GBIF

INFORMATION FACILITY

GBIF data architecture

Francisco Pando, GBIF Spain

Based on presentations by Tim Robertson (GBIF Information Systems Architect), David Remsen (ECAT PO)

NATIONAL MUSEUMS OF KENYA
WHERE HERITAGE LIVES ON

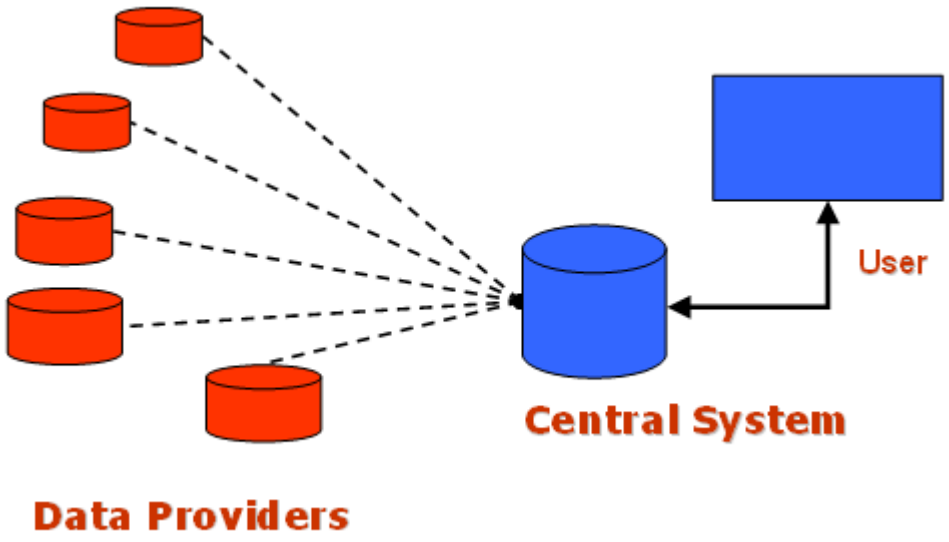
KENBIF TRAINING WORKSHOP

Nairobi, 6th to 7th June 2011

About GBIF

- An operational network
- Connecting hundreds of institutions
- Thousands of data sources
- Free and open access to information

Centralized model



Distributed model (network)

