



Deliverable 2.1 (D2.1)

Architectural design, review and guidelines for using standards

M14

Project acronym:	EU BON
Project name:	EU BON: Building the European Biodiversity Observation Network
Call:	ENV.2012.6.2-2
Grant agreement:	308454
Project duration:	01/12/2012 – 31/05/2017 (54 months)
Co-ordinator:	MfN, Museum für Naturkunde - Leibniz Institute for Research on Evolution and Biodiversity, Germany
Delivery date from Annex I:	M14 (January 2014)
Actual delivery date:	M14 (January 2014)
Lead beneficiary:	UEF, University of Eastern Finland
Authors:	Hannu Saarenmaa (UEF, University of Eastern Finland) Antonio García Camacho (CSIC, Consejo Superior de Investigaciones Cientificas, Spain) Éamonn Ó Tuama (GBIF, Secretariat of the Global Biodiversity Information Facility) Donat Agosti (PLAZI, Plazi Inc., Switzerland) Nicolas Bailly (FIN, Fishbase Information & Research Group, Inc., Philippines) Marco Calderisi (TerraData, Italy) Debora Drucker (FDB-INPA, Fundação Amazônica de Defesa da Biosfera, Brazil) Anton Güntsch (BGBM, Freie Universität Berlin, Germany) Kim Jacobsen (MRAC, Musée royal de l'Afrique centrale, Belgium) Matúš Kempa (IBSAS, Botanický Ústav Slovenskej Akadémie Vied) Urmas Kõljalg (UTARTU, University of Tartu, Estonia) Patricia Mergen (MRAC, Musée royal de l'Afrique centrale, Belgium) Israel Peer (GlueCAD, GlueCAD Ltd. – Engineering IT, Israel) Lyubomir Penev (Pensoft Publishers Ltd., Bulgaria) Nils Valland (NBIC, Norwegian Biodiversity Information Centre) Aaike de Wever (RBINS, Royal Belgian Institute of Natural Sciences)

This project is supported by funding from the specific programme 'Cooperation', theme 'Environment (including Climate Change)' under the 7th Research Framework Programme of the European Union.

Dissemination Level:

PU	Public	✓
PP	Restricted to other programme participants (including the Commission Services)	
RE	Restricted to a group specified by the consortium (including the Commission Services)	
CO	Confidential, only for members of the consortium (including the Commission Services)	

All intellectual property rights are owned by the EU BON consortium members and protected by the applicable laws. Except where otherwise specified, all document contents are: “© EU BON project”. This document is published in open access and distributed under the terms of the Creative Commons Attribution License 3.0 (CC-BY), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.



Executive summary

Introduction

The goal of the Deliverable is to provide direction for the data integration and interoperability work of the EU BON project. It makes recommendations on the design of the EU BON and GEO BON network as a system of systems, following the GEOSS agenda. It reviews state of the art in data standards, charting the way forward in integrating biodiversity and ecosystem data, and addresses better use of data standards by the various biodiversity monitoring networks.

Progress towards objectives

The objectives of EU BON work package 2 “Data Integration and Interoperability” are the following:

- Establish an information architecture for the EU BON project that will be compatible with the global GEO BON, INSPIRE, other European projects, and the LifeWatch research infrastructure
- Develop data integration and interoperability between the various networks, and, with the new generation of data sharing tools, enhance linking between observational data, ecosystem monitoring data, and remote sensing data
- Develop new web service interfaces for data holdings using state-of-the-art standards and protocols. Register the networks on the GEOSS Common Infrastructure (GCI) using harmonised metadata
- Develop a new portal to enable fast access to EU BON integrated data and products by researchers, decision makers and other stakeholders
- Ensure global coordination of development efforts through an international data interoperability task force and adoption of the results through helpdesk and a comprehensive training programme

The current deliverable presents a review and design that form the basis for fulfilling each of these objectives.

Achievements and current status

This document consists of two parts. The first part is concerned with review of the various biodiversity e-infrastructures. Importance to EU BON of major European networks (e.g., ALTERNet, EBONE, LifeWatch, SeaDataNet, EModnet), and international ones (GBIF and DataONE), is summarised. As guiding use scenarios for user needs, the Essential Biodiversity Variables and EU and Aichi Targets are being used. Elements of the information architecture are then identified. Data sharing is already well advanced in the biodiversity domain, but the use of an enhanced repository infrastructure, data publishing, and development of a community specific portal are suggested as mechanisms to further advance data sharing.

With particular reference to the European context, the report outlines the reporting obligations relating to biodiversity under the INSPIRE directive and highlight key EU projects and frameworks relevant for EU BON. Taking a mixed thematic approach, a state-of-the-art summary is provided in several key areas, referencing appropriate standards from bodies such as TDWG, OGC, BioCASE, GBIF, LTER-Europe, and PESI. Gaps in the available standards are highlighted, and recommendations are made for their use in the platform under development by EU BON for integrating different data layers (e.g., genetic data, primary occurrence data, monitoring data, ecological measurements, remote sensing data). In the conclusion, we identify some of the main reasons for heterogeneity of biodiversity data and suggest how it may be overcome.

Future developments

This document will guide further work into detailed planning and construction of the biodiversity observation system of systems. Several milestones are due in the next 1-2 years to provide detailed construction plans. Outreach and training will be needed in order to promote wider use of data standards, in particular, in ecological research.

Table of Contents

Executive summary	3
1. Information architecture	7
1.1. Background and the description of work.....	7
1.2. Use scenarios.....	9
1.2.1. Data requirements for the EU and Aichi Targets.....	9
1.2.2. Essential Biodiversity Variables.....	10
1.2.3. Data sharing through a common platform.....	12
1.3. Requirements vis-à-vis existing networks.....	13
1.3.1. GEO BON.....	13
1.3.2. ALTER-Net.....	13
1.3.3. DataONE.....	14
1.3.4. LifeWatch.....	15
1.3.5. GEOSS Common Infrastructure.....	16
1.3.6. GBIF.....	19
1.3.7. Marine projects and networks.....	20
1.4. General requirements for the architecture.....	23
1.5. Functional requirements of EU BON components.....	24
1.5.1. Requirements for the EU BON Portal.....	24
1.5.2. Requirements for the registry and semantic mediation.....	25
1.5.3. Requirements for data/metadata providers and their services.....	26
1.6. Component architecture.....	28
1.7. Application architecture.....	29
1.8. Runtime architecture.....	30
2. Review of data standards	31
2.1. Introduction.....	31
2.2. The European context.....	31
2.2.1. Biodiversity data management at EU institutions.....	33
2.2.2. Biodiversity standards in the INSPIRE Directive.....	35
2.3. Generic data families.....	39
2.4. Ecological measurements.....	40
2.4.1. Standards for ecological data.....	40
2.4.2. INSPIRE and ecological data.....	41
2.4.3. Unifying data standards of biodiversity and ecological monitoring.....	42
2.4.4. Observations and measurements model.....	44
2.5. Genetic/genomic data.....	45
2.5.1. Major genomic data related initiatives.....	45

2.5.2. Data heterogeneity – reasons and overcoming	47
2.5.3. Recommendations for EU BON	48
2.6. Nomenclature and checklists	48
2.6.1. Pan-European Species Infrastructure	48
2.6.2. Catalogue of Life and Species 2000	49
2.6.3. Integrated Taxonomic Information System	49
2.6.4. Global Names Architecture	50
2.7. Linking in-situ and remote sensing data	50
2.7.1. The PPBio example	50
2.7.2. Standards and tools adopted by PPBio.	51
2.8. Scholarly publishing, data papers, and digital literature	52
2.8.1. Workflows for data publishing	53
2.8.2. EU BON and scholarly data publishing.....	54
2.8.3. Markup formats	55
2.9. Vocabularies and ontologies	57
2.10. Stable identifiers.....	58
3. Conclusions.....	59
3.1. Reasons for data heterogeneity.....	59
3.2 Overcoming impediments for data sharing	59
3.3. Need for new and enhanced data standards.....	60
3.4. EBVs, modelling and data flows	60
3.4 Supporting GEO and IPBES processes	61
References.....	62
Annex I: Related European projects.....	64
BioFresh	64
BIO_SOS.....	64
BioVeL.....	65
EBONE.....	66
EMODnet	67
EUDAT	67
EUMON	68
EuroGEOSS	69
KNEU.....	69
MS.MONINA.....	70
PESI.....	70
pro-iBiosphere.....	70
SeaDataNet.....	71

ViBRANT 71

Annex II: Key standards for EU BON..... 73

Annex III: Acronyms..... 85

1. Information architecture

1.1. Background and the description of work

The description of work defines the task as follows:

Task 2.1 Design of information architecture for EU BON

Starting from the information architectures of relevant infrastructures, i.e., GBIF, LTER, GEOSS, GEO BON, LifeWatch, and INSPIRE, adopt a coherent architecture that will guide the development, integration and interoperability efforts within the EU BON project. The architecture will highlight the relevant components of registry, portal, semantic mediation, workflows, and e-services as envisaged in the GEO BON Detailed Implementation Plan and open access as recommended by the GEOSS Data Sharing Principles. Link to, and adopt informatics components and approaches of other relevant EU projects. The task will address heterogeneity of projects and networks by ensuring that the developments of EU BON can be migrated to permanent infrastructures. In particular, the architecture will map GCI components to European and global biodiversity infrastructure. (Lead CSIC; UTARTU, UEF, GBIF, MRAC, GlueCAD, IBSAS, NBIC, TerraData; Months 4-14)

Hence, the information architecture needs to consider 1) central services such as portals, 2) enabling services such as registries and semantic mediation, and 3) distributed e-services of the data sources. The available services from external projects need to be considered in all these areas. However, we should note that there is no clearly stated provision for building a dedicated network for EU BON. This is on purpose, as it would be futile for a time-limited project to do so. Instead, EU BON will follow the “system of systems” approach of GEOSS, and integrate existing networks and advance interoperability functions that will become part of permanent infrastructures.

That said, EU BON will build a portal and registry functions, and make available data sharing tools in order to implement the interoperability mechanisms. How these will interoperate is explained in the present document. Detailed specifications for each of these will be made separately later (MS231, MS241, MS251), but in the later parts of this document we will review the data and interoperability standards.

Although the task definition does not explicitly mention use cases or use scenarios, these have been called for in the GEO BON Detailed Implementation Plan and a focus on useful products is high in the GEO BON agenda. Therefore, an investigation of important use cases and end products needs to be considered in order to scope the functional requirements.

The DoW also describes some background for the work as follows:

“At the moment there are still considerable interoperability problems to be overcome. EU BON will build on the “GEO BON Detailed Implementation Plan”, and, in particular, apply the “Principles of the GEO BON Information Architecture” as prepared by the GEO BON Working Group for Data Integration (WG 8). More seamless interoperability will be achieved by moving towards cloud computing and enabling web service access to different existing biodiversity data sources (i.e. genetic diversity, species occurrence, ecological traits, habitats, remote sensing). Also applications for modelling, such as trend analysis and geographic visualisation, need to be equipped with web service access the same way as for ecological niche modelling (Muñoz & al. 2009). To make this happen, dedicated actions for data mobilisation, helpdesk, and deployment of a new generation of data sharing tools is planned for EU BON.

The catalogues of the GEOSS Common Infrastructure (GCI) will be the glue that enables these linkages in practice (cf. Nativi et al. 2009). The GCI includes the GEO Portal and the GEOSS Clearinghouse. The latter is a “registry of registries” that can be used for searching data through the component and services registries, the standards and interoperability registry, and community catalogues. EU BON will develop an informatics architecture that can tap into the GCI, contribute content to it, and build functions that serve the GEO Biodiversity Societal Benefit Area in Europe. EU BON will therefore have its own registry, which in the view of the GCI is a “community catalogue”. The EU BON registry will be initiated by combining GBIF and LTER registries, and then expanding further.

EU BON will also build on the work being currently done around the LifeWatch infrastructure. LifeWatch has an elaborate design (Hardisty 2009, Hernandez-Ernst & al. 2010) which will materialise only gradually and aims to be permanent. EU BON, on the other hand, is a time-limited project, which can and will make components and services available for LifeWatch to use. In order to reduce the current heterogeneity of

biodiversity data, it is crucial to ensure that EU BON contributions will find a place in permanent infrastructures. Therefore, information architecture will be developed for EU BON that is compatible with LifeWatch. This will be ensured, among other things, by the fact that the task leader for the EU BON architecture and EU BON portal is also responsible for building the LifeWatch ICT Core.

It is clear that EU BON will need to liaise more broadly on technical issues in order to fully benefit from all new developments. Close cooperation with the GEO Infrastructure Implementation Board will be essential here, as well as ongoing work on data standards (INSPIRE Data Specifications; TDWG: Life Science Identifiers – LSIDs, Structured Descriptive Data - SDD; LifeWatch:Reference Model), which will all be invited to participate in a dedicated EU BON informatics task group.”

This basically calls for determining an interoperability strategy for GEOSS and LifeWatch, initially building on the available GBIF and LTER content.

For advancing data integration, the DoW states this:

“Data integration covers the next steps beyond simple interoperability and data access. It includes aggregation and harmonisation of data, standardisation, semantic interoperability, and building search and download functions for human end users and machine search engines. For EU BON, this work will closely link to other ongoing work on data standards (e.g., INSPIRE Data Specifications, Biodiversity Information Standards (TDWG)), and will progress by bringing experts together in the informatics task force drawn both from consortium partners and associated institutions and organisations. Many data standards, including Darwin Core, are currently undergoing active development and will evolve during the EU BON project. We will take part in this process.

For EU BON, three areas will be targeted: (1) the unification of the taxonomic backbone; (2) integrating ecological/habitat and species data; and (3) linking remote sensing with in situ data. The EU BON taxonomic backbone will be built on the Pan-European Species directories Infrastructure (PESI, www.eu-nomen.eu), and aims to provide an integrated view on nomenclatural and taxonomic information across all organism groups in Europe. At present, PESI integrates information from Euro+Med, Fauna Europaea, and Index Fungorum, thus covering a large proportion of European Biodiversity. EU BON will provide the data through a novel web-service interface based on the EDIT Platform for Cybertaxonomy (<http://wp5.e-taxonomy.eu>), which will make it possible to use the taxonomic backbone as the core of all EU BON tools and services. The machine interface will also support systems developed to allow peer-reviewed community participation in both maintaining and developing this taxonomic resource for the future.

Integrating species-level occurrence data from the GBIF portal and ecological monitoring information from LTER sites will be enabled through cross-mapping relevant metadata descriptions (e.g., ABCD, EML – Ecological Metadata Language). This will be a major breakthrough enabling unified search/discovery across species and ecological resources.

The availability of new sensors with different spatial and spectral resolutions (Kerr and Ostrovsky, 2003), and new analytic techniques based on machine learning algorithms (Bradtner et al. 2011) narrows the gap for integrating remote sensing and on-ground data. EU BON will advance by remote sensing based diversity estimates (Gillespie et al., 2008), in addition to generalized additive models (Parviainen et al., 2009), and by promoting the wider application of neural networks predicting species richness and abundance (Foody and Cutler, 2003).

The European Earth monitoring programme GMES (Global Monitoring for Environment and Security) which is the recognised European contribution to the Global Earth Observation System of Systems (GEOSS), offers distinct opportunities for EU BON, particularly with its upcoming GMES/ESA Sentinel2 satellite which will be launched in 2013 with a spatial resolution of 10m and a potential update rate of 3 days. Among ongoing (FP7) GMES downstream service projects, MS.MONINA (Multi-scale service for monitoring NATURA 2000 habitats of European Community interest) and BIO-SOS (BIOdiversity multi-Source monitoring System: from Space TO Species) are currently developing remote sensing based methodologies for NATURA2000 monitoring, as a contribution and support to in-situ and model based monitoring.”

1.2. Use scenarios

1.2.1. Data requirements for the EU and Aichi Targets

At the Berlin kick-off meeting it was decided that EU BON would, as a contribution to the GEOSS 2015 implementation target, undertake an assessment of six to eight big databases that can support the six EU Aichi Targets¹. There is therefore a need to focus on data requirements and standards that support the latter. The six EU targets are listed in Table 1 together with some suggested indicators drawn from SEBI (Streamlining European Biodiversity Indicators)². The GEO BON review³ of the adequacy of biodiversity observation systems to support the 2020 Targets is also a primary source for suitable data sets.

Table 1: The six EU targets to halt the loss of biodiversity and ecosystem services by 2020 are in line with the Aichi targets.

	Target	Indicators
1	Fully implement the birds and habitats directives	Trends in <u>abundance</u> , <u>distribution</u> and extinction risk of species (SEBI 03) Trends in coverage, condition, <u>representativeness</u> and effectiveness of protected areas and other area-based approaches (SEBI 05)
2	Maintain and restore ecosystems and their services	Trends in <u>abundance</u> , <u>distribution</u> and extinction risk of species (SEBI 01) Trends in coverage, condition, <u>representativeness</u> and effectiveness of protected areas and other area-based approaches (SEBI 07) Trends in pressures from habitat conversion, pollution, <u>invasive species</u> , climate change, overexploitation and underlying drivers (SEBI 14)
3	Increase the contribution of agriculture and forestry to maintaining and enhancing biodiversity	Trends in <u>abundance</u> , <u>distribution</u> & extinction risk of species (SEBI 03) Trends in coverage, condition, <u>representativeness</u> and effectiveness of protected areas and other area-based approaches (SEBI 05)
4	Ensure the sustainable use of fisheries resources	n/a
5	Combat invasive alien species	Trends in pressures from habitat conversion, pollution, <u>invasive species</u> , climate change, overexploitation and underlying drivers (SEBI 10)
6	Help avert global biodiversity loss	n/a

¹ <http://ec.europa.eu/environment/nature/biodiversity/comm2006/2020.htm>

² <http://www.eea.europa.eu/publications/streamlining-european-biodiversity-indicators-2020>

³ http://www.earthobservations.org/documents/cop/bi_geobon/2011_cbd_adequacy_report.pdf

1.2.2. Essential Biodiversity Variables

The Essential Biodiversity Variables (EBVs)⁴, under development by GEO BON, provide another critical resource for deriving what data will be required for EU BON. An EBV is defined as “a measurement required for study, reporting, and management of biodiversity change” and GEO BON aims to identify EBVs that are relevant for the CBD Aichi Targets and indicators. EBVs help in two important ways:

- promote harmonised monitoring by stipulating how variables should be sampled and measured;
- facilitate integration of data by acting as an abstraction layer between the primary biodiversity observations and the indicators.

For example (Fig.1), we could build up an aggregated population trend indicator (for multiple species and locations) from an EBV which estimates population abundances for a group of species at a particular place and which, in turn, is derived from the primary, raw data which can involve different sampling events and methodologies.

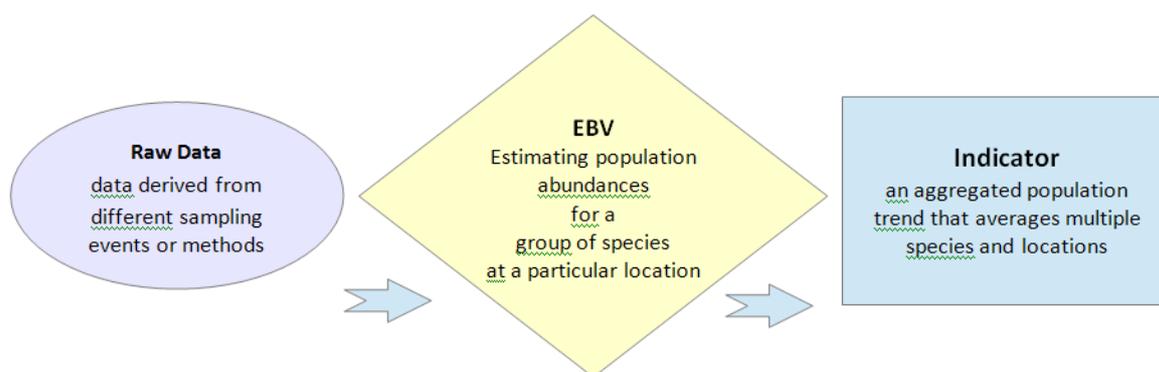


Fig. 1 An EBV acts as an intermediate layer between raw data and indicators.

GEO BON has identified six EBV classes. These are listed in Table 2 with some EBV examples. By analysing the variables/measurements associated with each EBV, appropriate data standards can be proposed or recommended, or new and enhanced standards proposed. To achieve this, it will be important to collaborate closely with the GEO BON work groups and build on the outcomes⁵ of the GEO BON EBV workshop which began collating EBVs and their characteristics (e.g., definition, how to measure, scalability, temporal sensitivity, feasibility, relevance to indicators and targets). Of particular relevance to WP2 are the EBV definitions and how an EBV is measured, e.g., the three EBVs listed for the Species Populations class, can be broken down as illustrated in Table 3. In fact, the Species Population class EBVs are possibly the most tractable given the current status of biodiversity informatics, and could act as the initial test case for EU BON.

In addition to suitable data exchange standards, there is a need to identify appropriate communication protocols for messaging and data flow between systems, and, as part of the architecture design, how to automate the data flows for the EBVs.

⁴ http://www.earthobservations.org/documents/cop/bi_geobon/ebvs/201301_ebv_paper_pereira_et_al.pdf

⁵ http://www.earthobservations.org/geobon_docs_20120227.shtml

Table 2: EBV classes with examples (adapted from Pereira et al. 2013).

EBV Class	Genetic composition	Species populations	Species traits	Community composition	Ecosystem structure	Ecosystem function
EBV example	Allelic diversity	Abundances and distributions	Phenology	Taxonomic diversity	Habitat structure	Nutrient retention

Table 3: The three EBVs of class Species Populations with their definitions and variables/measurements.

Class	EBV	Definition	How to measure in marine, terrestrial, freshwater (spatial, temporal, taxonomic)
Species populations	Species occurrence	Presence/absence of a given taxon or functional group at a given location	Quantify number/biomass/cover at a sample of selected taxa (or functional gps) at extensive suite of sites (selected from stratified random sample or building on existing networks)
	Population abundance	Quantity of individuals or biomass of a given taxon or functional group at a given location	
	Population structure by age/size class	Quantity of individuals or biomass of a given demographic class of a given taxon or functional group at a given location	

Recommendation 1. Define the data requirements and associated standards for the Species Population EBV class.

Recommendation 2. Design an automated data flow for the Species Population EBV class and test within the EU BON network.

The EBV on abundances and distributions would need to be measured from “counts or presence surveys for groups of species easy to monitor or important for ecosystem services, over an extensive network of sites, complemented with incidental data”. Such EBV would be updated at intervals from 1 to 10 years. EBVs have not yet been implemented, but need to be piloted. Piloting would naturally need to start from such organism groups and parts of the world for which sufficient data is available. Birds, butterflies, and vascular plants have earlier been targeted by indicator development, so these would be obvious choices.

This calls for integrating data from sites such as those of LTER, and other regular surveys, and from GBIF. Integration would happen through processing services that would compute abundance trends and changes in distribution for these two types of data: surveys and incidental. These are shown in Fig. 2 as “ecological” and “occurrence” domains. Software tools and web services are available to do these computations, for instance from the TRIM⁶, BioVeL⁷, and EUBrazilOpenBio⁸ projects.

⁶ www.cbs.nl/en-GB/menu/themas/natuur-milieu/methoden/trim

⁷ www.biovel.eu

⁸ www.eubrazilopenbio.eu/

The computation of an EBV of this class would be visualised on the portal, which would allow selecting the data sources and species in question, showing the intermediate steps, and presenting the trend and change of distribution for individual species or whole groups of organisms.

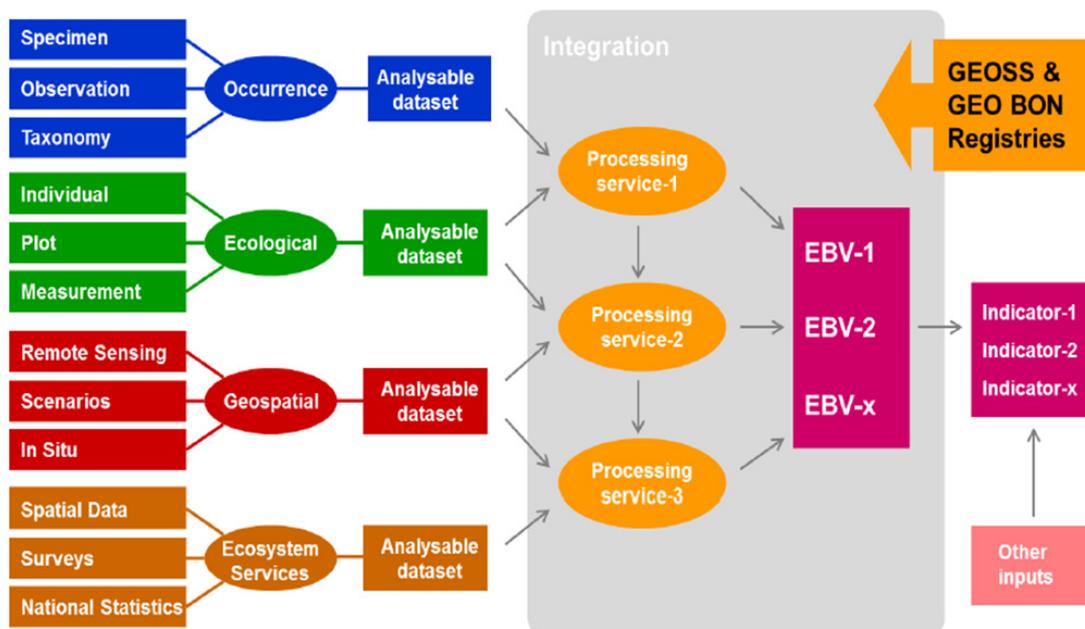


Fig. 2 (from Hoffman et al. 2014). EU BON will be implementing the GEO BON vision of automated, streamlined data flow, end-to-end, from observations to Essential Biodiversity Variables (EBV), using a plug-and-play service-oriented approach, coordinated through the GEO BON registry system and linked to the GEOSS Common Infrastructure, and transparent to users through portals.

1.2.3. Data sharing through a common platform

In its first informatics meeting in Trondheim 2013-05-29/31, the EU BON project prioritised building a common platform for data sharing for the willing biodiversity observation networks. This would allow integrating biodiversity and ecosystem data, possibly also linking to remote sensing data.

This responds to the following statement in the DoW: “EU BON will provide integration between social networks of science and policy and technological networks of interoperating IT infrastructures, resulting in a new open-access platform for sharing biodiversity data and tools, and greatly advance biodiversity knowledge in Europe.” It was agreed that a new platform that integrates primary biodiversity data and ecosystem data is needed, and can be built on existing solutions. The new platform would technically consist of the solutions of the DataONE⁹ network, which may be augmented with semantic mediation for primary biodiversity data. In order to start piloting, willing EU BON Partners will implement DataONE Member Nodes, as appropriate. These will be populated with data from the networks and sites in which each of the Partners is involved. A DataONE coordinating node in Europe may be established towards 2015, and data will be later integrated *via* the EU BON Portal.

DataONE uses Metacat¹⁰ as its data sharing platform. However, to make data integration possible, the EU BON platform would not support just any kind of data to be shared. The data would need to comply with standards like EML, ABCD or Darwin Core with its new extensions for ecological measurement data. This would allow Metacat to be used in similar ways to the GBIF data provider tools. The portal would be able to download raw data tables from Metacat providers and index the data.

⁹ <http://www.dataone.org/> -- Also see the section 1.3.3. for more details.

¹⁰ <http://www.dataone.org/software-tools/metacat>

1.3. Requirements vis-à-vis existing networks

1.3.1. GEO BON

GEO BON (Group on Earth Observations Biodiversity Observation Network) aims at building a network of networks for understanding biodiversity change on earth (Scholes *et al.* 2012). The information components of GEO BON have been outlined in the GEO BON Detailed Implementation Plan¹¹ and its associated technical document (Ó Tuama *et al.* 2010)¹². The main objective for EU BON is to build a substantial part of GEO BON. Therefore it should comply with the set of principles laid out in “The GEO BON Manifesto” below (Hugo *et al.* 2013). As a component of GEOSS, GEO BON is required to conform to the GEOSS Common Infrastructure. Thus, to ensure interoperability with GEOSS, the GEO BON infrastructure must implement the Service Oriented Architecture (SOA) model and its associated international standards and Earth system science multidisciplinary standards and best practices, e.g. ISO 211/Open Geospatial Consortium Reference Model.

Regarding data types and content interoperability, the GEO BON implementation plan prioritises mediation (many data content standards in use; interoperability achieved through mapping of concepts at consumer end) rather than harmonisation (common data exchange schema for all data providers).

The GEO BON Manifesto

GEO BON data and services are described properly, preserved properly, and are discoverable.

Once discovered, their utility, quality and scope can be understood, even if the data sets are huge.

Once understood, they can be accessed freely and openly.

Once accessed, they can be included into distributed processes and collated - preferably automatically, and on large scales.

Once processed, the knowledge gathered can be re-used.

All against a backdrop of a move to extend formal metadata with emerging semantic web technology, increased focus on cross-domain interoperability, and the construction of knowledge networks.

Recommendation 3. EU BON should strive to comply with the principles of the GEO BON manifesto.

Recommendation 4. Open Data, should be normal practice and should embody the principles of being accessible, assessable, intelligible and reusable. Furthermore, data should be made available with a proper license or copyright waiver that allows full reuse.

Recommendation 5. Data encoding should allow analysis across multiple scales, and such encoding schemes need to be developed. Individual data sets will have applications over a small fraction of these scales, but the encoding schema needs to facilitate the integration of various data sets in a single analytical structure.

1.3.2. ALTER-Net

As a long-term biodiversity and ecosystem research network, the main objective of ALTER-Net¹³ is to develop lasting integration amongst its partner institutes, and others, all of whom are involved in biodiversity research, monitoring and/or communication. ALTER-Net established and now supports

¹¹ http://www.earthobservations.org/documents/cop/bi_geobon/geobon_detailed_imp_plan.pdf

¹² http://www.earthobservations.org/documents/cop/bi_geobon/geobon_information_architecture_principles.pdf

¹³ <http://alter-net.info/>

LTER-Europe (European Long-Term Ecosystem Research Network). LTER-Europe consists of both the main LTER web site and national LTER networks. The implementation efforts by LTER-Europe have shown that integration of data and metadata managed by the different sites and platforms is often hampered by technical and legal difficulties and by a lack of the harmonisation and semantic translation of the contents. Regarding the latter, LTER aims to integrate data and databases through an ontology covering any kind of object and its parameters in the field of LTER research.

