D3.3: Specific implementation roadmap

Project acronym: *GLOBIS-B*

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

Grant agreement no.: 654003

Due-Date:	31st May 2018
Actual Delivery:	29th May 2018
Lead Partner:	Cardiff University
Dissemination Level:	PU
Status:	Public version
Version:	1.0

DOCUMENT INFO

Date and version no.	Author	Comments/Changes
April / May 2018, v0.1, v0.2	**Alex Hardisty, CU**	**Structuring and outlining of content, filling details. Draft for project approval.**
29th May 2018	**Alex Hardisty, CU**	**Final version, following internal review.**

TABLE OF CONTENTS

LIST OF TABLES AND FIGURES

1 Executive summary

Exploiting primary biodiversity and other data to produce and manage Essential Biodiversity Variables (EBV) data products depends on cooperation, practicality and interoperability among multiple stakeholders, including those collecting and mobilising data with EBV potential (EBV-usable data), those making data 'EBV-ready' and those producing, publishing and preserving EBV data products. Ten principles encapsulated as 'The Bari Manifesto' serve as specific implementation actions needed for participating research infrastructures to fully support the emerging EBV operational framework based on transnational and cross-infrastructure scientific workflows. They are best current practice guidance formulated to allow data and infrastructure organisations to enhance their ability to contribute towards production of global EBV data products and to achieve interoperability, whilst retaining autonomy and flexibility to achieve what is needed in ways appropriate to the organisations' own business. These ten principles cover: Data management plan; Data structure; Metadata; Services; Data quality (fitness-for-use); Workflows; Provenance; Ontologies / vocabularies; Data preservation; and, Accessibility. For each principle, a desired outcome, short-term goals and an aspirational goal have been formulated.

The Bari Manifesto is the agenda for further infrastructure development, which has been the principle goal of the GLOBIS-B project. It acts as the specific implementation roadmap supporting cooperation among infrastructure providers working closely together with each other and with responsible GEO BON Working Groups to make the necessary translational steps from proof-of-concept case studies (today's situation) to the future factory-scale processes needed to support EBVs.

2 Contributors

The author of the present document is Alex Hardisty (Cardiff University, UK), WP3 leader. Other members of the GLOBIS-B project team and the participants of the GLOBIS-B workshops have provided contributions shaping the outcome of the work into the Bari Manifesto for Essential Biodiversity Variables (EBV) data products, April 2018. We acknowledge the assistance and contributions of the following individuals for their help in formulating the principles and the precise words used to express those principles: Enrique Alonso, Lucy Bastin, Anne Bowser, Renato De Giovanni, Rui Figueira, Quentin Groom, Rob Guralnick, Wim Hugo, Daniel Kissling, Dimitris Koureas, Wouter Los, Jeffrey Manuel, William Michener, Jorrit Poelen, Hannu Saarenmaa, Dmitry Schigel, and Paul Uhlir.

3 Background

Essential Biodiversity Variables (EBV) are information products located between primary biodiversity data (e.g., occurrence records, sampling events) and statistical indicators that help scientists, managers, politicians and citizens to understand the state of biodiversity [Pereira 2013, Brummitt 2017, Navarro 2017]. Bringing together biodiversity scientists and biodiversity informatics experts from data and research infrastructures, the GLOBIS-B project uses the need to implement Essential Biodiversity Variables (EBV) to drive improvements in interoperability between different e-Infrastructures supporting biodiversity science. In this context, the important question has been how multi-lateral cooperation at the global level can be achieved between data collectors, data providers, monitoring schemes, and biodiversity research infrastructures to support harmonised implementation of EBVs. Further background is provided in [Kissling 2015].

Informatics experts in the GLOBIS-B project have focussed on the technical Information Technology (IT) challenges associated with supporting EBVs, including elaboration of steps and tools needed to move from data collection, integration, filtering, through modelling, testing and validation to the final presentation and publication of an EBV data product (Figure 1, green, red).

