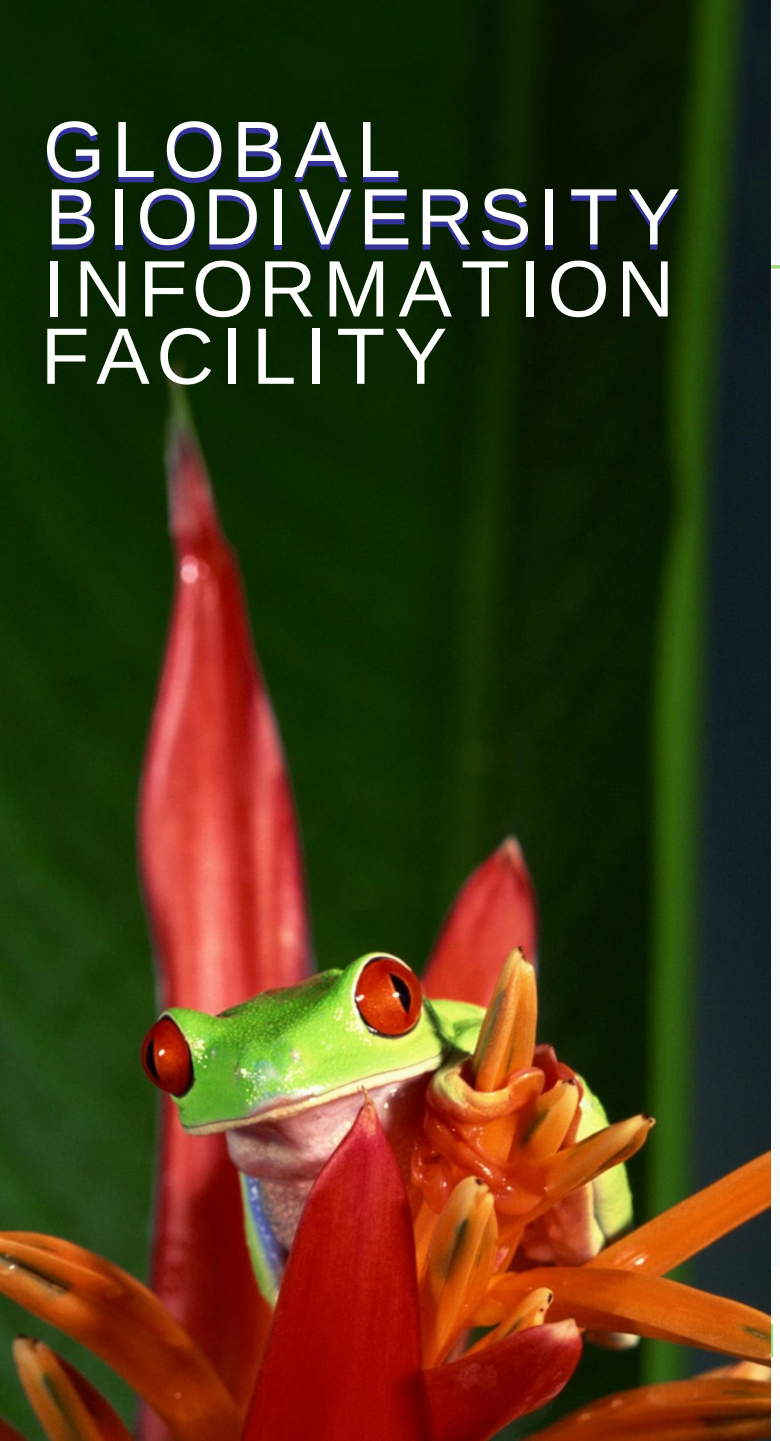


GLOBAL BIODIVERSITY INFORMATION FACILITY

Information Technologies and GBIF

Francisco Pando

COSTECH Training Event

Dar Es Salaam (Tanzania)
25-29 February 2008

WWW.GBIF.ES

Summary

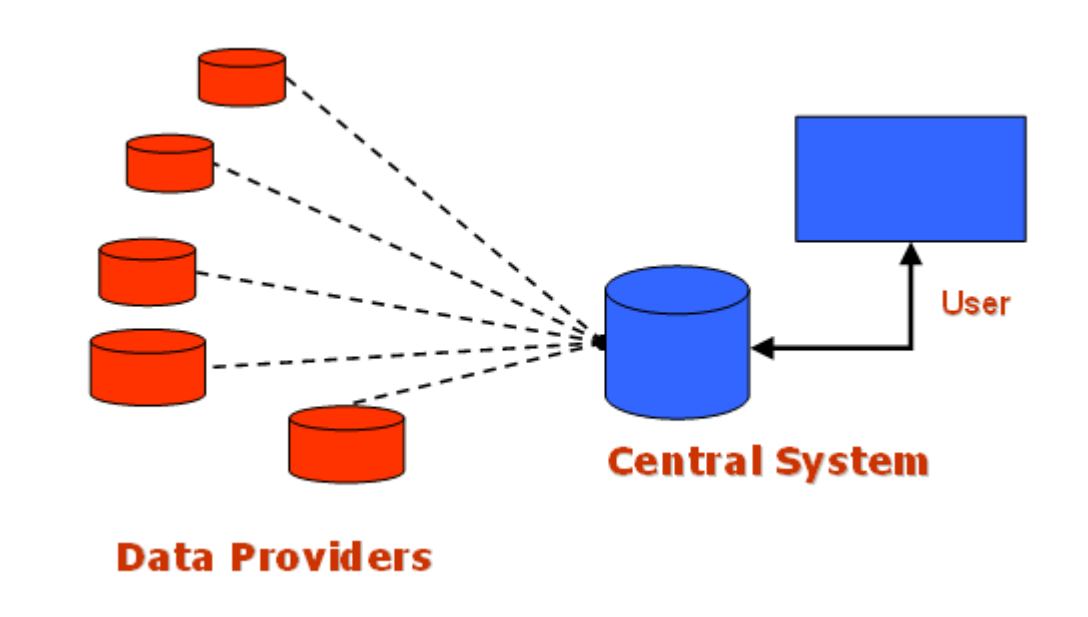
- GBIF objectives and vision
- Data network
- Data schemas
- Points and names
- Protocols
- Database registration
- Portals
- Strategies, options and implementations when sharing data

GBIF Objectives

GBIF has the objective of making all information about all known living organisms available on the Internet at a global level.

In other words, GBIF tries to change this old idea of “My data is mine, look at my results” (in science and management) into “Everybody’s data for everybody”

Data model: centralised networks



Distributed data network

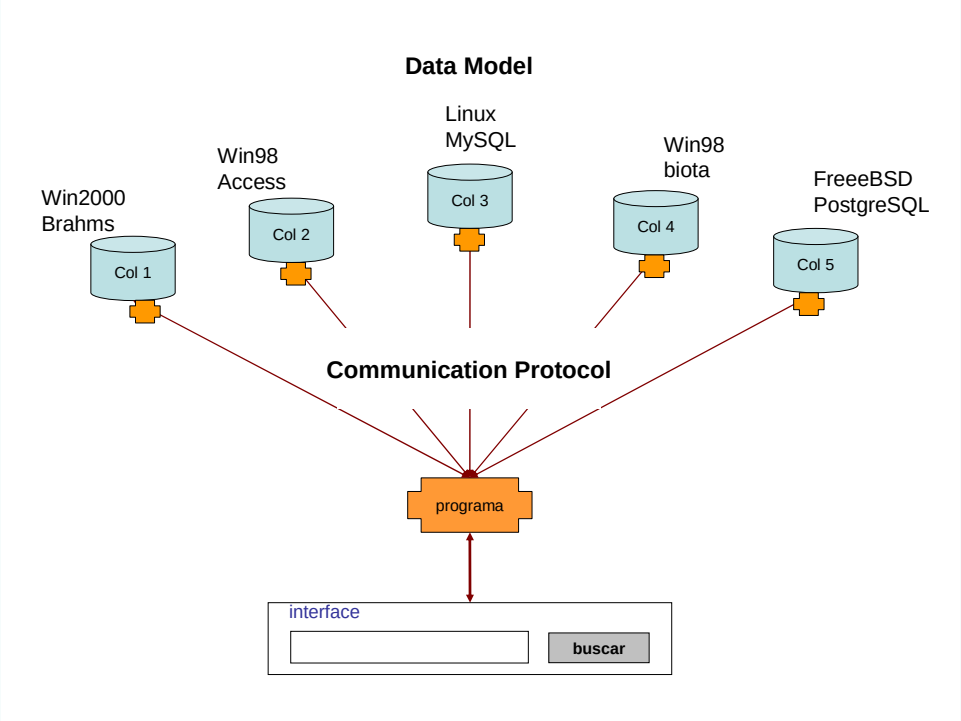


Figure 4. Diagram showing the complexity of integrating data from biological collections