The *Drupal Ecological Information System* (DEIMS)¹⁴ was developed as a portal for sharing metadata within the LTER network. In DEIMS, end users can describe, discover, view and download information about data sets, research and observation websites, bibliographic references and personnel information. Metadata associated with data sets is modelled using EML and, of particular interest, a form-based tool is provided in DEIMS for uploading data sets and easily generating associated metadata in EML.

Recommendation 6. EU BON should trial the use of the LTER ontology for annotating data sets in a semantic manner in order to enhance their discovery and re-use.

1.3.3. DataONE

Interoperability and data integration with the DataONE network has been prioritised in the EU BON project. The architecture of DataONE network needs therefore to be reviewed in order to ensure compatibility. DataONE is a distributed network of repositories (Member Nodes) and currently four search facilities (Coordinating Nodes), which contain resource descriptions of the Member Nodes.

The Member Nodes¹⁵ maintain a preservation-oriented repository. Different repository products may take different approaches to data preservation, but in general they i) use persistent identifiers for data products, ii) ensure access to these data products over the long term; and iii) ensure that metadata documents exist alongside the data products. Resource Maps provide a common format for describing the bidirectional relationship between a metadata object and the data object(s) it documents. DataONE expects all content submitted *via* a primary system that must have associated Resource Maps. If data owners are not doing this alongside content submission, this should be a service provided centrally by the Member Node. Once published, end users expect the data product to remain the same over the long term. Curation practices should be compatible with data-preservation and data reproducibility ideals in mind. Specifically, content update and archiving activities should be transparent to DataONE end users.

A data package in DataONE is composed of at least one science metadata document describing at least one data object with the relationships between them documented in a resource map document. Resource maps are RDF documents that conform to the Open Archives Initiative Object Reuse and Exchange (OAI-ORE) specification. Resource maps are generated by Member Nodes to define data packages¹⁶.

Member Nodes may use any software, but the majority of them are based on Metacat. There are currently 14 Member Nodes. Their content can be searched directly in each repository, or through the metadata indexes of the Coordinating Nodes. ONE Mercury search currently reveals 168,027 datasets. Keyword “biodiversity” is present in 894, and “species” in 899 datasets. Most of these originate from the Member Node of the Partnership for Interdisciplinary Studies of Coastal Oceans (PISCO¹⁷). Much more data is available on keyword such as “ecosystem” 17,760, “sensor” 19,026, “water” 50,516. The largest Member Node with over 120,000 data products is the Merritt Repository of the University of California Curation Center (UC3) which is part of the California Digital Library. 535 data products cover Europe.

For EU BON, participating in DataONE would, at minimum, mean implementing the OAI-ORE resource maps for the central EU BON Registry. In addition, test sites could directly become

¹⁴ <http://data.lter-europe.net/deims/>

¹⁵ http://www.dataone.org/member_node_requirements

¹⁶ <http://mule1.dataone.org/ArchitectureDocs-current/design/DataPackage.html>

¹⁷ <http://www.piscoweb.org>

DataONE Member Nodes, if they use the Metacat repository software. Building the OAI-ORE interface for GBIF IPT, TAPIR, and BioCASE ABCD data providers is also a possibility to be considered.

The central EU BON Registry or the EU BON Portal could, at some point, become a DataONE Coordinating Node.

1.3.4. LifeWatch

LifeWatch¹⁸ (*e-Science infrastructure for biodiversity and ecosystem research*) is a research infrastructure proposed by the European Strategy Forum on Research Infrastructures (ESFRI)¹⁹. ESFRI serves as facilitator to identify urgent European level research infrastructures. The ESFRI roadmap was first published in 2006, covering 35 projects relating to research infrastructures. It was updated in 2008 bringing the number of research infrastructures to 44. The latest update, published in December 2010, raised this number to 48. For the next few years, ESFRI will focus more on their implementation. The next update of the roadmap will be carried out in 2015.

LifeWatch was included in the first release of the ESFRI roadmap as the result of the collaboration of several biodiversity networks. LifeWatch itself is not an infrastructure generating data, but an environment, mainly for the biodiversity community, for discovering, processing, modelling and collaborating on data (including software tools). The LifeWatch Reference Model²⁰ provides guidelines and specifications for an infrastructure based on proven concepts and standards. Based on the ORCHESTRA Reference Model²¹ which itself extends the OGC Reference Model and complies with the INSPIRE Directive and Implementation Guidelines, the LifeWatch Reference Model addresses both syntactic and semantic interoperability for managing information across data models, data sets, services and workflows. Thus, the reference model requires the use of ontologies to support semantic interoperability.

LifeWatch aims to conform to several published standards whenever feasible. In particular, the following organisations and standards are of interest:

- Biodiversity Information Standards (TDWG): Darwin Core, Access to Biological Collection Data (ABCD), Structure for Descriptive Data (SDD), Taxonomic Concept Transfer Schema (TCS)
- Open Geospatial Consortium (OGC): Open Distributed Processing, POSIX Open System Environment
- Organisation for the Advancement of Structured Information Standards (OASIS): OASIS SOA reference model and reference architecture, OpenGIS Service Architecture
- Open Grid Forum standards
- Geospatial information standards, e.g., ISO 19101:2004, ISO 19111:2003 or ISO 19115:2003
- World Wide Web Consortium standards and XML-based languages as SOAP, WSDL, BPEL, RDF, OWL, SPARQL, etc.

Both EU BON and LifeWatch have to deal with the heterogeneity of data and related metadata provided from external sources. Translation services between models will be required to ensure future interoperability between systems.

Recommendation 7. EU BON will adopt the LifeWatch concepts and Service Model to enhance interoperability for data, protocols and services and to promote syntactic and semantic interoperability between services and applications.

EU BON and LifeWatch share common objectives, mainly the integration of biodiversity networks through a common platform. The LifeWatch architecture (Hernandez-Ernst *et al.* 2009) is based upon the reference model of ORCHESTRA framework, hence following the Service-Oriented Architecture approach.

¹⁸ <http://www.lifewatch.eu>

¹⁹ http://ec.europa.eu/research/infrastructures/index_en.cfm?pg=esfri

²⁰ http://www.eubon.eu/getatt.php?filename=LW-RMV0.5_4310.pdf

²¹ <http://www.eu-orchestra.org/TUs/RMOA/en/html/index.html>

However, the last revision of ORCHESTRA reference model is from 2007. Since then, many technological improvements related to SOA have arisen, mainly concerning the integration bus, mediation services, business process management (workflows, subprocesses) and business activity monitoring.

EU BON architecture will follow ORCHESTRA as a reference, but there is a need to consider at least the following changes and improvements:

- The Enterprise Service Bus (ESB) will act as the common platform in which services will be deployed and managed. In that way, the ESB which will act as a Business Process Management (BPM) solution.
- Service mediation will be achieved through service orchestration executable languages, such as BPMN 2.x or WS-BPEL. WS-BPEL is an OASIS standard executable language for specifying actions within business processes with web services, whilst BPMN 2.x is an Object Management Group standard that provides a graphical notation for business processes and a set of executable instructions that provide service orchestration capabilities. EU BON will delegate a set of common functionalities in the ESB, for example authentication, authorization, transaction management or secure web services management.
- EU BON is intended to act with an institutional focus and a citizen focus, therefore a set of services as well as the common user portal need to be accessible in an open and anonymous way.

As is recommended by the INSPIRE directive, LifeWatch and EU BON are guided by standardisation, therefore EU BON architecture, at a messaging level, will use at least the same standard formats proposed in the LifeWatch Reference Model. Biodiversity data and metadata will be shared or provided using several standards, i.e., Darwin Core, ABCD, and EML. Geospatial information will be included in the aforementioned standards and by providing dynamically generated GML outputs whenever it may be necessary, according to INSPIRE specifications and ISO 19136:2007. Additionally, concerning geospatial grid representations of species distribution, INSPIRE requires the use of the grid ETRS89-LAEA as the specific geodetic Cartesian reference frame.

1.3.5. GEOSS Common Infrastructure

The Global Earth Observation System of Systems (GEOSS)²² is an initiative lead by the Group on Earth Observations (GEO) to link together already existing observation systems in the service of several “Societal Benefit Areas” including biodiversity and ecosystems. The GEOSS Common Infrastructure (GCI)²³ provides the architectural framework essential to implementing the GEOSS concept, supporting the GEOSS Data Sharing Principles²⁴ and enabling “*full and open exchange of data, metadata and products*”. EU BON will be integrated with GEOSS through the GCI using its set of core services that promote the integration of disparate systems as a functional “System of Systems”. GCI is formed of the following elements (as illustrated in Fig. 3):

- Component and Service registry
- Standards and Interoperability registry
- User requirements registry
- Best practices Wiki
- GEO Web Portals
- GEOSS Clearinghouse

²² <http://www.earthobservations.org/geoss.shtml>

²³ http://www.earthobservations.org/gci_gci.shtml

²⁴ http://www.earthobservations.org/geoss_dsp.shtml

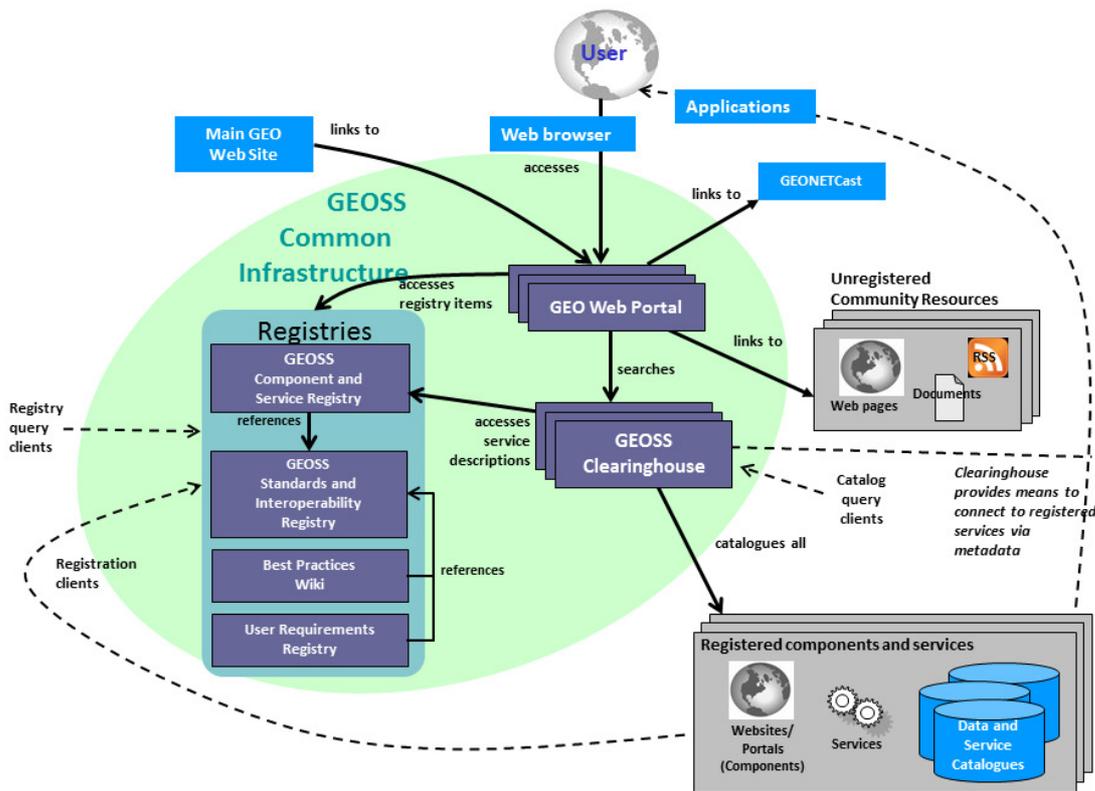


Fig.3 Components of the GEOSS Common Infrastructure.

In the proposed GEOSS architecture, EU BON will, at a minimum, act as a data provider. It will be exposed to the GEOSS Component and Service Registry so that the GEOSS Clearinghouse can catalogue the information and provide access to EU BON metadata *via* an API, e.g., based on the standard OGC CSW 2.0.2²⁵.

The *GEOSS Component and Service Registry* is linked to the *GEOSS Standards and Interoperability Registry*, a collection of standards and community practices that are nominated through a registration process by GEO individuals. Currently, there are more than 290 standards registered (approved or pending). If necessary, EU BON work groups can propose new standards for inclusion in the registry.

Recommendation 8. EU BON should comply with the GEOSS Common Infrastructure requirements, registering components and standards as necessary.

However, it is possible to go beyond the minimum linking mechanism described above, and build a distinct biodiversity community within GEOSS (Fig. 4). The Detailed Implementation Plan of GEO BON and the associated Principles of Information Architecture document have outlined how such a community infrastructure can be built. It is this community infrastructure that EU BON will be contributing. No such formal community infrastructure currently exists.

²⁵ Open Geospatial Consortium – Catalogue Services for the Web: http://portal.opengeospatial.org/files/?artifact_id=20555

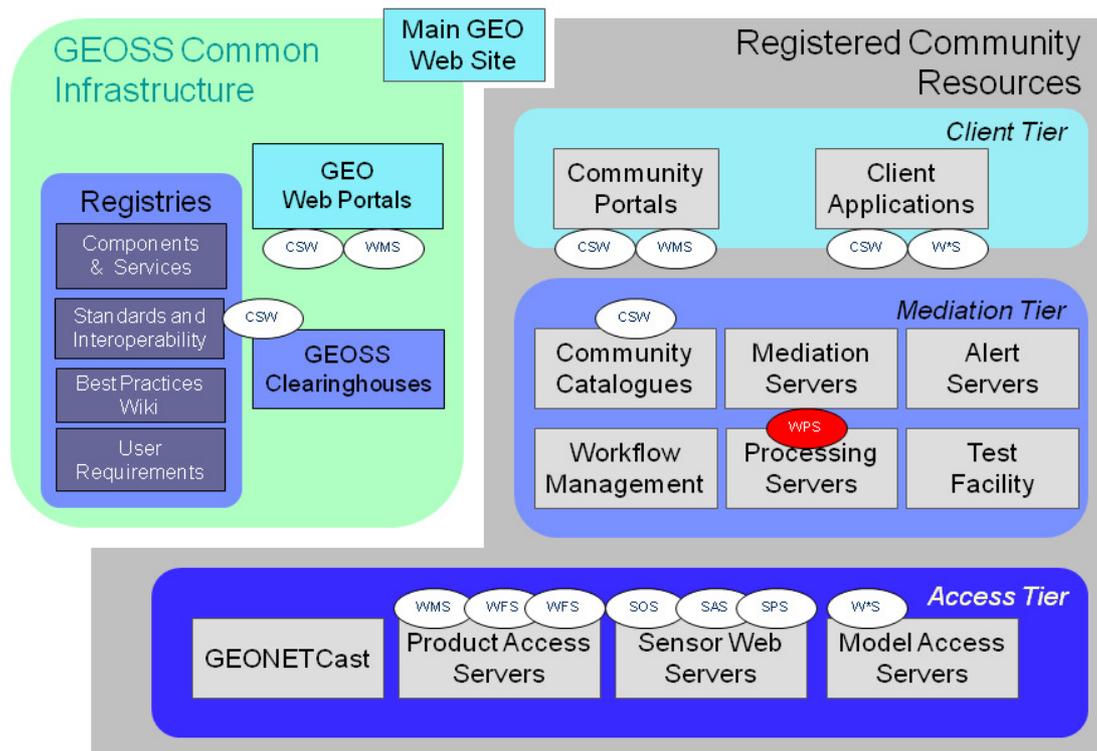


Fig. 4 The GEOSS Common Infrastructure consists of web-based portals, clearinghouses for searching data, information and services, registries and other capabilities supporting access to GEOSS components, standards, and best practices. It will link to the components from community networks.

Currently, the GEOSS Portal can be used to search information resources across about 30 earth observation catalogues through a brokering architecture (Nativi *et al.* 2012). The GEOSS Clearinghouse is one of these registries. Others are national or thematic registries, corresponding to distinct communities. Currently, 891 resources on biodiversity are discoverable this way, although none of the catalogues is dedicated to biodiversity. Five resources can be found when searching for “Darwin Core”, and none for “ABCD” or “EML”. Among these five matches are GBIF, OBIS, the British and Spanish networks, and the Continuous Plankton Recorder Survey²⁶. It is notable that individual biodiversity datasets are not currently discoverable through the GEOSS Portal, only certain, but not all, aggregators are. There is no registry interoperability in GEOSS, but only distributed query from the GEOSS Portal to the selected registries.

Fig. 4 identifies “community resources” and some possible components. From a technical standpoint, biodiversity does not currently constitute a distinct community within GEOSS, because there is no coherent architecture that would link together these or other components, and would have registered them in GEOSS. Many of these components are available, though, as follows:

- Community Portals
 - GBIF
- Client Applications
 - Map of Life
- Community Catalogues
 - GBIF, KNB/DataONE, LTER-Europe, EUMON, BiodiversityCatalogue
- Mediation Servers
 - Catalogue of Life, PESI
- Alert Servers

²⁶ <http://www.sahfos.org/>

- GMES and GEOSS alert services and feeds.
- Workflow Management
 - BioVeL
- Processing Servers
 - EUBrazilOpenBio, D4Life, ...
- Test Facility
 - To test conformance and compliance with service specifications, none yet.
- Product Access Servers
 - Data providers of GBIF and KNB
- Sensor Web Servers
 - Sierra Nevada Global Change Observatory, SAEON, ...
- Model Access Servers
 - openModeller

In this situation, EU BON can contribute by building a community registry, directly searchable from the GEOSS Portal, where all known “Product Access Servers” can be found. EU BON will also build a portal to have a specialised interface to the data and metadata of these resources. The portal can also be used to organise the other community resources and products into a functional whole. We should also note the ongoing discussion within GEO about reorganising the GEOSS Portal so that each of the nine communities built their own portal. This might be necessary, since the GEOSS Portal at the moment is merely a distributed query system to the resource metadata, and building a richer user interface across nine communities has shown to be difficult. The biodiversity community could perhaps pioneer that approach.

1.3.6. GBIF

GBIF is a major integrator of biodiversity occurrence data and thus a top priority for integration in EU BON. The GBIF informatics architecture provides unified access *via* a web portal and web services to a global network of data publishers. Through appropriate standards and tools, the infrastructure is designed to serve three main types of data: metadata (i.e., descriptions of datasets), primary biodiversity data and names data. These data are made available online by numerous GBIF Nodes for central harvesting by GBIF which indexes and integrates the content and makes it available for discovery, retrieval (e.g., via web services) and analysis.

Data can be published to the GBIF network using a number of tools and protocols. The GBIF Secretariat developed the Integrated Publishing Toolkit (IPT) and the associated Darwin Core Archive format as a software platform to facilitate the efficient publishing of biodiversity data on the Internet using the GBIF network. Although the IPT is the solution recommended by the Secretariat for publishers, data published using other software platforms such as DiGIR, BioCASE, or TapirLink will continue to be harvested and indexed by GBIF.

Since the conception of the GBIF network, a registry has been a key component of the informatics infrastructure. The Registry and associated metadata catalogue system has recently been expanded to ensure its suitability as a central discovery tool for datasets holding all classes of biodiversity data, through collaboration with other global and regional biodiversity informatics projects. Implemented as a service-oriented architecture, all interaction with the registry is via a RESTful JSON based API, and the new GBIF portal itself is the first and principal client of the registry. The registry provides a number of functions (available now or under development). Following are the main characteristics of the registry.

The registry acts as an authoritative source of information (metadata) on institutions, datasets, technical services and other key entities as required by registry partners. GBIF uses a profile of EML for describing the datasets.

- The registry is a source of information on inter-relationships between datasets, institutions and other entities according to the needs of the registry partners.

- The registry provides a discovery mechanism for users and machines for: a) registered network entities, b) technical endpoints, c) data definitions (e.g. standards) such as the extensions and vocabularies used in the Darwin Core Archive format. Discovery is provided through indexing of metadata, and through flexible tagging of entities using simple key value pairs of tags, optionally in a restricted namespace.
- The registry can become a trustworthy identifier assignment (minting) service for institutions and datasets. Identifiers are currently allocated as Universally Unique IDentifiers (UUID) on first registration and, for external use, in addition to the UUID format of the identifier, there are plans to also make them available as Digital Object Identifiers (DOI) .
- The registry can become an identifier resolution service allowing external clients to submit a known identifier and resolve this to the registry assigned identifier.
- The registry can help coordinate distributed system activities by: a) providing preferred technical access points where multiple routes exist, b) offering stable identifiers for registered entities, and c) providing notification services of significant events such as a dataset being registered.
- The registry acts as a technical endpoint monitoring and alerting service to notify technicians of servers going offline.

The GBIF API, in addition to the registry, also covers four other categories: species, occurrence, maps, news feed. The categories are summarised below.

- Registry: Provides means to create, edit, update and search for information about the datasets, organisations (e.g. data publishers), networks and the means to access them (technical endpoints). The registered content controls what is crawled and indexed in the GBIF data portal, but as a shared API may also be used for other initiatives.
- Species: Provides services to discover and access information about species and higher taxa, and utility services for interpreting names and looking up the identifiers and complete scientific names used for species in the GBIF portal.
- Occurrence: Provides access to occurrence information crawled and indexed by GBIF and search services to do real time paged search and asynchronous download services to do large batch downloads.
- Maps: Provides simple services to show the maps of GBIF mobilised content on other sites.
- News feed: Provides services to stream useful information such as papers published using GBIF mobilised content for various themes.

Potentially, the EU BON portal could hold a cached version of the GBIF index to support discovery and access. However, the rich GBIF API offers a more efficient route for federated querying, whether for network entities, species names and other taxonomic information, taxon occurrences, or for map generation. It also seems desirable for any EU BON Metacat instances that output Darwin Core Archives to be published *via* the GBIF network and thus available to EU BON, leaving the EU BON portal to focus on integration across networks.

1.3.7. Marine projects and networks

Marine data flow globally and in Europe is complex and partially intertwined (Fig. 5). Some networks are internationally scientific networks (ICES or OBIS – the Ocean Biogeographic information System) or initiated by UNICEF or IOC (Intergovernmental Oceanographic Commission) such as IODE (International Oceanographic and Information Exchange). Some of these marine networks are connected to other international networks that are collecting and distributing other datasets than purely marine, like GBIF or GEOSS.

Some of the European networks and initiatives are European nodes in these international bodies, like EUROBIS and specific European marine data portals. SeaDataNet (Pan-European infrastructure for ocean & marine data management) is a European network connecting both ICES and OBIS/EUROBIS with an own dataportal. SeaDataNet is part of EU's Sixth and Seventh Framework Programme. SeaDataNet is being followed up, extended and further developed in EMODNET (European Marine Observation and Data Network). More details of these networks can be found in Annex I.

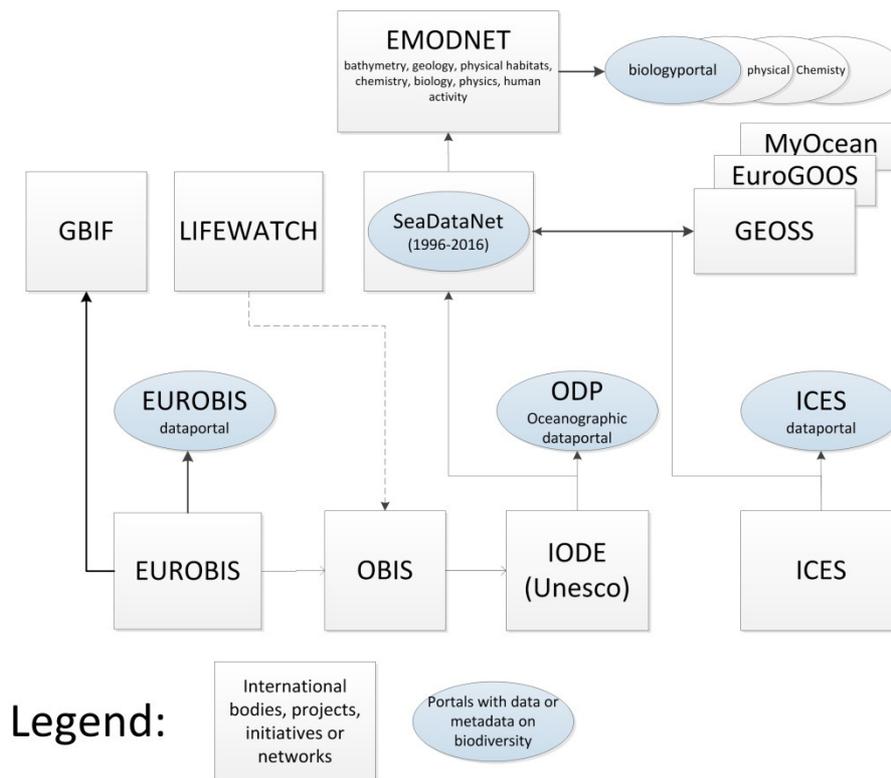


Fig. 5 Marine infrastructure – an overview.

EUROBIS is a European node in OBIS²⁷. EUROBIS currently holds 516 datasets, provided by 159 institutes. Compiling data from different sources collected under different circumstances and with various purposes requires a minimum of standardisation and quality control before sound and useful integration becomes possible. EUROBIS follows a number of international standards and runs a number of quality control procedures on each received dataset, in order to be able to estimate the quality of the provided data and to define the fitness for purpose of the data for our various users.

All datasets within EUROBIS are described in the Integrated Marine Information System (IMIS) developed by the Flanders Marine Institute (VLIZ)²⁸. The advantage of IMIS is that it not only stores the metadata of the datasets, but it can also capture and interlink information on persons, institutes, projects and publications. Within IMIS, existing international standards are taken into account, such as ISO 19115 - the international standard for geographic information - and EDMED, the European Directory of Marine Environmental Data. In addition, EUROBIS use the thesaurus of the Aquatic Sciences and Fisheries Abstracts (ASFA thesaurus) to assign searchable keywords to datasets. EUROBIS uses the World Register of Marine Species (WoRMS) as a standard list for taxonomic names. WoRMS is an authoritative taxonomic list of species occurring worldwide in the marine environment.

EUROBIS aims to centralise biogeographic data on marine species collected by European institutions. The data can either be collected within or outside European waters. As long as the data providing institute is within Europe, EUROBIS acts as the responsible node to make these data available to the OBIS community. EUROBIS receives its data through different pathways:

²⁷ <http://www.EUROBIS.org>

²⁸ <http://www.vliz.be/en>

- Individual providers can send their data to the EUROBIS data management team e.g. through email
- In addition, the two European subnodes - OBIS Black Sea and MedOBIS - provide their data to EUROBIS, thus capturing all the marine European data in one system
- EUROBIS also aims to mobilise the marine distribution data that are stored and hosted in large data networks, foundations and national data centres or institutes. So far, EUROBIS is actively working together to make the data from the International Council for the Exploration of the Sea (ICES) and the Continuous Plankton Recorder (CPR) data available to the scientific community, with regular updates. In the future, EUROBIS intends to expand its connections with other such institutes and organisations
- EUROBIS is in close communication with OBIS-SeaMap. OBIS-SeaMap - the Ocean Biogeographic Information System Spatial Ecological Analysis of Megavertebrate Populations - is a spatially referenced online database, aggregating marine mammal, seabird and sea turtle observation data from across the globe. Datasets from OBIS-SeaMap containing European data are also made available to users.

The EUROBIS system and its data are part of two large European initiatives, EMODnet Biology and LifeWatch. A strong collaboration exists, resulting - amongst others - in the active growth of available datasets within EUROBIS.

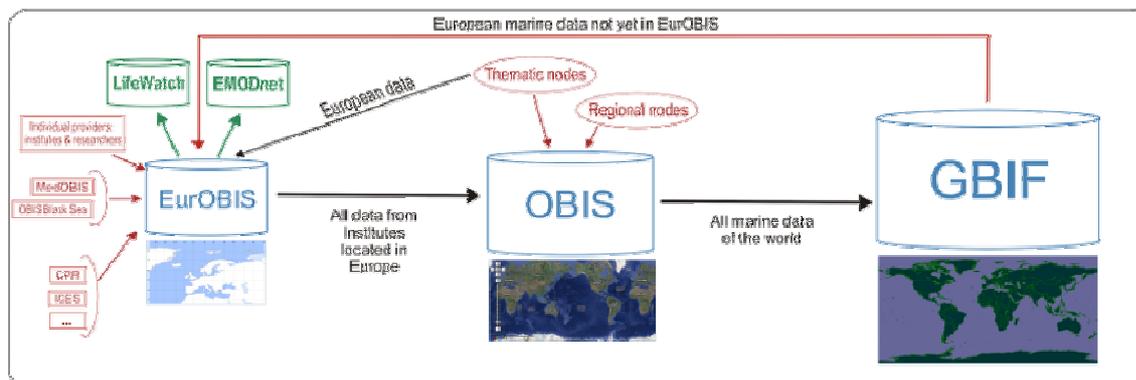


Fig. 6 Data flow from EurOBIS to GBIF.

Together with the host institutes or organisations of the other OBIS nodes, the Flanders Marine Institute (VLIZ) has committed to a sustained support of OBIS. This has resulted in making distribution information of European marine species freely available online, and in transferring these data to OBIS on a regular basis. EUROBIS is currently one of the largest data providers to OBIS. In its turn, OBIS publishes its data through GBIF (Fig. 6). OBIS is recognised as the marine thematic sub-network of GBIF. There is also a 'back-flow' of data from GBIF to EUROBIS. If European marine data is available through GBIF but not through EUROBIS, GBIF will notify EUROBIS to add these data to its system and make the EUROBIS inventory more complete.

The International Council for the Exploration of the Sea (ICES)²⁹ is a global organisation for enhanced ocean sustainability. It is a network of more than 4000 scientists from almost 300 institutes, with 1600 scientists participating in activities annually. The ICES Secretariat has been based in Copenhagen, Denmark, since 1902. ICES has a well established Data Centre, which manages a number of large dataset collections related to the marine environment. The datasets are organised so that it is easy to find what you are after, whether you are interested in a particular geographic area, an effect on the environment or a group of species.

The European Directory of Marine Environmental Data (EDMED) is a comprehensive reference to the marine data sets and collections held within European research laboratories, so as to provide

²⁹ <http://www.ices.dk/>

marine scientists, engineers and policy makers with a simple mechanism for their identification. It covers a wide range of disciplines including marine meteorology; physical, chemical and biological oceanography; sedimentology; marine biology and fisheries; environmental quality; coastal and estuarine studies; marine geology and geophysics; etc. Data sets are described in EDMED irrespective of their format (e.g. digital databases or files, analogue records, paper charts, hard-copy tabulations, photographs and videos, geological samples, biological specimens etc).