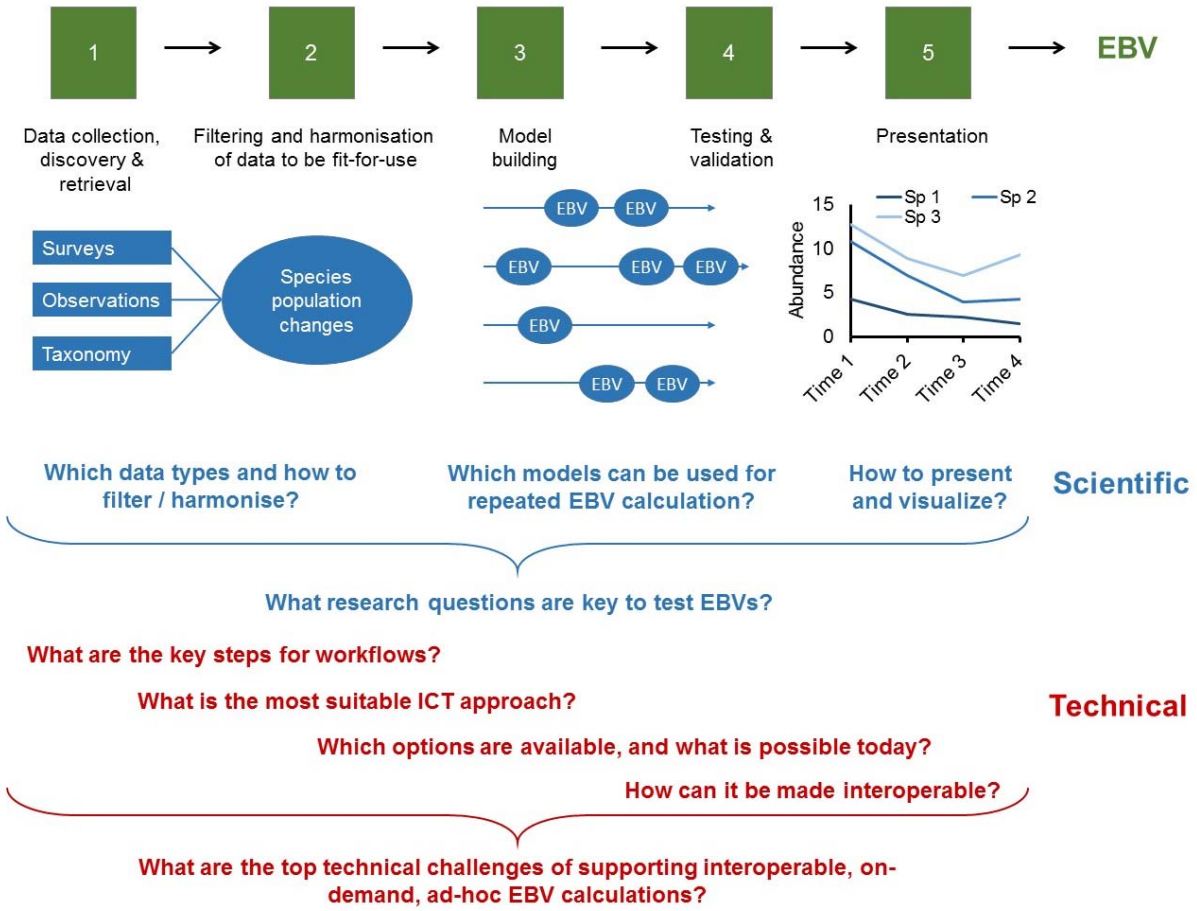

Figure 1: Potential steps for producing Essential Biodiversity Variables (EBVs, green); and related scientific (blue) and technical (red) questions and challenges.

Specifically, experts have examined how to provide data, workflows and computational services for supporting the production of EBVs data products against a vision with multiple assumptions:

- 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.

This work carried out through multiple workshops (February 2016 – March 2018) has been reported in several project deliverables, of which the present deliverable, D3.3 is the concluding one.

