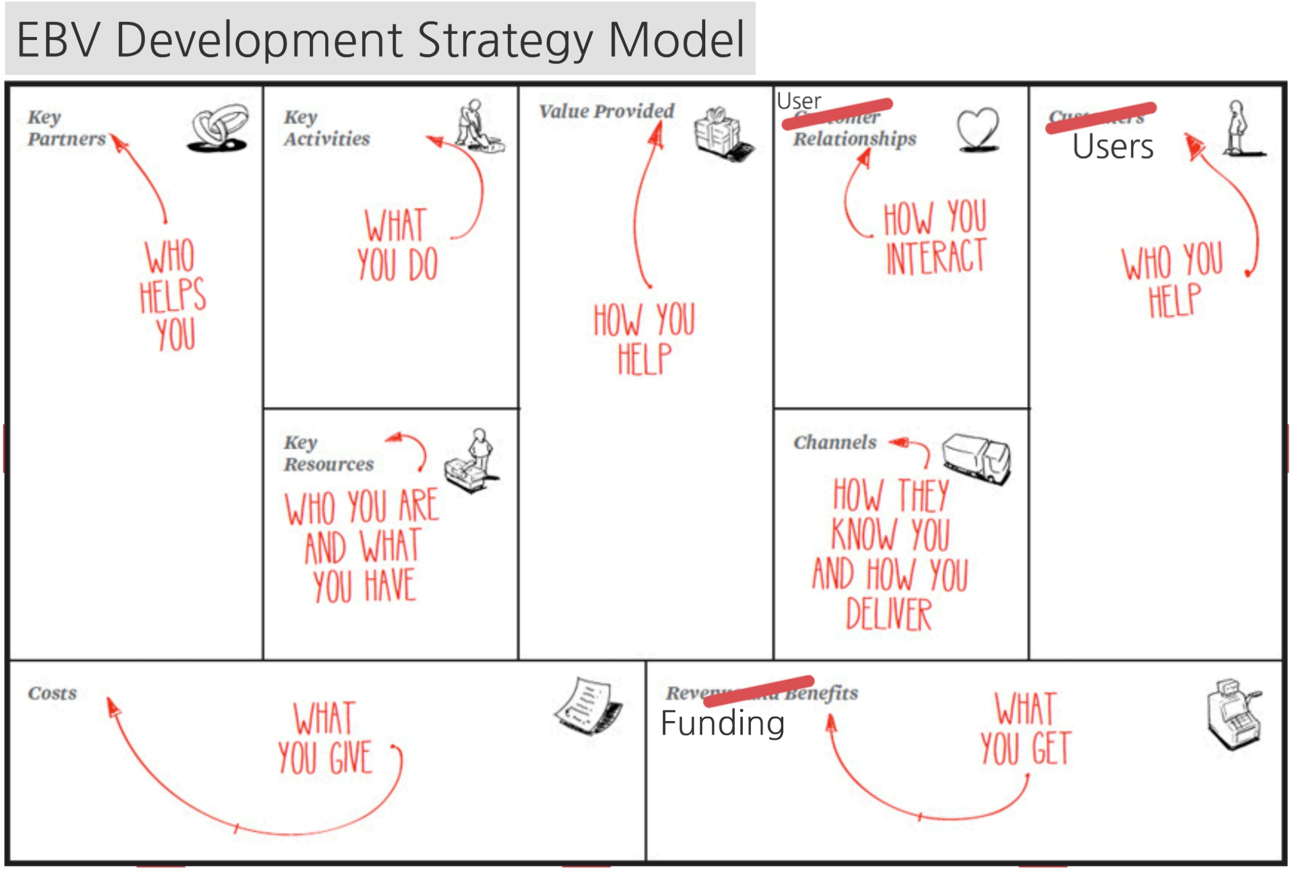
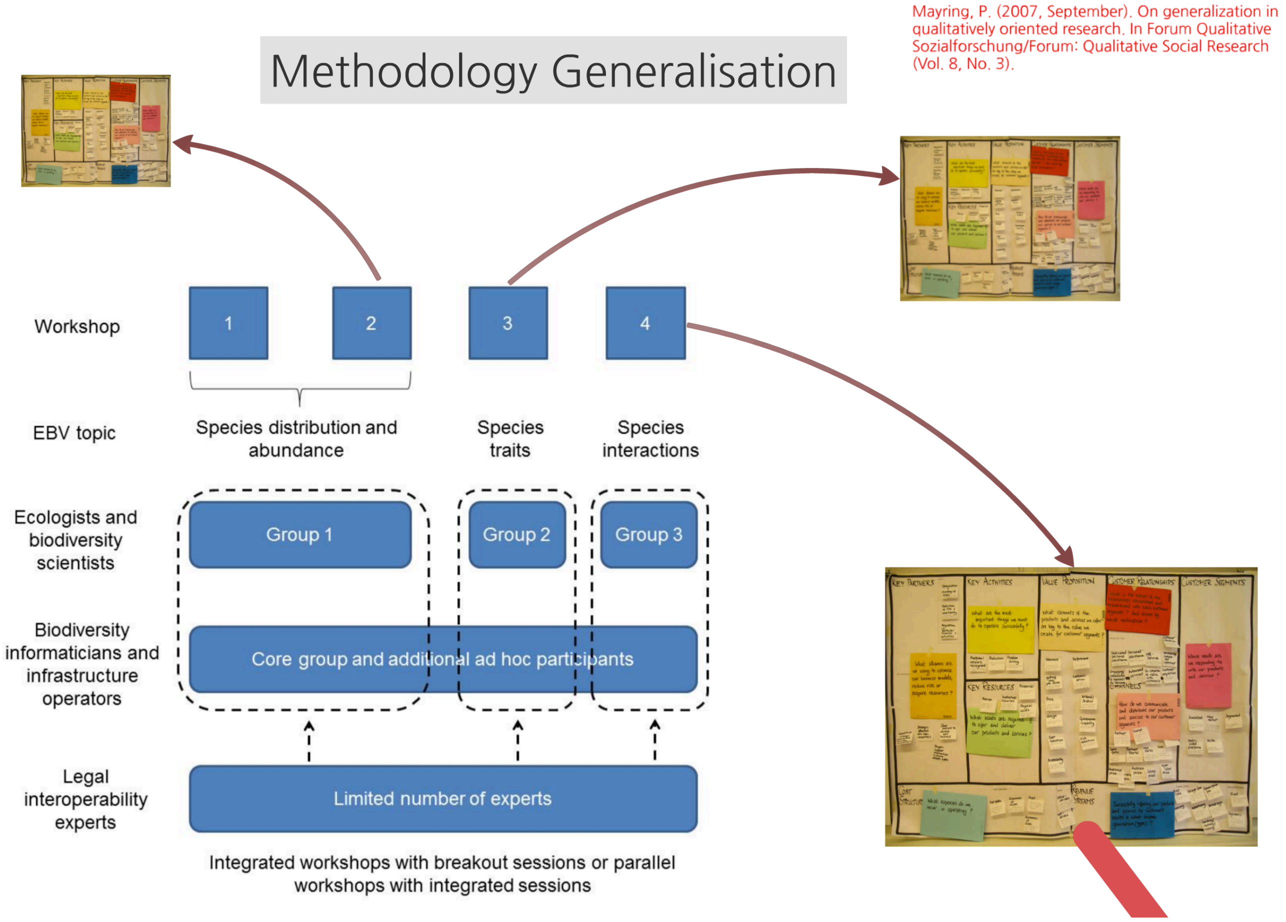


Figure 4: GLOBIS-B's ~~Business~~Strategy Model Canvas

In the resulting strategy model canvas, we use the terms “users” and “funding”. The canvas can be used as a simple 1-page summary of the value provided by the target EBV, its users, how it relates to them and through what channels. The canvas makes it possible to formalise associated key activities, resources and partners to implement the EBV. The canvas also considers costs and sources of funding. Overall, the canvas is a focal point for discussion where participants can place and position elements in the different sections, while thinking about the “flow”, e.g., from the partners to the resources, to associated added value and finally users. It must be seen as a compass and a map at the same time, which once completed gives an overall picture.

As illustrated in Figure 5 below, a first canvas has been presented at workshop 2 that will be used to document work with Atlas of Living Australia on the species distribution EBV. Workshop 3 will develop this further.



Methodology Generalisation

Mayring, P. (2007, September). On generalization in qualitatively oriented research. In Forum Qualitative Sozialforschung/Forum: Qualitative Social Research (Vol. 8, No. 3).

Workshop
1
2
3
4
EBV topic
Species distribution and abundance
Species traits
Species interactions
Ecologists and biodiversity scientists
Group 1
Group 2
Group 3
Biodiversity informaticians and infrastructure operators
Core group and additional ad hoc participants
Legal interoperability experts
Limited number of experts
Integrated workshops with breakout sessions or parallel workshops with integrated sessions

Figure 5: Methodology generalisation

The documented canvases will be available for interested EBV groups. These groups could then take over associated developments and benefit from a single and shared understanding of their EBV strategy.

10 Technical issues for EBV implementation

10.1 On-demand, on-the-fly production versus periodic cyclic production

As introduced in section 8.2, there are two possible technical approaches to producing EBV data products – on-demand, on-the-fly as needed; versus a periodic cyclic approach where EBV data products are produced, published, updated and extended, for example annually, quarterly or monthly.

On-demand, on-the-fly production requires ready access to relevant raw data, the workflow and processing capacity to transform raw data to the selected EBV product for the indicated area (local, regional, national) at the time of interest. Processing capacity "at the touch of a button" is necessary to service the instantaneous demand of the request (and of simultaneous requests). The size and complexity of requests are not known in advance although geographical area and resolution can be constrained. Repeatability is a key requirement, such that if the EBV is again requested on-demand for the same place and time, with the same data then the same answer should be delivered[5]. EBV data production is *ad-hoc*, responding to

[5] Note, of course that when more data for the place and time becomes available this could lead to a different answer.

demands of the moment, with the quality assurance checks in-built in the procedure. Archiving of the EBV products is not required.

In the periodic cyclical approach, EBV data production is more systematic, can be aggregated over large areas (potentially, the whole globe) and published/archived as an ever extending database(s) of information to be queried to provide the data for the indicated place or area of interest (local, regional, national) at the time of interest. Processing capacity can be estimated in advance. Periodicity of the production cycle for an EBV data product can be tuned to the available processing capacity and to the expected temporal sensitivity of that EBV. The information is generated once, archived and then available semi-permanently to be re-used as needed. The anytime, anyplace requirement is met not by on-demand computation but by querying previously computed data products that have undergone a post-production quality assurance assessment.

This choice of on-demand or periodic production is fundamental to the way production processes are defined and for how infrastructures are organised and optimised for calculating, archiving and serving EBVs data products. The choice has to be feasible, efficient, and affordable. Global cooperation is needed to ensure consistency, serving comparable raw data sets and processing capabilities for production and maintaining appropriate archives. The workflows for producing EBVs data products have to be capable of being executed in any infrastructure, and from anywhere in the world. The choice raises issues for permissions to use primary data, for secondary data, for citation and attribution and for provenance tracking.

10.2 Quality, integrity and accessibility of data

10.2.1 Data quality

Knowing whether data are of sufficient quality is a matter of knowing what the data is to be used for and of knowing the requirements that data has to meet for that purpose. In the EBV context, the challenge is to define the requirements the source data has to meet to be fit for the purpose of producing EBV data products. Such data is known as 'EBV usable data'. This is not an ICT challenge.

The technical ICT challenges come from filtering out non-compliant data, or enhancing data to meet the requirements. Modifying data to make it fit-for-purpose conflicts with the need to maintain integrity of the data (section 10.2.2 below). However, reconciling data collected by different methods over time is essential to provide a like-for-like basis for further processing. Automated workflows for data discovery, access, retrieval, preparation and quality assurance can help to minimise and manage the consequences for data integrity.

Data quality assertions or flags based upon tests and corrections made by data publishers at the time data is deposited in their repository can go a long way towards helping potential users (whether human or machine) to know when the requirements are being met. This is dependent on the requirements for EBV usable data being expressed in comparable terms. Quality assurance can be based on 'errors by exception' checking; either manually by inspection or (more desirably) in an automated way. Such assertions and flags may also be associated with data after its publication, using emerging tools and standards for annotation.

10.2.2 Data integrity

Data integrity[6] is fundamental if EBV data products are to be used for scientific research or for environmental policy and decision-making

[6] According to the USA's FDA more than 20 years ago, data integrity refers to completeness, consistency and accuracy of data; also including elements of being attributable, legible, contemporaneously recorded, original or a true copy and accurate.

Data should be secure, traceable and free from manipulative modification (either accidental or deliberate) during the EBV production process. Preserving data integrity becomes challenging when integrating data from multiple sources and passing data through multiple steps/transformations/processes, each perhaps operated by different service providers. There is a need to ensure that data are passed through consecutive steps in the processing chain in their entirety i.e., with their contextual metadata. To reduce the possibility of introducing errors, there is a need to avoid manual manipulations of the data at all stages. Data management systems and processing workflows have to assist in preventing mistakes from being made and in alerting human operators when problems are detected.

Fixity provisions protect datasets that are too expensive to reproduce in the event of loss, as well as preventing them from changing after publication. Persistent Identifiers (section 10.12) have a role to play here. Decisions have to be taken on the level of fixity needed, balancing cost of providing it against likelihood of change in the data products, potential for their loss and liability arising from that loss or from manipulative modification.

Integrity goes hand-in-hand with data quality. It’s difficult for data to be complete, consistent and accurate if quality requirements are not being met for the data. Integrity is also related to adequate metadata, to provenance and traceability (section 10.9) and to standardization (section 10.11), especially of data formats.

10.2.3 Accessibility of data

Availability of data refers to having the right data in sufficient quantity and quality that is capable of being used for the task at hand. Data that is appropriate and sufficient is not the technical issue here but access to these data is. This is a technical, procedural and legal issue.

Accessibility of the available data means:

- Can we discover the data? Has it been published and catalogued somewhere? Is there sufficient metadata and can we discover it by interrogating the catalogue (i.e., from a software program) to find the data we need?
- Can we retrieve the data? Is that retrieval process manual or automated? Can it be done by a piece of software? What metadata information is needed to retrieve the data?
- Are we allowed to use the data (i.e., due to any licensing restrictions)? Is there a price to pay? What are the mechanisms for signifying our agreement with the licensing conditions and for making payment (if any)?

Accessibility can also mean: Is the data we want to use ‘open data’[7] or closed data?

Giving workflows and other software direct access to validated collections of open data (i.e., through standardized Web service APIs) minimises inconsistencies, the dangers of accidentally incorporating either poor quality data or closed data, whilst also maintaining integrity of the data into the processing chain. Working on the technical interface between the EBV data production process and the relevant data publishers is a priority.

Although the provision of standardised data services (Web service / REST APIs) makes it easier for software to discover, retrieve and utilise data, and in the long run reduces development and integration effort, this is by no means an essential requirement to be placed on data publishers.

[7] In the sense of the definition at http://opendefinition.org/

10.3 Taxonomic matching and reconciliation

A ‘complete taxonomic backbone as a service’ is mentioned as being a missing part of the infrastructure for producing EBV data product. The need for taxonomic look-up, matching and reconciliation capabilities is mentioned several times as required. What is meant by these needs? What is necessary for data preparation for EBVs?

The EU BON Unified Taxonomic Information Service (UTIS)[8] is described as being a ‘taxonomic backbone’. According to EU BON, the UTIS allows a federated search to be run on multiple European checklists to return a unified result set of the individual responses of the various checklists. UTIS connects the web services of the Pan-European Species directories Infrastructure (EU-Nomen), EUNIS which fully covers Natura 2000, the Catalogue of Life (CoL), and the World Register of Marine Species (WoRMS). UTIS can be used in full compliance with Appendix 3 of the INSPIRE directive. Currently, it is possible to search for taxa and synonyms by scientific name or vernacular name strings. Where the search string matches a synonym, the accepted taxon is resolved. Furthermore, UTIS also provides a search mode for resolving taxon identifiers. The responses of the search web service always include information on the classification and optionally on related taxa as far as this data is delivered by the checklist providers.

Similarly, VLIZ (Belgium) builds a central Taxonomic Backbone (TB) for LifeWatch[9]. According to LifeWatch Belgium, this TB includes species information services - taxonomy access services, a taxonomic editing environment, species occurrence services and catalogue services. TB brings together different component databases and data systems. In addition to taxonomic information (taxonomic databases, species registers and nomenclatures), the TB will also include biogeographical data (species observations), ecological data (traits), genomic data and links to the available literature.

A further category of tools is those that provide semi-automated help with matching names and cross-mapping between taxon concepts. Chinese Academy of Sciences Beijing Insitute of Zoology has its “Taxonomic Tree Tool” (TTT)[10] while Cardiff University offers its Cross-mapping Tool[11]. TAXAMATCH[12] supports fuzzy searches (e.g., typos in names) and is widely acknowledged and used.

10.4 Balancing automation and expert input

There is a need to achieve a balance between having an automated approach based on common tools in a sequence of steps (a workflow) and the essential application of expert human input at multiple points to verify and assure the correct progress of the procedure.

Various factors assume differing importance according to the phase of EBV work (see 12.1 below).

During the research phase of EBV development, access to intermediate outputs is essential. Traceability of the inputs, algorithms and parameters is key. The researcher must also have the opportunity to adjust parameters of each step to achieve the desired effect. This phase is an iterative approach to developing the methods.

When the method becomes stable, reviewed, and accepted by the community, there is less need for human or expert intervention. Automation for efficient use of computational resources, speed and performance, etc. becomes more important than manual intervention. Human intervention is needed for occasional quality

[8] http://cybertaxonomy.eu/eu-bon/utis/1.0/
[9] http://www.lifewatch.be/en/taxonomic_backbone
[10] http://ttt.biodinfo.org/TF/
[11] <url needed>
[12] http://www.cmar.csiro.au/datacentre/taxamatch.htm

control purposes or selection of choices from a restricted number of agreed parameter ranges – time period and geographical area being obvious ones.

10.5 Standardised workflow approach

10.5.1 Rationale for adopting a standardised workflow approach

Workflows support automation of routine tasks. They are robust and reliable and the work they perform is reproducible[13]. Regardless of precise implementation mechanism, workflows are ideal when a standardised, repeatable approach with traceability (provenance) is needed (Davidson 2008, Goble and De Roure 2009). Workflows are a standard way of doing something and can be an approved procedure in a regulatory environment.

Production of EBV data products will inevitably require a substantial level of data integration across a large and dispersed number of data providers, as well as complex data preparation and processing steps. Historically, such work has involved a combination of manual work and bespoke computing 'scripts' coded in a variety of programming languages, with multiple and different user interfaces. If such processing steps are meant to be maintained and repeated over many years, it is unlikely that the experts, their skills, or the costs to support them remain the same. Hence it is advantageous to develop and preserve all data access, integration, and processing steps of a method for preparing EBV data products as a service in a workflow-oriented infrastructure [Hardisty 2016]. However, a single and fully automated EBV production workflow may not be realistic, as such processing will always need some degree of flexibility that can be expensive to develop and maintain in a workflow infrastructure. However, the most generic key steps of an EBV calculation could be provided by workflow systems.