1.4. General requirements for the architecture

EU BON's main mission is to build a substantial part of the GEO BON network, focusing on advancing the technological/informatics infrastructure and establishing new integration techniques for GEO BON. The EU BON information architecture is obliged to build on the work of existing information network structures to allow discovery and access to biodiversity data. EU BON test sites already included in the project will interoperate and share their data sets through the EU BON platform using supported services.

The Service Oriented Architecture (SOA) model which has achieved "best practice" status within the Open Geospatial Consortium (OGC) is at present the most acceptable architecture for EU BON. In an SOA, different functionalities are packaged as component services that can be orchestrated together for specific tasks. It is proposed to implement EU BON using an Enterprise Service Bus (ESB) to connect external data sources using various SOA standards (WSDL³⁰, SOAP³¹, REST³² and BPEL³³, among others). The use of an ESB facilitates the interactions among data sources, working in a message-centred interaction and providing the ability to orchestrate web services through the use of workflow handling technology (e.g. Kepler³⁴, Taverna³⁵). Within the European context, EU BON should incorporate and build on the work of ALTER-Net, LifeWatch and other existing frameworks and networks.

Recommendation 9. EU BON should adopt the Service Oriented Architecture model.

The architecture will be focused on the Service-Oriented Architecture paradigm. It will be based on web services, using WSDL/SOAP and RESTful web services. There are/were other ways to implement the SOA approach rather than using XML or REST web services, e.g., using Java Message Service API or CORBA. However, WSDL/SOAP is the W3C recommended standard, and REST is a *de facto* standard that uses other W3C standards.

Scalability, access, security, user concurrency and data reliability must be considered. For scalability, it is expected that tens of thousands of data sources will ultimately be integrated. They will be hosted in a smaller number of data repositories. In fact, it is likely that during the lifetime of the EU BON project, only a small number of new data repositories will be established by the project itself, such as for test sites and for hosting data centrally or at few places. All other repositories will be those of existing networks.

GEOSS Data Sharing Principles shall to be followed. There are three provisions:

- *There will be full and open exchange of data, metadata and products shared within GEOSS, recognising relevant international instruments and national policies and legislation;*
- *All shared data, metadata and products will be made available with minimum time delay and at minimum cost;*
- *All shared data, metadata and products being free of charge or no more than cost of reproduction will be encouraged for research and education.*

Although open access is expected, there can be exceptions. Certain data may be set under embargo, or be available for authorised users only, or there may be a cost involved. In order to maximise the

³⁰ Web Services Description language (WSDL); <http://www.w3.org/TR/wsdl20/>

³¹ Simple Object Access Profile; <http://www.w3.org/TR/soap12-part1/>

³² Representational State Transfer; <http://www.ibm.com/developerworks/webservices/library/ws-restful/>

³³ Business Process Execution Language; <http://docs.oasis-open.org/wsbpel/2.0/OS/wsbpel-v2.0-OS.html>

³⁴ <https://kepler-project.org/>

³⁵ <http://www.taverna.org.uk/>

amount of data that is being shared, it is critical that the data owners will be enabled to control release of data along these lines. They also need to be informed when their data is being used. Therefore, a system of authentication and authorisation of users is necessary to access certain data. However, it should not be mandatory to log in for all uses.

GEOSS Data Core is a distributed pool of documented datasets, contributed by the GEO community under full, open, or unrestricted access principles, and using the Data Core should not require logging into a system. EU BON is developing more detailed data sharing guidelines, which will support these requirements.

Data reliability needs to be considered. The software architecture must be prepared to be clustered or virtualised if it is needed, minimising the efforts needed to achieve further improve activities.

User concurrency needs to be ensured, by high availability and load balancing server clustering abilities. At a software architecture level, this applies to the Enterprise Service Bus and the application server in which the main application and the ESB will be deployed.

To fulfil the requirements and recommendations of the INSPIRE directive, open source technologies and international standards and techniques will be adopted for the construction of the EU BON Portal software development.

Quality assurance needs to be considered, applying methodologies, code quality analysis, automated testing, etc.

1.5. Functional requirements of EU BON components

1.5.1. Requirements for the EU BON Portal

The EU BON Portal's first priority is to connect and access data from GBIF, LTER, testing sites databases, and other data providers, allowing users to search biodiversity data through a public web interface. The search engine will look for this information by querying each data provider or aggregator connected through the Enterprise Service Bus. As a network of networks, EU BON will not connect directly to the original sources of data, if these are available through existing aggregation services. Instead, the EU BON Portal will use already aggregated data. This means, for instance, connecting to the GBIF index through its API or by mirroring.

The software architecture has to provide the foundations and structure on which to deploy functional building blocks, interconnect them and implement the functionalities summarised by the following uses cases:

- User authentication and authorisation:
 - Sign up.
 - Log in.
 - Roles and permissions.
- Upload data: dataset and related metadata.
 - Dataset values by XLS, CSV or TXT import.
 - Data and metadata import from standard formats: DwC, EML, ABCD.
 - Edit data and metadata (authorised users only).
- Search biodiversity occurrence data (returning data from any connected provider):
 - By taxon vernacular name or scientific name.
 - Filters: dates, geospatial, etc
 - Enable advanced searches.
 - Save users searches.
- Visualise data and metadata:
 - Grids, forms and maps.
 - Charts, statistics and reports.
- Export and download:
 - Dataset: XLS, CSV, plain text, etc.
 - Data and metadata: EML, DwC, ABCD, etc.

- Execute pre-built workflows:
 - Based on available remote services.
 - Using datasets exported from above functions within the portal.
 - Using background data from GEOSS services.
 - Produce statistics and estimate EBVs from related data.

Functional requirements summary for the portal include the following:

- The EU BON Portal should serve as an integration hub of multiple biodiversity networks, allowing users to discover and access existing information within datasets and metadata, as well as providing new information to a central database.
- Therefore, the EU BON Portal must be able to present through browsing and searching resource metadata from other portals, in particular from GBIF and LTER.
- The EU BON Portal will not aggregate or index raw data which has already been aggregated by other portals. However, it may add other data to these indexes, if necessary. The mechanism of doing this will be determined later in MS251.
- Data and metadata must be discoverable in terms of geo-positioning or variables/species identification.
- Users should be able to edit uploaded metadata using a web form based EML editor.
- Users should be able to annotate data uploaded by others. Annotations could trigger notifications.
- The portal must provide a dataset loader that should support importing data from Excel or CSV files.
- The system has to provide a metadata import/export engine, using data/metadata exchange formats, i.e. Darwin Core, EML and ABCD.
- The INSPIRE directive recommends using EU-Nomen to provide species names³⁶. The next option might be provided by EUNIS, and in the third place by Natura 2000. Although EU-Nomen includes European Register of Marine Species, it is possible to broaden this information connecting to the World Register of Marine Species (WoRMS³⁷). Both EU-Nomen and WoRMS provide WSDL web services.
- The EU BON Portal must provide a REST API of web services based on the EDIT Platform for Cybertaxonomy in order to allow other portals or services interact with EU BON information.
- The portal has to provide user authentication and authorisation and a public search interface as well.
- The portal needs to visualise the Essential Biodiversity Variables through time and space. It should make transparent and offer drill-in into the data flows and data sources used as basis. It should provide provenance of the modelling steps that have been applied.

1.5.2. Requirements for the registry and semantic mediation

It is well known that one does not get very far with the service-oriented architecture without a registry. The overview of GEOSS above shows clearly that biodiversity data is not well covered in the GEOSS Portal because the leading biodiversity registries are not included. EU BON needs to address this. For the functioning of the EU BON Portal, registry functions are needed as well.

The EU BON registry needs to have, at the minimum, the following requirements for the data sources and services it keeps track of:

- Identifiers
- Access points
- Owning organisation and contacts
- Information on last update

It would also be useful to know of the following:

³⁶ Data Specification on Species distribution – Draft Technical Guidelines

http://inspire.jrc.ec.europa.eu/documents/Data_Specifications/INSPIRE_DataSpecification_SD_v3.0rc3.pdf

³⁷ World Register of Marine Species: <http://www.marinespecies.org>

- Types of data, such as taxonomic, specimen, occurrence, monitoring, raster, etc.
- Data standards used and mappings
- Size and number of records
- Spatial, temporal, and taxonomic coverage
- Access rights and terms of use
- Whether the data been cached elsewhere

The registry could be initially built by merging the available GBIF and LTER registries. These need to be synchronised periodically. Later on the EU BON registry could be expanded with the EDMED marine directory, LTER-Europe DEIMS, and EUMON Database³⁸ of monitoring networks.

The registry would also contain descriptions of those data repositories and services, which will be registered directly with EU BON. These include test sites and locally hosted datasets.

Once the combined registry is in place, it would need to be registered among the registries of the GEOSS Portal. This would make all shared biodiversity datasets available through the GCI.

The semantic mediation layer of EU BON needs to keep track of the range of the interoperability protocols and standards that the network can deal with. These functions can be adopted from the GEOSS Components and Services Registry, and we need to comply with its requirements. Using the GEOSS services for this also is a possibility. Other possibilities include the BiodiversityCatalogue of the BioVeL project³⁹.

Whether EU BON should adopt a brokering architecture (Nativi *et al.* 2012) should be discussed. An ESB is actually a message-broker approach that facilitates a common layer between servers/providers and clients. However, it does not work exactly in the way described by Nativi *et al.*, because an ESB works using a process-level approach, orchestrating services. But in that sense we can say that EU BON will be using a brokering approach.

Registration and discovery, including the capabilities of each biodiversity data provider connected to EU BON platform, will be achieved in the registration service through the ESB. The ESB will act as the web service orchestration layer, being capable of interconnecting several data providers, their search engines and databases through messages, operate with them and apply mediation processes. Data providers' capabilities will be established and registered at the time each one is connected to EU BON platform. Unlike for the GBIF infrastructure, EU BON will not be focused on indexing data or crawling the network. Therefore it is not necessary to include automatic discovery processes.

1.5.3. Requirements for data/metadata providers and their services

Each data provider has to expose its data and metadata through web services that will be connected to EU BON through the ESB. In the cases that a provider may not comply with this requirement, it is possible to connect databases, files or even third party services through binding components provided by the ESB implementation.

Data and metadata management through existing biodiversity networks or portals will be compatible with EU BON approach. In first place, EU BON will be able to connect to the following services:

- GBIF central index
- DataONE/KNB central index
- LTER-Europe DEIMS central index

There are other central repositories that EU BON may also need to connect to, such as the EUMON database and Scratchpads⁴⁰ datasets.

Each biodiversity network has its own publishing toolkits or applications. The main ones that EU BON needs to work with are:

³⁸ http://eumon.ckff.si/about_daeumon.php

³⁹ www.biodiversitycatalogue.org

⁴⁰ <http://scratchpads.eu/>

- GBIF IPT repositories
- LTER Metacat repositories

Details of these and requirements for possible enhancements for these will be considered in MS231.

A very basic, almost a fundamental question, is the following: *"How does EU BON envisage its own central database?"* And related to this, where in information architecture do we see the function of this database and the associated publishing facility (repository with a dataset loader, and a dataset/EML editor)? This database would not be a mere cache of the external databases as these would either be directly linked to or separately cached. But then, the question is, which datasets do we envisage to be published through EU BON? Do we see any major partners which are not already making data available through GBIF and/or LTER and if not, wouldn't it make more sense to ensure that these partners can join existing networks and EU BON accesses the data through external services?

Fact is that most biodiversity data is still *not* being shared through any of these networks, but remain proprietary commodities. There are various reasons for this, such as limitations of data standards, lack of resources and funding, lack of technical support, and lack of understanding of the benefits of data sharing.

One obstacle is the lack of a trusted, controlled, but shared environment, where data custodians can control release of their data. EU BON needs to consider building such an environment. This would be a large scale hosting environment for distributed, exported datasets, such as DataONE. Such a facility could be built using the current Metacat as repository software. However, EU BON would need to introduce strong support for biodiversity data standards in that environment. This is not yet available anywhere. A dataset editor or mapping tool in the EU BON Portal could take care of this function. A strong outreach activity would also be needed towards existing monitoring networks that are not yet sharing their data. These are tasks for the EU BON Helpdesk.

It is therefore recommended that EU BON consider building such a repository or repository infrastructure, enhanced with controlled use of biodiversity data standards, and in cooperation with DataONE.

Metacat is a static repository, basically an archive system. There may also be need to connect to online databases from sensor networks and citizen observatories where updates are continuous. This will be the realm of LifeWatch. However, for the time being there are no concrete plans available that could be considered here.

Integrating species-level occurrences from the GBIF portal with ecosystem information from LTER / ILTER sites (DataONE, KNB, SEAON) will be achieved through a mapping process of Metacat networks/metadata repositories. Metacat uses Morpho⁴¹, in a similar way of GBIF IPT, to enable users to create and manage EML metadata and to share and publish those metadata and their associated data with others, thus provides a cross-platform application for accessing and manipulating Metacat on the network. When mapping Metacat data, the following notions were considered:

- Mapping the data to an extended DwC schema
- IPT embedded into Metacat
- Extending IPT to support Metacat mapping
- Mapping data series as mediated Web Context Documents⁴²
- Tagging/ annotating LTER data sets in a semantic manner in order to enhance their discoverability

⁴¹ <https://knb.ecoinformatics.org/morphoportal.jsp>

⁴² <http://www.opengeospatial.org/pressroom/pressreleases/1851>

1.6. Component architecture

This and the following sections on application and runtime architecture are brief on purpose. They will be elaborated in the MS251 report when this document has been thoroughly reviewed and discussed with the stakeholders.

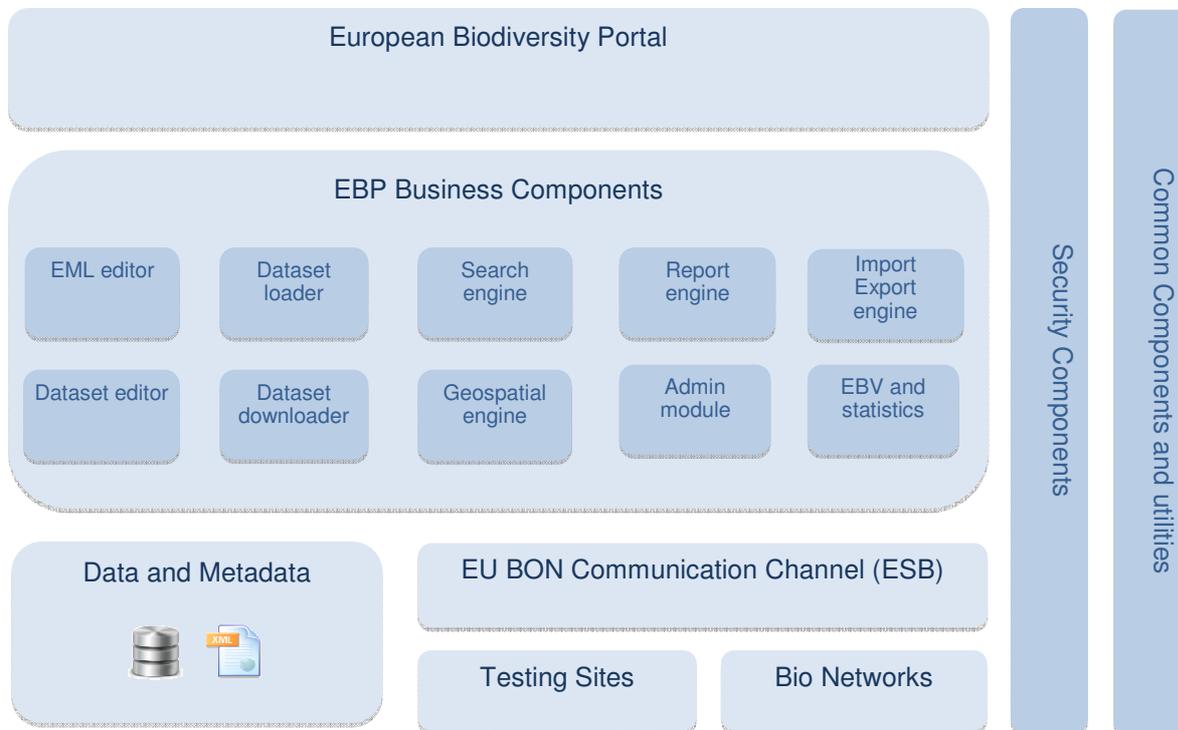


Fig. 7 System architecture: components and containers.

The system architecture (Fig. 7) is divided into the following layers and containers regarding different levels of abstraction:

- European Biodiversity Portal: public and administration websites, with different user roles and functionalities. The portal is considered the main user interface, being an implementation client for the broader middleware architecture.
- Business components, including:
 - Search engine: provides search capacities for biodiversity information across different sources and within the EBP database as well.
 - Dataset loader, downloader and editor: provide interaction with all kind of input or output information.
 - EML Editor: EML based metadata editor.
 - Geospatial engine: to present geospatial information to the users through OGC standard formats.
 - Report engine: to provide reports to the users or administrators.
- EBP Data and Metadata repository.
- Enterprise Service Bus to handle message exchanges between testing sites and biodiversity networks, relying on SOA based standards and service orchestration.
- Transversal components: security, scalability, common utilities, etc.

Fig. 8 illustrates the search information use case. Users interact with the EBP website, asking for some information. The portal, using the search engine, queries the information, first in the EUBON repository and later within different shared sources of datasets, which are accessible to the portal through the ESB by web services. Later, the information is integrated and made available to the user by the dataset downloader component.

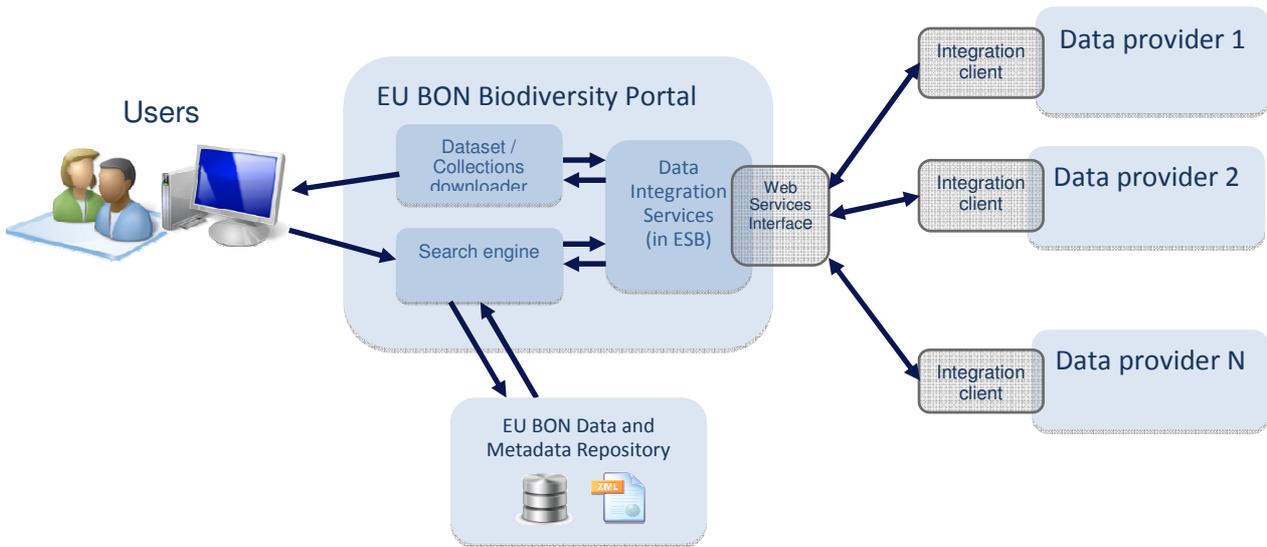


Fig. 8 Search use case.

1.7. Application architecture

The proposed solution is based on Java EE 6, using a Model-View-Controller architectural design pattern. Fig. 9 exemplifies the different frameworks and technologies implementing the different layers of functionality:

- Java Server Faces to create the website
- Enterprise Java Beans to implement business models
- JPA (Java Persistence API) as Object-Relational Mapping, to obtain persistence classes as a representation of the database model in terms of business logic.
- WSDL and SOAP to connect web services. We will also include REST web services as external API.

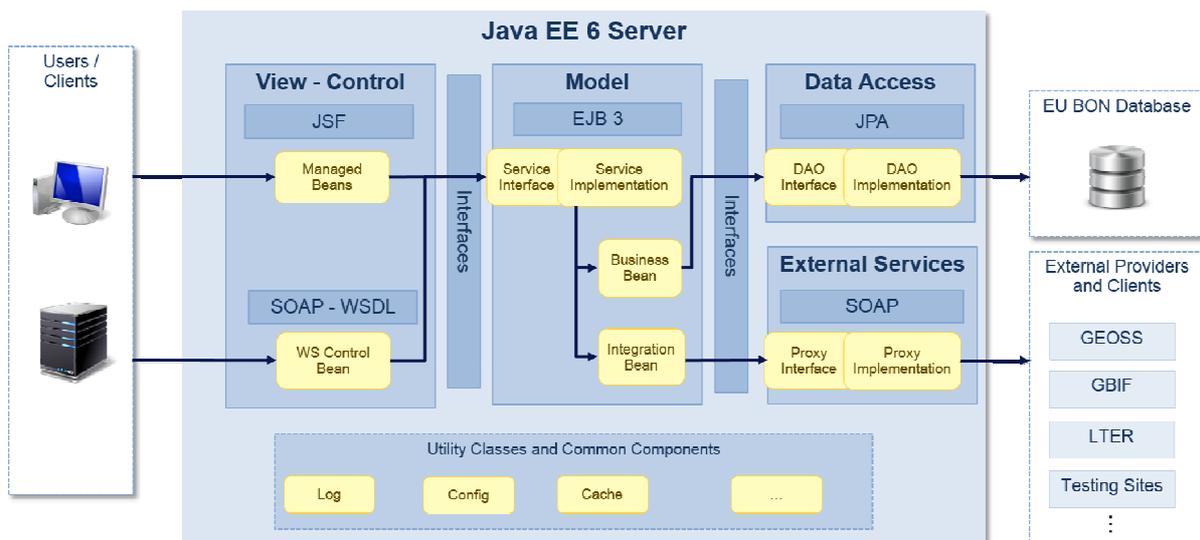


Fig. 9 Application Architecture.

1.8. Runtime architecture

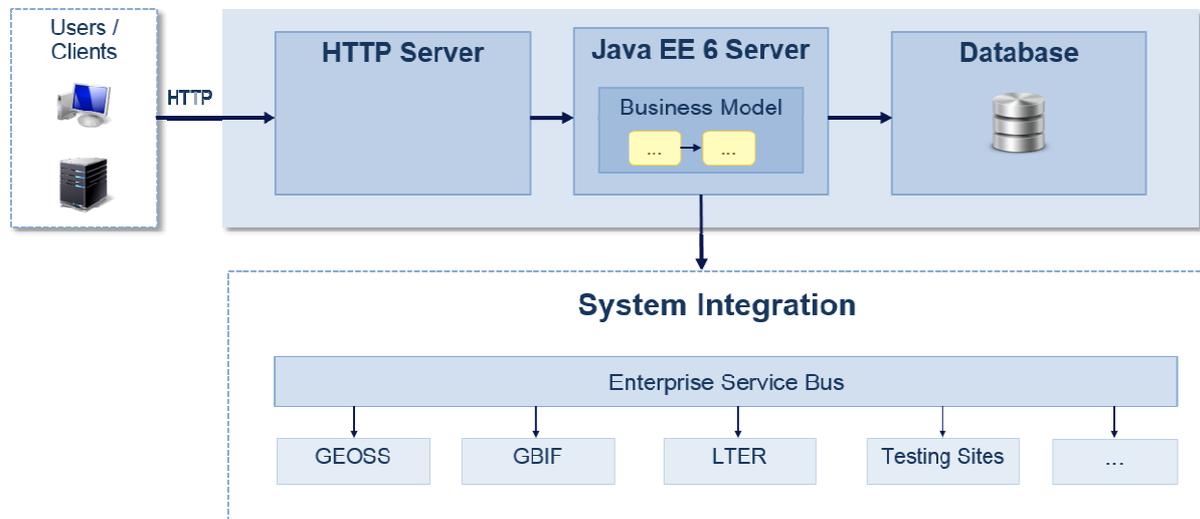


Fig. 10 Runtime architecture.

The runtime architecture refers to the servers and the management facilities that will interact with the execution of the functional use cases:

- User and machine clients interact with the portal using HTTP protocol, connecting to the HTTP server.
- HTTP Server acts as a proxy for Java EE 6 Application Server, which stores all the business logic and models.
- The Java Application Server facilitates connectivity with the Database Management System and with the system integrator (ESB).
- The Enterprise Service Bus provides communication services for the different data providers, orchestrating web services and returning messages to the Java Application.

2. Review of data standards

2.1. Introduction

The goal of EU BON is to provide a new open-access platform for sharing, integrating and analysing biodiversity data. As data will originate from many sources, adoption of standards is crucial to enable interoperability and this section of the report provides a review and guidelines for using standards within the EU BON platform. To begin, the architectural design underpinning the informatics infrastructure of the platform was outlined in the previous chapter, as this provides the essential context for identifying the standards that will be required to realise the goal of EU BON. In its role as the European contribution to the Group on Earth Observations Biodiversity Observation Network (GEO BON)⁴³, EU BON will be informed by the GEO BON manifesto (see page 12). This entails consideration of the GEO BON Detailed Implementation Plan⁴⁴ and the GEOSS Common Infrastructure (GCI) and a model based on a Service Oriented Architecture (SOA) that enables interoperability across disparate, loosely connected systems. More specifically, based on the recommendation of the EU BON Informatics Task Force, EU BON will also adopt the DataONE architecture⁴⁵ for the platform by implementing a European coordinating node and network of member nodes. It will also promote the uptake of the GEO BON Essential Biodiversity Variables (EBVs) which define a set of measurable parameters that can support indicators of progress for meeting both EU and Aichi targets relating to the status of biodiversity loss. The parameters associated with the EBVs help to constrain and define data requirements and one of the main tasks for EU BON is to develop an integrated system of systems for automating the flow of data from the original field observations through intermediate aggregating and other processing services that generate the EBVs which are used in the composition of the biodiversity indicators. Coordination is achieved through a portal with a central registry that provides unified discovery and access to disparate data sources, services, applications, etc.

In this chapter, we review the data standards required for implementing EU BON. The DoW sets out the requirements as follows:

Task 2.2 Improving data standards and interoperability

Starting from the GEO BON Detailed Implementation Plan and the architecture (task 2.1) as well as relevant European projects (ALTERNet, EBONE, LifeWatch), review the state-of-the-art and needs for improvement of the current data standards of TDWG, OGC, BioCASE, GBIF, LTER-Europe, PESI, and INSPIRE. Consider how the available protocols and mechanisms for interoperability can be best used for integrating different data layers (i.e., genetic data, primary occurrence data, monitoring data, ecological measurements, remote sensing data) in the European context. Consider reasons for heterogeneity of biodiversity information and make recommendations for use of standards by the various networks. (Lead GBIF; UTARTU, UEF, CSIC, Pensoft, MRAC, Plazi, GlueCAD, INPA, IBSAS, NBIC, TerraData; Months 4-51)

2.2. The European context

EU BON has been charged with addressing data integration, particularly within the European context. In this section, we introduce the INSPIRE directive and provide an overview of several key EU funded initiatives and projects relating to biodiversity informatics, excluding networks and frameworks such as ALTER-Net and LifeWatch which have been mentioned under the EU BON architecture (Chapter 1).

The establishment of the Group On Earth Observations (GEO) and the numerous intergovernmental initiatives (e.g., GBIF), projects and plans relating to biodiversity information networks (GEOSS, GEO BON, EU BON) reflect the growing recognition that an informed decision-making process

⁴³ <http://www.earthobservations.org/geobon.shtml>

⁴⁴ http://www.earthobservations.org/documents/cop/bi_geobon/geobon_detailed_imp_plan.pdf

⁴⁵ <http://mule1.dataone.org/ArchitectureDocs-current/>

benefits from access to, and use of, a wide range of related data (biological, environmental, economic, etc.) within a common framework.

The European Union has invested in multiple international research efforts, covering projects, data infrastructure and supporting mechanisms, through the framework programs (FP5, FP6 and FP7), the European Science Foundation (ESF) and the European Cooperation in Science and Technology (COST). Since 1998, there have been 324 EU research projects focusing on biodiversity and ecosystems (see full list⁴⁶). The duration ranges from 1-45⁴⁷ years, with the majority of projects lasting from 2-4 years, with a peak for 3-year projects (Fig. 11).

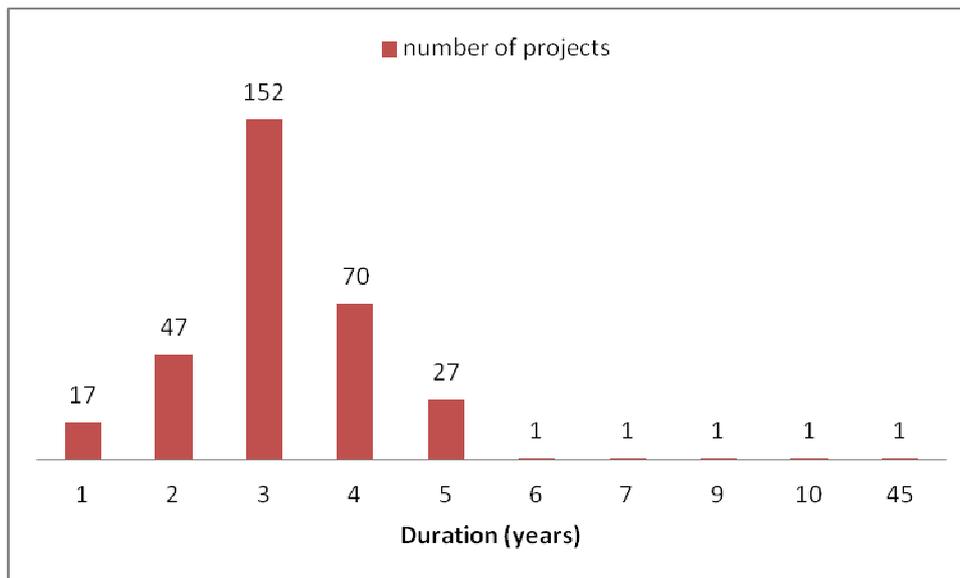


Fig.11 Most common duration of EU research projects about biodiversity and ecosystems. Source: <http://www.eea.europa.eu/>. Copyright holder: European Environment Agency (EEA).

