GBIF

INFORMATION
FACILITY

The GBIF Metadata Framework - Implementation

Éamonn Ó Tuama

PIIB Metadata Workshop,
Bogotá, 13-16 September, 2010

Outline

- GBIF architecture – main components

- Related standards

- ISO 19139 and Data Quality

- Metadata and languages

- Knowledge Organisation Systems

- Regional Networks – expectations?

Architecture: main components

A central metadata catalogue integrated in portal

A web based interface for searching/browsing

Many catalogues in network contributing metadata

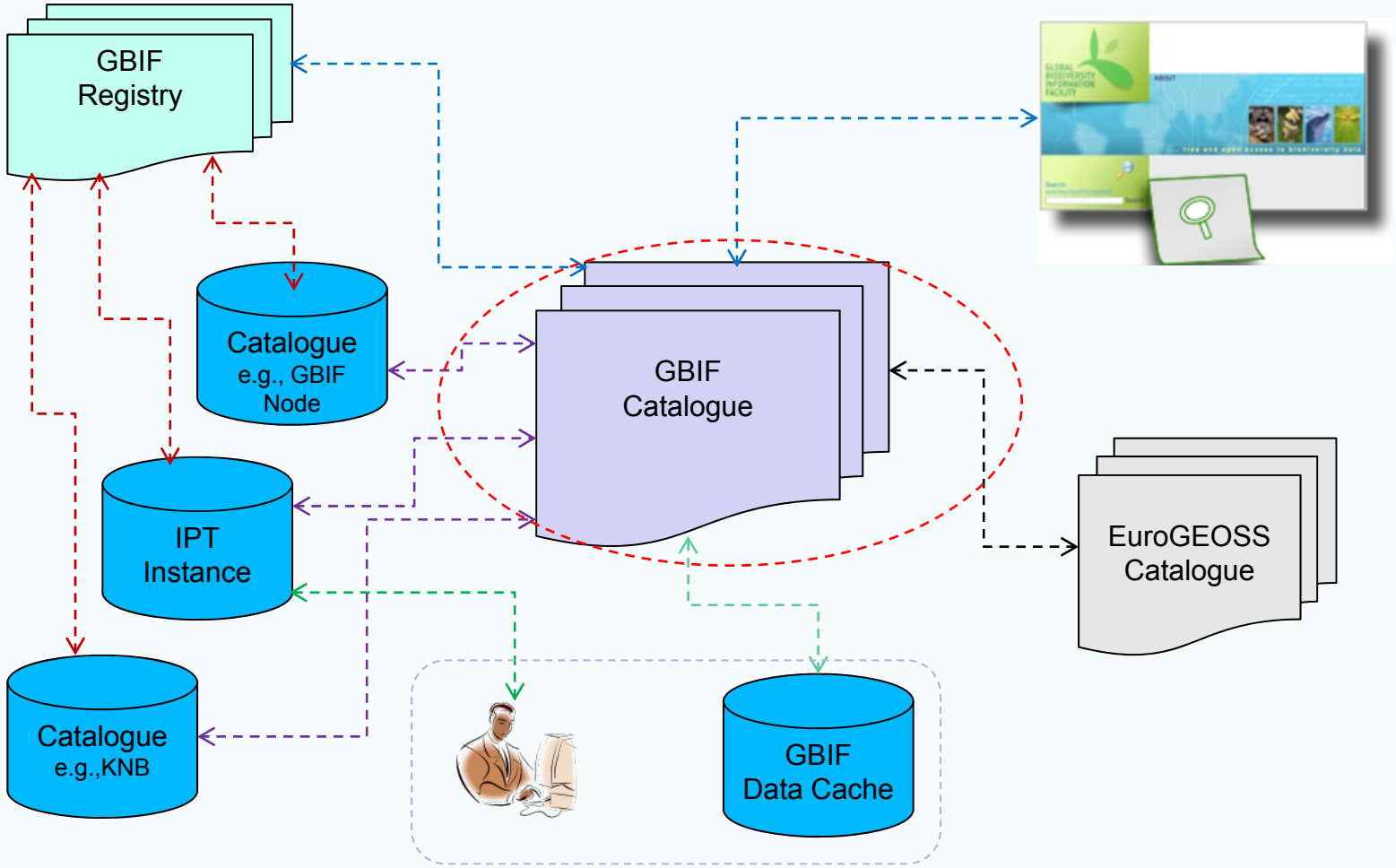
A registry to manage network entities

Standards based protocol for communications

Tools for editing metadata

Goal:
A gateway to enable discovery and re-use of data on the GBIF and other biodiversity networks

Architecture: main components



Current Catalogue Implementation

Based on Metacat version 1.9.2

Reviewed two other systems: GeoNetwork, Mercury

Cassia unavailable at time of review

Plan report on Metacat based on implementation experience

Exploring use of lightweight replacement (Solr/Lucene)

What kinds of Metadata?

Primary focus on metadata for datasets mediated via GBIF (collections, observations, names)

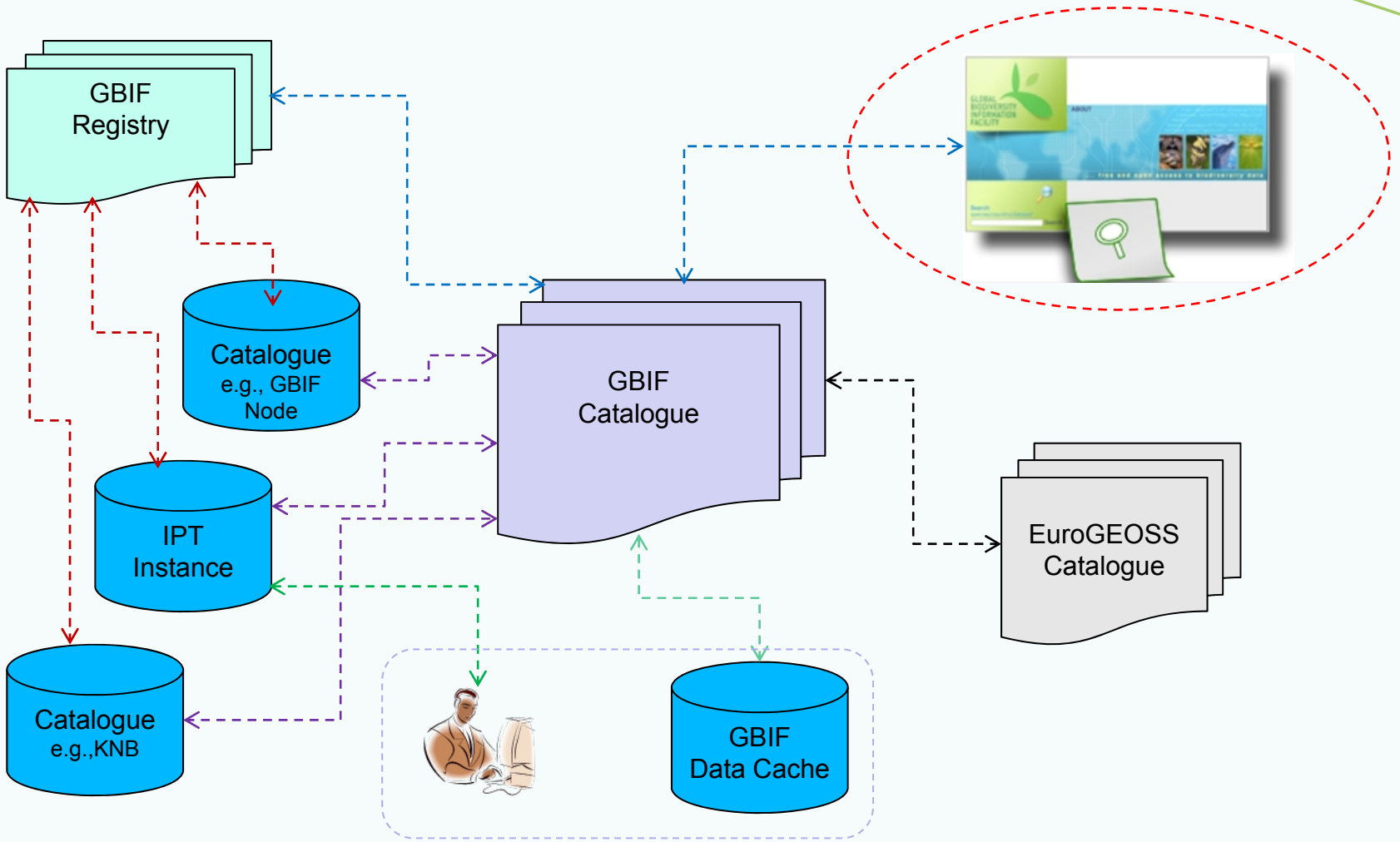
Metadata for inventory (prioritisation for digitisation)

Metadata for richer datatypes (e.g., associated ecological, environmental data)

- e.g., LTER and KNB networks with rich EML descriptions

Metadata for GBIF products (e.g., aggregated data, maps)?

