GROUP ON EARTH OBSERVATIONS | GEO BON

Group on Earth Observations Biodiversity Observation Network (GEO BON)

Principles of the GEO BON Information Architecture

Version 1.0 – 14 June 2010

GROUP ON EARTH OBSERVATIONS | GEO BON

Authors

Éamonn Ó Tuama, GBIF, Denmark (co-lead)

Hannu Saarenmaa, University of Helsinki, Finland (co-lead)

Stefano Nativi, IMAA-CNR, Italy

Mark Schildhauer, NCEAS, USA

Nicolas Bertrand, NERC, UK

Edward van den Berghe, IOBIS, USA

Lori Scott, NatureServe, USA

Meredith Lane, NBII, USA

Gladys Cotter, NBII, USA

Dora Canhos, CRIA, Brazil

Roman Khalikov, ZIN-RAS, Russia

Document Version History

Version	Date	Modified by	Comment
V1.0	2010-06-14		

GROUP ON EARTH OBSERVATIONS | GEO BON

Table of Contents

1. Introduction

While preparing the detailed implementation plan (GEO BON 2010) for the Group on Earth Observations Biodiversity Observation Network (GEO BON) in December 2009-March 2010, it became obvious that all the biodiversity networks that will make up the GEO BON boast such a multitude of data and employ such a wide variety of access mechanisms that all their pertinent features which affect data integration and interoperability cannot be sufficiently covered in the plan. Of necessity, the plan had to focus on activities and deliverables. Yet there is a need to gather in one place and briefly document the "diversity of biodiversity networks" and their chief characteristics. This is the purpose of this *Companion Document* to the implementation plan. When implementation of GEO BON actually begins, more detailed surveys and design documents will be produced.

Working Group 8 (WG8) of the GEO BON is concerned with data integration and interoperability. It is a significant challenge to coordinate, standardize, and manage *in situ* data that are collected by disparate institutions and individuals for differing purposes. As envisaged in the Concept Document (GEO BON 2008), GEO BON, building on existing networks and initiatives, should develop an implementation plan *for an informatics network* in support of the efficient and effective collection, management, sharing, and analysis of data on the status and trends of the world's biodiversity, covering variation in composition, structure and function at ecosystem, species and genetic levels and spanning terrestrial, freshwater, coastal, and open ocean marine domains. It is probably safe to say that all this will be impossible to achieve in the short term. However, several activities are being done by various research groups all the time. GEO BON can leverage this, by connecting them and supporting their work so that integrated and novel products can be produced more efficiently.

WG8 thus has a mandate that is somewhat different from other GEO BON working groups. It is not directly aiming at certain products about biodiversity. Instead, it will focus on building permanent structures and linkages that will support producing those and similar products more efficiently. Close interactions with the other GEO BON groups is thus necessary in order to understand existing infrastructures and ascertain requirements. At the Asilomar meeting 22-25 February 2010, preliminary interviews were held with representatives of each of the thematic working groups. Interview notes are available (http://imsgbif.gbif.org/CMS/DMS_.php?ID=1056). The template for a more detailed survey of requirements for the thematic areas is presented in Annex 1.

The GEO BON Concept Document covers data integration and interoperability only in very general terms. Alone, it is not a sufficient basis for the aspects of the implementation plan that WG8 is concerned with. More extensive guidance can be found in the documentation from the GEOSS Architecture and Data Committee (ADC) and in existing interoperability pilot projects that have been prototyping the GEOSS information system (Figure 1). WG8 has therefore taken as its goal to introduce these concepts into the design of GEO BON.

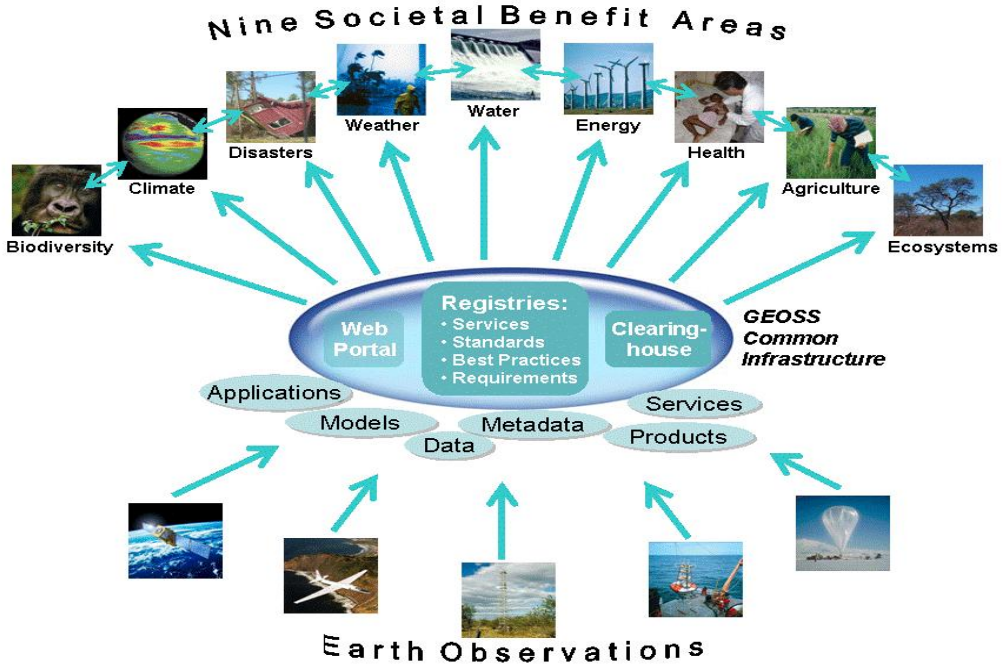

Figure 1 – Conceptual operational view diagram of the GEOSS Common Infrastructure (GCI) and its relationship to observations and end-users in the nine Societal Benefit Areas (SBAs).

2. Review of the concepts of GEO BON

The presentation by Scholes provides a general introduction to GEO BON (Scholes 2009). According to the GEO BON Concept Document, the network should make use of existing resources, including data, data systems and catalogues; it should be comprehensive, dealing with all aspects of biodiversity on a global scale; it should provide a framework that is scientifically robust, that enables setting of priorities, gap analyses, and facilitates modelling of biodiversity change in a changing environment. The GEO BON system will largely be built from contributing systems that have their primary responsibility at regional, national or sub-national scales. GEO BON will add value by connecting such networks together. At the global level, GEO BON will

build on the experience of GBIF, ILTER, IODE, and others, but fill gaps in data and extend the coverage to other types of data such as genetic and ecosystem levels. Data sources will encompass field observations (including those by volunteer networks of citizen observers), specimen and image collections, and remote sensing imagery. Work will be needed to harmonise observation standards, to promote use of multidisciplinary interoperability standards, and to define and update interoperability arrangements –applying the *System of Systems* approach promoted and implemented by GEOSS (Nativi, 2010).

GEO BON will help to promote data publication principles in support of full and open availability of data and information, recognizing relevant international instruments and national policies and legislation.

One more aspect that is of interest to WG8 is the end-to-end concept of GEO BON. The network is built to deliver major products based on integrated data and information. The system will, for instance, enable quantifying and mapping the drivers of biodiversity change, including threats; recording the impacts of biodiversity change with a focus on vital ecosystem functions and resulting services; and reporting the current state and changes in biodiversity over time. GEO BON will enable integrated assessment across scales: from extensive surveys (e.g., remote sensing of land cover, productivity) to intensive site-based observations (e.g., *in situ* long term ecosystem research). Products across all these areas are being identified by various GEO BON working groups. It will be necessary for WG8 to review them, and adjust its plan accordingly so that production of these deliverables can be supported by the information infrastructure.

3. Main approach of the information architecture

In keeping with the GEOSS conceptual approach, the informatics infrastructure for GEO BON should be based on a decentralised and distributed architecture. Service Oriented Architecture (SOA) is the leading approach to build such networks. It has been solidified through formal definitions by several organisations and networks, for instance, in the LIFEWATCH Reference Model (Hernandez-Ernst 2009), and has been followed by GBIF from its inception (cf. Saarenmaa 2005). By adopting an SOA approach, inventory and discovery via a system of metadata catalogues and registries becomes a core component of the GEO BON network and provides the foundation for integration with other community clearinghouse systems. The design facilitates development of complex systems implementing interoperability at the enterprise level: services establish a high form of abstraction encapsulating both application and process logic.

For interoperability within GEOSS, the GEO BON infrastructure must implement the SOA international standards and Earth system science multidisciplinary best practices, e.g. the GEOSS

Standards and Interoperability Forum (SIF) interoperability arrangements. In fact, Earth system information is usually encoded using one or more common (generally agreed upon) representations (models) such as the ISO 211/OGC Reference Model and the Orchestra Framework (Percivall, 2010), while SOA standards are built using a combination of industry specifications.

However, the SOA pattern presumes that any service producer and consumer share both a distributed computing protocol and a semantic domain which is comprised of a data and metadata model. In heterogeneous and complex systems (like a system of systems), this is generally not the case. Thus, the introduction of broker components, implementing mediation services, has proven to be a good solution to implementing interoperability for a number of issues including discovery services. Experiments on this type of solution were successfully demonstrated in the context of GEOSS IP3 (Interoperability Process Pilot Project) (Khalsa 2009) and AIP-2 (Architecture Implementation Pilot –phase 2) pilots for Climate Change and Biodiversity (GEOSS AIPa; GEOSS AIPb; IPCC 2007).

GEO BON will need to contribute to the GEOSS Common Infrastructure (GCI). The GCI consists of a web-based portal, a clearinghouse component for searching data, information and services, and registries containing information about GEOSS components, standards, best practices, and requirements (Figure 2).