The focus of these projects broadly covers ten main categories, which are seldom mutually exclusive. Fig. 12 illustrates the number of EU projects that fall within each category. The output generated by these projects is initially stored on project websites. However, as project funding periods come to an end, there is a need for sustainable information management services. Several websites now provide access to this information, including the Biodiversity Information System for Europe⁴⁸ (BISE), and the European Environment Agency⁴⁹ (EEA). BISE is a partnership between the European Commission (DG Environment, Joint Research Centre and Eurostat) and the EEA. The EEA aims to ensure that decision-makers and the general public are kept informed about the state and outlook of the environment. Furthermore, in early 2005, a joint meeting was held between eight major EU Networks of Excellence⁵⁰ to propose a European research infrastructure (LifeWatch) that would enable large-scale EU-wide co-operation on the frontiers of biodiversity science (see section 1.3.4).

⁴⁶ <http://www.eea.europa.eu/data-and-maps/data/eu-research-projects-on-biodiversity/projects/research/view>

⁴⁷ The longest EU-funded project, i.e. 45 years, is the FP5 project entitled: "Long-term comparative study of oligotrophication process in four European lakes (LCSOOPIFEL)" which started in 2001 and is foreseen to end in 2046.

⁴⁸ <http://biodiversity.europa.eu/>

⁴⁹ <http://www.eea.europa.eu>

⁵⁰ Terrestrial Biodiversity – ALTER-Net, European Distributed Institute of Taxonomy – EDIT; Marine Biodiversity and Ecosystem Functioning – MarBEF; Marine Genomics Europe – MGE; Ocean Ecosystems Analysis EUR-OCEANS; Infrastructure network SYNTHESYS; Biological Collection Access Service for Europe – BioCASE; European Network for Biodiversity Information – ENBI

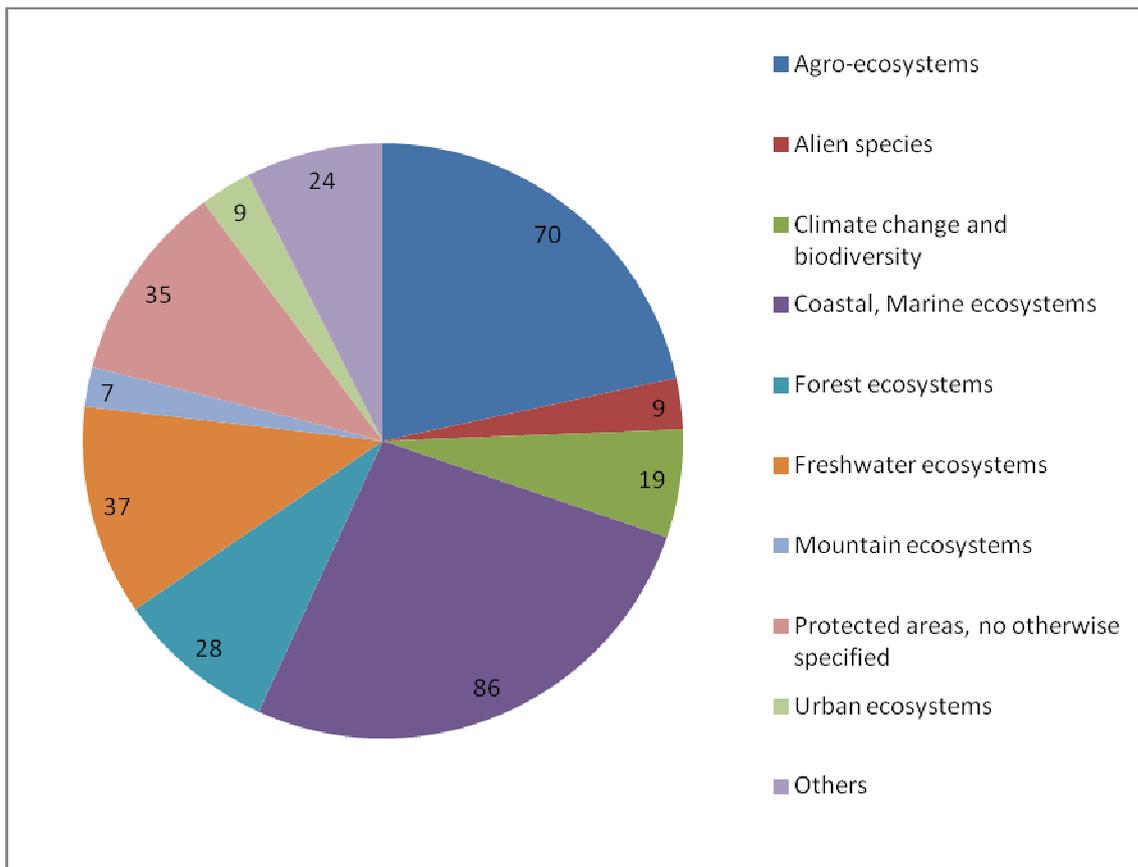


Fig. 12 Focus of EU research projects about biodiversity and ecosystems. Source: <http://www.eea.europa.eu/>. Copyright holder: European Environment Agency.

2.2.1. Biodiversity data management at EU institutions

European Environment Agency

The EEA's mandate is to help the Community and member countries make informed decisions about improving the environment, integrating environmental considerations into economic policies and moving towards sustainability, and to coordinate the European environment information and observation network (Eionet⁵¹). Eionet provides timely and quality-assured data, information and expertise to assess the state of the environment in Europe and the pressures acting upon it. This enables policy-makers to decide on appropriate measures to protect the environment at national and European level and to monitor the effectiveness of policies and measures implemented. The EEA also has developed the EUNIS Database to manage data on species, habitats, and sites across Europe⁵². SEBI is an initiative to streamline data flows for the European biodiversity indicators⁵³, which has a similar background to the current thrust to develop the Essential Biodiversity Variables.

In the EEA, data sets, tables and elements in the data dictionary are defined by a set of attributes, the core set of which corresponds to the ISO 11179 standard for describing data elements. The criteria requested by the EEA for the integration of data into the Environmental data centres (see below) and their hosting infrastructure include the following:

⁵¹ <http://www.eionet.europa.eu/>

⁵² <http://eunis.eea.europa.eu/>

⁵³ <http://biodiversity.europa.eu/topics/sebi-indicators>

- Use of open standards and formats, which implies that data must allow processing with freely available tools
- Microsoft Access Databases should be delivered in mdb Access 2002-2003 file format
- Data set structure and fields should be well-documented
- Methodology for production of the data should be well-documented
- INSPIRE compliant metadata for spatial data; guidelines are available⁵⁴
- Guidance for countries on geospatial data reporting for national experts⁵⁵

Guidelines⁵⁶ are available for co-operator providing data for maps, graphs and metadata.

Eurostat

Eurostat processes and publishes comparable statistical information at the European level through use of a common statistical ‘language’ that embraces concepts, methods, structures and technical standards. Data is collected, verified and analysed by Member States’ statistical authorities and sent to Eurostat. Eurostat’s role is to consolidate the data and ensure they are comparable, using harmonised methodology. To this end, Eurostat works with Member States to define common methodologies when gathering national data.

JRC – Institute for Environment and Sustainability

The Institute for Environment and Sustainability (IES) is one of seven scientific institutes that constitute the Joint Research Centre⁵⁷ (JRC), which is a Directorate-General of the European Commission that provides customer-driven scientific and technical support for the conception, development, implementation and monitoring of European Union policies. Made up of eight scientific Units, the Institute is engaged in the following fields of activity:

- Sustainable Use of Strategic Resources: Water, soils, forests, air, land, biodiversity
- Sustainable Agriculture and Rural Development: Crop production, food security
- Climate Change and Air Pollution: Reduction of greenhouse gas emissions
- Environmental Risks and Natural Hazards: Fire, droughts, floods, desertification
- Environmental Dimension of Development Cooperation: Focus on Africa
- Environmental Monitoring and Information Systems: Global Monitoring for Environment and Security (GMES) and INSPIRE
- Sustainability Assessment: Integrated socio-economic and environmental assessments; indicator development.

Environmental Data Centres

Three European Union (EU) bodies managing information on environmental pressures and the state of Europe's environment (Eurostat, EEA and JRC) are currently setting up Environmental Data Centres in order to centralise and more efficiently present data on ten specific topics.

Eurostat is responsible for two of these Environmental Data Centres that cover three topics:

- Environmental Data Centres on Natural Resources and Products (EDCNRP)⁵⁸
- Environmental Data Centre on Waste⁵⁹

Other Environmental Data centers are managed by the EEA:

- European air pollution data centre⁶⁰
- European Biodiversity Data Centre (BDC)⁶¹

⁵⁴ https://taskman.eionet.europa.eu/projects/sdi/wiki/Cataloguemetadata_guidelines

⁵⁵ <http://www.eionet.europa.eu/gis/nationaldeliveries>

⁵⁶ <http://www.eionet.europa.eu/gis/>

⁵⁷ <http://ec.europa.eu/dgs/jrc/index.cfm>

⁵⁸ <https://webgate.ec.europa.eu/fpfis/mwikis/edcnp/>

⁵⁹ <http://epp.eurostat.ec.europa.eu/portal/page/portal/waste/introduction>

⁶⁰ <http://www.eea.europa.eu/themes/air/dc>

The BDC contains information reported by European countries which is maintained or hosted by the EEA. The BDC will be developed further within the framework of the Biodiversity Information System for Europe (BISE) and the agreement entered into by DG Environment, DG Eurostat, DG Joint Research Centre and the EEA. The BDC uses web-based tools allowing access to quality-controlled spatial and tabular data sets, interactive maps, static maps and graphs, statistics and indicators. Supporting documents, code lists and standards are also available for the data sets provided.

- European Climate Change Data Centre⁶²
- Environmental Data Centre for Land Use⁶³
- European Water Data Centre⁶⁴

The JRC-IES manages:

- European Soil Data Centre (ESDAC)⁶⁵
- European Forest Data Centre (EFDAC)⁶⁶

The Shared Environmental Information System (SEIS)⁶⁷, launched in January 2013, is the outcome of an on-going activity of the European Commission and the EEA, and supported by the EU Member States, aimed at modernising and simplifying the availability, exchange and use of data and information required for the design and implementation of environment policy. Within this approach the current, mostly centralised systems for reporting and information (including the exploitation thereof) are progressively replaced by systems based on access, sharing and interoperability. At the Pan-EU level, GEO and GEOSS are part of the ongoing efforts to build SEIS.

In order to integrate information systems, EU BON will be required to link into existing frameworks and support mechanisms for biodiversity information from on-ground and remote sensing data sources. Annex I (Related European projects) provides a listing of relevant documents and information from EU projects operating in similar research domains. EU-BON may benefit through learning, integration and association with these and other complementary research networks. The list of complementary projects is non-exhaustive. Information was collated from project websites; for more details on each project's objectives and output, please visit the project website.

2.2.2. Biodiversity standards in the INSPIRE Directive

Directive 2007/2/EC of the European Parliament and of the Council of 14 March 2007, known as the INSPIRE Directive, establishes an infrastructure for spatial information in Europe to support Community environmental policies, and policies or activities which may have an impact on the environment. The Directive addresses 34 spatial data themes⁶⁸ needed for environmental applications, with key components specified through technical implementing rules including the development of data models⁶⁹ and data specifications⁷⁰ based on the models to support interoperability of spatial data sets and services. One of the themes is Environmental Monitoring Facilities which supports the location and operation of environmental monitoring facilities including observation and measurement of emissions, of the state of environmental media and of other ecosystem parameters (biodiversity, ecological conditions of vegetation, etc) on behalf of public authorities.

The thematic areas affected by the Directive are listed in the Annexes of the Directive including the three themes related to biodiversity in Annex III: i) Species Distribution, ii) Habitats and Biotopes and

⁶¹ <http://www.eea.europa.eu/themes/biodiversity/dc>

⁶² <http://www.eea.europa.eu/themes/climate/dc>

⁶³ <http://www.eea.europa.eu/themes/landuse/dc>

⁶⁴ <http://www.eea.europa.eu/themes/water/dc>

⁶⁵ <http://esdac.jrc.ec.europa.eu>

⁶⁶ <http://efdac.jrc.ec.europa.eu>

⁶⁷ <http://ec.europa.eu/environment/seis/>

⁶⁸ <http://inspire.jrc.ec.europa.eu/index.cfm/pageid/2/list/7>

⁶⁹ <http://inspire.jrc.ec.europa.eu/index.cfm/pageid/2/list/datamodels>

⁷⁰ <http://inspire.jrc.ec.europa.eu/index.cfm/pageid/2/list/2>

iii) Bio-geographical regions. All these biodiversity data specifications and recommendations are thoroughly described in comprehensive documents.

The implementation programme for INSPIRE will decide upon the proposed implementation rules in October 2013. A roadmap has already been agreed for member states to provide all data sets, including *species distribution* and *habitats and biotopes*, within a common infrastructure that includes a European map portal for search, view and download. By October 2015, newly collected and extensively restructured Annex II and III spatial data sets must be available. By October 2020, other Annex II and III spatial data sets must be available in accordance with the implementation rules for Annex II and III.

Species distribution

The INSPIRE Directive defines *Species Distribution* as geographical distribution of occurrence of animal and plant species aggregated by grid, region, administrative unit or other analytical unit [Directive 2007/2/EC]. The definition refers to a distribution of occurrence of a given species and is not intended to cover the raw field observation data. The definition interprets occurrence as the spatial representation of a species at a specific location and a specific time period, rather than being equivalent to an observation. Due to some use cases of local authorities and scientists, however, an extended model is provided with the possibility to link to the original observations used as sources for the aggregations.

The term “aggregated” most commonly means to form into a class or cluster. It is closely related to (but not synonymous with) the term amalgamated, which means to combine to form one structure. Both terms are used throughout the data specification as being suitable for describing the process of converting raw observations into a distribution of occurrence. As a result of the definition, distributions may be represented in a wide range of formats such as points, grid cells at different scales, or polygons of specifically defined areas (analytical units).

Species distribution data specification

The Species Distribution data specification⁷¹ is mainly divided into three sections: the Data Set description, the Distribution Information description and the Source Information description.

As most of the thematic community currently appear to encode their species distribution data as feature collections, i.e., sets of individual features such as polygons represented in a data set, the model is based on distribution units and collections of those constituting a distribution. Each unit specifies a referenceSpeciesScheme which refers to a choice of three widely known reference lists and a referenceSpeciesID which refers to an ID from that reference list for the given species of interest. EU-nomen⁷² is the preferred reference list to be used. If a taxon is listed in EU-nomen, this reference must be used as first choice. If it is not listed in EU-Nomen, the second choice is EUNIS⁷³, if not EUNIS, Natura2000⁷⁴ can be used.

An extended schema allows for associating metadata with each unit *via* the featureType

SourceInformation. There exist a multitude of approaches and methodologies both for collecting data on species observations and actually deriving the species distribution from these. In order to ascertain whether a distribution for a given species from a given country is directly comparable with a distribution for the same species for a different country, it is necessary to know the details of the methodologies used. It is important, therefore, that this information is adequately described in the associated metadata.

SourceInformation is feature-level metadata allowing the description of methodology information about each specific instance of distribution information. These metadata can be shared among several species distributions but, when downloaded by a user, they appear as part of the data set, encoded in

⁷¹ http://inspire.jrc.ec.europa.eu/documents/Data_Specifications/INSPIRE_DataSpecification_SD_v3.0rc3.pdf

⁷² <http://www.eu-nomen.eu/>

⁷³ <http://eunis.eea.europa.eu/>

⁷⁴ http://ec.europa.eu/environment/nature/natura2000/index_en.htm

Geography Markup Language⁷⁵ (GML), rather than with the data set-level metadata in the associated XML. The extended schema also provides the possibility to link to observation data specified within the Environmental Monitoring Facilities specification (Annex III: EF) and, in addition, includes a Darwin Core triplet⁷⁶ attribute which allows linking to the original observational data that can be accessed from GBIF data providers.

The EU BON partners and other scientific institutions throughout Europe are potentially major providers of species data to INSPIRE. The amount of species distribution maps and their quality will be significant for users of the INSPIRE mediated data. INSPIRE provides a unique possibility for contributions from scientific institutions. This distribution channel will promote a broader awareness among European decision makers of the importance of scientific institutions in environmental knowledge development.

Recommendation 10. Apply the Environmental monitoring Facilities specification⁷⁷ and the Darwin Core Triplet to allow users to access raw data applied in species distribution.

Recommendation 11. Use the INSPIRE Specification on Geographical Grid Systems⁷⁸ for the distribution maps based on grid cells.

Recommendation 12. Use EU-Nomen⁷⁹ for species code lists.

Habitats and Biotopes

The INSPIRE Directive defines *Habitats and Biotopes* as geographical areas characterised by specific ecological conditions, processes, structure, and (life support) functions that physically support the organisms that living there. They include terrestrial, fresh water and marine areas distinguished by geographical, abiotic and biotic features, whether entirely natural or semi-natural. Common to all spatial data that fall under this category is a characterisation of the distribution of geographical areas as functional areas for living organisms: biotopes are the spatial environment of a biotic community; habitats are the spatial environment of specific species.

Different countries or communities have different habitat classification systems. As such, there may be difficulties in mapping accurately certain habitat classes between national nomenclatures and also between national and European nomenclatures. Harmonisation needs to take into account local, national and international habitat classification systems. Harmonisation can be achieved, if there is one habitat classification system, which serves as “first among equals” to which all other classification systems can be mapped. The EUNIS habitat classification system serves this purpose. In addition a set of habitat types has been drawn up for the Marine Strategy Framework Directive. Local or national habitat classification can be used as well as long as a link is provided to these references. As a result, all habitat features will have one or more habitat type encodings, obligatory one(s) from, most frequently, the EUNIS habitat classification code list and optional one(s) from a registered code list related to an international, national or local habitat classification system.

Habitats and biotopes data specification

Data are needed on the geographic location and extent (area, length and/or volume) of habitats as well as on the geographic distribution of species. The distribution of habitats and biotopes has been added as a separate specification⁸⁰ because of the reporting obligation under article 17 of the Council Directive 92/43/EEC on the conservation of natural habitats and of wild fauna and flora. As the boundaries of the distribution of the habitats and biotopes are not based on the habitat and biotope

⁷⁵ <http://www.opengeospatial.org/standards/gml>

⁷⁶ In the absence of a bona fide globally unique identifier for a record, the Darwin Core triplet has been proposed as a way of generating an ad-hoc globally unique identifier from a combination of three other Darwin Core fields: [institutionCode]:[collectionCode]:[catalogNumber] <http://rs.tdwg.org/dwc/terms/index.htm#occurrenceID>

⁷⁷ http://inspire.jrc.ec.europa.eu/documents/Data_Specifications/INSPIRE_DataSpecification_EF_v3.0rc3.pdf

⁷⁸ http://inspire.jrc.ec.europa.eu/documents/Data_Specifications/INSPIRE_Specification_GGS_v3.0.1.pdf

⁷⁹ <http://www.eu-nomen.eu/portal/>

⁸⁰ http://inspire.jrc.ec.europa.eu/documents/Data_Specifications/INSPIRE_DataSpecification_HB_v3.0rc3.pdf

themselves, but on the boundaries of other geographic features, two different application schemas are provided: one on the characteristics and one on the distributions of habitats and biotopes.

Habitat feature descriptions usually carry a range of information, e.g., structural traits, lists of species, management proposals, to name just a few. However, for Annex III themes, the data specification is restricted to some basic attributes related to biotic features, such as vegetation types and species. Attributes that are related to abiotic features (e.g., water chemistry for freshwater or marine habitats) are not included yet, but may be added later in an extended application schema.

The application schema dealing with attributes of habitats and biotopes treats habitats and biotopes as geographic areas with their own specific boundaries. Habitat maps fall under this application schema. Habitats and biotopes are classified and mapped based on their specific characteristics, e.g., species composition and vegetation structure which are important for environmental impact assessments. Only the most basic characteristics have been considered in the present application schema. The assessments, e.g., of the conservation status of habitats, are dealt with in the framework of other EU directives, so this type of information is not included in the application schema.

The application schema for distribution of habitats and biotopes is similar to the application schema for the distribution of species and concerns the geographic distribution of habitats and biotopes. The distribution of habitats and biotopes is depicted in relation to a reference data set, e.g., gridded data. Source information is added in order to include metadata about specific instances of habitats.

Recommendation 13. EU BON should adopt the EUNIS habitat classification system. Local, national or other habitat classification can be used as well as long as they reference the EUNIS habitat classification code list.

Recommendation 14. EU BON should contribute to the creation of a set of ecological indicators to be included in a hybrid ontology model (mixing local and other ontologies) so that indicators can be stored in a shared vocabulary. Moreover, EU BON should promote the formalisation of indicators and classification outputs in defined formats, e.g., RDF/XML.

Bio-geographical regions

The Directive defines bio-geographical regions as “areas of relatively homogeneous ecological conditions with common characteristics”. The most important guiding document in regard to bio-geographical regions in Europe for the data specification is the Habitats Directive (EEC/92/43) which contains a list of bio-geographical regions (Article 1.iii). The Habitats Directive was the first EU legislation to introduce the concept of bio-geographical regions. There are currently 9 regions, covering the 27 Member States of the EU, and an additional 2 bio-geographical regions have been added through the Bern Convention. The regions have been modified to make them easier to use administratively and they form ecologically coherent units of similar environmental conditions. Despite their importance for the data specification, these bio-geographical regions will be represented by only one distinct data set which will be provided by EEA.

Bio-geographical regions data specification

While the legally mandated bio-geographical regions fulfil administrative needs, there is further need amongst users for other types of ecological regions for various analyses at a European scale or for use at a regional, national or sub-national level. The needs of these users for a more detailed or conceptually different set of ecological regions are covered under the use of code lists such as the Environmental Stratification Classification values. Another example is the *European Map of Natural Vegetation* which uses a specific vegetation type classification. These more detailed ecological regions may also include local sub-categories of the bio-geographical regions outlined in the Habitats Directive. The Bio-geographicalRegions data model⁸¹ thus provides a generic means for a common pan-European representation of bio-geographical regions. The spatial object Bio-geographicalRegion is the key spatial object of the application schema for representing regions or areas of relatively

⁸¹ http://inspire.jrc.ec.europa.eu/documents/Data_Specifications/INSPIRE_DataSpecification_BR_v3.0rc3.pdf

homogenous ecological conditions with common characteristics. It is this spatial object type that will allow for a proper description of the bio-geographical classification that has been applied to identify and classify the bio-geographical region each feature represents.

With this in mind it should be emphasised that the application schema not only supports the classification of bio-geographical regions as mandated by the European Habitats Directive, but also meets the requirements raised by INSPIRE stakeholders with regard to alternative or more precise ecological regions.

Recommendation 15. EU BON should adopt the European Habitats Directive bio-geographical regions classification. Other more detailed ecological regions should be covered by the use of code lists.

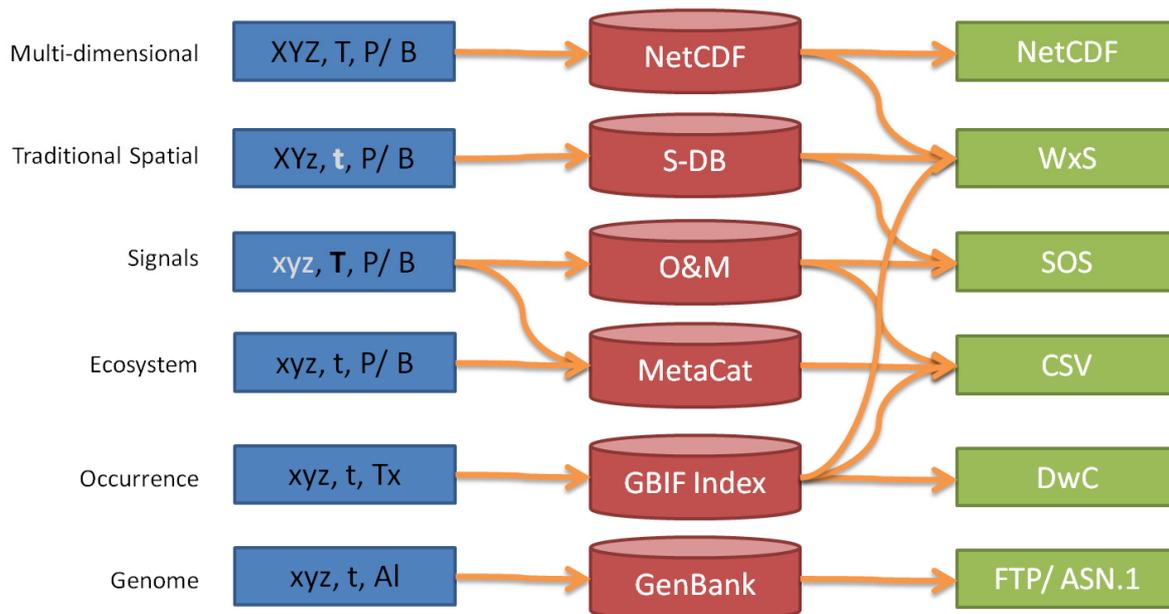


Fig. 13 Example generic data families and interoperability requirements (Hugo et al. 2013). The abbreviations are: S-DB: spatial database; WxS: OGC (Open Geospatial Consortium) web services; O&M: OGC Observations and Measurements model; SOS: OGC Sensor Observation Service; CSV: comma separated value; DwC: Darwin Core.

2.3. Generic data families

The GEO BON working group on data integration and interoperability has developed (Hugo et al. 2013) a classification of generic data families⁸² and their interoperability requirements (Fig. 13) all of which are applicable and relevant for EU BON. Data families are grouped according to their spatial, temporal and semantic coverages with each unique combination of these, supported by a vocabulary/ontology, considered a generic data family. Thus the occurrence, genome, and ecosystem data families all include a reference to a particular place and time, but differ in that occurrence also references a taxon, genome references a sequence and ecosystem references biological phenomena. The different types of coverage (spatial, temporal and semantic) and their attributes are:

- Spatial Coverage
- XYZ
- Temporal Coverage: **T** (continuous or near-continuous); **t** (discrete)
- Topic or Semantic/ Ontological Coverage
- **P**: Phenomenon
- mostly physical, chemical, or other contextual data

⁸² <http://meetingorganizer.copernicus.org/EGU2013/EGU2013-6968.pdf>

- **B:** Biological
- **Tx:** Species and Taxonomy (with some extensions)
- **Al:** Allele/ Genome/ Phylogenetic

Each unique combination of these, supported by a vocabulary/ ontology is a generic data family.

2.4. Ecological measurements

“When writing my electronic monography (e-monograph) in 2007-9 I wished to link the plant species to other organisms within the ecological food chains/food web. However, I could not even find an e-monograph on birds at the time or have the software programming knowledge to create interspecies relationships between electronic monographs and/or electronic floras. Ultimately I wish to see a ‘virtual life on Earth’ where cross-linking of data can be explored, for example, how shifting species distribution in light of climate change will affect food webs. Consequently the results can be used to drive conservation management and placement on the IUCN Red Data List”. -- Prof. Fiona Young, University of Reading (UK)

The quotation above provides an insight regarding the challenge for EU BON and biodiversity informatics in the near future: to develop an infrastructure to allow the available data to be brought into a coordinated modelling environment, e.g., for the carbon cycle, information is required from the molecular level over seconds, to information on tree growth per year and per species, in addition to much more information. First of all, we need to understand how the ecosystem works⁸³. Then, we need a systems approach to understanding biodiversity that moves significantly beyond taxonomy and species observations. Such an approach needs to look at the whole system to address species interactions, both with their environment and with other species. Moreover, data on all these ecological aspects needs to be standardised in order to support integration.

Numerous biodiversity informatics projects have been funded in the Framework Programmes (see Annex I: related European projects). Globally, there are more than 650 projects known. Many of these projects directly address the challenges of deploying e-Infrastructure for biodiversity science, but from the variety of approaches, it is clear that consensus is lacking about how best to do so. Data integration and analysis critically require semantic consistency as well as syntactic standardisation. As terms are rarely independent of one another, a vocabulary list evolves into a thesaurus and, once formal relationships between terms are agreed, an ontology. However, for the most part, stable, widely used ontologies are lacking in the biodiversity domain and their provision is currently an area of active investigation through TDWG and other organisations (see section 13 Vocabularies and Ontologies).

2.4.1. Standards for ecological data

The TDWG standard Structured Descriptive Data (SDD) is designed for the expression and transport of descriptive information about biological specimens, taxa, and similar entities such as diseases or ecosystems. However, SDD does not currently accommodate certain types of data and is thus not suitable for ecological measurements. For instance the following cannot be included:

- Molecular sequence and other genetic data,
- Occurrence and specimen data (e. g., distribution maps),
- Complex ecological data such as models and ecological observations,
- Organism interactions (host-parasite, plant-pollinator, predator-prey, etc.),
- Nomenclatural and formal systematic (rank) information.

Another TDWG standard, the Darwin Core (Wieczorek et al. 2012) glossary of terms, is designed for sharing data about biodiversity – *“the occurrence of life on earth and its associations with the environment”*. Darwin Core (DwC) can be seen as an extension of Dublin Core for the biodiversity domain. It is amongst the most widely deployed of biodiversity vocabularies (e.g. on the GBIF

⁸³ See, e.g., the GoMexSi project which is developing an open source tool for recording, archiving and analysing species interaction data for the Gulf of Mexico (<http://gomexsi.tamucc.edu/>).

network), and while its main use is for publishing specimen and observation records, it continues to evolve to meet the needs for sharing more complex sample-based and ecological data. For example, a recent workshop⁸⁴ explored the extension of DwC for encoding sample based data.

A third TDWG standard, Access to Biological Collection Data⁸⁵ (ABCD) uses a more comprehensive model than DwC and is thus more expressive. ABCD covers metadata (data set descriptions), everything related to the collecting or observing event (who, why, where, when, how), everything related to identifications (who, when, as what, according to etc.), biological observations (pathogen, pollinator, parasitic and other relationships, sex, stage, etc.), and freely chosen measurements and categorised observations and their methodology. ABCD is used in the BioCASE network and readily integrates with the GBIF network through a mapping of ABCD to DwC. Significantly, DwC has adopted some properties such as measurementOrFact from ABCD (see section 8: Unifying data standards). ABCD also provides extensions for Earth/geosciences (Access to Biological Collection Data Extended for Geosciences, ABCDEFG⁸⁶) and genomic data (ABCDDNA – see section 9.1.4).