As we have seen (section 9.2, Figure 2) a generalised set of workflow steps for producing EBV datasets can be formalised to take into account multiple sequential activities, such as aggregation of various raw data sources, data quality checks, identifying duplicate data, taxonomic name / cross-mapping, choosing/using appropriate statistical models and calculating uncertainty. These kinds of steps are likely to be needed in most EBVs procedures.

10.5.2 Main issues

The main technical issues around workflows implementation are:

- Choice of appropriate workflow representation language, and the associated execution mechanism;
- Implementation of a distributed service model in which workflows orchestrate execution of multiple independent services exploiting data from multiple sources;
- To ensure workflows are sufficiently transparent in their description and operation to be open to scrutiny and peer review;
- To ensure ease of maintenance and enhancement over many years;
- The inclusion of flexibility to permit human expert inspection of intermediate results and input to control the execution procedure.

The choice of appropriate workflow representation language must acknowledge that R is widely embraced as a *de facto* programming language for ecology today, mainly because it is open, extensible and has the widest variety of ecology-specific R packages available. On the other hand, languages such as Python, C or Java, with their supporting programming and execution environments can bring different advantages not available from current R environments. EBV production has to be executed in and across a variety of different

[13] Provided that steps are taken to preserve not only the data used but also the processing steps encapsulated by the workflow. This is the notion of 'research objects'. See, for example http://www.researchobject.org/.

infrastructures. Thus, factors such as abstraction and separation from underlying computational resources arrangements and comprehensibility have to be taken into account to ensure workflows are sufficiently easy to use, transparent and maintainable. An ability to interface to and interact with a wide range of heterogeneous localised and remote service types is essential.

Some execution and data services are available today but there is a need to move towards common and more robust ones that are infrastructure oriented rather than desktop oriented. There should be no software installation dependencies for the user. Capabilities have to be flexible enough ('elastic') to support fluctuating levels of computational and storage demand that cannot be known in advance. Fundamentally, tools and services deployed for light, local, benign use by (a few) well-known or internal users are not sufficient. Services have to support heavy, remote, and possibly malicious use by 'risky strangers', as well as use by multiple users simultaneously. That is to say, service providers, although acting locally have to think globally.

Practitioners have to be sufficiently skilled to develop and test the necessary workflows for such an environment. They have to have knowledge of the infrastructure(s) and of the available services e.g., by discovering them from the Biodiversity Catalogue[14]. They have to understand how to optimally exploit computing and storage resources for highly scalable applications working in conjunction with large quantities of distributed data.

10.6 Data product storage and publishing

Large amounts of derived data delivered by the EBV production process will have to be stored and managed on a long-term, semi-permanent basis. These data products are likely to be produced in a distributed manner so how should they be published for the community? This will be especially challenging if we decide to take a distributed hypercube approach (8.3 above).

Whatever the approach, we must define the common specifications for each EBV data product. What does the data product contain? How is it structured? What are the dimensions of the product? Are any dimensions common across all or a subset of EBV data products? In what units is the data presented? What metadata is needed? How is the data represented and stored? An appropriate body, such as GEO BON, TDWG or RDA has to be responsible for these specifications. See section 10.11 below on Standardization.

For each family of EBVs, some agreement on common dimensions seems required. For the overall collection of EBVs, a few common dimensions are needed. Three common dimensions needed by every EBV are time, space and species name / taxa.

A rough calculation is needed to determine how much physical storage capacity is required. Initiatives are needed to take on the burden of the storage commitment, which has to be funded, of course. Such calculations have to take into account whether the semi-permanent storage of EBVs data will be done centrally (i.e., in a single repository, perhaps with duplicated mirror sites) or as a network of distributed, independent but interoperable storage services.

10.7 Implementing the hypercube approach

nDimensional cubes, implemented using traditional relational databases have long been a feature of data warehouses, where relational queries are used to "slice, dice, drill-down and roll-up" subsets of the data for "OnLine Analytical Processing" (OLAP) in business intelligence applications[15]. Today, array-oriented

[14] Biodiversity Catalogue is the Web services registry for the biodiversity sciences. It can be found at www.biodiversitycatalogue.org.

[15] For an insight into directions in the hypercube paradigm as the basis for business intelligence and decision-support applications, see this 2008 blog post: http://blogs.forrester.com/james_kobielus/08-07-14-olaps_cube_crumbling_around_edges.

databases like SciDB, Rasdaman, and MonetDB are increasingly used for their ability to easily handle n-dimensional arrays.

A hypercube approach is proposed as the mechanism for warehousing EBV data products (see section 8.3).

The technical challenge is to manage a collection of hypercubes when they are physically distributed, owned and loaded by different organisations, with roll-up across hypercubes being necessary to achieve a regional or global view of the data. Cataloguing these hypercubes is required for discovery and effctive use..

10.8 Standardised APIs and meta-description of them

As ‘software eats the world’ [Andreessen 2011] and software development itself comes to rely more and more on component-based approaches, Application Programming Interfaces (API) become by far the most important mechanism for achieving interoperability between services. RESTful API Modelling Language (RAML)[16] is a standardized mechanism for designing and describing APIs.

Well-known APIs with standardized meta-descriptions make it easier to access data resources and associated services. When details of these APIs (services) are available in a well-known catalogue such as the Biodiversity Catalogue[14], it is easier to link across infrastructures (11.4 below).

Data publishers, tool providers, service providers and infrastructure operators should all encourage and embrace a cultural shift toward well designed, well described, standardised APIs that are easy to find, to use and manage.

10.9 Provenance and traceability, reproducibility and repeatability

Needs related to provenance and traceability recur often in conversations about using and processing data. Tracing data back to its original source is frequently stated as necessary for evaluating results and conclusions. “What data was this based on?” is the question most often asked. Tracking of data citations and re-use of data is necessary for the funding of a data observation/collection activity. Confirmation that data licensing conditions have not been breached is required. Information about the tools and capabilities that were used to do an analysis is needed so that the analysis can be reproduced.

To what extent is it necessary to be able to reproduce the entire sequence of data collection, processing and events that leads to a particular EBV data product? If the process itself is replicated elsewhere (e.g., in multiple infrastructures) will it reproduce the same outputs? How important is reproducibility and why? What should happen in the event of a disaster (fire in a data storage facility, for example)?

Reproducibility is not the same as repeatability[17]. Repeatable ‘computer actionable’ workflows / processes / procedures are essential to ensure that the process can be carried out over and over again i.e., whenever new data product is need.

Tools for traceability have to be built using standard mechanisms (such as the W3C PROV family) for collecting provenance information, which itself has to be enabled throughout the tool-chain. To capture the relationships that constitute provenance (e.g., composition, transformation, filtering, etc.) persistent

[16] http://www.raml.org/

[17] Reproducibility is the ability to duplicate or replicate a sequence of actions, elsewhere and/or by another person working independently. Repeatability is the ability to perform the same task/workflow, by the same person on the same item, under the same or very similar conditions (e.g., with slightly different input parameters) and within a reasonably short period of time after the first time it was done.

identifiers (section 10.12) are key for unambiguously identifying the entities (e.g., datasets, algorithms, intermediate products, etc.) linked by those relationships.

10.10 Distributed versus centralised processing

Distributed versus centralised responsibility in the procedures for producing EBVs is important. We refer to this as the "warehouse choice" because of its similarity to the choice between a single central warehouse to supply for example, all food supermarkets, or a set of distributed autonomous warehouses, each serving a smaller number of supermarkets, or possibly just one.

Distributed or centralised processing is not a computer or group of computers in a single place versus using computers distributed across multiple locations. Distributed or centralised processing is also not a choice between using an organisation's own computers, or outsourcing to a 3rd party (such as a cloud computing provider). Those choices are separate from the warehouse choice. See section 10.13 on ICT scenarios.

In the warehouse choice we are concerned with allocation of responsibility for performing the task of producing EBV data products, considering:

- Location and ownership of the data needed for the task;
- Geographical area to which the required EBV data product relates;
- The need to ensure consistency of calculation across all producers of EBV data products;
- The level of resourcing needed at, for example national level versus the need to support major resources more centrally to carry out the task;
- The risk of errors arising, either through limited understanding of the data in a centralised operation, or through inconsistent application of the procedures in a distributed operation.

The warehouse issue is a socio-political-technical issue, to be considered in relation to organisation, financing and governance of the EBV procedures, taking into account national and global reporting requirements and likely use of EBV data.

One scenario, illustrated in Table 5 foresees primary data observations and initial processing organised along political administrative divisions at a national (or lower unit) level, with aggregation and roll-up to the level of regions and continents[18]. Within such a two-level scheme, the lower levels can organise their own centralised or distributed computing while the upper aggregation level can also organise itself in a centralised or distributed manner. The upper level can harvest automatically from the lower level.

Purpose and level of data processing	Unit responsibility for data management and processing	Computing and storage provision
Primary data observations, initial processing and publishing	National or lower levels of administration	Can be centralised at or distributed across the unit level
Aggregation (roll-up), interpolation, extrapolation - for wider use	Supra-national / regional	Centralised for storage and publication; Single global source, perhaps replicated at regional level for security. Processing per roll-up region Automated harvesting from lower level. Within region processing can be centralised or distributed.

Table 5: Example structuring for hierarchical processing

[18] IUCN Protected Areas, for example.

Section 11.6 suggests adopting the nomenclature of NASA data processing levels as the means to define EBV data products more precisely. Using such an approach could also help to determine issues related to centralised or distributed processing and associated responsibilities.

10.11 Standardization

10.11.1 Choice of standards

Standards applicable to EBV data, data products, formats, exchange and services fall into three main categories.

The first category is concerned with selecting from those more general-purpose standards, recommendations and specifications already published by widely recognised organisations such as ISO/IEC, W3C, Open Geospatial Consortium, OASIS, RDA, etc. Such standards are independent of the EBV-specific nature of the business. They may cover for example, standard means of persistently identifying data, standard means of transferrring it over networks and standard means of recording and tracking its provenance.

The second category is concerned with selecting from domain-specific standards such as those under the responsibility of TDWG (see below) for representing and exchanging various types of biodiversity data. Again, these are likely to be pre-existing specifications such as Darwin Core and TAPIR.

Finally, there is a category of specifications specific to the nature of the EBV business that do not yet exist. The members of this category have to be identified, defined, written and agreed (10.11.4 below).

10.11.2 Issues of standardization

Specification of data formats has to be based on an understanding of how the data is most likely to be used across different categories of end-users. Similarly, specifications of data transfer protocols will be based on how, what and when data is transferred. The way that data is to be stored and made available (published) and how that is organised (8.3 and 10.6 above) has also to be taken into account.

The issue starts with the relevant primary biodiversity data that is already being collected and published, and ends with the EBV data products for researchers and policy and decision-makers.

It has been mentioned that safe data transfer protocols are needed. What does 'safe' mean? Safe from what? This has to be further studied.

10.11.3 Compliance with requirements of the INSPIRE Directive

In Europe, EBV data are part of the "Infrastructure for Spatial Information in the European Community"[19] and have to be discoverable through the INSPIRE Geoportal[20]. Thus, compliance with the requirements of the INSPIRE Directive [EU Parliament 2007] is mandatory for data products and services related to EBVs in the EU. These requirements, for which many technical specifications have already been published include specific provisions relating to metadata; harmonisation of spatial data; unique identification of spatial objects; services for discovery, viewing, downloading and transforming data; interoperability of services and data sharing, etc.

There are unlikely to be any significant incompatibilities arising when aspects of the specifications put in place to meet European mandatory requirements are applied more widely in a 'best practice' sense for all

[19] http://inspire.ec.europa.eu/

[20] http://inspire-geoportal.ec.europa.eu/

EBV data, products and services. By this we mean, for example that metadata defined for specific purposes in Europe is likely to also be useful elsewhere.

10.11.4 Responsibility for standardization

Which body(s) should be responsible for new EBV data standards?

GEO BON[21] the Biodiversity Observation Network of GEO, is a part of GEO, The Group on Earth Observations. GEO BON is building the pathways to link biodiversity data and metadata to GEOSS, the Global Earth Observation System of Systems, which will provide decision-support tools to a wide variety of users. GEO BON is a response to the lack of organised or coherent infrastructure to collect the biodiversity information necessary to monitor progress towards objectives of the Convention on Biological Diversity (CBD) Strategic Plan. The **GEO BON Secretariat** coordinates work on Essential Biodiversity Variables and is the most likely body to coordinate activities related to standardization. However, it is not necessarily the appropriate body to carry out the actual work of drafting and agreeing the necessary standards; an activity that requires a high degree of technical skill. Such a task could be delegated, for example to **GEO BON Working Group 8 on data integration and interoperability, informatics and portals**[22].

'**Biodiversity Information Standards (TDWG)**'[23] focuses on developing and promoting standards for recording and exchanging biological/biodiversity data about organisms. TDWG is responsible, for example for the Darwin Core (DwC) standard and has a future role to play in EBV related standards.

The **Research Data Alliance (RDA)**[24] is a community-driven organization that has the goal of building the social and technical infrastructure to enable open sharing of data. It was initiated in 2013 by the European Commission, the United States Government's National Science Foundation and National Institute of Standards and Technology, and the Australian Government's Department of Innovation. Working in close cooperation with the International Council for Science (ICSU) World Data System (WDS)[25] and its Committee on Data for Science and Technology (CODATA)[26], RDA is supported by members from more than 110 countries and has a wide remit covering all aspects of data standardization. Many of the recommendations emerging from RDA are applicable to environmental sciences and hence to EBVs.

10.12 Persistent identifiers

Persistent identifiers (PID) are the means by which long-lasting (i.e., semi-permanent) references to important artefacts such as datasets are created. Given that there are several different PID mechanisms in present-day use, there should ideally be a global agreement on a single mechanism for identifying all EBV data products. The most ubiquitous PID is the Digital Object Identifier (DOI). DOIs have been widely used for traditional publications and more recently, datasets. Adopting the DOI mechanism allows third-party services for PID assignment, registration and resolution, such as those provided by DataCite and Crossref to be leveraged.

Global agreement will also be needed on the detailed procedures for using a PID mechanism to issue, maintain and track globally unique PIDs. Such an agreement has to include specification of the detailed metadata content to be stored alongside a PID in a PID registry, so that unambiguous citation, discovery and re-use of EBV data products is possible.

[21] http://www.geobon.org/

[22] http://geobon.org/working-groups/working-group-8-data-integration-and-inter-operability-informatics-and-portals/

[23] http://www.tdwg.org/

[24] https://www.rd-alliance.org/

[25] https://www.icsu-wds.org/

[26] http://www.codata.org/

Provision has to be made to accommodate the use of PID mechanisms that are used for persistently and uniquely identifying the source data that contributes to making EBV data products and those used for identifying software and other resources that form part of the production process. Workflows have to accommodate the heterogeneity of identification mechanisms in use around the sources of their input data. These aspects are essential to maintain data integrity (section 10.2.2), for tracing provenance (section 10.9) and for publication, citation and tracking data re-use.

Unambiguous referencing of subsets of data - so-called 'slice, dice, roll-up and drill-down' subsets may also be needed.

Many further considerations relating to PIDs are given in [Atkinson 2016].

10.13 Different ICT scenarios

As suggested in section 10.10, multiple ICT scenarios are possible, balancing different compute and storage options against one another in different ways. Different organisations playing in the EBV data products game are likely to adopt different ICT solutions. To ensure consistency of production of EBV data products, and to foster interoperability of both data products and services for producing them, these differences have to be accommodated by providing a common 'EBV platform architecture' that can be supported by the different organisations involved. By common platform architecture we mean a set of components (microservices[27]) with low variety and high reusability potential that can be adopted by all players[28] and deployed to their preferred ICT solutions.

10.14 Software enforcement of data licensing terms and conditions

Workflow-oriented processing chains have to enforce data licensing terms and conditions from beginning to end. Even using open data carries with it a moral obligation to acknowledge the source of the data and how it has been used to create the final data products. In an automated process, this means carrying the relevant information (or a pointer to it) for any particular data record through the entire chain of processing so that it can be presented as part of the final presentation of the EBV data product.

An EBV data product could be prepared on the basis of some confidential or sensitive data (not commercially licensed, but protected). The EBV data product itself might be public information but not the source data.

11 Technical risks for EBV implementation

The sub-sections below are a set of risks identified so far.

11.1 Readiness / capability of infrastructures to support EBVs production

As noted in section 8.1, it is not clear even that RIs are the responsible infrastructures for EBV production. A few advanced RIs may be capable of supporting research into EBVs and associated procedures themselves, but are not able to support production at scale. RIs today are likely to be providing support at the proof-of-principle stage (see 12.1 and Figure 6). We may see first forays into experimental integration in the coming year or two.

[27] Microservices are separately deployed units, each representing a service component. Each microservice can be administered separately, whilst being used in conjunction with another. The concept of microservices is becoming established as a new pattern of application deployment in response to increasingly continuous 'DevOps' styles of application delivery, especially in relation to 'Software as a Service' (SaaS) style of application delivery on Cloud infrastructures. It is an overall simplification of the distributed Service-Oriented Architecture (SOA) pattern. Microservices work well with workflow management systems such as Apache Taverna.

[28] See also paragraphs on page 36 relating to 'integrated platform for production'.

It is critical to gauge the timescale within which responsible infrastructures can reach the required readiness versus that of the political demand to deliver EBV data products reliably and as needed.

11.2 Open access for service dependencies

In a distributed informatics approach, workflows for producing EBV data products might rely on services operated and provided by third parties. Google Maps is one such example. Another is Google Earth Engine, where the use of a powerful service can require the workflow operator to relinquish some ownership/copyright over the resulting product.

Some organisations and/or countries may not permit access to some Internet based services that are considered essential for correct operation of a workflow or procedure. Therefore, when establishing critical and perhaps long-term service dependencies, it is necessary to consider the factors that may render that service inaccessible or unusable for some operators of a workflow or procedure.

11.3 Multilingualism

All aspects of the EBV products and application(s) have to be translated into multiple languages.

Multiple sources of data for the EBV process may imply multiple language metadata. A computer-assisted thesaurus capability could be necessary to assist automated data integration.

11.4 Linking across infrastructures

Achieving interoperability across infrastructures is the mechanism by which an integrated means of producing and delivering harmonised EBV data products at a variety of resolutions and for different geographic areas can be achieved. When faced with the probability that different ICT approaches can be adopted by each of the different infrastructures, we have to be clear about how interoperability can be achieved. It is the main aim of the GLOBIS-B project to determine how to achieve interoperability, using EBVs as the motivating use case.

The preceding CReATIVE-B project already considered the different kinds of technical interoperability, and the level at which interoperability has to optimally take place [Hardisty 2014].

Achieving interoperability means supporting transactions at the service level across infrastructures - supporting service level transactions between users and / or services. Users or services of one infrastructure should be able to access and utilise resources and services of another infrastructure to achieve their goal.