Architecture: main components



Catalogue Search Interface

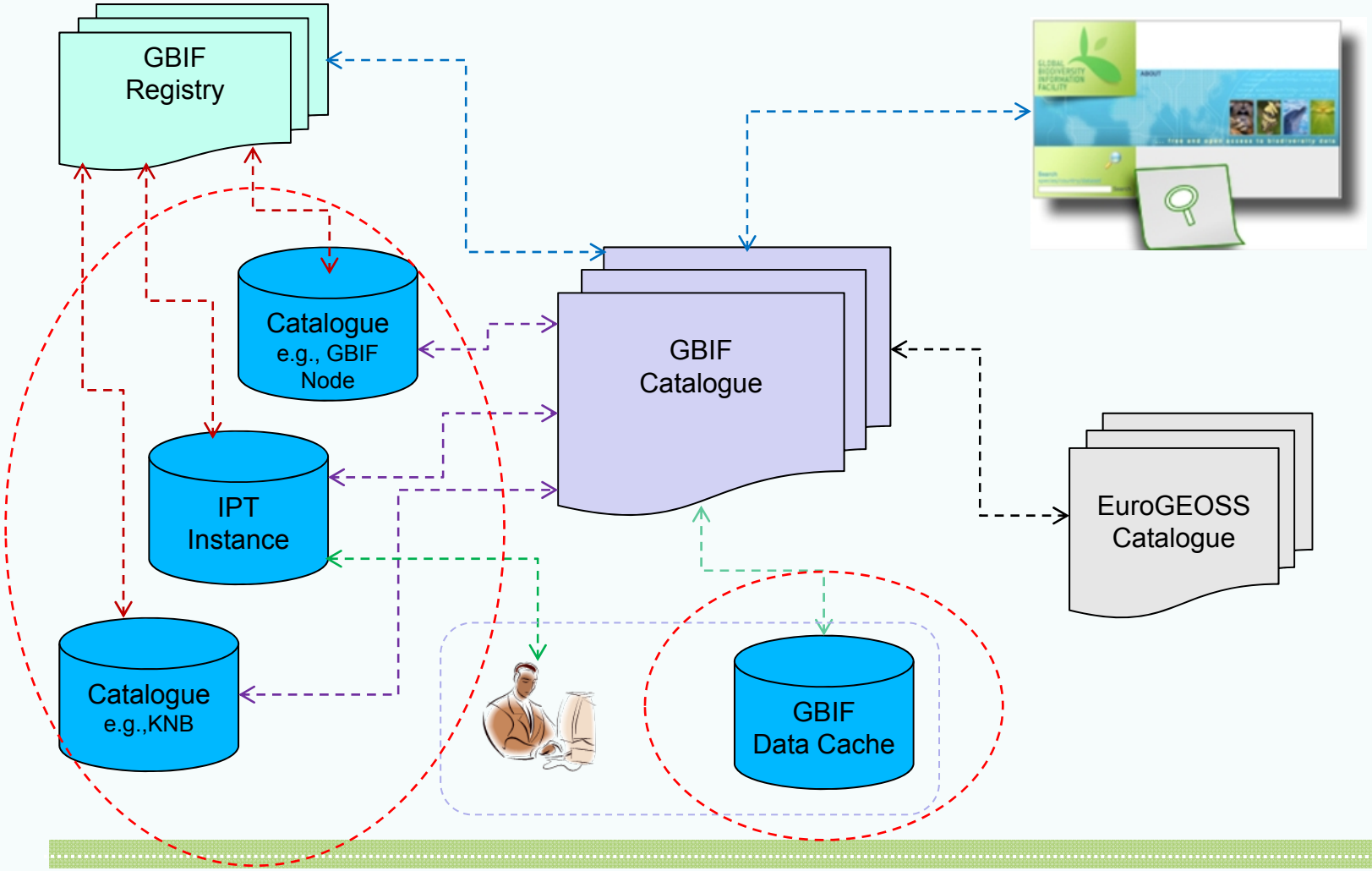
Simple and structured search; based on Spring framework (modification to Metacat)

Harvested metadata mapped to common search model (title; abstract; keywords; temporal, taxonomic, spatial coverages)

Returns full metadata as harvested

Links to corresponding datasets in GBIF portal

Architecture: main components



Sources of Metadata

GBIF Data Cache

GBIF Participants

- National/regional/organisation level catalogues
- Thematic catalogues, e.g., OBIS

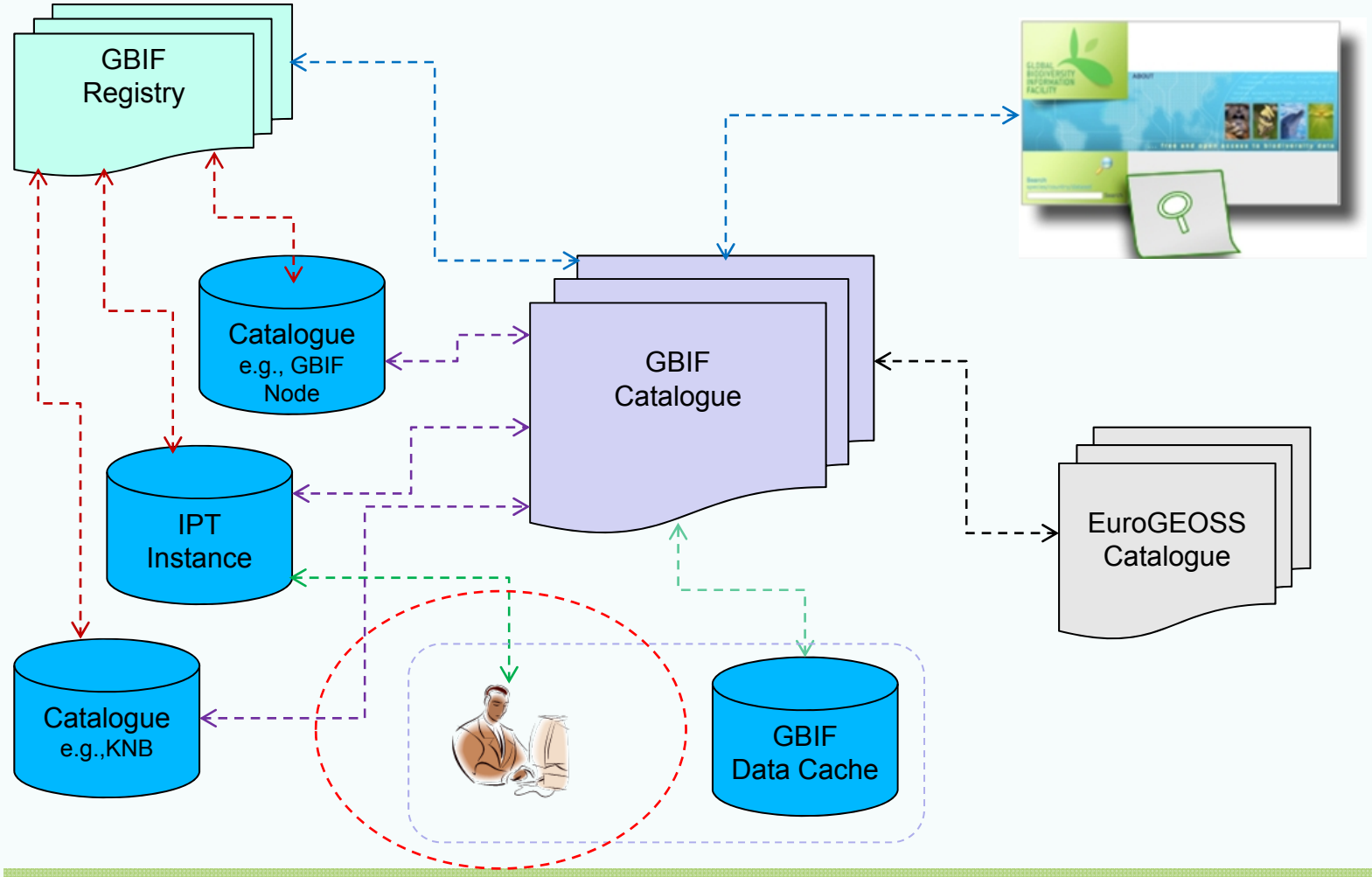
External networks

- e.g., Knowledge Network for Biocomplexity (KNB)

