Chapter 11
Global Infrastructures for Biodiversity Data and Services

Wim Hugo, Donald Hobern, Urmas Kõljalg, Éamonn Ó Tuama and Hannu Saarenmaa

Abstract GEO BON regards development of a global infrastructure in support of Essential Biodiversity Variables (EBVs) as one of its main objectives. To realise the goal, an understanding of the context within which such an infrastructure needs to operate is important (for instance, it is part of a larger drive towards research data infrastructures in support of open science?) and the information technology applicable to such infrastructures needs to be considered. The EBVs are likely to require very specific implementation guidelines once the community has defined them in detail. In the interim it is possible to anticipate the likely architecture for a GEO BON infrastructure, and to provide guidance to individual researchers, institutions, and regional or global initiatives in respect of best practice. The best practice guidelines cover general aspects applicable to all research infrastructures, the use of persistent identifiers, interoperability guidelines in respect of vocabularies, data services and meta-data management, and advice on the use of global infrastructure services and/or federated, standards-based implementations.

Keywords Interoperability · Research · Infrastructure · Architecture · Best practice · Guideline · Persistent identifier · Biodiversity · Informatics

W. Hugo (✉)
South African Environmental Observation Network, P.O. Box 2600,
Pretoria 0001, South Africa
e-mail: wim@saeon.ac.za

D. Hobern · É.Ó. Tuama
Global Biodiversity Information Facility, Universitetsparken 15,
2100 Copenhagen, Denmark
e-mail: dhobern@gbif.org

U. Kõljalg
Institute of Ecology and Earth Sciences, University of Tartu, Ülikooli 18,
50090 Tartu, Estonia
e-mail: urmas.koljalg@ut.ee

H. Saarenmaa
Digitarium/University of Eastern Finland, P.O. Box 111, 80101 Joensuu, Finland
e-mail: hannu.saarenmaa@helsinki.fi

M. Walters and R.J. Scholes (eds.), *The GEO Handbook on Biodiversity Observation Networks*, DOI 10.1007/978-3-319-27288-7_11

11.1 An Emerging Culture of Data Sharing, Publication and Citation

It has been widely accepted that the future usability and availability of research outputs, and specifically data, will be enhanced by proper description of these outputs using standardised metadata schemes, supplemented by deposit of the data in trusted repositories. Despite this, such outputs continue to be poorly described in practice. In addition, it is also commonly reported that the data supporting scholarly publication quickly becomes inaccessible or lost (Vines et al. 2014; Goddard et al. 2011). This disparity between what is seen as desirable behaviour, and reality is about to change, due to three significant drivers:

- Data publication and citation is gaining momentum (Chavan and Penev 2011). For a comprehensive review, see the report by a CoDATA[1] Task Group (Socha 2013).
- Funders are increasingly demanding the preservation of and continued open access to tax-funded research outputs.[2,3,4]
- Controversy in respect of reproducibility of scientific claims[5] have led to insistence by journals[6] that the data underpinning articles should be made available.

We believe these drivers will rapidly increase the availability of well-described, well-preserved, and sometimes standardised data services in the future.

11.1.1 Research Infrastructures

The drive towards data publication and citation requires support, hence the growth and proliferation of Research Data Infrastructures. These are supplemented strongly by voluntary, community-driven initiatives, and by member-funded bodies that support standardisation and interoperability.

Infrastructure operates on several levels: it provides governance and collaboration infrastructure (for example, the Belmont Forum[7] and Future Earth[8]),

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

[2]Berlin Declaration: http://www.berlin9.org/about/declaration/.

[3]OECD: http://www.oecd.org/sti/sci-tech/oecdprinciplesandguidelinesforaccesstoresearchdatafrompublicfunding.htm.

[4]USA: http://www.whitehouse.gov/sites/default/files/microsites/ostp/ostp_public_access_memo_2013.pdf.

[5]http://www.economist.com/news/briefing/21588057-scientists-think-science-self-correcting-alarming-degree-it-not-trouble.

[6]PLOS: http://blogs.plos.org/everyone/2014/02/24/plos-new-data-policy-public-access-data-2/.

[7]Belmont Forum: http://igfagcr.org/index.php/about-us.

[8]Future Earth: http://www.icsu.org/future-earth/media-centre/relevant_publications/future-earth-initial-design-report.

architecture and standards infrastructure (e.g., Research Data Alliance[9]—RDA, TDWG,[10] OGC,[11] GEO[12]), and physical, centralised or federated infrastructure (GBIF,[13] EUDAT,[14] and GEOSS[15]). Some global and regional initiatives span all of these (for example, the ICSU World Data System,[16] and GEO itself), and some are focused more narrowly on regional or domain-specific infrastructures (for example, DataOne,[17] EU BON,[18] Lifewatch,[19] and others).

It is worth noting that one of the motivations for the Research Data Alliance is to provide a cross-disciplinary, global exchange to minimise duplication of effort and divergence. Hence the landscape is at once characterised by divergent initiatives resulting from the nature of competitive grant funding and efforts to converge the impacts of funding these efforts. This is necessary, since divergence results in multiplicity of approaches, standards, protocols, and vocabularies—not supportive of interoperability.

11.1.2 Persistent Identifiers and Linked Open Data

Establishment of access to research outputs, either directly or via standardised services, requires a critical element: the ability to reliably find such objects in the web. This implies a *persistent identifier*, and several mechanisms are available to achieve this.

The biodiversity informatics community requires an identifier architecture that is capable of resolving two overlapping requirements—that of permanently identifying resources (data, services, and other web-based resources), and that of permanently identifying concepts (taxons, biomes, etc.).

There are several services available for either hosting or providing a minting framework for persistent identifiers (PIDs). Services that are general in nature, and allow hosting of PIDs on behalf of anyone, include the foundational Handle System.[20] This service can be used directly, but is also packaged and mediated, for

[9]RDA: https://rd-alliance.org/about.html.
[10]TDWG: http://www.tdwg.org/about-tdwg/.
[11]OGC: http://www.opengeospatial.org/.
[12]GEO: https://www.earthobservations.org/index.shtml.
[13]Global Biodiversity Information Facility: http://www.gbif.org/.
[14]EUDAT: http://www.eudat.eu/.
[15]GEOSS: https://www.earthobservations.org/geoss.shtml.
[16]http://www.icsu-wds.org.
[17]DataONE: http://www.dataone.org/.
[18]EU BON: http://eubon.eu/.
[19]LifeWatch: http://www.lifewatch.eu/.
[20]Handle System: http://www.handle.net/factsheet.html.

example by the members of the International Digital Object Identifier (DOI) Consortium[21]—allowing value-added services. DOI-based services that are important to our community include DataCite (linking published data sets and meta-data through DOIs to journal articles for purposes of citation tracking) and, CrossRef (more focused on linking DOI-based references across different journals), and GBIF (allocating DOIs for all published datasets and for search results). Several other biodiversity-focused initiatives exist, and these are discussed in the section on 'Specific Implementation Guidelines' (Barcode of Life,[22] Life Sciences Identifier, and similar, with identifiers.org[23] providing an aggregation of such services).

The availability of persistent identifiers assists the construction of Linked Open Data[24] (LOD) networks—making a significant contribution to the Semantic Web.[25]

11.1.3 Free and Open Data: Licensing and Policy

Delivering interoperable, open access to data and services involves (1) the implementation of applicable policies and (2) appropriate supporting licenses.

There are likely to be as many policies as there are data custodians and providers, but this is not really an issue as long as there is general compliance with the principles of free and open access—as documented by various global programmes such as the ICSU World Data System,[26] GEO,[27] and others.

Licenses, however, do need to be standardised, since machine-readability is a prerequisite for automated processing of data and services in the web. The most widely adopted candidates for this are the Creative Commons[28] family of licenses. These have been tested in multiple jurisdictions. Note that issues still under discussion include:

- 'Legal Interoperability' (how different licenses combine in automated processes, and what the resulting license is) (Uhlir 2013),
- Conditions or exceptions to be added to licenses to address legitimate concerns in respect of privacy, ethics, publication embargoes, endangered species, and similar.

[21]Digital Object Identifier: http://www.doi.org/doi_handbook/1_Introduction.html.

[22]http://www.barcodeoflife.org/.

