NoSQL working group
Use case: Network of Life

Mario David (LIP)

With contribution from Miguel Porto and Rui Figueira (CIBIO Portugal)

Outline

- GBIF and Atlas of Living Australia Web portal
- From GBIF to Network of Life
- Graph DBs - ArangoDB
- Current status and first tests

Challenges of GBIF biodiversity data

Global Biodiversity Information Facility

- 570 million records with many dimensions.
- Need to support different spatial scales, information detail, in the same platform.
- Ensure confidence, users need to be able to scrutinize all details of information.
- The rate of new data addition is not fully predictable.
- Crossing data with other types of information (remote sensing, climatic) is also resource-demanding.

Atlas of Living Australia

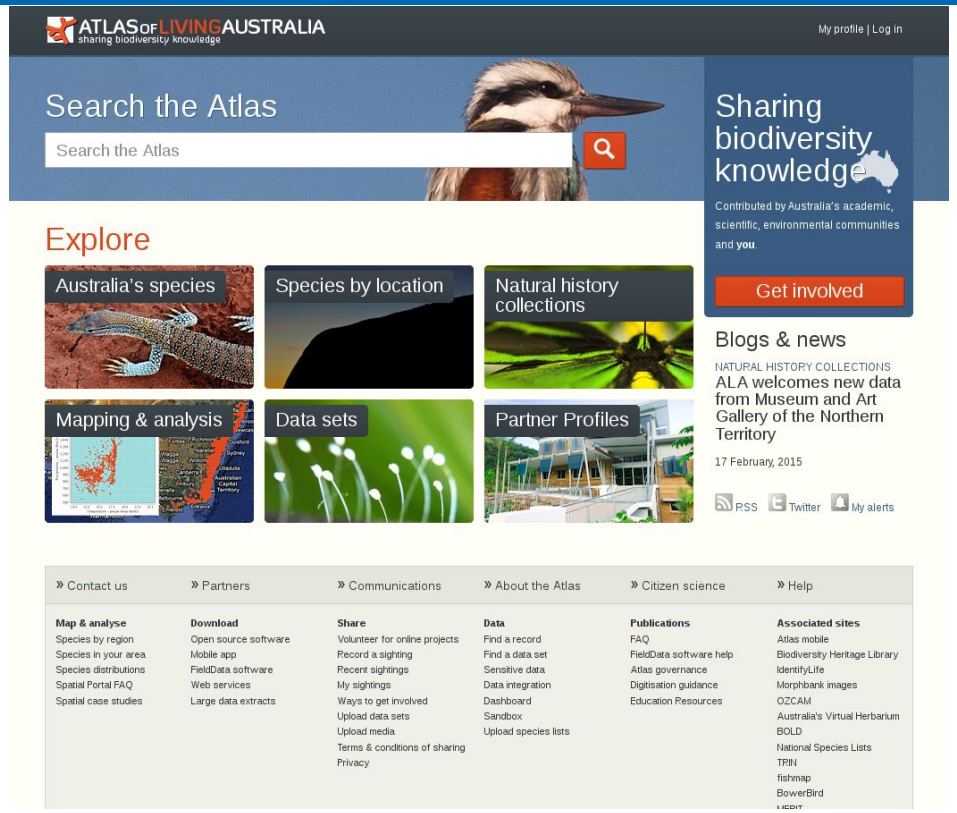
Platform for web portals and services for societal uses in biodiversity

IBERLIFE action!