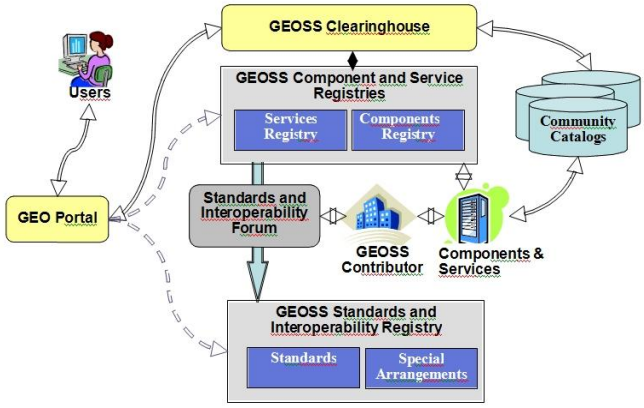

Figure 2. Interactions of GEOSS Registries, Portal and Clearinghouse[1]

Various community portals and clearinghouse components may contribute to GEOSS by implementing the necessary international standards to contribute to the GCI.

[1]Extracted from: *GEOSS Core Architecture Implementation Report (http://portal.opengeospatial.org/files/?artifact_id=24315)*

The primary resources 'outside' of the common infrastructure are the web sites, services, data, and portals operated by GEO Members and Participating Organizations (Figure 2).

GEO BON will contribute to the GCI by registering these resources and implementing interoperability solutions mediated via the GCI Web Portal, Clearinghouse Catalogue (using a Distributed Community Catalogue) and registries. Whether GEO BON will need a portal or portals (i.e. Community Portal) needs to be investigated, as well as how such components would relate to existing ones (e.g., the GEO Portal and GBIF Data Portal). GEO BON will build on, not duplicate, existing systems.

Each of the candidates and building blocks of such infrastructure will be discussed below.

4. Data types and data content

GEO BON will need to support a rich observations' information model such as that described in the OGC Observations and Measurements specification[2] , and cover such data types as species occurrences based on points (e.g. point locations), lines (e.g., transects), polygons (e.g., range distributions); imagery/gridded data (remotely sensed images, coverages); population and time series data (e.g., density, abundance, age stratification, trends).

Strategies for traversing across data from the different levels of organisation of biodiversity (genes -> species -> ecosystem) are of interest to GEO BON. Spatial and taxonomic referencing are two main ways for linking across levels and a key concern for GEO BON, to enable integration, will be to ensure that genetic sequence data are documented with a georeference and the environmental parameters of the extraction environment using the appropriate standards.

Because of the diversity of data types across levels of organisation, GEO BON will need to adopt a broker model architecture (see section 3) in which data interoperability is achieved through mediation (many data content standards in use; interoperability achieved through mapping of concepts at consumer end) rather than harmonisation (data providers agree up-front on a common data exchange schema). This is in the scope of interoperability arrangements which characterize a system of systems approach.

It is likely that the data providers will be required to serve more complex, integrated data. For instance, the early focus of GBIF providers was on museum collection specimens and field observations treated as "primary data" (i.e., what was observed, where and when) (GBIF 2009). However, many of these data actually were part of much richer datasets but those were not in the

[2] http://www.opengeospatial.org/standards/om

interest of GBIF at that time, and were left out. Conversely, data providers to the LTER network, which primarily deals with rich ecological datasets, also have primary occurrence data that is of interest to GBIF. However, LTER providers do not have interfaces for serving these data in a way that GBIF can use. If these networks would agree on common interfaces, and a supporting framework for each other's data types, much more data would end up being made available.

5. Networks and their information resources

5.1. Existing global networks

GEO BON builds on existing networks such as those summarised in the tables below. Not all existing networks are based on a similar architecture. Some operate centralised databases while others follow the SOA model. Some kind of open access principle is common, but standards-based discovery (via metadata catalogues) and access mechanisms have not been implemented widely.

In Table 1 we compare major examples of the existing global networks that carry biodiversity data. For each network, a concise description is provided that includes what aspects of biodiversity are addressed by the network (e.g., genetic, species, ecosystem, terrestrial, freshwater, marine), the data types managed, and the standards used.

Table 1. Characteristics of some examples of major global networks that make biodiversity observation data and information available.

Name	Eco-system Coverage	Taxonomic or Topical Coverage	Data or Information Types Covered	Data and Metadata Standards	Archi-tecture	Access and Protocols
BOLD (Barcode of Life Database; project of Consortium for the Barcode of Life - CBOL)	Any	All organisms	DNA barcode and specimen records	Darwin Core; GCMD DIF		Website-based search; REST services
DataONE (Data Observation Network for Earth)	Any	Ecology, earth science	Data, metadata, workflow	FGDC, EML, Dublin Core, Darwin Core, NETCDF, GCMD	Distributed	Various

Name	Eco-system Coverage	Taxonomic or Topical Coverage	Data or Information Types Covered	Data and Metadata Standards	Archi-tecture	Access and Protocols
Discover Life	Any	All organisms	Species data and information			
EOL (Encyclopedia of Life)	Any	All organisms	Species data, information, images	TDWG	Aggregator	DiGIR, TAPIR, web services
FAO GeoNetwork	Any	Geospatial data	Maps, satellite imagery, spatial datasets	GeoNetwork Catalogue		
FishBase	Marine, aquatic	Fish	Fish data and information		Centralized	Portal/ Search
GBIF (Global Biodiversity Information Facility)	Any	All organisms	Organism occurrence, Names data, provider and dataset metadata	TDWG, EML	SOA: portal, registry, providers	Open access: DiGIR, TAPIR, web services
GCMD (Global Change Master Directory)	Any	Any	Datasets, documents, tools	FGDC	Registry, providers	Portal/ Search
GISIN (Global Invasive Species Information Network)	Any	All invasive species	Organism occurrence data, species information	TDWG, EML	GBIF IPT with GISIN extension	Open access: TAPIR, web services
GCOS (Global Climate Observing System)	Any	physical, chemical and biological properties of climate system	atmospheric, oceanic, terrestrial, hydrologic, and cryospheric components			

Name	Eco-system Coverage	Taxonomic or Topical Coverage	Data or Information Types Covered	Data and Metadata Standards	Archi-tecture	Access and Protocols
GOOS (Global Ocean Observing System)	Marine	Physical oceanographic data	Temperature, salinity, etc.	ESRI Shapefile; ESRI GeoDatabase	Distributed	
GTOS (Global Terrestrial Observing System)	Terrestrial	Biodiversity, climate, coastal, Forest and Land Cover, Glacier, Hydrology, Land, Permafrost, Water	Biological, Physical			GOSIC Portal (Global Observing Systems Information Center)
ILTER Network (International Long-term Ecological Research Network of Networks)	Terrestrial	All	Ecological	EML	Registry, providers	
IUCN Red List (International Union for the Conservation of Nature / CITES)	Any	CITES species	Species information; range maps	EML; no standard data model	Internal database	Website-based search
KNB (Knowledge Network for Biodiversity)	Any	All	Ecological	EML	Distributed metadata catalogue (Metacat)	Metacat protocol
OBIS (Ocean Biogeographic Information System; project of Census of Marine Life - CoML)	Marine	All organisms	Organism occurrence data	TDWG, ISO 19115/19139	Aggregator	Portal/ web services

Name	Eco-system Coverage	Taxonomic or Topical Coverage	Data or Information Types Covered	Data and Metadata Standards	Archi-tecture	Access and Protocols
speciesLink	focus on Brazil	all organisms	Organism occurrence data, integrated with species information	DarwinCore, DiGIR, Tapir	distributed	Open access: DiGIR, TAPIR, web services
UNEP -WCMC (United Nations Environment Programme World Conservation Monitoring Centre)	Any	Habitats (e.g. reefs, mangroves) and species of conservation and protection concern	Interactive maps (protected areas), organism and habitat-type atlases)	OGC	Internal databases; links to external tools	Portal/ search of internal databases

5.2. National and regional networks

Because there is a multitude of national and regional networks, listing them all is simply not feasible. Table 2 lists examples of some of the most advanced or largest networks.

These networks can be incorporated in GEO BON if they adhere to the data exchange standards and protocols that will be adopted by GEO BON, e.g., by developing specific interoperability arrangements where required. Some of these networks are quite advanced in their development. Mapping all of their specialized functions to the global level may not therefore be possible, although adherence to, or adoption of, common standards for data and metadata would allow access to the data they share. Harmonisation or mediation, and ensuring compatibility of approaches will be key to the successful development of GEO BON across regions.

In Europe, for example, LIFEWATCH has adopted an SOA model based on Orchestra and therefore fits closely with the GEOSS and GEO BON information architectures. Provided LIFEWATCH becomes a legal entity funded by EU member states, GEO BON will have a clear mechanism for mobilising key biodiversity data across Europe and across all the GEO BON themes.

Table 2. Characteristics of some examples of major national or regional networks that make biodiversity observation data and information available.