[23]http://identifiers.org/.

[24]Linked Open Data: http://linkeddata.org/.

[25]https://www.w3.org/standards/semanticweb/.

[26]ICSU-WDS Data Policy: http://icsu-wds.org/services/data-policy.

[27]GEO Data Sharing Principles: https://www.earthobservations.org/geoss_dsp.shtml.

[28]Creative Commons and Data: http://wiki.creativecommons.org/Data.

GEO BON, being part of GEO, will adopt the GEOSS Data Sharing Principles (currently under review and likely to be modified slightly). In short, these are:

- 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.

11.1.4 Data Citation and Publication

Many of the institutional, technical, and legal hurdles that impeded the growth of data citation and publication have been addressed, and there is a broad consensus amongst journal publishers, data centres, and scientists in general on implementation (Socha 2013). CoDATA[29] and RDA[30] have played (and continue to play) a significant enabling role in this process.

Scientists should note that future research would be subject to:

- Planning for deposit and description (through metadata) of research output in a Trusted Digital Repository[31]—increasingly required by funders;
- Allocating persistent identifiers to such outputs, as appropriate.

Global coordinated research programmes, such as Future Earth, also attempt to align their funded outputs with the requirements of free and open access, and to promote a culture supportive of data publication and citation.

11.1.5 Big Data, Citizen Science, Crowdsourcing, and Proliferating Sensors

The field of biodiversity observation and monitoring is subject to rapid change both in regard to the variety of sources and to the volume size of the data that needs to be described, visualised, understood, preserved, and processed. This is due to a number of interrelated factors:

[29]CoDATA Task Group: http://www.codata.org/taskgroups/TGdatacitation/index.html.

[30]RDA Working Group: https://rd-alliance.org/working-groups/data-citation-wg.html.

[31]Trusted Digital Repository Checklists: http://www.crl.edu/archiving-preservation/digital-archives/metrics-assessing-and-certifying-0.

- *Growing Diversity and Productivity of Observation Channels*: Increasing availability of sensor channels lead to larger volumes of usable data. Traditional channels (remote sensing, gene sequencing, field observation) are increasingly supplemented by crowd-sourced observations, and the rapidly growing number of connected *smart devices* in the internet (Hugo et al. 2011).
- Methods using automated markup for metadata and data mining of existing or future publications contribute to increasing volumes (Agosti and Egloff 2009).
- *Storing Observations*: It is becoming increasingly affordable to store and process large volumes of data.
- *Less Expensive Platforms*: It is becoming very affordable to deploy observation platforms such as aerial drones[32] and underwater guided cameras, leading to large, multidimensional data sets at low cost of acquisition. Similarly, cost reductions are set to deliver significant and growing volumes of environmental genomic data addressing aspects of biodiversity which until now have been inadequately recorded.

These factors all combine to put pressure on the traditional architecture, standards, and infrastructure arrangements that have evolved to deal with a less demanding situation. The implications of this growth need to be accommodated in requirements for a scalable architecture.

11.2 The Network of the Future

GEO BON is by definition a network, and it is important to recognise that the concept of a network applies on multiple levels: on an institutional and personal level; as a collaboration network; and with the support of an infrastructure network. This infrastructure includes networks defined physically through protocols, schematically and syntactically through registries and catalogues, and semantically in emergent knowledge networks, ontologies, and vocabularies.

Any future networks, and resulting research data infrastructure, will likely be a combination of all of these and require governance, best practice conventions, standards, and reference implementations to work.

11.2.1 A Vision for Future Data and Services

The vision for a future network extends work done earlier by GEO BON (Scholes et al. 2012), and includes ideas about the generic use cases that it should support. This is summarised largely in the GEO BON Manifesto[33] (Hugo et al. 2013), which

[32]UNEP: http://www.unep.org/pdf/UNEP-GEAS_MAY_2013.pdf.

[33]Agreed by GEO BON Workgroup 8 at the Asilomar All Hands meeting, December 2012.

highlights a set of functions that are expected to be available. These, in turn, influence architecture and standards that are required to support such a network. The GEO BON Data Working Group (Working Group 8) has focused on these, and on developing a working implementation demonstrating the generic use case.

The Manifesto, as set out in updated form below, addresses description, discovery, assessment, access, analysis, and application or reporting, by stating that it is the interest of any specific community to do the following:

- Ensure that scientific data and services are described properly, preserved properly, and discoverable;
- Once discovered, the utility, quality, and scope of data can be understood, even if the data sets are large;
- Once understood; the data can be accessed freely and openly;
- Once accessed, the data can be included within distributed processes, and collated—preferably automatically (Hernandez et al. 2009a, b), and on large scales (the 'Model Web') (Nativi et al. 2013);
- Once processed, the associated mediations and annotations, usefulness, and knowledge gathered can be re-used.

All of this needs to be implemented against the backdrop of:

- Due recognition to the creators of the data, models, and services;
- The push to extend formal metadata with Linked Open Data and persistent identifiers;
- The increased availability of crowd-sourced and citizen contributions;
- A proliferation of devices and sensors; and
- The construction of knowledge networks.

11.2.2 The Role of Standards and Specifications

Standards and specifications are intended, from a formal systems engineering perspective, to *reduce the risk of failure*. The basic aim of this approach is 'Predictable Assembly from Certifiable Components' (Wallnau 2003). The risk of failure is lowered because assembly is made from components certified to meet the specifications and standards. In the type of scalable, open architecture envisaged for GEO BON, the ability of *third parties* to assemble larger systems from components using well-defined interfaces is critical as a contributor to the goal of interoperability and scalability.

Data standards in biodiversity are primarily defined by the Biodiversity Informatics Standards organisation. It is better known by its earlier name 'Taxonomic Databases Working Group'[34] (TDWG). TDWG works with other

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

standards bodies, such as Open Geospatial Consortium (OGC), and has been recognised by them.

11.2.3 A Scalable, Interoperable Architecture

A realistic, shorter-term expression of the goals implied by the manifesto can be summarised as follows (Saarenmaa et al. 2014):

- Allow for data flow from observations through various aggregation and processing/modelling services, supporting evaluation of EBVs and derived indicators;
- Automated and streamlined, as appropriate;
- Using a plug-and-play (service-oriented) approach, supported by robust service provider organisations;
- Coordinated through a GEO BON registry system and linked to the GEOSS Common Infrastructure;
- Transparent to users through multiple channels, portals and applications.

11.2.3.1 General Requirements for a Biodiversity Information Architecture

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 through GEO BON. They will be hosted in a smaller number of data repositories. Additionally:

- The infrastructure must incorporate a federated architecture which will allow many data centres, initiatives, and infrastructures to co-exist and participate;
- While a minimum set of standards is desirable, pragmatism and reliance on brokering and mediation will be the norm for a considerable time to come;
- Human resource, financial, scalability, and institutional constraints will necessitate building the infrastructure using many small contributions in addition to a few large, global ones.

The main components in the information architecture can be divided into three main functions, corresponding to the tasks of (i) data publishing, (ii) data discovery, and (iii) data access. As a fourth function, various applications and uses can be envisaged, and for all functions mediation may be required between services and clients in cases where standardisation of services and vocabularies are not perfect.

There are two options for interoperability architecture, both essentially 'service-oriented', with varying degrees of rigour required for implementation. Firstly, the model proposed by EU BON and others, is based on an Enterprise

Service Bus (ESB), and allows automation of asynchronous workflow and distributed processing as envisaged by the Model Web. Secondly, one can serve a significant proportion of needs with less complex synchronous orchestration, using mostly RESTful Services. These architectures are not mutually exclusive and are likely to co-exist within a systems-of-systems environment.

11.2.3.2 Option 1: SOA and ESB

The Service Oriented Architecture (SOA) is a model, which has achieved 'best practice' status within the Open Geospatial Consortium (OGC). Building on SOA has been recommended also for GEO BON (Ó Tuama et al. 2010) and EU BON (Saarenmaa et al. 2014). In an SOA, different functionalities are packaged as component services that can be orchestrated for specific tasks. An Enterprise Service Bus (ESB), which is a virtual private connector over the Internet, would connect external data sources using various SOA standards (WSDL,[35] SOAP,[36] REST[37] and BPEL,[38] 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., Kepler,[39] Taverna[40]).