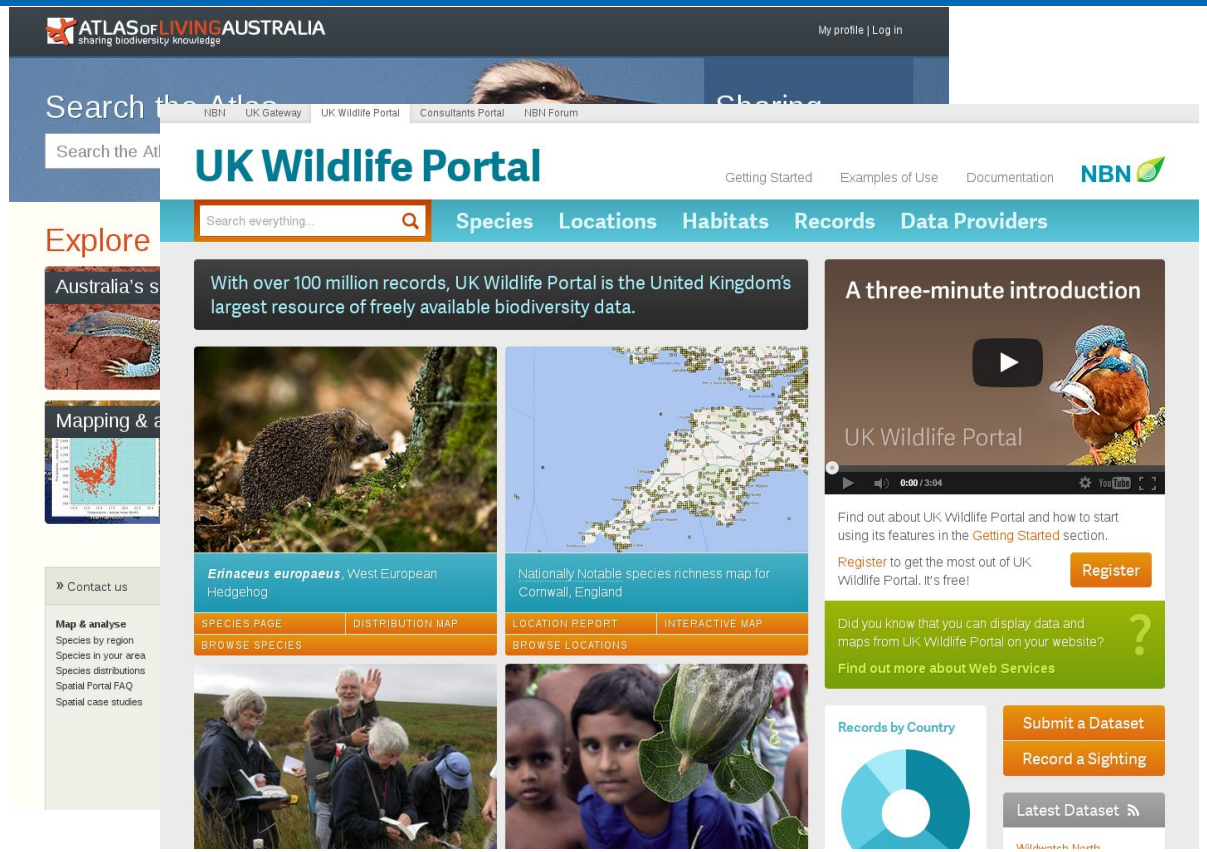
Provide:

- Efficient organization and management of biodiversity information, including to find, access and visualize data;
- Integration with genetic, habitat, ecosystem and geographical data;
- Building different facets, e.g., for Invasive Alien Species, threatened species, nature conservation
- Web data services through API.