Ecological Metadata Language (EML)⁸⁷ is a metadata specification developed by the ecology discipline and for the ecology discipline. It is based on prior work done by the Ecological Society of America and associated efforts (Michener et al., 1997). EML is implemented as a series of XML document types that can be used in a modular and extensible manner to document ecological data. Each EML module is designed to describe one logical part of the total metadata that should be included with any ecological data set. The top level modules include descriptions of data sets, literature, citations, software, and research protocols. Supporting modules cover access control, file formats, people and organisations, spatial and temporal coverage, research context, and methodological information. EML is suitable for describing information resources, but it does not provide vocabulary or schema for the actual data. In fact, EML is mostly about general project data, while its ecology substance is rather limited. Because of this, EML nicely complements the domain specific standards mentioned above.

Biodiversity informatics is inherently a global initiative. With a multitude of organisations from different countries publishing biodiversity data, the foremost challenge is to make the diverse and distributed participating systems interoperable in order to support discovery and access to data. A common exchange technology, e.g. the widely used XML or JSON, may allow the syntactic exchange of data blocks, but participating systems also need to understand the semantics of the data being delivered to process it meaningfully. However, unless the data share a common reference model, the exchange requires some brokering or other semantic processing.

2.4.2. INSPIRE and ecological data

Data interoperability is also essential in European spatial and ecological harmonisation efforts such as INSPIRE. Especially for the Natura2000 framework with its strict reporting obligations, achieving harmonisation and comparability of spatial biodiversity also meets the requirements for at least four of the EU Aichi Targets (1, 2, 3 and 5). Remote Sensing based assessment of Natura2000 habitats (with their associated physical and ecological information) should provide automated algorithms for generating comparable biodiversity information across Europe. Ecological measurements should be used and transformed in ecological and biodiversity indicators. In the case of biodiversity monitoring that is based on the evaluation and correlation with different ecological indicators, the need for a shared ontology is urgent. Such ontologies are used to improve communication and interoperability by specifying the semantics of the data. And while standard ontologies may exist within specific domains (e.g., in the field of taxonomy or genetics), building a new shared ontology will be required for collaboration across domains, as in the case of the multifaceted domain of ecology and environmental science.

⁸⁴ http://www.gbif.org/orc/?doc_id=5424

⁸⁵ <http://wiki.tdwg.org/ABCD/>

⁸⁶ <http://www.geocase.eu/efg>

⁸⁷ <http://knb.ecoinformatics.org/software/eml/>

Recommendation 16. EU BON should promote the development and uptake of shared ontologies for ecology and environmental science.

2.4.3. Unifying data standards of biodiversity and ecological monitoring

The two largest domains of biodiversity observation are specimen occurrences and biological (natural resource) surveys. The former usually is based on sporadic, opportunistic collection or observation activity, while the latter consists of repeated sampling at known sites, locations and follows a known protocol. Hence, the latter method is most appropriate for observing change, but the former can also be used, if the number of data are large enough and sampling biases can be eliminated by computation. Data potentially available through both of these domains are very large. GBIF, which represents the occurrence domain, currently has mobilised more than 14,000 data sets. ILTER, which represents the ecosystem monitoring domain, has 25,000 data sets. Both have the potential of growing at least ten-fold. In particular, for ecosystem monitoring, much data exists in government agencies for forestry and agriculture which have not yet started any data sharing activities.

The status of data standards and data sharing in these two domains is somewhat different. Common ground has been found in recent years through adoption of Ecological Metadata Language (EML) by the occurrence domain. EML originates in the ecosystem monitoring domain, and is suitable for describing sampling protocols. Through development of a profile⁸⁸, it has now been extended to also cover descriptions of museum collections and databases. However, this still needs to be directly supported in EML through a dedicated class of elements.

Recommendation 17. Extend EML to contain a class of elements for collections and databases.

Commonalities among data standards end here. The occurrence domain has developed the ABCD and Darwin Core standards, but the ecosystem monitoring domain has nothing similar, probably due to the perceived complexity of the task. However, this notion probably could be revisited, since the standardisation process around Darwin Core centres on the development of a flat vocabulary. Apart from the provision of the Simple Darwin Core schema⁸⁹, it does not take a position on how the elements are structured with regard to each other in a schema or other data model. This allows tackling complex domains with simple statements about facts.

In ABCD a class of elements for abstract MeasurementOrFacts was included early on. This is capable of capturing any data, and comes rather close to RDF's basic triple structure object:attribute:value (or subject:predicate:object). MeasurementOrFacts has since been added to Darwin Core. However, its use still is not widespread, possibly because what goes in its elements is not standardised and requires that some kind of specific profile is developed for each project. Not much has been publicised about these profiles. The question is, should MeasurementOrFacts be developed further and its content standardised, or should similar capabilities be sought by using RDF?

Recommendation 18. Assess the state of using MeasurementOrFacts, and develop best practices for its use, if possible. Assess whether RDF would be a better alternative.

In the occurrence domain, the basic observation always is about a species in a location at a time. Further data elements can be recorded, but the observation is useful with only such basic data. This can be represented in Simple Darwin Core which is a flat table. Additional repeated elements, such as MeasurementsOrFacts can be represented in auxiliary flat tables where multiple records can be linked to individual records in the Simple Darwin Core table as described in the Darwin Core Text Guide⁹⁰. This so-called star schema can potentially be nested further, but with concomitant, added complexity. Examples of such nested schemas have yet to be publicised.

In the ecological monitoring domain, no such concept of a basic observation exists. No standardised vocabulary exists, either. Each project has their own way of representing data and coding it. For instance, when the Knowledge Network for Biocomplexity⁹¹ (KNB) had 14,000 data sets, there were

⁸⁸ <http://rs.gbif.org/schema/eml-gbif-profile/>

⁸⁹ <http://rs.tdwg.org/dwc/terms/simple/index.htm>

⁹⁰ <http://rs.tdwg.org/dwc/terms/guides/text/index.htm>

⁹¹ <http://knb.ecoinformatics.org>

11,000 different schemas (Matt Jones, pers.comm. in 2007). It might be possible to tackle this now, following the approach of Darwin Core.

Much of ecological monitoring data comes from vegetation and natural resource surveys. There, a basic observation is made on a plot. Most elements of a plot come from the location class of terms in Darwin Core. The geographic extent of the plot can be fully represented with Darwin Core, although some standardisation of encoding the shape of the plot may be needed. A plot is more than a location, though, as it is always connected to a sampling protocol, which is an available Darwin Core term (*samplingProtocol*)⁹². A fixed trap or other capture device has many of the attributes of a plot, although its geographic extent is more vague. The capturing method can be represented in Darwin Core using *samplingProtocol*. Capturing methods may change from event to event, which is a consideration in whether to combine *samplingProtocol* and location in practical data management.

Recommendation 19. Examine with real data how sampling plots can be represented in Darwin Core.

There is an important shortcoming in the way DwC is organised with regard to species densities and presence/absence data. Such information is often required for a large number of model algorithms. While theoretically the fields for providing “Event” information (including *eventID*⁹³) and the *individualCount*⁹⁴ field could be used for providing the necessary detail, this is somehow regarded as tricky and this information is rarely provided when publishing data sets through the BioFresh and GBIF network. Abundances can be expressed as number of individuals, densities, abundance ranks, etc. To accommodate the inclusion of such information in data sets exposed in DwC, one could envisage the need for, e.g. a “speciesDensity” and “speciesDensityType” field (whereby the latter field is meant to record the unit in which the density is recorded).

However, as this issue is clearly relevant for all primary biodiversity data and its use in modeling, there is a need for discussion and consensus within the community. GBIF has already initiated this process through convening a recent workshop on extending DwC for sample data (report available⁹⁵). As such EU BON could take a leading role in continuing the investigation with the Biodiversity Informatics (TDWG) community and prototyping actual use cases. A plot is usually not standalone, but is part of one or more sampling schemes with many plots. These are field protocols. This is wider than Darwin Cores sole term *samplingProtocol*, and can be described in EML.

From each plot a number of measurements will be made at a recording event. These correspond to primary biodiversity data and typically include assessing the coverage (percent) of each species or compound species such as “broadleaved bushes” at various strata (vertical layers). Compound species can probably be encoded in Darwin Core with *originalNameUsage*⁹⁶, but this needs to be verified with real data. Depending on the survey type, individuals may or may not be counted, or their numbers just approximated.

Recommendation 20. Darwin Core has no terms for coverage or stratum, which need to be included. Alternatively, consider abandoning *IndividualCount* and replacing it with a more general term such as quantity and quantityUnit (closed vocabulary with values individualCount, coveragePercentage, basalArea, etc.). Also, a range for the quantity would be useful.

It is clear that observations other than those on biodiversity will be recorded in a recording event on a plot. Examples are soil type, habitat class, etc. These are standalone MeasurementOrFact records, which are linked to the plot and event.

An important difference to flat Darwin Core is that the location elements need not to be repeated for each record made on a plot. Instead, the locations need to have a unique *locationID*⁹⁷ (or *plotID*) which then is included in the records. Also, the unique IDs of the recording event and of the protocol need to be included. Some of this information can be stored in the accompanying EML document.

⁹² <http://rs.tdwg.org/dwc/terms/index.htm#samplingProtocol>

⁹³ <http://rs.tdwg.org/dwc/terms/index.htm#eventID>

⁹⁴ <http://rs.tdwg.org/dwc/terms/index.htm#individualCount>

⁹⁵ http://www.gbif.org/orc/?doc_id=5424

⁹⁶ <http://rs.tdwg.org/dwc/terms/index.htm#originalNameUsage>

⁹⁷ <http://rs.tdwg.org/dwc/terms/index.htm#locationID>

Such separation of the location, event, and the species recorded might also be useful for occurrence data, as it is well known that collectors repeatedly visit the same locations. Why digitise and georeference these repeatedly?

Recommendation 21. Develop a shared repository of known and named observation localities, i.e. plots.

There is a variant of the ABCD schema which already supports separation of location and records (FMNH 2008⁹⁸) by lifting the *gatheringLocation* to the top level of the schema, where it is then used for all records (units). Using primary biodiversity data this way would pave the way for unifying the two large data domains.

Many of the above issues have already been addressed by the Botanical Information and Ecology Network (BIEN)⁹⁹. The BIEN working group is developing a standardised workflow and informatics engine for the integration, access, and discovery of disparate sources of botanical information. The BIEN 3 schema is based in part on VegBank¹⁰⁰, with modifications to support herbarium specimens and a broader diversity of inventory data. Data is transferred to VegBank from specimen databases using the existing Darwin Core (DwC) exchange schema; plot data is transferred using VegCSV, a new plain text schema developed by the BIEN project and derived from Veg-X. It is a vegetation plot exchange schema, which enables organising vegetation data and making them available to the entire ecological community. Veg-X has been developed by TDWG's Vegetation Observations Data Exchange Task Group, which is a subgroup of the Observations and Specimen Records Interest Group. Veg-X and VegCVS have not yet been standardised by TDWG. Veg-X contains selected modules from EML, some terms from Darwin Core, and its own schemas for plots, plot observations, organism observations, and miscellaneous items. The latter ones may be possible to include in Darwin Core terms in future

Recommendation 22. Investigate inclusion of Veg-X elements in Darwin Core.

There is one more concept in the ecological monitoring domain which has no counterpart in occurrence domain. "Site" is a facility such as an experimental station with certain capabilities for carrying out projects. Perhaps a biological collection is a corresponding concept. We do not need to consider supporting the site concept in Darwin Core at this stage, but will need to appreciate that progress has already been made by ALTER-net¹⁰¹ to describe sites in a standardised way.

Much work is also being done in habitat classification. Darwin Core currently has just one property (*habitat*) for describing the habitat in which an event occurred but a proposal¹⁰² has recently been submitted to expand Darwin Core by adding three new properties (environmental material, environmental feature, and biome), with the recommended values for these and habitat drawn from the equivalent Environment Ontology (EnvO)¹⁰³ classes.

In summary, Darwin Core vocabularies can probably be used to capture, if not all of the ecological monitoring data, at least the biotic part. This needs to be tested at various sites with different types of records.

Recommendation 23. Test with real data from EU BON test sites and in workshops with willing cooperating projects the use of Darwin Core for capturing broader ecological data. Consider with practical examples how and whether the standard can be extended to cover such data.

2.4.4. Observations and measurements model

In the longer term, for describing observations and measurements associated with biological sampling, the biodiversity community will benefit from adopting a comprehensive conceptual model,

⁹⁸ <http://www.luomus.fi/info/schemas/fmnh2008/documentation.html>

⁹⁹ <http://bien.nceas.ucsb.edu/bien>

¹⁰⁰ <http://vegbank.org/>

¹⁰¹ <http://data.lter-europe.net/deims/>

¹⁰² http://www.gbif.org/orc/?doc_id=5424

¹⁰³ <http://environmentontology.org/>

such as the OGC and ISO O&M^{104 105}. The model is an essential underpinning for the related OGC Sensor Observation Service (see Section 2.7.2). The model defines an observation as an activity that results in a measurement, obtained using a particular procedure, of the value of a property associated with a feature-of-interest. A sampling feature can be, e.g., a station, transect or specimen and a set of related observations can be grouped together in the same *samplingEvent*. The model is, by its nature, high level and abstract, and although an XML implementation¹⁰⁶ exists, the challenge remains¹⁰⁷ for any community of practice to develop community based vocabularies through identifying the important features and their properties within a particular domain and express these using GML application schemas.

2.5. Genetic/genomic data

Generally the term genetics is used for the study of single genes and genomics covers study of all genes and their interactions including environmental. Standards developed for this subject cover both areas and there is no clear demarcation line between them. Therefore we use here the term genomic *sensu lato* which also covers genetics. Metagenomics refers to the study of sequence data derived directly from environmental samples without first undertaking DNA isolation and culture steps. Such studies are set to revolutionise our understanding of biodiversity by enabling investigation of microbial diversity in relation to community structure, habitat and environment at the fundamental level of the genome (Wooley 2010).

Genomic data is one type of many which are used to study taxa and their function in different environments. Other major data types include morphological/anatomical, physiological, chemical, environmental, etc. Exhaustive understanding of taxa, their function and distribution related to the environment and climate change is possible if all data types are stored and managed in conjunction. This is now a major driving force and most organisations developing biodiversity standards are trying to merge or link standards developed originally for a specific data type only.

2.5.1. Major genomic data related initiatives

INSDC

The International Nucleotide Sequence Database Collaboration¹⁰⁸ (INSDC) includes three collaborating partners, viz. National Center for Biotechnology Information¹⁰⁹ (NCBI), European Nucleotide Archive¹¹⁰ (ENA) and DNA Database of Japan¹¹¹ (DDBJ) who, together, developed a common standard and exchange format for genomic data. Documentation includes a feature definition table¹¹² and sample record¹¹³.

GSC

The Genomic Standards Consortium¹¹⁴ (GSC) is the principal organisation for the development of genomic standards: founded in 2005, its mission is the implementation of new genomic standards as well as methods to capture and exchange associated metadata. GSC collaborates with the INSDC in order to implement genomic standards in their system. The GSC standard “Minimum Information about any (x) Sequence” (MIxS) (Yilmaz et al. 2011) includes three separate checklists which are sometimes also called standards: MIGS for genomes (“minimum information about a genome sequence”), MIMS for metagenomes (“minimum information about a metagenome sequence”) and MIMARKS for marker genes (“minimum

¹⁰⁴ <http://www.opengeospatial.org/standards/om>

¹⁰⁵ http://portal.opengeospatial.org/files/?artifact_id=41579

¹⁰⁶ http://portal.opengeospatial.org/files/?artifact_id=41510

¹⁰⁷ <https://teamwork.niwa.co.nz/display/NZEIIF/Biodiversity+Interoperability+through+Open+Geospatial+Standards>

¹⁰⁸ <http://www.insdc.org>

¹⁰⁹ <http://www.ncbi.nlm.nih.gov/>

¹¹⁰ <http://www.ebi.ac.uk/ena/>

¹¹¹ <http://www.ddbj.nig.ac.jp>

¹¹² <http://www.insdc.org/documents/feature-table>

¹¹³ <http://www.ncbi.nlm.nih.gov/Sitemap/samplerecord.html>

¹¹⁴ <http://gensc.org>

information about a marker genome sequence”). MiXS also includes so called environmental packages for describing the environment from where the organism(s) or DNA sample was taken. There are currently 14 environmental packages with new, additional packages under development. The list of environmental packages as well as shared and specific descriptors in the checklists are shown on Fig. 14.

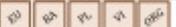
Specification projects	MIGS	MIMS	MIMARKS	New checklists
Checklists		metagenomes	survey	specimen
Shared descriptors	collection date, environmental package, environment (biome), environment (feature), environment (material), geographic location (country and/or sea, region), geographic location (latitude and longitude), investigation type, project name, sequencing method, submitted to INSDC			
Checklist specific descriptors	assembly, estimated size, finishing strategy, isolation and growth condition, number of replicons, ploidy, propagation, reference for biomaterial		target gene	
Applicable environmental packages (measurements and observations)	Air Host-associated Human-associated Human-oral Human-gut Human-skin Human-vaginal		Microbial mat/biofilm Miscellaneous natural or artificial environment Plant-associated Sediment Soil Wastewater/sludge Water	

Fig. 14 Overview of the MiXS checklists and environmental packages.

Source: <http://gensc.org/index.php?title=File:Fig1.png>

The Genomic Biodiversity Working Group¹¹⁵ (GBWG) of the GSC was formed to review existing biodiversity standards and bridge the gaps between researchers working in molecular biology, taxonomy, ecology, and biodiversity informatics. The GBWG collaborates with the Biodiversity Information Standards (TDWG) consortium with annual meetings at TDWG conferences. In addition, a series of workshops funded largely through the US National Science Foundation and GBIF brought together experts from the genomics and traditional biodiversity communities to address the aligning of their respective standards. In February 2012, a hackathon brought together several experts to continue the alignment of the Darwin Core and MiXS standards (Ó Tuama 2012). In May 2012, at the Semantics of Biodiversity meeting¹¹⁶, term definitions in biodiversity informatics were addressed and, in September 2012, at the bioCollections Ontology Hackathon¹¹⁷, a prototype bioCollections Ontology was developed. These workshops gave significant input to two initiatives:

- Darwin Core DNA and Tissue Extension which aims to track DNA extracts, and any biological samples as they relate to occurrence records, harvested by GBIF. Two primary use cases were proposed for this extension – a) barcoding, producing 1:1 mapping between sample and taxonomy, and b) metagenomics / molecular community ecology giving typically 1-to-many mapping between sample and taxonomy.
- BiSciCol¹¹⁸, a linked data project with a goal of tracking biological collection objects and their derivatives, across distributed databases, multiple domains and information standards. BiSciCol provides a method for determining allowable relationships and traversing graph-based data derived from multiple standards for biological collections.

These initiatives led to the creation of two extensions to the Darwin Core Standard (DwC), viz. MiXS sample¹¹⁹ and Taxon Abundance¹²⁰ which are still under development. There are many adopters of the

¹¹⁵ http://gensc.org/index.php?title=Biodiversity_Working_Group

¹¹⁶ <http://biocodecommons.org/workshops/sob.html>

¹¹⁷ <http://biocodecommons.org/workshops/bioCollections/>

¹¹⁸ <http://biscicol.org>

¹¹⁹ <http://tools.gbif.org/dwca-validator/extension.do?id=http://gensc.org/ns/mixs/terms/Sample>

¹²⁰ <http://tools.gbif.org/dwca-validator/extension.do?id=http://rs.gbif.org/terms/1.0/TaxonAbundance>

MIxS standards including INSDC, the Quantitative Insights Into Microbial Ecology¹²¹ (QIIME) software package, EBI Metagenomics Portal, Genomes Online Database (COLDB), etc. and the number continues to grow. The GSC also have their own journal “Standards in Genomic Sciences”¹²² and several core projects: 1) GCDML¹²³ - Genomic Contextual Data Markup Language, an XML Schema for generating MIxS compliant reports for data entry, exchange and storage. This sample-centric, strongly-typed schema provides a diverse set of descriptors for describing the exact origin and processing of a biological sample, from sampling to sequencing, and subsequence analysis; 2) Genomic Rosetta Stone - a registry of identifiers describing complete genomes across a wide range of relevant databases (Genome Catalogue) and allowing to automatically track down all related metadata for these published genomes. Their end goal is to make this physical mapping available in multiple formats (e.g. relational schema / spreadsheet / webservice) to facilitate the discovery of genomic information on the web, comparative genomic studies, and the population of databases with hyperlinks and metadata;; 3) Habitat-Lite, which is a light-weight , easy-to-use set of terms that captures high-level information about habitat while preserving a mapping to existing Environment Ontology (EnvO). The main motivation is to meet the needs of the majority of users by generating enhanced list of terms based on already existing data submitted to INSDC. EnvO terms are used in MIxS specification. GSC also participates in many projects on the community level.

BOL

The Barcode of Life¹²⁴ (BOL) includes three major consortia, viz. the International Barcode of Life Project¹²⁵ (iBOL), the Consortium for the Barcode of Life¹²⁶ (CBOL) and the European Consortium for the Barcode of Life¹²⁷ (ECBOL). Their Database Working Group (DBWG) published BARCODE Data Standard “Data Standards for BARCODE Records in INSDC (BRIs)”¹²⁸. BRIs set five major components which secure integration between DNA barcode sequences and other biodiversity information (data on specimens, taxonomy, biogeography, etc.).

GGBN

The Global Genome Biodiversity Network¹²⁹ (GGBN) is a global network of well-managed collections of genomic tissue samples across the Tree of Life, which also develops standards for sharing DNA and tissue information. The DNA Bank Network¹³⁰, initiated by GBIF Germany in 2007 is one of the founding organisations of the GGBN. It maintains a central web portal, upon which the GGBN portal will be built, providing DNA samples of complementary collections and has developed and uses in its network, ABCDDNA¹³¹, a DNA extension for the ABCD standard, and submitted it to TDWG for ratification. GGBN is also involved in creating and testing the DNA and tissue extension for DarwinCore Archive¹³², and planning to use it in parallel with their ABCDDNA schema.

Protocols and mechanisms for interoperability in European context

Genomic data are well organised and there is a high level of interoperability because most genetic data are published in the INSDC databases (NCBI, ENA, DDBJ) using common standards.

2.5.2. Data heterogeneity – reasons and overcoming

There is almost no data heterogeneity and lack of interoperability in genomic data. However there are limitations around associated “contextual” taxonomic and geographical information. Based on

¹²¹ <http://qiime.org/>

¹²² <http://www.standardsingenomics.org>

¹²³ <http://gensc.org/projects/gcdml/>

¹²⁴ <http://www.barcodeoflife.org>

¹²⁵ <http://ibol.org>

¹²⁶ <http://www.barcodeoflife.org/content/about/what-cbol>

¹²⁷ <http://www.ecbol.org>

¹²⁸ http://www.barcodeoflife.org/sites/default/files/legacy/pdf/DWG_data_standards-Final.pdf

¹²⁹ <http://www.ggbn.org/>

¹³⁰ <http://www.dnabank-network.org/>

¹³¹ <http://wiki.bgbm.org/dnabankwiki/index.php/ABCDDNA>

¹³² <http://rs.tdwg.org/dwc/terms/guides/text/>

different studies nearly 20% of fungal DNA nucleotides in INSDC are insufficiently identified including misidentifications. This may apply to other kingdoms as well. Analyses of DNA sequences in INSDC also show that only 60% have information on country of origin and approximately 10% have geographic coordinates. In most cases, this information is available in published research papers but there is no direct link between these data. Country information is now mandatory in INSDC but coordinates and GSC checklists are not. It is important that coordinates and specific environmental information based on GSC checklists be made mandatory as well. Insufficiently identified INSDC DNA sequences are easily amended by third-party annotations. This is done, for example, for the Fungal Kingdom by the UNITE¹³³ community. Their annotations are also linked back to the ENA and NCBI.

2.5.3. Recommendations for EU BON

The storage of genomic data by INSDC is a guiding light for how all types of the taxon occurrences should be stored and managed. Most research journals require that authors upload their genomic data into INSDC databases as a condition of publication of research results. This ensures that all genomic data are stored in the same way and are accessible from one site. EU BON could promote central storage for published biodiversity data. It is probably comparably easy to implement if research journals and funding agencies are involved. There is rapid growth of environmental studies where the species richness is measured by sampling DNA from the different biological samples like soil, water, air, etc. Such genomic data are usually accompanied by rich environmental data which demand fast reaction from all major players because these data must be stored for future studies in standardised way. However, it is not yet mandatory.

2.6. Nomenclature and checklists

In this section, the principal frameworks, standards and services for nomenclature and classification are outlined. Even though the main focus of EU BON is necessarily at the European level, it must be emphasised that we also need global names data sets because of the need to address species introductions and invasive alien species. However, as all regions in the world have this need, it is probably better solved at GEO BON rather than EU BON level.

2.6.1. Pan-European Species Infrastructure

The Pan-European Species directories Infrastructure¹³⁴ (PESI) provides standardised and authoritative taxonomic information by integrating and securing Europe's taxonomically authoritative species name registers and nomenclators (name databases) and associated expert networks that underpin the management of biodiversity in Europe. PESI integrates three European Focal Points Networks: Fauna Europaea, European Register of Marine Species (ERMS), and Euro+Med PlantBase and is now part of the broader initiative on taxonomic data standards known as EU-nomen (see section 6.1.1 Species distribution; also Recommendation 12).

PESI offers web services¹³⁵ based on the platform-independent SOAP/WSDL standard. It consists of nine functions that allows a programmer to check the spelling of a taxon, to get the authority for a taxon, to get the full classification for a taxon, to resolve an unaccepted name to an accepted one, to get the citation for a taxon, to get the scientific name for a given common name/vernacular, to get the common name(s), to fuzzy match one or more scientific names. Some functions use a GUID (Globally Unique Identifier) as a parameter which can be obtained by the function getGUID.

Every record retrieved from PESI has a Globally Unique Identifier (GUID). In general, it consists of 32 hexadecimal numbers separated into five groups. Many records in PESI have an LSID¹³⁶ (Life

¹³³ <http://unite.ut.ee>

¹³⁴ <http://www.eu-nomen.eu/>

¹³⁵ <http://www.eu-nomen.eu/portal/webservices.php>

¹³⁶ <http://www.ipni.org/lsids.html>

Sciences Identifier). However, it is now accepted that HTTP URIs can perform a similar naming task, are less technically complex to set up, and follow W3C architecture best practices.

2.6.2. Catalogue of Life and Species 2000

The Catalogue of Life¹³⁷ (CoL), a product developed through the partnership of Species 2000 and the Integrated Taxonomic Information System (ITIS), is the most comprehensive and authoritative global index of species currently available. It consists of a single integrated species checklist and taxonomic hierarchy. The catalogue holds essential information on the names, relationships and distributions of over 1.4 million species. This figure continues to rise as information is compiled from diverse sources around the world. The key features of the CoL are the species checklist, management classification, and integration of global species databases. It provides critical species information on synonymy, higher taxa and distribution. Two versions of the checklist are available. The Dynamic Checklist is always up to date and anything can change as the list develops. The Annual Checklist is a snapshot of the entire catalogue.

CoL provides a web service¹³⁸ for retrieving data from both versions of the checklist. The service is PHP-based with a simple GET request. It uses any one of two mandatory parameters: name, id. Moreover, response format, type of response and record start can be specified by optional parameters. When a request is sent, the service searches the name or id in the database and returns the appropriate response. If the name/id is found, each result can be an accepted (infra)species name, an (infra)species synonym, a common name for an (infra)species, or a higher taxon.

The response is issued in XML by default and contains the search parameters and, if an error occurred, an error message. The response can be either terse or full as specified in request parameter *response*. As the name suggests, terse response contains only basic important data and is used when run time is an issue while the full response contains complete information. The results are also available in PHP format, as an array with serialised string format, which can be converted back to an array in PHP using the *unserialize* method. In the context of i4Life and BioVel, an additional CoL service layer has been implemented on the basis of an instance of the EDIT Platform for Cybertaxonomy. The services are particularly strong in performing fast fuzzy searches on taxonomic names.

Species 2000¹³⁹ is a network of database organisations that engages with taxonomists around the world in order to develop a uniform and validated index of the world's species (plants, animals, fungi and microbes) by integrating several global databases that deal with the major groups of organisms.

2.6.3. Integrated Taxonomic Information System

The Integrated Taxonomic Information System¹⁴⁰ (ITIS) provides authoritative taxonomic information on plants, animals, fungi, and microbes of North America and the world. ITIS is meant to serve as a standard to enable the comparison of biodiversity data sets, and therefore aims to incorporate classifications that have gained broad acceptance in the taxonomic literature and by professionals who work with the taxa concerned.

ITIS uses a few data standards. Data conform to the International Code of Botanical Nomenclature and the International Code of Zoological Nomenclature. Ranks in the animal kingdom below *subspecies* are not included as these ranks are not allowed in the zoological code. The botanical code allows the ranks *variety*, *subvariety*, *forma*, and *subforma*. ITIS adopted a five kingdom system - *Morena*, *Protista*, *Plantae*, *Fungi*, *Animalia*. ITIS makes practical decisions as to the placement of protists within the five kingdom framework.

¹³⁷ <http://www.catalogueoflife.org/>

¹³⁸ <http://www.catalogueoflife.org/content/web-services>

¹³⁹ <http://www.sp2000.org>

¹⁴⁰ <http://www.itis.gov/>

The ITIS SOAP web service¹⁴¹ provides 51 functions to retrieve various data. These include common functions like getting full record, accepted name, or hierarchy, as well as uncommon functions like getting credibility rating, taxon currency, or jurisdiction values. JSON¹⁴² and JSON-P¹⁴³ based services are also available.

2.6.4. Global Names Architecture

The Global Names Architecture¹⁴⁴ (GNA) is a system of databases, programs, and web services - a cyberinfrastructure - that will be used to discover, index, organise and interconnect on-line information about organisms and their names. It is a communal open environment that manages names so that we can manage information about organisms and serve the needs of biologists. The main component of the GNA is the Global Names Index¹⁴⁵ (GNI) that provides a list of all names that have been used for organisms. Within this list lie all of the nomenclaturally correct names, all of the names that are accepted as tokens for taxa, and all of the taxonomic metadata for biodiversity informaticians.

GNA offers access to various services and tools either as web services or Ruby implementations. The Global Names Recognition and Discovery (GNRD) service accepts text documents, images, and other files, performs OCR and discovers names in these files. The Global Names Index, as a service, resolves names against known sources. It uses exact or fuzzy matching as required. Its second version is in development. The Biblio service is a parser for discovery of bibliographic citations. All these services provide their output in JSON or XML format.