Network elements

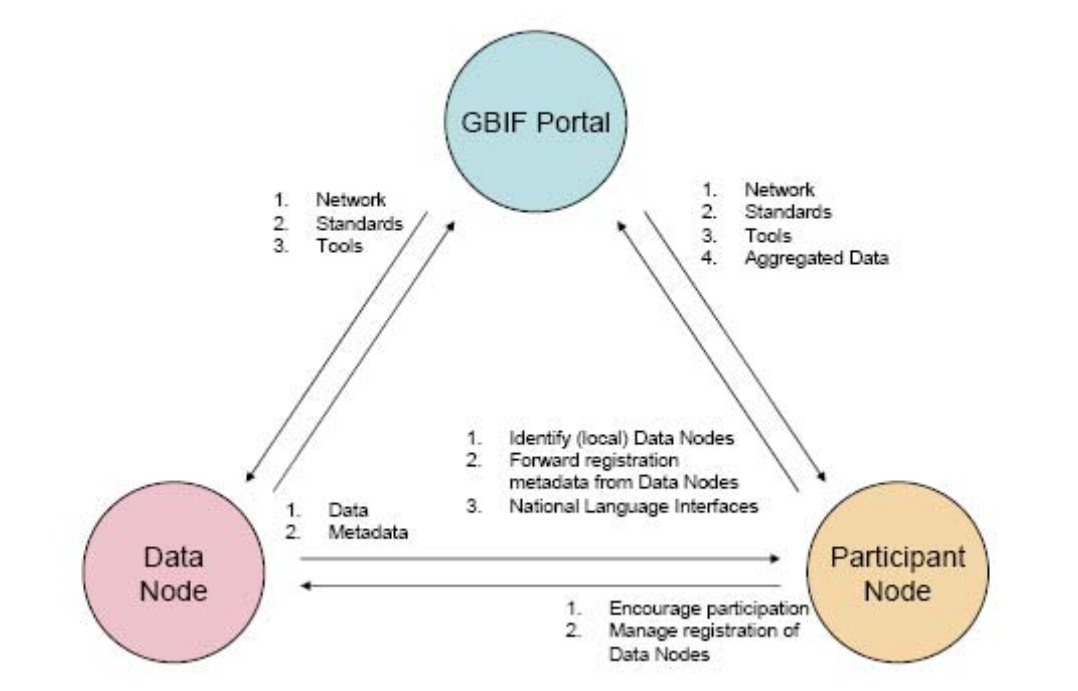
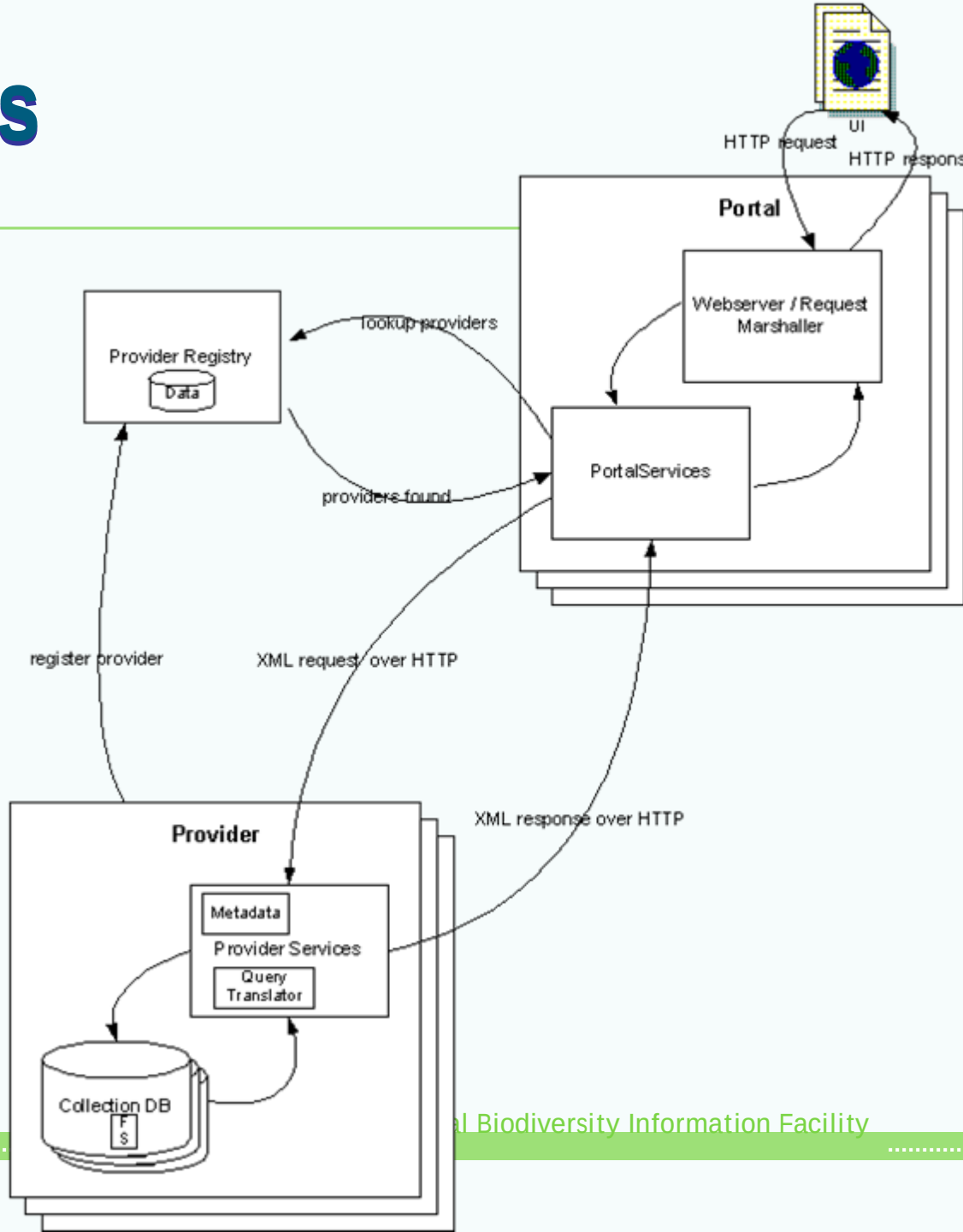


Figure 5. GBIF Network: major classes of nodes
GBIF is responsible for running the network, establishing standards, and developing tools. The portal is the hub for the development of any service that must be centralized such as the registry of metadata and for serving data from the biodiversity data index to the end user.