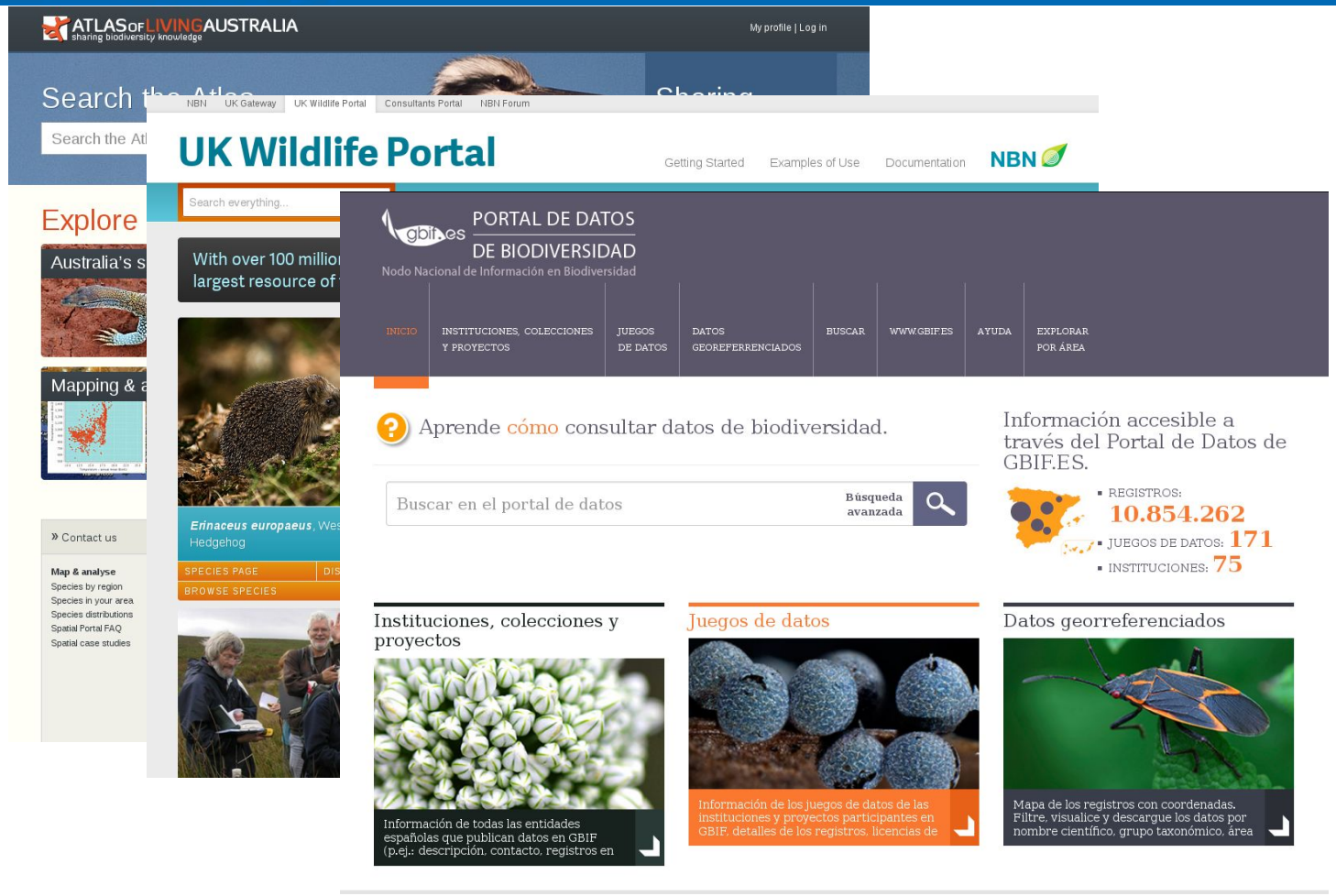
One platform, many facets (thematic, regional, national), different user communities



One platform, many facets (thematic, regional, national), different user communities