Name	Geographic Coverage	Topical Coverage	Data or Information Types	Data and Metadata Standards	Architecture	Access and Protocols
AKN (Avian Knowledge Network)	Western Hemisphere	Birds	Occurrence observations	TDWG	Aggregator	Web services
ALA (Atlas of Living Australia)	Australia	All organisms, molecular to ecological	Specimen, observation, ecological, sequence	TDWG, LSID	Distributed and federated	TAPIR, web services
Artdatabanken / Artportalen ("art" = species in Swedish)	Sweden	All organisms	Occurrence observations	TDWG	Centralized	TAPIR, web services
ASEAN BISS (ASEAN Biodiversity Information Sharing Service)	ASEAN Member States	Protected areas, wetlands; all organisms	Maps (protected areas, wetlands); species information	Catalogue of Life (names)		
Biota-Africa (Biodiversity Monitoring Transect Analysis in Africa)	Africa	Terrestrial ecology and biodiversity	Weather, remote sensing, geospatial, soil, vegetation, species richness, animal diversity, socioeconomic	Not stated	Centralized	
DAISIE (Delivering Alien Invasive Species Inventories for Europe)	Europe	Terrestrial, aquatic; all organisms	Species information		Centralized	
EBONE (European Biodiversity Observation Network)	Europe	Terrestrial	Species indicators, vegetation and habitat quality measures		Not available online	

Name	Geographic Coverage	Topical Coverage	Data or Information Types	Data and Metadata Standards	Architecture	Access and Protocols
GAP (Gap Analysis Program)	USA	Terrestrial	Land cover datasets	FGDC	Centralized	Portal/ search; download
GOS (Geospatial One Stop)	USA	Geospatial	Maps, satellite imagery, spatial datasets	FGDC		
IABIN (Inter-American Biodiversity Information Network)	Western Hemisphere	Invasive species, pollinators, ecosystems, protected areas	Specimen data, species information, metadata, publications	TDWG	Distributed	Portal/ Search, web services
INBio (Instituto Nacional de la Biodiversidad)	Costa Rica	Species, ecosystems, conservation	Specimen data, species information, ecosystems	TDWG	Centralized	Portal/ Search
IPANE (Invasive Plant Atlas of New England)	USA northeast	Invasive plant species	Specimen data, maps			
LIFEWATCH	Europe	Any	Any			Not online yet
NatureServe	Western Hemisphere	Species, ecosystems, terrestrial, aquatic	Species and ecosystem information, distribution and status, population viability, invasiveness impact ranks, climate change vulnerability	FGDC Biological Profile	Federated (aggregated and synthesized by NatureServe from its members)	Website search, web services, download
NBII (National Biological Information Infrastructure)	USA (to global)	Terrestrial, aquatic, marine; species data and information	Landcover, occurrence observation, and ecological data and information; documents	FGDC, TDWG	Distributed	Portal/ Search, web services

Name	Geographic Coverage	Topical Coverage	Data or Information Types	Data and Metadata Standards	Architecture	Access and Protocols
NBN (National Biodiversity Network)	UK	Terrestrial	Occurrence observation, habitat, geospatial	TDWG	Aggregator	Portal/ Search, web services
PBIF (Pacific Biodiversity Information Forum)	Oceania	Marine, terrestrial	Specimen, occurrence observation; protected areas	TDWG	Distributed (links)	Portal/ Search
REMIB (World Network for Biodiversity)	Mexico	Any	Specimen; maps	TDWG, Z39.50	Centralized	Portal/ Search
SIBIS (South African National Biodiversity Institute's Integrated Biodiversity Information System)	South Africa	Threatened and other species; area checklists	Species data and information; maps			Portal/ Search of internal databases
SCAR-MarBIN (Scientific Committee on Antarctic Research Marine Biodiversity Information Network)	Antarctica	Biology, Ecology, Geography	Species surveys, research results; metadata; documents			Portal/ Search, web services
SERVIR (Regional Visualization and Monitoring System)	Meso-America, East Africa	Earth observations (satellite); forecast models	Visualizations, analytical tools, images			Portal/ Links
SpeciesLink	Brazil	all organisms	Organism occurrence data, integrated with species information	DarwinCore, DiGIR, Tapir	distributed	Open access: DiGIR, TAPIR, web services
ZooInt (Zoological Integrated Retrieval System)	Russia	Animals	Specimen data; classifications		Centralized	Website / internal databases

6. Discovery services and registries

As GEO BON will be based on a SOA featuring loosely coupled components that can be joined in arbitrary ways, discovery of available resources via a system of registries and community metadata catalogues is an essential requirement for the GEO BON network and provides the foundation for integration with other clearinghouse systems. The GEO ADC has designed the GEOSS Core Architecture as a system for exchange and dissemination of observations and guided implementation of its initial operating capability: the GEOSS Common Infrastructure (GCI). Consisting of a GEO Web Portal, a Clearinghouse, and Registry components (Figure 2), GCI provides a process to register, discover and use services accessible using the Interoperability Arrangements recognized by GEOSS –see the SIF activity.

GEO BON, as part of the wider GEOSS can exploit the functionality provided by the GCI. It is expected that all components, services, standards and special interoperability arrangements contributing to GEO BON will be discoverable through the GEOSS Clearinghouse and should therefore be entered in the appropriate registry (Components Registry, Services Registry, Standards and Special Arrangement Registry - http://geossregistries.info/). The GEOSS portal provides web-based entry forms for populating the registries. One of the main tasks for GEO BON is thus to identify the main components contributing to the network, list the services that they provide and the standards or special interoperability arrangements used by those services. Metadata catalogues and registries are two particular types of service fundamental for resource discovery and will normally be maintained by individual communities of practice (e.g. GEOSS Distributed Community Catalogues).

As with metadata catalogues, communities of practice may also maintain their own specialist registries. It is not yet clear how much integration is envisaged between such community registries and the GEOSS Registries, but any GEOSS-equivalent components, services and standards should certainly be registered. The next sections list some of the main metadata catalogues and registries that are expected to play a role in GEO BON. It is envisaged that these catalogues would connect to the GEOSS Clearinghouse by implementing the required interface based on the OGC CSW (Catalog Service for the Web) specification.

6.1. GBIF UDDI, GBRDS, and Metadata Catalogue

GBIF maintains a UDDI Registry at http://registry.gbif.net/uddi/web for participating nodes on its network to advertise their data and services. Their web service access points are registered thereby enabling GBIF to harvest data into a dynamic, regularly refreshed cache which is fronted by a web portal providing unified search and retrieval across the whole network (http://data.gbif.org).

At the time of writing, over 300 data providers and some 10,000[3] resources have been included in GBIF registry. The GBIF Global Biodiversity Resources Discovery System (GBRDS), now at the alpha version (http://gbrds.gbif.org/), will significantly enhance the GBIF UDDI service by creating a single annotated index of publishers, institutions, networks, collections, datasets, schemas and services. The GBIF metadata infrastructure, planned for 2010, will feature a centralised, indexed cache of harvested metadata documents derived, through reciprocal sharing agreements, from both GBIF Participants' metadata catalogues and other participating networks' catalogues. It will support multiple metadata models natively including Ecological Metadata language (EML), ISO 19115/19139, Natural Collections Descriptions (NCD), and FGDC Biological Profile.

6.2. International Long Term Ecological Research (ILTER) Network

ILTER is developing a distributed system of metadata catalogues and has recommended a metadata profile for the network based on EML. Some of the ILTER member nodes already contribute metadata through the KNB, e.g., Taiwan Ecological Research Network (TERN) and Japanese Long Term Ecological Research Network (JaLTER).

6.3. Knowledge Network for Biocomplexity (KNB)

The Knowledge Network for Biocomplexity (KNB) (http://knb.ecoinformatics.org/) has created a set of open source software tools for use by the ecological community including Metacat, a metadata database, and EML. About 20 other groups/networks in several countries and regions of the world use KNB software and their data holdings are accessible via the KNB catalogue, and through individual portals.

6.4. NASA GCMD

NASA's Global Change Master Directory (GCMD) (http://gcmd.nasa.gov/) is a web based catalogue that enables users to locate and obtain access to more than 30,000 descriptions of Earth science datasets and services covering all aspects of Earth and environmental sciences. The GCMD goals are closely aligned with those of the International Directory Network (IDN) of the Committee on Earth Observation Satellites (CEOS) which works to foster the exchange of information among international agencies. The metadata standard used is Directory Interchange Format (DIF) although other standards are supported through cross mappings, e.g., ISO 19115, North American Profile ISO 19115, FGDC, ESRI profile of FGDC. By providing subset views of the full metadata catalogue, the GCMD enables participating organisations to maintain and

[3] http://www.gbif.org/participation/data-publishers/who-is-publishing/

document their data in one place without having to create their own online directory while at the same time contributing to the GCMD general search pages for scientists in other disciplines to access and use. Over 100 portals are in use.

6.5. NBII

The main metadata profile in use in the NBII is the NBII Biological Profile of the FGDC Content Standards for Digital Geospatial Metadata, but EML is also accepted. Participating nodes make their metadata records available through a weekly harvesting process. NBII also participates in two other FGDC clearinghouse initiatives, the National Spatial Data Clearinghouse (NSDI) and the Geospatial One-Stop (GOS) sharing those biological databases that are geospatially referenced.

The NBII Clearinghouse uses Mercury (http://mercury.ornl.gov/), a web-based system for search and retrieval of data, developed by the Oak Ridge National Laboratory (ORNL). Mercury harvests metadata and data from distributed participating servers, builds a centralized index and provides search interfaces to allow users to perform simple, fielded, spatial and temporal searches across these metadata sources. Several metadata standards are supported including FGDC, Dublin-Core, EML and ISO-19115. Mercury is based on a SOA and supports various services such as Thesaurus Service, Gazetteer Web Service, UDDI Directory Services, RSS, Geo-RSS and OpenSearch.

7. Interoperability and information management services

Developing interoperability arrangements for sharing varied and complex data types is a demanding process. Not all can be implemented in the short term and therefore some priorities will need to be agreed on. What these priorities are will depend on the kind of early products that GEO BON is aiming for. We identify two categories of data for GEO BON which ideally will be combined: i) data required for the delivery of the specific products identified by the thematic working groups, and ii) data for supporting the vision of GEOSS as an informatics network, i.e., integration of existing biodiversity observation networks for long-term benefits.

