Data Sharing and Repositories in the GBIF network

Tim Robertson
GBIF Head of Informatics
Global Biodiversity Information Facility (GBIF)

trobertson@gbif.org

2014 EU BON General Meeting
Greece, April 2014

Session aim:

Participants improve their understanding of the GBIF network architecture

The nature of repositories within the GBIF network is understood

The GBIF registry functionality is understood

The GBIF Integrated Publishing Toolkit is demonstrated in “near real time” publication

GBIF
www.gbif.org

This presentation is available (temporarily) on
http://tinyurl.com/gbif-eubon-1

GBIF: A distributed network

- **Publishers**: Those institutions that are sharing datasets through the GBIF network. (A GBIF publisher **may** be analogous to a repository to other networks)
- **Aggregators**: Those institutions that are assembling datasets from publishers, often for providing thematic services (e.g. regional / taxonomic queries or reports)
- **Registry**: The coordinating component in the network
- **GBIF Portal**: A service platform that provides discovery of, access to and reporting of data shared through the network. The registry is a critical component of the portal.

GBIF
www.gbif.org

Publishers: Protocols and tools

Servers that expose databases through XML protocols. Clients can query using custom filters (similar to SOAP)

- Biological Collections Access Services (**BioCASe**) and the BioCASe pywrapper
- Distributed Generic Information Retrieval (**DiGIR**) with a PHP implementation *(legacy)*
- TDWG Access Protocol for Information Retrieval (**TAPIR**) with a PHP implementation