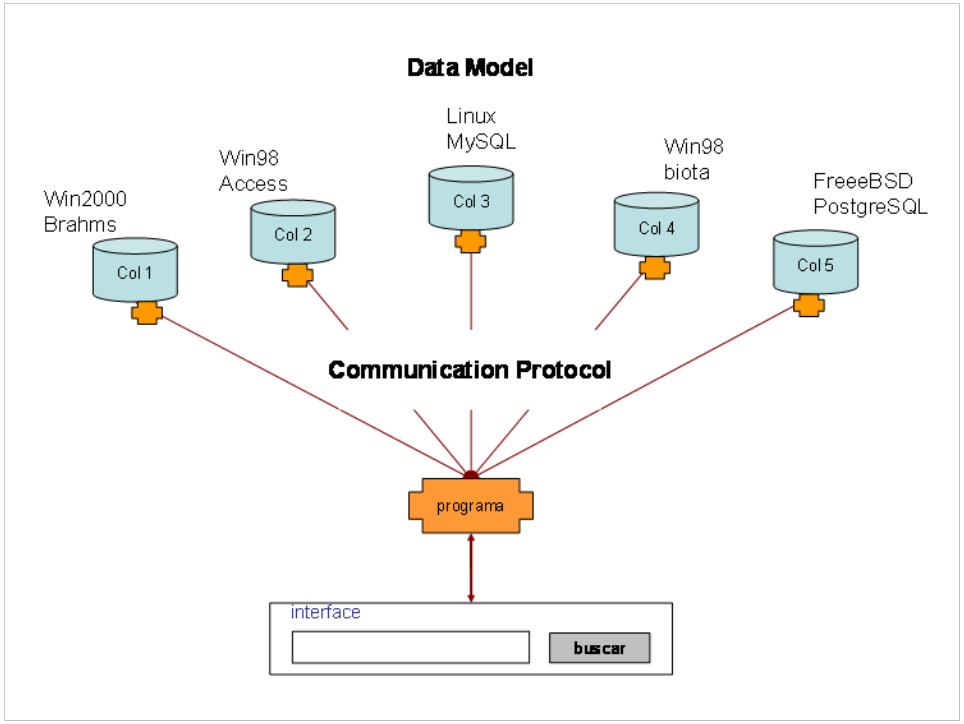
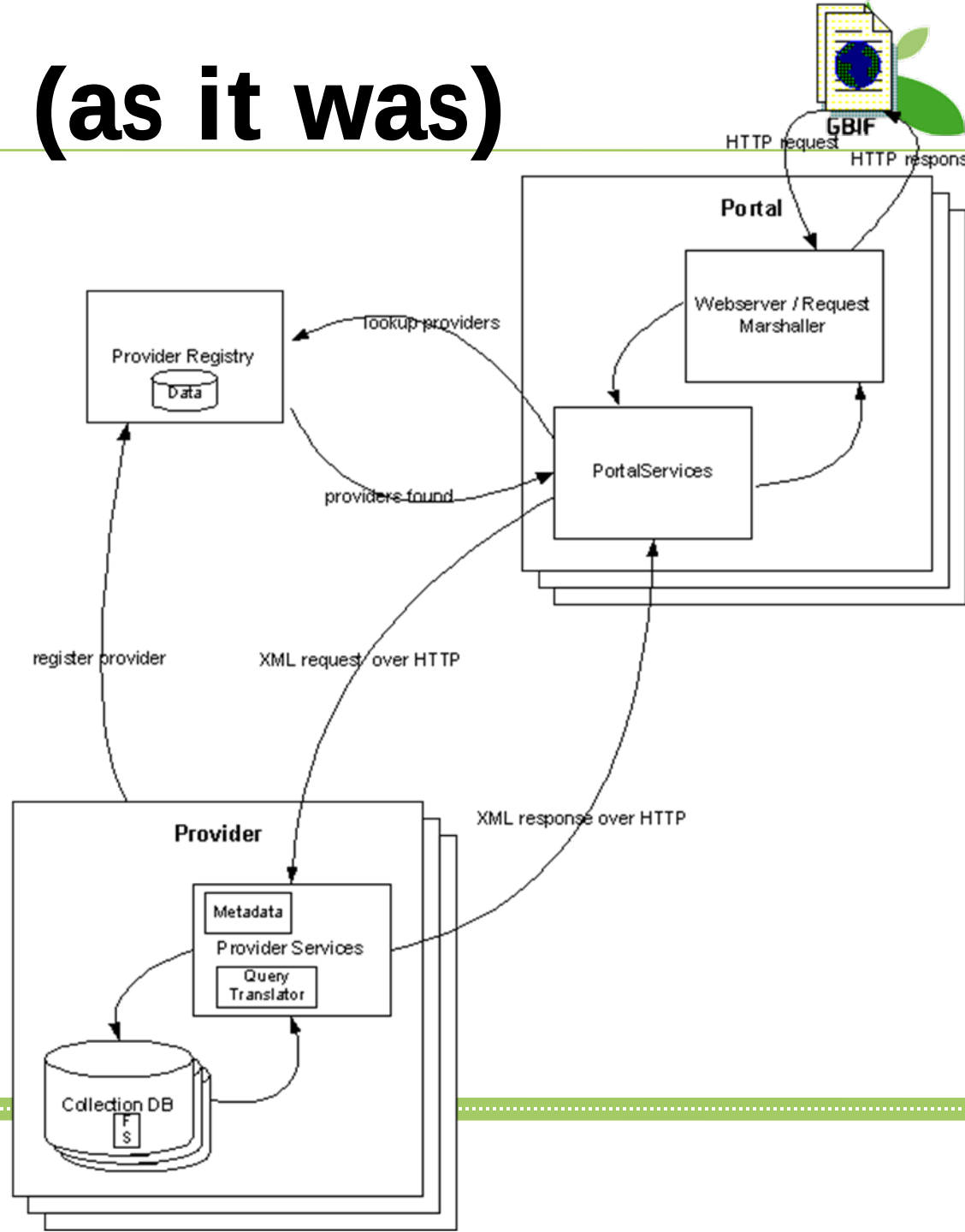


Figure 4. Diagram showing the complexity of integrating data from biological collections

In some detail (as it was)