11.2.3.3 Option 2: Synchronous, RESTful Services

Some applications do not require orchestration of services to take account of long-running, asynchronous processes, and may not require authentication if data services are in the public domain. In these cases, RESTful HTTP calls, stored in OGC Web Context Documents (XML files defining a collection of RESTful services and their roles) should be adequate to collate information in support of a user requirement. The role that each service plays to achieve the collective outcome will have to be captured for future use, and can potentially be stored in OGC Web Context Documents (XML files defining a collection of RESTful services and their roles), but other methods may also be used.

[35]Web Services Description language (WSDL); http://www.w3.org/TR/wsdl20/.

[36]Simple Object Access Profile; http://www.w3.org/TR/soap12-part1/.

[37]Representational State Transfer; http://www.ibm.com/developerworks/webservices/library/ws-restful/.

[38]Business Process Execution Language; http://docs.oasis-open.org/wsbpel/2.0/OS/wsbpel-v2.0-OS.html.

[39]https://kepler-project.org/.

[40]http://www.taverna.org.uk/.

11.3 Considerations in Respect of Best Practice

11.3.1 Sources of Data and Its Classification

11.3.1.1 Essential Biodiversity Variables

The Essential Biodiversity Variables (EBVs) (Pereira et al. 2013), under development by GEO BON, provide a critical use case for determining requirements for information systems. An EBV is defined as 'a measurement required for study, reporting, and management of biodiversity change'. EBVs provide focus 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. 11.1), 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.

GEO BON has identified six EBV classes. These are listed in Table 11.1 with some candidate 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. Of particular relevance are the EBV definitions and how an EBV is measured. For example, the three EBVs listed for the Species Populations class, can be broken down as illustrated in Table 11.2. 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.

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.

The EBV on abundances and distributions would need to be measured using 'counts or presence surveys for groups of species easy to monitor or important for

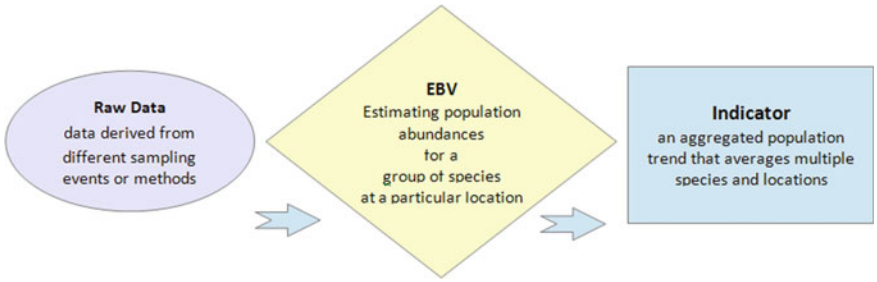

Fig. 11.1 An EBV acts as an intermediate layer between raw data and indicators

Table 11.1 EBV classes with examples

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

Source Adapted from Pereira et al. (2013)

Table 11.2 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 groups) 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	

ecosystem services, over an extensive network of sites, complemented with incidental data'. Such an EBV would be updated at intervals from 1 to 10 years. EBVs have not yet been implemented, but need to be piloted.

Implementation of these specific EBVs calls for integration of data from sites such as those of LTER, and other regular surveys, and from historical and recent data published through GBIF. Integration implies 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. 11.2 as 'ecological' and 'occurrence' domains. Software tools and web services are available to do these computations, for instance from the TRIM,[41] BioVeL,[42] and EUBrazilOpenBio[43] projects. Recent developments within GBIF include support for additional core data elements from survey data,[44] indicating the possibility of incorporating all of these data sources within a single access infrastructure.

[41] www.cbs.nl/en-GB/menu/themas/natuur-milieu/methoden/trim.

[42] www.biovel.eu.

[43] www.eubrazilopenbio.eu/.

[44] www.gbif.org/sites/default/files/gbif_IPT-sample-data-primer_en.pdf.

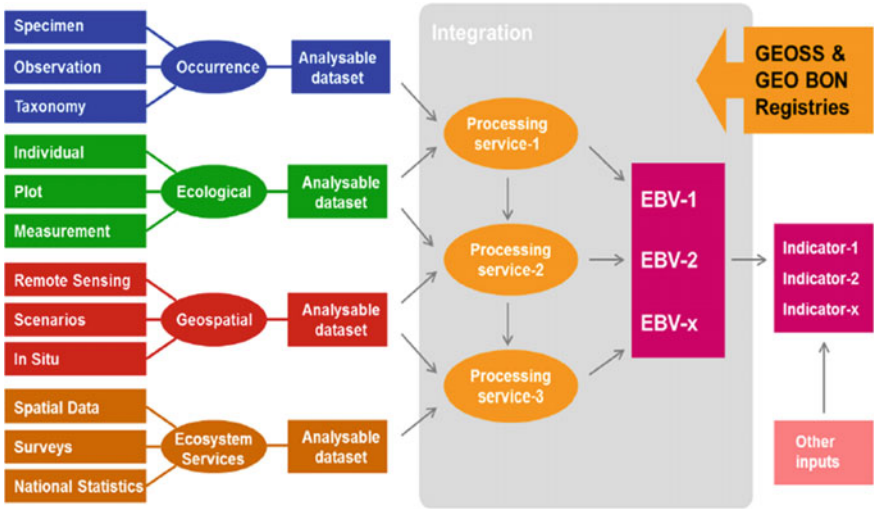

Fig. 11.2 The GEO BON vision of automated, streamlined data flow, end-to-end, from observations to Essential Biodiversity Variables (EBVs), using a plug-and-play service-oriented approach, coordinated through the GEO BON registry system and linked to the GEOSS Common Infrastructure, and transparent to users through portals. *Source* Hugo et al. (2013); modified by Hoffman et al. (2014)

The computation of an EBV of this class involves data cleansing and normalisation and interpolation of values to offer a modelled data surface. Such EBVs could be visualised in a 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.

11.3.1.2 Protocols for Observation

The two largest domains of biodiversity observation are specimen occurrences and biological (natural resource) surveys. The former is frequently based on sporadic, opportunistic collection or observation activity, while the latter consists of repeated sampling at known sites, locations and follows a known protocol from which quantitative estimates of abundance, and at times additional information, can be derived. Hence, the latter method is most appropriate for observing change, but the former can also be used, if the observations sets are large enough and sampling biases can be eliminated by computation (Ariño 2010). Data potentially available through both of these domains are very large. GBIF, which already represents the occurrence domain, currently has mobilised more than 15,000 data sets and is

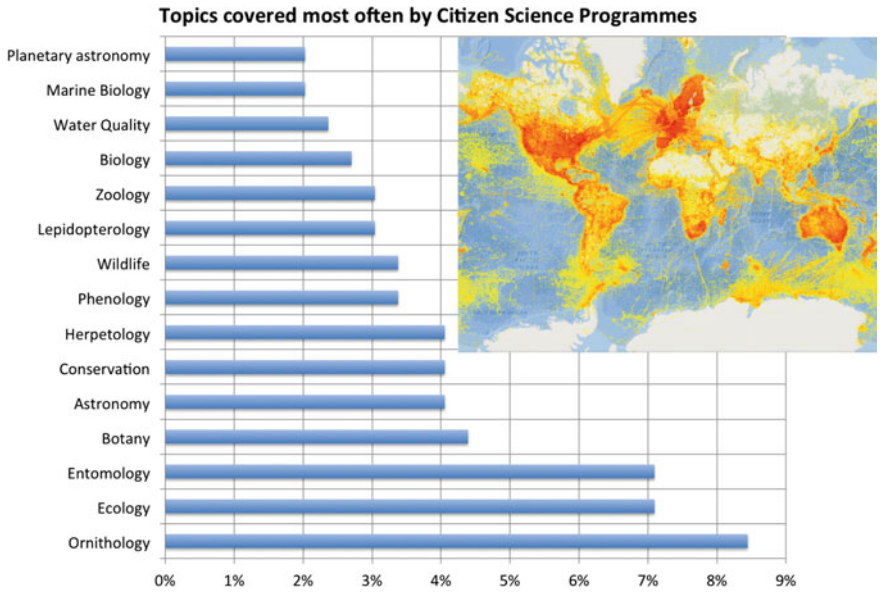

Fig. 11.3 Topics covered most often by Citizen Science Programmes (https://en.wikipedia.org/wiki/List_of_citizen_science_projects#Active_citizen_science_projects). Inset—distribution of GBIF observation data, a large proportion of which originates from volunteer contributions (http://www.gbif.org/occurrence)

expanding to index and integrate data from survey datasets. 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 the environment, forestry, fisheries, and agriculture, which in many cases have not yet started any data sharing activities.