Project deliverable D3.1 [GLOBIS-B D3.1 2016] reported on the technical issues and risks associated with the general challenges of provisioning research infrastructures to deliver capabilities for EBV processing. A practical case study, reported in deliverable D3.2 [GLOBIS-B D3.2 2018], has provided an experimental context for identifying specific detailed issues. Project deliverable D2.4 [GLOBIS-B D2.4 2018, Annex section 5.5] identified several important steps involved for infrastructure providers to become ready to provide

support for producing EBVs data products, and more than 100 ideas for solving some of the specific technical challenges. The present deliverable, D3.3 draws on the gathered information to propose ten principles to enhance the ability of data and infrastructure organisations to contribute towards production of global EBV data products and to improve their interoperability. These have been worked out jointly by representatives of those organisations and other experts and prepared as a draft manifesto offering best current practice guidance.

A manifesto approach to describing a specific implementation roadmap has been selected because it allows experts to set the direction of technical infrastructure developments needed, without being prescriptive about how or when infrastructure provider organisations achieve the goals. This approach, which focuses on desirable outcomes rather than specific technical solutions supports the organisations' strong wish to maintain autonomy in the way they carry out their business. It does not constrain their flexibility to respond to fast-changing circumstances.

Collectively the experts consider that adopting and working towards the principles outlined in the manifesto (section 5, page 11 below) can improve interoperability between different biodiversity infrastructures overall.

4 Developing the manifesto

4.1 The manifesto vision and scope

Recalling that the generalised use case of supporting Essential Biodiversity Variables acts as a driver for thinking about how to generally improve informatics interoperability between diverse cyber / e-Infrastructures supporting biodiversity science and ecology, the manifesto vision derives from that which we set out at the beginning of the project i.e., *of enhancing the multi-lateral cooperation of biodiversity research infrastructures worldwide to be able to support measurement and production of EBV data for any geographic area, covering time-period(s) of interest, for any desired species / assemblage / ecosystem / biome of interest, with data that is held in any or multiple repositories, by appropriately skilled persons anywhere in the world*.

To give context, we took a simpler form of the above sentiment and recast the vision behind the manifesto as: "*Achieving global co-operation on interoperability between infrastructures*". To give it a practical emphasis we set the scope as covering those principles necessary to achieve deployment and execution of standard workflows for preparing, publishing and preserving fit-for-use EBV data products that are comparable with one another.

4.2 Aggregated view from prior work

Guided by the main elements for an EBVs IT technical framework (see box) and drawing on results from earlier project workshops – specifically, identified IT technical challenge topics and more than one hundred solution ideas – infrastructure experts developed an aggregated view of important topics to be covered by a manifesto (Figure 2, Figure 3).

Main elements for an EBVs IT technical framework must include the following:

1) A multi-step workflow outline that appears workable in practice;
2) Common dimensional structure across EBV data products;
3) Specific metadata descriptions for EBV data products;
4) Common workflow representation, independent of underlying computing platform;
5) Consistent quality checking and assertion across data from different sources; and,
6) Use of standard mechanisms for provenance recording and packaging.