To achieve such interoperability requires coordinated implementation of selected standards and specifications, using precise definitions ("profiles") of how each standard or specification can be adopted and implemented to solve the specific transaction need. Standardized APIs with standardized meta-descriptions, well-known by being registered in a catalogue such as the Biodiversity Catalogue[29] is one example. Standard data brokering services are another.

A mechanism (i.e. organisation, people) is needed whereby each infrastructure can be represented and can contribute to and make agreements on standards, specifications and their implementation.

11.5 The diversity of tools and techniques

Many possible approaches to producing EBVs have been discussed. What has been suggested is based on workshop participants' expertise, their experience with tools and personal preferences.

[29] Biodiversity Catalogue is the Web services registry for the biodiversity sciences. It can be found at www.biodiversitycatalogue.org.

The best tools for the job have to be selected, based on an objective evaluation of how particular tools meet the specific needs of the procedures. In some cases, several tools can fulfil the requirements and, so long as the output produced meets the needs of the process, there is no reason to exclude one in favour of another. Tools will evolve with experience and research.

The key considerations risks are based on the reliability, ease of maintenance and use of appropriate tools. Focussing on a small number of tools around which the community congregates, and putting effort into improving those tools seems an effective strategy.

11.6 Precise nature of EBV data products

As noted at the end of section 4.2, it's clear from the workshop discussions that there is a lack of understanding and therefore consensus on the precise nature of EBV data products. We know (see, for example the top part of Figure 2) that EBV data products start from primary observation datasets, moving through a process of harmonisation eventually to produce interpolated/extrapolated datasets. This is however a vague definition that has to be substantially improved via a more formal specification of the details for each EBV data product.

One possibility could be to adopt the nomenclature of NASA data processing levels (0 – 4)[30], attempting to define: a) the characteristics of the EBV data at each level; and b) the processing needed to move data from one level to the next. This would help to determine what has to be processed (and where – see also section 10.10) as well as what has to be stored/retained, and published as usable data products.

12 Translational risks of EBVs methods research

12.1 The innovation chain

The steps from where we are today towards quality-controlled regular EBV production are stages of what is known in the commercial/industrial world as an 'innovation chain' [Kline 1986]; extending from fundamental laboratory research, to applied methods, a product or service development phase, field testing, possible regulatory approval, in-use evaluations and ultimately to widespread adoption and use (Figure 6).

[30] https://en.wikipedia.org/wiki/Remote_sensing#Data_processing_levels

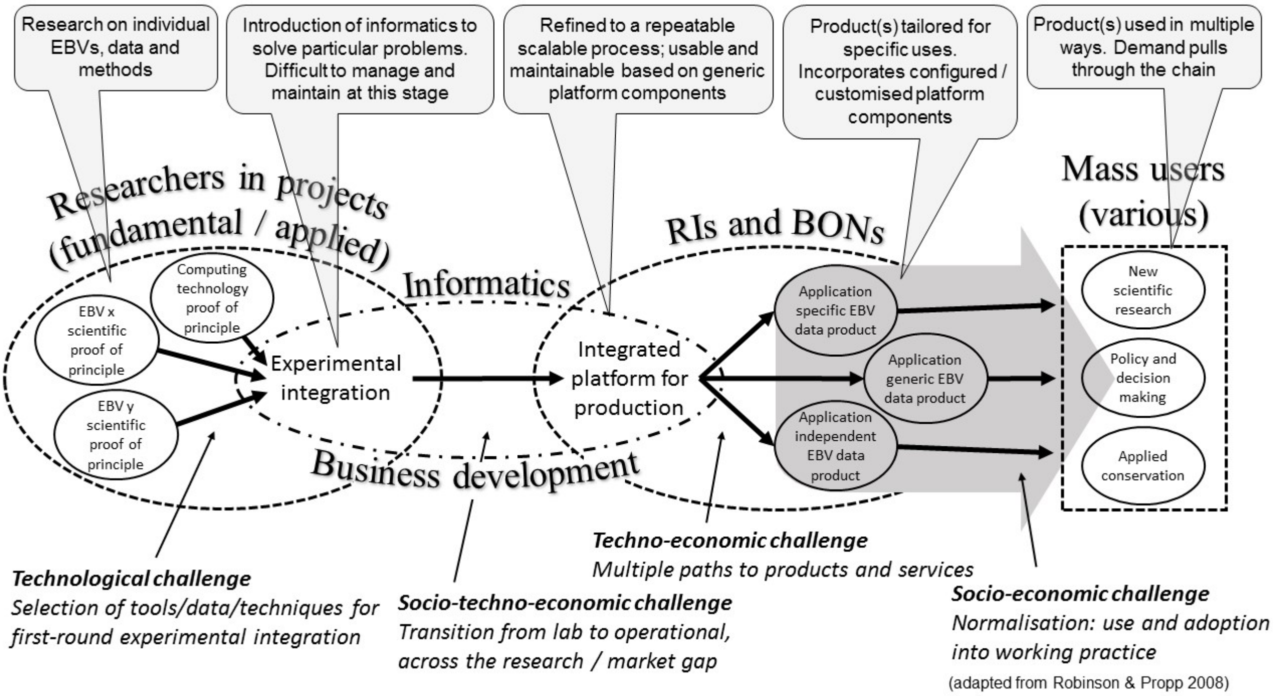

Research on individual EBVs, data and methods
Introduction of informatics to solve particular problems. Difficult to manage and maintain at this stage
Refined to a repeatable scalable process; usable and maintainable based on generic platform components
Product(s) tailored for specific uses. Incorporates configured / customised platform components
Product(s) used in multiple ways. Demand pulls through the chain
Researchers in projects (fundamental / applied)
Informatics
RIs and BONs
Mass users (various)
Computing technology proof of principle
EBV x scientific proof of principle
EBV y scientific proof of principle
Experimental integration
Integrated platform for production
Application specific EBV data product
Application generic EBV data product
Application independent EBV data product
New scientific research
Policy and decision making
Applied conservation
Business development
Technological challenge
Selection of tools/data/techniques for first-round experimental integration
Socio-techno-economic challenge
Transition from lab to operational, across the research / market gap
Techno-economic challenge
Multiple paths to products and services
Socio-economic challenge
Normalisation: use and adoption into working practice
(adapted from Robinson & Propp 2008)

Figure 6: The innovation chain - transition from research to productisation to mass use
Adapted from (Robinson and Propp 2008)

Reading Figure 6 from left to right, with terms in *italics* being first reference to elements in the figure: *Researchers in projects* carry out research on individual EBVs, and the data and methods needed to support them, leading to *scientific proof-of-principle* of the specific EBV (EBV x or EBV y in the figure). *Computing technology proof-of-principle* can also be researched at this time to solve specific problems, dealing with data quality issues, or investigating how best to store massive amounts of data, for example. A particular technological challenge at this stage is to select the right tools, data and informatics techniques to support a first round of *experimental integration*. That is to say, to find informatics approaches that can be applied in common to implement the scientific methods for production across multiple EBVs. Adopting, for example, a workflow approach towards the generation of each EBV data product would show some level of integration insofar as some steps in the workflow are likely to be the same or quite similar for different EBVs. However, because of the experimental nature of the work at this stage, the steps may not work well without software adaptation or manual intervention during execution of the workflows. Such a prototype is suitable for operation only by persons with sufficient expertise and knowledge to check that everything proceeds as expected through the various steps. Because it's a prototype, some elements may be missing. The prototype may not be easily adaptable for all scenarios, combinations of parameters, etc. nor will all error cases will be properly catered for. Prototype infrastructure at this stage is often difficult to manage, maintain and support.

Transition from laboratory to market system involves *informatics* development and deployment and *business development* i.e., service model and business model definition, planning and deployment, as well as understanding of how the products will actually be used. It is where research ends, and where development into usable, scalable, flexible data products and their associated robust and reliable 'commercial-grade' production and support processes begins. This is the stage that converts the experimental integration into an *integrated platform for production*. By 'platform' we mean a set of stable, reusable core components [Baldwin 2009] that support the production processes for EBV data products. By 'integrated' we mean unified as a whole, with the component elements combined harmoniously[31]. These core components may include both ICT elements and business elements. Together the platform components offer the essential functions for the intended business need.

An integrated platform for production and management of EBV data products may, for example offer generalised components for each of the generalised workflow steps outlined in section 9.2 above that can be used for all EBVs. These components can work in general-purpose utility (cloud) computing environments, such as those offered by Amazon, Microsoft, Google, etc., making it possible for any Biodiversity Observation Network to produce EBV data products from the data collected by the network. These workflow components will interact and work seamlessly with a large-scale EBV data repository database, quality assurance, publishing and archival service offered by the integrated platform that permits different BONs and other actors to contribute to, manage and publish quality assured EBV data products.

The integrated platform, with additional custom components on top makes it possible to derive and produce different kinds of EBV data product to serve a variety of different application purposes. *Application specific EBV data products* are likely to be precise and tailored towards one particular application or type of application. *Application generic EBV data products* are less specialised and useful for a broader range of applications. General-purpose *application independent EBV data products* are neutral as regards any particular intended use and have very broad application. Examples of each are given in Table 6 below.

Example EBV data products	Application class of data product	Typical applications or uses
Phytoplankton primary productivity	Application independent	Monitoring trends in carbon fixation, food web modelling, valuation of ecosystem services for fisheries
Size structure of phytoplankton assemblage	Application generic	Early warning of nutrient enrichment, trends in mixing and potential toxic events
Presence / abundance of particular zooxanthellae clades	Application specific	Assessment of coral reef resistance to temperature change
Wetland extent / recurrence	Application independent	Estimation of habitat for amphibian & other species, climate circulation models, flood/drought resilience and food security for human populations.
Seasonality of wetland	Application generic	Assessing effectiveness of protected areas for migrating mammals and birds.
Wetland plant species composition	Application specific	Identifying succession and drying, invasive species monitoring.

Table 6: Examples of different types of EBV data product

There is the need to identify and understand what the EBV data products and related outputs are likely to be used for in *mass user* (various societal) contexts. Uses can include *new scientific research* based on EBV data, *policy and decision-making*, and *applied conservation* (for example). There are also multiple paths by which the integrated platform can lead to different kinds of EBV data product. As suggested in section 7 above, responsibility for maintaining and operating the integrated platform and for producing the EBV data products most likely lies with the individual Research Infrastructures and Biodiversity Monitoring Networks (bubble labelled '*RIs and BONs*' in the figure). Actors such as data publishers also have roles to play. Finally, an understanding of how EBV data products get adopted into everyday working practices of the mass users is required.

[31] A well-known example of an integrated platform is the Microsoft Windows operating system, which contains a defined set of core valuable components (such as, for example the ability to display multiple windows on a screen, the ability to interact with the user through dialog boxes, and file handling) that can be exploited in a broad range of different applications (email, word processing, etc.).

As illustrated by the large grey arrow in Figure 6, demand from and needs of mass users create a 'pull' from the market – the right-hand side of the innovation chain – that can expedite resolution of the many other challenges earlier in the chain. Those challenge areas ('arenas of development' is the term coined by Robinson) shown in the bottom part of the figure (and already referred in the explanations above) become the foci for targeted interventions and activities that 'oil the chain' of innovation.

Roadmaps help us to understand and plan what needs to be done. The knowledge needed to optimally support progress from *proof-of-principle* through to reliable, repeatable, sustained delivery of EBV data products to the *mass users* can be determined and then applied. Gaps identified in the necessary knowledge can be turned into recommendations for intervention. Such interventions may include, for example new research, capacity and capability building, and policy actions or market stimulations[32] to create a more favourable environment for the technology to emerge into.

12.2 Particular gaps in translation

Previously [Hardisty 2011, Elwyn 2012] demonstrated particular gaps in how new technologies translate from the research laboratory into healthcare settings, and how new technologies become embedded as a routine part of everyday working practice [May 2009, May 2013].

Research and development, policy-making, decision-making, service planning, delivery and action are as complex in environmental and biodiversity management settings as in healthcare settings. The design and exploitation of technology, particularly software technology is beneficial to both arenas. In both cases there is a mix of technical design issues, data availability and quality issues, legal and regulatory issues, economic issues; as well as the sociological and psychological factors at play. There is no reason to suppose that the two gaps in translation uncovered by [Hardisty 2011, Elwyn 2012] in one particular example of healthcare technology translation do not also exist elsewhere.

The first gap in translation occurs once a scientific proof-of-principle has been established. The gap is represented by the need to actively involve stakeholder actors (in the EBV case this is research scientists, policy-makers, decision-makers, conservation managers, etc.) in the design, production and exploratory use of prototype data products. We call this the 'setting of first use'. This setting is where the proof-of-principle is turned into something that can be properly used (trialled) for the first time. Such trials are often carefully controlled and limited in scope, e.g., in terms perhaps of EBV type, chosen species, selected data, area, time period, resolution, etc. The actors (users) in that setting of first use are known, knowledgeable and well supported, limiting the things that can go wrong. Nevertheless, there is some risk, arising mainly from lack of iterative collaboration to act on early feedback and adjust to build what is really needed.

The second gap in translation arises when moving beyond first use to extend, scale and embed the data products more widely. For example, embedding use of the data products into the live business processes operated by the various actors, or into business services offered by those actors as intermediaries to their own end users, for example environment agencies or conservation bodies. In this case, the specific end users are likely unknown and may not be supported. The risk at this stage is failure to properly understand the implementation work that actors carry out in order to adopt new data products into their routine business practices. This work involves elements of education, building communities of practice, activity to make the data product operational in their working practices and appraisal of the effect and impact of the new data

[32] As suggested during Workshop 1 in Leipzig, one such 'market stimulation', to be taken at the level of the G7+20 could be to establish national-level biodiversity bureaux or ecosystem services bureaux as a public good; essential assets for collecting and acting on biodiversity data. Compare these, for example with the role of weather bureaux in modern society. Another possibility is to leverage on assessments from IPBES.

product on work outcomes [May 2009]. Building confidence and creating trust are a large component of such work.

The same kinds of actors are implicated in both activities but they are often insufficiently involved at the earlier stages, leading to lack of understanding and hence lack of confidence and trust. Therefore, understanding how the data products are likely to be adopted and embedded into live business processes, service delivery and routine working practices is crucial knowledge for the earlier product development phase.

The gaps we have explained and hence the risks presented are intensified when there is insufficient understanding of the translational process by those involved in and affected by it. The process has to be orchestrated as a clear programme of work with accompanying interventions. In this sense it comes back to proper and widespread understanding of the innovation chain explained above and the challenge areas where attention has to be focussed.

13 Conclusions and recommendations

Research Infrastructures (RI) have to respond, either individually or as a coordinated community (e.g., through the GLOBIS-B project) to [Pereira 2013] on how RIs can support interoperable workflows for EBV measurement. They have to say, for example how they can:

- Support interoperable workflows for measuring essential biodiversity across the tree of life;
- Support the traceability of data through the workflows i.e., provenance;
- Maintain metadata; and,
- Demonstrate a multi-lateral cooperation among existing RIs.

It is clear that Biodiversity Observation Networks (BON) also have to be included in making such a response.

We are currently not advanced enough to provide definitive answers. Mapping the issues (section 10) and risks (section 11) against the various steps of the generalised workflow for EBVs (section 9.2) can help to identify the likely areas of difficulty and prioritise the issues and risks for further attention. Key factors affecting decisions rely on responsibility and governance for producing EBVs (section 8.1), warehouse choice (section 10.10) and EBV data product publication and storage (sections 8.3 and 10.6).

Decisions have to be made with political and social elements in mind and from a better developed understanding of the translational approach to follow (section 12). Like climate variables, a periodic cyclic production process for delivering EBV data products is most likely what is needed (section 10.1). This has to be verified among the stakeholders.

We need to **develop the technical strategy towards the future with finer details on a roadmap for the next 3 – 5 years**. To create that roadmap is the main recommendation arising from the present report. There are two main threads to such a roadmap, reflecting the informatics and business developments needed to bridge the two key gaps of the innovation chain (section 12.2).

Firstly, the experimental integration has to evolve from, develop from and illustrate what can be done in each infrastructure from a set of scientific and technical case studies. This process will identify key bottlenecks (scientific, technical, legal), addressing many of the specific technical issues highlighted.

Second, a generic solution and recommendations for solving these bottlenecks has to emerge, leading to specification and deployment of the integrated platform for production, necessary to make the transition to mass production of EBV data products.

Pre-requisites to the roadmap are:

- Having a better understanding of the likely overall ICT deployment scenario;
- Knowing how to organise the workflow to meet EBV data product requirements[33];
- Knowing which organisations and infrastructures are responsible for each aspect of EBV data products mass production; and
- Making the warehouse choice.

In parallel we should **aim to gain more practical experience by enabling and supporting RIs and data publishers to recognise the steps and the issues involved in supporting a particular EBV or set of EBVs and their current gaps**. This will enable some stakeholders to identify their support for some classes of EBVs. Identification of abilities and gaps will lay groundwork for standard approaches (e.g., to data quality tests and assertions) that infrastructures can converge to. It will support them from the bottom up as they consider their value proposition strategies (section 9.3.2).

14 References

[Andreessen 2011] Andreessen, M. (2011). Why software is eating the World. Wall Street Journal 20th August 2011.

[Atkinson 2016] Atkinson, M., Hardisty, A., Filgueira, R., Alexandru, C., Vermeulen, A., Jeffery, K., Loubrieu, T., Candela, L., Magagna, B., Martin, P., Chen, Y. and Hellström, M., A consistent characterisation of existing and planned Research Infrastructures, Technical report D5.1, ENVRIplus project, 203 pages, May 2016. http://www.envriplus.eu/wp-content/uploads/2016/06/A-consistent-characterisation-of-RIs.pdf.

[Baldwin 2011] Baldwin, CY., and Woodard, CJ. (2011). The architecture of platforms: A unified view. In: Gawer, A, (ed.) (2011). Platforms, markets and innovation. Edward Elgar Publishing. ISBN 978 1 84844 070 8.

[Belbin 2013] Belbin, L., Daly, J., Hirsch, T., Hobern, D. and LaSalle, J. (2013). A specialist's audit of aggregated occurrence records: An 'aggregators' response. ZooKeys 305: 67–76. doi: 10.3897/zookeys.305.5438.

[Bojinski 2014] Bojinski, S., Verstraete, M., Peterson, T. C., Richter, C., Simmons, A., & Zemp, M. (2014). The concept of essential climate variables in support of climate research, applications, and policy. Bulletin of the American Meteorological Society, 95(9), 1431-1443.

[Chaudhuri 1997] Chaudhuri, S., and Umeshwar, D. (1997). An overview of data warehousing and OLAP technology. ACM Sigmod record 26.1 (1997): 65-74. doi: 10.1145/248603.248616

[Elwyn 2012] Elwyn, G., Hardisty, A., Peirce, S., May, C., Evans, R., Robinson, D., Bolton, C., Yousef, Z., Conley, E., Rana, O., Gray, A., and Preece, A. (2012). Detecting deterioration in patients with chronic disease using telemonitoring: navigating the 'trough of disillusionment'. Journal of Evaluation in Clinical Practice Volume 18, Issue 4, August 2012, pages 896–903. doi: 10.1111/j.1365-2753.2011.01701.x

[EU Parliament 2007] EU Parliament. (2007). Directive 2007/2/EC of the European Parliament and of the Council of 14 March 2007 establishing an Infrastructure for Spatial Information in

[33] Currently we are mainly dealing with legacy data collected for other purposes.

the European Community (INSPIRE). Official Journal of the European Union, vol. 50, no. L108, April 2007. http://data.europa.eu/eli/dir/2007/2/oj.