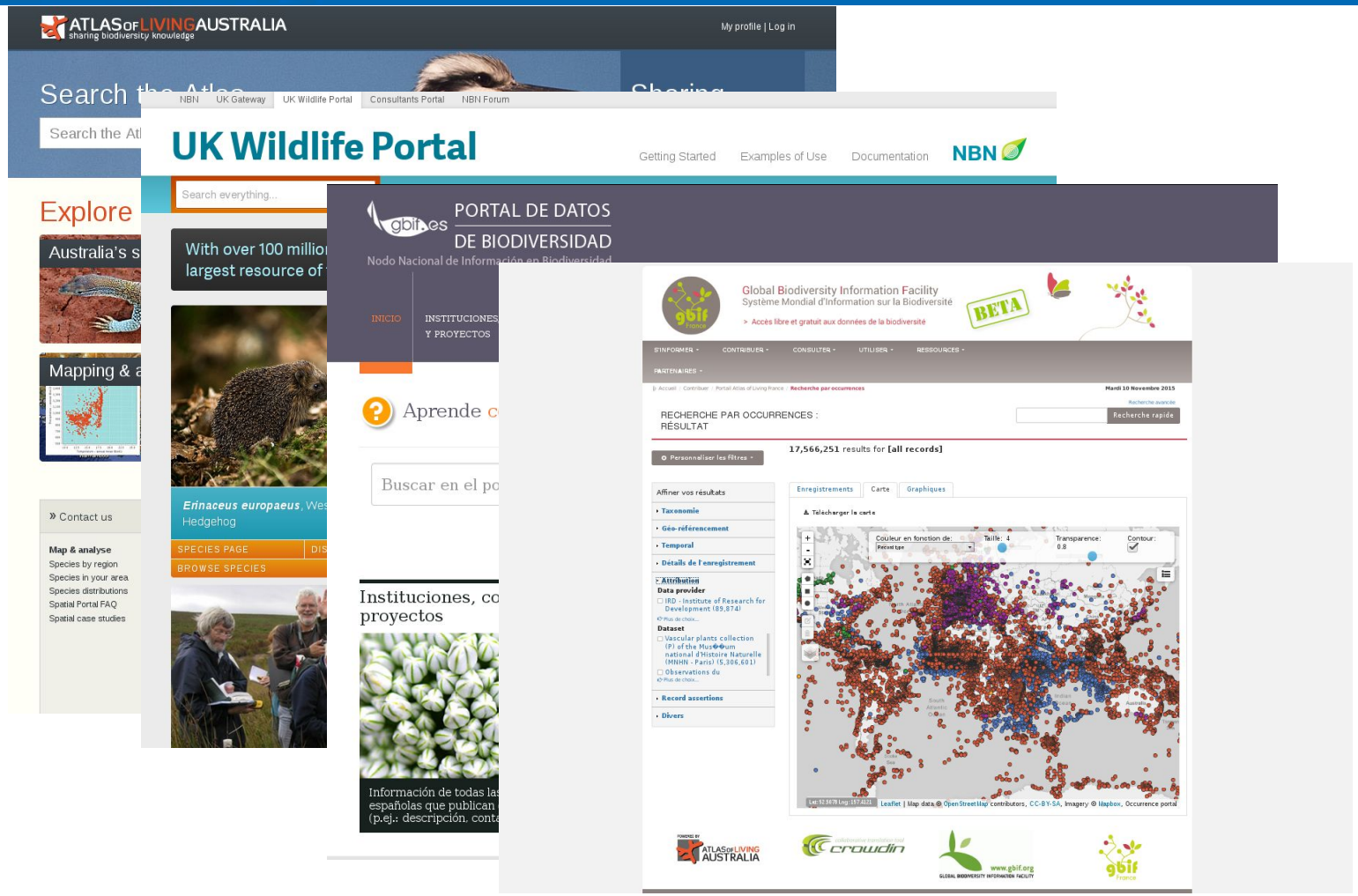


One platform, many facets (thematic, regional, national), different user communities