Our approach:

- *no imposed metadata standard or preferred catalogue implementation for participants*
- *avoidance of lossy conversions in submitting metadata*

Architecture: main components



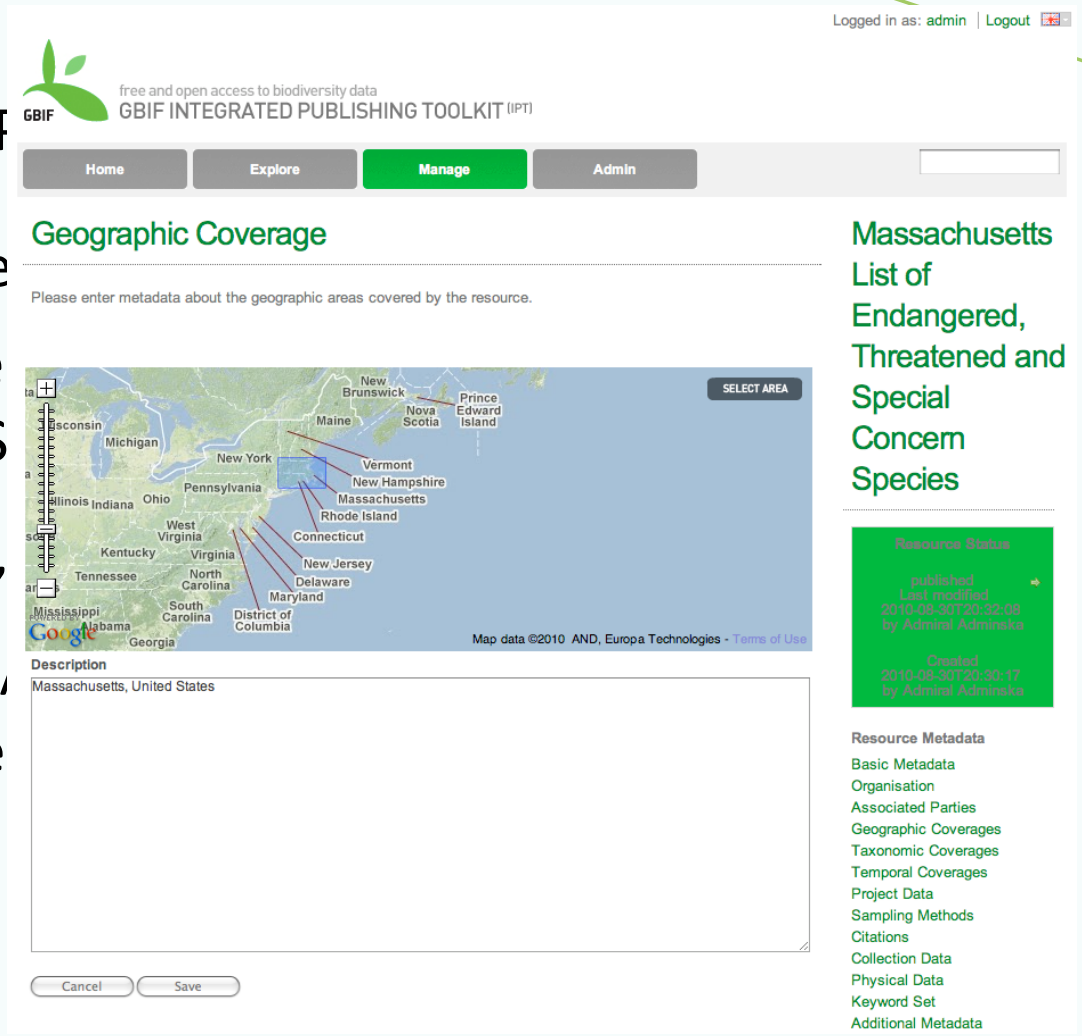
Metadata Preparation

Integrated Publishing Toolkit (IP

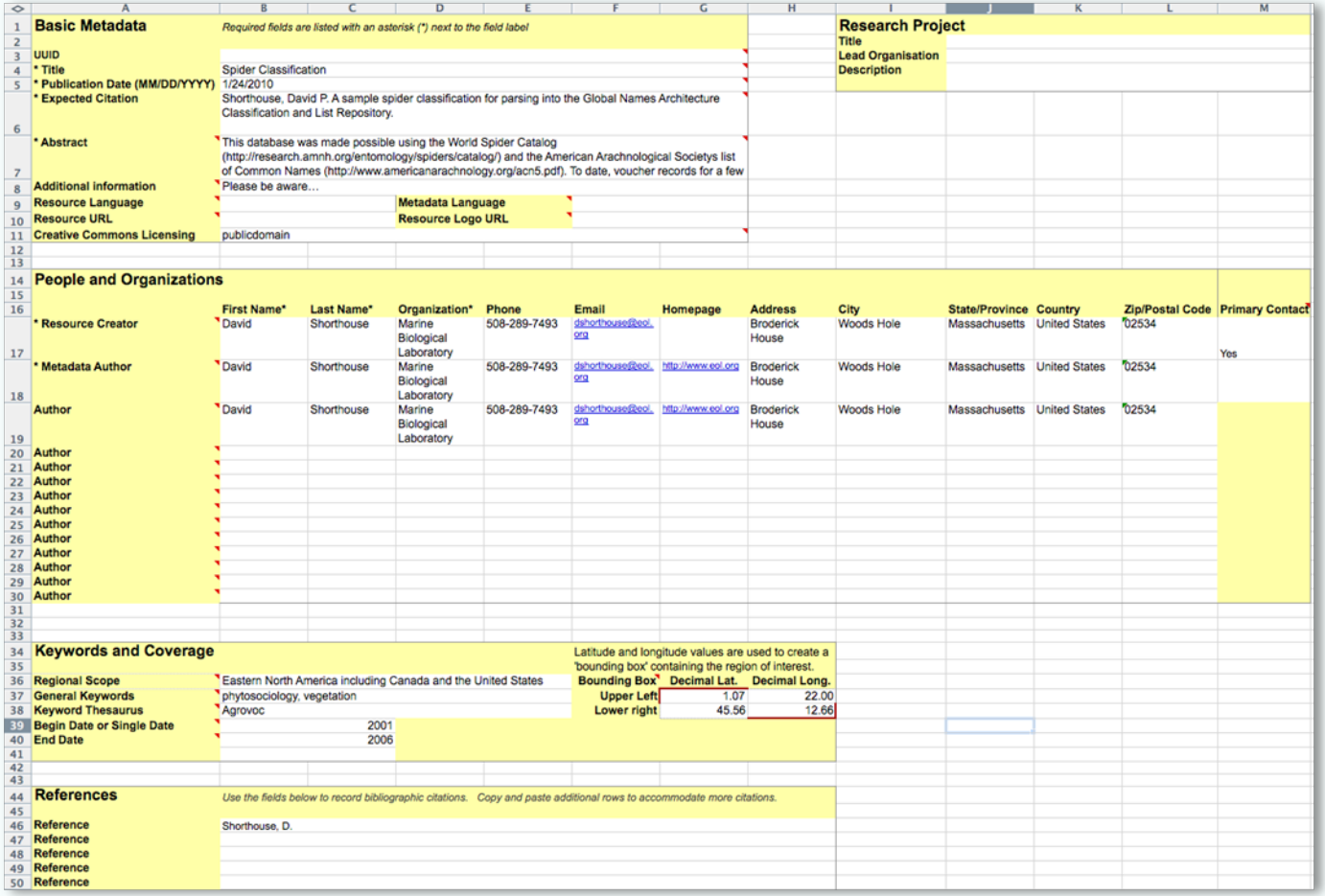
Includes integrated metadata e

- metadata profile base
  extensions for NCD, IS
- document collections,
- dataset (Darwin Core A
  submitted as package

IPT still under development



Darwin Core Archive Assistant Tool



Basic Metadata — *Required fields are listed with an asterisk (*) next to the field label*

UUID	
* Title	Spider Classification
* Publication Date (MM/DD/YYYY)	1/24/2010
* Expected Citation	Shorthouse, David P. A sample spider classification for parsing into the Global Names Architecture Classification and List Repository.
* Abstract	This database was made possible using the World Spider Catalog (http://research.amnh.org/entomology/spiders/catalog/) and the American Arachnological Societys list of Common Names (http://www.americanarachnology.org/acn5.pdf). To date, voucher records for a few
Additional information	Please be aware…
Resource Language	
Resource URL	
Creative Commons Licensing	publicdomain

Metadata Language	
Resource Logo URL	

Research Project

Title	
Lead Organisation	
Description	

People and Organizations

	First Name*	Last Name*	Organization*	Phone	Email	Homepage	Address	City	State/Province	Country	Zip/Postal Code	Primary Contact*
* Resource Creator	David	Shorthouse	Marine Biological Laboratory	508-289-7493	dshorthouse@eol.org		Broderick House	Woods Hole	Massachusetts	United States	02534	Yes
* Metadata Author	David	Shorthouse	Marine Biological Laboratory	508-289-7493	dshorthouse@eol.org	http://www.eol.org	Broderick House	Woods Hole	Massachusetts	United States	02534	
Author	David	Shorthouse	Marine Biological Laboratory	508-289-7493	dshorthouse@eol.org	http://www.eol.org	Broderick House	Woods Hole	Massachusetts	United States	02534	
Author												
Author												
Author												
Author												
Author												
Author												
Author												
Author												
Author												
Author												

Keywords and Coverage

Regional Scope	Eastern North America including Canada and the United States
General Keywords	phytosociology, vegetation
Keyword Thesaurus	Agrovoc
Begin Date or Single Date	2001
End Date	2006

Latitude and longitude values are used to create a 'bounding box' containing the region of interest.