[García 2014] García EA, Bellisari L, De Leo F, Hardisty A, Keuchkerian S, Konijn J, et al. (2014) Flock Together with CReATIVE-B: A Roadmap of Global Research Data Infrastructures Supporting Biodiversity and Ecosystem Science. http://orca.cf.ac.uk/88151/.

[GBIF 2016] Licensing milestone for data access in GBIF.org. 18th August 2016. http://www.gbif.org/newsroom/news/data-licensing-milestone. Accessed: 7th September 2016.

[Hardisty 2014] Hardisty, A., and Manset, D. (2014). CReATIVE-B Deliverable D3.2: Guidelines for Interoperability for Biodiversity and Ecosystem Research Infrastructures. [Technical Report]. Cardiff: Cardiff University, School of Computer Science and Informatics. http://orca.cf.ac.uk/92562/.

[Hardisty 2011] Hardisty, A., Peirce, S.C., Preece, A.D., Bolton, C.E., Conley, E.C., Gray, W.A., Rana, O.F., Yousef, Z., Elwyn, G. (2011). Bridging two translation gaps: a new informatics research agenda for telemonitoring of chronic disease. International Journal of Medical Informatics Volume 80, Issue 10, Pages 734–744. http://dx.doi.org/10.1016/j.ijmedinf.2011.07.002.

[Hobern 2013] Hobern, D., Apostolico, A., Arnaud, E., Bello, J.C., Canhos, D., Dubois, G., Field, D., Alonso Garcia, E., Hardisty, A., Harrison, J. and Heidorn, B. (2013) Global biodiversity informatics outlook: delivering biodiversity knowledge in the information age. Global Biodiversity Information Facility, Copenhagen. ISBN 87-92020-52-6. http://imsgbif.gbif.org/CMS_ORC/?doc_id=5353&download=1.

[Kissling 2015] Kissling, W.D., Hardisty, A., García, E.A., Santamaria, M., De Leo, F., Pesole, G., Freyhof, J., Manset, D., Wissel, S., Konijn, J. & Los, W. (2015) Towards global interoperability for supporting biodiversity research on essential biodiversity variables (EBVs). Biodiversity, 16, 99–107.

[Kline 1986] Kline, S. J. and Rosenberg, N. (1986). An overview of innovation. The positive sum strategy. R. Landau and N. Rosenberg. Washington, National Academy Press: 275-304.

[May 2009] May, C. and Finch, T. (2009). Implementation, embedding, and integration: an outline of Normalization Process Theory. Sociology 43 (3): 535-554.

[May 2013] May, C. (2013) Towards a general theory of implementation. Implementation Science 8(1):18.

[Osterwalder 2004] Osterwalder, A. (2004) The Business Model Ontology - A Proposition In A Design Science Approach. PhD thesis University of Lausanne.

[Pereira 2013] Pereira, H.M., Ferrier, S., Walters, M., Geller, G.N., Jongman, R.H.G., Scholes, R.J., Bruford, M.W., Brummitt, N., Butchart, S.H.M., Cardoso, A.C., Coops, N.C., Dulloo, E., Faith, D.P., Freyhof, J., Gregory, R.D., Heip, C., Höft, R., Hurtt, G., Jetz, W., Karp, D.S., McGeoch, M.A., Obura, D., Onoda, Y., Pettorelli, N., Reyers, B., Sayre, R., Scharlemann, J.P.W., Stuart, S.N., Turak, E., Walpole, M. & Wegmann, M. (2013) Essential Biodiversity Variables. Science, 339, 277-278.

[Robinson 2008]	Robinson, D. K. R. and Propp, T. (2008). "Multi-path mapping for alignment strategies in emerging science and technologies." Technol. Forecast. Soc. 75(4): 517-538.

Annex 1: Responses to pre-workshop questions Q5-Q8

Question 5

What are the key steps of a workflow(s) for calculating species distribution and/or abundance EBVs, starting from accessing the raw data to presenting a visual result? What are the complexities involved? What data preparation is needed?

ANSWER:

- Many possible approaches have been described in the literature, ranging from direct detection from RS through to SDM (see above).
- Main challenge is avoiding developing categorical data – stick to continuous EBVs

ANSWER:

- Raw data should be uploaded together with a metadata file mentioning the important characteristics of the data set (name of the monitoring program, type of data (occurrence or count), time duration of the monitoring period (start – end or dates of first and last records), indicative location (e.g. national or regional level), sampling protocols (number of species and taxonomic group, sampling design and frequency, observation techniques (binocular, satellite, microscope, etc), etc.) as well as the name and contact of the data provider (name of the institute or staff member).
- The data provider may also need to commit to data sharing agreements: either by agreeing to the direct access to / download of the raw data so that the uploaded data can be (re-)used and (re)analysed by any other user, or at least the use of the data set and the access to / download of the derived EBV calculations from his data set by other users.
- Importantly, data preparation to fit to a given format would be a pre-requisite. The format of the datasets should match with specific requirements detailed in some guidelines describing e.g. the minimum number of field categories and the information which are supposed to be recorded (species name, site ID, Latitude-Longitude, quantities (filled with 0/1 for distribution and integer above or equal 0 for abundance, "NA" for no sampling or gaps in the sampling design, etc), unit of the quantities (e.g. number of individuals, densities, catch per unit effort) as well as a nomenclature for naming each field categories so that each dataset should have the exact same name for each field categories. A thesaurus of species names and the geographical reference system (e.g. WGS84) should also be provided. Immediately after the upload of the data set, an automatic quality control could be performed to check whether the name of the field categories, their content, the name of all species in the data set, the geographical coordinates, etc. fit rigorously to the format guidelines. If any recommendation of the guidelines is not respected, the data set could not be considered for further analysis and would need to be modified accordingly. If the data set is respecting the guidelines, it could be allowed to go further.
- Then, the user would be proposed to perform some analysis by himself or not, whether his interest is to make use of his data or only provide them to others. Analysis could be performed through a simplified interface by choosing statistical methods and visualisation tools proposed to the user. Applying these methods and tools for calculating and visualizing EBVs would consist in linking the data set to pre-written scripts or software that would have be made available, selected and reviewed by experts. The user might also be offered the opportunity to access to other datasets already uploaded in the database for calculating and visualising EBVs from other species or places. The user could also download files (csv, text, etc.) with the outcomes of the analysis (EBV estimates, trends, maps, etc.).

ANSWER:

- Availability of a consistent suite of data

- Data Quality tests are the first and most important step to eliminate non-relevant data.
- A consistent methodology for surveying abundance and distributions.
- I leave the details of SDM and GDM steps to Jane Elith and Kristen Williams. It is no longer my expertise.

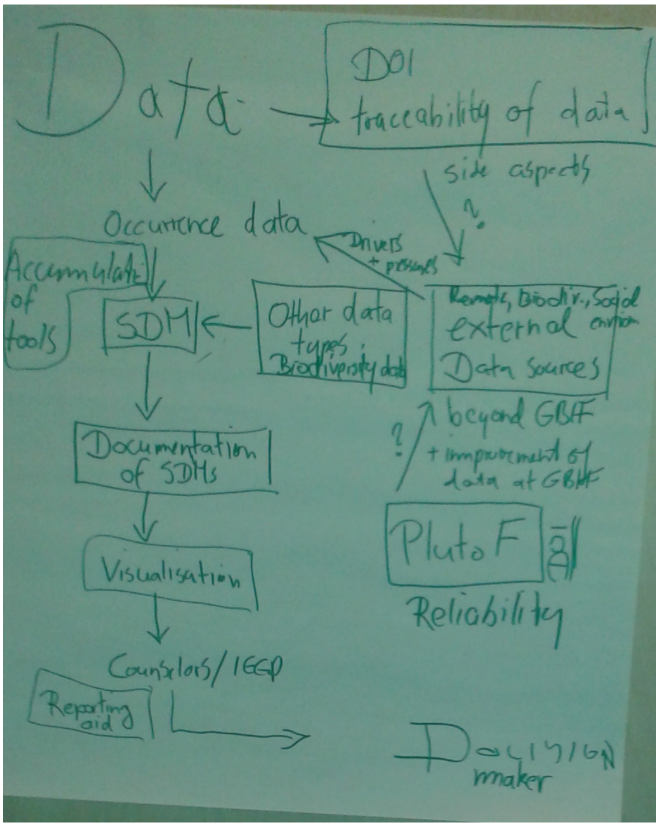
ANSWER:

- For all data, the level of supporting evidence for the observation/measurement/trait should be known or estimated and data should be incorporated based on an appropriate minimum confidence level.
- For all data, precision and accuracy should be understood for key dimensions (spatial, temporal, taxonomic, environmental) and data should be incorporated based on appropriate precision and accuracy for the model in question.

ANSWER:

- Establish a data documentation and a quality assurance plan for the workflow.
- Determinate source of data: research projects, involving other sector as health, hydrocarbons, defense, agronomy, etc; biological collections: citizen science. It means a cultural change for most and a capacity enhancement in this matters. Also data use licences must be clear in order to avoid legal issues.
- Integration: Once the data is published the mechanism for updating and indexing the data should be clean and fully interoperable, that's why data standardization is so important.
- Processing: data can be processed with the relevant variables for research. Here, the computational capacity and development of scripts are components can make this step the most automatized of the workflow.
- Analyse: Once the data is processed can be analysed for getting information. In this step informatic tools are useful but the expert judgment determinates what can be concluded from the analysis, whether if something can be concluded or if there is not enough information for.
- Visualization: All should be shown on the web, whether it can be presented as a map, table, charts of statistics or a publication. A content management system will facilitate that with some developments and design to enhance the way it can be given to the stakeholders.

ANSWER:



ANSWER:

- The key steps will be in the front-end of the workflow. Issues of integrating data by standardizing reference systems and spatial and temporal coverage are most time consuming and may be problematic in historic data where contextual information is lacking. I feel the next most important issue is traceability between the evolving concepts of EBVs, the data and algorithms used to calculate these metrics, and evidence for the validity of the approach in the science literature. I think formal references to scientific evidence, validated data and algorithms are increasing important when displaying the final metric to decision makers.

ANSWER:

- For molecular data (DNA barcode or metagenomic sequences):
- Step 1: development and implementation of a query system that integrates the information from different world wide available databases / infrastructure. The central searching criterion could be the name of the taxonomic class, possibly of the species, combined with other criteria such as the sampling geographical location or time;
- Step 2: development and implementation of a system to manage any data submitted by the user for the investigated species. There should be a temporary or definitive data and analysis results storage system;
- Step 3: Implementation in the infrastructure of tools and reference databases for taxonomic analysis.
- Step 4: Implementation in the infrastructure of tools for comparative statistical analysis of presence / abundance of species in different geographical areas and time points.

ANSWER:

- Accessing the relevant data at global scale is one of the biggest challenges and it is of highest importance to find a good balance between bottom-up and top-down approaches when facing this challenge. Top-down approaches such as GBIF implement an infrastructure and then ask countries/regions to share data to feed the system. These approaches are interesting because of their large spatial scale, but they may prevent from easily controlling for important issues such as uneven sampling effort, data storage/sharing and mobilization among countries (Beck et al., 2014). Bottom-up, network-of-network approaches such as the EuroBirdPortal (http://www.eurobirdportal.org/ebp/en/) that intend to connect different systems to each others are initiated by the countries themselves and have the potential to provide information associated with a lower level of spatial bias because sampling/recording effort and/or "sharing willingness" may be estimated in a more straightforward way. Reconciling the two approaches is a big challenge ahead.

ANSWER:

Depends on the data, but I will outline for presence/absence data (camera trap images or recordings):

- Import data from camera traps/recorder from a sampling season and ensure each image/recording has the following info: Project name, site name, date, time, spatial coordinates, species name, and other metadata (person identifying the image, sampling period name, sampling period dates, etc.).
- Basic data consistency check. For example, are all dates and times within the expected time frame? Are species names consistent across the data.
- For each species in the data set create a matrix of sampling points (rows) vs. time (columns) in days or any other meaningful time interval. Fill this matrix with 1,0, or NA depending on whether the species was seen at this point on this time (1), not seen (0) or the point was not sampled on this day. This matrix is the input of basic occupancy analysis. A matrix of number of sampling events of a species can also be created (defined as the number of times the species was sampled at a point at this time) for abundance analysis. User needs to determine what constitutes a sampling event (e.g. series of images that are at least 5 min apart in time). From this event matrix, point abundance analysis can be performed using binomial mixed models.
- If species observations are sparse (species detected at less than 5% of the points per sampling period) and detection probability is low (< 0.05), most models will have difficulty converging. In this case, combine species together, or do naïve analyses (without correcting for detection probability) without covariates.
- Choose a series of covariates that can be: spatial (value of covariate changes with space), temporal (value of covariate changes with time), both (value of covariate changes with space and time). These covariates can be used to model occupancy/abundance as well as detection probability.
- Temporal analysis will require fitting a dynamic occupancy or dynamic binomial mixture model (for abundance). Software already exists for this (Presence, package unmarked in R, or TEAM's Bayesian analysis using JAGS in R).
- Ensure that model recovers patterns adequately by checking model consistency.

ANSWER:

- Need a good computational infrastructure
- Need standardization of computational analysis pipelines

ANSWER:

- Getting and processing raw data are the key step. At present, raw data were collected by different agencies with different standards and different data quality, and distributed in different places and lack of data sharing.

ANSWER:

- This is general without thinking about whether it needs to be automated / rolled out globally:
- First think about what you are trying to achieve with the modelling and what sort of data that requires (don't start with the available data; think first). Monitoring change is much more challenging than a one-off species distribution model, and – despite published examples to the contrary – I don't view presence-only data as suitable.
- Collect relevant species observation data. If this is not your own data needs some time to understand it – to get to grips with the survey design, to understand survey effort, to get a feeling for whether species identification is reliable etc. Check whether the data meet the requirements for the sort of modelling that is appropriate. If predictions are to be made across landscapes, do the samples cover the main environmental gradients likely to be important to the species? Check for any errors in the data (terrestrial records in the sea; mismatch between textual descriptions and lat/long coordinates; records for riverine fishes on land; etc).
- Assess the coverage of the samples, of the environmental and geographic gradients in the region of interest – this is relevant for understanding whether predictions to unsampled sites are likely to be well informed. Ref: Cawsey, E.M., Austin, M.P. & Baker, B.L. (2002) Regional vegetation mapping in Australia: a case study in the practical use of statistical modelling. Biodiversity and Conservation, 11, 2239-2274.
- Gather covariate data, for both the observation process (detection) and the state process (occupancy/abundance) – want variables relevant to the species at a grain that represents those environments properly (e.g. aspect is irrelevant at coarser grain but may be important to the temperatures experienced by the species). Evaluate correlations between covariates and decide whether to reduce the covariate set, and how.
- Fit a model (what method is going to be used for model selection? – big issue), test its fit and predictive ability (the latter e.g. with cross-validation). At this stage need to decide what minimum predictive ability is acceptable for monitoring change.
- If required, predict occupancy / abundance over whole region.
- Repeat in time 2. Question: do you use same set of candidate covariates, or same set of covariates selected in the final model at time 1 (intuitively: the first)
- Calculate change. Include uncertainty throughout.

ANSWER:

The main complexities are:

- Data accessibility, which includes people not wanting to share their data.
- Data harmonization, which is related to data and metadata standards

ANSWER:

- Selecting a species or group of species from a taxonomy.
- Fetching, providing or getting automatic suggestions (e.g. possible misspellings) for alternative names for the species.
- Retrieving what you call "raw data" from each possible data source using all names.
- Manually and/or automatically filtering retrieved data based on data quality, geographical, and/or temporal parameters.
- Calculating the EBV.
- Storing results, preferably with all data used.
- Organizing and presenting results.

- There can be lots of complexities depending on the details of the workflow. For example, EBV calculation may be as simple as calculating the extent of occurrence based on point data, but it may also involve complex ecological niche modelling techniques with pre and post processing steps. Data quality filters can be numerous. Additionally, many steps could involve human interaction, which could seriously affect workflow efficiency. And if the level of interaction in certain steps is too high, it may be necessary to either create intermediary databases/services or change the original data repositories to accept some specific tagging mechanism through web services, so that further workflow runs do not require duplicate work to review the same data.

ANSWER:

- Identification of the taxonomic, spatial and temporal scales where best available biodiversity data are enough to make reliable distribution and abundance predictions. Taxon and locations specific verifications and calibrations.

ANSWER:

- I think that the workflow should begin with protocols for data collection and verification. At this stage, it will be important to collect the metadata required to determine legal and policy interoperability (perhaps a generic standard like Dublin Core could be expanded by drawing on other standards and frameworks, including the European Interoperability Framework (http://ec.europa.eu/isa/documents/isa_annex_ii_eif_en.pdf). Of course, all raw data sets without sufficient metadata will need to be re-visited, and new metadata added, before these data sets can be used.
- Having accurate metadata to document things like data policies, the amount and type of PII included in a data set (if any), and the legal jurisdiction of data collection is a first step towards supporting interoperability. Automated matchmaking could be sufficient to determine key aspects of legal interoperability, for example by using information including national provenance, data/database structure, and licensing to determine whether two data sets are compatible from an intellectual property perspective.
- But, automated metadata matchmaking between data sets is not a complete solution. For some data sets, including those collected by citizen scientists for a specific purpose, the initial goals of data collection may be compatible with some forms of re-use but not others. This could be thought of as a form of political or ethical interoperability, and isn't always given the same weight as legal and policy concerns. There should be some way for the creators of a data set to indicate (upon uploading) their preferences for re-use (is automated matchmaking OK, or does the system need to send the request to a human researcher for review?).
- While all necessary steps of the workflow should be uncovered through the design process (see Q6), a few features that may appeal to citizen scientists include the ability to save an in-progress or completed analysis/ visualization, the ability to export a visualization with references to source data and metadata, the ability to work collaboratively, and the ability to ask questions of experts or others through a discussion feature.

ANSWER:

- Data preparation:
  - Collect the raw data (potentially from a range of sources)
  - Identify duplicate data
  - Quality-check the spatial and temporal reference if present, filter data which cannot be pinned down in time and space.

- Look for obvious outliers or possible misidentifications using modelled species ranges, and handle these according to a consistent strategy.
- Investigate and interpret any flags which may indicate breeding / migratory / etc status.
- Ensure that taxonomic definitions, units etc match or can be transformed between the different datasets.
- Optional: If working with abundances, transform observations to inferred abundance where sampling effort / strategy is known.
- Perform all necessary transformations and harmonise the data.