DiGIR

- Protocol
- Provider
- Portal
- Registry



Architecture Strategic Big Picture

GBIF'S ARCHITECTURE

Services

Integration

GBRDS REGISTRY

DATA STORES

METADATA

SPECIMEN & OBSERVATION

CHECKLIST

SPATIAL

VOCABULARY REPOSITORY

HARVESTING & INDEXING TOOLKIT

DiGIR

BioCASE

Other tools

DarwinCore archive

INTEGRATED PUBLISHING TOOLKIT

Legend

Data store

data flow

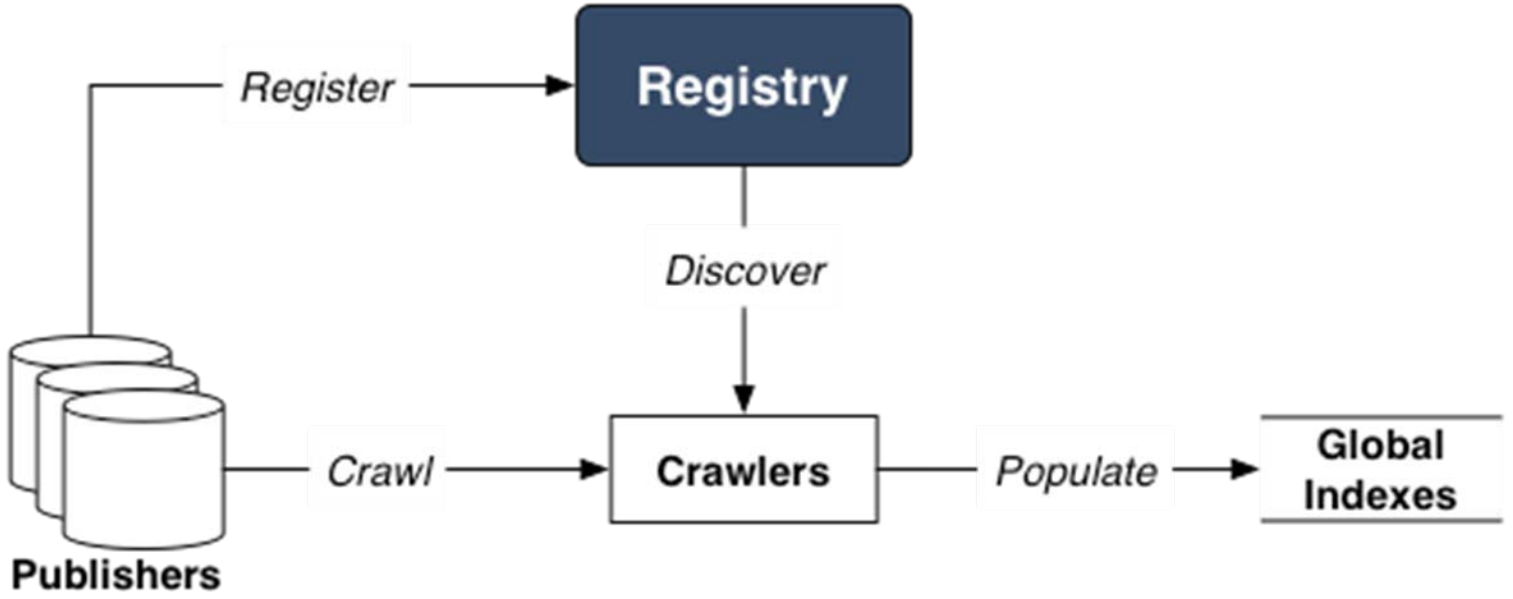
registration and resolution

work in progress

- Data profile standard (Darwin Core)
- Registry
- Index, Cache databases & protocols
- Portal