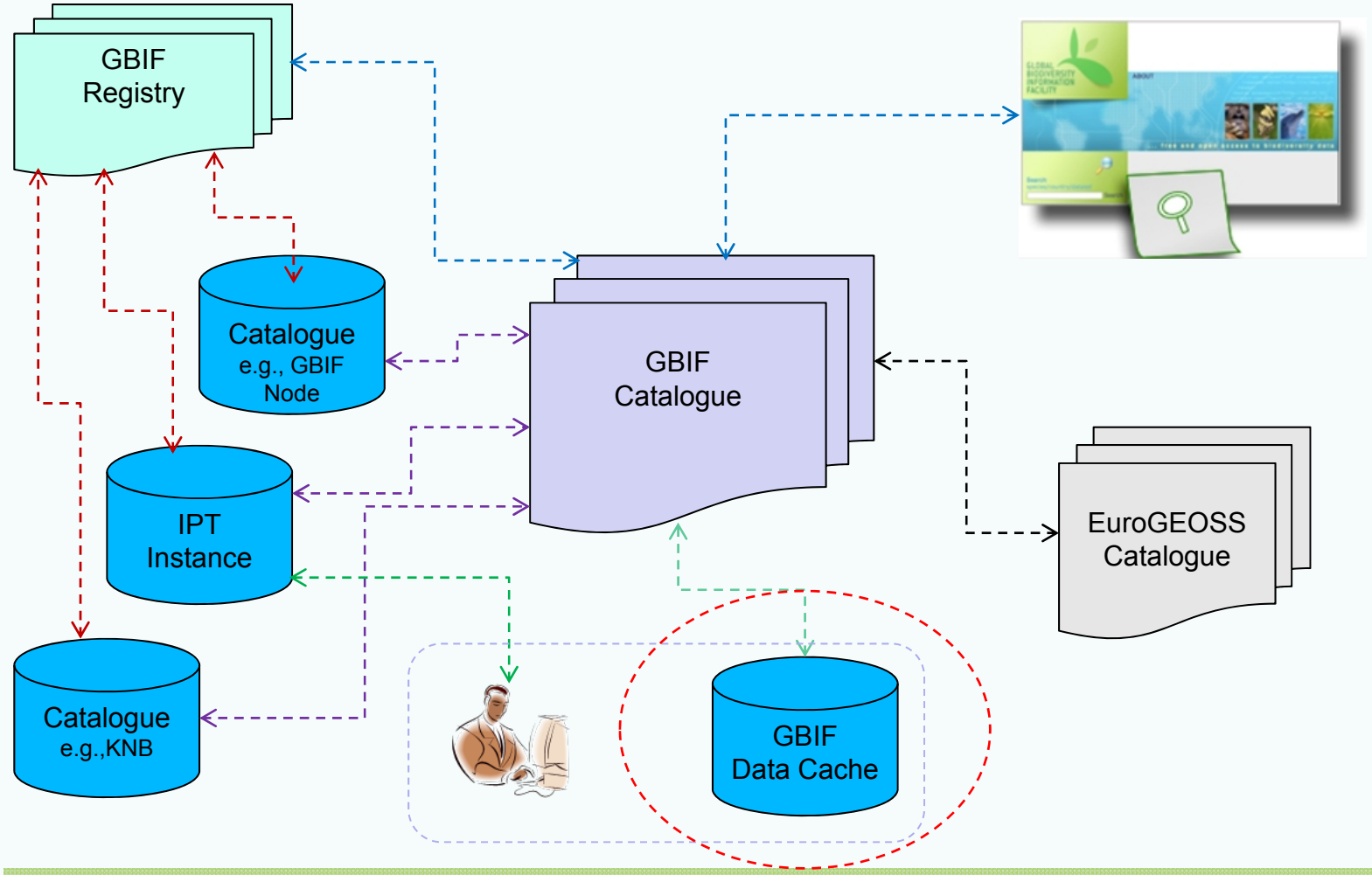
Bounding Box	Decimal Lat.	Decimal Long.
Upper Left	1.07	22.00
Lower right	45.56	12.66

References — *Use the fields below to record bibliographic citations. Copy and paste additional rows to accommodate more citations.*

Reference	Shorthouse, D.
Reference	
Reference	
Reference	
Reference	

- outputs dataset and metadata as package (DwCA)
- EML 2.1.0
- submit via email or web application
- due in Q4 2010

Architecture: main components



Generating dataset metadata

via the UDDI registry, DiGIR, TAPIR BioCASe

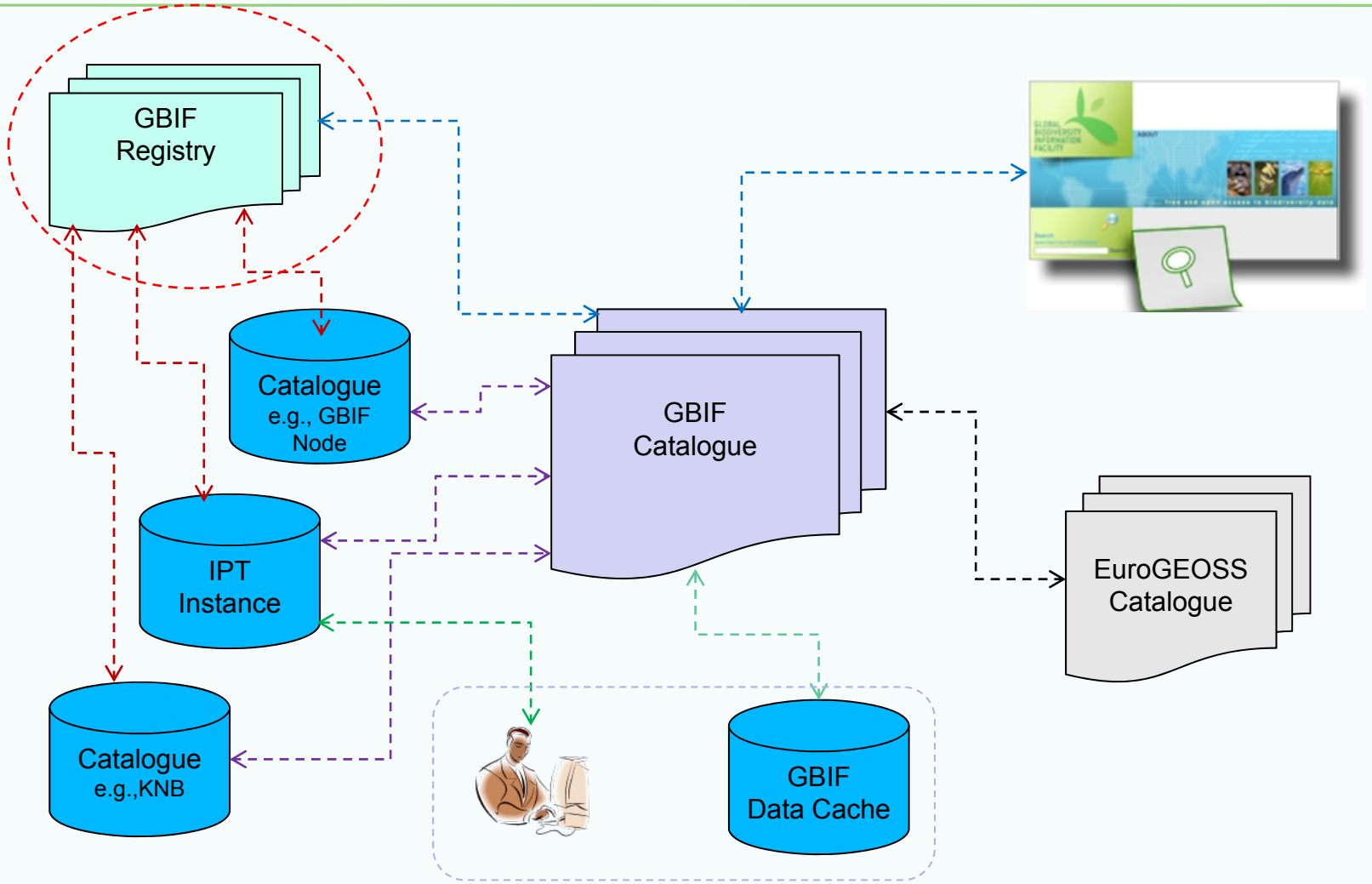
- Data provider details (Name; Website; GBIF participant; Description; Country; Added to portal; Information updated)
- Provider (DiGIR, BioCASe, TAPIR) binding
- Name
- Website
- Description
- Citation
- How to cite this dataset
- Basis of record
- Access point URL
- Added to portal
- Information updated
- Contacts (Name, Role, Address, Email, Telephone)
- Data networks