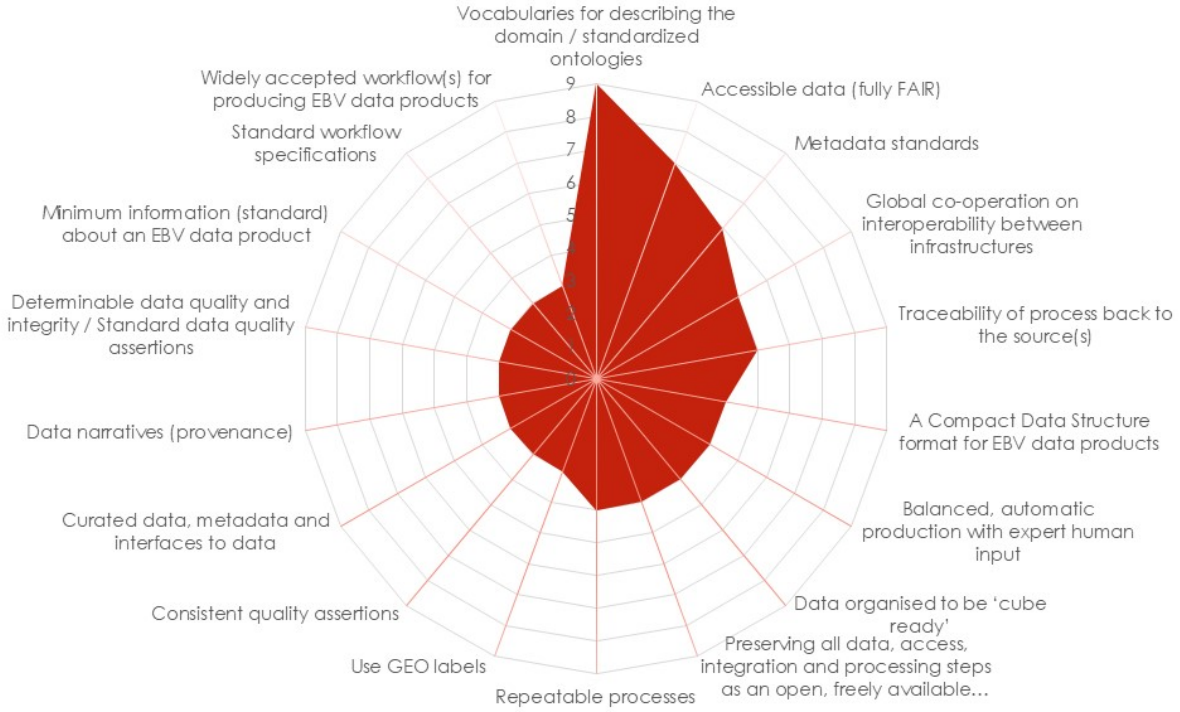

Figure 2: Top selected outcomes to promote

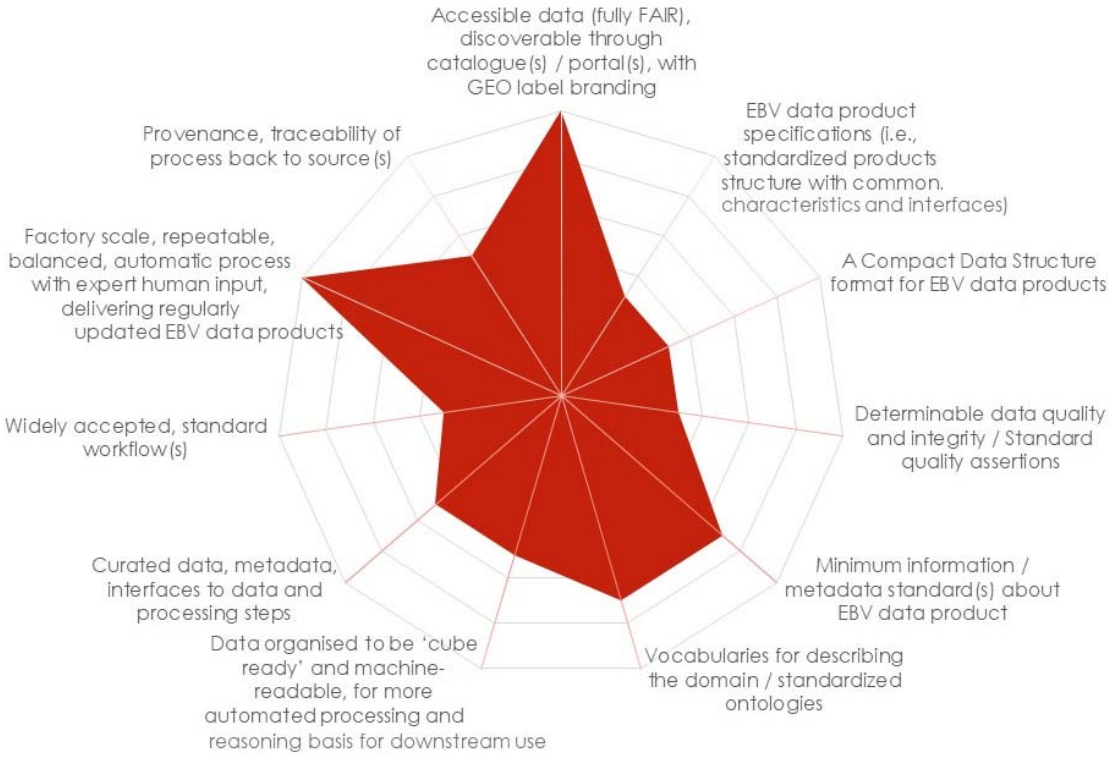

Figure 3: Outcomes, consolidated and reduced

Based on this, ten principle headings have been agreed as most important for guiding work towards the vision of achieving global co-operation on interoperability between infrastructures for EBV data products, and thus are the basis for the manifesto. The ten principles cover:

- Data management plan;
- Data structure;
- Metadata;
- Services;
- Data quality (fitness-for-use);
- Workflows;
- Provenance;
- Ontologies / vocabularies;
- Data preservation; and,
- Accessibility.

For each principal, a desired outcome, short-term goals and an aspirational goal have been formulated and agreed during Workshop 4 of the project, which was held in Bari, Italy during 26 – 28 February 2018, hence the name given to the manifesto. The full text of 'The Bari Manifesto' is in section 5 below, with a summary of main points contained in Annex 1.

5 The Bari Manifesto for Essential Biodiversity Variables (EBV) data products

Introduction

Exploiting primary data to produce and manage Essential Biodiversity Variables (EBV) data products depends on cooperation, practicality and interoperability among multiple stakeholders, including those collecting and mobilising data with EBV potential (EBV-usable data), those making data 'EBV-ready' and those producing, publishing and preserving EBV data products. Ten principles offer best current practice guidance to data and infrastructure organisations to enhance their ability to contribute towards production of global EBV data products, whilst retaining autonomy and flexibility to achieve what is needed in ways appropriate to the organisations' own business.

This 'Bari manifesto', April 2018 has been prepared by a collective representation of the global biodiversity informatics research and data infrastructures community as an outcome of the European Commission Horizon 2020 funded "GLOBIS-B" project (GLOBal Infrastructures for Supporting Biodiversity research), 2015 – 2018.