Biodiversity observation is unique in that for species occurrence, most observations are made by volunteers. The EUMON project[45] estimates that 80 % of biodiversity monitoring data comes from volunteers. In Finland, for example, there are 60 different biodiversity monitoring programmes in which 250 person years are spent annually, and 70 % of this is voluntary work. This pattern is similar to some extent many other countries—a summary prepared based on a listing of such volunteer programmes is shown in Fig. 11.3. In the top 15 topics, only astronomy is unrelated to biodiversity.

Volunteer contributions pose a special challenge in respect of introduction of observer bias and strict adherence to observation protocols, and may be used in special circumstances to derive additional EBVs (Kery et al. 2010; Hui and McGeogh 2014).

[45]http://EuMon.ckff.si/index1.php#2.

11.3.1.3 Generic Data Families

The GEO BON working group on data integration and interoperability has developed a classification of generic data families and their interoperability requirements (Fig. 11.4). Data families are grouped according to variations in their spatial, temporal and semantic coverages with each unique combination of these, supported by a vocabulary/ontology, is considered a generic data family. As an example: occurrence, genome, and ecosystem data families all include a reference to a particular place and time, but differ in that occurrence data also references a taxon, genome data references a sequence and ecosystem data 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

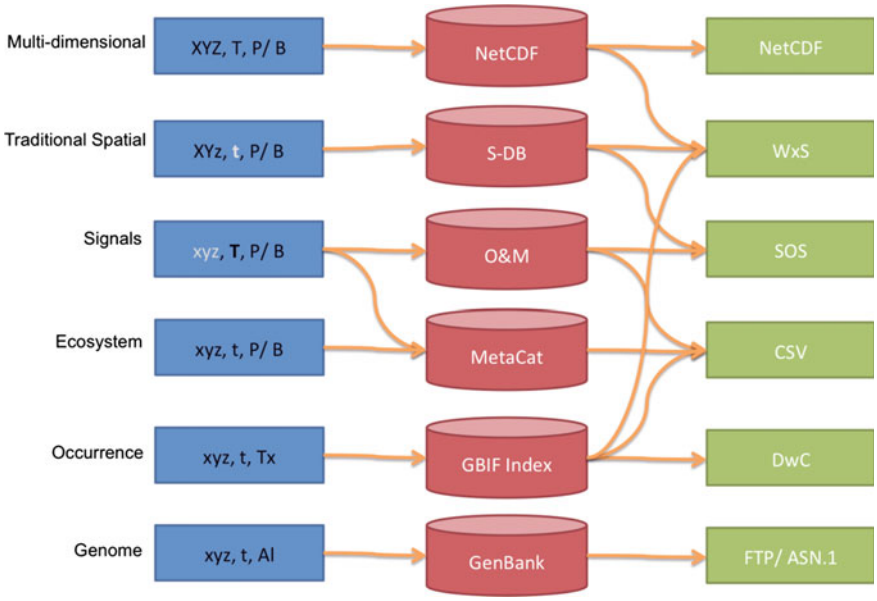

Fig. 11.4 Example generic data families and interoperability requirements. The abbreviations are: S-DB: spatial database; WxS: OGC (Open Geospatial Consortium web services); O&M: OGC Observations and Measurements model; SOS: OGC Sensor Observation Service; CSV: comma separated value; DwC: Darwin Core. The *leftmost* boxes represent typical data families and their dimensions, the *centre* shows typical data storage technology, and the *rightmost* boxes typical services whereby such data is exchanged. Some data storage technologies support multiple service standards. *Source* Hugo et al. (2013)

– **P**: Phenomenon

  mostly physical, chemical, or other contextual data

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

The dimension of a sampling event or specimen applies to all data families.

11.3.2 Published Advice and Guidance

The recommendations from published material discussed here have been incorporated into the 'Specific Implementation Guidance' later in the chapter, as appropriate.

Recent advances in the availability of standards include the development of 'Biological Collections Ontology' (BCO) and the 'Population and Community Ontology' (PCO) (Walls et al. 2014)—bridging a gap in the availability of vocabularies derived from formal ontology to describe the collection of biodiversity data, and to formulate more complex relationships between primary data elements such as evolutionary processes, organismal interactions, and ecological experiments.

11.3.2.1 Research Data Alliance (RDA)

The Research Data Alliance (RDA) produces community consensus on important aspects of research data infrastructure in general, and includes representation from biodiversity and ecosystem data infrastructures.[46] This interest group envisages work in respect of name (vocabulary) services standardisation, with a focus on taxonomy, and the support of improved interoperability. In more general terms, RDA has recently endorsed its first sets of formal outputs, and some of these have a bearing on biodiversity informatics:

- The Data Citation Working Group[47] has produced a clear set of guidelines in respect of implementation of persistent identifiers for data sets.
- The Data Type Registries Working Group[48] aims to standardise the description of complex data types—which in principle includes the 'data families' that can be identified for GEO BON EBVs. This enables processes, visualisations, and other tools to reliably be linked to data services.

[46] https://rd-alliance.org/groups/biodiversity-data-integration-ig.html.

[47] https://rd-alliance.org/groups/data-citation-wg.html.

[48] https://rd-alliance.org/groups/data-type-registries-wg.html.

- The Metadata Standards Catalog Working Group[49] has produced a set of principles, and aim in future to develop a canonical set of metadata elements that can serve as a broker between different metadata schemas in use by communities.
- The Practical Policies Working Group[50] has published its first recommendations in respect of 11 important practical policies for repository management, based on a survey of the research repository community.

11.3.2.2 Global Biodiversity Informatics Conference (GBIC)

The Global Biodiversity Informatics Conference (Copenhagen, 2012)[51] assessed the state of Biodiversity Informatics across four focus areas (Understanding, Evidence, Data, and Culture), and provided a community consensus on the desirable futures for the elements in each of these focus areas (Hobern et al. 2012).

11.3.2.3 GEO Data Management Principles

The GEO Data Management Principles[52] were adopted in short form by the organisation in April 2015, and in full form by the GEO Plenary in November 2015. The 10 principles deal with aspects of discoverability, accessibility, usability, preservation, and curation.

11.3.2.4 EU BON

EU BON published a review and guidelines for its proposed architecture (Saarenmaa et al. 2014) that contains a portfolio of recommendations. These recommendations (39 in all) are strongly supportive of existing projects and initiatives (Lifewatch, BioVEL, EBONE, INSPIRE, LTER, GBIF, to name a few) and provide guidance in respect of service-bus type implementation in a service-oriented architecture.

[49] https://rd-alliance.org/groups/metadata-standards-catalog-working-group.html.

[50] https://rd-alliance.org/groups/practical-policy-wg.html.

[51] http://www.biodiversityinformatics.org/.

[52] https://www.earthobservations.org/documents/dswg/201504_data_management_principles_long_final.pdf.

11.3.2.5 CReATIVE-B and GLOBIS-B

The CReATIVE-B project[53] (2011–2014) dealt with the 'Coordination of Research e-Infrastructures Activities Toward an International Virtual Environment for Biodiversity'. CReATIVE-B enabled collaboration between the European LifeWatch/ESFRI Research Infrastructure and other large-scale research infrastructures on biodiversity and ecosystems in other parts of the world. The project published an integrated Roadmap in 2014 and this serves as high-level guidance in respect of biodiversity infrastructure and data management activities.

GLOBIS-B has as its main aim the definition of research needs and infrastructure services required to calculate EBVs, and will do so by fostering collaboration between scientists, global infrastructure operators, and legal interoperability experts. GLOBIS-B has produced its first outputs, and a recent publication (Kissling et al. 2015) details thoughts on interoperability in support of EBVs. GLOBIS-B correctly identifies a scientific challenge (definition of EBVs) and a technical one (legal and information technology considerations) that need to be addressed.