Registry component

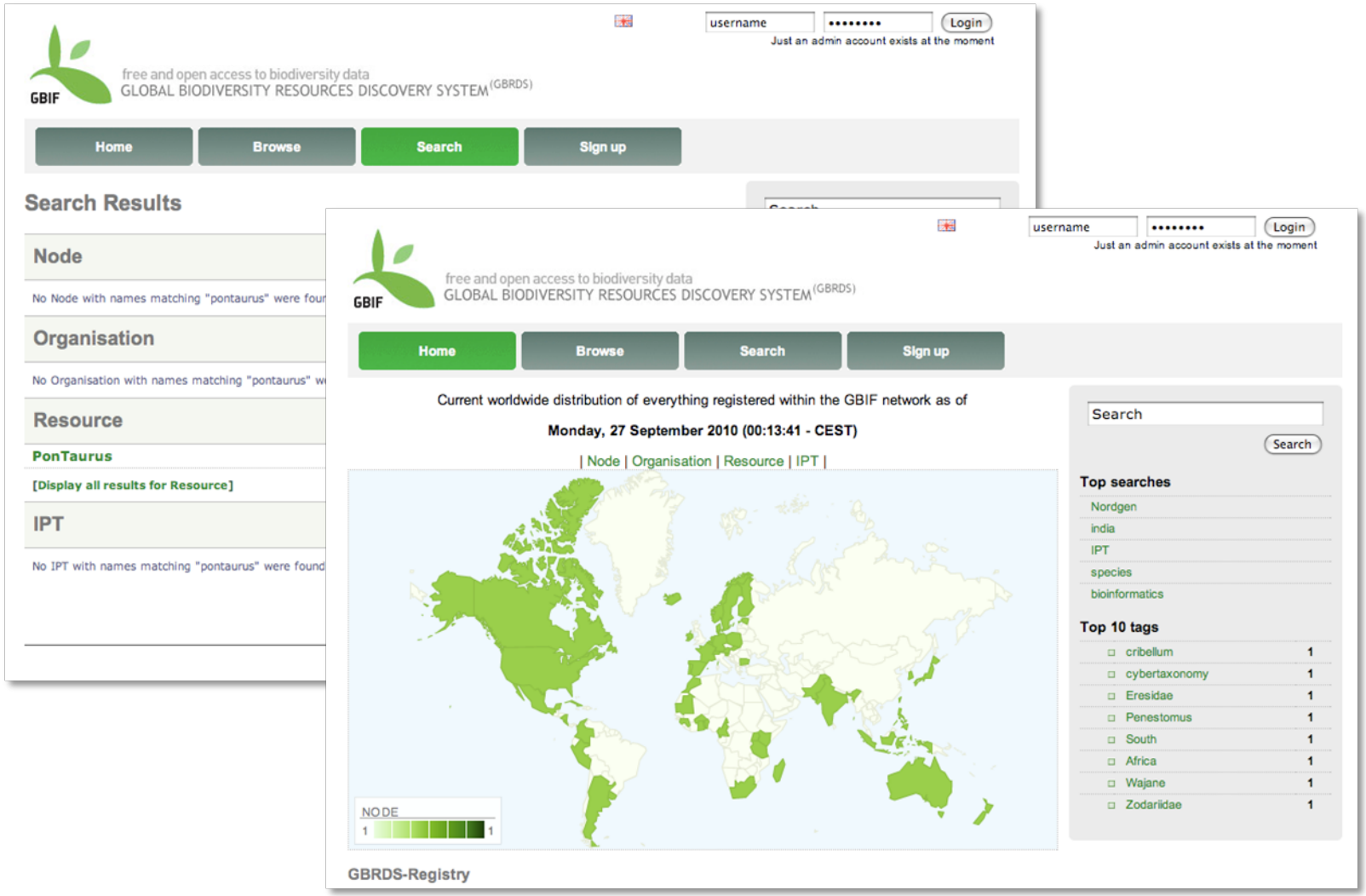
- Provides the information to determine the participating institutions in GBIF and the technical end-points to access their datasets, along with contact information.



Registry component

- Previously implemented using an open industry business registry known as UDDI
  - 2-tier model of “data publisher having several datasets”
- The GBIF network is more complicated than this.
  - Datasets are shared or published through multiple channels. Results in complex attribution chains.