The Ten Principles

1. Data Management Plan

Projects developing EBV data products should have comprehensive data management plans.

Components of the plan should include information about: data structures and packaging; data formats and standards; metadata standards and tools; workflows; provenance; the data quality control and quality assurance processes; referenced vocabularies and ontologies; policies that will be adhered to; and the budget and resource requirements to produce and curate an EBV data product and its requisite EBV-ready datasets (people, systems, training, software and services, repositories, maintenance).

2. Data Structure

EBV data products should adhere to agreed-upon minimum dimensions for each product (i.e., time, space, name/taxonomy (where applicable), etc.). All EBVs should be accommodated in a common framework that conforms (as far as is practically possible) with content and schema standards for representation formats and exchange protocols.

Each EBV class / variable is likely to have its own distinct data model that should be part of the overall conceptual data structure. Clear definition of these data models will help to identify what vocabulary definitions and relations are needed (see below). The use of standard content and schema standards (e.g., NetCDF, JSON and other compact data structures) encourages interoperability with the widest possible range of processing and visualization tools.

Digital objects (DO) should be used as the means of wrapping and structuring information associated with the production and maintenance of EBV data products, including the steps necessary to generate EBV-ready datasets from documented sources. Operations acting on DOs and EBV data products should be unified at both the DO level and the EBV data product level, allowing interactions with the EBV data product through DO level operators (create, move, copy, update, etc.). Typed EBV DOs should carry minimum level kernel information, including, but not limited to DO type, version, capabilities list and access control information.