11.3.2.6 EarthCube and DataONE

These are primarily US-based initiatives, though DataONE has participating data providers from outside the US, and EarthCube has formal collaboration with EU partners. DataONE publishes and maintains best practice in respect of data management,[54] which was reviewed for inclusion into our guidance, and EarthCube has recently published a roadmap[55] and a supporting architecture[56] that also contributed input by way of principles.

11.4 Specific Implementation Guidelines

References quoted in the following sections are available in the supplementary materials on the Springer Website. Supplementary materials are also hosted and maintained on the GEO BON website at http://dataintegration.geobon.org/guidance.

[53] http://www.slideshare.net/dmanset/20140909creativeb-roadmap-interactive.

[54] https://www.dataone.org/sites/all/documents/DataONE_BP_Primer_020212.pdf.

[55] http://earthcube.org/sites/default/files/doc-repository/ECRoadmapv6%203%201.pdf.

[56] https://docs.google.com/document/d/10OhZntRpizn-KaYECXtGY_tcVbanG2kR0OFJ7JZpnWw/edit#.

11.4.1 Recommended Data Management Approaches

This section proposes guidelines for biodiversity data management from three perspectives: that of (1) individual researchers, (2) institutions, projects, or initiatives (such as regional BONs), and (3) from the broader community and GEO BON's perspective. It focuses on the information technology aspects of the challenge to provide an infrastructure in support of EBV calculation. The guidelines support both architectures described above.

For all of these end user categories, we recommend that

- General guidelines in respect of data management be followed (Section A below, and elaborated in supplementary materials), with indications of deficiencies that may exist;
- Specific guidelines to foster semantic interoperability are followed (Section B below). These are also supplemented by online materials and deficiencies are highlighted;
- As a first choice, data be shared in global repositories that serve a specific data family and is well established (Section C below);
- Other data be published and catalogued using widely adopted interoperable service standards and content schema—while recognizing that the community, and especially GEO BON, should play a role in extending such content schema where deficiencies exist (Section D).

Content schema and vocabularies in support of specific EBVs will be required once the community has adopted definitions—GEO BON has a critical role in developing these, and the GLOBIS-B project will make a direct contribution to this effort.

11.4.2 Section A: General Considerations

These considerations apply to all research data infrastructures (Table 11.3).

11.4.3 Section B: Semantic Interoperability

Guidelines in respect of the use of name services (vocabularies, ontologies, and persistent identifiers), and development of a knowledge network as it applies primarily to biodiversity informatics (Table 11.4).

Table 11.3 General guidelines applicable to all research infrastructures

Aspect	Guidance			Reference
	For individual researchers	For institutions and projects	For the community/GEO BON	
Open access	Select open licenses, preferably the most open suitable Creative Commons license by preference possible (CC0, otherwise CC-BY or, if necessary CC-BY-NC—avoid the ND, no-derivates and SA, share-alike options), for all published data sets unless one of a specific set of exceptions apply	Develop data policies in support of open access and open science, and standardise on Creative Commons licenses for all but specific exceptions	Support Creative Commons licenses, and work towards machine-readable, multi-jurisdiction licenses for the valid exceptions not supported by Creative Commons	[1, 4, 5, 8, 10, 39, 40, 42, 43, 46–48]
Federated identity	Make use of globally available resources in this regard, such as EduRoam	If available, use EduRoam as a basis for service and system authentication, and ensure that researchers have access to it	Work with RDA to facilitate a globally available identity resolution framework that can be used by system and service developers	[1, 41, 42]
Data citation	Ensure that data sets are published with a persistent identifier, and make use of persistent identifiers when citing others	Ensure that mechanisms are available for persistent identifiers to be minted for data set publication, and that sufficient infrastructure spending is available for implementation of RDA guidelines in respect of data citation	Promote a culture of data citation and license respect/maintenance. Contribute use cases to RDA working groups on data citation to ensure that the needs of the biodiversity community are included in guidance	[2, 8, 10, 40, 42]
Data types (data families)	Use the guidance below to select a data family appropriate to the data being published. Bear in mind that the publication format is not the same as the format in which the data is best applied in your own context	Make an effort to ensure that data type registries are supported once these become available. Ensure that tools and processing routines are designed and implemented in such a way that the data type registry can be supported		[3, 8, 10]

(continued)

Table 11.3 (continued)

Aspect	Guidance			Reference
	For individual researchers	For institutions and projects	For the community/GEO BON	
Metadata interoperability	Use the guidance below to select an appropriate metadata standard for the data to be published, and ensure that maximal use is made of name services (vocabularies, ontologies, registries of permanent identifiers)	Work towards adoption of metadata standards within the institution. Ensure that catalogues of institutional data offer harvesting end points supported by widely accepted protocols (see below for guidance)	Work towards adoption of metadata standards within the community and develop best practice/guidelines in respect of name service usage, mandatory elements, quality, protocol and lineage description, and other elements supporting re-use of the data	[4, 8, 42, 43]
Name service interoperability	Support appropriate name services (vocabularies, ontologies, and PID registries) wherever these are indicated. Refer to the guidance below in this regard	Develop institutional best practice/guidance in respect of name service usage	Contribute to an support RDA efforts to improve name service interoperability, and actively promote the use of such services by the community	[4, 9, 42, 43]
Data interoperability	Use the guidance below to select a data family appropriate to the data being published. Specifically work towards making sure that the data is not only available in a standardised schema (format), but that it is also available via a standardised protocol (web service)	Ensure that infrastructure exists for implementation of appropriate web services to enable access to standardised data sets	Seek consensus and endorsement within standards bodies of especially content standards for all data families identified below as requiring attention	[4, 8, 42, 43]
Trusted repositories and reliable future access	Make sure that your data is published and archived in a Trusted Digital Repository, with long-term curation policies and contingency planning in place	Take steps to accredit your institutional repository/data service with one of the recommended global initiatives, and register the repository with re3data. Implement recommendations of the RDA Practical Policies Working Group	Support the principle of deposit in Trusted Digital Repositories	[4–8, 42]

Table 11.4 Guidelines in respect of semantic interoperability

Aspect	Guidance			Reference
	For individual researchers	For institutions and projects	For the community/GEO BON	
Use of persistent identifiers	Use persistent identifiers for identification of data sets (see Data Citation above), and for referencing of important dimensions as described below	Ensure that the mechanisms for obtaining PIDs for data sets are available and affordable for researchers in the institution	Assist, within RDA and other initiatives, with the development of a suite of integrated services for PID resolution	[1, 2, 4, 10, 42]
Knowledge networks	Ensure that dimensions of data (see below) make use of recommended name services, and use such name services for provision of keywords in metadata	Adopt institution-endorsed name services and best practice in respect of implementation	Develop standards and infrastructure that allows individual data element annotation, and encourage the use of formal vocabularies and ontologies for such annotation	[4]
Persons and individuals	Obtain an ORCID for use as a persistent identifier in metadata and data	Encourage the use of ORCID within the institution and community		[1]
Sample	Ensure that individual samples (physical samples or biological specimens/tissue, video, audio, images, signals) are assigned a persistent identifier so that analysis and resulting data from multiple sources can be collated. Consider the use of BCO (Biological Collections Ontology)	Develop institutional best practice in respect of sample identifiers, and make use of global identifier services appropriate to the sample type	Support international efforts, such as now emerging in RDA, EU BON, and in GBIF, to explicitly identify and link samples to observations. Assist with the adoption of BCO as a community standard	[4, 10–13, 45]
Protocols and lineage of data	Make use of published and citable protocols and methodology where possible. Publish own protocols independently and assign a persistent identifier. Use these as references in metadata and describe data lineage properly	Encourage the use of published protocols and the publication of institutional or community of practice protocols	Within GEO BON, work towards the development of published and peer-reviewed protocols for monitoring of all EBVs at all relevant scales. Consider hosting a registry of protocols for EBVs	[4, 10, 42]

(continued)

Table 11.4 (continued)