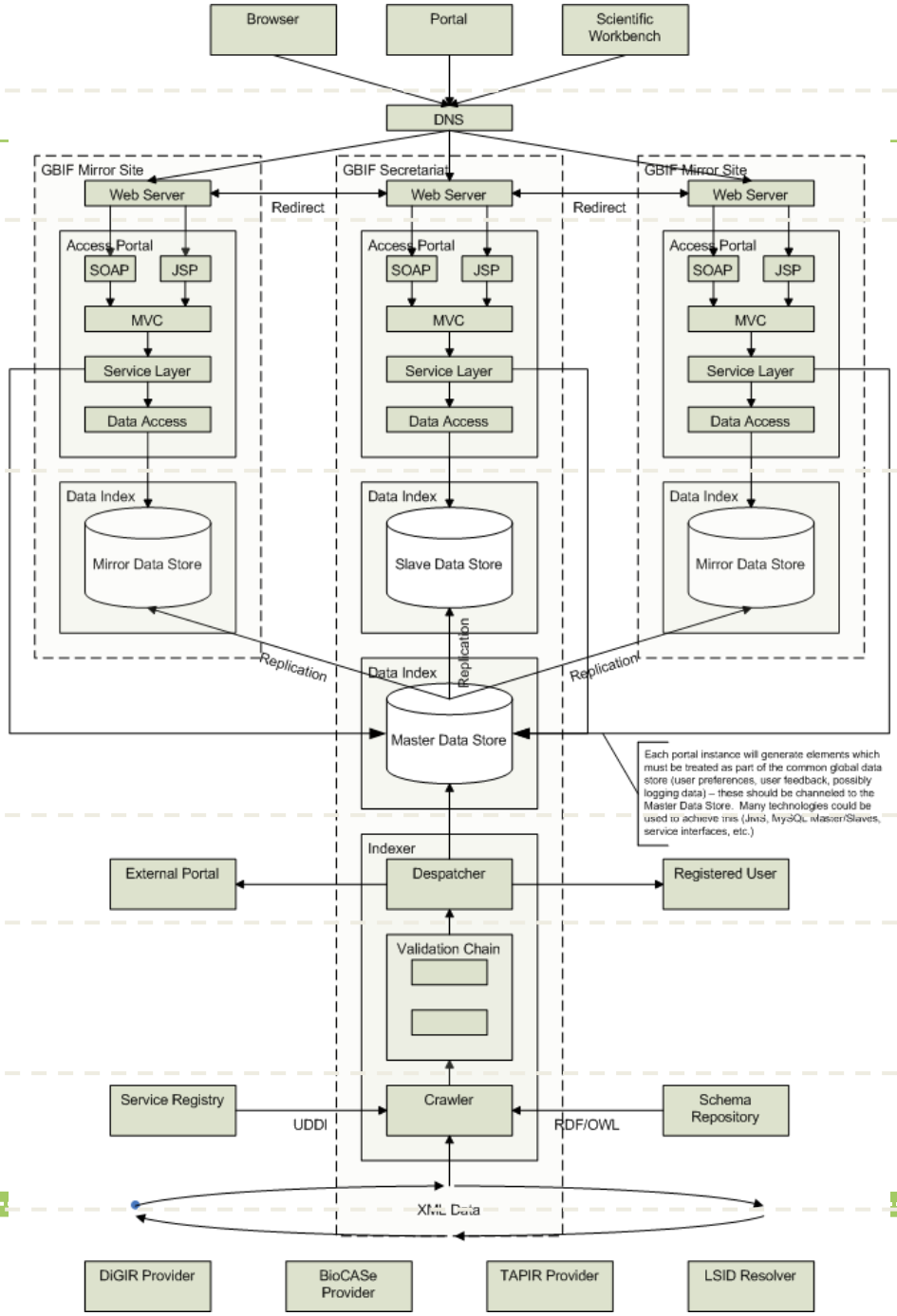
Registry component



Status: Prototype available

http://gbrds.gbif.org

Portal architecture



Clients

Mirrored access

Web applications

Synchronised data stores

Data despatcher

Interpretation and validation

Resource crawler

Data resources

Challenge: performance

- Post-harvesting stage:

215,000,000 records refreshed in a month	
7,000,000 per **day**	
300,000 per **hour**	(24/7 is a challenge in itself!)
5000 per **minute**	
83 per **second**	(with no growth…)