An EML metadata document is generated for each dataset in the GBIF data cache

and via the indexing process -

- Taxonomic coverage
- Temporal coverage
- Spatial coverage

Architecture: main components



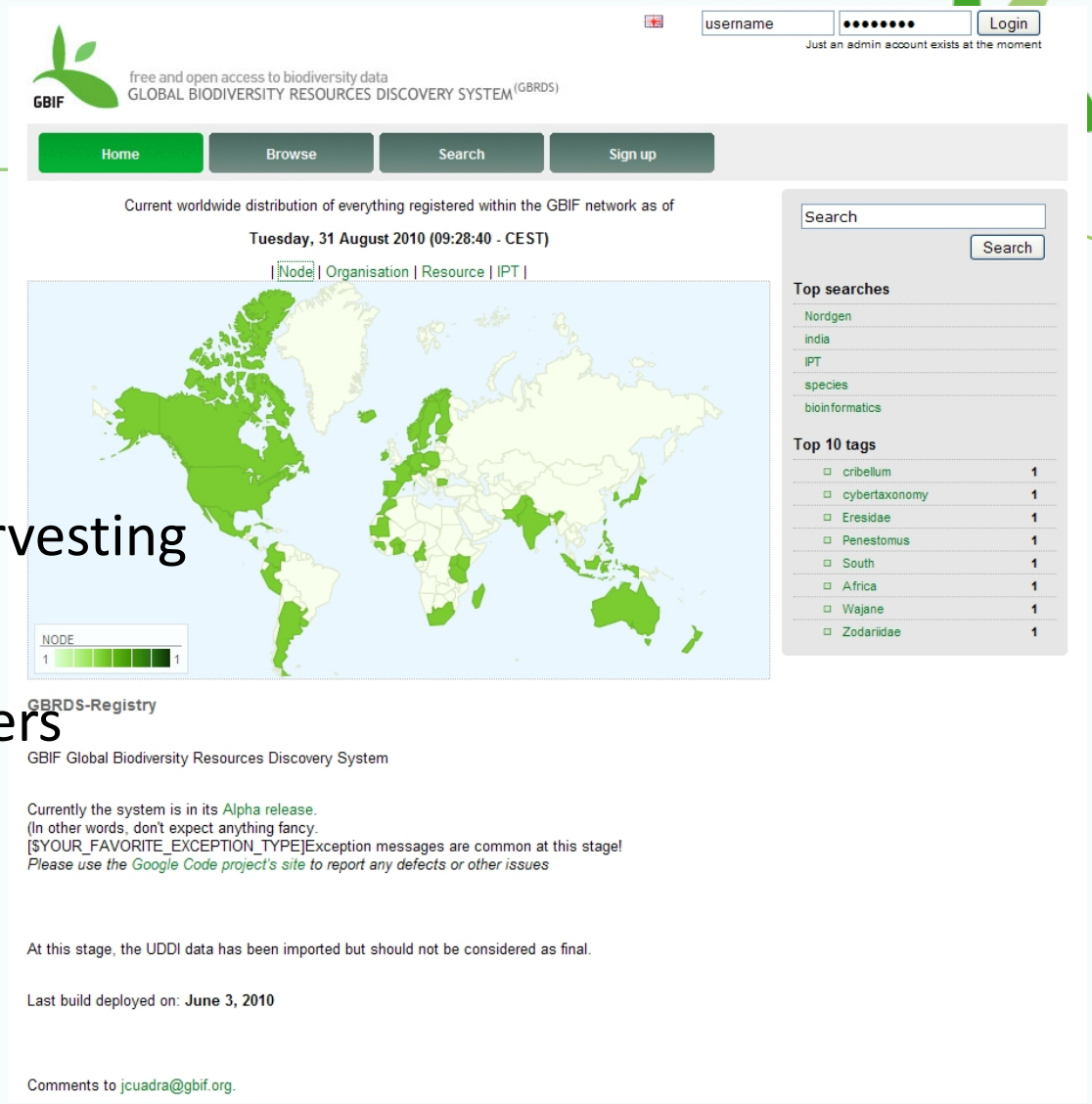
The GBIF Registry

Extension of current UDDI

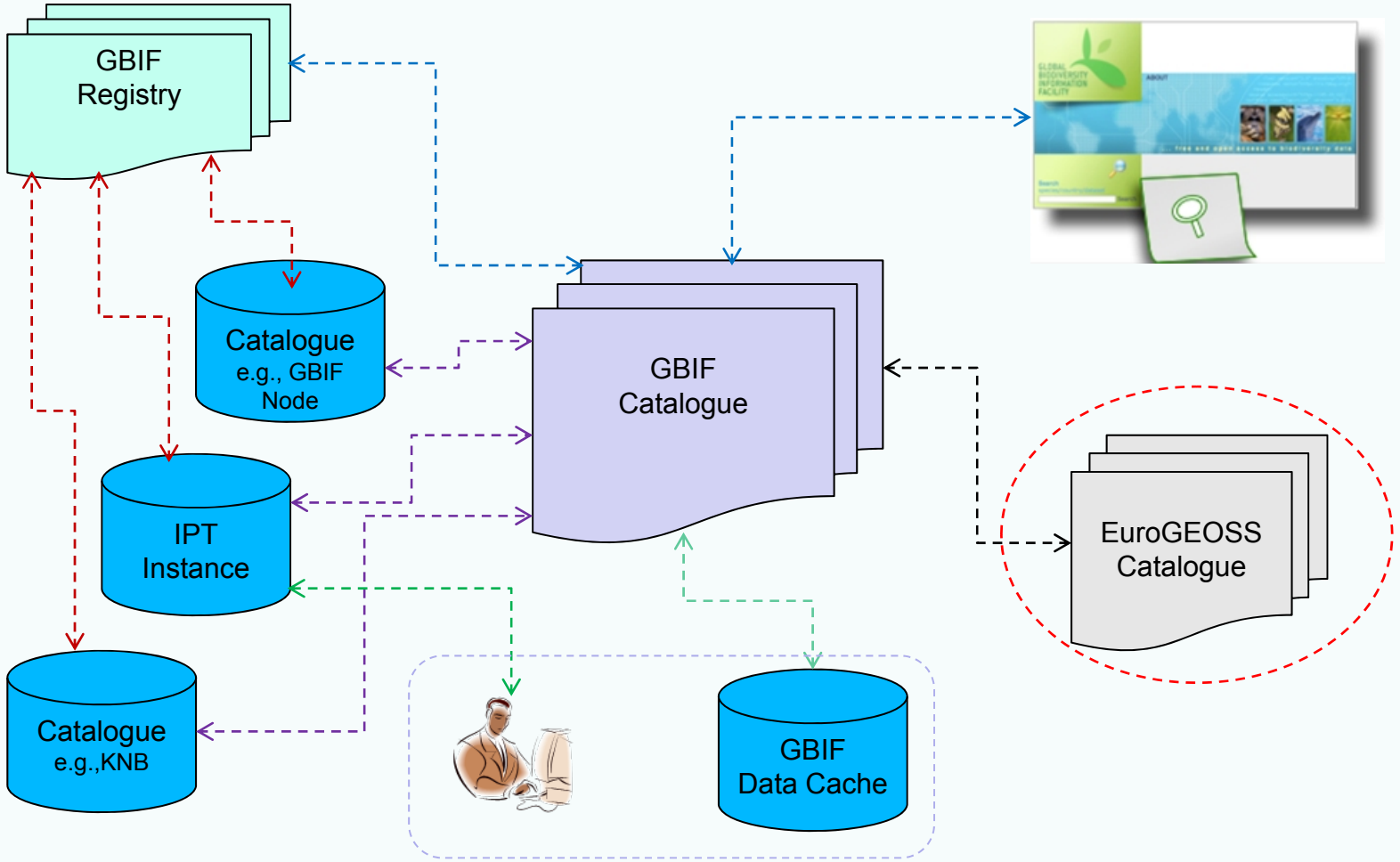
Several functions

- Stores end-point URLs for harvesting
- Rich model of GBIF network
- Reconciles persistent identifiers

