Towards Interoperable Infrastructure for GEO BON by 2015

Workgroup 8 Framework Document

## Document Information

Document Information 1

The Aim of WG8 3

Long-Term Goals 3

Main Objectives 4

Alignment with Programmes and Initiatives 5

Resources and Task Groups 5

Work Programme and High-Level Milestones 6

Important Events 6

Milestones 6

Annexure A: Real-World Applications 8

A.1 Use Cases 8

A.2 Content Workgroup Deliverables 9

A.3 Essential Biodiversity Variables 9

A.4 Data Upload, Aggregation, Indexing, and Dissemination 10

A.5 Indicators and Targets 11

A.6 Tasks and Resource Allocation 11

Annexure B: Conceptual Models 12

Range of Conceptual Models 12

Discovery and Resource Description 12

Content and Abstract Data Models 12

Services, Orchestration, and Architecture Models 12

Tasks and Resource Allocation 12

Annexure C: The Biodiversity Standards Landscape 13

Meta-Data and Resource Description Options 13

Concepts of Meta-Data Liberalisation and Extension 13

Data and Content Standards 13

Service and Processing Standards 13

Ontologies, Thesauri, and Controlled Vocabularies 13

Gaps and Overlaps 13

Tasks and Resource Allocation 13

Annexure D: Implementation Programme 14

Programme 1: Development of Content Standards 14

Programme 2: Conceptual Models and Scientific Foundations 14

Programme 3: Consensus on Use of Ontologies 14

Programme 4: Working Implementations 14

Programme 5: Compliant and Supporting Deliverables 14

Programme 6: Best Practice and Book Chapter Contribution 14

Annexure E: Proposed Table of Contents – Book Chapter 15

Annexure F: Formal Deliverables from Contributing Initiatives 17

GBIF 17

EU-BON 17

DataOne, ILTER, and NCEAS 17

SAEON 17

NIWA 17

BioVel 17

# The Aim of WG8

According to the Detailed Implementation Plan of GEO BON, the *“WG focuses on data integration and interoperability to help coordinate, standardize and manage data collected by a variety of disparate institutions and individuals for many different purposes. The WG has a mandate that is somewhat different from other GEO BON Working Groups. It is not directly aimed at generating certain products about biodiversity, rather, it will focus on building permanent structures and linkages that will support the delivery of such products.”*

This aim was defined in 2010, and still stands. In this document, we scope the delivery mechanism of the required permanent structures and linkages. The aim of the WG is to put the most important pieces in place by 2015. This can only happen by working through existing, funded thematic, regional and national initiatives such as AP BON, CRIA, DataOne, EU BON, GBIF, LifeWatch, SAEON, and by linking them together through the GEOSS Common Infrastructure, as appropriate.

# Long-Term Goals

The long-term goals for GEO BON in respect of service and data interoperability are summarized in the generalized use case described below. Each of the statements in this use case, which broadly covers the life cycle of data or data-driven services, has many implications in respect of the supporting standards, specifications, and architecture that will to some extent be realized by GEO, but will have to be supplemented by initiatives within the Biodiversity community and GEO BON specifically.

From a purely technical perspective, the generic workflow spans phases of description, discovery, assessment, access, analysis, and application or reporting, by stating that it is the interest of any specific community to do the following:

* Ensure that scientific data and services are described properly, preserved properly, and discoverable;
  + *Meta-data standards implied.*
  + *Harvesters, brokers, and meta-data interoperability implied.*
  + *Persistent identifiers implied.*
  + *Protocols and standards for data exchange/ uploads (IPT, Tapir, BioCASE, DiGIR, etc.) implied.*
  + *Preservation standards and formats implied.*
  + *Tools and approaches to make searches more efficient (vocabularies, ontologies, dealing with massive meta-data collections, …).*
  + *Sustainable data centers and long-term archives are implied.*
* Once discovered, its utility, quality, and scope can be understood, even if the data sets are huge;
  + *Implies: Visualisations, feedback on quality, quality metrics and standards, viewing search results in relation to referenced spatial, temporal, and ontological/ taxonomic coverages, ability to dynamically extract 'thumbnail' views of large datasets, …*
* Once understood; it can be accessed freely and openly;
  + *Implies: standardised services, licenses and policies, simplified distribution channels, even if costs are involved, …*
* Once accessed, it can be included into distributed processes, and collated - preferably automatically, and on large scales;
  + *Implies: persistence of mash-ups and mediations, web context documents, web processing services, standards and guidelines for grid computing, ability to construct indicators and standardized, interoperable final products,  …*
* That due recognition is afforded to the creators of the data and services;
  + *Implies: data publication and citation, linking to scholarly articles, …*
* Once processed, the knowledge gathered can be re-used.
  + *Implies: defining and storing templates and examples of finished work, processes, mash-ups, …*

All of this needs to be implemented against the backdrop of

* The push to extend formal meta-data with linked open data;
* The increased availability of crowd-sourced and citizen contributions;
* A proliferation of devices and sensors;
* And the construction of knowledge networks.

A realistic, shorter-term expression of this goal can be summarized as follows[[1]](#footnote-1):

* data flow from observations through various aggregation and processing/ modelling services, supporting evaluation of EBV’s and derived indicators;
* automated and streamlined, as appropriate;
* using a plug-and-play (service-oriented) approach;
* coordinated through a GEO BON registry system and linked to the GEOSS Common Infrastructure;
* transparent to users through multiple channels, portals and applications.

An additional, supplementary data flow exists in the case of GBIF – specifically addressing the process of aggregating data from nodes distributed globally, and then making the aggregated data available as a range of services. This is explained in more detail in Annexure A.4, and implies an additional set of biodiversity-specific protocols and standards in support of the process.

# Main Objectives

While the long-term goals will enable the generalised use case, objectives closely aligned to the process of infrastructure establishment need to be formulated. By evaluating the implications of the generalized use case, we believe that the following objectives will assist with achievement of the goals:

1. **Concentrate on Real Applications**: Based on real-life use cases, the Essential Biodiversity Variables identified by GEO BON, and the deliverables envisaged by other GEO BON working groups.
2. **Solid Foundation**: Provide sound scientific grounding for extensions to meta-data, content, and services specifications. One of the likely outputs of this process is a set of abstract/ conceptual models.
3. **State of Play**: Assessment of the existing range of meta-data, service, and content standards and specifications, and how these align with the portfolio of use cases, EBVs, and deliverables identified by GEO BON and associated initiatives. Gaps need to be addressed within the framework of the abstract data model.
4. **Think Big and Do Small**: A range of content standards, based on a common abstract model is foreseen, but it is important to start with one of the simpler content specifications, and to develop this into a community-endorsed standard through traditional processes, involving TDWG and OGC – with minimum delay.
5. **Immediate Benefits**: An implementation programme is required to ensure that WG delivers tangible, useful results as soon as possible. For this reason, one needs to plan for three deliverable cycles, one each at the end of the calendar year – coinciding roughly with the GEO Plenary and the demonstration of the Architecture Implementation Pilots in each of these.
6. **Community Consensus**: Important aspects of consensus include the best practice in respect of application of controlled vocabularies and thesauri, and continued improvement of the ontologies on which these are based. In addition, recommendations in respect of new or extended standards have to follow an adoption path through two (possibly more)
7. **Best Practice**: Contribution of a Chapter on Biodiversity informatics to the planned GEO BON book – summarizing the standards and specifications landscape, resources and services that are already available, and best practice in respect of creating infrastructure for data/ service cataloguing, publication, and preservation. A new perspective includes the emerging field of protocols for crowd-sourced and citizen-contributed observations.