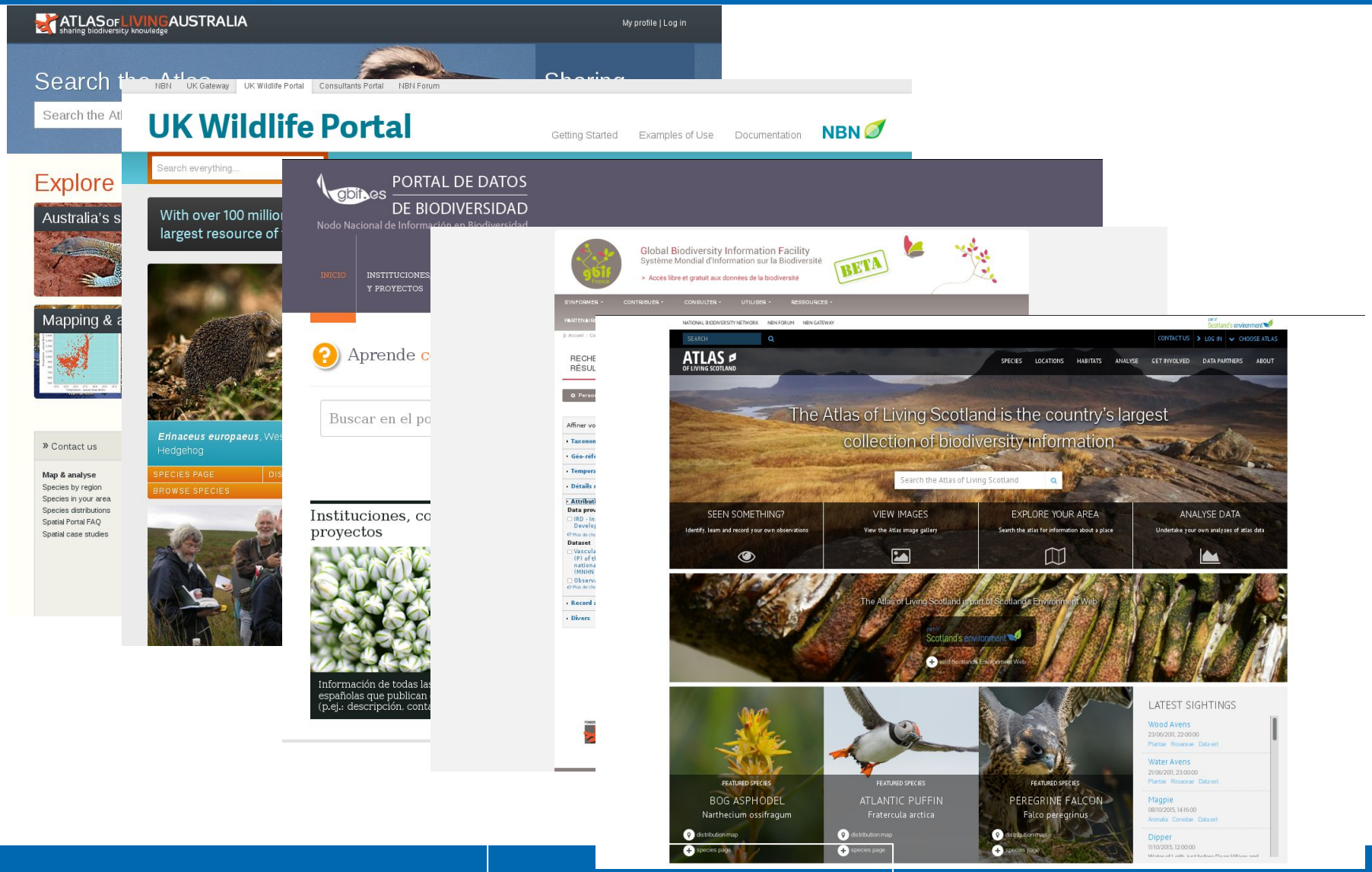


Rui Figueira (CIBIO)

One platform, many facets (thematic, regional, national), different user communities



One platform, many facets (thematic, regional, national), different user communities



Rui Figueira (CIBIO)

Advantages of cloud solutions

Provide:

- Scalability of the allocation of resources.
- Sharing infrastructure and capacity between members of GBIF network.
- Persistence and availability of big volumes of data.

GBIF ⇒ Net of Life

Biologists POV

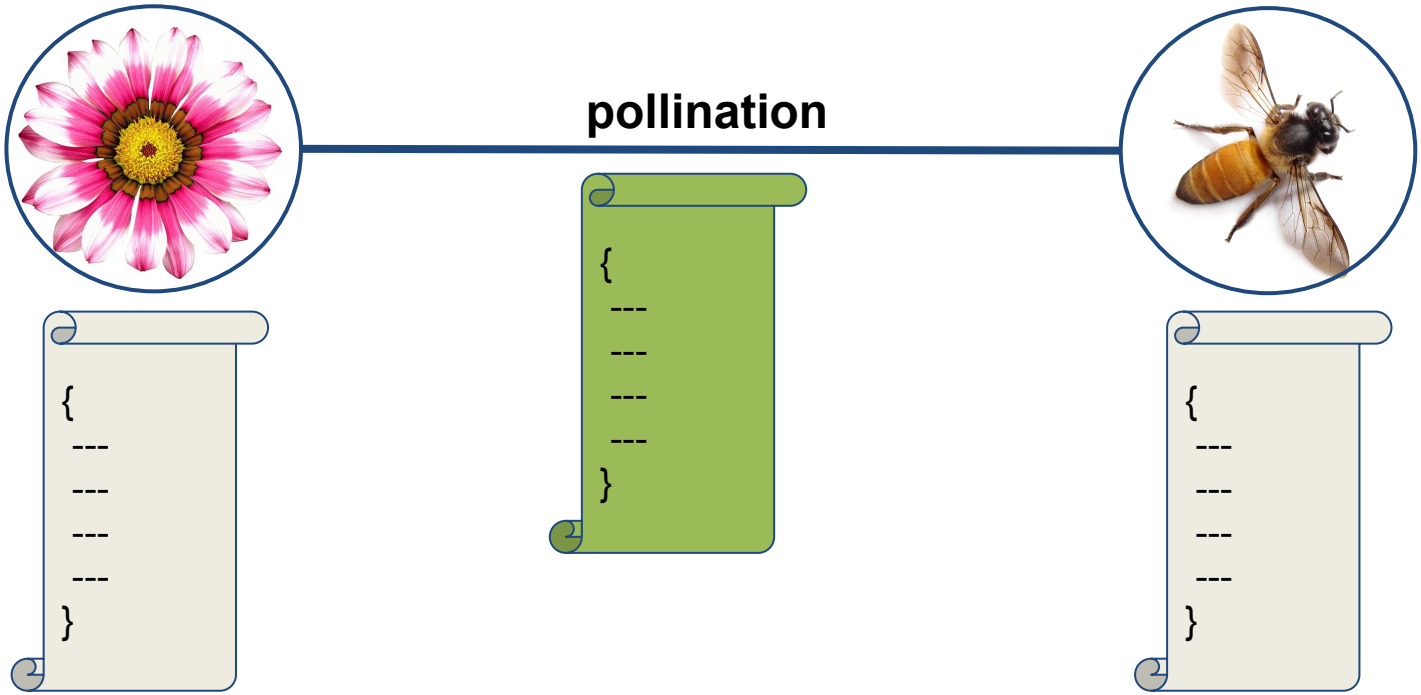
GBIF

```
{
 ---
 ---
 ---
 ---
}
```

```
{
 ---
 ---
 ---
 ---
}
```