Aspect	Guidance			Reference
	For individual researchers	For institutions and projects	For the community/GEO BON	
Location, spatial coverage, and stratum	Use a standardised vocabulary for referencing locations in data. If institutional or community guidance is not available, use GeoNames.org as a definitive reference for locations on earth. Provide point or bounding box coordinates—preferably in WGS 84 Lat-Long projection—for study areas defined in metadata	Develop institutional guidelines aligned with national or regional directives, while taking cognisance of international standards that may emerge in this respect	Collaborate towards specific community standards (for example using extensions to Darwin Core) to explicitly indicate and reference plots and their coverages/strata in data and metadata. Develop a robust guideline for location, spatial coverage, sampling plot, and stratum references	[1, 10, 54]
Time	Use UTC (Coordinated Universal Time) to denote events within the present, recent past or future (±100 years). Adhere to guidelines for denoting time on historical, paleo/geological and far future scales	Promote institutional guidelines in respect of time, and implement protocols for synchronisation of automated data sensor date and time stamps	Work towards a definitive community consensus for referencing time in the immediate (±100 years) observation space, historical, paleo/geologic, and far future time scales	[1]
Molecular sequence and genetic data	Implement the guidelines and standards promoted by the Genomics Standards Consortium (GSC), including MIGS, MIMS, and MIMARKS—depending on the data type	Promote the guidelines published by GSC within the institution	Continue the current collaboration between GEO BON and GSC with a view to widespread adoption of the standards and its continuous improvement	[4, 10]
Taxonomy	Use the services registered with the Global Names Architecture (GNA) in the first instance to verify taxonomy. Use widely reviewed sources such as Catalogue of Life. Make use of automated services, such as Plazi for taxonomic data mining	Ensure that taxonomy guidelines for data and metadata are aligned with regional directives and guidelines	Develop best practice in respect of taxonomy referencing, considering use cases that involve changes in taxonomic reference	[1, 4, 10, 14, 37, 38, 50, 51, 52, 57]

(continued)

Table 11.4 (continued)

Aspect	Guidance			Reference
	For individual researchers	For institutions and projects	For the community/GEO BON	
Traits and functional diversity	Use one of a number of ontologies/vocabularies aiming to standardise descriptions of traits (Structured Descriptive Data (SDD), the Plinian Core, the Phenotypic Quality Ontology, and the Animal Natural History ontology)	Agree on institutional use of a specific vocabulary or ontology	Mobilise the community to develop interoperability or brokering between the main trait vocabularies and ontologies. Encourage publication of trait datasets via Encyclopaedia of Life TraitBank	[4, 10, 45]
Habitat, biome, biogeographic and biotope classification	Use the descriptions of biomes and biogeographic regions as promoted or directed by national or regional authorities, or ENVO	It is likely that institutional guidelines will be subject to national or regional directives in this regard	Work towards a brokering or interoperability arrangement to align regionally and nationally adopted biome and bioregion descriptions, and define relationships between them	[10, 33]
Life stage	No definitive vocabulary or ontology for life stages is available. Use the approach proposed by MorphoBank to create a checklist of characteristics and states that cannot be duplicated within your own body of work	Develop institutional best practice to guide data published by researchers	Develop an authoritative vocabulary within a standards body such as TDWG as a community consensus	[1]
Species relationship and biological interaction	Use high-level classification—such classification is less contentious—as well as lower level classifications pertinent to the data at hand. Authoritative services in this regard are not yet available. Consider use of PCO (Population and Community Ontology). Make use of new techniques in metagenomics	Develop institutional best practice to guide data published by researchers	Develop an authoritative vocabulary within a standards body such as TDWG as a community consensus. Use the PCO as a basis of such development	[1, 45, 55]

(continued)

Table 11.4 (continued)

Aspect	Guidance			Reference
	For individual researchers	For institutions and projects	For the community/GEO BON	
Ecosystem functions and services	No specific guidance available at present. The best general ontologies to use include the NASA SWEET Ontology and ENVO, if applicable		Work towards development of standardised vocabularies and ontologies for description of ecosystem functions and services	[1, 32, 33]

Table 11.5 Guidelines applicable to data families for which global infrastructures exist

Data family	Metadata and catalogue services	Sustainable international infrastructures	Schematic and syntactic interoperability—service protocols and content standards	Reference
Electronic samples and specimens	A globally available publication platform for audio, video, and image media used as the basis of species identification and traits/character annotation	MorphoBank	Metadata: Site-specific Data Content: SDD or NEXUS/TNT/NeXML Data Deposit: Any valid media file Services: Portal-based search and discovery	[10]
Presence/absence, occurrence data, species survey data	GBIF Indexing is the most appropriate metadata and discovery mechanism, although INSPIRE in Europe also makes provision for such data	GBIF OBIS	Metadata: Darwin Core/ABCD Data Content: Darwin Core/ABCD Data Deposit: IPT/BioCASE Services: Multiple API options provided by GBIF	[10, 19–21, 54, 56]
Allele/genomic	Services as provided by INSDC Consortium members	GenBank DNA Databank of Japan European Molecular Biology Laboratory	MetaData: MIxS compliant Data Content: Multiple upload tools are available, GCDML	[10, 16]

(continued)

Table 11.5 (continued)

Data family	Metadata and catalogue services	Sustainable international infrastructures	Schematic and syntactic interoperability—service protocols and content standards	Reference
Functional genomics/transcriptomics	Services as provided by GEO and ArrayExpress	Gene Expression Omnibus (GEO) ArrayExpress	MetaData: MIAME-compliant Data Content: MIAME-compliant Data Deposit: FTP Upload Services: JSON/FTP	[10, 17, 18]
Phylogenetic data	Metadata is provided by way of a peer-reviewed article—in other words all data submissions are supported by a published article. An OAI-PMH interface is available for metadata harvesting	TreeBASE	Metadata: Published Article Data Content: NEXUS Data Deposit: via web portal Services: OAI-PMH for metadata, portal and RESTful API for data access	[10, 23]
Micro-CT	Service-specific metadata is gathered on submission, no harvestable or machine-readable endpoints	MorphoSource	Metadata: gathered manually on submission Data Content: files produced by scanners Data Deposit: via portal Services: portal search and browse facility	[10, 24]

Table 11.6 Guidelines applicable to data families for which distributed systems and federated access will apply

Data family	Metadata and catalogue services	Aggregating global or regional infrastructures	Schematic and syntactic interoperability—service protocols and content standards	Reference
Traditional spatial data (raster and vector)	OGC Catalogue Services for the Web (CS/W) or OAI-PMH Aggregation to GEOSS Broker	EU BON GEOSS GCMD Biodiversity Catalogue Consider IPT feed to GBIF in respect of species occurrence	Metadata: ISO 19115 preferred, FGDC supported Data Content: domain-dependent Data Deposit: not required—distributed Services: Publish data via OGC WxS services	[19, 25, 26, 34, 35]
Signals and time series observation data	OGC Catalogue Services for the Web (CS/W) or OAI-PMH Aggregation to GEOSS Broker	EU BON GEOSS Biodiversity Catalogue	Metadata: ISO 19115 preferred Data Content: domain-dependent but based on Sensor Markup Language Data Deposit: not required—distributed Services: Publish data via OGC Sensor Observation Services	[25, 26, 34, 35]
Model outputs and multidimensional data	THREDDS and OPeNDAP Aggregation to GEOSS Broker	EU BON GEOSS Biodiversity Catalogue	Metadata: THREDDS crosswalk to ISO 19115 preferred Data Content: domain-dependent Data Deposit: Not required—federated Services: NetCDF/OPeNDAP queries or mapping to WMS	[25–28, 34, 35]
All other tabular data	OAI-PMH serving Dublin Core or EML Metadata	EU BON GEOSS DataOne, KNB, LTER Consider IPT feed to GBIF in respect of species occurrence	Metadata: EML Data Content: domain-dependent DataDeposit: any compatible format Services: download via API	[19, 29, 31, 34, 53]

(continued)

Table 11.6 (continued)

Data family	Metadata and catalogue services	Aggregating global or regional infrastructures	Schematic and syntactic interoperability—service protocols and content standards	Reference
Any other digital object	Media files, grey literature, code, and similar: provide a DataCite metadata record to DataCite and obtain a DOI	DataCite	Metadata: DataCite Data Content: any digital object Data Deposit: not required—distributed Services: DataCite API	[30]

11.4.4 Section C: Specialised Global Infrastructure

For some data types and families, it is best practice to publish data and make it available via established global infrastructures (Table 11.5).