GBIF
www.gbif.org

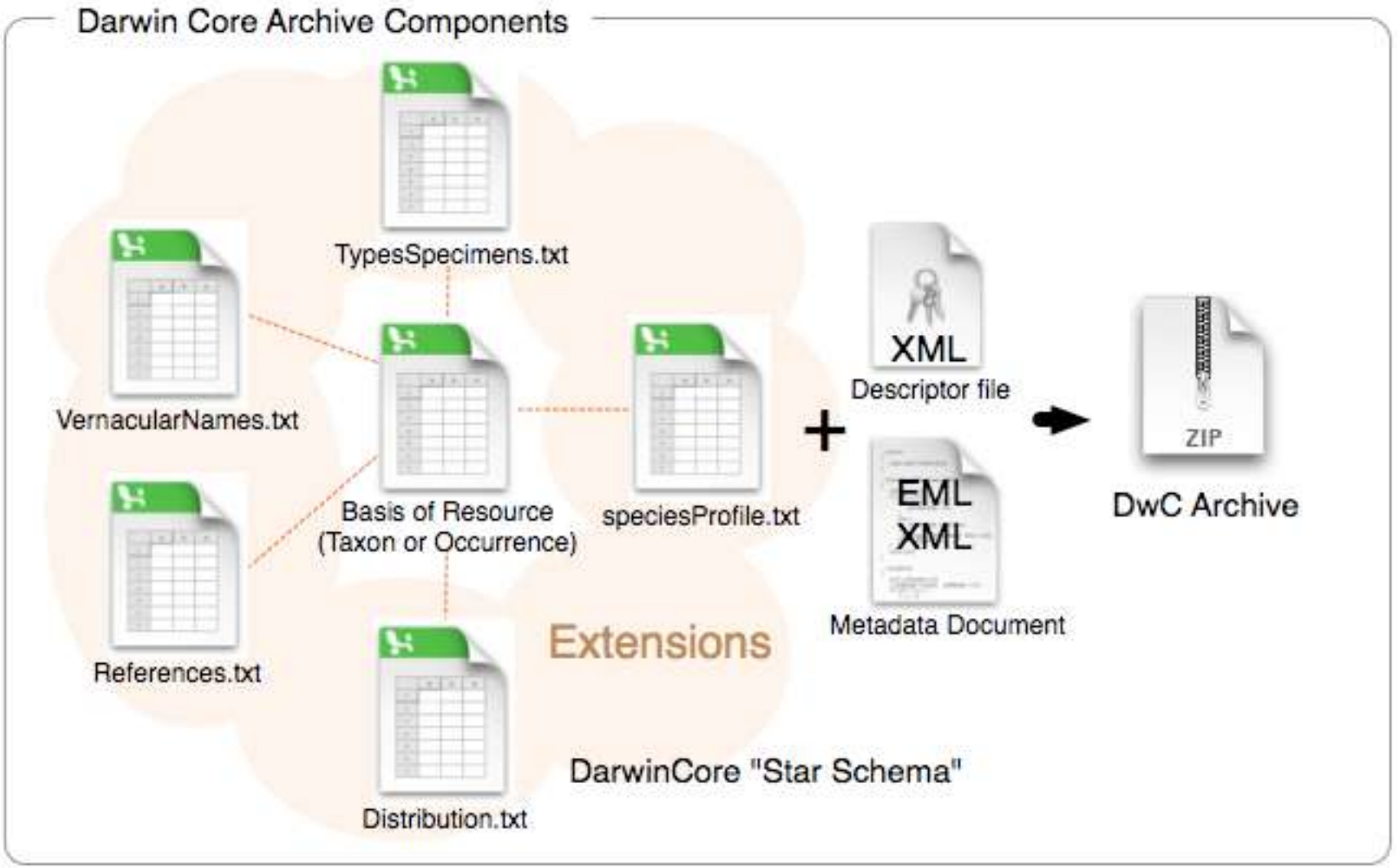
Publishers: Protocols and tools

Servers that expose a single data archive, known as the Darwin Core Archive

- The GBIF Integrated Publishing Toolkit (IPT)
- Custom solutions, served by generic web servers

GBIF
www.gbif.org

Publishers: Protocols and tools



GBIF: A distributed network

596 Publishing institutions on the GBIF network

"Decentralization and distribution of IT resources begets centralized services"

(author unknown)

GBIF
www.gbif.org

Component: Coordinating registry

Administration console
(for GBIF staff)

GBIF
www.gbif.org

Component: Coordinating registry

A centrally hosted database, accessed through RESTful interface

- http://www.gbif.org/developer/registry
- Registration of Institutions, Datasets, Networks, Technical Installations, GBIF Participant Nodes and means to access them
    - Technical service access points
    - Human resource contacts
- Organization, user and application based access control (security)

GBIF
www.gbif.org

Component: Coordinating registry

- Ecological metadata language for dataset. GBIF metadata profile: www.gbif.org/resources/2559
- A central unique identifier minting service (UUIDs) for registered entities
- Alternative identifiers for an entity (e.g. DOIs for a dataset)
- Repository for data standards http://rs.gbif.org/
  - Darwin Core Archive (DwC-A) extension definitions
  - Darwin Core Archive vocabulary definitions

GBIF
www.gbif.org

Component: Coordinating registry

- Supports generic tagging API
- User scoped “Private tagging”
- Enables datasets to be associated
  - Dataset duplication (happens regularly)
  - Dataset composition (E.g. Catalogue of Life)
- Models institutional relationships
  - Endorsement model for GBIF
  - Relationships of data ownership and data hosting (for credit and attribution)

GBIF
www.gbif.org

Component: Coordinating registry

- Coordinates GBIF indexing schedule
  - Emits messages (RabbitMQ) on change
  - Supports API for crawlers to store history
  - Background scheduler tasks interrogate and detect datasets eligible for update

- Dataset access history
  - GBIF index download queries stored, the filter used and number of records returned

- Database enables ad hoc reporting for GBIF
  - Data access by country of user
  - Publishing metrics (e.g. datasets by GBIF country)

Component: Coordinating registry

Technologies

- Jersey for REST (Java)
- PostgreSQL for database
- Apache SOLR for indexing
- RabbitMQ for messaging
- AngularJS for administration console
- Yammer for metrics
  - *Integrates with ganglia for monitoring*
  - *Integrates with central log management using logstash and kibana*

(All open source technologies)