Biologists POV

Network of Life

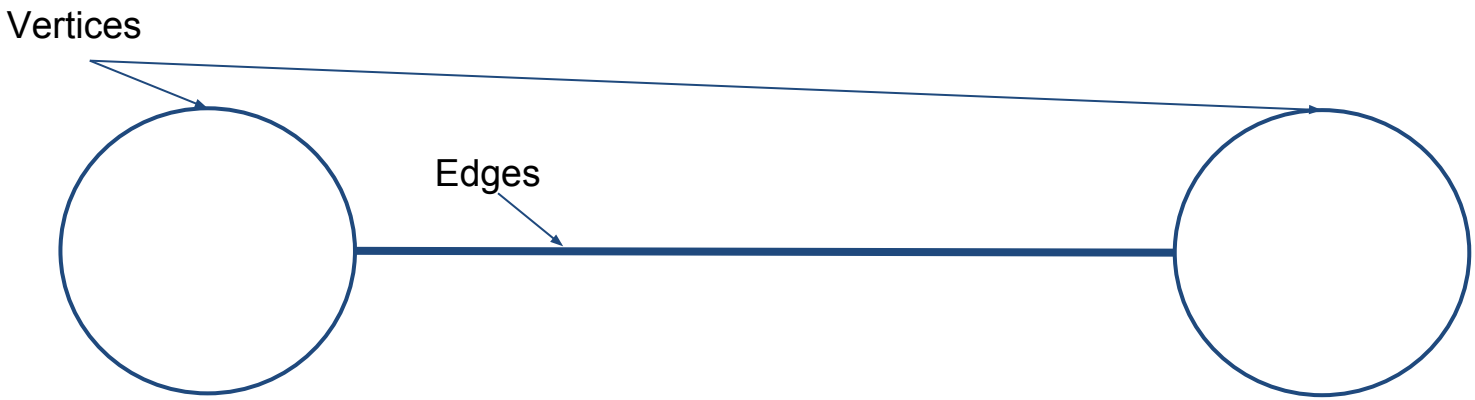


GBIF ⇒ Net of Life

Maths/Comp.Scient POV

G = (V, E)
V = {v1, v2, …}
E = { {v1, v2}, {v1, v3},... }

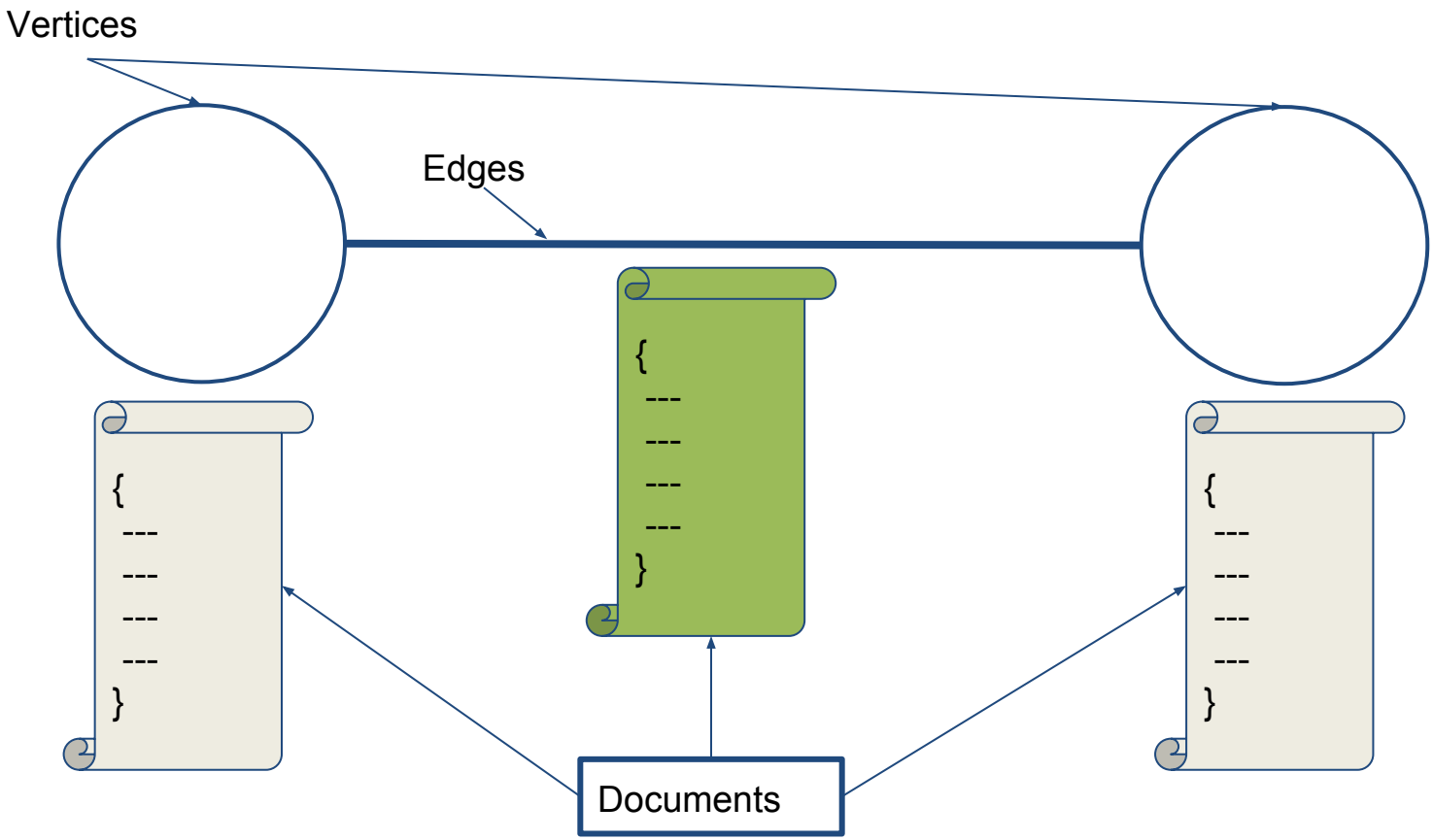
Graph ⇒ GraphDB



GBIF ⇒ Net of Life

Maths/Comp.Scient POV

GraphDB + Documents ⇒ ArangoDB



ArangoDB - I

ArangoDB

- Multi-model database: document, graph, key-value
- Open source: https://github.com/arangodb/arangodb
- *Document model*:
  - Data stored as linked JSON-like documents, organized in collections
  - No schema enforced, but set of indexes can be defined for each collection
  - Fields can store other subdocuments and pointers to independent documents