- Compute species distributions or abundance values / maps for specific time steps. A particular complexity here is identifying where an absence of records actually indicates absence or loss of the species.
- Combine the values / maps to get an idea of CHANGE between the epochs. Here, a complication is that the uncertainty at both steps may drown out any actual change signal.
- Generate clear, usable maps / tables / charts (ideally accessible via the web, with links to full clear metadata on how the computation was carried out, and links to the source data). Raw results should also be made available (along with metadata) so that users can present the results in their own chosen way.

ANSWER:

- Change of distribution: Download data from GBIF. Clean it by merging synonyms and duplicate records. Choose meaningful variables, run principal component analysis (PCA), choose the right calibration area, and calibrate the ecological niche model. The main difficulty is to get modern environmental data layers, because WorldClim is now rather outdated. There are data gaps, fo instance in Eastern Europe. Data management is a challenge in general. OpenModeller does not offer PCA, etc. Lots of technical hurdles and dealing with data gaps.
- Change in abundance: Download data from GBIF. Clean it as above. Cross-tabulate it into a OLAP hypercube, aiming at reasonable data density at each cell. Compute trends in various spatiotemporal resolutions. The complexity is to maintain performance when using hundreds of millions of records. Data cleaning at that scale is challenge, because it can only be done automatically.

ANSWER:

- In Sparta, we start with a set of records from multiple species that we believe to be recorded as an assemblage, by which we mean that records of one species can be used to infer absences of others (see Isaac & Pocock 2015 for a UK-centric exposition of the data issues). Sparta interfaces with a package called rnbn that collects data from the British National Biodiversity Network (the UK node of GBIF) and it would be trivial to link it directly with rGBIF. The challenge is to identify the assemblage, i.e. which set of records can be considered to be co-recorded. The rest of the workflow is well-defined in Sparta, but involve converting these records into a 'visit matrix' where each row is a unique combination of site and date. Our estimate of sampling intensity is the list length (the number of species recorded, following Szabo et al 2010). We use 1km^2 and day precision, but coarser resolutions are quite possible. Each species' detection history is appended to this matrix: the modelling thereafter is described in the scientific literature (e.g. van Strien et al 2013, Powney et al 2015). There are issues of spatial coverage and spatial autocorrelation that we have yet to master.

ANSWER:

- Knowing the fitness of the data for a specific question is essential. Select the data based on proper documentation. Raw data can only be re-used for other applications if properly standardized and quality controlled. For EurOBIS and OBIS we use taxon matching tools, spatial end environmental outlier detection in an automated system to assign quality flags to the records.

- Relevant publication: Vandepitte et al. 2015 Database

ANSWER:

- Drawing on experience from Essential Climate Variables (Bojinski et al 2014) it is clear the process and methodology for generating EBV data products is multi-stage.
    - Assembling the relevant raw data is the first step. In some cases this can be quite straightforward; retrieving relevant occurrence records from GBIF, for example. In other cases, less so; perhaps requiring additional processing and transformation of satellite images (for example) or establishment of new observing protocols.
    - A second step could be concerned with adjusting the assembled data to account for heterogeneities in it; perhaps in the observing methods or to fill gaps where data is absent.
    - Depending on the nature of the EBV, a modelling and correlation step may come next, leading to the EBV product itself.
    - Thereafter, comes post-production quality assurance - a check necessary to ensure the uniformity of the product when compared to the same product calculated for a different place or at a different time or using different data.
    - All of this has to be carried out in a standard, repeatable, open and transparent manner with clear and accessible documentation of each step, such that it can be subject to expert scrutiny and peer review.
    - And finally, the EBV product needs to be updated regularly, perhaps in near real-time so that the information can be used as the basis for monitoring change.

ANSWER:

- Key steps:
    - Knowledge of scientific question, sampling design, sampling procedure
    - Data management, including protocols, standards, etc.
    - Data standardization/normalization
    - EBV calculation method assessment and knowledge of its statistical properties
    - Identification of the appropriate workflow
    - Identification of the statistical package
    - Identification of the web service which provides access to the package
    - Assessment of the computational limitations of the offered web service
    - Data uploading and massaging
    - Execution of the EBV calculation

  - Visualization of results

- Complexities:

  - Infinite number of complexities starting from the design and sampling procedure to EBVs calculation. Unless every single step of the above steps of the process is crystal clear, bias may come at any stage and may have considerable effect on our calculation and estimation.

ANSWER:

- I'm assuming this question addresses what to do after the raw data has been collected and isn't asking about how to go about doing the monitoring. The workflow depends on the data being used and the intended output. I'll describe one example for abundance data.
- Raw abundance data are logged in order to make trends comparable between populations and discount the size of the population units (if appropriate). This is essential if different types of population counts have used e.g. sightings per km of transect versus biomass.
- Use an appropriate model to process the time series data. This will depend on the type of data and whether it is structured or unstructured. Generalised additive models can be used for longer time series, linear models for shorter ones (Collen et al 2009; Buckland et al 2005). Other models are used when using structured survey data (Gregory et al 2005).
- Use ancillary information attached to the population data to disaggregate the results in meaningful ways and answer more specific questions.
- Geometric mean is the most common method of combining a multi species abundance indicator. Usually each species is weighted equally but this can be altered if there is an imbalance in the data set.
- Complexities
  - Addressing bias in the data. Weighting can be applied to address under-representation of species or regions.

  - Other issues of representation include how much of a global species population should be monitored in light of the fact that we will always be limited in how much we can monitor.

  - Other issues when monitoring abundance particularly for migratory species can be in understanding if a population has moved or shifted its range rather than genuinely decreased in numbers.

  - Understanding what makes up a population – is it defined in an ecological sense or does it refer to the geographical extent of a study site and all the individuals of a species located there.

ANSWER:

- See answer to Q8 (the first 2 priorities). Certainly reliable metadata is the key issue.

- In more abstract and general terms, my answer is simple. I would test whether the data management life cycle analysis of GEO (if they are adaptable to biodiversity data –which I am not so sure about) and others are well articulated or they are still wishful thinking. Too many people have been thinking about it not to take their outputs seriously.

ANSWER:

- Species distribution:

    - Name resolution and reconciliation
    - Data (occurrence) harvesting from across resources based on taxon concept queries
    - Fit for purpose evaluation
    - Data cleaning algorithms (confidence level thresholds)
    - Data aggregation (occurrence records)
    - Projection over multi-layered environments
    - Ad-hoc geo-correlation services

ANSWER:

- Dealing with data from different data sources the integration and analysis mainly focuses on two complexities: a) the changing taxonomy and b) changing methods in the estimation of abundance and distribution. Information on the underlying methods applied and the related uncertainties in the quantification needs to be taken into account for the calculation. The estimation of the overall uncertainty is one of the important issues.
- The main issue is to get data on a national scale for threatened species. While monitoring for certain species works well e.g. also using tracking mechanisms (e.g. whales or birds), for others it is more complicated to get consistent figures.
- Issues to be addressed:
    - Taxonomic reference
    - Method comparison for estimation of abundance and distribution
    - Estimation of the single and overall uncertainty
    - Dealing with data gaps (both in temporal as well as in spatial terms)
    - Data availability
- The issue on using provenance and quality information in automated workflows is an issue for the integration of information from different sources.

ANSWER:

- The workflow will most likely be ad-hoc depending on the specific question being addressed.
- Any workflow will include at least the following steps:
    - It is necessary to facilitate the access to (meta-) data reliable and identified resources;
    - Data gap analysis to identify potential holes affecting distributions (see e.g. http://www.gbif.org/resource/82566);
    - Data cleaning (see e.g. http:// http://www.gbif.org/resource/80528) to detect suitable, fit-for-use data

    (http://www.gbif.org/resource/80623,http://www.unav.es/unzyec/papers/ztp720_postprint.pdf)

    - To provide the proper e-Tools to also guarantee their (semantic) inter-operability;
    - Niche modelling (http://press.princeton.edu/titles/9641.html);
    - Some type of visualization ranging from the most basic (e.g. GIS) to complex, network/relationship graphs and their further integration into proper Virtual Research Environments-VRE (see also Question 6).

- An effective approach based on COMMON tools as simple as possible (for example Python-based) but flexible to allow the users detailed analysis, could be encouraged.

ANSWER:

- Data prep, having the source data evaluated to see if it is fit for purpose (https://www.youtube.com/watch?feature=player_embedded&v=eVYwt86mC_4Q and Complexity is in agreeing how to calculate the measure of the EBV mathematically so that it can be implemented in software.
- Once have that algorithmically would suggest that the data is processed into a multi-dimensional cube (OLAP – see https://en.wikipedia.org/wiki/OLAP_cube) to enable interactive dashboard at multiple scales.

ANSWER:

- There are likely to be many variations of a particular workflow, depending on the question being asked. The general approach needs to be scientifically well-established but also sufficiently flexible.
- The existing BioVeL environment implements many of these steps already:
    - Accessing, preparing and cleaning the data to remove erroneous records are fundamental steps. Conveying changes in editing existing data back to the data providers is also very important. For example, we routinely utilise GBIF data but often edit existing co-ordinates more accurately or add co-ordinates where there is detailed locality information but no geo-reference.
    - For species distributions, ecological niche modelling is a widely-used tool for individual species, or by making species richness surfaces by stacking model outputs. Models run against different time-stamped datasets can reveal changes in a species distribution EBV for multiple species, if compared against an appropriate baseline.
- Again, the interesting questions are not so much changes in particular EBVs over time but in inferring mechanisms and predicting future patterns by integrating approaches for different EBVs into shared indicators based on a common mathematical framework.

ANSWER:

- I think the main issue here is the broad deployment of the Extended Darwin Core with Event Core. This extended core allows us to manipulate the structured data coming from systematic monitoring efforts, something that was not possible to do with the original Darwin Core. The next step will be to combine data from various sources, including opportunistic and systematic sampling, to generate species populations and distribution EBVs. This may also include the use of proxies such as land-cover and climate to expand from the sample points to a continuous surface (wall to wall monitoring).

Question 6

What is a suitable technical (ICT) approach to perform this workflow(s) for calculating EBVs (any place, any time, using data anywhere, by anyone)? What special considerations have to be taken into account?

ANSWER:

- The workflow for calculating EBVs would ideally be accessible at any time by anyone. For this purpose, any data provider should commit to a data sharing agreement (see Question 5). Any user, either providing data or not, would need to commit to another agreement for accessing, downloading or using data sets available in the database. This user agreement would specify e.g. non-commercial use of the data, EBV calculations or any outcomes arising from the use of the database. Besides, the user should commit to cite all the data sources that would have been used, analysed or downloaded whenever the outcomes will be published or communicated (citation of the name of each data the provider and the name of the scheme from who data as been).

ANSWER:

- Monitoring abundance and distribution EBVs is unlikely to be properly determined, unless specific project teams are appropriately qualified.

- Obviously a standard workflow is required, with each step justified as efficient and robust (Best Current Practice).

- While anyone, anywhere could use the approach, it is unlikely that a totally ‘canned’ approach could be optimal given current state of knowledge.

ANSWER:

- This question is a distraction. Depending on the repeatability and necessary parameterisation of the models in question, ICT solutions may be based on workflow engines (where this a suitable rapid exploration/prototyping approach) or on robustly implemented algorithms. Parallel processing technologies like Hadoop may be important. However, all of this is frankly relatively trivial once we determine what we hope to model and how those models relate to the source data.

ANSWER:

- Workflow can be performed in different conditions, the technologies can be adapted to a well-defined software architecture. Establishing an appropriate solution architecture to perform this workflow means that we can address the principal concerns according to the requirements. That also means that requirement gathering should be an exercise that must be done with the best judgement.

ANSWER:

- Data must come from several sources and should be able to be dynamically interlink different databases and do corrections, pointing out discrepancies etc. We need to mobilize all data and not focus only on open access data. This brings copyright issues.

ANSWER:

- The data, contextual information and algorithms used in workflows will be widely distributed and heterogeneous in nature. There will need to be agreement on a number of standards to enable these resources to be brought together. I would expect these to include wrapping these resources as web services with standard vocabularies to describe them when interrogated. How these services are then marshalled into a workflow can be carried out in any workflow engine that adhere to the service standards

ANSWER:

- Implementation of analysis systems in a Workflow Management System such as Galaxy or Taverna.
- The systems should be usable even by non-experts and include a user-friendly interface.

ANSWER:

- TEAM offers an analytical engine to perform occupancy analysis of camera trap data with or without covariates (wpi.teamnetwork.org) The analysis can also be performed by running R scripts for pre-processing and model fitting. In the near future we will have the capability of doing abundance analyses as well.

ANSWER:

- The Creative-B document “D3.1 Comparison of technical basis of biodiversity e-infrastructures” documents the conceptual architecture (figure 26) for how applications, service logic, and resources can be stacked and interfaced. Workflows are incorporated into the framework as part of the service logic. What’s missing from that conceptualization, which would be useful for the computation of EBVs, is to associate technical standards or implementation technologies that are “GLOBIS-B recommended”. For example, the “Data Resource” component needs to have associated with it standards and technologies (for metadata, for catalog services, for discovery services, for access services, etc) that is not just a laundry list of standards and technologies, but a constrained, vetted, set: too much and wide of a set of options, and it becomes useless. Technology implementers / system integrators appreciate a “sanctioned”, controlled suite of options given to them, with perhaps a reference architecture of how one such combination of standards and technologies is used to implement a solution.
- I feel that developing solutions at the conceptual level (like the conceptual architecture above, but supplemented with a small suite of “sanctioned” standards and technologies) coupled with a documented instance of one such implementation of the concept would encourage other EBV projects to try to adopt a solution that would be interoperable with each other, at least at various spots in the different implementations.
- An example of a constrained set of options for standards and technologies has very recently (late 2015) been proposed by the US Group on Earth Observations, US GEO, which is the US body to the worldwide GEO. The data management subcommittee of US GEO has a draft version of the “Common Framework for Earth-Observation Data” (https://www.whitehouse.gov/blog/2015/12/09/improving-access-earth-observations), which should be finalized sometime in 2016.

ANSWER:

- I don’t think any workflow for this purpose is available, now. Quality and standard of row data have to be taken into account.

ANSWER:

- (unsure about the meaning of this question). “by anyone” ? I doubt that it’s safe/possible to automate this to the extent that someone with no modelling experience could do it. Maxent (the species distribution modelling software) is a good example of something made available to non-experts that is then used with poor choices by many users (because it's hard for newcomers to understand the nuances of the choices, and most people go for defaults and don't understand the implications of their choices). Needs a reasonable level of commitment for someone to develop the expertise to run the models properly. I believe the models needed for change detection are a step more difficult and need trained people to run them.

- Presumably some steps in data prep can be automated. Tools can be developed to report on available covariates. Modelling: E-bird (USA) is a good example of sophisticated modelling analyses applied to vast quantities of data. Has taken a team with considerable statistical and computing skills to set it up and to continually evaluate model output (with input from species experts). Hasn't been left for others to run it (i.e. the analytical team is always there).

ANSWER:

- Assuming that the whole process needs to be replicable by anyone, the workflow would need to interact with publicly accessible data repositories and not depend on specific hidden/private data from certain researchers/institutions. Time and geographic range could easily be workflow parameters if the underlying raw data contain these dimensions. Things can get trickier if you also want to handle incomplete raw data, such as occurrence data without coordinates, having only a description of a place in natural language. Another critical step is to handle uncertainty. For example, occurrence records with high spatial uncertainty would probably need to be discarded in local/regional scale calculations, but could still be suitable for continental/global scale calculations.

ANSWER:

- I think this question can only be answered through the process of cooperative (or at least user-centered) design. The GLOBIS-B team could begin by brainstorming a set of personas representing relevant stakeholders, including scientists, policymakers, and different types of citizen science volunteers. Then, users from each stakeholder group could be recruited to inform the design and development of an EBV calculation platform.

ANSWER:

- Data preparation:
    - Collection of raw data requires the user to be able to easily discover all the possible sources of data related to a particular taxonomic group. To some degree this is possible through catalogue searches but is still challenging.
    - Processing and formatting the data requires that it is originally available in an easily-transformed digital format where each dataset has at least some common tags, fields etc.
    - QA requires the user to have a clear idea of what constitutes a reasonable or impossible value/observation.
    - Transformation of datasets may be computationally intensive.
    - Most importantly, even if there are shared open-source libraries (e.g., in python and R) for performing the above tasks, there will always be an element of user parameterization and tweaking, meaning that potentially huge amounts of effort could be expended to produce EBVs from the same datasets which are inconsistent. The ideal would be that aggregation, quality checking, harmonization, catalogue harvesting / metadata publication etc were all carried out

before the user gets to the data. The agency which comes closest to doing this job at the moment is GBIF.

- Computation of EBVs including gap-filling / inference / interpolation. If uncertainty in terms of detection / misidentification is to be quantified, some Monte Carlo simulations / random permutations of the data would be necessary. If positional accuracy is likely to be a problem, this should also be acknowledged and the impact of problematic observations assessed.

- Computing change between time steps – probably the simplest step – plain maths or map algebra, though if uncertainty is taken into account there are more calculations necessary to calculate lower / upper bounds or quantiles.

- Presenting visual results – recommend open interoperable web services for maps / standard formatted data (see below for how this can be done)..

- What special considerations have to be taken into account?

    - Technical capacity of users, necessary investment in training / effort, access to data (or unit tests) for verification, testing and validation – accessibility of software (i.e., open source / freeware vs. corporate). Legal / copyright restrictions on component data, and whether these percolate through to derived products. Necessity to aggregate or obfuscate sensitive records. One big question – how will 'any user, anywhere' be able to get good advice on when the available data is too sparse or inaccurate for use in their chosen context?

ANSWER:

- Computations need to be performed in portals using OLAP.

ANSWER:

- We rely heavily on cluster computing. Some of our datasets are reaching the limits of available RAM limitation? We are also finding that many invertebrate groups lack sufficient data to work at 1km^2 and date precision.

ANSWER:

- [Note: Don't understand why everyone should be able to calculate EBV's. It does require some skill's..]

- Virtual labs that make data and algorithms available through webservices offers many possibilities, this is the approach taken by Lifewatch and Biovel, Biodiversity catalogue.

- Overview of virtual labs for the marine world: http://marine.lifewatch.eu/

- Statistical packages can easily harvest data from webservice, we built several interfaces based on R, Rstudio, Rshiny.

- A good, scalable infrastructure using OGC standard webservices (WMS/WFS/WCS/WPS) is geoserver. EMODNET makes all data products available as OGC compliant webservices.

- The different data sources can be queried simultaneously.