Further details

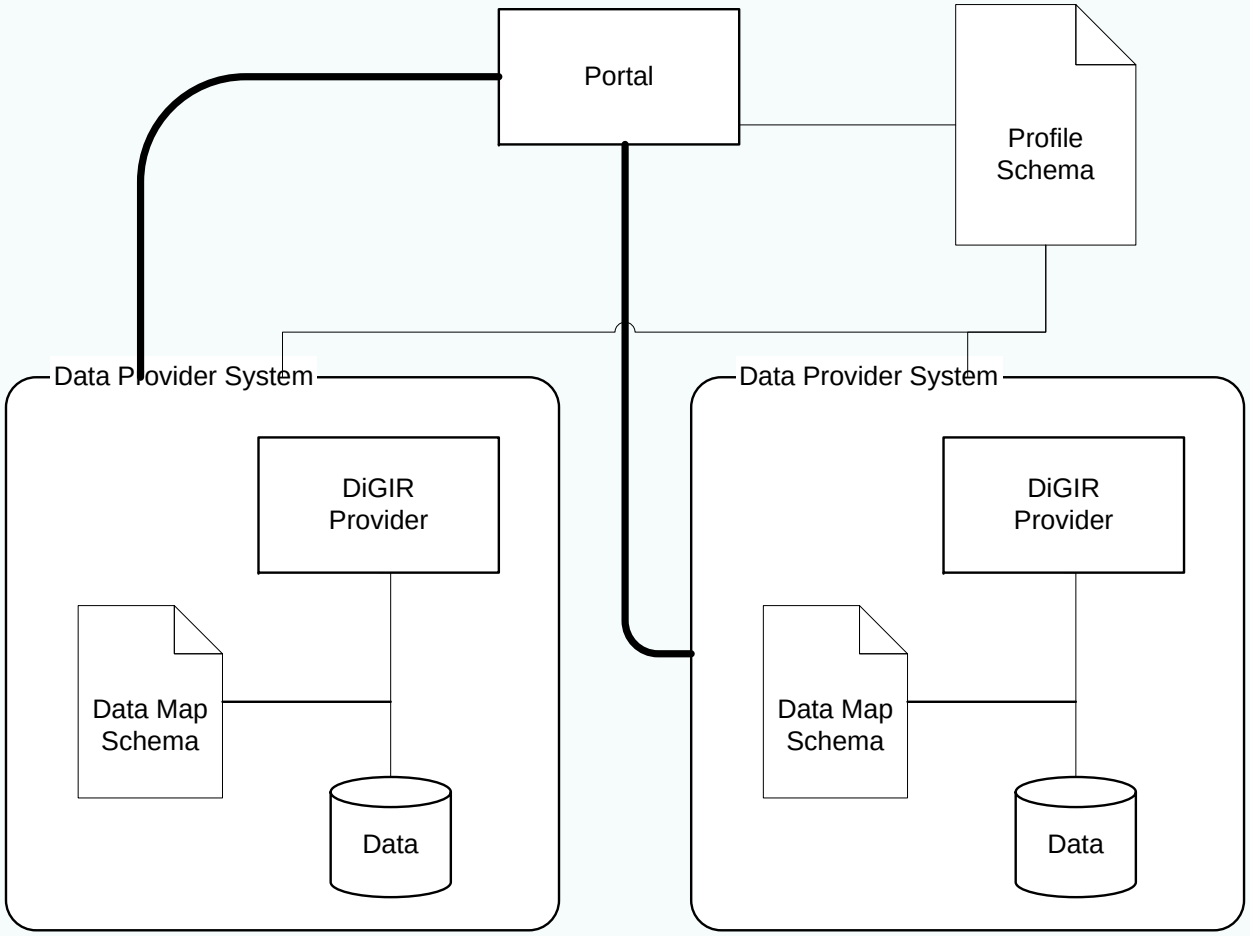
- Protocol
- Provider
- Portal
- Registry



Technologies used

- XML data exchange based on Providers, Services and Biodiversity Data Records
- UDDI registry for technical (access) metadata
- Descriptive metadata retrieved through service interfaces
- Specimen/observation exchange using DiGIR-Darwin Core or BioCASE-ABCD
- Taxonomic name data from Catalogue of Life (annual checklist for first release, moving to service-based approach as possible)
- Java (and JSP) components being developed centrally for GBIF Portal
- Current portal development using Tomcat, Xerces, Log4J, MySQL
- Components to be packaged for reuse as appropriate

Data mapping



Data schemas

- Darwin Core
  - Simple (50 elements)
  - The basic unit is the record
  - 500 databases in data.gbif.org
- ABCD
  - Elaborated and detailed content (+500 elements)
  - The basic unit is the file
  - 170 databases in data.gbif.org

The base for unified access:

- Common profile:

  Every database is translated into a "common profile", which is a table with a standardised field list that can be queried in a unified manner

- Standards:
  - "Darwin Core"
  - ABCD Schema
  - www.tdwg.org

The Species Analyst - Database Scan Results - Microsoft Internet E

File Edit View Favorites Tools Help

Address http://habanero.nhm.ukans.edu/TSA/scanDB.asp?tar Go

Attr#	Name	
1	ScientificName	Genus [+ " " + species [+ " " + subspecies]]. This
2	Kingdom	The kingdom to which the organism belongs.
3	Phylum	The phylum (or division) to which the organism
4	Class	The class name of the organism.
5	Order	The order name of the organism.
6	Family	The family name of the organism.
7	Genus	The genus name of the organism.
8	Species	The species name of the organism.
9	Subspecies	The subspecies name of the organism.
10	InstitutionCode	A unique identifier for you institution.
11	CollectionCode	Unique identifier for the collection within the ins
12	CatalogNumber	Unique identifier for the specimen record within
13	Collector	The name of the collector or collectors that were observation) from the field.
14	Year	The year (four digit) in which the specimen was
15	Month	The month of the year (1..12) in which the speci
16	Day	The day of month that the specimen was collec
17	Country	The country or major political unit (ocean) from
18	StateProvince	The state, province or region (i.e. next political r collected.
19	County	The county (or shire, or next political region sma
20	Locality	The locality description (place name plus option specimen was collected.
21	Longitude	The longitude in decimal degrees of the locatio
22	Latitude	The latitude in decimal degree of the location fr

Darwin Core, current developments

- Digital Image support

References Elements	
ImageURL	A Universal Resource Locator reference to digita associated with the specimen or observation.
RelatedInformation	Free text references to information not delivere conceptual schema, including URLs to specimen publications, etc.

- Extensions

Darwin Core, extensions

Geospatial Extension

Schema document and element definition table for the geospatial extension to the Darwin Core 2.

Curatorial Extension

Schema document and element definition table for the curatorial extension to the Darwin Core 2.

Paleontology Extension

Schema document and element definition table for the paleontological extension to the Darwin Core 2.