3. Metadata

EBV data products and the EBV-ready datasets from which they are generated should have associated human- and machine-readable metadata, compliant with accepted community standards and sufficient for purposes of data discovery, access, fitness-for purpose evaluation, citation, interpretation and use.

Accepted community metadata standards include those from bodies such as: TDWG, ISO/IEC, RDA, OGC and W3C. EML and MIxS standards are also relevant. OWL or SPARQL traversal and interpretation of metadata may be enabled by linking the metadata to reference ontologies.

4. Services

EBV data products, EBV-ready datasets, digital objects and other related services should expose their capabilities and be accessible through common, standardized Application Programming Interfaces (APIs).

Disaggregating programmatic functionalities into discrete services and operations offered through standardized APIs makes it easier to implement, maintain and evolve services that are common across multiple infrastructures. As a first step, the community should agree upon and adopt standard services for discovery of, and access to, EBV data products and EBV-ready datasets. These services can evolve to a broader range of community tools that cover processing, brokering, visualization and workflow execution.

5. Data Quality (Fitness-for-use)

Each EBV data product and EBV-ready dataset should include data quality documentation, sufficient to identify fitness-for-use of the data for specific purposes.

The data quality decisions made during production of an EBV should be fully documented, along with applied thresholds and criteria. Standard tests (for example, as defined by TWDG-DQIG) should be automated and widely implemented. It is desirable that assertions resulting from data quality tests should be available as standard annotations at the record level wherever appropriate. The generation of EBV data products from EBV-ready datasets may involve sub-setting and filtering based on record-level quality assertions, and aggregation of quality assertions to produce a quality evaluation at the product level. Report-back of quality assertions to data providers should promote corrections at the source.

6. Workflows

It should be possible to execute published, standard workflows for preparing, publishing and preserving EBV data products and the EBV-ready datasets from which they are produced. Ideally, such workflows should be defined and represented in a non-proprietary manner.

Standard workflows are needed to ensure that data products are both reproducible and consistent over time. They should be non-proprietary in terms of their representation language so that they are portable across underlying execution mechanisms in different infrastructures. The discovery, selection, harmonisation and quality assurance steps necessary to make datasets ‘EBV-ready’ should also be documented in a reproducible manner as an integral component of this workflow.

7. Provenance

It should be possible to trace the EBV production process from the product back to the primary data and to reproduce the process. Provenance information must be readable both by humans and by machines.

In the short-term this implies that details of all elements used in production of the EBV data product, such as the source data, the software tools, and the workflows should be packaged together; for example, as a digital object with a persistent identifier. Various technology alternatives are available to achieve this. In the longer term, tools supporting automated provenance generation and tracking should be employed throughout the production process, leading to the potential for provenance graphs to be automatically traversed to understand dependencies.

8. Ontologies / Vocabularies

EBV data products and EBV-ready datasets should be described by standard, openly accessible and machine-readable key vocabulary terms and conceptual relations (ontologies) presented in a simple way to enable wide usage.

An extensible EBV Application Ontology (EBVapp), covering the main components of an EBV semantic layer should be developed as an interoperable and complementary part of the OBO Foundry of ontologies (www.obofoundry.org). This ontology should, to the greatest extent possible, inherit from and coordinate with terms and concepts from existing sources, such as those of TDWG and the biodiversity science domain (e.g., Darwin Core (DwC), Biological Collections Ontology (BCO), Environment Ontology (ENVO), Population and Community Ontology (PCO), etc.), the OBO Foundry and the SWEET collection (https://sweet.jpl.nasa.gov/). Persistent efforts should be made to converge or align the descriptions of primary data resources used in the production of EBV data products. Through training, the wide adoption of vocabularies by research communities should be promoted.

9. Data Preservation

EBV data products and associated underlying data should be preserved with an associated persistent identifier in a community supported and trusted repository.

Data underlying EBV data products must be preserved in the form of a snapshot of the raw data (and any transformations necessary to make that data 'EBV-ready') at the time the data product was produced.