    - http://www.emodnet.eu/dataservices/

ANSWER:

- Behind the above explanation lies a significant issue that has to be addressed early-on by scientists and the potential end-users of EBV products. It concerns a fundamental choice between calculating an EBV data product on-demand on-the-fly, versus a more periodic systematic production cycle where EBV data products are produced, updated and extended, for example annually, quarterly or monthly.
    - Simplistically, on-demand, on-the-fly production requires ready access to relevant raw data, and to the workflow and processing capacity to transform this raw data to the selected EBV product for the indicated place or area of interest (local, regional, national) at the timestamp of interest. Processing capacity "at the touch of a button" is necessary to service the instantaneous demand of the request (and of simultaneous requests). Size and complexity of requests is not known in advance (although this can be controlled by limiting geographical area and resolution). Repeatability is a key requirement, such that if the EBV is again requested on-demand for the same place and time, the same answer has to be delivered. EBV data production is ad-hoc, responding to demands of the moment, with the quality assurance checks in-built in the procedure. Archiving of the EBV products is not required.
    - In the cyclical approach, EBV data production is systematic, aggregated over large areas (potentially, the whole globe) and archived as an ever extending database(s) of information to be queried to provide the data for the indicated place or area of interest (local, regional, national) at the timestamp of interest. Processing capacity can be estimated in advance. Periodicity of the production cycle for an EBV can be tuned to the available processing capacity and to the expected temporal sensitivity of that EBV. The information is generated once, archived and then available forever (or a set period of time) to be used and re-used as needed. The any time, any place requirement is met not by on-demand computation but by querying previously computed data products that have undergone a post-production quality assurance assessment.

ANSWER:

- First part of the question not entirely understood
- Special considerations: unlimited computational capacity; transparency (leads to adequate repeatability of any observation and analysis)

ANSWER:

- Several techniques exist to process abundance data, for example the software package TRIM, the method behind the Living Planet Index (Collen et al 2009) both of which could probably be developed into an online platform for use by anyone. The former is used for abundance data using a standardised monitoring protocol for a set of species e.g. birds, butterflies. The latter approach can incorporate abundance data from any species, method and unit of data

ANSWER:

- Combination of existing "official" data management systems (e.g. digitized national biodiversity inventories) with quality controlled citizens´ science based apps, and adequate VREs familiar to biodiversity science actors. I have not at all enough informed knowledge to describe them (with the exception of very specific marine species estimations). What I am sure of is that a cluster of very sophisticated policies concerning all the life-cycle data flows and reuses is an unavoidable development that necessarily has to be in place.

ANSWER:

- The implementation of web processing services (WPS) with defined input interfaces including quality information will be a possible solution. The implementation of these standardised WPS could be done on different platforms.

- Implementation of data services for species distribution and abundance data including a semantic taxonomy mapping tool. Enhancing service based availability of data on species is a pre-requisite for the modelling approaches. As earlier stated information on the quality and uncertainty of certain methods needs to be provided with the data. Here is a limitation in the current data services which either focus on spatial data services (e.g. species distribution maps) or sensor based observations (e.g. a single species observation). Further development on the services for these kind of data is needed.

ANSWER:

- This has already been approached with occurrence data:

    - See the GBIF portal (http://www.gbif.org) and developments built around it: for example WALLACE (http://protea.eeb.uconn.edu:3838/wallace2/). In general, web services and REST able to milk large databases and produce subsets of data already condensed according to criteria supplied by the user will be the preferred method, as most scientists or practitioners are likely to prefer experimentation on the data (as opposed to final products such as ready-made niche models)

    - See also the LifeWatch Marine Virtual Research Environment (Virtual Lab) developments (lifewatch.eu) which common construction blocks are also being used for the implementation of the LifeWatch Freshwater Virtual Research Environment

- In general terms, these developments could be integrated (before being adapted accordingly) in LifeWatch ICT distributed e-Infrastructure, as the European Reference Platform (ESFRI) in order to further compose some e-Services to offer these EBVs values in a visual way through the development of in turn proper Virtual Research Environments-VRE. All this process involves analysing in-detail the final users ("customers": researchers, decision makers-environmental managers) requirements.

- Therefore, all of this should be performed through the design, establishment, deployment and maintenance of an OPENESS and Big Data paradigms-based Conceptual Framework such as LifeWatch e-Infrastructure is offered at the disposal to this purpose.

ANSWER:

- Use data warehousing techniques through the ETL (Extract Transform Load) process to aggregate the data and build the OLAP cube.

ANSWER:

- Again, many of the individual steps in carrying out such a workflow have been tackled already, with several running sequentially e.g. in the BioVeL environment. Having such workflows open source, or making use of repositories of pre-written code, so that analyses can be adapted to particular research questions is an important factor in their successful implementation.

ANSWER:

- As stated above, a combination of structured data in Darwin Extended Core, and some statistical inference (e.g. correction for sampling bias, trend detection), will be the first targets. Use of SDM's or habitat suitability models with remote sensing of proxy variables (e.g. land cover) may also be used to expand the data from the monitoring points to continuous surfaces.

Question 7

What are the technical options available and what is possible to achieve today or within the next 12 months? What data and/or workflows, software etc. are available today? Where is it and how can it be used?

ANSWER:

- Global analysis using the freely available RS is central – postage stamp approaches of joining many local studies result in inconsistent output

ANSWER:

- Data available: see examples in Question 1.
- Software / methods available for calculating or visualizing EBVs: TRIM (software), PRESENCE (occupancy software for distribution EBVs) or R-script of occupancy models, n-mixture models or visualisation tools that could be made available from publications or any expert contributor. Q-GIS or GRASS are open source software that can support the mapping and the visualisation of the EBVs.

ANSWER:

- There are Data Publishers that have a good foundation of distribution data, e.g., GBIF, ALA, CRIA, BISON, SANBI etc.
- Methods such as MaxEnt and GDM are well known and robust.
- Workflows for SDM are widely available, e.g., R, BCCvL, BioVel,
- Methods for estimating abundance are well established.

ANSWER:

- For data, GBIF is probably the most complete occurrence data pool and we should all work together to aggregate all possible data on occurrence and sample events in one place, and collaborate in data quality improvements to the whole.
- Significant existing GIS and remote-sensed environmental datasets exist and should be made accessible through a consistent discovery and access catalogue.

ANSWER:

- There are many tools and technologies that can speed up the development of this workflow:
    - Standards as PlinianCore and DarwinCore.
    - Publishing tools, such GBIF IPT.
    - Queue messages technologies: Apache Kaftka, Amazon SQS.
    - Indexing technologies: Solr, Elastic Search
    - Powerful relational and non-relational databases: PostgreSQL, MongoDB, Hadoop.

  - Map technologies: Mapbox, CartoDB.
  - Stats visualization: Kibana.
  - And several data quality tools: http://community.gbif.org/pg/pages/view/39746/list-of-data-quality-related-tools-in-the-gbif-catalogue

ANSWER:

- We are working a lot with OGC EF and O&M standards at present to bring together the description of the origins, configuration and accessibility of environmental monitoring data. This is being used to deliver SOS web services. There are reference implementation (e.g. 52N) for these standards which many groups are working with. I would like to explore how these standards could be applied to biodiversity (e.g. from GBIF etc) to not only deliver species data but the environmental context around them. There would then be many ways to assemble these services into workflow from simple python scripts to Taverna style systems.

ANSWER:

- Data:
  - As concerns DNA-barcoding data, a lot of resources are available online, such as GenBank or BOLD. In BOLD each barcode sequence is associated with a well curated taxonomic description, with the collection site and date, the organism picture, etc.
  - As concerns metagenomic data, among the most used reference databases there are RDP, GreenGenes, Silva, ITSoneDB e PR2/HMaDB. Previous metagenomic project sequences can be explored from various online archives such as EBI metagenomics, MeganDB, iMicrobe, etc
- Workflows, software:
  - For taxonomic assignment of DNA-barcoding sequences some of the online available phylogenetic pipeline are SAP e RaxML. Also in the BOLD site the taxonomic assignment is available but it requires a preliminary registration and not more than 100 sequences can be analysed each time.
  - For taxonomic assignment of metagenomic datasets some of the online available pipeline are BioMaS, QIIME, Mothur, MetaShot, Kraken, Sparta.
  - Metagenassist, DESeq2, Phyloseq and Metagenomeseq packages are among the commonly used package for the statistical and comparative analysis.

ANSWER:

- We can do occupancy analysis of camera trap data on a massive scale now using TEAM's wildlife picture analytics system. This will calculate population trends and combine these on a flexible biodiversity index (the wildlife picture index). Within the next months we could also accommodate other sources of data (acoustic) and perform abundance based analysis.

ANSWER:

- The analysis pipeline can be standardized in the next 12 months

ANSWER:

- I don't think any workflow for this purpose is available, now.

ANSWER:

- Analysts currently run these models using specialised software packages, with R (the free statistical software), etc.

- A relevant issue: there is currently quite a push towards "reproducible science" – e.g. https://zoonproject.wordpress.com/ . The ability to trace analyses could be excellent.

ANSWER:

- First of all you need to choose what EBVs should be calculated and how (in many cases the same EBV can be calculated in different ways). You could start by listing the possible EBVs, assigning each one a rank of "importance/impact" and a rank of associated data availability, then list the possible ways to calculate them, assigning each way a level of complexity to finally decide what can be done in the given timeframe. Scientists will also need to define which kind of data will be used. For instance, if the whole GBIF database will be used, you may consider a specific partnership with them to build the new application directly on top of their database. On the other side, if only specific parts of it will be used, you may create a separate application, still with significant web service interaction to retrieve data and a local database to store results. An interface on top of that database could be used to display results. There are many possibilities for implementation, including workflow management tools and other software frameworks – it's hard to tell at this point what could be the best options.

ANSWER:

- In addition to time stamped species occurrence and static environmental data, it would be good to involve species interaction, physiological adaption and dynamic habitat loss & quality layers – if there are models that are ready to consume such data

ANSWER:

- Within the next 12 months it is possible to a) write the specifications for an EBV platform by working with different user groups, and b) in parallel, conduct a survey of major existing software used by biodiversity experts and technical experts.

ANSWER:

- As stated above, GBIF is the agency currently performing many of the identified data preparation tasks. Software libraries and tools exist for most of the steps (commercial GIS, Quantum GIS, PostGIS spatial queries, R / python / Matlab libraries… but the question is whether it makes sense for this data preparation effort to be duplicated, or whether it would be possible to set up a toolbox / framework for this workflow which could be shared. If so, python could be a useful language since it has many statistical, data manipulation and spatial libraries, and can optionally interact with ArcMap / QGIS where those are installed on a user's machine. Technically, I think that at least a prototype workflow for data preparation could be produced in the next 12 months, though the scraping and discovery of all relevant input data is a big challenge.

- Computation - Open-source libraries such as Sparta (https://github.com/BiologicalRecordsCentre/sparta) may be useful for this process. For identification and handling of problematic positional referencing, see e.g. http://onlinelibrary.wiley.com/doi/10.1111/j.1600-0587.2013.00205.x/abstract.

- Presenting visual results – many accessible and interoperable ICT solutions are available, e.g., OGC WMS / WFS / WCS (tools like Geonode, CartoDB and Mapbox have lowered the entry barrier), and for raw / tabular results, REST services which return JSON or other easily usable formats that don't require corporate software to visualize / chart.

ANSWER:

- The Swedish LifeWatch Analysis Portal https://www.analysisportal.se/ is a good example of what we need. It just needs to be scaled up to any country and (sub-)continental, and global scales. They do not speak of EBVs, although they are computing something similar. EU BON is working on this.

ANSWER:

- We are beginning to explore supercomputing options, including through Microsoft Azure.

ANSWER:

- The choice - on-demand production or periodic production - is fundamental because of its implications for the way that production processes are defined, and for how infrastructures are organised and optimised for calculating, archiving and serving EBVs data. The choice has to be feasible, efficient, and affordable. Global cooperation is needed to ensure consistency, serving comparable raw data sets and processing capabilities for production and maintaining appropriate archives. The workflows for producing EBVs data have to be capable of being executed in any infrastructure, and from anywhere in the world. The choice raises issues for permissions to use primary data, for secondary data, for citation and attribution and for provenance tracking.

- Now and within 12 months, on-demand calculation is possible using the BioVeL infrastructure and, for example an adapted generic ENM workflow.

    - generic ENM workflow on BioVeL portal: https://portal.biovel.eu/workflows/440

    - myExperiment: http://www.myexperiment.org/workflows/3355.html

    - Documentation: https://wiki.biovel.eu/x/ooSk

ANSWER:

- Technical options:

    - Mach of the infrastructure already in place: e-infrastructures, such as LifeWatch, BioVel, iMarine, ViBRANT, that could serve as building blocks of the infrastructure required

- Next twelve month target:

    - Registry of the e-infrastructures in place

    - List their technical specs and features

    - Deliver a plan by which the services they provide can be interoperable

- Data and/or workflows:
    - Most of data needed are cite above
    - Workflows and statistical software operational in the context of several e-Infrastructures: BioVel, LifeWatch, ViBRANT, iMarine, Aquamaps, etc.

ANSWER:

- I could only provide a wild guess. I´d rather decline to answer. I have the feeling though that open reusable data (with appropriate metadata) is simply not available yet. Certainly the departure point are some already existing systems such as GBIF or OBIS

ANSWER:

- Currently the main sources of information will be GBIF on the one side and data from the European FFH directive on the other. An issue with the estimation on population status and trends is that the underlying data are not provided together with the estimation. In addition due to lacking information a part of estimation are based on expert judgement backed-up by in-situ data.

- In the short term focuses (next 12 month), the integration of information on status and trends as a compilation of estimations will be the most suitable procedure for a wide range of species. For some already (e.g. certain whale and bird species) estimation models on a global scale exist.

- Options: Implementation of WFS/WCS services for species data observation. Extension of SOS services for species observations (issue of complex monitoring schema).

- The availability of data to estimate species population including changes in time seems to be a greater issue than the technical limitations. Nevertheless the automatic integration of uncertainties along the chain of methods is still an issue to be solved.

ANSWER:

- Taking up Question 6 story line, GBIF is possibly the most advanced, already-available portal and is seeding a growing community of developers that are producing services based on its API. Also, niche-calculating services are very useful as entry points. Other global or theme-specific datasets (e.g. OBIS) are equally important, although they are generally becoming increasingly interoperable.

- To this regard, an integration of some of these relevant developments in LifeWatch ICT distributed e-Infrastructure should be feasible within the next 12 months' time period.

ANSWER:

- The issue is more the source data, evaluating if it is fit for purpose and expressing the measure algorithmically. Any number of off-the-shelf software providers and open source solutions exist for data warehousing.

ANSWER:

- See above (e.g. BioVeL). For species distributions most of the steps are already in place, save perhaps explicitly visualising these results in the context of change in a particular EBV. Analyses of change in species abundance are more constrained by available data; abundance data is more spatially and taxonomically biased, the analyses more complex and varied, but also potentially more informative of the causes of genuine changes in population abundance.

ANSWER:

- There are a lot of challenges in doing anything in 12 months. I think the main goal should be mobilizing population abundance and atlas datasets into the Darwin Extended Core, and deploy a couple of apps that can perform statistical analyses on those

Question 8

What are the top 3-5 technical challenges of supporting interoperable EBV calculations on a global basis? How can these be addressed and in what time period? Who has to do something?

ANSWER:

- Funding
- Suitable high resolution imagery (hyperspectral, lidar, hypertemporal)
- Link between policy and space agencies

ANSWER:

- Some of the main challenges would be:
  - To define the data format guidelines in a way that they can be applied to any kind of monitoring (standardized survey and / or opportunistic data) and any taxonomic groups (see Question 5)
  - Once the EBV abundance and distribution would be clearly defined, to agree on robust and suitable statistical methods for calculating both distribution and abundance EBVs with respect to the requested data format and the heterogeneity of the monitoring.
  - To define critical steps of the workflow as well ad key linkages in order to implement it and make it operational.
- How it can be addressed:
  - Organising workshops
  - Engaging participative contribution of data owners / statisticians / biodiversity experts / technical IT experts
  - Learning and getting inspired by previous successful endeavours (e.g. see the excellent publication by Barker et al 2015 detailing a very efficient workflow of large scale ecological data)
  - Barker et al. 2015. Ecological Monitoring Through Harmonizing Existing Data: Lessons from the Boreal Avian Modelling Project. Wildlife Society Bulletin 9999:1–8; 2015; DOI: 10.1002/wsb.567
- Who has to do something:
  - Statistical and biodiversity experts need to define robust and suitable methods for EBVs calculation together, as well as providing software or scripts.
  - Technical experts of database management / IT experts need to support the implementation of the workflow.
  - Biodiversity experts and technical experts need to work together for defining the guidelines and the data sharing / data use agreements.

- Time period: Within the next 2-5 years seems reasonable.

ANSWER:

- What species? They can't be consistent internationally.

- Lack of systematic data. Lack of consistent, internationally agreed systematic surveys of key/indicator/target species.

- While not technical, it is the 'political' that is likely to be the most limiting factor in achieving effective abundance and distribution change monitoring. Long-term, ongoing funding is required for regular surveys.

ANSWER:

- The patchiness and sparseness of data especially across continental scales.

- Lack of clarity around priority species for organising and delivering EBVs.

- Absence of consistent modelling approaches for delivering at least best-available EBV data (or even a clear and consistent vision for a global modelled EBV component in e.g. GEOSS)

- Lack of clarity around GEO BON's place in delivering the modelled data layer to sit between e.g. GBIF on the one side and IPBES and onwards (to CBD, etc.) on the other.

ANSWER:

- By default the computational capacity is a challenge, but it can be overcome easily. Data quality issues are very important to address in order to have more trustful results, that involves georeferences for historical data that can be hard to determinate. In my opinion, the most difficult challenges are actually social, since the culture of making data open is not always well received, so assertive approximation to data holders will be key for enrich the system content.

ANSWER:

- Skill and training – without the people with the skills and knowledge to deal with interoperability issues at a global level, this cannot happen – Research funding bodies need to address this (see Belmont forum report http://www.bfe-inf.org/document/community-edition-place-stand-e-infrastructures-and-data-management-global-change-research)

- Standards development and adoption – In order to provide globally interoperable resources requires agreement on standards. The internet is the obvious place to start this and many science communities now use this as the basis of their research collaboration. This will have to run through to agreement on vocabularies and ontologies to link information together – There are existing standards for some of this but there needs to be some mechanism / governance to drive adoption of these within day-to-day work.

- Data Policy – there clearly needs to openness in the availability of data (and algorithms) so support workflow operations. These legal frameworks exist but are not always enforce as the conflict with researchers expectations of "ownership" of the data they have created. This is still a big cultural issue in that needs to be addressed at research agency level within legal jurisdictions and by scientific rewards within scientific journals.

- Long-term funding for informatics R & D and systems operation – researchers and decision making will not trust systems that have uncertain lifespans. We cannot convince researchers to entrust data to systems which have no long-term funding. If decision support systems are seen as important for dealing with environment challenges, they must be seen as part of national / international infrastructure. The stability required to establish these systems and ways of working cannot be achieved through a series of 3 to 5 years research grants. There needs to a rolling funding and review of what is essential infrastructure for development of essential biodiversity indicators – international agreement required between funding bodies(??) – see Belmont or RDA??

ANSWER:

- Data format: the informatics format of data to be correlated can be highly variable. First of all, unique formats must be selected for each type of data. Then the infrastructure should be designed to manage and integrate these formats.

- Access to data: it would be necessary to define if access to data and tools available in the infrastructure is unrestricted, restricted to certain users or under a simple registration.

- The data transfer protocols must be safe: for example, the user who submits his data does not want they become public.

- A storage system should be implemented and the following questions should be addressed: which data to keep? How long?