2.7. Linking in-situ and remote sensing data

All decisions about biodiversity are made using some measure that we think represents the essence of the problem. In the context of EU-BON, a process for linking raw data to Essential Biodiversity Variables and to Indicators is being designed. To successfully reach the goal of building a Biodiversity Observation Network, it will be necessary to access data from diverse providers, including field survey data, remote and field data sensors, which will cover different temporal and spatial scales. Integrating such heterogeneous sources of data, typically documented using discipline specific vocabularies, impose challenges for data management and interoperability. In this section, we describe previous efforts on promoting biodiversity data integration in Brasil within the Program for Planned Biodiversity Research¹⁴⁶ (PPBio), focusing on standardised ecological field surveys. We discuss challenges and opportunities for integrating field and remote sensing data and list some of the existing standards for data documentation, as well as opportunities for data improvement.

2.7.1. The PPBio example

PPBio's approach to connect biodiversity monitoring and decision making is based on spatial standardisation of plot surveys that is critical to answer most of the questions raised by decision makers. Although standardised at a spatial perspective, new scientific questions are generated constantly, which ultimately generate new variables to be collected and/or monitored in the field. PPBio's strategy allows flexibility and innovation and is supported by a data management workflow. PPBio's data policy is based on the concept that publicly funded research data should be disseminated and all partners are aware of the data workflow. Within such workflows, researchers are required to provide metadata 30 days after its collection. Metadata is provided using the Ecological Metadata Language (EML) standard and is stored at the PPBio Metacat instance¹⁴⁷. Data is stored one year after its collection and made publicly available two years after its collection. One person is exclusively responsible for data quality assurance and control, revising all metadata and data submitted before

¹⁴¹ http://www.itis.gov/ws_develop.html

¹⁴² <http://json.org/>

¹⁴³ <http://json-p.org/>

¹⁴⁴ <http://www.globalnames.org/>

¹⁴⁵ <http://gni.globalnames.org/>

¹⁴⁶ <http://ppbio.inpa.gov.br>

¹⁴⁷ <http://ppbio.inpa.gov.br/knb/style/skins/ppbio/>

they are inserted into the repository. This data curation procedure has proven to be crucial to allow anyone to understand, interpret and reuse the data. Data can go back and forth between the researcher and the curator up to three times before entering the database. Further developments to improve PPBio information management are focused on the standardisation of terms to avoid ambiguity through use of ontologies and controlled vocabularies, and on the documentation of analyses (using, e.g., R scripts) and workflows to allow reproducibility. From the conceptual perspective, linking field and remote sensed data is a challenge to be tackled to improve monitoring. There are initiatives already generating results, such as integrating LiDAR¹⁴⁸ and plot studies to assess light limitation in tropical forests and its consequences for biomass dynamics. Such complex approaches, linking remote sensing, hydrological and biological data are fostering the improvement of PPBio's capacity to document and organise data and workflows. While field survey data is mainly generated as tables, and best practices are oriented to generating logical schemas and correctly documenting table attributes and data variables, remote sensed data management requirements bring spatial reference systems information and processing and analytical workflows as crucial components to allow understanding and reproducibility.

Recommendation 24. EU BON should promote the establishment of institutionalised data policies, and also promote best practices through documents, training activities, web, etc.

Recommendation 25. EU BON should accept metadata in multiple formats that are in common use. When approached for a recommendation about which metadata standard to use, EU-BON should recommend a standard most appropriate for the data being described.

Recommendation 26. Metadata in the EU BON portal should be able to describe multiple types of primary biodiversity data, to support data discovery, provenance, interpretation and analytical reuse. It should be possible to search by space, time, taxa and theme. Institutions and custodians must be searchable. Related publications should be discoverable.

2.7.2. Standards and tools adopted by PPBio

Following is a summary of the standards and tools adopted by PPBio.

EML¹⁴⁹ and Metacat¹⁵⁰ are recommended for storing and disseminating field survey data. EML is a metadata specification for ecological data implemented as a series of XML document types that can be used in a modular and extensible manner. Allied to EML, PPBio proposed minimal variables related to place and time, a series of template tables with important information to be delivered such as sampling effort and attribute standardisations to facilitate data integration. Data and metadata curation by a human has also proven to be an important investment.

Regarding geospatial data, the OGC compliant ISO 19115:2003¹⁵¹ standard defines the schema required for describing geographic information and services. It provides information about the identification, the extent, the quality, the spatial and temporal schema, spatial reference, and distribution of digital geographic data. It is applicable to geographic data sets, data set series, and individual geographic features and feature properties. ISO 19115-2:2009¹⁵² extends the existing geographic metadata standard by defining the schema required for describing imagery and gridded data. It provides information about the properties of the measuring equipment used to acquire the data, the geometry of the measuring process employed by the equipment, and the production process used to digitise the raw data. This extension deals with metadata needed to describe the derivation of geographic information from raw data, including the properties of the measuring system, and the numerical methods and computational procedures used in the derivation. The metadata required to address coverage data in general is addressed sufficiently in the general part of ISO 19115.

¹⁴⁸ <http://en.wikipedia.org/wiki/Lidar>

¹⁴⁹ <http://knb.ecoinformatics.org/software/eml/>

¹⁵⁰ <http://knb.ecoinformatics.org/knb/docs/>

¹⁵¹ http://www.iso.org/iso/home/store/catalogue_tc/catalogue_detail.htm?csnumber=39229

¹⁵² http://www.iso.org/iso/home/store/catalogue_tc/catalogue_detail.htm?csnumber=26020

GeoNetwork¹⁵³ is a catalogue application to manage spatially referenced resources. It provides powerful metadata editing and search functions as well as an embedded interactive web map viewer. It is currently used in numerous Spatial Data Infrastructure (SDI) initiatives across the world, including Brazilian SDI.

The OGC and ISO Observations and Measurements¹⁵⁴ (O&M) standard specifies an XML implementation for the conceptual model (OGC Observations and Measurements v2.0 also published as ISO/DIS 19156), including a schema for Sampling Features. This encoding is an essential dependency for the OGC Sensor Observation Service (SOS) Interface Standard. More specifically, this standard defines XML schemas for observations, and for features involved in sampling when making observations. These provide document models for the exchange of information describing observation acts and their results, both within and between different scientific and technical communities.

Recommendation 27. EU BON should build its infrastructure by adopting and adapting existing technology where it meets the needs of the EU-BON infrastructure. If any software needs to be developed, it should be made available as open source.

Recommendation 28. EU BON should use controlled vocabularies and ontologies to enhance interoperability.

Recommendation 29. To enable accurate identification, versioning and citation, EU BON should promote the uptake of stable identifiers for data.

Recommendation 30. Considering the fact that EU BON is a global endeavour, it should, from the outset, address language and internationalisation issues.

Recommendation 31. Biodiversity EU BON should ensure that data curation is recognised as an important role.

2.8. Scholarly publishing, data papers, and digital literature

In terms of operational solutions, a variety of data publishing models is employed or being tested today. These models include (1) data published as supplementary files to articles, (2) data that is deposited and published through data repositories or data centres – either in standard community-agreed formats, or as generic files with a certain level of metadata – and then linked to journal articles, (3) data that is published in the form of marked up, structured and machine-readable text, and (4) data published retrospectively through markup of legacy publications (see also EU BON Task 3.4).

Most publishers nowadays clearly separate data from narrative (text). Moreover, data publishing through data centres and repositories has almost become a separate branch of the scholarly publishing industry preparing the landscape to publish data sets in a way similar to scholarly articles. This trend leads to formation of two groups that are closely interlinked to each other:

- Journal publishers
- Data publishers (data repositories, data centres)

Notwithstanding this diversity, these models have some common problems to solve – in particular to adopt a unified approach to cross-reference data and articles in a standardised, unique, and persistent manner, and to improve interoperability between different platforms through commonly accepted data and metadata standards.

Journal publishers use several means to publish data:

- Supplementary data files that underpin graphs, hypotheses, results, etc., are uploaded on the journal's website and published with the article.

¹⁵³ <http://geonetwork-opensource.org/>

¹⁵⁴ <http://www.opengeospatial.org/standards/om>

- Links to large and complex data sets in established international repositories (e.g. GBIF Integrated Publishing Toolkit (IPT)¹⁵⁵, Dryad¹⁵⁶, INSDC (GenBank/EMBL/DDBJ)¹⁵⁷, PANGAEA¹⁵⁸, TreeBASE¹⁵⁹, Morphbank¹⁶⁰, and others).
- Descriptions of large data sets – usually deposited in trusted international repositories - in the form of data papers.
- Detailed markup of the text to facilitate machine harvesting and automated dissemination of atomised content.
- Import and embedding of data into article text for several kinds of small data sets (e.g., Darwin Core occurrence data, checklists, tables of measurements, etc.). These can be downloaded from the article by users or harvested by machines. This format of data publishing is still in rudimentary form and is pioneered by the Biodiversity Data Journal¹⁶¹.

Data publishers provide stand-alone publication of data not necessarily underpinning a journal article. The examples below list the most important repositories for biodiversity-related data:

- Large primary biodiversity data sets (e.g., institutional collections of species-occurrence records): GBIF IPT.
- Genomic data are published with INSDC (GenBank/EMBL/DDBJ), either directly or *via* a partnering repository, e.g. Barcode of Life Data Systems (BOLD)¹⁶². Transcriptomics data are deposited and published in Gene Expression Omnibus (GEO)¹⁶³ or ArrayExpress¹⁶⁴.
- Phylogenetic data are published through TreeBASE.
- Morphological images related to taxa are deposited and published at Morphbank.
- MicroCT data could be deposited at emerging Morphosource¹⁶⁵ or other appropriate repository for this kind of data.
- Earth science and environmental data (including biodiversity data) are deposited at PANGAEA.
- Any other large data sets (e.g., ecological observations, environmental data, morphological and other data types) could be deposited and published in the Dryad Data Repository or other specialised institutional or international data repositories.
- Treatments that are deposited at Plazi and EOL.

2.8.1. Workflows for data publishing

There are a few workflows that attempt to streamline scholarly data publishing. We shall not analyse here the quite conventional mean of publishing of data files supplementary to an article, or standalone data publications at repositories, but focus on workflows that integrate scholarly narrative (text) and data publishing.

Amongst several impediments to facilitate access to data, one of the major bottlenecks is a lack of incentives for data publishers to publish their data resources. One of the solutions recommended by the GBIF Data Publishing Framework Task Group in order to overcome this impediment is the ‘data paper’ concept and associated workflow developed by GBIF and Pensoft (Chavan and Penev 2011). Earlier, this concept has been implemented by the Ecological Society of America through their journal Ecological Archives¹⁶⁶, and Earth System Science Data (ESSD)¹⁶⁷ journal of Copernicus¹⁶⁸.

¹⁵⁵ <http://ipt.gbif.org/>

¹⁵⁶ <http://www.datadryad.org/>

¹⁵⁷ <http://www.insdc.org/>

¹⁵⁸ <http://www.pangaea.de/>

¹⁵⁹ <http://treebase.org>

¹⁶⁰ <http://www.morphbank.net/>

¹⁶¹ <http://biodiversitydatajournal.com/>

¹⁶² <http://www.boldsystems.org/>

¹⁶³ <http://www.ncbi.nlm.nih.gov/geo/>

¹⁶⁴ <http://www.ebi.ac.uk/arrayexpress/>

¹⁶⁵ <http://morphosource.org/>

¹⁶⁶ <http://esapubs.org/archive/>

¹⁶⁷ <http://www.earth-system-science-data.net/>

¹⁶⁸ <http://www.copernicus.org/>

The data paper is a scholarly journal publication whose primary purpose is to describe a data set or a group of data sets, rather than to report a research investigation. As such, it contains facts about data, not hypotheses and arguments supported by the data, as found in a conventional research article. Its purposes are three-fold: (a) to provide a citable journal publication that brings scholarly credits to data publishers; (b) to describe the data in a structured human-readable form; and (c) to bring the existence of the data to the attention of the scholarly community.

The workflow, established by GBIF and Pensoft, generates data paper manuscripts automatically, at the “click of a button”, from the extended metadata descriptions in the IPT (based on EML). Then manuscripts are submitted to the journal and undergo a peer review and editorial process. To date, over 20 data papers¹⁶⁹ have been published, mainly in the journals ZooKeys, PhytoKeys and Nature Conservation.

Another scholarly data publishing network has been attempted through the BioFresh¹⁷⁰ FP7 project. Seventeen editors of journals dealing with freshwater biota and environment agreed to encourage authors to publish the raw data that underpin the submitted manuscripts. As far as is known, this workflow has not yet produced visible results.

An alternative workflow targets the still predominant traditional publications in which data is not marked up, thus not easily re-usable. To convert such unstructured publications into structured, semantically enhanced publications, Plazi¹⁷¹ developed a specific editor (GoldenGate¹⁷²) that allows semi-automatic markup of taxonomic publications to a high degree of granularity (treatments to materials citations), a repository and export facility to GBIF, EOL and other data recipients (Agosti and Egloff, 2009). Task 3.4 is dealing specifically with the extraction of materials citation data from legacy publications.

2.8.2. EU BON and scholarly data publishing

EU BON should encourage and support two means of scholarly data publishing:

- Development of scholarly *data paper* publishing workflows between the major repositories of data types relevant to EU BON goals (e.g., genomic, species occurrences, species populations, functional traits, environmental observations, etc.) with academic publishers. Data papers are appropriate for publishing of large data sets (for example institutional collections, long-term environmental monitoring data, etc.).
- Development of workflows that shorten the distance between data and narrative (text) publishing through advanced mark up and data import/export technologies. This workflow is appropriate for mobilisation of small isolated data, usually collected by individual researchers, citizen scientists, research groups for a particular study, etc.

Recommendation 32. EU BON should develop, demonstrate and promote scholarly data publishing using workflows adapted to both large and small data sets respectively.

Recommendation 33. EU BON should demonstrate mobilisation and re-usability of data, these being either digitally born/published or extracted from legacy publications through markup and databasing.

Implementation of these two approaches will be piloted by means of sample papers, involving repositories for different types of data, and journals that are able to provide highly automated ways for accepting and peer-reviewing manuscripts through XML-based editorial workflows and domain-specific markup. To the best of our knowledge, currently the Biodiversity Data Journal¹⁷³ seems to be the only candidate for this.

The main aim of these pilots would be to provide a streamlined mechanism for a high-level of automation of manuscript generation and submission from rich metadata descriptions. Vice versa, the

¹⁶⁹ <http://www.pensoft.net/page.php?P=23>

¹⁷⁰ <http://data.freshwaterbiodiversity.eu/submitdata.html>

¹⁷¹ <http://plazi.org/>

¹⁷² <http://plazi.org/?q=GoldenGATE>

¹⁷³ <http://biodiversitydatajournal.com/>

pilot should also demonstrate the reverse link that is export of data from journals to repositories and aggregators. For example, small peer-reviewed data sets (e.g. Darwin Core occurrence data for different taxa within an article) published in a computer-readable, XML markup format, could be made available for automated harvesting by data aggregators (e.g., GBIF).

To illustrate the above approaches, we propose to pilot and test the following workflows with sample papers for different data types:

- **Use case 1:** Fully automated manuscript generation and submission of data paper manuscripts in EML format, from the GBIF IPT to the Biodiversity Data Journal. This will be a step forward in comparison to the current mechanism of creating manuscripts in RTF from EML metadata through the GBIF IPT publishing module. The RTF files are further submitted to the journals in a conventional way which leads to loss of markup and return mechanisms from the journal to the IPT.
- **Use case 2:** Multiple submissions of data paper manuscripts describing data sets of a common type, e.g., distributions of various major freshwater taxa in Europe. This workflow will be elaborated by the BioFresh data publishing network, based on the GBIF IPT and the Biodiversity Data Journal.
- **Use case 3:** Data mobilisation from legacy literature through re-publishing of data-rich volumes of national flora/fauna or regional floristic, faunistic or taxonomic revisionary works. The workflow will be piloted by (1) Flora of Slovakia volume in collaboration with the EU BON partners SAVBA, PENSOFT, GBIF and Plazi, and (2) Flora of Northumberland and Durham (1838), in collaboration with the EU BON partners NBGB, PENSOFT, GBIF and Plazi.
- **Use case 4:** Development of advanced publishing model for large metagenomic data sets (in collaboration with WP 1 and WP2). The EU BON partners UTARTU and PENSOFT will produce two pilot projects through a novel UNITE¹⁷⁴/PlutoF¹⁷⁵ - Biodiversity Data Journal workflow that will (1) automate the export of metadata of fungal species that are identified as new to science into a manuscript template for Pensoft's Writing Tool (PWT) that will facilitate their formal description through peer-review and publishing; (2) produce a data paper describing ca. 100,000 fungal Species Hypotheses (SP) of UNITE version 6.0, to be used for the purposes of environmental metagenomics and monitoring.
- **Use case 5:** Mobilisation and aggregation of verified, peer-reviewed, small, occurrence data sets through a harvesting mechanism established by GBIF to extract Darwin Core Archives published in the Biodiversity Data Journal.
- **Use case 6:** Mobilisation and aggregation of taxon treatments through a harvesting mechanism established by EOL to extract Darwin Core Archives published in the Biodiversity Data Journal.

2.8.3. Markup formats

Marked-up text is more amenable to machine processing. Currently, two different XML schemas for mark-up (TaxonX¹⁷⁶ and TaxPub¹⁷⁷) are available. TaxonX and TaxPub both are used for taxonomic literature but for different corpora. TaxonX has been developed to markup legacy publications with the goal of data extraction. Furthermore only elements have been created that are unique to taxonomic literature (e.g., treatment) and all the other elements are imported from existing schemas. It does not aim to model an entire publication. On the other hand, TaxPub is self-contained and does not refer to external schemas, although all the elements can be mapped to other existing vocabularies (e.g., Darwin Core). It contains all the elements to model not only the semantics of a publication but all the publishing artifacts. TaxonX is an independent schema, whilst TaxPub is a domain specific flavour of the widely used JATS (Journal Archiving Tag Suite of the US National Library of Medicine) allowing import of TaxPub based articles into PubMed and PubMed Central.

¹⁷⁴ <http://unite.ut.ee/>

¹⁷⁵ <http://plutof.ut.ee/>

¹⁷⁶ <http://www.taxonx.org>

¹⁷⁷ <http://www.ncbi.nlm.nih.gov/books/NBK47081/>

TaxonX

TaxonX is a lightweight (with only 30+ elements) and flexible schema for mark-up of treatments which can be quickly learned and may be applied to the wide variety of formatting present in legacy documents as well as new publications. The goal of TaxonX is to model taxon treatments in publications to provide a basis for data mining and extraction, while generic textual features are given marginal importance. In many cases, it relies on use of external schemas for modelling certain kinds of information, e.g., the use of MODS¹⁷⁸ (Metadata Object Description Schema) for file level bibliographical metadata and Darwin Core for observation data. It has loose content requirements that allow for a wide variety of instances to be encoded over time and at many levels of granularity, while maintaining validity through iterations. Additionally, TaxonX contains mechanisms for semantic normalisation of the data contained in treatments.

Development of TaxonX began at the American Museum of Natural History (AMNH) and continued through the duration of a subsequent NSF/DFG grant (see below). As the project was concluding, participants established Plazi, a Switzerland-based independent not-for-profit organisation aimed at helping to remove technological, social, and legal barriers to the creation of, and access to, taxonomic literature. Among its many activities, Plazi maintains the TaxonX schema and a repository of XML-encoded publications and develops the semi-automatic mark-up tool GoldenGATE (Sautter *et al.* 2007).

TaxonX provides for the encoding of taxon treatments, with elements for the major structural components of treatments (e.g., Nomenclature, Materials examined, Description, etc.) and phrase-level features of interest in taxonomy (e.g., scientific names, locality names, characters, etc.) as well as mechanisms for linking to external resources and the semantic normalisation of terms mentioned in the source document. The TaxonX instances encoded by Plazi contain a moderate degree of mark-up. Bibliographic metadata for the source documents are provided in each instance. Other sections of treatments are identified and named when they occur, but are not always present due to the wide variability of the structure of the source documents. Many scientific names are marked and associated with a Life Science Identifier (LSID), but other features may not always be identified. The section “Materials examined” can be broken down to individual materials citations, which in turn may be normalised and linked to external resources such as a type specimen through LSIDs or other links.

A special emphasis has been given to link data to external resources. Tools in GoldenGATE have been developed to communicate automatically with external sources such as nameservers to retrieve LSIDs for taxonomic names in case they have already been entered, or to enter names upon discovery in an article, create the record and subsequently retrieve the generated LSID (e.g., in collaboration with the Hymenoptera Name Server¹⁷⁹), or on a manual basis with Zoobank.

TaxPub

Prior experiences with retrospective conversion showed that any schema attempting to model the broad range of stylistic, editorial, and formal variation of legacy taxonomic literature would be so loose as to greatly challenge interchange as well as development of consuming applications. TaxPub, conversely, was designed to be adequately constrained to facilitate interchange and application development. It was hoped that such constrained mark-up could be applied more easily during either the authorial or editorial stages rather than after publication. TaxPub is an extension of the Journal Publishing Tag Set of the U.S. National Library of Medicine’s Journal Archiving Tag Suite¹⁸⁰. For more details see Catapano (2010).

Starting in 2008, TaxPub was designed and developed by members of Plazi with the assistance of experts from the U.S. National Center for Biotechnology Information. The TaxPub extension¹⁸¹ is maintained as an open source project at SourceForge, inheriting from the base DTD an extensive and

¹⁷⁸ <http://www.loc.gov/standards/mods/>

¹⁷⁹ http://ptp.pensoft.eu/external_details.php?type=1&query=Hymenoptera

¹⁸⁰ <http://dtd.nlm.nih.gov/>

¹⁸¹ <http://sourceforge.net/projects/taxpub/>

robust set of elements for generic textual structures while adding a small number of elements relevant to taxonomy. A few phrase-level elements are made available at relevant places throughout the DTD. There are elements for scientific names, <tp:taxon-name>, citations of specimens and other materials, <tp:material-citation> and descriptions of organisms' physical characteristics, <tp:descriptive-statement>. The "taxon treatment" is the focus of TaxPub. Following publishing traditions in taxonomy, a taxon treatment is a formal description of a taxon, including sections on nomenclature, morphological characteristics, behaviour, ecology, distribution, and specimens examined. TaxPub primarily models these taxon treatment features, providing (within a namespace with the prefix "tp") a <tp:taxon-treatment> element with a required <tp:nomenclature> element which is highly structured and contains the essential data about the named species and an <tp:treatment-sec> element for other sub-sections for which a treatment-sec-type attribute provides specific semantics. Beyond these elements TaxPub relies on the elements in JATS "Blue" DTD for all other features. In particular, the <named-content> element is intended to be used for the wide range of phrase level data which may be of interest in taxonomy (e.g., locality information such as latitude, longitude, elevation, etc...).

Since July 2009, TaxPub has been routinely implemented in the everyday publishing practice of Pensoft for its journal ZooKeys, and later PhytoKeys, to provide: (1) Semantically enhanced, domain-specific XML versions of articles for archiving in PubMedCentral (PMC); (2) Visualisation of taxon treatments on PMC; (3) Export of taxon treatments to various aggregators, such as Encyclopedia of Life, Plazi Treatment Repository, and the Wiki Species-ID.net (Penev et al. 2010).

Use case 7: Mobilisation and aggregation of occurrence data published in legacy literature through harvesting mechanism established by GBIF to extract Darwin Core Archives produced by Plazi.

Use case 8: Mobilisation and aggregation of taxon treatment data published in legacy literature through harvesting mechanism established by EOL to extract Darwin Core Archives produced by Plazi.

Recommendation 34. EU BON should work towards integration of digitally born data and legacy data from historical literature.

Recommendation 35. EU BON should support extensions of TaxonX and TaxPub schemas for linking to other domains (ecology, genomics, etc.) through taxon names and other important data elements.

2.9. Vocabularies and ontologies

Common vocabularies are essential for interoperability, helping to ensure that data are understandable and usable across systems. The Global Biodiversity Informatics Outlook¹⁸² (GBIO) identified the need to develop such vocabularies as a priority in the short term, with the long term goal of their evolution into ontologies that can support semantic reasoning and multi-disciplinary integration. This is an area undergoing active development in the biodiversity community, e.g., through the work of the TDWG Vocabulary Management Task Group (VoMaG) and RDF Task Group, and the ongoing development of the Biological Collections Ontology¹⁸³, the Environment Ontology¹⁸⁴, and the Population and Community Ontology¹⁸⁵ (Walls et al. 2013). The VoMaG¹⁸⁶ report which built on earlier vocabulary related work in the ViBRANT project reviews the status of the TDWG ontologies and provides suggestions for advancement, including best practices for development, maintenance and use of vocabularies and evaluation of Semantic MediaWiki¹⁸⁷ as a community platform for defining terms.

Recommendation 36. EU BON should promote the standards and tools needed to structure data into a linked format by using the potential of vocabularies and ontologies for all biodiversity facets, including: taxonomy, environmental factors and ecosystem functioning and services.

¹⁸² <http://www.gbif.org/resources/2251>

¹⁸³ <http://bco.googlecode.com/svn/trunk/src/ontology/bco.owl>

¹⁸⁴ <http://purl.obolibrary.org/obo/envo.owl>

¹⁸⁵ <http://purl.obolibrary.org/obo/pcio.owl>

¹⁸⁶ <http://www.gbif.org/resources/2246>

¹⁸⁷ <http://terms.gbif.org>

Recommendation 37. EU BON should track developments concerning vocabularies and ontologies in TDWG and other fora and ensure best practices are adopted in implementing the EU BON architecture.

Recommendation 38. To support semantic interoperability, EU BON should, where possible, re-use established vocabularies, thesauri and ontologies.

2.10. Stable identifiers

Without stable, persistent identifiers for biodiversity resources (e.g., data sets, individual records, specimens, taxon names) data integration becomes an extremely difficult, if not impossible, task. Stable identifiers are essential for tracking resources on a distributed network, as an aid to understanding data provenance and data quality, resolving duplications, and for supporting citation and data annotation. Several identifier systems exist ranging from simple HTTP based URIs to Life Sciences Identifiers (LSIDs) and Digital Object Identifiers (DOIs). Most recently, the community appear to be converging on the use of http-URIs because of their relative simplicity and added advantage of supporting Linked Open Data, and DOIs for data sets because they support “citable data sets” through mechanisms such as DataCite¹⁸⁸. GBIF has consulted with the community on the use of the latter, including the possibility of creating stable URIs for individual records within data sets by appending a local identifier to the data set DOI. Likewise, the Consortium of European Taxonomic Facilities (CETAF) and pro-iBiosphere are encouraging the use of http-URIs as stable identifiers for collection specimens¹⁸⁹ with initial uptake among such institutions as BGBM, Paris-MNHN, Kew, RBGE Edinburgh, and MfN Berlin.

Recommendation 39. EU BON should promote the use of stable, persistent identifiers particularly at the data set and record/specimen levels.

¹⁸⁸ <http://datacite.org>

¹⁸⁹ <http://stories.rbge.org.uk/archives/3846>

3. Conclusions

3.1. Reasons for data heterogeneity

There are many causes of data heterogeneity in the biodiversity domain. Foremost, probably, is the traditional fragmentation into sub-disciplines such as terrestrial, freshwater and marine that creates communities that tend to be isolated from each other. Also, the majority of biodiversity studies are typically carried out by small teams or even single researchers in a culture where sharing and archiving data for reuse is not yet the norm. Thus, most scientists are free to develop or use ad hoc encodings and file formats for their data. By contrast, genetic sequence data is standardised because journals require, as a condition of publication, that the data must be archived in an INSDC repository (e.g., GenBank) following a standard format.

Different sampling protocols and field techniques also contribute to data heterogeneity and need to be documented (in metadata) to understand and potentially draw comparisons across data sets, e.g., those involving grid-based sampling and surveillance data with different grid sizes.

Overcoming data heterogeneity thus requires the adoption of standards for both the data and metadata, in a balance of harmonisation and mediation activities (see Section 1.3.1): particular communities of practice will typically develop common standards which they apply to their data, but for cross discipline data integration, mapping between community standards (i.e., mediation) using Knowledge Organisation Systems such as thesauri and ontologies is required.

A fundamental task in combating data heterogeneity, therefore, is development and uptake of suitable vocabularies and ontologies and their translation to accommodate multilingualism, coupled with development of a range of tools and applications that work with them. The various types of data EU BON will need to deal with are of very different structure and, most importantly, are currently at different levels of standardisation. In fact, the only two major types of data that can be published, indexed and collated are genomic data [INSDC¹⁹⁰ (GenBank/EMBL/DDBJ)] or occurrence data (Darwin Core, GBIF). Close to standardisation are taxon treatment data, mostly based on structured narrative but harvestable text. Most other biodiversity-related data types (e.g., various ecological data or environmental measurements) are still being published in non-interchangeable formats and, in fact, their “publication” means making them openly accessible online with accompanying metadata of differing levels of detail and integrity.

The EU BON platform must, in turn, be built on these standards and, *via* standardised web services and protocols, provide discovery, access and analytical workflows. In addition, a stable, long term infrastructure can counter data heterogeneity. EU BON will support this by developing its information architecture in compliance with the LifeWatch architecture. Thus, whilst the efforts of the informatics community produces better and smarter tools and standards, parallel efforts need to be made to encourage researchers and potential data providers to actively engage in the implementation process, and ensure their data are in compliance.

3.2 Overcoming impediments for data sharing

GEO has just been renewed for another 10 years¹⁹¹. “Open data” for benefits to society is the grand theme. The biodiversity community has been practicing this already for more than ten years, mainly through the efforts of GBIF, and we can arguably say that biodiversity is further ahead in its efforts than most GEO Societal Benefit Areas. Still, those 420 million data records openly accessible through GBIF represent no more than 20-40% of the already existing, digital data records, of which we know (cf. Ariño 2010). Open access is prescribed in principle, but implementation is lacking. Still, museums, agencies, research groups and citizen science associations can afford not to share their data. The reason is simple: There is no funding or effective credit mechanism for data sharing, least so for digitisation. It is not likely that such mechanisms appears anytime soon, but what EU BON can do is

¹⁹⁰ <http://www.insdc.org>

¹⁹¹ http://www.earthobservations.org/documents/pressreleases/pr_20140117_geneva_ministerial.pdf

to lower the technical barriers for data sharing, and promote agreements that support data sharing. Open access is a fine principle, but many research groups do not believe that they get the credit for their efforts by that way. EU BON needs to analyse the situation and gain trust of data owners. By trust we mean here a mechanism that generates scientific credit and enhanced opportunity for those who share their data. EU BON will approach this with a repository infrastructure tailored for the needs of monitoring networks, and by standardising ecological data so that it can be better integrated. In order to succeed with this, EU BON needs to understand the particular situations of the various monitoring networks and projects. This is not a small task, as alone in Europe, there are up to 2000 such data holders. The EU BON Helpdesk has a huge challenge in reaching out to them.