11.4.5 Section D: Aggregators and Open Federated Infrastructures

The data families and types listed below are best published in a federated manner, using standardised service protocols and content standards, with reliance on aggregation of standard metadata implementations to improve accessibility. GEO BON might consider hosting its own metadata aggregator as a component of the GEOSS Common Infrastructure (Table 11.6).

11.5 Conclusions

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 over HTTP, may allow the syntactic exchange of data blocks, but participating systems also need to understand the schema and semantics of the data being delivered in order to process it meaningfully. Unless the data share a common reference model, the exchange implies brokering, mediation, or other semantic processing.

The challenge, then, from the perspective of GEO BON, is largely one of agreeing appropriate content (schematic and semantic) standards for the main data

families appropriate to each EBV. This will not address all requirements, but should go a long way towards creating successful interoperability precedents and simplify the broadening of the scope of application.

11.5.1 What Is Already Achievable?

Researchers, institutions, and regional or global infrastructures or initiatives that follow the guidelines published in the chapter will already make an immense contribution to the components of an interoperable, federated system of systems as envisaged by GEO.

11.5.2 What Needs to Be Improved?

The guidance has indicated for each aspect what role GEO BON can play in coordinating the solutions to non-ideal situations and development of community-endorsed standards, and in general this remains a significant requirement.

If one considers the more specific goal of EBV interoperability: the majority of EBVs still need to be defined by the GEO BON community, and guidance in respect of interoperability standards and software to support these is dependent on these definitions. In practical terms, the tasks at hand are:

- Review the guidance presented here as more EBVs are formalised;
- Identify the main deficiencies in respect of the available interoperability standards that can be used for GEO BON supported EBVs across data families;
- Define extended content standards for the major data exchange service protocols (IPT, OGC WxS, NetCDF, Sensor Observation Services), using patterns and resources that already exist;
- Build mediation tools for mapping of non-standardised data sets, such as those found routinely in MetaCAT and PlantNet repositories, to services that are schematically and semantically interoperable; and
- Build schematic translation tools to serve any content standard over any service syntax.

It remains unclear how large data sets will be made available and included into an interoperable, orchestrated workflow in an open, free environment—the costs and time involved in sub-setting and processing the data may prove to be prohibitive, and it should be appreciated that the concept of having a suite of EBVs available within a distributed, interoperable global system of systems is constrained in many countries by availability of data sets and resources to gather and maintain such data sets.

Despite these constraints, GEO BON hopes to make steady progress in respect of extending the scope of content standards and services that implement them—leading to a set of EBVs available to a variety of end users from a variety of distributed contributors.

References

Agosti, D., & Egloff, W. (2009) Taxonomic information exchange and copyright: The Plazi approach. *BMC Research Notes*, *2*, 53. doi:10.1186/1756-0500-2-53

Ariño, A. H. (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 *Proceedings of the Journal Article Tag Suite Conference (JATS-Con) 2010 [Internet]*. Bethesda, MD: National Center for Biotechnology Information (US). 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.

Goddard, A., Wilson, N., Cryer, P., & Yamashita, G. (2011). Data hosting infrastructure for primary biodiversity data. *BMC Bioinformatics, 12*(Suppl 15), S5. Published online December 15, 2011. doi:10.1186/1471-2105-12-S15-S5.

Hagedorn, G., Mietchen, D., Agosti, D., Penev, L., Berendsohn, W., & Hobern, D. (2011). Creative Commons licenses and the non-commercial condition: Implications for the re-use of biodiversity information. *ZooKeys, 150*, 127–149.

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, V., Poigné, A., Giddy, J., & Hardisty, A. (2009a). Data and modelling tool structures reference model. *Lifewatch Deliverable 5.1.3.*

Hernandez, V., Poigné, A., Giddy, J., Hardisty, A., Voss, A., & Voss, H. (2009b). Towards a reference model for the Lifewatch ICT infrastructure. *Lecture Notes in Informatics 154*.

Hobern, D., et al. (2012). Global Biodiversity Informatics Outlook. *GBIF*. http://www.biodiversityinformatics.org/download-gbio-report/.

Hoffman, A., Penner, J., Vohland, K., Cramer, W., Doubleday, R., Henle, K., et al. (2014). The need for an integrated biodiversity policy support process—Building the European contribution to a global Biodiversity Observation Network (EU BON). 20 p.

Hugo, W., Saarenmaa, H., & Schmidt, J. (2013). Development of extended content standards for biodiversity data. European Geosciences Union (EGU) General Assembly, Vienna, April 8–12, 2013. In *Geophysical Research Abstracts, EGU 2013* (Vol. 15, p. 6968).

Hugo, W., Jensen, S., Onsrud, H., & Ziegler, R. (2011). White Paper 3: Crowdsourcing and Environmental Science. *Eye On Earth Summit, Abu Dhabi*, September 2011. http://www.eyeonearthsummit.org/sites/default/files/WG3_WP3_formatted_Dec5_Final%20check_.pdf.

Hui, C., & McGeoch, M. A. (2014). Zeta diversity as a concept and metric that unifies incidence-based biodiversity patterns. *American Naturalist*, *184*(5), 684–694. doi:10.1086/678125

Inspire Thematic Working Group Species Distribution. (2013). Data specification on species distribution—*Draft Technical Guidelines v.3.0rc3*.

Kery, M., Royle, A., Schmid, H., Schaub, M., Volet, B., Häfliger, G., et al. (2010). Site-occupancy distribution modeling to correct population-trend estimates derived from opportunistic observations. *Conservation Biology, 24*, 1388–1397.

Kissling, W., Hardisty, A., Alonso García, E., Santamaria, M., De Leo, F., Pesole, G., et al. (2015). Towards global interoperability for supporting biodiversity research on essential biodiversity variables (EBVs). *Biodiversity, 16*(2–3), 99–107. doi:10.1080/14888386.2015.1068709.

Michener, W. K., Brunt, J. W., Helly, J. J., Kirchner, T. B., & Stafford, S. G. (1997). Nongeospatial metadata for the ecological sciences. *Ecological Applications, 7*(1), 330–342.

Nativi, S., Craglia, M., & Pearlman, J. (2012). The brokering approach for multidisciplinary interoperability: A position paper. *International Journal of Spatial Data Infrastructures Research, 7*, 1–15.

Nativi, S., Mazzetti, P., & Geller, G. (2013). Environmental model access and interoperability: The GEO Model Web initiative. *Environmental Modelling & Software, 39*, 214–228, January 2013. http://dx.doi.org/10.1016/j.envsoft.2012.03.007.

Ó Tuama, É., Saarenmaa, H., Nativi, S., Bertrand, N., van den Berghe, E., Scott, L., et al. (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.

Pereira, H. M., Ferrier, S., Walters, M., Geller, G. N., Jongman, R. H. G., Scholes, R. J., et al. (2013). Essential biodiversity variables. *Science, 339*, 277–278. doi:10.1126/science.1229931.

Saarenmaa, H., et al. (2014). Architectural design, review and guidelines for standards. *Deliverable 2.1 (D2.1)—EU BON Project, FP 7 Grant 308454.* http://eubon.eu/documents/1/.

Scholes, R. J., Walters, M., Turak, E., Saarenmaa, H., Heip, C. H. R., Ó Tuama, É., et al. (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.

Socha, Y. (Ed.). (2013). Out of cite—Out of mind—The current state of practice, policy, and technology for the citation of data, CODATA-ICSTI task group on data citation standards and practices. *Data Science Journal, 12*, September 13, 2013. https://www.jstage.jst.go.jp/article/dsj/12/0/12_OSOM13-043/_pdf.

Uhlir, P. (2013). The Legal Interoperability of Data, National States Geographic Information Council, Public Resources. http://www.nsgic.org/public_resources/02_Uhlir_Legal-Interoperability-of-Data_NSGIC-Conf_Feb13.pdf.

Vines, T., et al. (2014). The availability of research data declines rapidly with article age. *Current Biology, 24*, 94–97, January 6, 2014. http://dx.doi.org/10.1016/j.cub.2013.11.014.