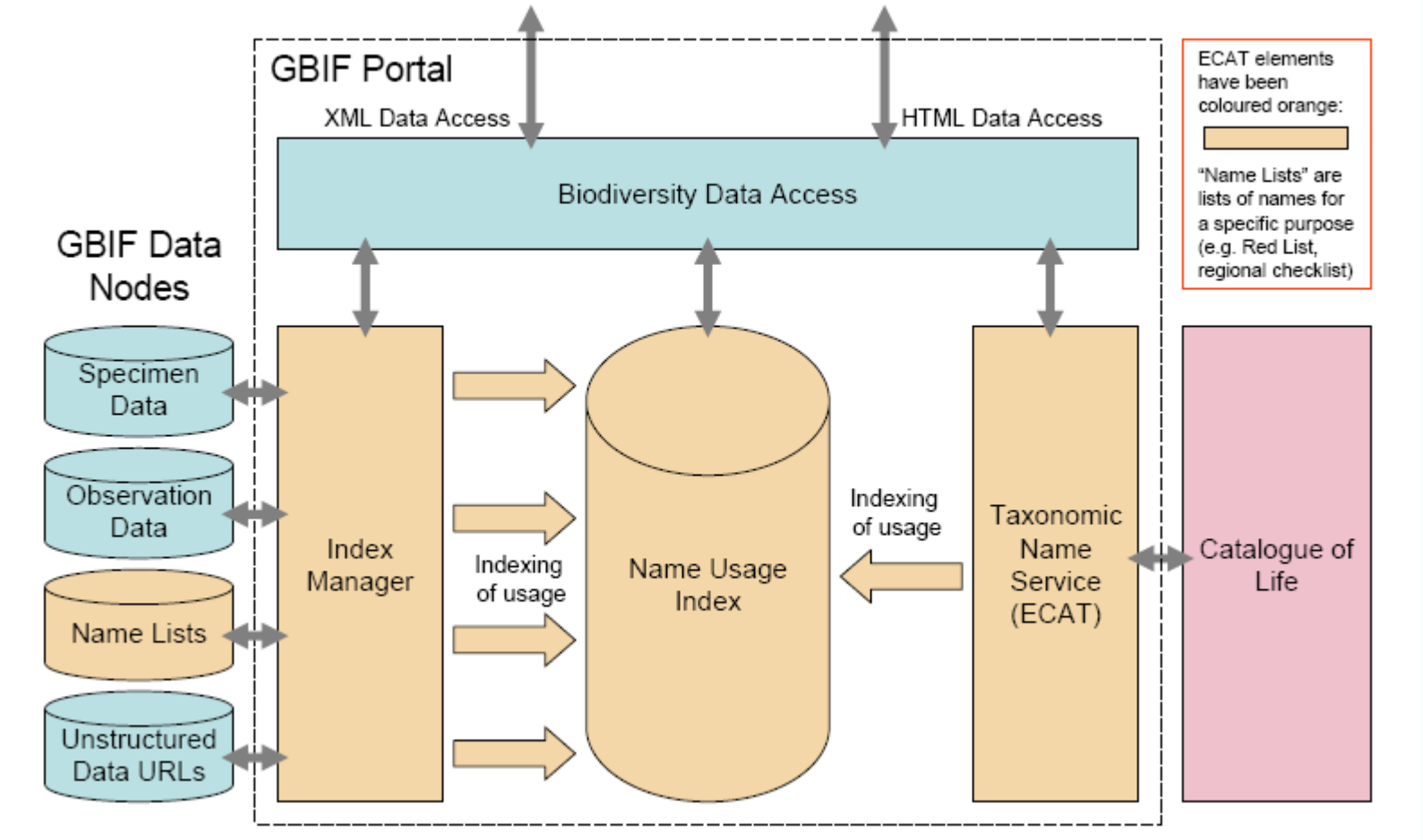
Microbial Extension

An extension to the Darwin Core 2 for microbiology culture collections. Posted by Renato de Giovanni. Integrated with the DiGIR protocol.

Observation/Monitoring Extension

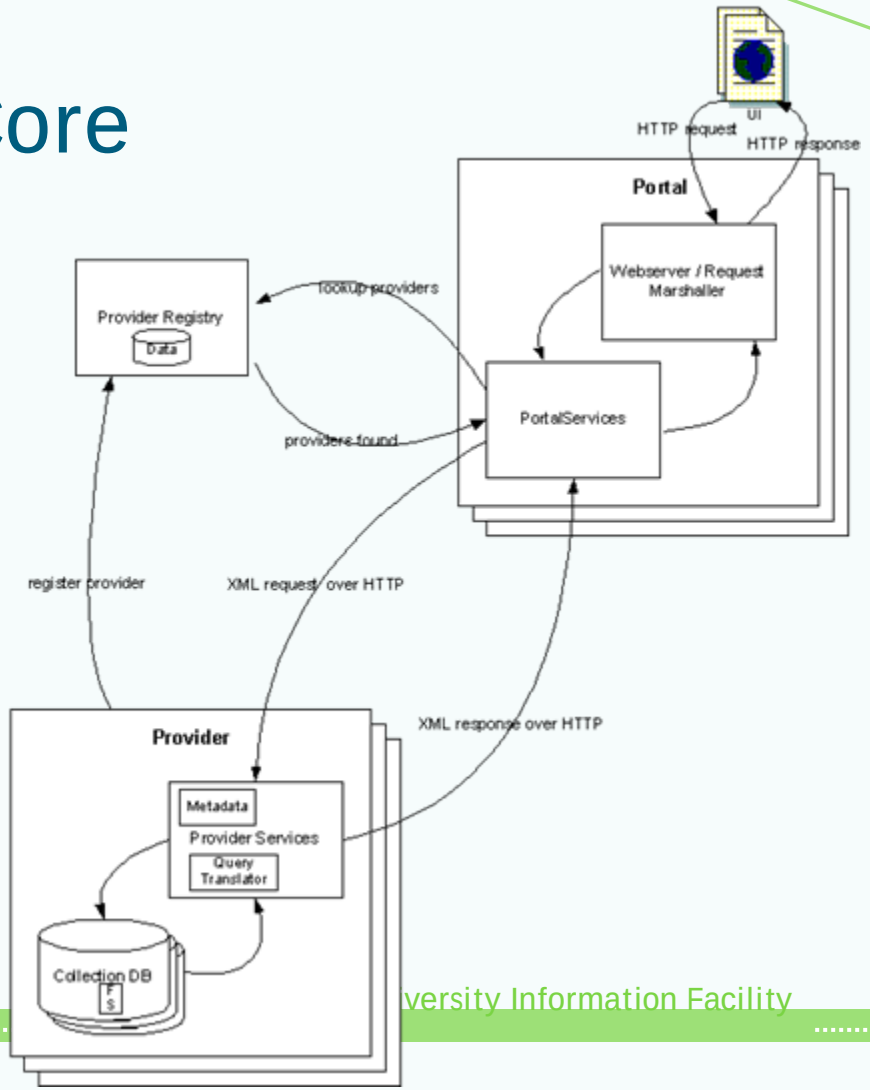
Link to the Avian Knowledge Network BMDE schema page which poses an extension to the Darwin Core 2 for observations and monitoring.