- Clearly parallelisation is key…

  … and database becomes a bottleneck

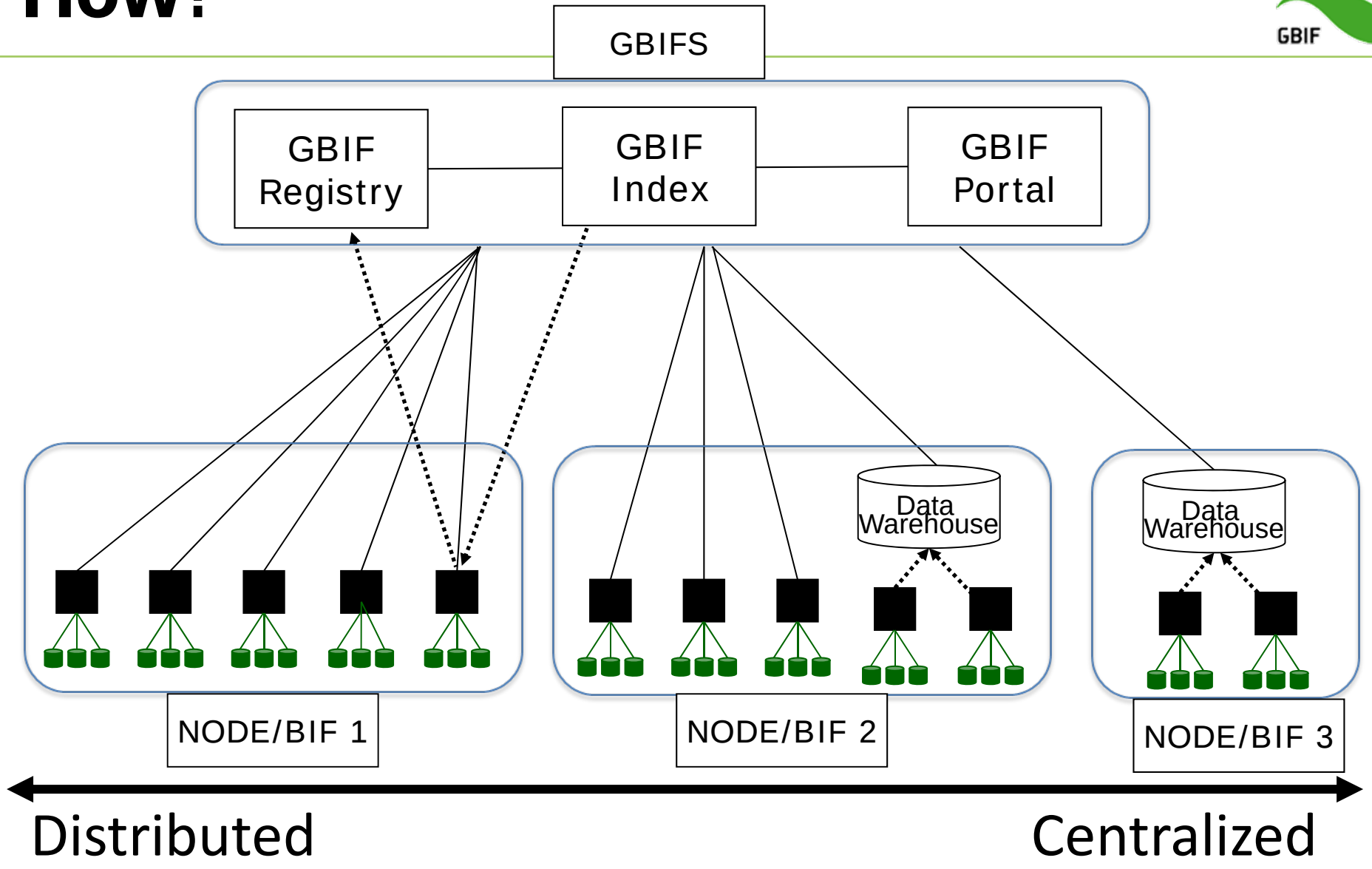
New approaches

- HIT (Harvesting Indexing tool
- Darwin Core Archive
- Cloud / Grid technologies

Protocols and other means of buliding the index (Cache database)

- May depend on the extent of your IT support
  - Via web services - data transfer long
    - BioCASE - Biological Collection Access Services
    - DiGIR - Distributed Generic Information Retrieval
    - TAPIR - TDWG Access Protocol for Information Retrieval
  - Via zipped text archives - data transfer short
    - Darwin Core Archives (GBIF)
    - Tab-separated values with appropriate headers (DiscoverLife.org)
    - GBIF's IPT (Integrated Publishing Toolkit)
  - XML - Guide includes
    - Simple Darwin Core Schema http://rs.tdwg.org/dwc/xsd/tdwg_dwc_simple.xsd
  - Additional resources (including Excel spreadsheet!)
    - http://code.google.com/p/darwincore/wiki/ToolsAndApplications

How?


GBIFS
GBIF Registry
GBIF Index
GBIF Portal
Data Warehouse
Data Warehouse
NODE/BIF 1
NODE/BIF 2
NODE/BIF 3
Distributed
Centralized

What makes GBIF work

- Standards for data and protocols (and their interaction via web services)
- Control and ownership of data remains with providers
- Registry for advertisement of data
- Integration at portals
- GBIF is multi-purpose open-ended cyber-infrastructure that enables taxonomists and others to serve the society in new ways

At your command

Francisco [Paco] Pando
GBIF Spain Node Manager
GBIF NODES Commitee Chair

GBIF Spain
Real Jardín Botánico - CSIC
Claudio Moyano 1
28014 Madrid, Spain

pando@gbif.es
www.gbif.es
Phone: + 34 91 420 3017
Fax: + 34 91 420 0157