Several standards are available to aid interoperability arrangements in GEO BON. These are listed in Table 3 and include standards for metadata, data exchange and transfer protocols.

Table 3. Standards for metadata, data exchange and transfer protocols.

Name	***Brief Description***
ABCD	"ABCD 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." http://www.tdwg.org/activities/abcd
BioCASe	The BioCASe protocol is based on DiGIR but adapted for use with ABCD encoded data. Its main user is the Biological Collection Access Service for Europe (BioCASE - http://search.biocase.org/). http://www.biocase.org/products/protocols
CSDGM	The Content Standard for Digital Geospatial Metadata (CSDGM), (FGDC-STD-001-1998), the US Federal Metadata standard, "provides a common set of terminology and definitions for the documentation of digital geospatial data." http://www.fgdc.gov/standards/projects/FGDC-standards-projects/metadata/base-metadata/index_html
CSDGM - Biological Data Profile	Biological Data Profile of the Content Standard for Digital Geospatial Metadata "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." http://www.fgdc.gov/metadata/geospatial-metadata-standards
CSDGM – Profile for Shoreline Data	Metadata Profile of CSDGM for Shoreline Data "addresses variability in the definition and mapping of shorelines by providing a standardized set of terms and data elements required to support metadata for shoreline and coastal data sets. The profile also includes a glossary and bibliography." http://www.fgdc.gov/metadata/geospatial-metadata-standards
Darwin Core	The Darwin Core is a set of standards including a glossary of terms intended to facilitate the sharing of information about biological diversity. It is based primarily on taxa, their occurrence in nature as documented by observations, specimens, and samples, and related information. http://rs.tdwg.org/dwc/index.htm
DiGIR	Distributed Generic Information Retrieval is a protocol that provides unified access to distributed databases allowing clients to retrieve information from distributed servers. It uses HTTP as transport mechanism with messages and data encoded in XML (Darwin Core). http://digir.sourceforge.net/
Dublin Core	"The Dublin Core metadata standard is a simple yet effective element set for describing a wide range of networked resources. The Dublin Core standard includes two levels: Simple and Qualified. Simple Dublin Core comprises fifteen elements; Qualified Dublin Core includes three additional elements (Audience, Provenance and RightsHolder), as well as a group of element refinements (also called qualifiers) that refine the semantics of the elements in ways that may be useful in resource discovery." (http://dublincore.org/documents/usageguide/). http://dublincore.org/documents/dcmi-terms/
EML	"Ecological Metadata Language (EML) 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... 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 dataset." http://knb.ecoinformatics.org/software/eml/
MIENS	MIENS – Minimum Information about an Environmental Sequence, 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 environment associated with a sequence. http://gensc.org/gc_wiki/index.php/MIGS/MIMS/MIENS
Natural Collections Descriptions	Natural Collections Descriptions (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. http://www.tdwg.org/activities/ncd/
OAI-PMH	The Open Archives Initiative Protocol for Metadata Harvesting (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. http://www.openarchives.org/pmh/
OGC CSW	The Open Geospatial Consortium Catalogue Services for the Web (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". Note that, from an interoperability perspective, this is not a single standard as it encompasses several non-interoperable search technologies. http://www.opengeospatial.org/standards/cat
OGC WCS	The Open Geospatial Consortium Web Coverage Service (WCS) "supports electronic retrieval of geospatial data as "coverages" – that is, digital geospatial information representing space-varying phenomena." http://www.opengeospatial.org/standards/wcs
OGC WMS	"The OpenGIS® Web Map Service (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." http://www.opengeospatial.org/standards/wms
OGC WFS	"The OpenGIS® Web Feature Service (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. " http://www.opengeospatial.org/standards/wfs
ISO 19115	"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." http://www.iso.org/iso/iso_catalogue/catalogue_tc/ catalogue_detail.htm? csnumber=26020 . 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, and a North American Profile (NAP) for the USA and Canada is under development. The new ANZLIC metadata standard used in Australia and New Zealand complies with ISO 19115.
ISO 19139	"ISO/TS 19139:2007 defines Geographic MetaData XML (GMD) encoding, an XML Schema implementation derived from ISO 19115."

	http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=32557
TAPIR	Designed as a generic tool that can be applied to domains other than biodiversity and natural science collections data, the TDWG Access Protocol for Information Retrieval (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. http://www.tdwg.org/activities/tapir/
Taxon Concept Schema	"The Taxon Concept Schema (TCS) provides a standard for taxon names and taxon concepts in the exchange and integration of biodiversity and natural history data." http://www.tdwg.org/activities/tnc/

8. Ontologies, thesauri, dictionaries, semantic mediation

GEO BON is a network of networks. Despite of all the efforts towards standardization of data and protocols described in the previous section, we must accept the fact that these networks use, and will continue to use, differently defined data and terminology. Therefore, additional layers of semantics that allow integration of data need to be adopted by GEO BON. This exercise is well known also in large corporations and governments where parallel information systems may exist. A term "Semantic Enterprise Architecture" has been coined to describe this dimension (http://www.mkbergman.com/859/seven-pillars-of-the-open-semantic-enterprise/). It seems clear to us that such an architecture must be designed also for GEO BON. Fortunately, technology is rapidly making this feasible.

The growth of the Internet, along with formal standards for the exchange of metadata through the Web, have created stunning new opportunities for enhancing collaborative research in biodiversity through the development of consistent ways of expressing the observations and measurements that constitute the basic data informing scientific researchers' models and analyses. Ontologies, thesauri, and dictionaries are no longer simply lists of terms (words, or concepts) to be read by humans, to help them understand the meaning and relationships among the values contained in some dataset. These modern variations of "controlled vocabularies" are now constructed in standardized syntaxes that allow for computers to rapidly exchange, search, manipulate, and even "reason" with these constructs, thereby providing some major conveniences to researchers in terms of powerful capabilities for operating on data. While HTML opened the doors to appealing exchange of graphical and natural language information via the Web, the new standards of OWL/RDF (ontologies) and SKOS (thesauri) will enable a "Semantic Web" where information is not just "rendered" as in a normal Web page, but rather can be conditionally purposed or transformed due to the additional rich content that is associated with it via its semantic links to

concepts expressed in W3C-sanctioned formats for ontologies and thesauri (Berners-Lee 2001).

Scientists will certainly agree that any understanding of natural phenomena depends upon our having some shared "model" of the various concepts essential to our disciplines. Our education and research experience naturally inculcates an ability to converse using specialized terms, with reasonable assurance that we have some common understanding what we mean (semantically) when we refer to "species", and how these are affected by "climate" or "habitat", etc. Specialized scientific terms such as these, while varying perhaps in their details and subtleties in interpretation by discipline or even individual preference, nevertheless have great utility in leading to efficient communication. However, while there is undoubtedly heuristic value for GEO BON in compiling a glossary or dictionary of scientific terms using some standard Web exchange syntax, our focus here will be on the application of these approaches in the service of data integration and interoperability.

Before demonstrating why we believe that ontologies in particular, but also thesauri and even simple controlled vocabularies or glossaries, will prove invaluable in facilitating data interoperability, it is useful to recall how data interoperability challenges will arise in the context of GEO BON.

GEO BON activities collectively investigate aspects of biodiversity that encompass a number of different themes, e.g., ranging from impacts on ecosystems services in the ocean, to understanding patterns in the genetic variability of microbes in soil, while also spanning broad spatial and temporal scales of focus, from plot level to regional and global levels. This heterogeneity in research interests inevitably entails heterogeneity in the data resources which inform the subsequent scientific analyses and models.

It is beneficial for the distinct GEO BON working groups to try to standardize their protocols and datasets to the greatest extent possible through discussion and agreement, as many of them are doing, since this will enhance the value of the data to a broader range of researchers by providing a common understanding of the contents of the data, and the data's appropriateness for investigating various research questions. Moreover, this standardization will enable the construction of generalized database schema that can serve large quantities of data in a consistent and scalable way. But there will undoubtedly still remain a large number of variations in the way data are collected due to specific research or logistical motivations, and this will lead to significant challenges in attaining optimal interoperability of the data for integrated analysis.

Furthermore, when GEO BON groups (and others) attempt to pass beyond the boundaries of the data collected using their own internally well-conceived design, as inevitably happens when studying the interactions of biodiversity phenomena within a complex, dynamic biotic and abiotic

environment, then data models conceived externally by other researchers or projects are essentially black boxes, where the specific details of the observations and measurements collected by "other researchers" - that is, critical metadata - are often lacking or presented using inconsistent formats. The quest for additional data to inform any given analysis typically leads to an arduous, inefficient, and error-prone process of trying to find the data, interpret whether it is appropriate for one's need, and then integrate it with one's existing data to accomplish a richer analysis.

The goal of Working Group 8 is to address the above issues by providing generalized solutions that can cut across working group activities in ways that simplify data interoperability with a minimum of effort. Efforts to develop a unified approach to ontology construction and deployment should provide immediate advantages with regards to data interoperability and integration within GEO BON. While the technical details of constructing proper ontologies are beyond the scope of this paper, it is relatively easy to understand how ontologies and thesauri will provide benefit, and also to understand how they differ in capability from each other, and even more so, from simple controlled vocabularies such as glossaries or dictionaries.