10. Accessibility

EBV data products and EBV-ready datasets must be sensitive, timely and FAIR (Findable, Accessible, Interoperable and Reusable).

Data should be mobilised and processed quickly from the point of production, ensuring they are available in a timely manner for research and policy needs, with appropriate attribution and the fewest possible limitations on use. Licenses (if any) should be both human- and machine-readable.

The FAIR data principles (doi: 10.1038/sdata.2016.18) cover requirements relating to metadata, identification, cataloguing and licensing. The principles aim to assist humans and machines in their discovery of, access to, integration and analysis of task-appropriate scientific data and their associated algorithms and workflows. EBV data products and the workflows necessary to create and use EBV-ready datasets must be findable and accessible via standard persistent identifier resolution mechanisms (for example, Digital Object Identifiers (DOI)) and their metadata must be openly available, and searchable via a catalogue maintained by an acknowledged authority, for example, GEO BON.

April 2018.

END OF MANIFESTO TEXT.

6 Conclusions

The ten principles of the Bari Manifesto serve as a set of specific implementation actions for participating research infrastructures to fully support the emerging EBV operational framework based on transnational and cross-infrastructure scientific workflows. Even with continuing uncertainty where too little is agreed about how the overall EBV production process will work and who will be the various actors, pursuing the principles with their desired outcomes can lead to improved informatics interoperability across many aspects of infrastructure for biodiversity science. Developing this manifesto as the agenda for further infrastructure development has been the principle goal of the GLOBIS-B project.

To make the translational steps from proof-of-concept case studies (today's situation) to the future factory-scale processes needed to support EBVs, infrastructure providers must work closely together with each other and with responsible GEO BON Working Groups. The Bari Manifesto serves as the specific implementation roadmap basis for such cooperation.

Translational IT steps towards factory-scale production of EBVs data products involves experiment and co-design to jointly address the specific problems of moving from experimental proof-of-concept type studies to first trials of producing and using real data products with real users. Beyond that, a transition must occur to more robust, general-purpose methods and solutions that scale out and up, providing the basis for the long-term support to GEO BON / GEOSS across a wide range of EBVs classes.

- Experimental integration towards first trials with users
  - As started in the ALA/GBIF invasive species case study
  - To illustrate and test what can be done in infrastructure
  - To start clearing the bottlenecks (scientific, technical, legal)
- Transition to factory-scale with integrated platform for production
  - A more generic solution, to extend, scale and embed
  - Recommendations and solutions to bottlenecks
  - Mechanisms for deployment, operations and support

7 References

[Brummitt 2017] Brummitt N, Regan E C, Weatherdon L V., Martin C S, Geijzendorffer I R, Rocchini D, Gavish Y, Haase P, Marsh C J and Schmeller D S 2017 Taking stock of nature: Essential biodiversity variables explained Biol. Conserv. 213 252–5 Online: http://www.sciencedirect.com/science/article/pii/S0006320716303652

[GLOBIS-B D2.4 2018] Konijn, J. (2018). GLOBIS-B Deliverable D3.2: Report of Workshop 3.

[GLOBIS-B D3.1 2016] Hardisty, A., and Manset, D. (2016). GLOBIS-B Deliverable D3.1: Technical issues and risks associated with general challenges of provisioning research infrastructures to deliver capabilities for EBV processing. url: http://orca.cf.ac.uk/100883/.

[GLOBIS-B D3.2 2018] Hardisty, A. (2018). GLOBIS-B Deliverable D3.2: 'Guinea Pig' EBV - Outcomes and conclusions of a thought experiment.

[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. 2015 Jul 3;16(2-3):99-107. doi: 10.1080/14888386.2015.1068709.