# Alignment with Programmes and Initiatives

To be completed – inputs required. Provisional listing:

1. Annual GEO Architecture Implementation Pilot
2. EU-BON and related initiatives
3. DataOne and its related initiatives
4. On-going extensions development for DwC driven by GBIF

# Resources and Task Groups

WG8, in addition to the individual participants in the activities of the group, can draw on the outputs and resources of the following to varying extent (Annexure F):

1. Ongoing development of GBIF services and facilities.
2. The EU-BON Project, as well as associated initiatives
3. SAEON’s contribution to the ICSU WDS, which can be directly useful to GEO-BON.

# Work Programme and High-Level Milestones

## Important Events

Year 1: 2013

1. GEOSS AIP-6 Kick-Off, January 2013
2. GEO Workplan Symposium, May 2013, Geneva (GWS 2103)
3. TDWG Conference, 28 Oct – 1 Nov, Florence (TDWG 2013)
4. GEO Plenary, End November 2013, Florence (GEOP 2013)

Year 2:

To be completed

## Milestones

Important Milestones:

1. 31 January 2013:o
   1. Finalisation of this document
   2. Agree participation in AIP-6 – scope and tasks
2. 7 February 2013:
   1. Finalise Table of Content
   2. Contributors to GEOBON Handbook Chapter
3. 15 April 2013:
   1. Finalise discussion paper on conceptual model(s) for biodiversity content.
   2. First draft of book chapter contributions, supporting online content, and case studies.
   3. Arrangements in place for workshop during GEO Workplan Symposium
   4. Call for papers: TDWG Conference Session on Biodiversity data. Paper on Biodiversity Content – conceptual model included.
4. Early May 2013:
   1. Work session at GEO Workplan Symposium to discuss and publish discussion paper on conceptual model for biodiversity content.
   2. Confirmation of three-year workplan at GEO Workplan Symposium
5. End of July 2013:
   1. Finalise first, and possibly second, set of Biodiversity Data Content Standards for discussion and confirmation at TDWG Conference.
   2. Final proofs for GEO BON Chapter are completed.
6. End of August 2013:
   1. Arrangements for workshop(s) at TDWG Conference are completed.
   2. First Iteration of AIP-6 contribution is available
7. End of October 2013
   1. TDWG Conference – Florence
      1. Workshop: Content Standards for Biodiversity Data
      2. Session: Practical Application of Standards in Biodiversity Systems and Initiatives.
   2. Final release for API-6 contribution
8. End of November 2013
   1. Release of Content Standard Process (TDWG/ OGC) at GEO plenary;
   2. Release of GEO BON Handbook (or Chapter) at GEO plenary;
   3. AIP-6 participation should it be demonstrated at GEO plenary.

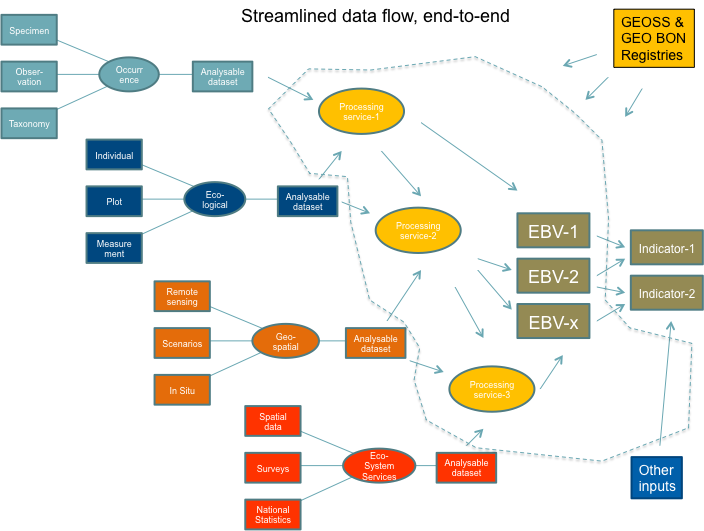
# Annexure A: Real-World Applications

## A.1 Use Cases

Workgroup 8 deals with three sets of use cases:

1. A generic use case (described under ‘[Long-Term Goals’](#_Long-Term_Goals)) - this explains the information-technology related tasks and standards that are implied by an ideal situation. WG8 will review this from time to time, and it functions as our technical goal.
2. This use case is a generalization of a more specific use case (the ‘End-to-End Use Case’), which details the data flows and services that need to be in place for a practical demonstration of our long-term goal. This use case was developed by WG8 during the Asilomar II work sessions in December 2012. See Figure A.1 below. We see this as an ambitious definition of our contribution to AIP-6.
3. Both these use cases are general cases or aggregates of the use cases identified by the other workgroups in their sets of deliverables. WG8 needs to verify that our high-level abstractions of these use cases remain representative of, relevant to, and accurate in respect of what the other workgroups are planning to deliver (See A.2).

Figure A.1: ‘End-to-End’ Use Case



The use case describes the intention to combine data from four generic services:

* Typical occurrence data, such as may be found in GBIF;
* Ecological data, which may include traits, relationships, and assessment of environmental drivers (physic-chemical, …);
* Additional context, derived from remotely sensed or in-situ data, and how these relate to different scenarios;
* Assessments of impacts within a framework of ecosystem service monitoring, pricing, and evaluation.

These disparate data services will, in, time, hopefully be standardized in terms of discovery, protocols, and content, and thus be able to combine in standardized processing in the web: these may have a wide range of objectives, ranging from simple composite presentations to modeling processes and scientific workflows.

The outputs of these, apart from directly useful contributions to science (as Essential Biodiversity Variables, but not limited to it), also need to have a wider societal benefit, and one of the simplest ways to accomplish this is to include the outputs into globally or regionally relevant indicators (Millennium Development Goals, Aichi Targets, etc.).

## A.2 Content Workgroup Deliverables

To be completed.

## A.3 Essential Biodiversity Variables

Essential Biodiversity Variables, as defined by GEO BON participants and recently published as a policy paper[[2]](#footnote-2), addresses the question of ‘what to observe’, and by extension, ‘what to manage, disseminate, and preserve’, and as such provides a valuable starting point for WG8’s stated objective of identifying ‘families of data’ – in an attempt to minimize the number of content specifications and standards that will be required to support the EBVs.

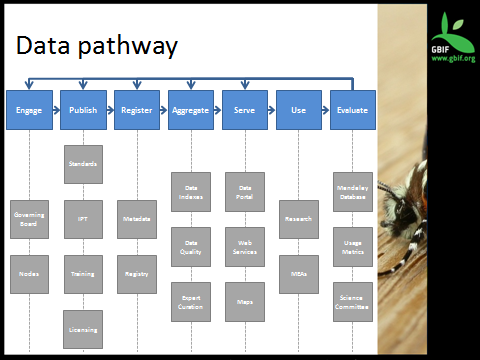
Figure A.3 summarises the scope of EBVs, based on inputs distributed at Asilomar II, and elaborated in the policy paper.

Figure A.3: Scope of Essential Biodiversity Variables

|  |  |
| --- | --- |
| **EBV Class** | **EBV** |
| Genetic Composition | Allelic Diversity for Selected Species |
|  | Breed and Variety Diversity |
| Species Populations and Ranges | Abundances for a selected set of species |
|  | Distributions for a representative set of species |
| Species traits | Phenology of selected functional groups |
|  | Body Mass for Selected Species |
| Community Composition and Interaction | Overall taxonomic diversity for selected locations |
|  | Species interactions |
| Ecosystem Extent and Structure | Ecosystem extent and fragmentation for a range of ecosystems |
|  | Ecosystem structure |
| Ecosystem function and processes | Net primary productivity |
|  | Nutrient retention |

## A.4 Data Upload, Aggregation, Indexing, and Dissemination

A special use case exists for a number of initiatives in respect of the process of data aggregation – of such initiatives, GBIF and DataOne are specific examples. The section below describes the GBIF process (‘Data Pathway’) in more detail:



## A.5 Indicators and Targets

## A.6 Tasks and Resource Allocation

# Annexure B: Conceptual Models

## Range of Conceptual Models

## Discovery and Resource Description

## Content and Abstract Data Models

## Services, Orchestration, and Architecture Models

## Tasks and Resource Allocation

# Annexure C: The Biodiversity Standards Landscape

## Meta-Data and Resource Description Options

## Concepts of Meta-Data Liberalisation and Extension

## Data and Content Standards

## Service and Processing Standards

## Ontologies, Thesauri, and Controlled Vocabularies

## Gaps and Overlaps

## Tasks and Resource Allocation

# Annexure D: Implementation Programme

## Programme 1: Development of Content Standards

## Programme 2: Conceptual Models and Scientific Foundations

## Programme 3: Consensus on Use of Ontologies

## Programme 4: Working Implementations

These contributions may or may not form part of the GEO AIP at some future point.

### Programme 4.1: GBIF Download Services

GBIF will be exposing download services in the near future. SAEON and EU-BON will assist with beta-testing, and SAEON will include value-added services (SensorWeb, WxS) into the AIP-6. See 4.1 below.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Responsible** | **Programme** | **Task** | **Description** | **Due Date** |
| OM | 4.1 | 4.1.1 | [Let others know when beta-testing services are available at http://uat.gbif.org](http://uat.gbif.org) | ASAP |
| WH | 4.1 | 4.1.2 | Set up data repository to work with GBIF data | End April |
| OM | 4.1 | 4.1.3 | Send representative data download (ZIP) to WH as test data set | End March |
| WH, OM | 4.1 | 4.1.4 | Follow up on progress | End June |

OM: Oliver Meyn (GBIF), WH: Wim Hugo (SAEON)

### Programme 4.2: SAEON Portals and Services

SAEON has defined a programme that aligns its activities in ICSU\_related initiatives to establish a biodiversity data center for Africa, its own data publication programme, and GEO-BON involvement in a programme to deliver a number of AIP-6 components and a framework infrastructure. A detailed plan can be found at <http://saeos.dirisa.org/communities/geo-bon/reports>.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Responsible | Programme | Task | Description | Due Date |
| WH, SN | 4.2 | 4.2.1 | Confirm AIP involvement and Scope | Mid April |
| WH | 4.2 | 4.2.2 | Establish Portal Environment(s) | Mid May |
| WH | 4.2 | 4.2.3 | SAEON Marine Data Sets online as standardised services with catalogue entries | Mid October |
| WH, SN | 4.2 | 4.2.4 | Confirm and test bi-directional links to GEOSS Broker | Mid October |
| WH | 4.2 | 4.2.5 | Harvest meta-data from other contributors: WG5, NIWA, EU-BON Partners | Mid October |
| WH | 4.2 | 4.2.6 | Create GBIF Value-Added Services | Mid November |

SN: Stefano Nativi (ESSI Lab), WH: Wim Hugo (SAEON)

### Programme 4.3: NIWA Services

## Programme 5: Compliant and Supporting Deliverables

## Programme 6: Best Practice and Book Chapter Contribution

# Annexure E: Proposed Table of Contents – Book Chapter

Title:

GEOSS Common Infrastructure –

1. How does this integrate?
2. Model template from Hannu
3. Mine the CGI Registry for inputs
4. Fill the gaps in respect of component and service registration
5. GEO-BON Profile of CGI

Engagement with other workgroups – review drafts against our processes, matrices, etc. – ask Michele

Link in one of the earlier chapters – editor? Bob?