In the simplest case, it will be useful if all data resources of interest to GEO BON researchers had some minimal information describing the resource. In earlier sections, we have described several well-established frameworks for capturing this information, e.g. in the case of specimen records, the Darwin Core standard, or for ecological data, the more generic Ecological Metadata Language (EML). In both cases (and there are many others, e.g., the FGDC's CSDGM for geospatial metadata, or the Dublin Core metadata element set) the metadata fields specify what type of information is desired. For example, information about "spatial location in geo-coordinates", or "data set creator", or more specifically, the content of some column of data in a spreadsheet, e.g., "attribute label", "attribute definition", and "attribute units". In this example, the attribute label might be "len" indicating the label as it might actually be listed in the header line of a spreadsheet, or attribute definition in a SQL table creation statement; while the attribute definition might be "standard length of fish", and the attribute units might be "centimetres". Note that often what the scientist or database developer has captured is simply the attribute label, which will often tend to be abbreviated, and sometimes cryptic, and rarely immediately interpretable without conferring with the creator of that data structure.

Highly standardized data, such as gene sequences, or biological specimen records, can derive great utility from agreeing upon a common metadata standard, and even a common standard for storing the data themselves. This is the case for biological specimen records, where a standard like Darwin Core is already providing major utility to researchers wanting to confederate records

from highly distributed and internally heterogeneous botanical and zoological collections. But beyond this "core", the ways that various associated contextual variables relating to the surrounding habitat's physical or ecological structure, such as co-occurring species, micro-climate characteristics, local hydrological features, etc., quickly complicate matters.

9. Organism names and habitat classifications

There are numerous taxonomic names databases and several notable initiatives to organize the naming systems in use by custodians world-wide. Obviously, they will be accessed by the various networks that make up GEO BON. Their services need to be registered as part of the GCI for seamless access by GEO BON applications.

The Integrated Taxonomic Information System (ITIS) is a partnership of federal agencies and other organizations from the United States, Canada, and Mexico, with data stewards and experts from around the world. The ITIS database is an automated reference of scientific and common names of biota of interest to North America. ITIS data are available through web services, described at http://www.itis.gov/ws_description.html. The KNB also provides APIs to search the ITIS database of taxonomic nomenclature, see http://knb.ecoinformatics.org/software/.

Species 2000 http://www.sp2000.org/ is an autonomous federation of taxonomic database custodians, involving taxonomists throughout the world. It provides access through web services, web portal download and CD-ROM to names of over 60% of world species. ITIS and Species 2000 work together to create The Catalogue of Life http://www.catalogueoflife.org, the goal of which is a comprehensive catalogue of all species on earth.

WoRMS, a combination of several species lists of marine groups is managed through the World Register of Marine Species, see http://www.marinespecies.org/.

The Global Names Index (GNI) s the first component of a semantic environment for biology called the Global Names Architecture (GNA). GNI itself is a fairly simple list of names, with reference to who holds the names, and links back to the sources of the names. See http://www.globalnames.org/data_sources for a list of participating scientific names repositories. GNI has been developed by GBIF and the Encyclopedia of Life (EOL). GBIF is currently extending the basic scope of the GNI to create a dynamic index of taxonomic catalogues and annotated species checklists. It serves as a global name service broker capable of serving multiple taxonomic resources through a single and consistent access point (see http://names.gbif.org/).

For habitat classifications, there are too many systems in the world to cover here. The U.S. Geological Survey (USGS) is leading the effort within the ecosystems societal benefit area of

GEOSS to classify and map global ecosystems in a standardized, robust, and practical manner at scales appropriate for on-the-ground management. The global ecosystems mapping task is creating a globally agreed, robust, and viable classification scheme for terrestrial, freshwater, and marine ecosystems and initiating a mapping approach to spatially delineate the classified ecosystems. See http://rmgsc.cr.usgs.gov/ecosystems/method.shtml for details on the conceptual approach and mapping methodology.

Observations of biodiversity in the field are made of biological concepts - not names. Hence the ability to uniquely identify the concept are paramount. Scientific names, unfortunately, are not alone sufficient for this purpose, but need to be accompanied by additional information about in what sense the name or classification has been used. Synonym lists that map historical names to current concepts are also needed. Such semantic information can best be attached to a globally unique identifier, such as LSID or URI, that is shareable between the various networks. Taxonomic services are increasingly offering such identifiers, and GEO BON needs to promote their use. Data integration, in particular, can benefit when datasets can automatically be united using shared identifiers.

10. Workflow of services and integration of applications

Between the acquisition and integration of data into a useful product, there usually is a long chain of transformations and analytical steps. WG8 will consider the needs of other working groups in order to find out what kind of services are required to build intermediate aggregated data (including workflows, etc), that might be needed to support them, and identify suitable networks such as GBIF offering primary occurrence data which, depending on fitness for use, can be transformed into secondary, derived aggregated products. This is closely related to plans for modelling by WG7. It might be possible to design a chain of various services, offered by different providers, which could be used to produce GEO BON deliverables.

In keeping with GEO IP3 and AIP-2 experiments (see Nativi 2009), a “system of systems” needs to manage and serve more than measurements and data: it must support modelling resources, allowing ad hoc, on demand service chaining. Three architecture patterns were recognized for this service chaining: a) User defined (transparent) chaining: the human user manages the workflow. b) Workflow-managed (translucent) chaining: the human user invokes a workflow management service that controls the chain (the user is aware of the individual services). c) Aggregate service (opaque): the user invokes a service that carries out the chain, with the user having no awareness of the individual services (see Figures 3, 4 and 5).

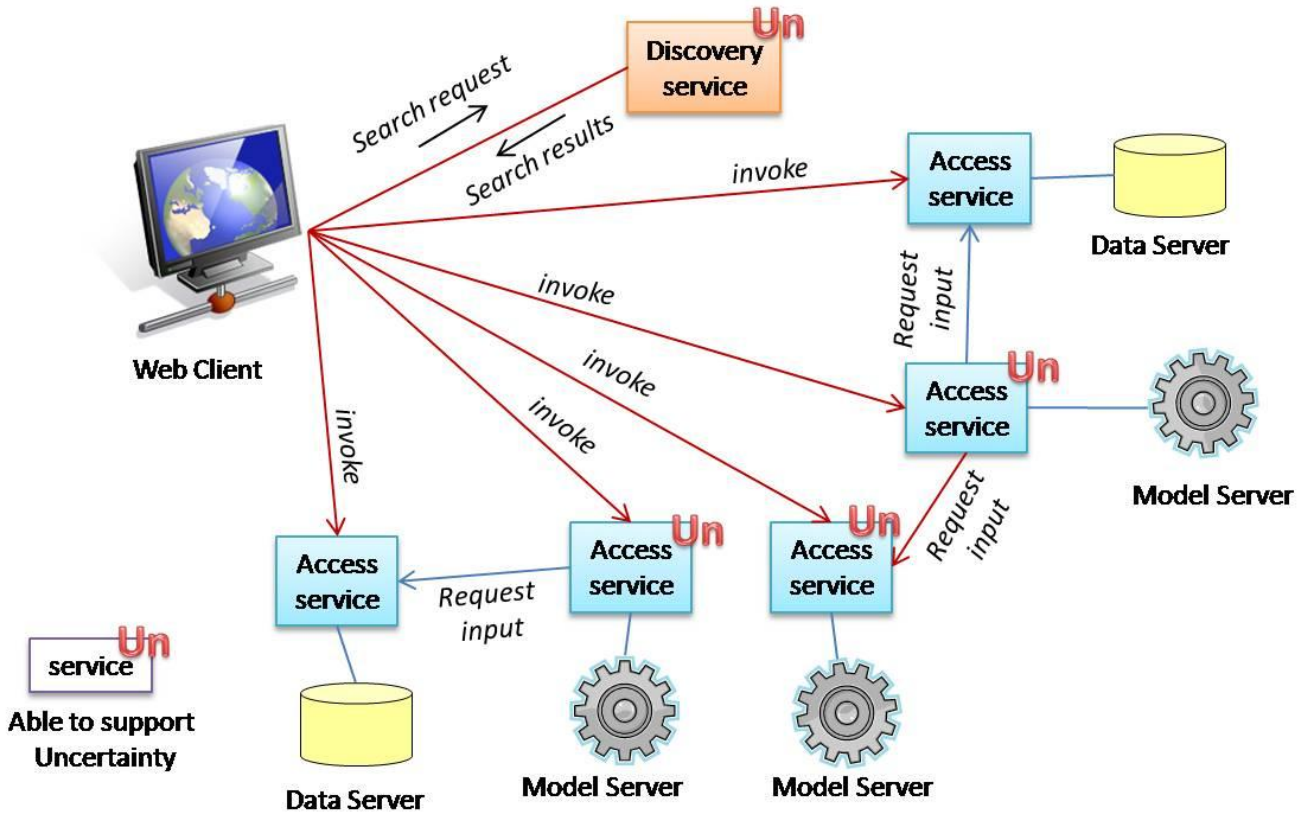

Figure 3. Service Chain: transparent architecture pattern.