- It could be useful to investigate other bioinformatic infrastructure (Elixir, Lifewatch, BioVeL, etc.) and platforms already available for storage and analysis of molecular biodiversity data.

ANSWER:

- Ensure that data collected under different protocols is standardized in some way and weighted accordingly

- Platforms to share databases in a federated way, so that data can be easily accessible on one site (for example see wildlifeinsights.org for camera trap data).

- Automation of analyses is a challenge; model building and construction requires some degree of user input unless a narrow class of models and covariates is run.

- Willingness of researchers, institutions and governments to share data on a global scale.

ANSWER:

- As part of ongoing quality assessments, even if one has an automated workflow to ingest data from a variety of sources, there should be a way to compute aggregated indicators of data quality for a given set of data sources. Suppose a workflow ingests from four three data streams:

    - Dataset 1: Subset of data retrieved with constraints A from repository X

    - Dataset 2: Subset of data retrieved with constraints B from repository X

    - Dataset 3: Subset of data retrieved with constraints A from repository Y

- Some suite of quality indicators should be computable against all datasets, to produce information like:

	Quality indicator 1	Quality indicator 2	Quality indicator 3
Dataset 1			
Dataset 2			
Dataset 3			

- The quality indicator could be of the type {High, Medium, Low} or a quantitative score. It may be a good idea to run audits of those quality indicators on some regular, or irregular, schedule, if you're running a service to compute EBVs for policy use. This will enable some level of certification of the quality of the EBV for policy use, which may ease any concerns about decision-making based on computed EBVs that use data from various sources.

- There is a recently NSF funded project called MetaDIG (lead by staff from DataONE and the US National Center for Ecological Analysis and Synthesis) looking into developing quality indicators for metadata and data. Their aim is to develop a suite of quality indicators that various communities of practice (e.g. the US Geological Survey, the Long-term Ecological Research network) can use.

- Provenance metadata standard. I am not sure how widespread is the community acceptance of the W3C PROV standard for provenance capture, but it is my hope that a body like GLOBIS-B plays a part in examining the applicability of the provenance schema that PROV recommends, and determines whether it is suitable for the computation of EBVs. I feel that like with quality indicators, making sure that workflows are accompanied by provenance metadata will be essential at some point down the road. This is especially true given the importance of reproducibility, which has been an issue discussed within the Belmont Forum e-infrastructure and data management cooperative research action.

ANSWER:

- Implementation of the workflow(s) in a distributed environment, assign tasks on nodes (servers/hubs) and link them together, 2. Efficiency of the workflow, 3. Updating of workflow and nodes, 4. Visualization for results

- If institutions in biodiversity informatics in the world work together, this job can be done in 36 to 60 months.

- The most important thing is find enough fund and the project be well designed.

ANSWER:

- substantial work assessing available species data and whether it can be massaged into a form suitable for change modelling (e.g. into a form that allows detection to be estimated)

- current online data often treat as presence-only data that are actually from structured surveys (i.e. absences aren't recorded; it can be hard to identify all sites from the one survey; information on survey methods can be lost). Improve this?

- geographical biases in collections data are already well known (e.g. Amano, T. & Sutherland, W.J. (2013) Proceedings of the Royal Society B 280. What are the priorities for monitoring change? Are priority regions the most poorly sampled? If so, can surveys be designed to satisfy several needs at once (so decisions needing data NOW are served, as well as longer term aims for monitoring change)

- environmental predictors – are they adequate, and at fine enough resolution to be useful for monitoring? – same for detection covariates.

- Modelling: how to manage the tradeoff between wanting it rigorous enough to enable reliable estimates of change, yet somehow widely available?

ANSWER:

- Living Planet Index from WWF-ZSL and Map of Life from Yale, but from both initiatives the underlying data that is used to derive the indices and models is not available.

ANSWER:

- Again, this would depend on the EBVs that need to be calculated and how they will be calculated, but potential challenges include:
    - Handling large volumes of data (fetching them remotely and processing them).
    - Designing for efficient human interaction, if this will be needed.
    - Depending on changes to be made on third-party systems (such as asking other initiatives to create new web services on top of their data or make other adjustments on their systems so that they can be integrated with GLOBIS-B).

ANSWER:

- Identifying and supporting key non-technical aspects of interoperability, including semantic, legal, policy, and political or ethical considerations (timeframe: 1-3 years; could be accomplished by a handful of workshops followed by period of comment and consultation).

- Finding a balance between automated matchmaking and matchmaking that requires human input. This challenge will be continually revisited through the EBV and platform design process (timeframe: 3 years+, depending on funding).

- Developing and documenting a system that is truly accessible to a range of stakeholder audiences, including professional researchers, amateur researchers, educators, and policymakers. Achieving this will require a clear statement of goals and purpose advanced by the GLOBIS-B team and collaborators followed by an inclusive design process (timeframe: 3-5 years+, depending on funding).

- Making sure that this project reaches the widest possible audiences, within and beyond the biodiversity and larger scientific community (3-5 years+).

ANSWER:

- Patchiness of the data: distinguishing gaps from absences. To tackle this, directed and systematic sampling is needed. High-quality citizen science projects have some potential but are of restricted value in geographically/politically inaccessible areas.

- Getting non-digitised / archived data into GBIF – already underway with new task force, and some well-designed citizen science projects based around naturalists' notebooks. Ensuring that these new observations also feed improved range modelling. Research councils and individual scientists also may be able to support this effort.

- Reproducibility and robustness of the EBV calculations - ensuring that they scale correctly when computed at smaller scales, that results are consistent and will be trusted by decision makers.

ANSWER:

- For distribution modelling, getting environmental data layers beyond WorldClim.

- For abundance, download and cleaning of full GBIF data into an OLAP cube.

ANSWER:

- SGDR (sui generis database right) - It is fundamental to keep in mind the distinction between data creation and data collection: only data collection (or presentation/verification) can lead to the existence of SGDR (if all the other requirements are met). If data are created there is no SGDR. To make things "easier" there is the unclear definition of data creation and data collection (a difference that not necessarily corresponds to the scientific/epistemological definition).

- Importance of correct labelling of data and metadata - It is mandatory that all data/dataset are properly labelled with the right tools: Public Domain Mark, CC0, CCPL.

- TDM, copyright, SGDR and licences: it is important to employ licences that address properly these considerations (e.g. CCPL v4.0 yes; CCPL v3.0 depends but usually no; CCPL v2.0 no).

- In order to licence data properly it is important that not only the right legal tools be available (to some extent they already are, e.g. licences) but that also the right set of incentives be available (i.e. if in order to obtain grants or get tenure researchers need to have high Impact Factors, then they will publish in high IF journals that not necessarily follow OA principles). Therefore, researchers cannot be "left alone" in dealing with copyright/assessment issues, but they need protective legislative interventions (like the German and Dutch, not like the Spanish or Italian; the UK solution is debatable) + the right set of incentives from funding bodies and employers (e.g. only papers/datasets self-archived in OA, aka green road, will be used for evaluation purposes).

- OA to be successful requires a new approach not only in the publication of science, but also in its evaluation/assessment.

ANSWER:

- Our contribution to EBV are from two angles, copyright and data sharing policies on the one hand, and form the published record.

- What we can contribute is to look at data, data quality and how this relates to open access to the data

- From the published record this only makes sense in two specific aspects: Publishing data sets so that they can be cited, e.g. using either GBIF or Pensoft publishing facilities, which is relevant in the longer term to set up monitoring schemes.

- Another aspect of the published record is that is often the only source for rare species beyond butterflies, or plants. This might add a special layer of taxa that represent a large part of biodiversity and are in most cases completely underrepresented. At the same time, the question might be raise, whether the known data is strong enough to contribute more than anecdotal evidence to EBV.

- For us the experience to participate in GLOBIS-B is that we are very interested to find out weak points in data, EBV workflow and data publishing, and how we can improve future data publishing.

ANSWER:

- High quality data from charismatic organisms in rich countries often not comparable with sparse data from elsewhere. We need to avoid the lowest common denominator. I see this essentially as a modelling problem, rather than a data availability problem.
- Metadata (see above): more sophisticated observation models will be computationally intensive.
- Multispecies models will be even more computationally-demanding.
- Spatial scale, spatial resolution and temporal resolution of the outputs.

ANSWER:

- Data generation is still the limiting factor: we need to measure faster, cheaper, automated.
- Lifewatch Belgium devotes a large part of budget to install biosensor networks for the measurement of phytoplankton, zooplankton, fish, bird, bats.
- Some examples: http://rshiny.lifewatch.be/
- The Jerico Next and Atlantos projects are examples at European and TransAtlantic scale.

ANSWER:

- There are multiple technical challenges but the main challenge lies in getting research infrastructures operators to work together at the global level to pursue an agreed roadmap (e.g., based on that coming from the CReATIVE-B project) that ensures that the various research infrastructures are interlinked and interoperable in both technical and legal terms. This requires investment funding. The responsibility should be taken up by the Belmont Forum, perhaps?

ANSWER:

- Challenges:
  - Secure unlimited computational capacity
  - Ensure transparency of the process
  - Provide web services by which viewing of data, using of data and workflow/software will be tracked and reported back to the developers
  - Mapping of EBVs at global scales
- Ways to address tech challenges:
  - Engaging grid and cloud infrastructure
  - Develop tools for the traceability of the entire process on the cyberspace
  - Develop the pipeline links between the data and workflows/software available, as well as with the available e-infrastructures
- Who has to do?

  - Scientific community from around the world- mobilizing the large Networks: e.g. MARS, WAMS, etc. for marine benthic biodiversity, there are many communities for other regimes
  - ICT community relevant with the biodiversity informatics
  - EU and other funding agencies at national, regional, continental and global scale, to create the appropriate funding instruments, at least for the coordination of the activities.

ANSWER:

- Agreed metadata standards (including rights statement metadata, which do not exist at this moment) for all the existing datasets under answer to Q1.
- Agreed standard to incorporate abundance-based data occurrences (Is there anyone on place with enough consensus and reliability?).
- Agreed GIS standards to facilitate the charts expressions of Q2 (and open source based)
- Assuming that there is minimum agreement on answer to the previous 7 Qs (which is a background minimal need, at least for some species or taxa) the main need I assume is conducting a real life testing in which digitized national inventories, GBIF data sets, and biodiversity species-related data mining of scientific publications and citizens´ science crowdsourced data, using multiple (or at least double) VRE based ITs to control reliability of results. It would need clear policy agreements and funding.

ANSWER:

- Data mobilisation across multiple sources (incl. legacy literature and collections)
- Use of common or interchangeable Standards that will allow data interoperability
- Open and well-documented web services to serve data
- Robust registries of data services and Standards
- Development of end-user services tailored to specific audiences

ANSWER:

- Technical challenges:
  - Data description including the uncertainties in a machine readable manner. One of the issues is the provision of this information which is not primarily a technical issue
  - Taxonomic references and mapping of species names and species groups – should be already be solved in the GBIF context, but still in the FFH directive it is an issue
  - Provision of time series of species observations including abundance information – with the issue on how to upscale from regional data to a global perspective
- The establishment of consistent monitoring schemes for biodiversity on national scale are an important pre-requisite for further activities. Methodological the use of high resolution EO data for habitat and species estimation needs to be evaluated. E.g. EcoPotential will focus on the identification of whales in one of the test areas based on EO Sentinel data.

ANSWER:

- The first and foremost challenge is to have a global database of occurrences that include abundance data. That's a major step from the current, presence-only datasets that make the bulk of globally available biodiversity data. A challenge that is currently being addressed is incorporating sample data. For this to work properly, there are still unresolved challenges that might be on track during the timeframe:
  - An effective system for unique identifiers (GUIDS) for biodiversity occurrences/objects;
  - An agreed-upon, proven standard to incorporate abundance-based and sample-based data to occurrence datasets;
  - a reliable way to represent/identify/describe absence data;
  - A clean, authoritative taxonomic backbone allowing easy identification of taxon concepts, duplications, synonyms, and overall deduplication of occurrence data. Some of these challenges, as well as many others, were identified by a large number of practitioners through a content needs assessment carried out a few years ago (see https://journals.ku.edu/index.php/jbi/article/view/4126 and https://journals.ku.edu/index.php/jbi/article/view/4094).
- Therefore, and in order to achieve these goals, a proper Organizational Knowledge Management Methodology (OKM) should be established and then refined-maintained by an OKM Committee. The OKM should be based on the following premises:
  - How to identify some practical cases from the perspective of relevant biotics and abiotics EBV indicators to be performed. This would largely depend on the "quality" of the (meta-)data resources above mentioned. This analysis should be performed by a Scientific Committee.
  - To this purpose, to further integrate-adapt existing Workflows developments into the LifeWatch ICT distributed e-Infrastructure, so that some essential "blocks" given in the form of e-Services can be offered in order to calculate EBVs values and then presented in a visual way through the development of proper Virtual Research Environments-VRE. This should be coordinated by a ICT Technical Committee.
- Therefore, not only we are talking about the creation and maintenance of a EBVs "ontology-driven" system, but also of how to guarantee the "engineering" mechanisms associated to their integration into the LifeWatch (and similar) e-Infrastructures from the ICT perspective.

ANSWER:

- It is obviously necessary to have a common data vocabulary, common standard and protocol for data exchange but this also depends on which step of the processing is done by whom. There are at least two broad alternatives and each of these have their own challenges for global assessment.
  - EBVs are calculated separately for each jurisdiction, region continent and then aggregated or reported globally .
    - Advantages of this is that much of the burden is distributed among countries/ jurisdictions do little needs to be done centrally. This would put greater onus on countries to coordinate national, monitoring, assessment and reporting of biodiversity

        and ensure a stronger link between biodiversity monitoring and management actions, policy and legislation

        - Challenges: To ensure consistency in calculation, data quality standards etc. among countries. Also some countries jurisdiction will simply not have the staff and resources to do these analyses so some of this will have to be done centrally

    - EBVs are calculated globally using data obtained from each jurisdiction

        - Advantages: transparency and consistency of calculation

        - Challenges: major resources may be required to chase up, acquire data. Any errors may not be easily recognised because the data will be processes by people who have limited knowledge of the data

ANSWER:

- Assuming questions on the definition of EBVs do not need to be further addressed, there needs to be a broad recognition that most biodiversity is currently un/under-represented by available data. The main technical challenges would then be that:

    - Accurate and widespread recording of abundance data with confirmed absences, rather than just presence-only data. This would sensibly build on the existing GBIF architecture.

    - The taxonomic backbone needs to be improved and made explicit i.e. synonymies made clear.

    - Over time, changes to existing point data sets (e.g. GBIF) need to be a) recorded and b) explicitly presented i.e. new specimen records, new geo-referencing of specimen localities, edits to the taxonomy and location details of existing records such as re-determinations and more precise geo-referencing. For plant specimens, duplicate records of the same collection from different institutions need to be explicitly linked; if geo-referencing is undertaken retrospectively this may differ between duplicates of the same collection.

    - An established but flexible workflow for species distribution modelling and stacking species extents would be imperative.

    - One of the outstanding conceptual challenges for the development of EBVs is agreement on common scales/units/indices of measurement to allow data on e.g. species distributions to be integrated sensibly with data on e.g. habitat extent, and in a way comparable for e.g. allelic diversity with e.g. habitat extent. Combining different EBVs in a standardised way is the real power of the whole conceptual approach. Alternative metrics such as effective numbers may be worth exploring.

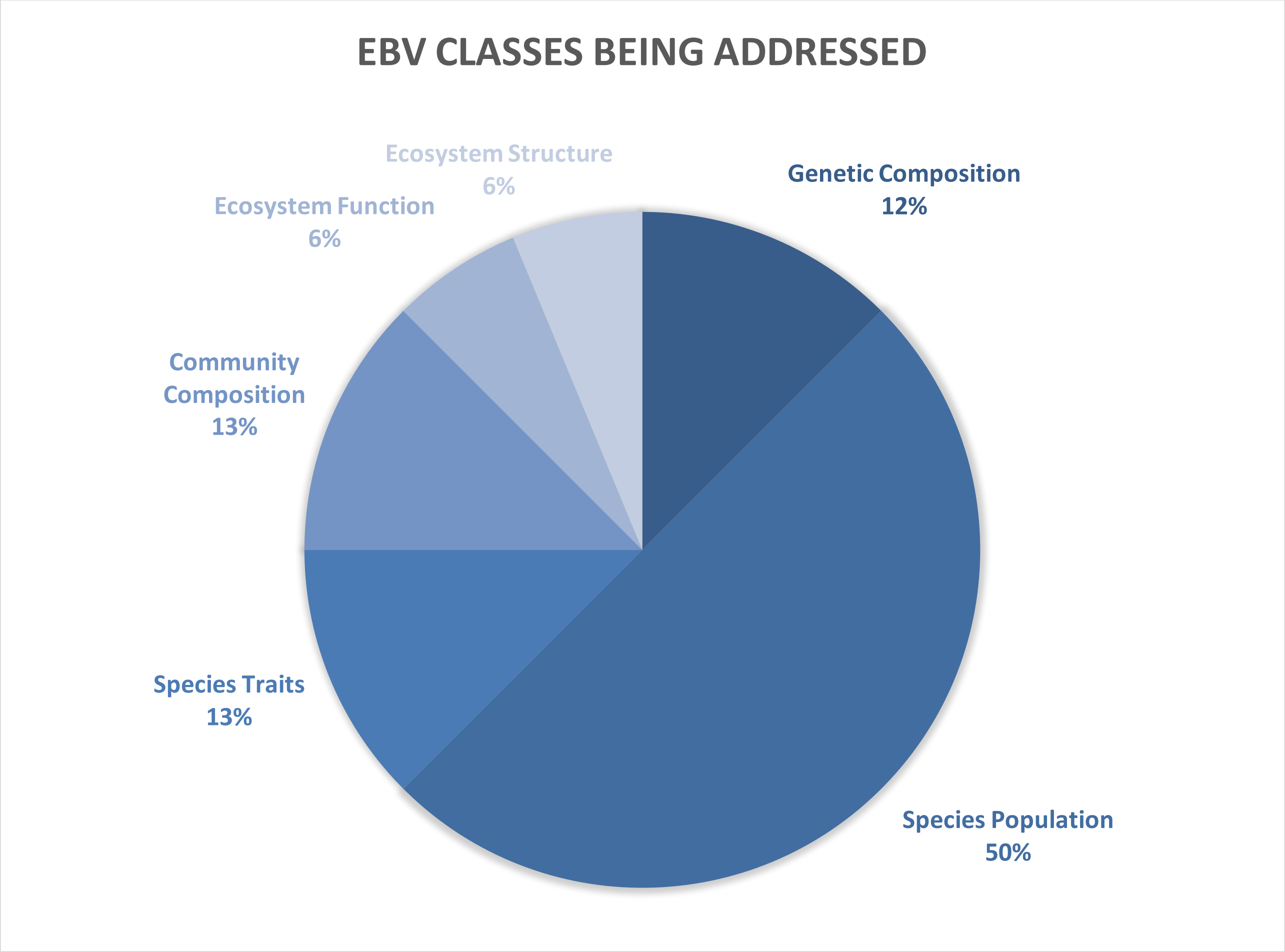
ANSWER:

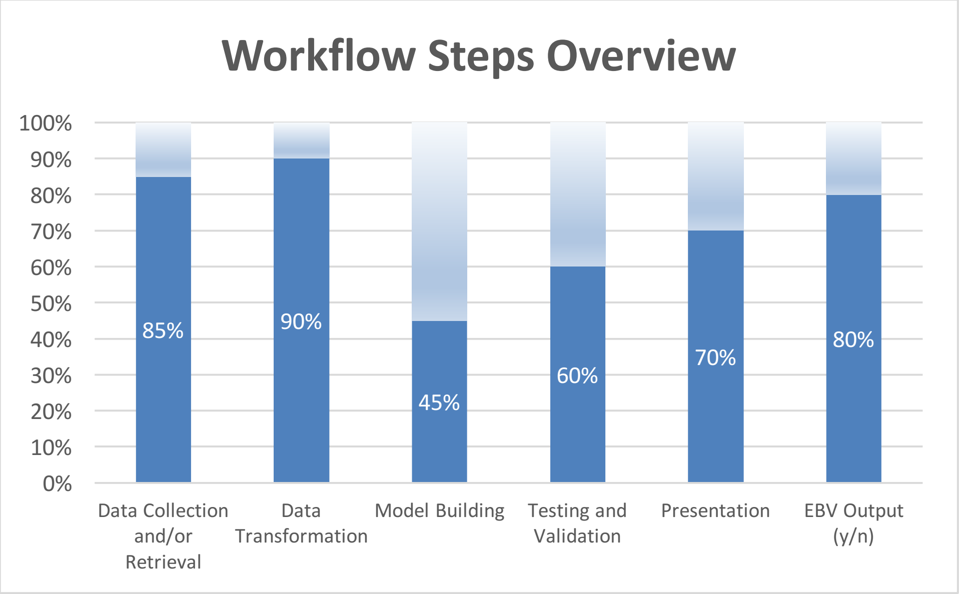
- The main problem is collecting the data and publishing the data openly. At least we could make a lot of inroads on the later.