Names and specimens integration



Protocols

- DIGIR for Darwin Core
- Biocase for ABCD



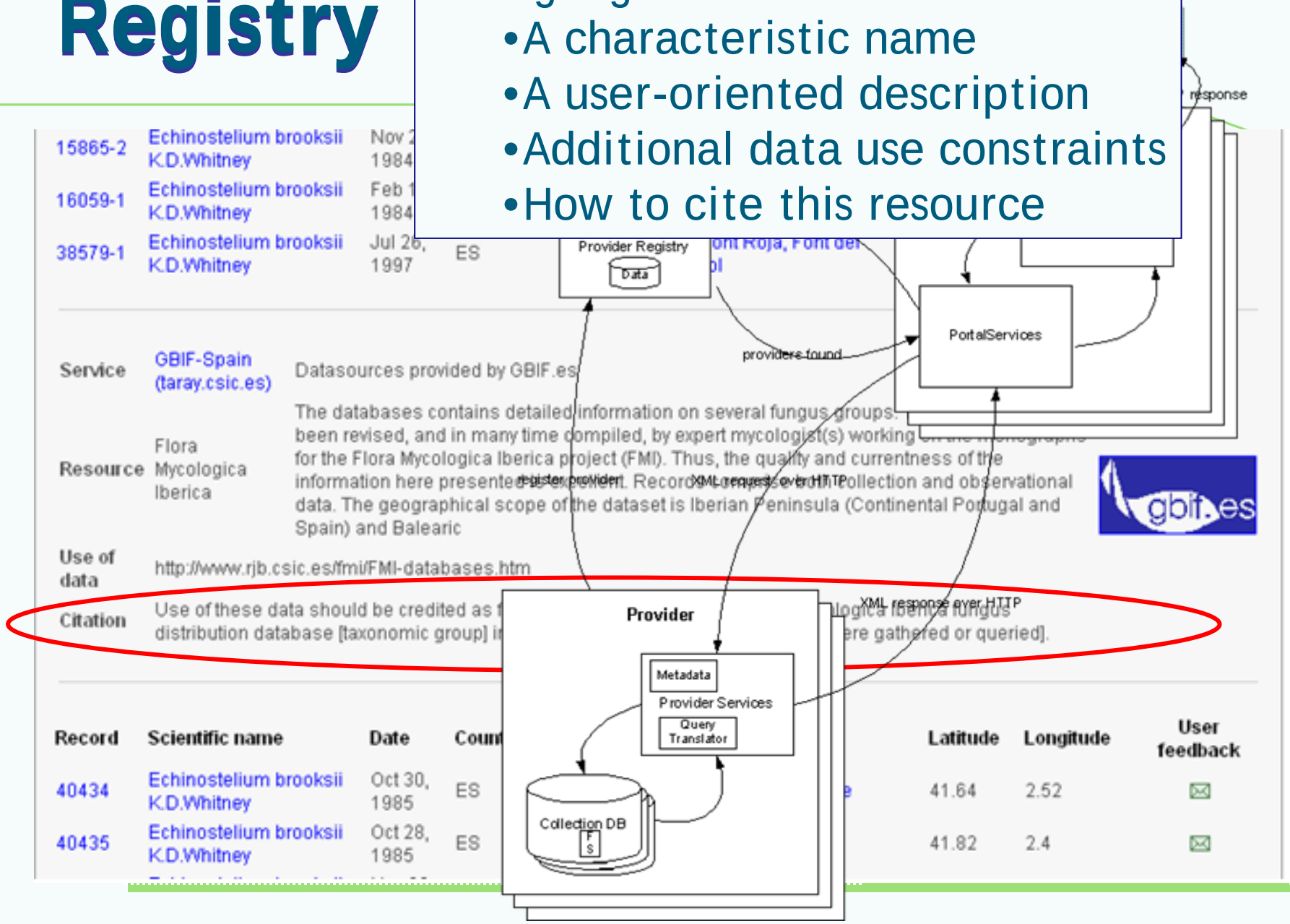
Life after DIGIR

- TAPIR
  - http://www.gbif.org/News/NEWS1129877273
  - http://ww3.bgbm.org/protocolwiki/
  - Unify protocols in GBIF data network
  - Registry expansion (UDDI), thematic networks support, national portals, Darwin Core extensions
  - Toolkit for data portals being developed

Registry

To highlight:

- A characteristic name
- A user-oriented description
- Additional data use constraints
- How to cite this resource



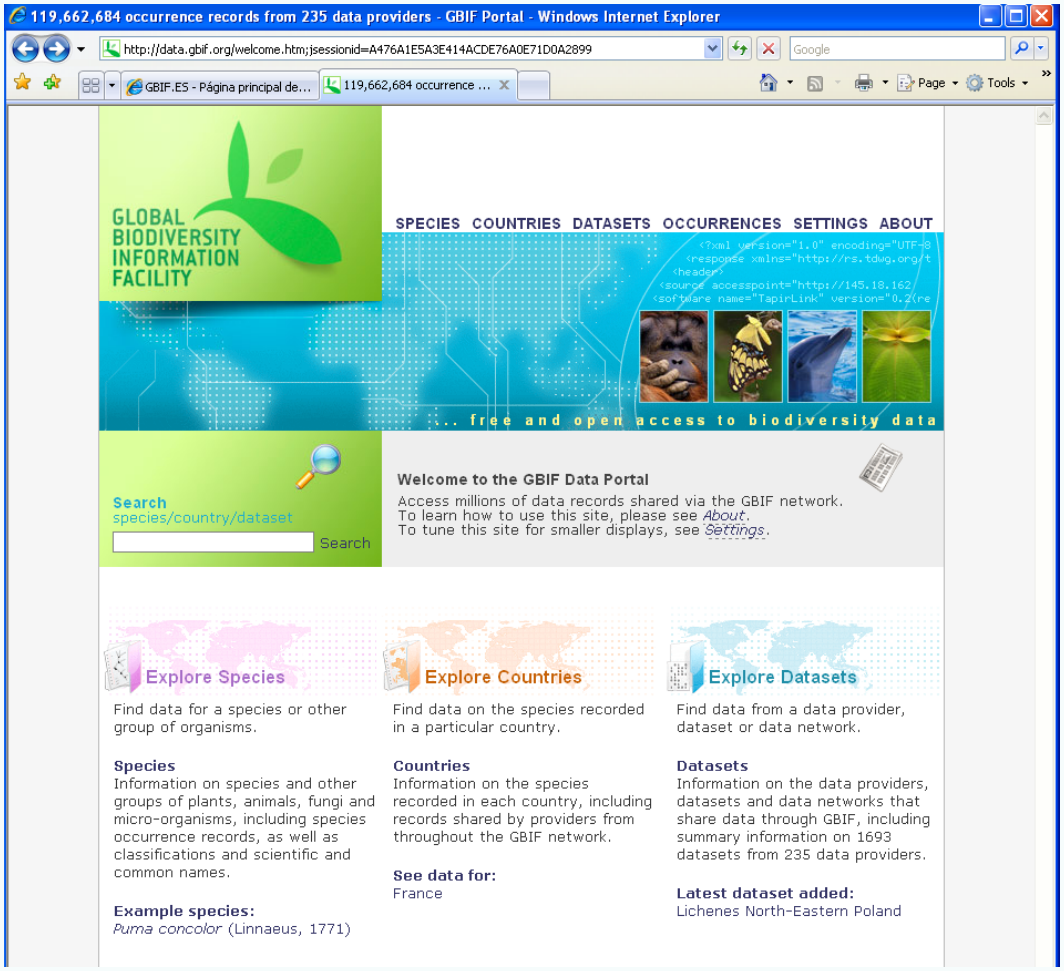
GBIF Data Portals

- New Data Portal - data.gbif.org
- Nodes Portal
- Old prototype - www.gbif.net
  (no longer available)