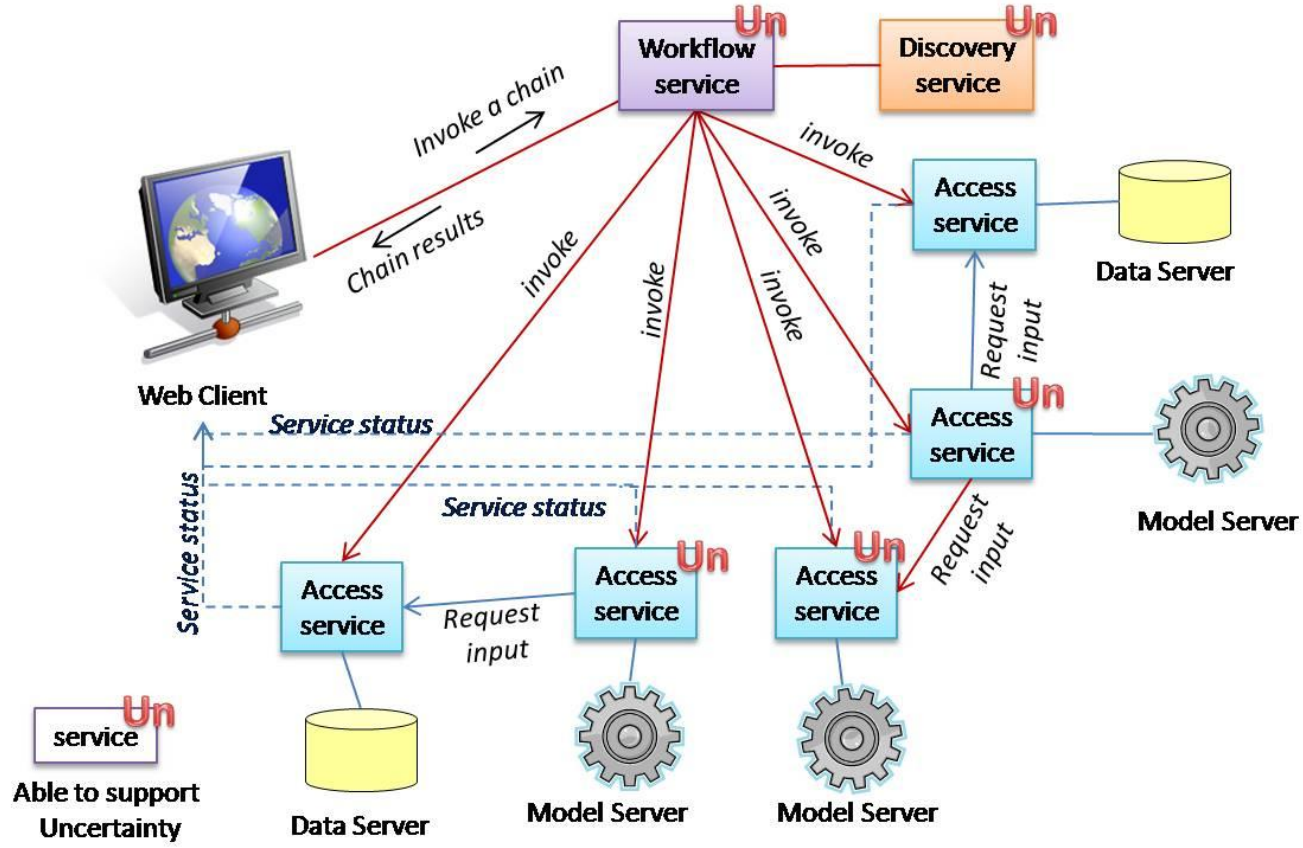

Figure 4. Service Chain: translucent architecture pattern.

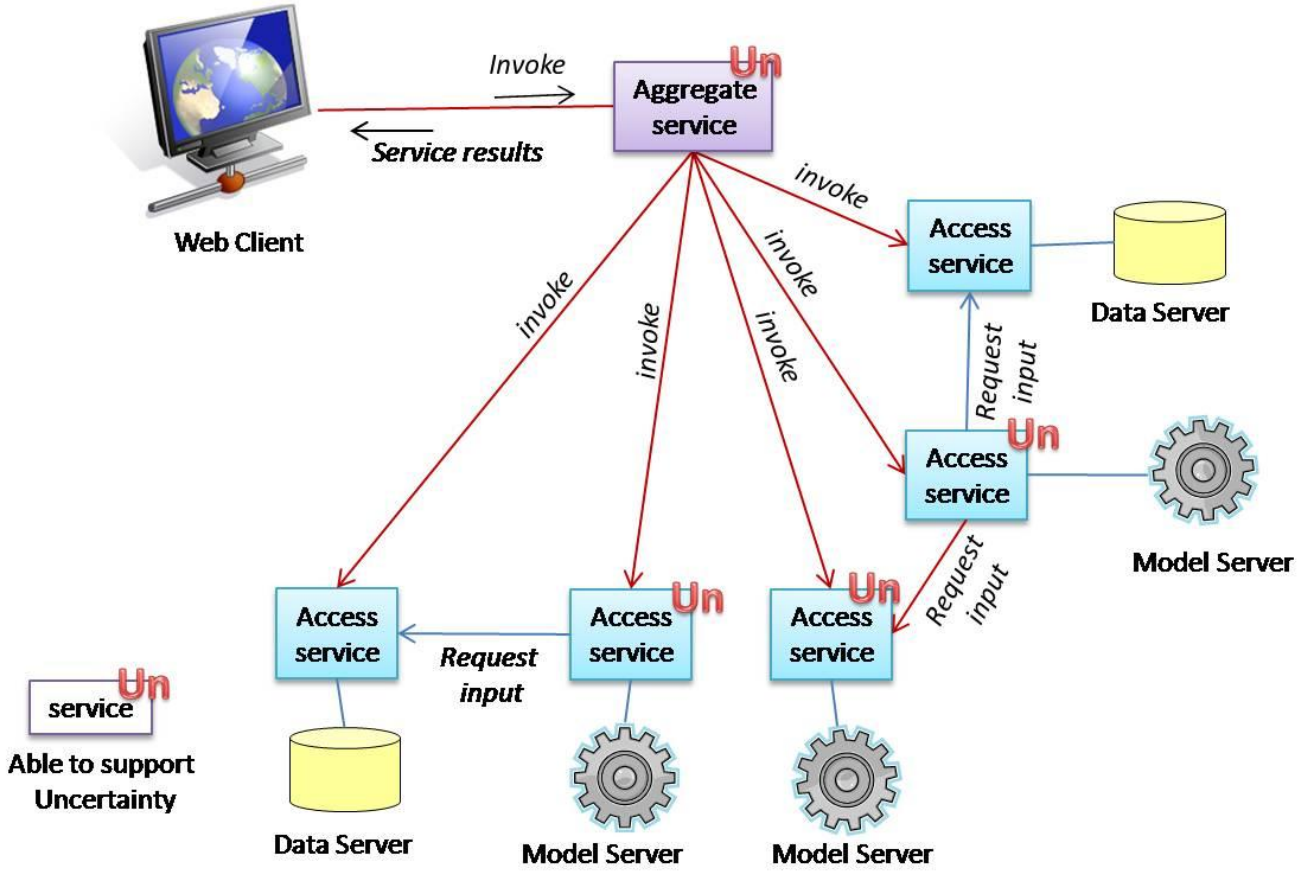

Figure 5. Service Chain: opaque architecture pattern.

These patterns differ primarily in the visibility of the services to the user, the chaining flexibility, and the user control of chaining. Presently, most of the chaining environmental services realise the opaque and translucent patterns. This is not only for technological reasons (e.g.the lack of effective and simple chain definition protocols and tools). In fact, human users who construct a new chain or invoke an existing chain of services should determine the semantic validity of the results of a service chain. The present technologies do not address whether the results of a chain are semantically valid. Indeed, the advanced semantic services (ontologies, thesauri, dictionaries, semantic mediation), discussed in the previous chapter, may contribute to address this challenge by providing the required knowledge.

Moreover, according to ISO 19119, important factors to be considered for the semantic evaluation of a chain result are uncertainty and error propagation. This information must be included in the

data models (specific metadata section) and must be considered by every service node which participates in a chain of services.

GEO BON must be able to manage all the three architecture patterns to enable discovery of biodiversity, ecological, and environmental service nodes on the Web and on-demand adaptive chaining of those nodes. These advanced services will take full advantage of international open standards, in particular those developed within the Open Geospatial Consortium (OGC), ISO TC211 and GEOSS contexts. Uncertainty and error propagation information should be supported by the architecture patterns. The new GEOSS AIP phases (e.g. the AIP-3 use scenarios on biodiversity and climate change area) and the GEO "Model Web" task will provide valuable experiences on these topics.

GEO BON will also need to consider various toolboxes/scientific workflow systems such as those offered by Kepler (https://kepler-project.org/), BioMoby (http://www.biomoby.org/) and Taverna (http://www.taverna.org.uk/) amongst others.

Climate change & biodiversity applications: A GEOSS Architecture Implementation Pilot

A typical biodiversity application scenario requires modelling the impact of climate change on species distribution (see Santana 2008, Nativi 2007, 2009). To achieve this, heterogeneous data resources (e.g. biodiversity, ecological, climatological and environmental resources) and processing services (e.g. implementing Ecological Niche Modelling (ENM) algorithms) are required to interoperate. An interoperability framework implements the required functionalities, e.g., modelling error propagation, uncertainty estimation, heterogeneous data sources mediation, service chaining, etc. In the framework of GEOSS AIP (phase 2), a general service chaining model for Climate Change and Biodiversity applications was designed and tested (see Figure 6). The results were successful; the new AIP climate change and biodiversity scenarios (phase 3) are based on this model.