Under development



Architecture: main components



EuroGEOSS Broker

	CSW	WMS	WFS	Others
Forest	-	4 services (38 datasets)	-	-
Biodiversity	-	-	1 service	1 GBIF-service
Drought	2 services (102 datasets)	6 services (161 datasets)	3 services (40 datasets)	1 WFS-G
Generic	2 services	-	-	-

CSW - « Catalogue Service for the Web »
Service used to request metadata catalogues of datasets and services.

WMS - « Web Map Service »
Service used to download geospatial information in a raster format. WMS are mainly view services.

WFS - « Web Feature Service »
Service delivering raw geospatial data (under GML). WFS are mainly downloading services.

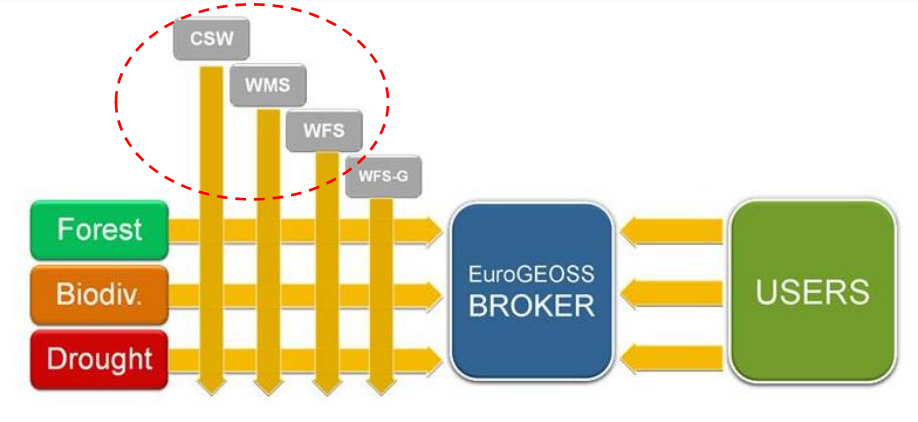
WFS-G
WFS used to deliver gazetteer service (producing Bounding Box from toponyms).

GBIF
Specific query interface connected to the GBIF (Global Biodiversity Information Facility) metadata catalogue.

GBIF is preparing services for the EuroGEOSS broker

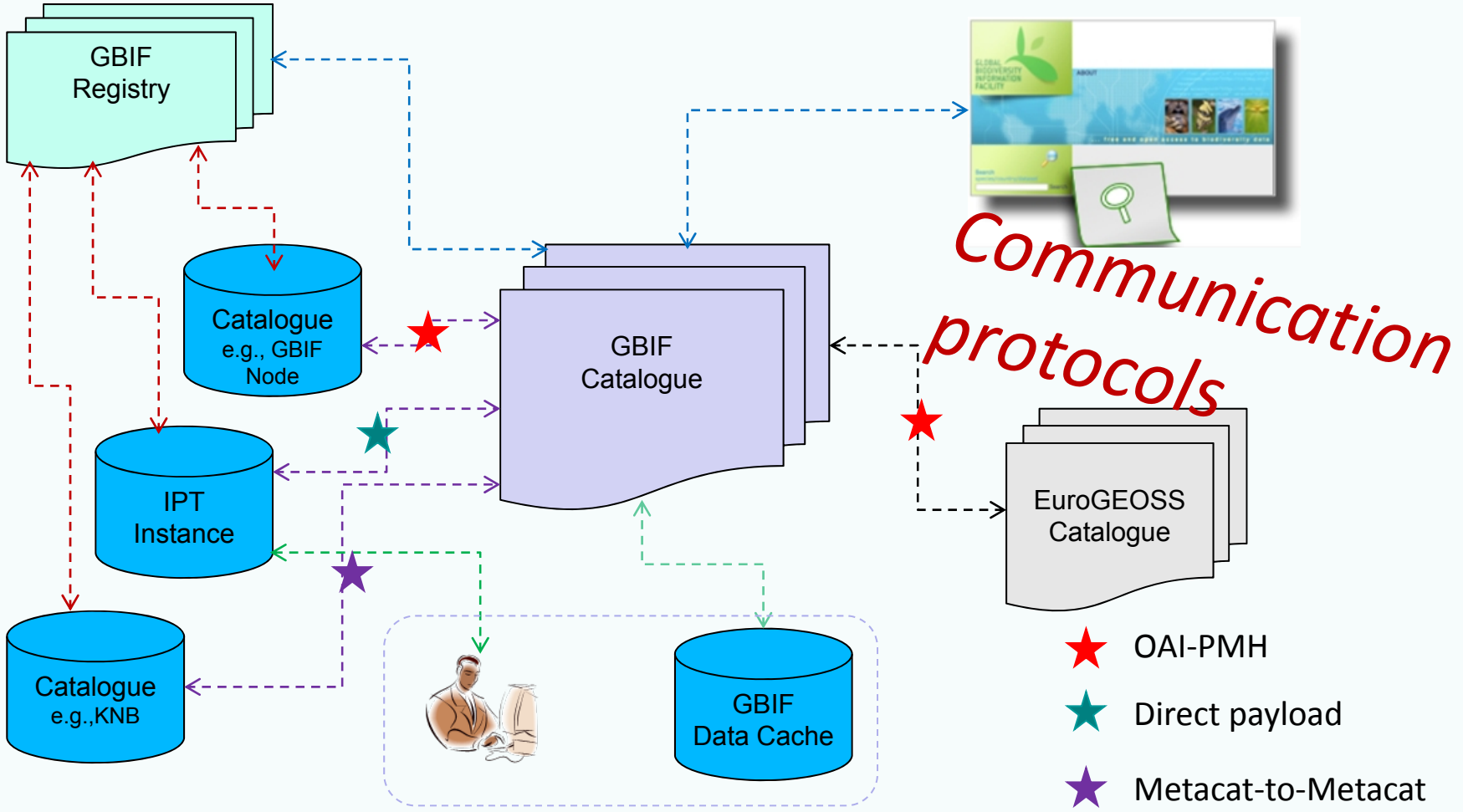
- *Web Map Service*
- *Web Feature Service*
- *~~CSW~~ -> OAI-PMH*

(Open Archives Initiative – Protocol for Metadata Harvesting)



Source: http://www.eurogeoss.eu/broker/default.aspx

Architecture: main components



OAI-PMH

- Open Archives Initiative Protocol for Metadata Harvesting
- Providing a low-barrier mechanism for interoperability across distributed metadata repositories
- Data providers expose metadata; Service providers consume metadata through a client application known as a harvester that issues OAI-PMH service requests over HTTP:

  1. GetRecord 2. Identify 3. ListIdentifiers

  4. ListMetadataFormats 5. ListRecords 6. ListSets

GBIF: role as harvester and producer

1. return individual record 2. retrieve information about repository 3. retrieve headers of records

4. return metadata formats available 5. return records from repository 6. retrieve set structure (groupings) of repository

http://www.openarchives.org/pmh/

Metadata Standards

Ecological Metadata Language (EML) v2.1.0
http://knb.ecoinformatics.org/software/eml/

Dublin Core (http://dublincore.org/documents/dcmi-terms/)

Directory Interchange Format (DIF)
http://gcmd.nasa.gov/User/difguide/difman.html

ISO 19115/19139 Geographic Metadata
ISO 19115: http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=26020
ISO 19139: http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=32557

Natural Collections Descriptions (NCD)
http://www.tdwg.org/standards/312/

Federal Geographic Data Committee Biological Profile*
http://www.fgdc.gov/standards/projects/FGDC-standards-projects/metadata/biometadata/

Multimedia Resources Metadata Schema
http://www.tdwg.org/charters/article/view/448/36

*An extension of the FGDC CSDGM (Content Standard for Digital Geospatial Metadata)