3.3. Need for new and enhanced data standards

Central to the success of GEO BON is increasing cooperation among the standards organisations with interests in the biodiversity science domain (Hardisty *et al.* 2013). When the EU BON project was started, many questioned the need for new and enhanced data standards, as such need was mentioned in the DoW. Arguably, there are already too many standards. The existing standards are being used in large scale, and Darwin Core is evolving continuously to fulfil the growing needs. However, during 2013 and while preparing this review, it became obvious that the standards we have today are not sufficient to meet the needs of EU BON. This is best manifested by the fact that in the ecosystem community no similar standard to Darwin Core exists, and subsequently, integrating ecosystem data is hard. There is now a proposal to extend Darwin Core towards ecosystem attributes (Wieczorek *et al.* 2013). First steps to that direction have been taken by GBIF, TDWG, and VegBank, which work has been reviewed here. We will need to put these extended standards to work at EU BON test sites (WP5). Based on these experiences, we will need to push them further. The authors of this document are convinced that promising, new developments for integrating biodiversity, ecosystem, and possible agriculture data are underway, and a breakthrough is near.

3.4. EBVs, modelling and data flows

“The grand challenge for biodiversity informatics is to develop an infrastructure to allow the available data to be brought into a coordinated coupled modelling environment able to address questions relating to our use of the natural environment that captures the ‘variety, distinctiveness and complexity of all life on Earth.’” (Hardisty *et al.* 2013). That is a proper citation to discuss the ongoing thrust towards use of Essential Biodiversity Variables to the science-policy interface. EBVs cannot be measured directly – they must be modelled and computed from a large number of inputs streaming in from a multitude of sources (see section 1.2.2). Unlike indicators, which are complex data sets with their interpretations, EBVs are just plain numbers. In that sense, an EBV is like a stock market index, which is continuously updated. Hence, EBVs will bring a new element into environmental discussion.

Computing EBVs will require orchestrating data flows from disparate, heterogeneous sources. There has been no shortage of initiatives for streamlining environmental data flows. Most of these have failed or are lingering, because re-engineering business processes in an international setting has many constraints. It took 15 years for the climate change community to agree on a set of fifty Essential Climate Variables. The biodiversity community can learn from this, and perhaps advance faster. The fact that there is no system to re-engineer, but the work can start from clean slate may help. EU BON will be prototyping the EBVs in coming years. This will require intense communication within the project and with the constituents of GEO BON and IPBES.

There is a challenge of integrating incidental data from collections and citizen science activities with that of systematic monitoring. When the number of observers is very large, it is possible to estimate collecting activity and normalise data so that collecting bias can be dealt with. Such data can then be verified from systematic surveys and monitoring. If both show the same trend, it can probably be used as evidence of real phenomena (cf. Kery *et al.* 2010). However, more research is required on this subject.

In any case, the extent and importance of citizen-collected data is going to increase. Extensive research is needed on how to make most use of these resources. Many citizen groups do not have proper information systems to support their work. The EU BON project could probably help in that regard.

3.4 Supporting GEO and IPBES processes

The biodiversity community needs to establish itself visibly in GEOSS. The community has a lot of experience in open data access and data integration, which will be a valuable contribution to the entire GEO process, when properly explained. We have come very far in big data, but the value of that achievement has not yet been fully understood and appreciated in a way it deserves. It is time to share that experience and become a formidable player in the GEO process. Biodiversity deserves that.

Yet the shortcomings must be admitted. Integration of remote sensing data and *in situ* biodiversity data has not been achieved, except in technical demonstrations. This document is sorely lacking detailed descriptions of this. Linking habitat data and species data will be required. During the EU BON project this void must be filled.

The EU BON project will establish a portal, registry and repository services, and provide support through a training programme and helpdesk which will respond to the needs brought forward in the GEO BON Detailed Implementation Plan. Yet the game has moved on since 2010 when those plans were made. Essential Biodiversity Variables, when made operational, will finally have crystallised the user needs that provide the focus for all the ICT developments.

Beyond GEO BON, the IPBES will require support in data management. Already in 2014, the IPBES is expected to prepare a number of assessments which all require large volumes of data. The work that will be done by distributed research groups needs a data repository service, where all data used in assessments is transparently available for verification and scrutiny. The repository structure suggested in this document can provide that function.

References

- Agosti D, Egloff W (2009) Taxonomic information exchange and copyright: the Plazi approach. *BMC Research Notes* 2009, 2: 53 doi:10.1186/1756-0500-2-53
- Ariño AH (2010) Approaches to estimating the universe of natural history collections data. *Biodiversity Informatics* 7:81–92.
- Catapano T (2010) TaxPub: An extension of the NLM/NCBI journal publishing DTD for taxonomic descriptions. In: *Journal Article Tag Suite Conference (JATS-Con) Proceedings 2010* [Internet]. Bethesda (MD): National Center for Biotechnology Information (US); 2010. Available from: <http://www.ncbi.nlm.nih.gov/books/NBK47081/>
- Chavan V, Penev L (2011) The data paper: a mechanism to incentivize data publishing in biodiversity science. *BMC Bioinformatics*, 12 (Suppl 15): S2. doi: [10.1186/1471-2105-12-S15-S2](https://doi.org/10.1186/1471-2105-12-S15-S2)
- Foody, G.M. & Cutler, M.E.J. 2003. Tree biodiversity in a region of conserved and differentially logged tropical rainforest and its measurement by satellite remotesensing. *Journal of Biogeography*, 30: 1053-1066.
- Gillespie TW, Foody GM, Rocchini D, Giorgi AP, Saatchi S (2008) Measuring and modelling biodiversity from space. *Progress in Physical Geography*, 32: 203-221.
- Hardisty A, Roberts D, The Biodiversity Informatics Community (2013) A decadal view of biodiversity informatics: challenges and priorities. *BMC Ecology* 13:16 (<http://www.biomedcentral.com/1472-6785/13/16>)
- Hernandez-Ernst V, Poigné A, Giddy J, Hardisty A, Voss A, Voss H (2009) Towards a reference model for the Lifewatch ICT infrastructure. *Lecture notes in informatics* 154
- Hernandez-Ernst V, Poigné A, Giddy J, Hardisty A (2009) Data and modelling tool structures reference model. *Lifewatch Deliverable 5.1.3*.
- Hoffman A, Penner J, Vohland K, Cramer W, Doubleday R, Henle K, Köljalg U, Kühn I, Kunin WE, Negro JJ, Penev L, Rodriguez C, Saarenmaa H, Schmeller DS, Stoev P, Sutherland WJ, Ó Tuama É, Wetzel F, Häuser C (2014) The need for an integrated biodiversity policy support process – Building the European contribution to a global Biodiversity Observation Network (EU BON). 20 p, submitted.
- Inspire Thematic Working Group Species Distribution (2013) Data specification on species distribution – Draft Technical Guidelines v.3.0rc3.
- Kerr JT, Ostrovsky M (2003) From space to species: ecological applications for remote sensing. *TRENDS in Ecology and Evolution* 18 (6): 299-305.
- Kery M., Royle A., Schmid H, Schaub M., Volet B, Häfliger G, Zbinden N (2010) Site-occupancy distribution modeling to correct population-trend estimates derived from opportunistic observations. *Conservation Biology* 24: 1388-1397.
- Michener WK, Brunt JW, Helly JJ, Kirchner TB, Stafford SG (1997) Nongeospatial metadata for the ecological sciences. *Ecological Applications* 7 (1): 330-342.
- Hugo W, Saarenmaa H, Schmidt J (2013) Development of extended content standards for biodiversity data. *European Geosciences Union (EGU) General Assembly 2013*. Vienna, 8-12 April 2013. *Geophysical Research Abstracts*, Vol. 15, EGU2013-6968, 2013.
- Nativi S, Craglia M, Pearlman J (2012) The brokering approach for multidisciplinary interoperability: A position paper. *International Journal of Spatial Data Infrastructures Research*, 2012, Vol.7, 1-15.
- Parviainen, M., Luoto, M. & Heikkinen, R.K. 2009. The role of local and landscape level measures of greenness in modelling boreal plant species richness. *Ecological Modelling*, 220: 2690-2701.
- Penev L, Agosti D, Georgiev T, Catapano T, Miller J, Blagoderov V, Roberts D, Smith V, Brake I, Rycroft S, Scott B, Johnson N, Morris R, Sautter G, Chavan V, Robertson T, Remsen D, Stoev P, Parr

- C, Knapp S, Kress WJ, Thompson F, Erwin T (2010) Semantic tagging of and semantic enhancements to systematics papers: ZooKeys working examples. *ZooKeys* 50: 1-16. doi: 10.3897/zookeys.50.538
- Penev L, Lyal C, Weitzman A, Morse D, King D, Sautter G, Georgiev T, Morris R, Catapano T, Agosti D (2011) XML schemas and mark-up practices of taxonomic literature. *ZooKeys* 150: 89-116. doi: 10.3897/zookeys.150.2213
- Pereira HM, Ferrier S, Walters M, Geller GN, Jongman RHG, Scholes RJ, Bruford MW, Brummitt N, Butchart SHM, Cardoso AC, Coops NC, Dulloo E, Faith DP, Freyhof J, Gregory RD, Heip C, Höft R, Hurr G, Jetz W, Karp DS, McGeoch MA, Obura D, Onoda Y, Pettorelli N, Reyers B, Sayre R, Scharlemann JPW, Stuart SN, Turak E, Walpole M, Wegmann M (2013) Essential Biodiversity Variables. *Science* 339: 277-278. doi: 10.1126/science.1229931
- Sautter G, Agosti D, Böhm K (2007) Semi-Automated XML Markup of Biosystematics Legacy Literature with the GoldenGATE Editor. Proceedings of PSB 2007, Wailea, HI, USA, 2007. <http://psb.stanford.edu/psb-online/proceedings/psb07/sautter.pdf>
- Scholes RJ, Walters M, Turak E, Saarenmaa H, Heip CHR, Ó Tuama É, Faith DP, Mooney HA, Ferrier S, Jongman RHG, Harrison IJ, Yahara T, Pereira HM, Larigauderie A, Geller G (2012) Building a global observing system for biodiversity. *Current Opinion in Environmental Sustainability* 4: 139-146. <http://dx.doi.org/10.1016/j.cosust.2011.12.005>
- Ó Tuama É, Saarenmaa H, Nativi S, Bertrand N, van den Berghe E, Scott L, Lane M, Cotter G, Canhos D, Khalikov R (2010) Principles of the GEO BON information architecture. Group on Earth Observations (Geneva), 42 p. http://www.earthobservations.org/documents/cop/bi_geobon/geobon_information_architecture_principles.pdf
- Walls R *et al.* [22 co-authors] (2013). Semantics in support of biodiversity knowledge discovery: an introduction to the Biological Collections Ontology and related ontologies. Submitted to PLOS.
- Wieczorek J, Bloom D, Guralnick R, Blum S, Döring M, De Giovanni R, Robertson T, Vieglais D (2012) Darwin Core: An evolving community-developed biodiversity data standard. *PLoS ONE* 7(1): e29715. doi:10.1371/journal.pone.0029715.
- Wieczorek J, Bánki O, Blum S, Deck J, Döring M, Dröge G, Endresen D, Goldstein P, Leary P, Krishtalka L, Ó Tuama É, Robbins RJ, Robertson T, Yilmaz P (2013) Meeting Report: GBIF hackathon-workshop on Darwin Core and sample data (22-24 May 2013). <http://www.gbif.org/resources/2245>
- Wooley JC, Godzik A, Friedberg I (2010) A primer on metagenomics. *PLoS Comput Biol* 6: 1000667. <http://dx.doi.org/10.1371/journal.pcbi.1000667>.
- Yilmaz P, Kottmann R, Field D, Knight R, Cole JR, Amaral-Zettler L, Gilbert JA, *et al.* (2011) Minimum information about a marker gene sequence (MIMARKS) and minimum information about any (x) sequence (MIXS) specifications." *Nat Biotech* 29(5): 415–420. doi:10.1038/nbt.1823.

Annex I: Related European projects

While it is impossible to discuss all 324 EU biodiversity initiatives, some other notable projects include the following projects and networks, in alphabetic order.

BioFresh

Full name: Biodiversity of Freshwater Ecosystems: Status, Trends, Pressures, and Conservation Priorities

Type: FP7 project

Duration: 2009 - 2014

Website: <http://www.freshwaterbiodiversity.eu/>

The EU FP7 project BioFresh aims to bring freshwater biodiversity data and information together and make it publicly available. The current focus of this work (relevant in the context of EU BON) is the creation of a metadatabase to document existing data sets relevant in the field of freshwater biodiversity and set up a freshwater occurrence database which contributes to the GBIF network.

For its metadatabase, BioFresh compiled an extensive selection of fields relevant for its network of freshwater biodiversity scientists. The information is organised in the following categories; 1) General information, 2) Technical specifications, 3) Intellectual property rights and citation, 4) General data specifications, 5) Site specifications, 6) Climate and environmental data, 7) Biological data, 8) Sample specifications/sample resolution, and 9) Other specifications.

While more extensive than any of the existing metadata (exchange) standards, mapping with the following standards was ensured during the design of the metadatabase; Directory Interchange Format (DIF; used e.g. in NASA's Global Change Master Directory (GCMD)), the GBIF Profile of the Ecological Metadata Language (EML), and the Federal Geographic Data Committee (FGDC) endorsed metadata standard ISO 19115 for geographic information. Export functionality to EML is available from the metadatabase and is used for publishing (meta) data to the GBIF network using the Integrated Publishing Toolkit (IPT).

For exchanging primary biodiversity in the freshwater realm, both within the BioFresh and GBIF networks, the Darwin Core (DwC) standard was adopted. Templates were developed with a selection of fields which are most relevant and useful specifically for freshwater related data sets. This work was inspired by the Apple Core¹⁹² - Darwin Core documentation and recommendations for herbaria - developed by Peter Desmet and similarly named "freshwater core". Templates and documentation are available¹⁹³ ¹⁹⁴. As part of this documentation, the use of one additional field for internal use, "waterBodyType", is recommended. This optional field for specifying the type of waterbody is adopted to indicate whether the locality is a river, lake, pond, wetland or groundwater sampling point. The currently proposed control vocabulary was chosen for consistency with the BioFresh metadatabase, but the use of an existing thesaurus (or ones being developed, e.g. through the European Environment Agency) could be considered when proposing this term for integration in the DwC standard.

BIO_SOS

Full name: Biodiversity Multi-Source Monitoring System from Space to Species

Type: FP7 project

Duration: December 2010 – 2013

¹⁹² <http://code.google.com/p/applecore/>

¹⁹³ <http://data.freshwaterbiodiversity.eu/submitdata.html#submspreadsheet>

¹⁹⁴ <http://code.google.com/p/freshwatercore/downloads/list>

Website: <http://www.biosos.eu/index.htm>

The main objective of BIO_SOS is the development of a knowledge-based pre-operational ecological modelling system suitable for effective and timely multi-annual monitoring of NATURA 2000 sites and their surrounding areas particularly exposed to different and combined type of pressures.

BIO_SOS aims to develop novel pre-operational automatic high spatial resolution (HR), very high spatial resolution (VHR) EO data understanding techniques and provide land cover/use (LCLU) map and LC/LU change map generation as Copernicus (GMES) core services; BIO_SOS will also develop a metadata geo-portal compliant with the Group on Earth Observations (GEO), GEOSS and INSPIRE initiatives.

Relevant documents and information for data standards

- D. Torri, J. Poesen [A review of topographic threshold conditions for gully head development in different environments](#). Earth-Science Reviews. (Accepted for publication). ([Download pdf file](#))
- Kosmidou, V., Petrou, Z.I., Bunce, R.G.H., Múcher, C.A., Jongman, R.H.G., Bogers, M.M., Lucas, R.M., Tomaselli, V., Blonda, P., Padoa-Schioppa, E., Manakos, I., Petrou, M., 2013. Harmonization of the Land Cover Classification System (LCCS) with the General Habitat Categories (GHC) classification system. Ecol. Indic. Accepted for publications on 26 July 2013.
- Damien Arvor, Laurent Durieux, Samuel Andrés, Marie-Angélique Laporte, 2013. [Advances in Geographic Object-Based Image Analysis with ontologies: A review of main contributions and limitations from a remote sensing perspective](#). ISPRS Journal of Photogrammetry and Remote Sensing. Volume 82, Pages 125–137.
- Valeria Tomaselli, Panayotis Dimopoulos, Carmela Marangi, Athanasios S. Kallimanis, Maria Adamo, Cristina Tarantino, Maria Panitsa, Massimo Terzi, Giuseppe Veronico, Francesco Lovergine, Harini Nagendra, Richard Lucas, Paola Mairota, Caspar A. Mucher, Palma Blonda, 2013. [Translating land cover/land use classifications to habitat taxonomies for landscape monitoring: a Mediterranean assessment](#). *Landscape Ecology*. DOI [10.1007/s10980-013-9863-3](https://doi.org/10.1007/s10980-013-9863-3).
- [From Space to species: Solutions for biodiversity monitoring](#) by Palma Blonda, Richard Lucas and João Pradinho Honrado – on behalf of BIO_SOS consortium. Window on GMES, Special Issue on Discover what GMES can do for European regions and cities, ISSN 2030-5419.

BioVeL

Full name: Biodiversity Virtual e-Laboratory

Type: FP7 project

Duration: 2011 –2014

Website: <http://www.biovel.eu/>

BioVeL is a virtual e-laboratory that supports research on biodiversity issues using large amounts of data from cross-disciplinary sources. BioVeL offers the possibility to use computerised "workflows" (series of data analysis steps) to process data, be that from one's own research and/or from existing sources.

Relevant documentation and information for data standards

BioVel Workflows:

- The [Taxonomic Data Refinement Workflow](#) provides an environment for preparing observational and specimen data sets for use in scientific analyses such as: species distribution analysis, species

richness and diversity studies, species occurrence studies, historical analysis, and other spatio-temporal analyses.

- The ecological niche modeling (ENM) workflow uses occurrence and environmental data to model species distributions using the openModeller Web Service¹⁹⁵. openModeller is an ecological niche modelling library providing a uniform method to model species distribution patterns with a variety of algorithms, including GARP, Climate Space Model, Bioclimatic Envelopes, Support Vector Machines and others. It combines species occurrence data with environmental data sets in the form of georeferenced raster layers (such as temperature, precipitation, salinity) to generate potential distribution models.
- The ENM statistical workflow (ESW) allows the computation of the extent and intensity of change in species potential distribution through computation of the differences between two raster layers using the R statistical environment. The difference file is computed from two input files (in this case present projection and 2050 projection). The difference between each corresponding raster cell value is computed and stored in the difference file, regardless of the input files' geographical extent and origin.
- Other workflows include the Metagenomic Workflows; Phylogenetic Workflows; Population Modelling Workflows; Ecosystem Functioning and Valuation Workflows

EBONE

Full name: European Biodiversity Observation Network

Type: EU FP7 project

Duration: April 2008 until March 2012

Website: <http://www.wageningenur.nl/en/Expertise-Services/Research-Institutes/alterra/Projects/EBONE-2.htm>

The EBONE project focused on the development of a cost effective system of biodiversity data collection at regional, national and European levels. The project developed a coherent system for data collection that can be used for international comparable assessments. EBONE acted as a pilot for GEO BON developing these networks in Europe and sharing the experience with other initiatives around the world. Specifically, EBONE focused on: (1) the provision of a sound scientific basis for the production of statistical estimates of stock and change of key indicators that can then be interpreted by policy makers responding to EU Directives regarding threatened ecosystems and species; (2) The development of a system for estimating past change and for forecasting and testing policy options and designing mitigating management strategies for threatened ecosystems and species.

Relevant documents and information for data standards:

- T. W. Parr, R.H.G. Jongman, M. Kùlvik. (2010) D1.1: The Selection of Biodiversity indicators for EBONE Development Work. Version 2.12¹⁹⁶.
- The Habitat Monitoring system is based on General Habitat Categories (GHCs). The definitions of the General Habitat Categories are based on the practical experience of the GB Countryside Survey adapted for Europe on the basis of the validation workshops. Further information¹⁹⁷.

The instructions for habitat mapping and recording advised by EBONE¹⁹⁸:

- R.G.H. Bunce, M.M. B. Bogers, P.Roche, M.Walczak, I.R. Geijzendorffer and R.H.G. Jongman

¹⁹⁵ <http://openmodeller.sf.net/>

¹⁹⁶ http://www.wageningenur.nl/upload_mm/f/5/c/1826d18f-6a54-4624-a9cb-3ff80661ac95_EBONED11_Indicators.pdf

¹⁹⁷ <http://www.wageningenur.nl/en/Expertise-Services/Research-Institutes/alterra/Projects/EBONE-2/Products/General-Habitat-Categories.htm>

¹⁹⁸ <http://www.wageningenur.nl/en/Expertise-Services/Research-Institutes/alterra/Projects/EBONE-2/Products/Habitat-Mapping-and-Recording.htm>

(2011) Manual for Habitat and Vegetation Surveillance and Monitoring: Temperate, Mediterranean and Desert Biomes. First edition. Wageningen, Alterra, Alterra report 2154. 1 06 pp .; 15 fig.; 14 tab.; 35 ref.

EMODnet

Full name: European Marine Observation and Data Network

Type: Direct Tender

Duration: 2009-2012; 2013-2016 (EMODnet2)

Website: <http://bio.emodnet.eu/>

EMODnet is an initiative from the European Commission Directorate-General for Maritime Affairs and Fisheries (DG MARE) as part of its Marine Knowledge 2020 strategy. It was established to improve access to quality-assured, standardised and harmonised marine data through building a consortium of relevant organisations within Europe. Presently, there are six sub-portals in operation that provide access to marine data under the following themes: hydrography, geology, physical parameters, chemistry, biology, and physical habitats. One further portal covering human activities is currently under construction. EMODnet Biology provides access to the marine biological data portal and metadata catalogue. In its current phase (EMODnet2), building on a set of preparatory actions undertaken in phase 1 (2009-2012), a consortium of 21 government agencies and research institutes with national and international expertise in marine biological data monitoring and data management will deliver data, metadata and data products of surveys in the water column and on the sea bed from phytoplankton, zooplankton, angiosperms, macroalgae, benthos, birds, mammals, reptiles and fish occurring in European marine waters. The project will identify and focus on biological data types, species, species attributes, sampling methods and biological indicators to support the variety of legislations, and will create biological data products to support environmental legislations including the Marine Strategy Framework Directive.

EUDAT

Full name: European Data Infrastructure

Type: FP7 project

Duration: 2011 - 2014

Website: <http://www.eudat.eu/>

EUDAT will support a Collaborative Data Infrastructure allowing researchers to share data within and between communities and enable them to carry out their research effectively. EUDAT's mission is to: (1) Help fulfil the vision of a European Data e-infrastructure by providing a sustainable platform for technologies, tools and services driven by user needs; (2) Engage users (including individual researchers along with representatives from universities, research labs, and libraries) in defining and shaping a platform for shared services that makes it possible for data-intensive research to span all the scientific disciplines; (3) Produce the common low-level services that are required to provide the level of interoperability and trust of data that is necessary to support both widespread access to data, and the long-term preservation of data for use and re-use; and (4) Ensure that the data infrastructure is sufficiently robust to keep pace with the expected acceleration of the scale and complexity of scientific data being generated within the European Research Area and beyond.

The EUDAT consortium includes representatives from research communities: CLARIN Linguistics; EPOS Earth sciences; ENES Climate sciences; LIFEWATCH Environmental sciences; VPH Biological and medical sciences; INCF International Neuroinformatics Coordinating Facility.

Relevant documents/information for data standards

- A report about the work done in accordance with the project plan outlining achievements gained following a EUDAT technology appraisal: D.5.5. Technology appraisal report¹⁹⁹.
- EUDAT is preparing the next phase of its common data services and is organizing a series of workshops with a view to establishing Working Groups on the following topics:
- common services in the area of **Dynamic Data** addressing real-time data such as data produced by remotely connected field sensors and massive crowd sourcing data generated *via* mobile devices,
- common services in the area of **Workflow Support** supporting researchers to orchestrate data processing chains and execute them on computers close to where their data is located,
- common services in the area of **Semantics** allowing researchers, for example, to check the correctness of incoming data against trusted ontologies and semantically annotate the data, and
- policies regarding **Data Access and Re-use** at community and service provider levels.

EU-BON should follow closely developments of EUDAT and, perhaps, collaborate where possible in the organisation of workshops.

EUMON

Full name: EU-wide monitoring methods and systems of surveillance for species and habitats of Community interest

Type: FP6 project

Duration: November 2004 – April 2008

Website: <http://eumon.ckff.si/summary.php>

EuMon focused on four major aspects important for biodiversity monitoring: the involvement of volunteers, coverage and characteristics of monitoring schemes, monitoring methods, and the setting of monitoring and conservation priorities. It further developed tools to support biodiversity monitoring.

EuMon developed three internet based support tools: [BioMAT](#) - the EuMon integrated Biodiversity Monitoring and Assessment Tool, the PMN database that contains information on organisations that carry out volunteer based biodiversity monitoring, and the database on European biodiversity monitoring schemes (DaEuMon). There are 643 datasets or monitoring networks in [DaEuMon](#), but the actual number is about three-fold. This is a valuable resource that would need to be updated and maintained also in future. EuMon has compiled methods to develop an efficient network of protected areas and has analysed gaps and biases in the NATURA 2000 network. In addition, EuMon developed key principles for biodiversity monitoring and for determining national responsibilities.

Relevant documents and information for data standards:

Henry P.-Y., Lengyel S., Nowicki P., Julliard R., Clobert J., Čelik T., Gruber B., Schmeller D.S., Babij V. & Henle K. (2008) Integrating ongoing biodiversity monitoring: potential benefits and methods. *Biodiversity and Conservation*, 17(14), 3357-3382, (IF = 1.4)

Grosbois V., Gaillard J.M., Barbraud C., Lambrechts M., Clobert J., Moller A.P. & Lebreton J.D. (2006) Selecting the most relevant climate indices to identify and predict climate impacts on bird population. *Journal of Ornithology*, 147(), 25-25 (IF = 1.0)

Lengyel S., Kobler A., Kutnar L., Framstad E., Henry P.-Y., Babij V., Gruber B., Schmeller D.S. & Henle K. (2008) A review and a framework for the integration of biodiversity monitoring at the habitat level. *Biodiversity and Conservation*, 17(14), 3341-3356 (IF = 1.4)

Schmeller D.S. (2008) European species and habitat monitoring: where are we now? *Biodiversity and Conservation*, 17(14), 3321-3326

¹⁹⁹ http://www.eudat.eu/system/files/EUDAT-DEL-WP5-D5.1.1-Technology_Appraisal_Report.pdf

EuroGEOSS

Full name: EuroGEOSS, a European approach to GEOSS

Type: FP7 large scale integrated project

Duration: May 2009 – April 2012

Website: <http://www.eurogeoss.eu/default.aspx>

Focussing on three strategic areas (Biodiversity, Forestry, Drought), EuroGEOSS developed an initial operating capacity for a European Environment Earth Observation System. It then carried out research necessary for developing an advanced operating capacity that would support inter-disciplinary interoperability allowing multi-scale modelling from heterogeneous data sources with the models expressed as workflows of re-usable components. Within the Biodiversity area, EuroGEOSS contributed to the ongoing development of the Joint Research Centre Digital Observatory for Protected Areas (DOPA) (<http://dopa.jrc.ec.europa.eu/>).

The project builds an initial operating capacity for a European Environment Earth Observation System in the three strategic areas of Drought, Forestry and Biodiversity. It undertakes research to develop this into an advanced operating capacity providing access to data and analytical models from different disciplinary domains.

This concept of inter-disciplinary interoperability requires research in advanced modelling from multi-scale heterogeneous data sources, expressing models as workflows of geo-processing components reusable by other communities, and ability to use natural language to interface with the models.

The extension of INSPIRE and GEOSS components with concepts emerging in the Web 2.0 communities in respect to user interactions and resource discovery, also supports the wider engagement of the scientific community with GEOSS as a powerful means to improve the scientific understanding of the complex mechanisms driving the changes that affect our planet.

Relevant documents and information for data standards

Environmental model access and interoperability: The GEO Model Web initiative” on the “Environmental Modelling & Software”, 39: 214–228, January 2013

Dubois, G., J. Skøien, M. Schulz, L. Bastin, S. Peedell (2013). eHabitat, a multi-purpose Web Processing Service for ecological modeling. *Environmental Modelling & Software*, 41: 123-133.

Skøien, J.O., G. Dubois, J. De Jesus (2011). Forecasting biomes of protected areas. *Procedia Environmental Sciences* 7: 44–49

Skøien, J., M. Schulz, G. Dubois, I. Fisher, M. Balman, I. May, É. Ó Tuama (2012). Climate change in biomes of Important Bird Areas – results from a WPS application. *Ecological Informatics* (In Press)

Bastin, L., G. Buchanan, A. Beresford, J-F Pekel, G. Dubois (2012). Open-source mapping and services for Web-based land cover validation. *Ecological Informatics* (In Press)

EuroGEOSS broker: <http://www.eurogeoss.eu/broker/Pages/AbouttheEuroGEOSSBroker.aspx>

EuroGEOSS documents and publications: <http://www.eurogeoss.eu/Pages/Publications.aspx>
<http://www.eurogeoss.eu/Pages/Publications.aspx>

KNEU

Full name: KNEU - Developing a Knowledge Network for EUropean expertise on biodiversity and ecosystem services to inform policy making economic sectors

Type: FP7 project

Duration: 2010 – 2013

Website: <http://www.biodiversityknowledge.eu/>

BiodiversityKnowledge (KNEU) is an initiative by researchers and practitioners to help all societal actors in the field of biodiversity and ecosystem services to make better informed decisions. The goal is an innovation called Network of Knowledge - an open networking approach to boost the knowledge flow between biodiversity knowledge holders and users in Europe.

MS.MONINA

Full name: Multi-Scale Service for Monitoring Natura 2000 Habitats of European Community Interest

Type: FP7 project

Duration: 2011-2013

Website: <http://www.ms-monina.eu/home-1>

MS.MONINA supports European, national and local authorities in monitoring the state of European nature sites of community interest. The project supports the GEO (Group on Earth Observations) societal benefit area of biodiversity and demonstrates the power of earth observation-based methods for monitoring sensitive ecological sites in general. The MS.MONINA geoportal is the central entrance point to view and download MS.MONINA products developed for the service cases on EU, State and Site level. All accessible datasets follow the INSPIRE metadata standards. The MS.MONINA Tools Catalogue provides a collection of web accessible tools and methods for monitoring Natura 2000 sites.