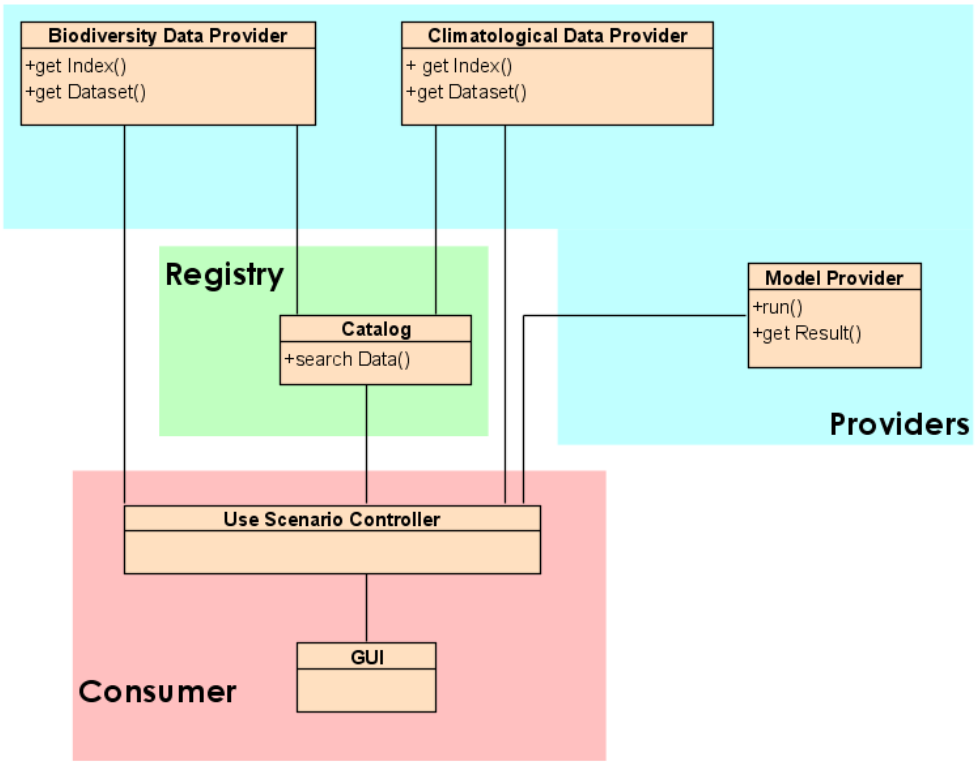

Figure 6. Service chaining framework for Climate Change and Biodiversity applications.

Figure 6 depicts the logical components chained by the service interoperability framework to build Climate Change & Biodiversity applications. Their functions are as follows:

- Biodiversity Data Provider: a component which is able to provide biodiversity data.
- Climatological Data Provider: a component which is able to provide climatological data.
- Catalogue: performs queries on the available biodiversity and climatological datasets. It allows filtering of datasets based on spatial and temporal metadata, data provider, keywords and so on. It may implement distribution and mediation functionalities through the same service interfaces.
- Model Provider: runs ENM on the selected biodiversity and climatological datasets.
- Use Scenario Controller: enables the running of a workflow implementing the business process of the typical biodiversity scenario described above. Generally, it is controlled by the user through the GUI (e.g. a Web client).
- Graphical User Interface (GUI): The component for user interaction.

11. Portals, search engines, querying and harvesting

The GEO Portal will provide a web-based interface for searching and accessing the data, information, imagery, services and applications available throughout GEOSS, including a user interface to databases, services and other portals. An assessment period was launched in 2008 to evaluate various prototypes for the GEO Portal. The prototypes have been mainly user interfaces to locate the services catalogued on the GEO Clearinghouse. No attempt to actually pull together information from the sources and present or analyse it has yet been made.

The search facilities of LTER/ILTER, NBII, KNB, and NASA GCMD work similarly. Datasets can be located based on metadata keywords. Thousands of datasets are available for download, but as they have thousands of different data models, data have not been integrated in any of these networks. Only metadata have been integrated.

The GBIF Data Portal works differently as it has a unified information model onto which data from various resources are mapped. It harvests and integrates selected information from a growing network, currently about 300 data providers with 10,000 datasets. However, it has been estimated that this is still only about 10-20% of existing primary biodiversity data. In particular, networks in non-GBIF countries such as Brazil, China and Russia have to be accessed through their national interfaces. On the other hand, the GBIF Data Portal does not yet provide a metadata catalogue for searching across datasets (but see section 6).

Various services providing aggregated information based on these integrated data systems are emerging, e.g., integration of GBIF occurrence data with IUCN / UNEP—WCMC Protected Areas (http://www.wdpa.org/) and the Global Register of Migratory Species (http://groms.gbif.org/); predictive maps of distribution (LifeMapper, http://specify5.specifysoftware.org/Informatics/informaticslifemapper.html; AquaMaps, http://www.aquamaps.org); predicting distribution of crop wild relatives (http://www2.gbif.org/PosterCCC31low-res.pdf).

How can GEO BON build on these experiences? There are several possibilities:

- Include metadata capabilities in GBIF data providers that are compatible with those of other networks. Then make GBIF data available for searching in these networks.
- Make selected LTER/ILTER data resources also GBIF data providers. Those resources that have occurrence data would need to be mapped to the GBIF data model (Darwin Core).
- Broker agreements that the GBIF Data Portal also becomes a GEO BON Data Portal by

harvesting data also from non-GBIF GEO member countries. These additional data can then be shown optionally.

- After the above exercise that will increase content, perform gap analyses and show in what geographic areas and organism groups analyses of various kinds (trends etc.) can be made.
- Promote registration of value-adding services such as LifeMapper in GEO Clearinghouse. Promote establishing more such services.
- Build chains of these value-adding services using workflow tools for the purpose of producing the outputs wanted by the other working groups. This will mean that these services become permanently available and can be reused to produce similar or related outputs later with less effort. Make these value-chains available on GEO Portal and biodiversity portals.
- Ensure registration of all biodiversity community endorsed standards and, in particular, those of TDWG, in the GEO Portal standards registry, so that they are widely promoted and available.

The above considerations are probably still too limited. GEO BON needs to go well beyond the present GBIF data model to support monitoring data and a richer observational model with geographic features such as polylines, polygons and associated attributes.

At a minimum GEO BON will need a place to gather the products and their underlying data from all WGs. In order to demonstrate GEO BON utility, it would be useful to show the underlying chain (workflow) from data management through analysis to report. Using the main GEO Portal for this would be the preferred alternative. Specific scenarios of use for each WG will help to scope this. The GEO BON portal would not be a toolbox for analysis, at least in the beginning, but discovery of datasets could possibly be supported.

12. Open access issues

The GEOSS 10-Year Implementation Plan explicitly acknowledges the importance of data sharing in achieving the GEOSS vision and anticipated societal benefits. GEO membership requires agreeing that 1) There will be full and open exchange of data, metadata, and products shared within GEOSS, recognizing relevant international instruments and national policies and legislation, 2) All shared data, metadata, and products will be made available with minimum time delay and at minimum cost, 3) All shared data, metadata, and products being free of charge or available at no more than cost of reproduction will be encouraged for research and education.

In 2006, GEO established Task DA-06-01, “Furthering the Practical Application of the Agreed GEOSS Data Sharing Principles.” A white paper has been written by a team commissioned by

CODATA to review the current practices and issues in data sharing, and alternatives for implementation have been laid out.

The GEO BON Concept Document states that governments, organizations and institutions that sign up to GEO BON will need to acknowledge and promote the principles of open access to scientific and monitoring data, fair use of data for educational and research purposes, and the development of international Intellectual Property Rights (IPR) laws that protect the investments of private industry but that are not so restrictive that societal benefit from scientific research on biodiversity is stifled.

All the major networks that are candidates to form the GEO BON already adhere to similar principles as those mentioned above. Many of them belong to the Conservation Commons (www.conservationcommons.net/), a community of practice promoting open access.

However, there is evidence that lack of scientific credit for data sharing activities is still hampering actual implementation of the open access principles. If GEO BON will be able to put the focus on the development of workflows and value chains, it will probably have an opportunity to connect data providers and data users closer to each other, and in this way help data providers to get the recognition they need. Formal citation of datasets could provide a mechanism to quantify the use of a dataset. Several working groups are looking into potential mechanisms (e.g. SCOR/IODE working group on Data Publishing, http://www.iode.org/index.php?option=com_content&task=view&id=110&Itemid=129). Data papers such as those accepted by the Ecological Society of America might assist in making data more rapidly available by providing a formal "publication" mechanism for data. The GenBank model, where sequence data are made available at the time of publication, could be employed for other biodiversity data publication.

Another mechanism to quantify the use of publicly available data is to keep detailed statistics on how often data from specific datasets are downloaded from portals redistributing those data. These statistics can be used by the original data provider in his/her reports to funding agencies, to demonstrate the relevance of the work done and data collected.

In an ideal world, all data would be publicly available. In particular, if data are used to underpin management decisions, it is important that all concerned parties have access to the data on which the decisions were based. It is clear, however, that strong adherence to openness of data might restrict the volume and timeliness of available data. Will GEO BON accept any data, or only data with fully open access? Breaking away from this simple principle would make the system very

complex, and for this reason many data integrators (including GBIF, OBIS and CRIA) only accept data 'without strings attached'. The decision on whether or not to allow restricted data into the system goes beyond the mandate of WG8, but has clear implications for the development of the data and information infrastructure.

13. Activities to implement GEO BON

The next steps required to start the actual implementation of GEO BON are described in the detailed implementation plan (GEO BON 2010). The main areas for action centre around: establishment of a working group and coordinating unit; review of existing provider networks and establishment of partnerships; review of the data processing needs of the thematic working groups; design of the information architecture of GEO BON; building the components such as portal, registry, ontologies; registration of data and services; provision of a helpdesk, and outreach and capacity building.

14. References

[Berners-Lee 2001] Berners-Lee, T., Hendler, J. & Lassila, O. 2001. The semantic web. Scientific American, May 2001.