Wallnau, K. (2003). Introducing Predictable Assembly from Certifiable Components (PACC), News at SEI, Library, Carnegie-Mellon Institute. http://www.sei.cmu.edu/library/abstracts/news-at-sei/architect2q03.cfm.

Walls, R. L., Deck, J., Guralnick, R., Baskauf, S., Beaman, R., et al. (2014). Semantics in support of biodiversity knowledge discovery: An introduction to the biological collections ontology and related ontologies. *PLoS ONE, 9*(3), e89606. doi:10.1371/journal.pone.0089606.

Wieczorek, J., Bloom, D., Guralnick, R., Blum, S., Döring, M., De Giovanni, R., et al. (2012). Darwin Core: An evolving community-developed biodiversity data standard. *PLoS ONE, 7*(1), e29715. doi:10.1371/journal.pone.0029715.

Wooley, J. C., Godzik, A., & Friedberg, I. (2010). A primer on metagenomics. *PLoS Computational Biology, 6*, 1000667. http://dx.doi.org/10.1371%2Fjournal.pcbi.1000667.

Web Links and References Used in the Guidance Tables 11.3, 11.4, 11.5 and 11.6

[1] Hardisty et al. (2013): See Reference section.
[2] Research Data Alliance: Data Citation Working Group: https://rd-alliance.org/groups/data-citation-wg.html.
[3] Research Data Alliance: Data Type Registries Working Group: https://rd-alliance.org/groups/data-type-registries-wg.html.
[4] Hobern et al. (2012): See Reference section.
[5] Refer to supplementary material for a review of licenses and exceptions to open licenses: http://dataintegration.geobon.org/.
[6] Research Data Alliance: Repository Audit and Certification DSA–WDS Partnership Working Group: https://rd-alliance.org/groups/repository-audit-and-certification-dsa%E2%80%93wds-partnership-wg.html.
[7] Research Data Alliance: Practical Policy Working Group: https://rd-alliance.org/groups/practical-policy-wg.html.
[8] GEO Data Management Principles: https://www.earthobservations.org/documents/dswg/201504_data_management_principles_long_final.pdf.
[9] Research Data Alliance: Vocabulary Services Interest Group: https://rd-alliance.org/groups/vocabulary-services-interest-group.html.
[10] Saarenmaa et al. (2014): See Reference section.
[11] Darwin Core and Archive Extensions: http://www.gbif.org/resource/80636.
[12] International Geo Sample Number (IGSN): http://schema.igsn.org/description/.
[13] Research Data Alliance: Management and Curation of Physical Samples: https://rd-alliance.org/bof-management-and-curation-physical-samples-ig.html.
[14] Research Data Alliance: Biodiversity Data Integration Interest Group: https://rd-alliance.org/groups/biodiversity-data-integration-ig.html.
[15] Genomic Standards Consortium: http://gensc.org/mixs/.
[16] Genbank and INSDC Members: http://www.ncbi.nlm.nih.gov/genbank/.
[17] Gene Expression Omnibus (GEO): http://www.ncbi.nlm.nih.gov/geo/.
[18] ArrayExpress: https://www.ebi.ac.uk/arrayexpress/.
[19] Global Biodiversity Information Facility (GBIF): http://www.gbif.org/.
[20] Map of Life: https://www.mol.org/about.
[21] Ocean Biogeographic Information System: http://www.iobis.org/.
[22] MorphoBank: http://morphobank.org/index.php/Documentation/Index#d5e1285.
[23] TreeBASE: https://treebase.org/treebase-web/home.html.
[24] MorphoSource: http://morphosource.org/.
[25] Open Geospatial Consortium Standards: http://www.opengeospatial.org/standards.
[26] Global Change Master Directory: http://gcmd.nasa.gov/.
[27] THREDDS—ISO 19115 conversion: https://geo-ide.noaa.gov/wiki/index.php?title=NetCDF_Attribute_Convention_for_Dataset_Discovery#Open_Geospatial_Consortium_Catalog_Service_for_the_Web_.28CSW.29.
[28] ncWMS—University of Reading: http://www.resc.rdg.ac.uk/trac/ncWMS/.
[29] DataOne: https://www.dataone.org/.
[30] DataCite: http://www.datacite.org.
[31] International Long Term Ecological Research Network (ILTER): http://www.ilternet.edu/.
[32] NASA SWEET Ontology: https://sweet.jpl.nasa.gov/.
[33] EnvO: http://www.environmentontology.org/.
[34] Biodiversity Catalogue: https://www.biodiversitycatalogue.org/services.
[35] GEOSS Broker: http://www.eurogeoss.eu/broker/default.aspx.
[36] EU BON: http://www.eubon.eu/.
[37] Catalogue of Life: http://www.catalogueoflife.org/.

[38] Encyclopedia of Life: http://www.eol.org/.
[39] Plazi: http://plazi.org/news/beitrag/data-sharing-principles-and-legal-interoperability-for-essential-biodiversity-variables/13f96ba8031d1c42c4519d3863e203e8/.
[40] Open Biodiversity Knowledge Management System (OBKMS): http://pro-ibiosphere.eu/getatt.php?filename=oo_4670.pdf.
[41] Research Data Alliance: Federated Identity Management Interest Group: https://rd-alliance.org/groups/federated-identity-management.html.
[42] Creative-B Roadmap: http://www.slideshare.net/dmanset/20140909creativeb-roadmap-interactive.
[43] EarthCube: https://docs.google.com/document/d/10OhZntRpizn-KaYECXtGY_tcVbanG2kR0OFJ7JZpnWw/edit#.
[44] GEO BON Information Architecture Principles: https://www.earthobservations.org/documents/cop/bi_geobon/geobon_information_architecture_principles.pdf.
[45] Walls et al. (2014): See Reference section.
[46] Traitbank: http://eol.org/info/516.
[47] Hagedorn et al. (2011): See Reference section.
[48] Canadensys—Open Licenses: http://www.canadensys.net/2012/why-we-should-publish-our-data-under-cc0.
[49] iNaturalist—Open Licenses: http://inaturalist.tumblr.com/post/138557593458/changes-to-gbif-licensing-requirements.
[50] Penev, L., Agosti, D., Georgiev, T., Catapano, T., Miller, J., Blagoderov, V., et al. (2010). Semantic tagging of and semantic enhancements to systematics papers: ZooKeys working examples. *ZooKeys, 50*, 1–16. doi:10.3897/zookeys.50.538.
[51] Penev, L., Lyal, C., Weitzman, A., Morse, D., King, D., Sautter, G., et al. (2011). XML schemas and mark-up practices of taxonomic literature. *ZooKeys, 150*, 89–116. doi:10.3897/zookeys.150.2213.
[52] Sautter, G., Agosti, D., & Böhm, K. (2007). Semi-Automated XML Markup of Biosystematics Legacy Literature with the GoldenGATE Editor. In *Proceedings of PSB 2007, Wailea, HI, USA, 2007*. http://psb.stanford.edu/psb-online/proceedings/psb07/sautter.pdf.
[53] Michener, W. K., Brunt, J. W., Helly, J. J., Kirchner, T. B., & Stafford, S. G. (1997). Nongeospatial metadata for the ecological sciences. *Ecological Applications, 7*(1), 330–342.
[54] Wieczorek, J., Bloom, D., Guralnick, R., Blum, S., Döring, M., De Giovanni, R., et al. (2012). Darwin Core: An evolving community-developed biodiversity data standard. *PLoS ONE, 7*(1), e29715. doi:10.1371/journal.pone.0029715.
[55] Wooley, J. C., Godzik, A., & Friedberg, I. (2010). A primer on metagenomics. *PLoS Computational Biology, 6*, 1000667. http://dx.doi.org/10.1371%2Fjournal.pcbi.1000667.
[56] Inspire Thematic Working Group Species Distribution. (2013). Data specification on species distribution—*Draft Technical Guidelines v.3.0rc3*.
[57] Catapano, T. (2010). TaxPub: An extension of the NLM/NCBI journal publishing DTD for taxonomic descriptions. In *Proceedings of the Journal Article Tag Suite Conference (JATS-Con) 2010 [Internet]*. Bethesda, MD: National Center for Biotechnology Information (US). http://www.ncbi.nlm.nih.gov/books/NBK47081/.