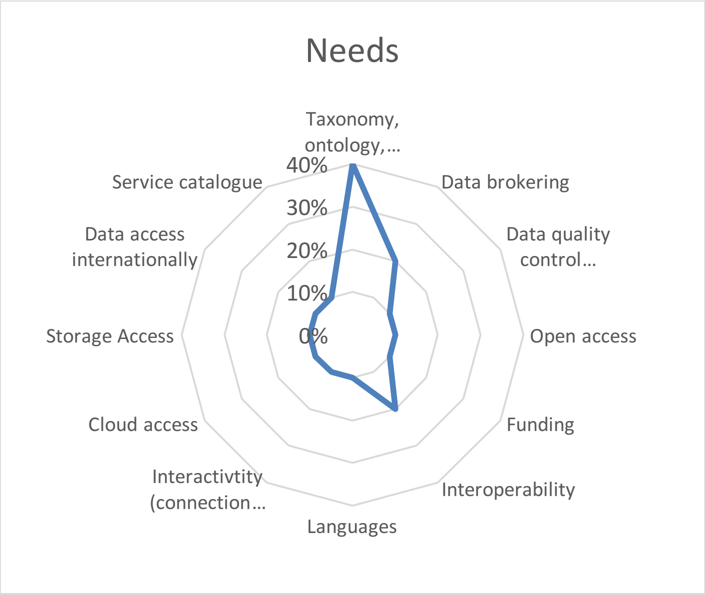
END.

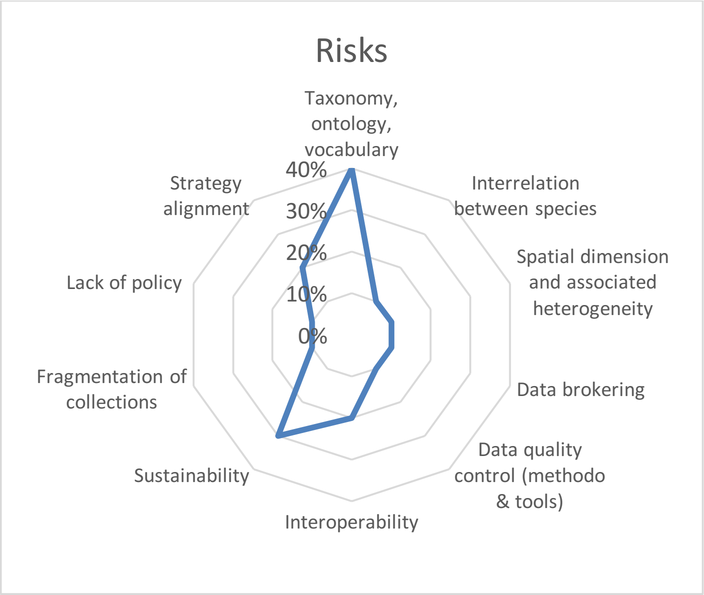
Annex 2: Figure 3 expanded for readability

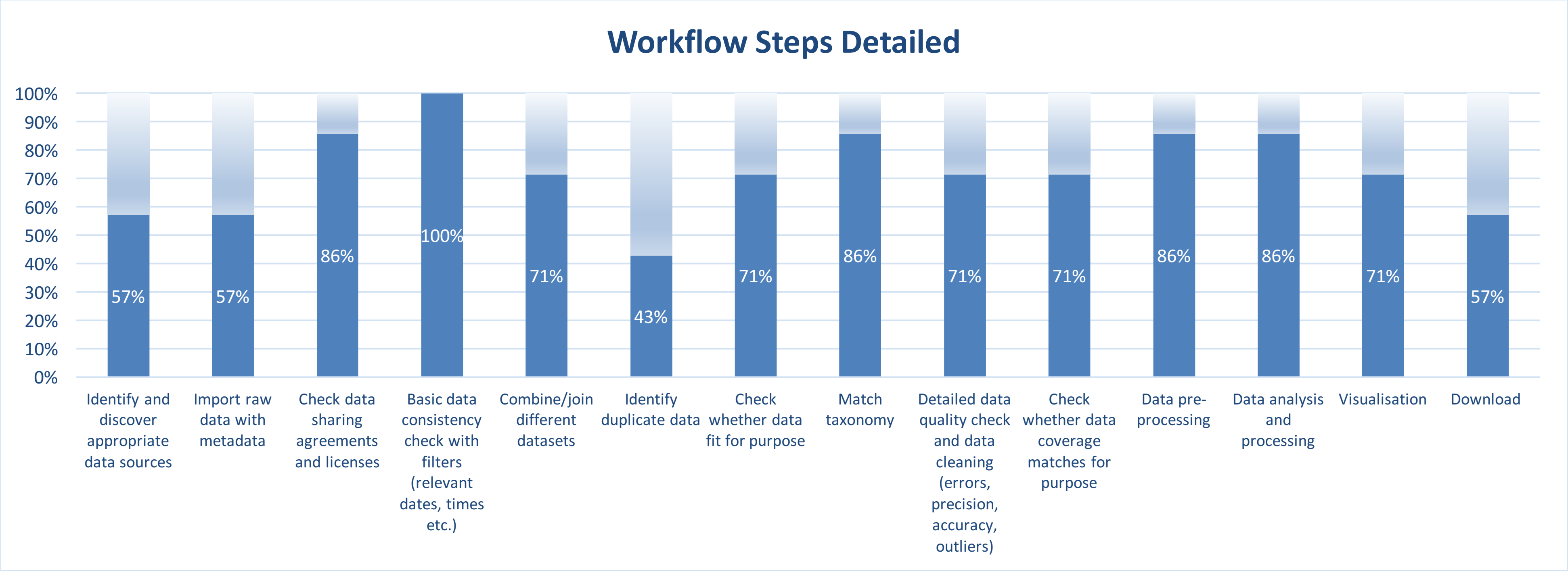
In this annex, Figure 3: Dashboard: Workflow steps, risks and needs has been expanded / broken apart into its constituent parts over several pages to make it easier to read. The charts illustrate EBVs vs RI/ACT Workflow Steps, Needs & Risks

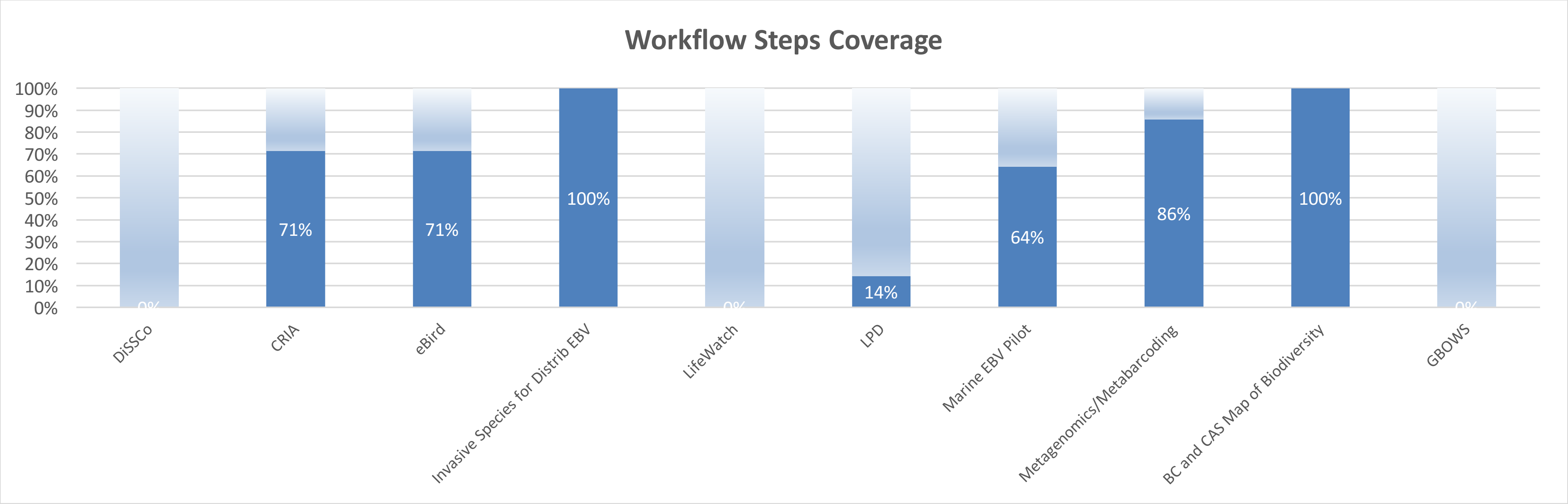


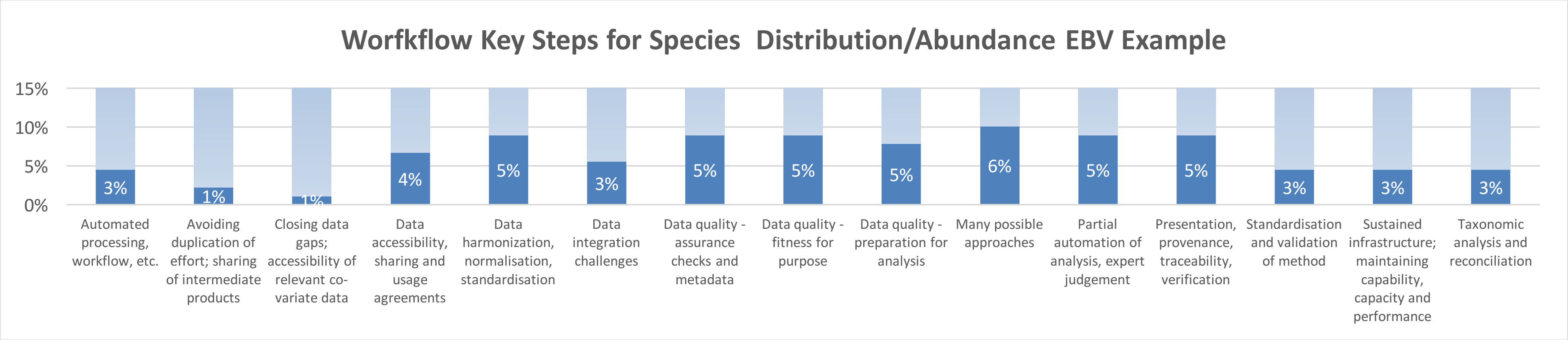

Workflow Steps Overview
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
85%
90%
45%
60%
70%
80%
Data Collection and/or Retrieval
Data Transformation
Model Building
Testing and Validation
Presentation
EBV Output (y/n)

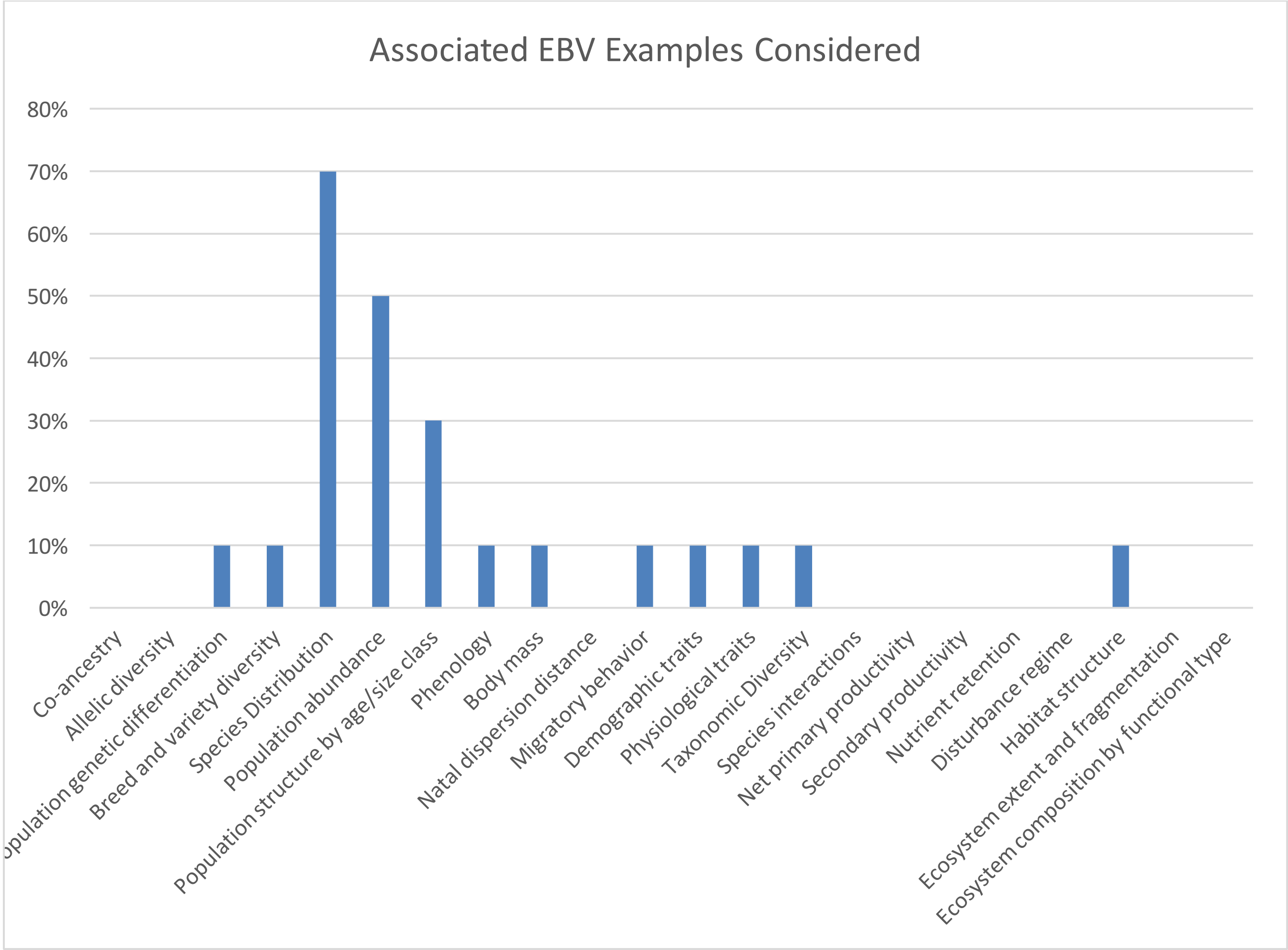

Needs
Taxonomy, ontology,…
Data brokering
Data quality control…
Open access
Funding
Interoperability
Languages
Interactivtity (connection…
Cloud access
Storage Access
Data access internationally
Service catalogue
40%
30%
20%
10%
0%


Risks
Taxonomy, ontology, vocabulary
Interrelation between species
Spatial dimension and associated heterogeneity
Data brokering
Data quality control (methodo & tools)
Interoperability
Sustainability
Fragmentation of collections
Lack of policy
Strategy alignment
40%
30%
20%
10%
0%


Workflow Steps Detailed
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
57%
57%
86%
100%
71%
43%
71%
86%
71%
71%
86%
86%
71%
57%
Identify and discover appropriate data sources
Import raw data with metadata
Check data sharing agreements and licenses
Basic data consistency check with filters (relevant dates, times etc.)
Combine/join different datasets
Identify duplicate data
Check whether data fit for purpose
Match taxonomy
Detailed data quality check and data cleaning (errors, precision, accuracy, outliers)
Check whether data coverage matches for purpose
Data pre-processing
Data analysis and processing
Visualisation
Download


Workflow Steps Coverage
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
0%
71%
71%
100%
0%
14%
64%
86%
100%
0%
DiSSCo
CRIA
eBird
Invasive Species for Distrib EBV
LifeWatch
LPD
Marine EBV Pilot
Metagenomics/Metabarcoding
BC and CAS Map of Biodiversity
GBOWS


Worfkflow Key Steps for Species Distribution/Abundance EBV Example
15%
10%
5%
0%
3%
1%
1%
4%
5%
3%
5%
5%
5%
6%
5%
5%
3%
3%
3%
Automated processing, workflow, etc.
Avoiding duplication of effort; sharing of intermediate products
Closing data gaps; accessibility of relevant co-variate data
Data accessibility, sharing and usage agreements
Data harmonization, normalisation, standardisation
Data integration challenges
Data quality - assurance checks and metadata
Data quality - fitness for purpose
Data quality - preparation for analysis
Many possible approaches
Partial automation of analysis, expert judgement
Presentation, provenance, traceability, verification
Standardisation and validation of method
Sustained infrastructure; maintaining capability, capacity and performance
Taxonomic analysis and reconciliation


Associated EBV Examples Considered
80%
70%
60%
50%
40%
30%
20%
10%
0%
Co-ancestry
Allelic diversity
pulation genetic differentiation
Breed and variety diversity
Species Distribution
Population abundance
Population structure by age/size class
Phenology
Body mass
Natal dispersion distance
Migratory behavior
Demographic traits
Physiological traits
Taxonomic Diversity
Species interactions
Net primary productivity
Secondary productivity
Nutrient retention
Disturbance regime
Habitat structure
Ecosystem extent and fragmentation
Ecosystem composition by functional type