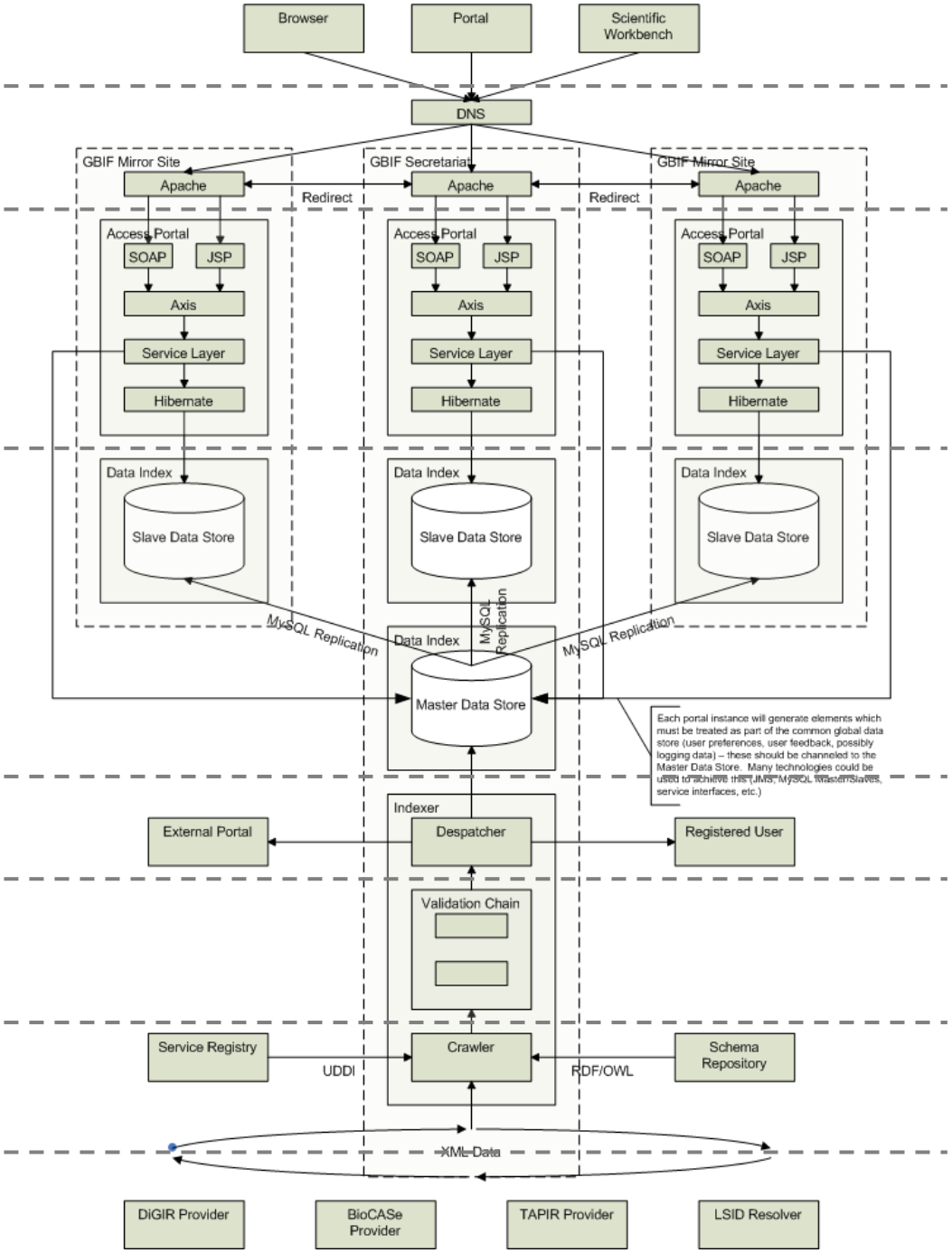
New Data Portal

- Similar to www.biologybrowser.com
- Web services
- API interface
- Added indexing and validation services

New Data Portal



Portal architecture (new version D. Hobern)



Clients

Mirrored access

Web applications

Synchronised data stores

Data despatcher

Interpretation and validation

Resource crawler

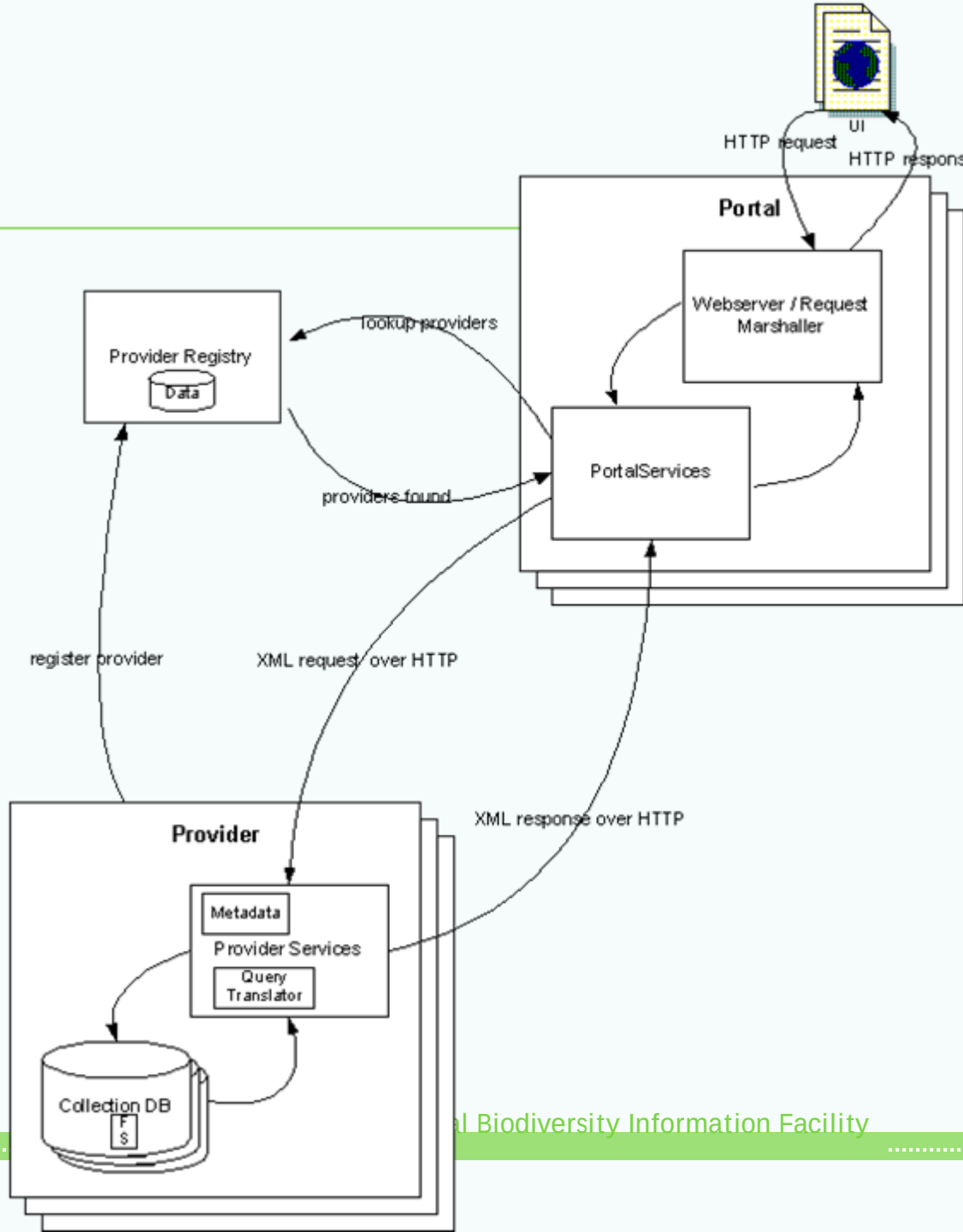
Data resources

Nodes Portal

- Under development
- Such as the nodes can provide:
  - Data from the collections in their area (country)
  - Data relevant under their scope
- Adaptable to their needs (language, common names,…)
- Help nodes to serve their communities

Clear enough?

- Protocol
- Provider
- Portal
- Registry



Old prototype

HOME | GBIF | BROWSE TAXONOMY | SEARCH | DATA PROVIDERS | COUNTRIES | DATA USE

Prototype data portal
Global Biodiversity Information Facility

BROWSE TAXONOMY
- Kingdom: Animalia
- Kingdom: Archaea
- Kingdom: Bacteria
- Kingdom: Chromista
- Kingdom: Fungi
- Kingdom: Plantae
- Kingdom: Protozoa
- Kingdom: Viruses

DATA PROVIDERS
- Taxonomic name
- Specimen/observation

COUNTRIES
- Data by country

GBIF REGISTRY
- Data providers 167
- Records 89656226
- Become a data provider
- Advanced search

ABOUT GBIF
- Press
- What is GBIF?
- Demonstration Projects
- GBIF's Current Members
- How to Join GBIF
- GBIF Directory
- Documents & History
- Ebbe Nielsen Prize
- Symposia & Workshops
- Work Programme

ABOUT GBIF

The Global Biodiversity Information Facility (GBIF) is an international non-profit organisation to provide free and universal access to data regarding the world's biodiversity. To learn more about GBIF itself, visit the GBIF Communications Portal: **www.gbif.org**

A wide range of countries and organisations participate in GBIF and have made their data available here. These bodies maintain ownership for all of the data they share. Any feedback provided through this web site will be passed back to the data provider concerned.