ISO 19115/19139

FGDC CSDGM → *ISO 19139*

North American Profile of ISO 19139
http://www.fgdc.gov/nap/metadata/napMetadataProfileV101.pdf

Several Resources available for crosswalk; transform; view

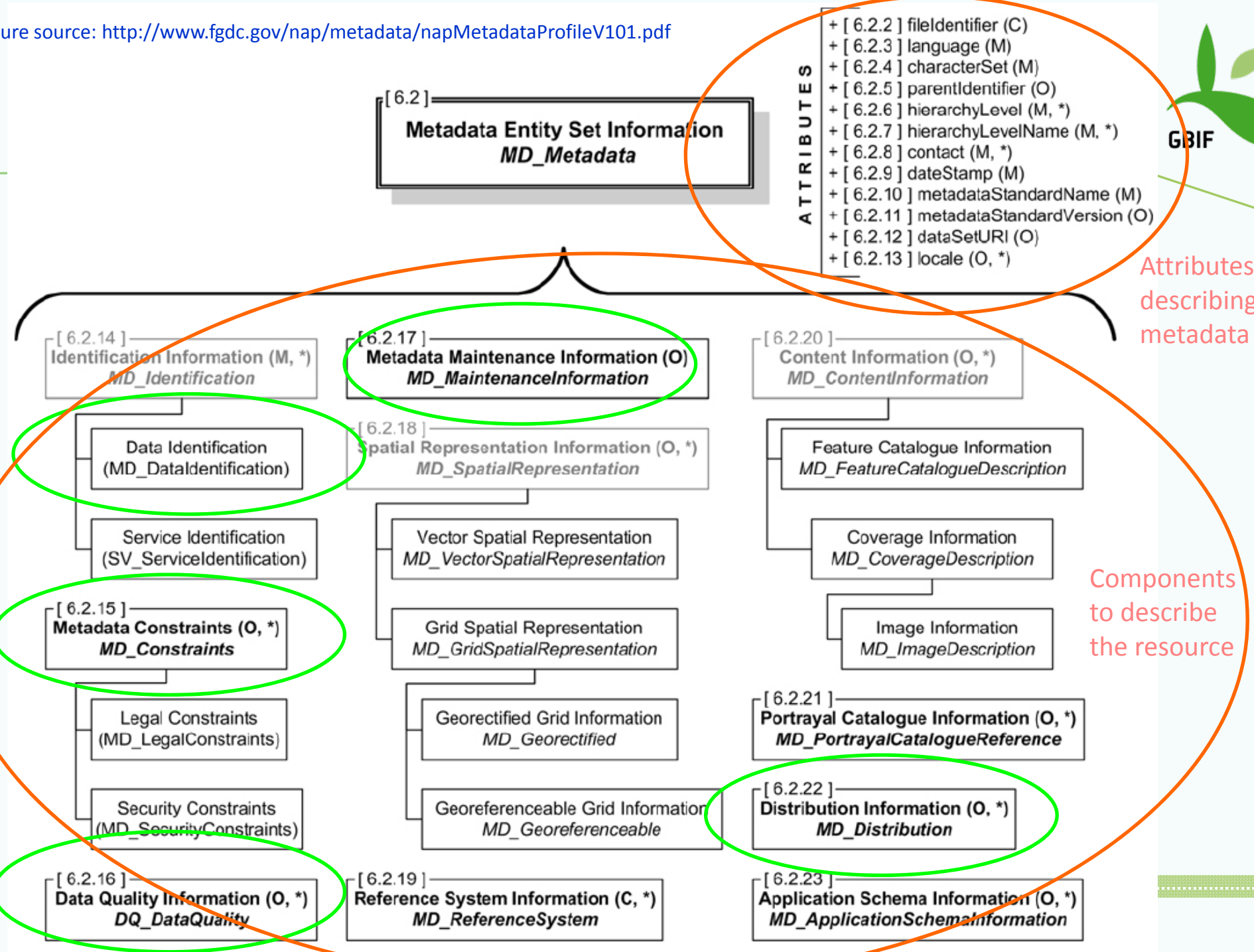
FGDC CSDGM to ISO Transform
FGDC CSDGM to ISO Crosswalk
ISO XML to HTML View:
FGDC BIO to ISO Transform
FGDC BIO to ISO Crosswalk
http://www.ncddc.noaa.gov/technology/metadataandxml/view

Open source INSPIRE-compliant MD editor
(planned multilingual functionality)
http://www.inspire-geoportal.eu/EUOSME/

EML to FGDC Biological Profile
https://code.ecoinformatics.org/code/eml/trunk/lib/eml2tonbii/

EML to ISO 19139
http://rs.gbif.org/schema/eml/eml2iso19139.xsl

Figure source: http://www.fgdc.gov/nap/metadata/napMetadataProfileV101.pdf



Attributes describing the metadata

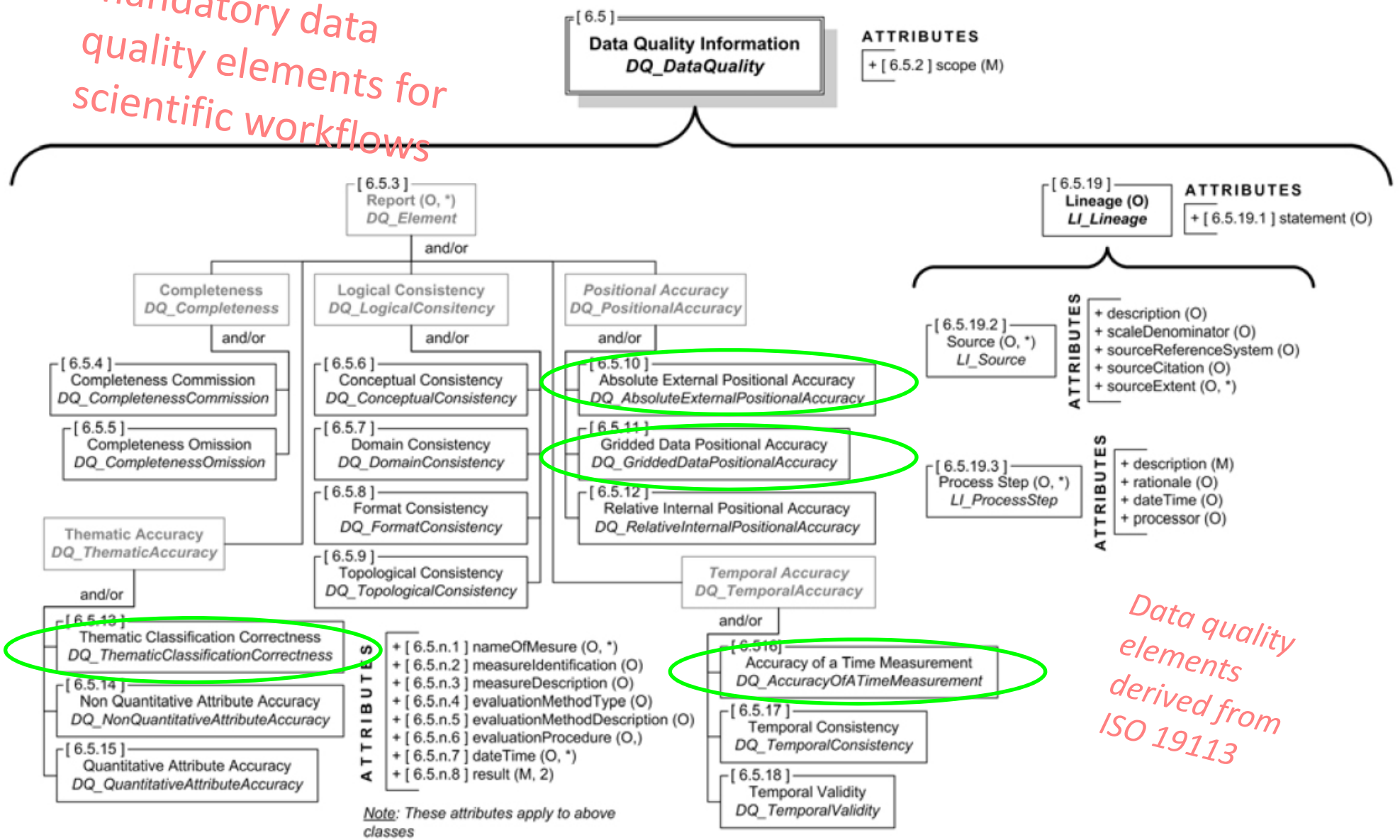
Components to describe the resource

This section describes attributes and components that provide information about data quality.