Relevant documents and information for data standards

[Technical synthesis on the possibilities and limits of remote sensing for mapping natural habitats \(Del 3.2\)](#)

[Report on methodological tools \(Del 6.1\)](#)

Spanhove, T., Vanden Borre, J., Delalieux, S., Haest, B., Paelinckx, D. (2012) Can remote sensing estimate fine-scale quality indicators of natural habitats?, *Ecological Indicators*, Volume 18, Pages 403-412, ISSN 1470-160X, <http://dx.doi.org/10.1016/j.ecolind.2012.01.025>.

PESI

Full name: Pan-European Species directories Infrastructure

Type: FP7 project

Duration: 2008-2011

Website: <http://www.eu-nomen.eu/portal/>

PESI integrates all-taxon registers in Europe into a single, authoritative checklist for plant and animal species in Europe. In ViBRANT, PESI will couple its networking activities with Scratchpad users to facilitate the production of regional checklists and taxonomic catalogues. An improved interoperability infrastructure is being built.

pro-iBiosphere

Full name: pro-iBiosphere

Type: FP7 project

Duration: 2012-2014

Website: <http://www.pro-ibiosphere.eu/>

The aim of pro-iBiosphere is to prepare (= pro), through a coordination action, the ground for an integrative system (= sphere) for intelligent (= i) management of biodiversity (= bio) knowledge.

Once it becomes operational, the European Open Biodiversity Knowledge Management System will play a major role in facilitating the synthesis of core biodiversity data by creating an authoritative framework including, discovery of new species, naming of specimens and species, identification tools, descriptions, and various other basic types of information. It will also facilitate the acquisition of high quality biodiversity data from various sources, including legacy data; the curation of the data; and at the same time it will optimize the delivery of those data to the various users.

Relevant documentation and information for data standards

- D3.3.1 Semantic integration of the biodiversity literature:
- http://wiki.pro-ibiosphere.eu/wiki/D3.3.1_Semantic_integration_biodiversity_literature
- D2.1.1 Report on ongoing biodiversity related projects, current e-infrastructures and standards:

http://wiki.pro-ibiosphere.eu/w/media/c/c9/Pro-iBiosphere_WP2_PLAZI_D2.1.1_VFF_30062013.pdf

SeaDataNet

Full name: Pan-European infrastructure for marine data management

Type: FP6 & FP7 project

Duration: 2011 – 2015 (SeaDataNet2)

Website: <http://www.seadatanet.org/>

SeaDataNet is developing a standardized system for managing marine data by creating a virtual network of the national oceanographic data centres of 35 countries that are active in data collection. This Pan-European network will provide on-line integrated databases of standardized quality. The on-line access to in-situ data, metadata and products is provided through a unique portal interconnecting the interoperable node platforms constituted by the SeaDataNet data centres. The development and adoption of common communication standards and adapted technology ensure the platforms interoperability. The quality, compatibility and coherence of the data issuing from so many sources, is assured by the adoption of standardized methodologies for data checking, by dedicating part of the activities to training and preparation of synthesized regional and global statistical products from the most comprehensive in-situ data sets made available by the SeaDataNet partners. The network has adopted the use of common vocabularies as a prerequisite for data interoperability based on the NERC DataGrid (NDG) Vocabulary Server Web Service API. A manual of quality control procedures is available²⁰⁰ and a harmonised scheme of quality control flags for labeling individual data values has been defined and adopted²⁰¹.

SeaDataNet has recently (2012) provided a data specification extension²⁰² for handling marine biological data sets and enabling better interoperability with other data sharing networks such as INSPIRE, LifeWatch, EurOBIS, GBIF and EMODnet.

ViBRANT

Full name: Virtual Biodiversity Research and Access Network for Taxonomy

Type: FP7 project

Duration: December 2010 – 2013

Website: <http://vbrant.eu/>

Virtual Biodiversity Research and Access Network for Taxonomy (ViBRANT) will support the development of virtual research communities involved in biodiversity science. Their goal is to provide a more integrated and effective framework for those managing biodiversity data on the Web. Amongst

²⁰⁰ http://www.seadatanet.org/content/download/18414/119624/file/SeaDataNet_QC_procedures_V2_%28May_2010%29.pdf

²⁰¹ http://seadatanet.maris2.nl/v_bodc_vocab/welcome.aspx/

²⁰² http://www.SEADATANET.org/content/download/14628/96004/file/SDN2_WP8_Extension_biology_KDeneudt.pdf

other services, ViBRANT will provide a standards compliant technical architecture that can be sustained by the biodiversity research community. Software is being developed to ensure that all data entered or managed in ViBRANT are compatible with, and available to other research and publishing infrastructures. Specifically, ViBRANT will target some current biodiversity information platforms including Scratchpads, CyberPlatform, EoL, PESI, GBIF and Species-ID. To enable the exchange of data between Scratchpads, EDIT Common Data Model, Encyclopedia of Life (EOL) and the Global Biodiversity Information Facility (GBIF), ViBRANT will use the Darwin Core Archive (DwC-A) format developed by GBIF. Default data content types include organism names, species information, factual data, distribution, media and literature.

Relevant documents and information for data standards

- Penev *et al.* (2011).
- M6.10 - Use cases of existing standards of XML mark up tagging and semantic enhancement collected and review²⁰³.
- M4.27 - Audubon Core standard on Mediawiki²⁰⁴.

²⁰³ [http://vbrant.eu/sites/vbrant.eu/files/Milestone 6 10-Review of mark up and tagging tools.pdf](http://vbrant.eu/sites/vbrant.eu/files/Milestone%206%2010-Review%20of%20mark%20up%20and%20tagging%20tools.pdf)

²⁰⁴ http://terms.tdwg.org/wiki/Audubon_Core

Annex II: Key standards for EU BON

Following is a list of the main standards for metadata, data exchange and transfer protocols.

Name **Access to Biological Collection Data (ABCD)**

Description ABCD - Access to Biological Collections Data - Schema is a common data specification for biological collection units, including living and preserved specimens, along with field observations that did not produce voucher specimens. It is intended to support the exchange and integration of detailed primary collection and observation data.

URL <http://wiki.tdwg.org/ABCD/>

Name **ABCDDNA – DNA extension for ABCD**

Description The DNA extension for ABCD, called ABCDDNA, is a product of the DNA Bank Network. ABCDDNA is a theme specific, XML Schema, created to facilitate storage and exchange of data related to DNA collection units, such as DNA extraction specifics, DNA quality parameters and data characterising products of downstream applications, along with the relation to the analysed voucher specimen. ABCDDNA offers only a rudimentary set of DNA-specific data sequences. Currently only BioCASE and ABCD can be used to provide data via the DNA Bank Network but work is underway to develop a Darwin Core Archive extension.

URL <http://wiki.bgbm.org/dnabankwiki/index.php/ABCDDNA>

Name **BioCASE Protocol**

Description The BioCASE protocol is used for communication between the Provider Software (database wrapper) and a client application on the BioCASE network, a transnational network of biological collections of all kinds. It is based on the DiGIR protocol.

URL <http://www.biocase.org/products/protocols/index.shtml>

Name **Biological Collections Ontology**

Description “The biological collection ontology includes consideration of the distinctions between individuals, organisms, voucher specimens, lots, and samples the relations between these entities, and processes governing the creation and use of "samples". Within scope as well are properties including collector, location, time, storage environment, containers, institution, and collection identifiers.”

URL <http://bioportal.bioontology.org/ontologies/BCO>

Name **Business Process Execution Language (BPEL)**

Description BPEL is a widely accepted (OASIS) standard for handling workflows. In the context of OGC, workflows are produced through "service chaining", which can be performed in a number of ways, one of which is orchestration of a service chain including one or more Web Processing Services (WPS) using a BPEL engine. Modern scientific applications, such as scientific workflows, increasingly rely on services, such as job submission, data transfer or data portal services and messages. Such services are referred to as Grid services.

URL <http://bpel.xml.org/>

Name **Content Standard for Digital Geospatial Metadata (CSDGM), (FGDC-STD-001-1998)**

Description CSDGM is a US Federal Metadata standard that "provides a common set of terminology and definitions for the documentation of digital geospatial data."

The standard establishes the names of data elements and compound elements to be used to determine the fitness , the means of accessing and transfer of geospatial data.

URL http://www.fgdc.gov/standards/projects/FGDC-standards-projects/metadata/base-metadata/index_html

Name **CSDGM - Biological Data Profile**

Description The standard "broadens the application of the CSDGM so that it is more easily applied to data that are not explicitly geographic (laboratory results, field notes, specimen collections, research reports) but can be associated with a geographic location. The profile changes the conditionality and domains of CSDGM elements, requires the use of a specified taxonomical vocabulary, and adds elements."

URL <http://www.fgdc.gov/standards/projects/FGDC-standards-projects/metadata/biometadata/biodatap.pdf>

Name **CSDGM - Metadata Profile for Shoreline Data**

Description Metadata Profile of CSDGM for Shoreline Data “addresses variability in the definition and mapping of shorelines by providing a standardised set of terms and data elements required to support metadata for shoreline and coastal data sets. The profile also includes a glossary and bibliography.”

URL <http://www.csc.noaa.gov/metadata/sprofile.pdf>

Name **Darwin Core**

Description The Darwin Core is body of standards. It includes a glossary of terms (in other contexts these might be called properties, elements, fields, columns, attributes, or concepts) intended to facilitate the sharing of information about biological diversity by providing reference definitions, examples, and commentaries

URL <http://rs.tdwg.org/dwc/>

Name **Digital Object Identifier (DOI)**

Description A Digital Object Identifier (DOI), i.e., a "digital identifier of an object" is a globally unique persistent identifier that can be resolved within the DOI system to obtain information about the object including descriptive metadata. The DOI system enables the construction of automated services and transactions.

URL http://www.doi.org/doi_handbook/1_Introduction.html

Name **Ecological Metadata Language**

Description Ecological Metadata Language (EML) is a metadata specification for the ecology discipline based on prior work done by the Ecological Society of America and associated efforts. EML is implemented as a series of XML document types that can be used in a modular and extensible manner to document ecological data. Each EML module is designed to describe one logical part of the total metadata that should be included with any ecological data set.

URL <http://knb.ecoinformatics.org/software/eml/>

Name **Distributed Generic Information Retrieval (DiGIR)**

Description DiGIR is a query and transfer protocol for distributed data sources based on HTTP, XML and UDDI.

URL <http://digir.sourceforge.net/>

Name Environment Ontology (EnvO)

Description An ontology of environmental features and habitats.

URL <http://bioportal.bioontology.org/ontologies/ENVO>

Name **Genomic Contextual Data Markup Language (GCDML)**

Description GCDML is an XML schema implementing the “minimum information about a genome” (MIGS) and “minimum information about a metagenome sequence” (MIMS) specifications.

URL <http://gensc.org/projects/gcdml/>

Name **Geography Markup Language (GML)**

Description An OGC standard, GML is an XML grammar and modeling language for geographic systems as well as an open interchange format for geographic transactions on the Internet. It covers not only conventional "vector" or discrete objects, but also coverages and sensor data. The ability to integrate all forms of geographic information is key to the utility of GML.

Note: KML made popular by Google, complements GML. Whereas GML is a language to encode geographic content for any application (by describing objects and their properties), KML is a language for the visualization of geographic information tailored for Google Earth.

URL <http://www.ogcnetwork.net/gml>

Name **Life Sciences Identifier (LSID)**

Description LSIDs are a type of persistent, globally unique identifier for Life Sciences entities. The specification covers a standardized naming schema, a service assigning unique identifiers complying with such naming schema, and a resolution service that specifies how to retrieve information associated with such identifiers.

URL <http://www.omg.org/cgi-bin/doc?dtd/04-05-01>

Name **Minimum Information about an Environmental Sequence (MIENS)**

Description MIENS is an extension to the minimum information about a genome/meta-genome sequence (MIGS/MIMS) specification of the Genomics Standard Consortium is a proposal for documenting the environmental parameters in the extraction

URL http://gensc.org/gc_wiki/index.php/MIGS/MIMS/MIENS

Name **Natural Collections Descriptions (NCD)**

Description NCD is a standard for facilitating the exchange of information on all kinds of collections of natural history material including specimens, original artwork, photographs, archives, published material.

URL <http://www.tdwg.org/activities/ncd/>

Name **Open Archives Initiative Object Reuse and Exchange (OAI-ORE)**

Description Open Archives Initiative Object Reuse and Exchange defines standards for the description and exchange of aggregations of Web resources (also referred to as compound digital objects). These include a collection of terms (vocabulary) for describing the objects and their inter-relationships.

URL <http://www.openarchives.org/ore/>

Name **Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH)**

Description OAI-PMH provides a “low-barrier” mechanism for interoperability across distributed

metadata repositories. Data providers expose metadata and service providers, in turn, consume the metadata through a client application known as a harvester that issues OAI-PMH service requests over HTTP.

URL <http://www.openarchives.org/pmh/>

Name **OpenGIS Catalogue Services for the Web (CSW)**

Description OGC CSW specification defines "the interfaces, bindings, and a framework for defining application profiles required to publish and access digital catalogues of metadata for geospatial data, services, and related resource information".

URL <http://www.opengeospatial.org/standards/cat>

Name **OGC and ISO Observations and Measurements (O&M)**

Description This international standard "defines a conceptual schema for observations, and for features involved in sampling when making observations. These provide models for the exchange of information describing observation acts and their results, both within and between different scientific and technical communities."

URL <http://www.opengeospatial.org/standards/om>

Name **OGC Web Coverage Service (WCS)**

Description OGC WCS "supports electronic retrieval of geospatial data as "coverages" – that is, digital geospatial information representing space-varying phenomena."

URL <http://www.opengeospatial.org/standards/wcs>

Name **OpenGIS Web Map Service (WMS)**

Description OGC WMS Implementation Specification provides three operations (GetCapabilities, GetMap, and GetFeatureInfo) in support of the creation and display of registered and superimposed map-like views of information that come simultaneously from multiple remote and heterogeneous sources.

URL <http://www.opengeospatial.org/standards/wms>

Name **OpenGIS Web Feature Service (WFS)**

Description OGC WFS Implementation Specification allows a client to retrieve and update geospatial data encoded in Geography Markup Language (GML) from multiple Web Feature Services. The specification defines interfaces for data access and manipulation operations on geographic features, using HTTP as the distributed computing platform.

URL <http://www.opengeospatial.org/standards/wfs>

Name **ISO 19101:2004**

Description ISP 19101 “defines the framework for standardization in the field of geographic information and sets forth the basic principles by which this standardization takes place”.

URL http://www.iso.org/iso/home/store/catalogue_tc/catalogue_detail.htm?csnumber=26002

Name **ISO 19111:2003**

Description ISO standard for geographic information- spatial referencing by coordinates. Establishes a common requirement for describing coordinate reference systems (CRSs) including the datum giving the relation to the Earth and the coordinate system used.

URL http://www.iso.org/iso/iso_catalogue/catalogue_ics/catalogue_detail_ics.htm?csnumber=26016

Name **ISO/DIS 19156**

Description This standard specifies an XML implementation for the OGC and ISO Observations and Measurements (O&M) conceptual model, including a schema for Sampling Features. This encoding is an essential dependency for the OGC Sensor Observation Service (SOS) Interface Standard. More specifically, this standard defines XML schemas for observations and for features involved in sampling when making observations. These provide document models for the exchange of information describing observation acts and their results, both within and between different scientific and technical communities.

URL <http://www.geoinformatics.com/tags/isodis-19156>

Name **ISO 19115**

Description ISO 19115:2003 defines the schema required for describing geographic information and services. It provides information about the identification, the extent, the quality, the spatial and temporal schema, spatial reference, and distribution of digital geographic data. It defines mandatory and conditional metadata sections, metadata entities, and metadata elements and is applicable to the cataloguing of data sets, clearinghouse activities, geographic data sets and more. Countries that are members of ISO are required to provide metadata in a profile of ISO 19115. The INSPIRE initiative in the European Union is recommending use of ISO 19115.

URL www.iso.org/iso/home/store/catalogue_tc/catalogue_detail.htm?csnumber=26020

Name **ISO/TS 19139:2007**

Description “ISO/TS 19139:2007 defines Geographic MetaData XML (GMD) encoding, an XML Schema implementation derived from ISO 19115.”

"ISO/TS 19139 is applicable to provide a common XML specification for describing, validating and exchanging geographic metadata. It is intended to promote interoperability, and exploit ISO 19115's advantages in a concrete implementation specification. "

URL http://www.iso.org/iso/catalogue_detail.htm?csnumber=32557

Name **Journal Archiving Tag Suite (JATS)**

Description JATS is “an application of NISO Z39.96-2012, which defines a set of XML elements and attributes for tagging journal articles and describes three article models”.

URL <http://jats.nlm.nih.gov/>

Name **Javascript Object Notation (JSON)**

Description JSON, based on a subset of the JavaScript Programming Language, is a lightweight data-interchange format. It is easy for humans to read and write and it is easy for machines to parse and generate. JSON is a text format that is completely language independent but uses conventions that are familiar to programmers of the C-family of languages, including C, C++, C#, Java, JavaScript, Perl, Python, and many others. These properties make JSON an ideal data-interchange language.

URL <http://www.json.org/>

Name JSON with Padding (JSON-P)

Description JSONP or "JSON with padding" is a communication technique used in JavaScript programs running in web browsers to request data from a server in a different server or domain, something prohibited by typical web browsers because of an important security concept called "same origin policy" (SOP). Within the context of data-exchanges and interoperability issues, JSON-P technique supports cross database/domains queries that uses content from more than one source ("mashup" in a single web-page). SOP policy permits scripts running on pages originating from the same site to access and interact with each other's objects in HTML, XHTML and XML documents with no specific restrictions, but prevents access to these objects on different sites. Same-origin policy also applies to XMLHttpRequest.

URL <http://www.json-p.org/>

Name Minimum information about a genome sequence (MIGS)

Description MIGS is a Genomics Standard Consortium (GSC) standard that defines a core set of properties (referred to as a checklist) for genome sequences.

URL <http://wiki.genesc.org/index.php?title=MIGS/MIMS>

Name Minimum information about a marker gene sequence (MIMARKS)

Description MIMARKS is a Genomics Standard Consortium (GSC) standard that defines a core set of properties (referred to as a checklist) for marker gene sequences.

URL <http://wiki.genesc.org/index.php?title=MIMARKS>

Name Minimum information about a metagenome sequence (MIMS)

Description MIMS is a Genomics Standard Consortium (GSC) standard that defines a core set of properties (referred to as a checklist) for metagenome sequences.

URL <http://wiki.genesc.org/index.php?title=MIGS/MIMS>

Name Minimum information about any (X) sequence (MIxS)

Description MIxS is a Genomics Standard Consortium (GSC) standard that defines the set of properties for describing any sequence data. It does this by unifying its other standards (MIGS, MIMS, MIMARKS).

URL <http://wiki.gensc.org/index.php?title=MIxS>

Name **Network Common Data Format (NetCDF)**

Description “NetCDF is a set of software libraries and self-describing, machine-independent data formats that support the creation, access, and sharing of array-oriented scientific data”.

URL <http://www.unidata.ucar.edu/software/netcdf/>

Name **OWL**

Description The W3C Web Ontology Language (OWL) is an ontology language for the Semantic Web providing classes, properties, individuals and data values. OWL facilitates greater machine interpretability of Web content than that supported by XML, RDF, and RDF Schema by providing additional vocabulary along with a formal semantics. OWL has three increasingly-expressive sublanguages: OWL Lite, OWL DL, and OWL Full.

URL <http://www.w3.org/TR/owl2-overview/>

Name **POSIX**

Description POSIX “defines a standard operating system interface and environment, including a command interpreter (or “shell”), and common utility programs to support applications portability at the source code level”. It was jointly developed by the IEEE and The Open Group.

URL <http://pubs.opengroup.org/onlinepubs/9699919799/>

Name **Resource Description Framework (RDF)**

Description RDF is a model for data exchange on the Web based on a directed, labelled graph where the edges represent the named link between two (named) resources, represented by the graph nodes.

URL <http://www.w3.org/RDF/>

Name **Representational State Transfer (REST)**

Description REST is a particular architecture style for the design of networked applications. In contrast to more complex mechanisms such as CORBA, RPC or SOAP for connecting between machines, simple HTTP is used. For example, with a network of Web pages

(a virtual state-machine), navigation through an application is by selecting links (state transitions), which results in the next page (representing the next state of the application) being transferred to the user and displayed. The World Wide Web itself, based on HTTP, can be considered as a REST-based architecture. RESTful applications use HTTP requests for all four CRUD (Create/Read/Update/Delete) operations used for persistent storage.

URL http://en.wikipedia.org/wiki/Representational_state_transfer

Name **Structured Descriptive Data (SDD)**

Description SDD is an XML-based TDWG standard for capturing and managing descriptive data about organisms.

URL <http://www.tdwg.org/standards/116/>

Name **SOAP**

Description SOAP is a protocol specification for exchanging structured information between programs running in the same or another kind of an operating system (such as Windows or Linux) by using the HTTP protocol and Extensible Markup Language (XML) as the mechanisms for information exchange. Web protocols are installed and available for use by all major operating system platforms and SOAP specifies how to encode and respond to an HTTP header and an XML file. An advantage of SOAP is that program calls through HTTP requests are usually allowed through firewall servers that screen out requests, thus allowing programs using SOAP to communicate with programs anywhere.

URL <http://www.w3.org/TR/soap/>

Name **SPARQL Protocol and RDF Query Language (SPARQL)**

Description SPARQL is a query language for databases, designed to retrieve and manipulate data stored in Resource Description Framework (RDF) format. It is an official W3C Recommendation and is recognized as one of the key technologies of the Semantic Web. Implementations for multiple programming languages already exist. The SPARQL Protocol consists of two HTTP operations: a "query operation" for performing SPARQL Query Language queries and an "update operation" for performing SPARQL Update Language requests. SPARQL Protocol clients send HTTP requests to SPARQL Protocol services that handle the request and send HTTP responses back to the originating client.

URL http://www.w3.org/2009/sparql/wiki/Main_Page

Name	TDWG Access Protocol for Information Retrieval (TAPIR)
Description	Designed as a generic tool that can be applied to domains other than biodiversity and natural science collections data, TAPIR is a specification for accessing structured data on distributed databases using HTTP for transport and XML for encoding messages and data. It combines and extends the features of DiGIR and BioCAsE protocols to create a new and more generic means of communication between client applications and data providers using the Internet.
URL	http://www.tdwg.org/activities/tapir/
Name	Taxon Concept transfer Schema (TCS)
Description	TCS provides a standard for taxon names and taxon concepts in the exchange and integration of biodiversity and natural history data.” The majority of elements in TCS are optional to allow for the variety of different approaches to defining and recording taxonomic names and concepts, hence TCS allows more choices if an expert simultaneously authors concepts AND asserts concept relationships.
URL	http://www.tdwg.org/activities/tnc/
Name	VegCSV
Description	VegCSV is an extension of Darwin Core for plots data.
URL	https://projects.nceas.ucsb.edu/nceas/projects/bien/wiki/VegCSV
Name	Veg-X
Description	Veg-X is an exchange schema for vegetation plot data under development through the TDWG Vegetation Observations Data Exchange Task Group.
URL	http://wiki.tdwg.org/Vegetation/
Name	Web Service Description Language (WSDL)
Description	WSDL is a proposal submitted to the W3C for an XML format “for describing network services as a set of endpoints operating on messages containing either document-oriented or procedure-oriented information. The operations and messages are described abstractly, and then bound to a concrete network protocol and message format to define an endpoint.” Different network messaging protocols can thus be used, e.g., SOAP, HTTP.
URL	http://www.w3.org/TR/wsdl

Annex III: Acronyms

ABCD Access to Biological Collection Data
ABCDEFG Access to Biological Collection Data Extended for Geosciences
ABCDDNA Access to Biological Collection Data DNA
ALTER-Net A Long-Term Biodiversity Research Network
AMNH American Museum of Natural History
API Application Programming Interface
ASFA Aquatic Sciences and Fisheries Abstracts
BDC (European) Biodiversity Data Centre
BGBM Botanischer Garten und Botanisches Museum Berlin-Dahlem
BIEN Botanical Information and Ecology Network
BioCASE Biological Collection Access Service for Europe
BioMAT EuMon integrated Biodiversity Monitoring and Assessment Tool
BISE Biodiversity Information System for Europe
BOLD Barcode of Life Data Systems
BPEL Business Process Execution Language
BRIs BARCODE Records in INSDC
BiSciCol Biological Science Collections
BOL Barcode of Life
CBOL Consortium for the Barcode of Life
CETAF Consortium of European Taxonomic Facilities
CoL Catalogue of Life
CORBA Common Object Request Broker Architecture
COST (European) Cooperation in Science and Technology
CPR Continuous Plankton Recorder
CSDGM Content Standard for Digital Geospatial Metadata
CSV Comma separated value
DaEuMon Database on European biodiversity monitoring schemes
DDBJ DNA Data Bank of Japan
DEIMS Drupal Ecological Information System
DG Directorate-General (of the European Commission)
DG MARE (European) Directorate-General for Maritime Affairs and Fisheries
DIF Directory Interchange Format
DOI Digital Object Identifier
DOPA Digital Observatory for Protected Areas
DoW Description of Work
DTD Document Type Definition
DwC Darwin Core
DwC-A Darwin Core Archive
EBV Essential Biodiversity Variables
ECBOL European Consortium for the Barcode of Life
EDCNRP Environmental Data centers on Natural Resources and Products
EDIT European Distributed Institute of Taxonomy
EDMED European Directory of Marine Environmental Data
EEA European Environment Agency

EFDAC European Forest Data Centre
Eionet European environment information and observation network
EMBL European Molecular Biology Laboratory
EML Ecological Metadata Language
EMODNET European Marine Observation and Data Network
ENA European Nucleotide Archive
ENBI European Network for Biodiversity Information
ENES European Network for Earth System Modeling
ENM Ecological Niche Model
ESW ENM statistical workflow
EnvO Environment Ontology
EOL Encyclopedia of Life
EPOS European Plate Observing System
ERMS European Register of Marine Species
ESDAC European Soil Data Centre
ESF European Science Foundation
ESFRI European Strategy Forum for Research Infrastructures
ESSD Earth System Science Data
EC European Commission
EUDAT European Data Infrastructure
EUNIS European Nature Information System
EUR-OCEANS Ocean Ecosystems Analysis
FGDC Federal Geographic Data Committee
GBIF Global Biodiversity Information Facility
GBIO Global Biodiversity Informatics Outlook
GBWG Genomic Biodiversity Working Group
GCDML Genomic Contextual Data Markup Language
GCI GEOSS Common Infrastructure
GCMD Global Change Master Directory
GEO BON Group on Earth Observations Biodiversity Observation Network
GEO Gene Expression Omnibus
GEOSS Global Earth Observation System of Systems
GGBN Global Genome Biodiversity Network
GMES Global Monitoring for Environment and Security
GML Geography Markup Language
GNA Global Names Architecture
GNI Global Names Index
GNRD Global Names Recognition and Discovery
GOLD Genomes Online Database
GSC Genomic Standards Consortium
GUID Globally Unique Identifier
ICT Information and Communication Technology
ICES International Council for the Exploration of the Sea
IES Institute for Environment and Sustainability
IFREMER Institut français de recherche pour l'exploitation de la mer
ILTER International Long Term Ecological Research network
IMIS Integrated Marine Information System
INCF International Neuroinformatics Coordinating Facility

INSDC International Nucleotide Sequence Database Collaboration
IODE International Oceanographic Data and Information Exchange
IPT Integrated Publishing Toolkit
ISO International Organisation for Standardization
ITIS Integrated Taxonomic Information System
IUCN International Union for the Conservation of Nature
JATS Journal Archiving Tag Suite of the US National Library of Medicine
JRC Joint Research Centre
JSON JavaScript Object Notation
JSON-P JSON-with-padding
KNB Knowledge Network for Biocomplexity
KNEU - Knowledge Network for European expertise
LSID Life Science Identifier
ILTER Long Term Ecological Research network
MarBEF Marine Biodiversity and Ecosystem Functioning
MfN Berlin Museum für Naturkunde, Berlin
MGE Marine Genomics Europe
MIENS Minimum Information about an Environmental Sequence
MIGS Minimum information about a genome sequence
MIMARKS Minimum information about a marker gene sequence
MIMS Minimum information about a metagenome sequence
MIxS Minimum Information about any (x) Sequence
MODS Metadata Object Description Schema
NCBI National Center for Biotechnology Information
NCD Natural Collections descriptions
NERC Natural Environment Research Council
NDG NERC DataGrid
NetCDF Network Common Data Format
NSF National Science Foundation
O&M Observations and Measurements
OAI-ORE Open Archives Initiative Object Reuse and Exchange
OAI-PMH Open Archives Initiative Protocol for Metadata Harvesting
OASIS Organisation for the Advancement of Structured Information Standards
OBIS Ocean Biogeographic Information System
OBIS-SeaMap Ocean Biogeographic Information System Spatial Ecological Analysis of Megavertebrate Populations
OCR Optical Character Recognition
OGC Open Geospatial Consortium
OGC CSW Open Geospatial Consortium Catalogue Services for the Web
OGC WxS Open Geospatial Consortium web services
OpenGIS WCS Open Geospatial Consortium Web Coverage Service
OpenGIS WFS Open Geospatial Consortium Web Feature Service
OWL Web Ontology Language
Paris-MNHN National Museum of Natural History, Paris
PESI Pan-European Species directories Infrastructure
POSIX Portable Operating System Interface
PPBio Program for Planned Biodiversity Research
QIIME Quantitative Insights Into Microbial Ecology

RBGE Royal Botanical Garden Edinburgh
RDF Resource Description Framework
REST Representational State Transfer
RTF Rich Text Format
SDD Structured Descriptive Data
SDI Spatial Data Infrastructure
SEBI Streamlining European Biodiversity Indicators
SISMER Systèmes d'Informations Scientifiques pour la Mer
SOA Service Oriented Architecture
SOAP Simple Object Access Protocol
SOS Sensor Observation Service
SPARQL Query Language for RDF
TAPIR TDWG Access Protocol for Information Retrieval
TCS Taxon Concept transfer Schema
TDWG Taxonomic Databases Working Group
URI Uniform Resource Identifier
Veg-X Vegetation plot exchange schema
VHR Very high spatial resolution
VoMaG TDWG Vocabulary Management Task Group
VPH Virtual Physiological Human
WoRMS World Register of Marine Species
WSDL Web Services Description Language
XML Extensible Markup Language