[GBIF 2009] GBIF 2009. Global Strategy and Action Plan for the Digitisation of Natural History Collections, 5 p. http://www2.gbif.org/GSAP_NHC.pdf

[GEO BON 2008] The GEO Biodiversity Observation Network Concept Document. http://earthobservations.org/documents/geo_v/20_GEO%20BON%20Concept%20Document.pdf

[GEO BON 2010] Group on Earth Observations Biodiversity Observation Network (GEO BON) Detailed Implementation Plan, Version 1.0 – 22 May 2010 http://earthobservations.org/documents/cop/bi_geobon/geobon_detailed_imp_plan.pdf

[GEOSS AIP2a] GEOSS AIP-2 Climate Change and Biodiversity WG, Arctic Food Chain, Use Scenario - Engineering Report.

[GEOSS AIP2b] GEOSS AIP-2 Climate Change and Biodiversity WG, The Impact of Climate Change on Pikas Regional Distribution, Use Scenario - Engineering Report.

[Hernandez-Ernst 2009] Hernandez-Ernst, V. & al. 2009. LifeWatch Reference Model, version 0.4, 230 p. Fraunhofer IAIS, Cardiff University. http://www.lifewatch.eu/index.php?option=com_content&view=article&id=69&Itemid=18

[IPCC 2007] IPCC, 2007: Summary for Policymakers. In: Climate Change 2007: The Physical Science Basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change [Solomon, S., D. Qin, M. Manning, Z. Chen, M. Marquis, K.B. Averyt, M.Tignor and H.L. Miller (eds.)]. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA.

[Khalsa 2009] Khalsa, S.J., Nativi, S., Geller, G., The GEOSS Interoperability Process Pilot Project (IP3), IEEE Transactions on Geoscience and Remote Sensing special issue on data archiving and distribution, vol. 47, num. 1. January 2009, pp. 80-91.

[Nativi 2007] Nativi, S., Mazzetti, P., Saarenmaa, H., Kerr, J., Kharouba, H. , Ó Tuama E. & Singh Khalsa, S. J. 2007. Predicting the Impact of Climate Change on Biodiversity – A GEOSS Scenario, *The Full Picture*, pp. 262-264; edited by the Group of earth Observation (GEO) secretariat, Tudor Rose, Leicester, UK.

[Nativi 2009] Nativi,S., Mazzetti, P., Saarenmaa, H., Kerr, J., Ó Tuama, É., "Biodiversity and Climate Change Use Scenarios framework for the GEOSS Interoperability Pilot Process", Ecological Informatics, Vol. 4 Issue. 1, January 2009, pp. 23-33.

[Nativi, 2010] Nativi,S., 2010,"The implementation of international geospatial standards for Earth and space sciences", Int. Journal of Digital Earth, Vol. 3, Supplement 1, 2010, pp. 2:13.

[Percivall 2010] Percivall, G., 2010, "The application of open standards to enhance the interoperability of geoscience information", Int. Journal of Digital Earth, Vol. 3, Supplement 1, 2010, pp. 14:30.

[Santana 2008] Santana, F.S., Siqueira, M.F., Saraiva, A.M., Correa, P.L.P. 2008. A reference business process for ecological niche modelling. Ecological Informatics 3(1): 75-86

[Saarenmaa 2005] Saarenmaa, H. 2005. Sharing and accessing biodiversity data globally through GBIF. 25th Annual ESRI International User Conference, San Diego, 25-29 July 2005. ArcUser Online January-March 2006. Environmental Systems Research Institute, Redlands, California. http://www.esri.com/news/arcuser/0206/biodiversity1of2.html

[Scholes 2009] Scholes, R.J. 2009. GEO BON - Group on Earth Observations Biodiversity Observing Network. Presentation 17 slides. IGOS-GEO Symposium, Washington DC, 19 November 2009. http://www.earthobservations.org/documents/cop/bi_geobon/200911_geobon.ppt

GROUP ON EARTH OBSERVATIONS | GEO BON

Annex 1: Data requirements template distributed to thematic work groups

Guidelines for thematic work groups from Work Group 8

To aid WG 8 in understanding data interoperability issues of the various thematic groups, we have set out some guidelines that may help you in organizing your data requirements.

You are requested to:

- Add your responses after each of the numbered points below and use the appended table to list resources where appropriate. Return to WG 8 (eotuama@gbif.org) by March 7th so that we can collate responses for the WG 8 section of the implementation plan.
- Appoint a technical liaison to WG8.

Information required:

1. What are the main existing data resources that are being used to support your WG theme? Please give us names of institutions, projects, people, specific information resources (URL), and types of data included (e.g. some major resources might be very focused-- e.g. biodiversity occurrence records, while others might be much broader-- monitoring data with lots of environmental context along with taxon info). The appended table can be used to list these resources.
2. Are there any major interoperability issues among these that you are aware of? Spatial or temporal resolution? Lack of critical variables? Difficulty in obtaining data records due to clumsy Web interfaces? Access restrictions?
3. What other resources are you aware of that are available that would assist your WG theme? E.g., are there lots of spreadsheets spread among researchers in your community that contain invaluable information?
4. What new types of information should be gathered to advance your WG theme? (Identify Data GAPS)
5. What are the main Data and metadata formats or standards used by researchers in your WG? Please provide formal names for these if known. Is there adequate metadata or other descriptive information to assist researchers in using these resources?

6. What are the computational approaches typical for researchers in your WG-- community models or individual models. Are people used to using community resources-- well-established data frameworks, or are these very fragmented-- individual analyses done on data contained within individual labs or people's PCs.
7. List the kinds of output products that would be generated by your network.

Data required by your network; also list data types that might not yet be available

	Data Type	Standard used (for data or metadata)	Organisation	Web Portal	Service Access protocol	Synergies (name other WGs that would also use these data)
1	Species occurrence	Darwin Core	OBIS	www.iobis.org	DiGIR	
2	Species monitoring	EML	ILTER	www.....	Metacat protocol	
3	Species occurrence	Darwin Core	GBIF	http://data.gbif.org	Several REST web services	
4						
5						
6						
7						

(examples highlighted in yellow)

Annex 2: Acronyms

ABCD: Access to Biological Collection Data

ADC: Architecture & Data Committee

AIP: Architecture Implementation Pilot

API: Application Programming Interface

BOLD: Barcode of Life Database

CEOS: Committee on Earth Observation Satellites

CoML: Census of Marine Life

CRIA: Centro de Referência em Informação Ambiental

CSGDM: Content Standard for Digital Geospatial Metadata

DIF: Data Interchange Format

DataOne: Data Observation Network for Earth

DiGIR: Distributed Generic Information Retrieval

EBONE: European Biodiversity Observation Network

EML: Ecological Metadata Language

ENM: Ecological Niche Modelling

EOL: Encyclopedia of Life

ESRI: Environmental Systems Research Institute

EU: European Union

FAO: Food and Agriculture Organisation

FGDC: Federal Geographic Data Committee

GAP: Gap Analysis Program

GBIF: Global Biodiversity Information Facility

GBRDS: Global Biodiversity Resources Discovery System

GCI: GEOSS Common Infrastructure

GCOS: Global Climate Observing System

GEOSS: Global Earth Observation System of Systems

GISIN: Global Invasive Species Information Network

GNA: Global Names Architecture

GNI: Global Names Index

GOOS: Global Ocean Observing System

GOS: Geospatial One Stop

GOSIC: Global Observing Systems Information Center

GTOS: Global Terrestrial Observing System

GUI: Graphical User Interface

HTML: HyperText Markup Language

IABIN: Inter-American Biodiversity Information Network

IDN: International Directory Network

ILTER: International Long Term Ecological Research

INBio: Instituto Nacional de la Biodiversidad

IODE: International Oceanographic Data and Information Exchange

IP3: Interoperability Pilot Process

IPANE: Invasive Plant Atlas of New England

IPT: Integrated Publishing Toolkit

ISO: International Organization for Standardization

ITIS: Integrated Taxonomic Information System

IUCN: International Union for Conservation of Nature

JaLTER: Japanese Long Term Ecological Research Network

KNB: Knowledge Network for Biocomplexity

LSID: Life Science Identifier

LTER: Long Term Ecological Research

MIENS: Minimum Information about an Environmental Sequence

NASA GCMD: NASA Global Change Master Directory

NBII: National Biological Information Infrastructure

NBN: National Biodiversity Network

NCD: Natural Collections Descriptions

NetCDF: Network Common Data Form

OAI-PMH: Open Archives Initiative – Protocol for Metadata Harvesting

OBIS: Ocean Biogeographic Information System

OGC: Open Geospatial Consortium

OGC CSW: Open Geospatial Consortium Catalog Services for the Web

OGC WCS: Open Geospatial Consortium Web Coverage Service

OGC WFS: Open Geospatial Consortium Web Feature Service

OGC WMS: Open Geospatial Consortium Web Map Service

ORNL: Oak Ridge National Laboratory

OWL: Web Ontology Language

PBIF: Pacific Biodiversity Information Forum

RDF: Resource Description Framework

REMIB: Red Mundial de Información sobre Biodiversidad (World Network for Biodiversity)

RSS: Really Simple Syndication

SBA: Societal Benefit Area

SCAR-MarBIN: Scientific Committee on Antarctic Research - Marine Biodiversity Information Network

SIBIS: South African National Biodiversity Institute's Integrated Biodiversity Information System

SIF: Standards & Interoperability Forum

SKOS: Simple Knowledge Organisation System

SOA: Service Oriented Architecture

SQL: Structured Query Language

TAPIR: TDWG Access Protocol for Information Retrieval

TDWG: Taxonomic Databases Working Group

TERN: Taiwan Ecological Research Network

UDDI: Universal Description Discovery and Integration

UNEP: United Nations Environment Programme

UNEP-WCMC: United Nations Environment Programme—World Conservation Monitoring Centre.

URI: Uniform Resource Identifier

WoRMS: World Register of Marine Species

ZooInt: Zoological Integrated System