Type: DQ_DataQuality

Figure source: http://www.fgdc.gov/nap/metadata/napMetadataProfileV101.pdf

EuroGEOSS mandatory data quality elements for scientific workflows



Data quality elements derived from ISO 19113

Documenting Data Quality

ISO 19113 “Geographic information – Quality principles”
http://www.iso.org/iso/catalogue_detail.htm?csnumber=26018

Lays out the principles for describing the quality of geographic data.

Used to assess -

- How well does a dataset meet its original specification?
- How well does the dataset meet the needs of a new application?

Essential for use of data in modelling

Positional Accuracy particularly relevant for current GBIF mediated data.

Absolute external positional accuracy
closeness of reported coordinate values to values accepted as or being true

Gridded data positional accuracy
closeness of gridded data position values to values accepted as or being true

Appropriate measures:
- circular error at 95%
- variance
- probability density function

Metadata and Languages

Example network: International Long Term Ecological Research Network (ILTER)

Data providers to provide discovery level metadata in English

Initial focus on tools for metadata capture and translation: localised metadata editors; multilingual environmental thesaurus

GEMET, the GEneral Multilingual Environmental Thesaurus;
27 languages. http://www.eionet.europa.eu/gemet/

AGROVOC; agriculture, forestry, fisheries, food and related domains
e.g., environment; 20 languages. http://www4.fao.org/agrovoc/default.htm

Long term solution: multilingual ontologies

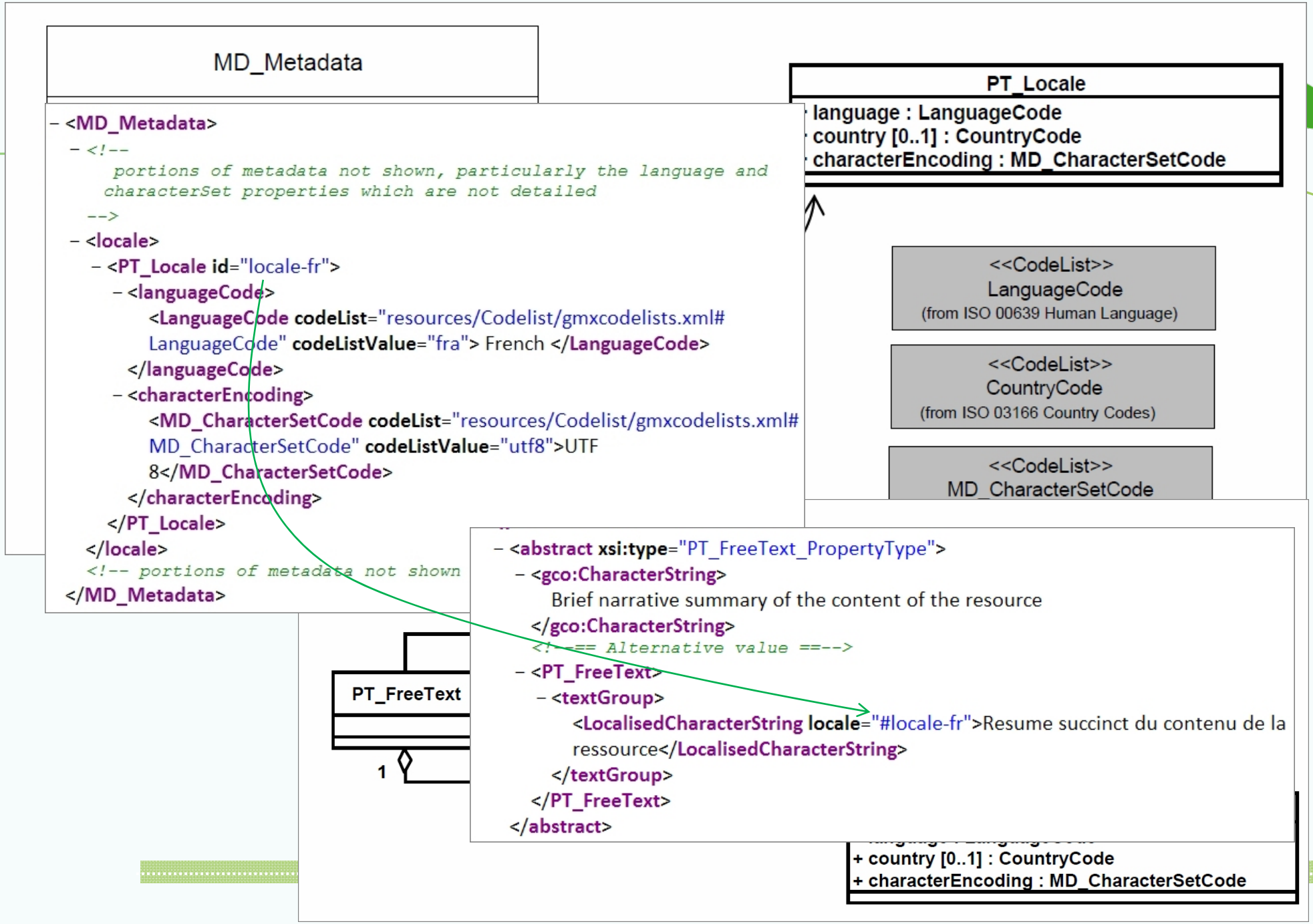
A Multilingual Metadata Catalog for the ILTER: Issues and Approaches.
Vanderbilt, K.L., et al., Ecological Informatics, Volume 5, Issue 3, May 2010, Pages 187-193,
doi:10.1016/j.ecoinf.2010.02.002

Metadata and Languages

Metadata standards vary in ability to allow use of multiple (natural) languages

EML – currently limited to a single language; investigations under way on how to “globalize” EML.
http://mercury.nceas.ucsb.edu/ecoinformatics/mailman/listinfo/eml-dev

ISO 19115/19139 provides comprehensive “locale” unit.
http://inspire.brgm.fr/Documents/MD_IR_and_ISO_20080425.pdf




```
– <MD_Metadata>
  – <!--
      portions of metadata not shown, particularly the language and
     characterSet properties which are not detailed
    -->
  – <locale>
    – <PT_Locale id="locale-fr">
      – <languageCode>
          <LanguageCode codeList="resources/Codelist/gmxcodelists.xml#
          LanguageCode" codeListValue="fra"> French </LanguageCode>
        </languageCode>
      – <characterEncoding>
          <MD_CharacterSetCode codeList="resources/Codelist/gmxcodelists.xml#
          MD_CharacterSetCode" codeListValue="utf8">UTF
          8</MD_CharacterSetCode>
        </characterEncoding>
      </PT_Locale>
    </locale>
    <!-- portions of metadata not shown
</MD_Metadata>
```

```
– <abstract xsi:type="PT_FreeText_PropertyType">
  – <gco:CharacterString>
      Brief narrative summary of the content of the resource
    </gco:CharacterString>
    <!-== Alternative value ==-->
  – <PT_FreeText>
    – <textGroup>
        <LocalisedCharacterString locale="#locale-fr">Resume succinct du contenu de la
        ressource</LocalisedCharacterString>
      </textGroup>
    </PT_FreeText>
  </abstract>
```

Knowledge Organisation Systems

The need for community supported dictionaries, vocabularies, thesauri and ontologies (Knowledge Organisation Systems - KOS) is a key issue for advancing interoperability.

KOS in 2011 GBIF Work Programme

Goal 2: Facilitate development and deployment of standards

… facilitate, in response to community needs, development and deployment of vocabularies and ontologies for biodiversity data

Activity 1: Scope requirements for a standard for annotating data/metadata records

Activity 2: Scope requirements for a standard for describing species distribution data

Activity 3: Review geospatial web services in GBIF portal/network

Activity 4: Commission a task group on primary biodiversity data associated with genomic data

Key questions for discussion

How can GBIF help? Infrastructure, tools, training

How are you using metadata? discovery; priority setting; automated scientific workflow; admin/stats

As a regional metadata network, what kind of network intersections do you envisage? within network, with GBIF, with others

What is the role of a central GBIF portal for metadata?

How do you propose to handle language issues?

What are your approaches to incentivising production of high quality, complete metadata?

Key activities for 2012-2016 work programme?

Help us design the GBIF metadata system

How to contact GBIF:

Web site: www.gbif.org
Data portal: www.gbif.net

GBIF Secretariat
Universitetsparken 15
DK-2100 Copenhagen Ø
Denmark

E-mail: info@gbif.org
Phone: +45 3532 1470
Fax: +45 3532 1480

GBIF Secretariat building, supported by a grant from the Aage V. Jensens Fonde