ABOUT THIS PORTAL

This portal is a prototype service providing access to the two types of data which are already being shared through the GBIF Network:

1. **Taxonomic names.** GBIF developing an 'Electronic Catalogue of Taxonomic Names'. This will provide access to authoritative information about both scientific and common names for all organisms, and will integrate data from a wide range of different organisations. The portal already includes data for over 983,000 scientific names and 253,000 common names from the Catalogue of Life Partnership Annual Checklist. Some names are listed with the words 'Tentative position in taxonomy'. This indicates that the name is only known to the portal from specimen/observation records and should not be treated as authoritative simply on the basis of being listed here.
2. **Specimens and Observations.** The GBIF Network already provides access to over 40 million records of occurrences of different organisms. Many of these relate to specimens in natural history museums and herbaria around the world, or to living cultures of micro-organisms, but at least a third come from observations of wild organisms. Wherever possible these records include information about the locality where the organisms were found and are used to generate maps of the distribution of these occurrences. Counts of occurrence records are listed against the organism names to which they apply. For genera and taxa of higher ranks, these counts include only those occurrences which have been identified to the taxon concerned. For species these counts include all occurrences for the species and also for any included infraspecific taxa, as well as for any known synonyms.

This is a work in progress. Please explore what is already present and visit us again in the coming months as we integrate more data and provide more flexible interfaces to search and browse the data. The following icons are used on many pages in the portal:

- Send feedback on a data item to the original data provider
- Get further details
- Download data

Search only by scientific or vernacular names (optional country filter)

USE OF DATA

GBIF Participants have made their data available for use according to the terms of the GBIF Data Use Agreement. Please understand these terms before using GBIF data: **GBIF Data Use Agreement**

SEARCH

Search for name: [ ] Submit Country/Territory: All countries

Selecting a country limits results to those taxa which specimen/observation records show have been identified from the country concerned.

- (•) Search by scientific name (any rank)
- ( ) Search by common name in any language
- ( ) Search by English name
- (•) Find names starting with search string (minimum 3 characters)
- ( ) Find names containing search string (minimum 3 characters)
- ( ) Find exact matches

Google Earth

You can search and visualise distributions using Google Earth link to GBIF data

Old prototype

Occurrence Detail

Georeferenced records

Countries from which species is recorded

Additional names from specimen/observation data (unreviewed)

Rank	Name	User feedback
Subspecies	*Balaenoptera musculus pribilofensis*	

Common names

Language	Name	Authority	Details	User feedback
English	**Blue whale**	Catalogue of Life: Integrated Taxonomic Information System		
French	**Rorqual bleu**			

Specimens/observations

Including records from: Antarctica; Australia; Canada; France; Mexico; Norway; Portugal; Spain; United Kingdom; United States

Service	Resource	All		Lat Long	
ABRS DiGIR Provider (www.deh.gov.au)	Australian Faunal Directory	3		0	
Australian Antarctic Data Centre (aadc-maps.aad.gov.au)	Seabirds of the Southern and South Indian Ocean	4		4	
Australian Antarctic Data Centre (aadc-maps.aad.gov.au)	Whale catches in the Southern Ocean	1968		1968	
Avian Knowledge Network (akn.ornith.cornell.edu)	Macaulay Library - Audio Data	34		12	
California Academy of Sciences (CAS) (www.calacademy.org)	Mammal Collection Catalog	8		0	
EUNIS 2 DiGIR Provider (woodpecker.eea.eu.int)	EUNIS	21		0	
Los Angeles County Museum of Natural History (LACM)	Vertebrate specimens	3		2	
MCZ-Harvard University Provider (digir.mcz.harvard.edu)	MCZ Mammalogy Collection	5		0	
Museum of Natural Science - Louisiana State University Mammal Collection (LSUMZ)	Mammal specimens	1		0	
Museum of Vertebrate Zoology (MVZ) (128.32.146.144)	Mammal specimens	1		0	
National Chemical Laboratory (digir.indobis.org)	IndOBIS, Indian Ocean Node of OBIS	1		1	
NatureServe (services.natureserve.org)	NatureServe Network Species Occurrence Data	1		0	
NLBIF (145.18.162.60)	Natural History Museum Rotterdam (NMR)	2		0	
OBIS/DIGIR Data Provider Server (www.iobis.org)	HMAP-History pf Marine Animal Populations	1137		1051	
OBIS/DIGIR Data Provider Server (www.iobis.org)	Seabirds of the Southern and South Indian Ocean	4		4	
OBIS/DIGIR Data Provider Server (www.iobis.org)	Bay of Fundy Species List	1		1	
OBIS/DIGIR Data Provider Server (www.iobis.org)	Canada Maritimes Regional Cetacean Sightings	171		171	
OZCAM Provider (digir.austmus.gov.au)	Online Zoological Collections of Australian Museums	9		9	
Royal Ontario Museum (digir.rom.on.ca)	Mammal specimens	4		0	
University of Washington Burke Museum (UWBM) (biology.burke.washington.edu)	Mammal specimens	7		1	
Total		**3385**		**3224**	

Contact info | Webmaster

Data resources with specimens or observations

Download records

Standards for data integration and interoperability

TDWG Taxonomic Databases WORKING GROUP

International Working Group on Taxonomic Databases

International Union of Biological Sciences Taxonomic Database Working Group

Select a topic:

- Home Page
- What's New
- Listservs
- Standards
- Subgroups
- Secretariat
- 2005 Officers
- Constitution
- Membership
- Related Links
- Comments
- History
- Archives

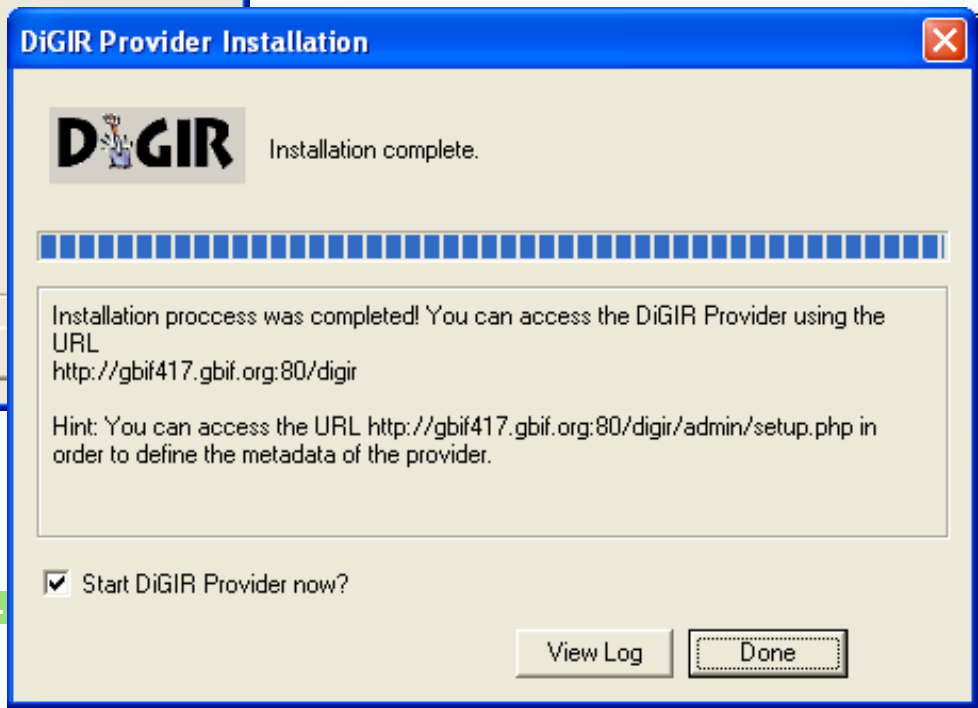
- 2004 Meeting
- 2005 Meeting

TDWG STANDARDS

http://www.tdwg.org/standrds.html

- Authors of plant names
- Botanico-periodicum-huntianum
- Botanico-periodicum-huntianum/supplementum
- Economic botany data collection standard
- Floristic regions of the world
- Herbarium information standards and protocols for interchange of data
- Index Herbarium, Part 1: The herbaria of the world
- International transfer format for botanic garden plant records
- Plant names in botanical databases