[Kissling 2018] Kissling W D, Ahumada J A, Bowser A, Fernandez M, Fernández N, García E A, Guralnick R P, Isaac N J B, Kelling S, Los W, Mcrae L, Mihoub J B, Obst M, Santamaria M, Skidmore A K, Williams K J, Agosti D, Amariles D, Arvanitidis C, Bastin L, De Leo F, Egloff W, Elith J, Hobern D, Martin D, Pereira H M, Pesole G, Peterseil J, Saarenmaa H, Schigel D, Schmeller D S, Segata N, Turak E, Uhlir P F, Wee B and Hardisty A R 2018 Building essential biodiversity variables (EBVs) of species distribution and abundance at a global scale Biol. Rev. 93 600–25 Online: http://doi.wiley.com/10.1111/brv.12359

[Navarro 2017] Navarro L M, Fernández N, Guerra C, Guralnick R, Kissling W D, Londoño M C, Muller-Karger F, Turak E, Balvanera P, Costello M J, Delavaud A, El Serafy G, Ferrier S, Geijzendorffer I, Geller G N, Jetz W, Kim E-S, Kim H, Martin C S, McGeoch M A, Mwampamba T H, Nel J L, Nicholson E, Pettorelli N, Schaepman M E, Skidmore A, Sousa Pinto I, Vergara S, Vihervaara P, Xu H, Yahara T, Gill M and Pereira H M 2017 Monitoring biodiversity change through effective global coordination Curr. Opin. Environ. Sustain. 29 158–69.
Online: https://www.sciencedirect.com/science/article/pii/S1877343517301665

[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.

Annex 1: Summary of the Bari Manifesto

Exploiting primary data to produce and manage Essential Biodiversity Variables (EBV) data products depends on cooperation, practicality and interoperability among multiple stakeholders, including those collecting and mobilising data with EBV potential (EBV-usable data), those making data 'EBV-ready' and those producing, publishing and preserving EBV data products. Ten principles offer best current practice guidance to data and infrastructure organisations to enhance their ability to contribute towards production of global EBV data products, whilst retaining autonomy and flexibility to achieve what is needed in ways appropriate to the organisations' own business.

1. Data Management Plan

Projects developing EBV data products should have comprehensive data management plans.

2. Data Structure

EBV data products should adhere to agreed-upon minimum dimensions for each product (i.e., time, space, name/taxonomy (where applicable), etc.). All EBVs should be accommodated in a common framework that conforms (as far as is practically possible) with content and schema standards for representation formats and exchange protocols.

3. Metadata

EBV data products and the EBV-ready datasets from which they are generated should have associated human- and machine-readable metadata, compliant with accepted community standards and sufficient for purposes of data discovery, access, fitness-for purpose evaluation, citation, interpretation and use.

4. Services

EBV data products, EBV-ready datasets, digital objects and other related services should expose their capabilities and be accessible through common, standardized Application Programming Interfaces (APIs).

5. Data Quality (Fitness-for-use)

Each EBV data product and EBV-ready dataset should include data quality documentation, sufficient to identify fitness-for-use of the data for specific purposes.

6. Workflows

It should be possible to execute published, standard workflows for preparing, publishing and preserving EBV data products and the EBV-ready datasets from which they are produced. Ideally, such workflows should be defined and represented in a non-proprietary manner.

7. Provenance

It should be possible to trace the EBV production process from the product back to the primary data and to reproduce the process. Provenance information must be readable both by humans and by machines.

8. Ontologies / Vocabularies

EBV data products and EBV-ready datasets should be described by standard, openly accessible and machine-readable key vocabulary terms and conceptual relations (ontologies) presented in a simple way to enable wide usage.

9. Data Preservation

EBV data products and associated underlying data should be preserved with an associated persistent identifier in a community supported and trusted repository.

10. Accessibility

EBV data products and EBV-ready datasets must be sensitive, timely and FAIR (Findable, Accessible, Interoperable and Reusable).

The Bari Manifesto, April 2018: Prepared by a collective representation of the global biodiversity informatics research and data infrastructures community as an outcome of the European Commission Horizon 2020 funded "GLOBIS-B" project (GLOBal Infrastructures for Supporting Biodiversity research), 2015 – 2018.

END.