GBIF
www.gbif.org

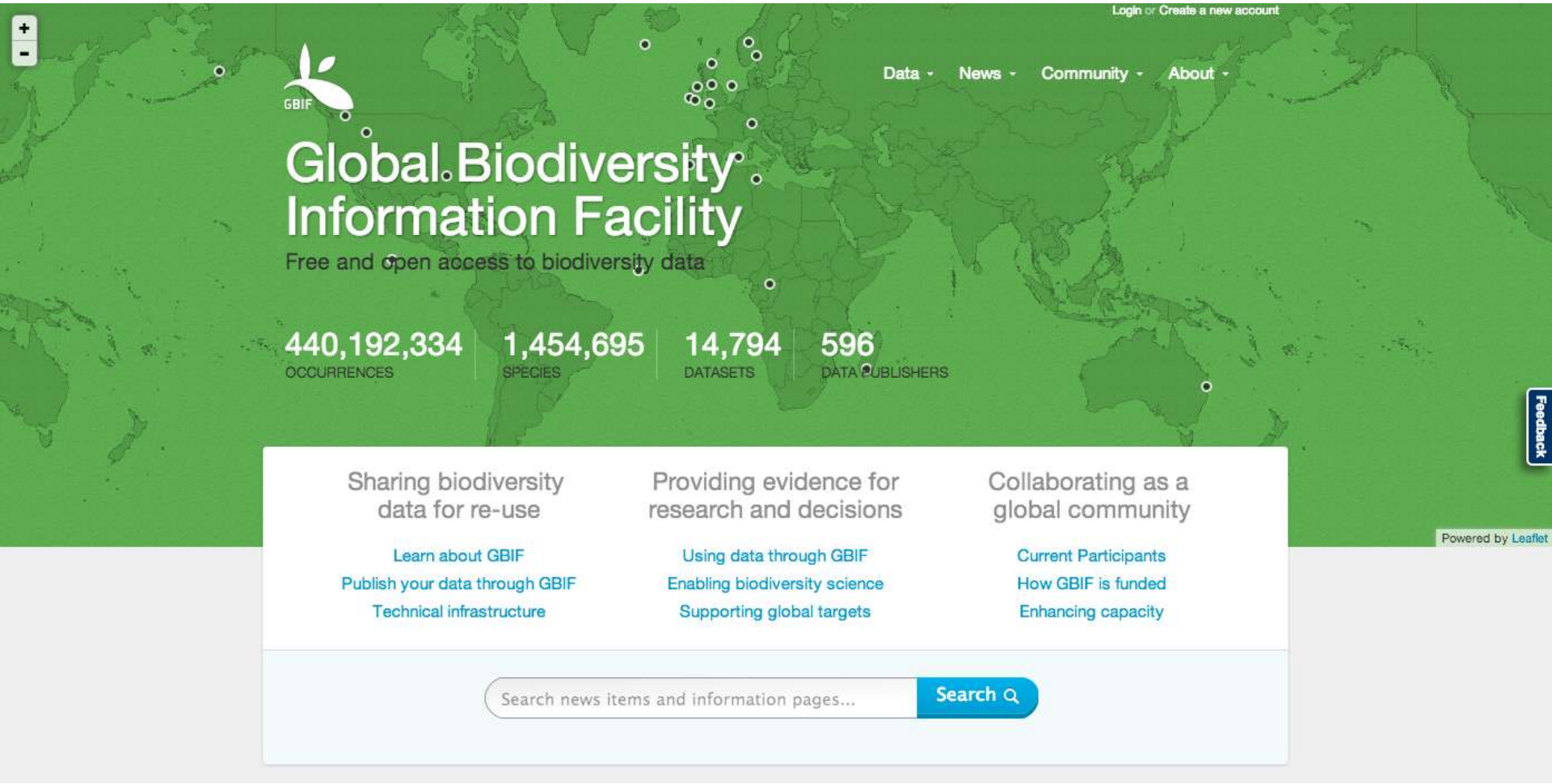
Component: Coordinating registry

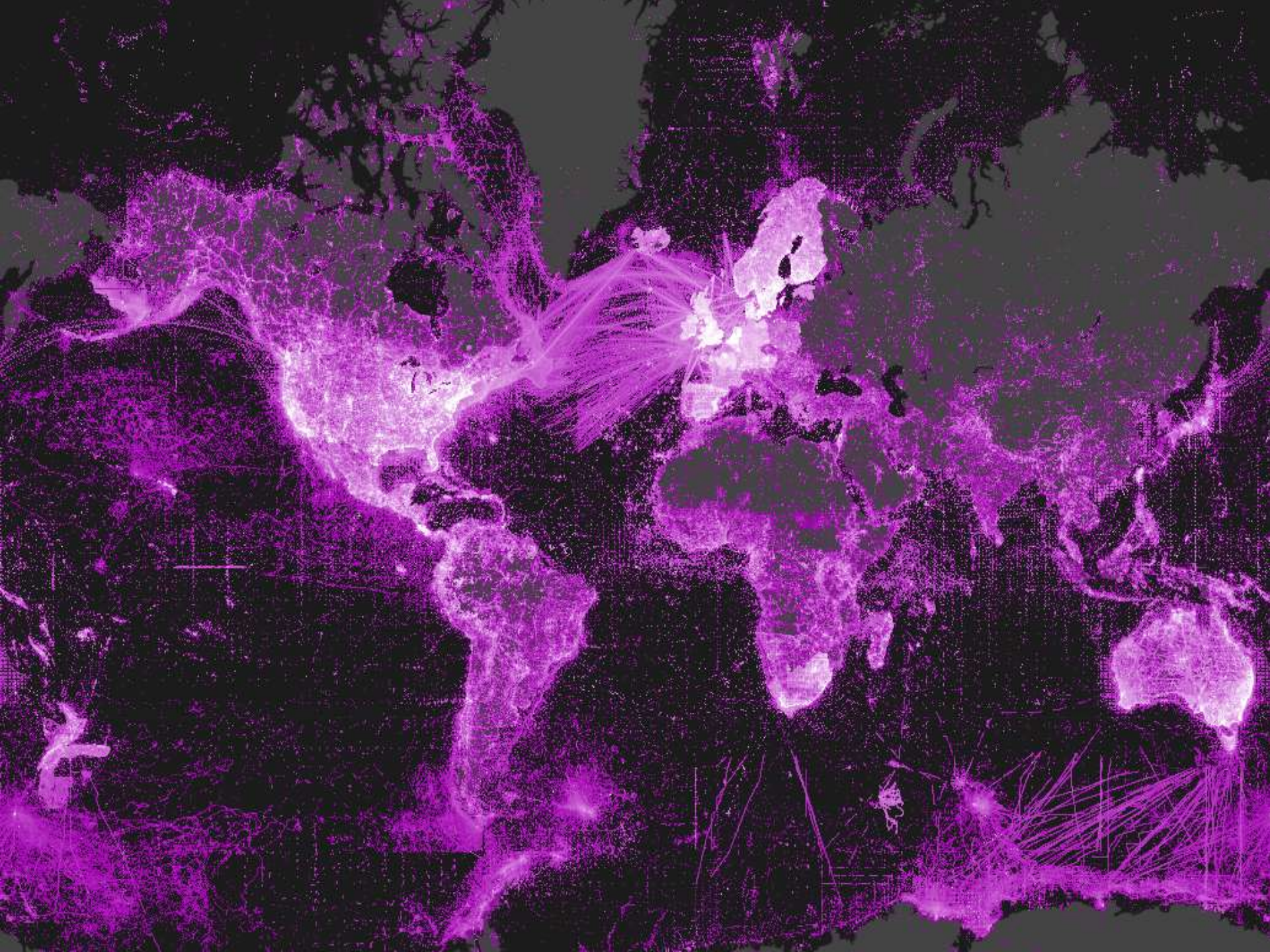
Forthcoming developments

- DOI minting service for datasets
  - proxy for DataCite
- Citation services
  - Built around DOIs
  - Complex citation generated, archived, and DOI assigned for the citation itself
- Coordination of redundant storage
  - Publishers opt to have redundant offsite replicas
  - Registry coordinates distributed archival (through partners, potential DataONE / EU BON)

GBIF
www.gbif.org

Component: GBIF Portal





Component: GBIF Portal

5,000 - 7,000 users per day

86,621,589,856 records downloaded
- 6 months
- Darwin Core Archive format

Live site: http://www.gbif.org/

Component: GBIF Portal

A customized Drupal implementation for CMS, and Java application for data

http://www.gbif.org/developer

Technologies at a glance:

- *Drupal for CMS*
- *PostgreSQL for checklists*
- *Hadoop, HBase, for occurrences*
- *SOLR for search*
- *Jersey for REST*
- *RabbitMQ for messaging*
- *Yammer for metrics*

GBIF
www.gbif.org

We’re hiring

http://www.gbif.org/newsroom/opportunities

- Programme Officer for Content Mobilization
- Programme Officer for Content Analysis / Use

GBIF
www.gbif.org