Annotation and Curation Standards/ Quality Metrics/ Coverage Assessments

Cost and Time Metrics

Processing and Protocol, Provenance

Best practice

1. Trends and an Emerging Culture of Data Sharing, Publication and Citation
   1. Data Citation and Publication
   2. Permanent Identifiers and Linked Open Data
   3. A Scalable, Interoperable Architecture
   4. Free and Open Data: Licensing and Policy
   5. Roles and Responsibilities
2. Scope of Biodiversity Observation
   1. Essential Biodiversity Variables
   2. Important Entities, Relationships, and Properties
      1. Species and Genomes
      2. Specimens, Samples and Occurrence
      3. Traits
      4. Ecosystem Structure
      5. Interdependencies
3. Domain-Specific Best Practice in Respect of
   1. Scales of Observation and Aggregation
      1. Appropriate Scales of Observation in Support of EBVs
   2. Protocols for Observation
      1. Methodology
      2. Provenance and Quality
      3. Special Case: Citizen Science
   3. Data, Services, and Supporting Standards
      1. Data Families
      2. Applicable Services, Content, Catalogue, and Meta-Data Standards
      3. Overview of Compliant Software and Services
   4. The Role of Vocabularies, Thesauri, and Ontologies
   5. Distributed Processing, Workflow, and Automation
   6. Non-Standard Data Challenges
   7. Data Curation, Annotation, and Preservation
4. Real Use Cases
   1. The GEO BON Interoperability Use Case
      1. Integrating GEOSS and GEO Common Infrastructure
   2. The GBIF Workflow Use Case
      1. Intelligent normalization
5. Challenges, Research Questions, and Gaps
   1. Development of Extended Standards
   2. Long-Term Access and Funding Models
   3. Dealing with ‘Big Data’

It is envisaged that the chapter will be supplemented by a number of online resources, which may include the following:

1. A number of case studies, reinforcing best practice and illustrating the benefit derived from standards. Several initiatives have informally expressed an interest in contributing material for this (ASEAN, Latin America).
2. Detailed ‘How-To’ guides obtained from practitioners in respect of standards implementation.

# Annexure F: Formal Deliverables from Contributing Initiatives

## GBIF

Contact Person: Eamonn O’Tuama (

## EU-BON

Contact Person: Hannu Saarenmaa (

## DataOne, ILTER, and NCEAS

Contact Person: Mark Schildhauer

## SAEON

Contact Person: Wim Hugo – wim@saeon.ac.za

SAEON has developed an implementation programme aimed at establishment of biodiversity-related portal facilities as part of the ICSU-funded World Data Centre for Biodiversity and Human Health in Africa. The biodiversity component of this data centre can be seen as a prototype Africa-BON portal.

SAEON will establish and maintain the portal infrastructure to end 2015, and source and contribute content within its reach or ownership. A comprehensive programme description can be found here:

A significant part of the generalized use case (Annexure A.1) will be realized in this programme.

## NIWA

Contact Person: Jochen Schmidt (

## BioVel

## Institute for Environment and Sustainability, Joint Research Commission

Contact: Gregoire Dubois - [gregoire.dubois@jrc.ec.europa.eu](mailto:gregoire.dubois@jrc.ec.europa.eu)

The institute has been developing a number of OGC compliant services to process remotely sensed data in support of biodiversity monitoring, with a focus on protected areas. These services are currently being catalogued as part of a programme called ‘ACP’ – see <http://acpobservatory.jrc.ec.europa.eu/> - and this can be used as an input to GEO BON and the AIP.

Ongoing work parallel to ACP includes e-Habitat (assessment of species habitats and the impact of change on these), which feeds into DOPA (Digital Observatory for Protected Areas). It is foreseen that both these programmes can make a contribution to GEO BON and the AIP.

See <http://dopa.jrc.ec.europa.eu/> and http://ehabitat-wps.jrc.ec.europa.eu/ .

Finally, unpublished work on configuration and publication of biodiversity indicators are also of interest, and can be aligned with the WG8/ WG9 objectives.

1. Annexure A.1 [↑](#footnote-ref-1)
2. Pereira et.al., 2013, “Essential Biodiversity Variables” , Science, Vol 339, AAAS, January 2013. [↑](#footnote-ref-2)