- ”Turn-key package”
- Based on PHP and code from DiGIR project
- For Linux and Windows
- Register with GBIF UDDI



- Technical support: helpdesk@gbif.org

About access to databases via Internet

- Whose data is this?
- All, Everything?
- How clean is it?
- How?

Whose data is this?

Address http://www.secretariat.gbif.net/portal/ecat_search.jsp?nextTask=ecat_search.jsp

Global Biodiversity Information Facility (GBIF)

Data Use Agreement

Background

The goals and principles of making biodiversity data openly and universally available have been defined in the Memorandum of Understanding on GBIF (MoU; see the relevant excerpts in Annex).

The Participants who have signed the MoU have expressed their willingness to make biodiversity data available through their nodes to foster scientific research development internationally and to support the public use of these data.

GBIF data sharing should take place within a framework of due attribution.

Therefore, using data available through the GBIF network requires agreeing with the following:

1. Data Use Agreements

1. The quality and completeness of data cannot be guaranteed. Users employ these data at their own risk.
2. Users shall respect restrictions of access to sensitive data.
3. In order to make attribution of use for owners of the data possible, the identifier of ownership of data must be retained with every data record.
4. Users must publicly acknowledge, in conjunction with the use of the data, the data providers whose biodiversity data they have used. Data providers may require additional attribution of specific collections within their institution.
5. Users must comply with additional terms and conditions of use set by the data provider. Where these exist they will be available through the metadata associated with the data.

2. Definitions

- GBIF Participant: Signatory of the GBIF-establishing Memorandum of Understanding (MoU).
- GBIF Secretariat: Legal entity empowered by the GBIF Participants to enter into contracts, execute the Work Programme, and maintain the central services for the GBIF network.
- GBIF network: The infrastructure consisting of the central services of the GBIF Secretariat, Participant Nodes and data providers. Making data available through GBIF network means registering and advertising the pertinent services via the GBIF central services..

All, everything?

- Data provider keeps control over what is made available:
  - Tests
  - Decides to dilute the record precision for endangered or economically relevant species
  - Does not publish data from current research

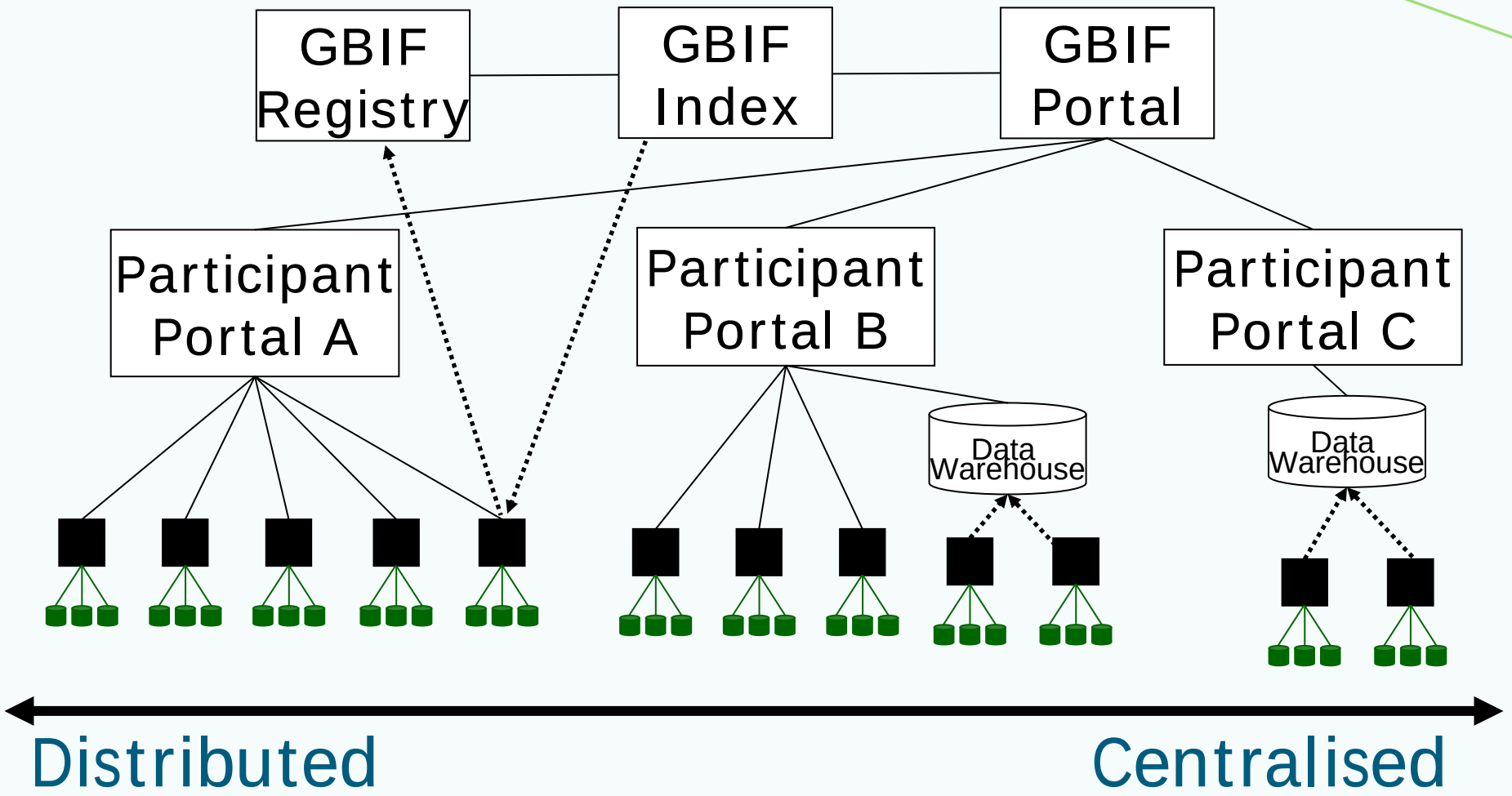
How clean is it?

- Perfection does not exist
- Neither bad data exist; a record is not good or bad intrinsically, its validity depends on its use
- Making data publicy available helps to improve them
- There are some tools to improve data and GBIF is working on this:

  http://www.secretariat.gbif.net/datatester/index.jsp

  http://www.gbif.org/prog/digit/data_quality

How?


GBIF Registry
GBIF Index
GBIF Portal
Participant Portal A
Participant Portal B
Participant Portal C
Data Warehouse
Data Warehouse
Distributed
Centralised

At your disposal:

Francisco Pando

Spanish GBIF Node

Royal Botanical Garden (CSIC)

Plaza de Murillo, 2

E-28014 Madrid, Spain

pando@gbif.es

Tél.: + 34 91 420 30 17