ArangoDB - II

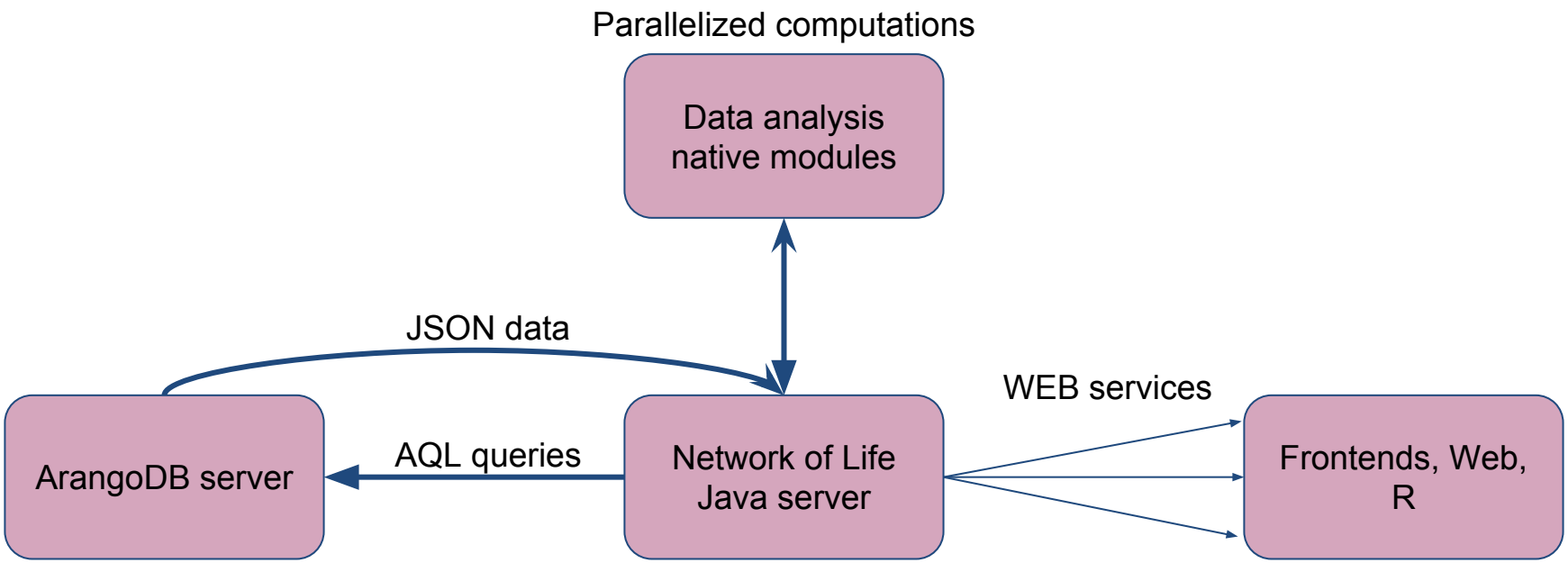
- *Graph model*:
  - An “interpretation” built upon the document model:
    - Defined by a set of document collections representing **vertices**.
    - Another set of collections representing the **edges** connecting the vertices.
  - Vertexes and Edges are documents.
  - Native support for traversal queries:
    - Highly customizable behaviour
    - No need for “infinite” JOINs.
- *Indexes*:
  - Graph traversal indexes (edge-vertex connections)
  - Geo indexes (constructed from latitude-longitude fields)
  - Full text, hash, etc.

ArangoDB - III

- *AQL query language*:
  - SQL-like but very different logic:
    - Entirely JSON-based.
    - No tables.
  - Rather complete set of functions to work with documents:
    - Data aggregation.
    - Filtering (including Geo functions), etc.
    - Document and array manipulation
  - Graph traversal and shortest path functions
  - Easy querying, processing and output results in the desired data format
  - Very flexible in chaining and nesting query sentences

"Powerful and Fast"

Network of Life: Architecture



Graph traversal
Data aggregation

Exposes services for:

- querying interaction data at different levels of aggregation
- downloading raw data
- submitting data analysis jobs
- uploading new data

Visualization
Network queries
Network data analysis
Hypothesis testing
Data downloading
...

Some first tests

- Simple ArangoDB instance running on the desktop
- Good query performance, in particular the ones involving geographic indexes and graph traversal
  - ArangoDB having integrated geo indexes matches nicely the use case
- The application logic should be implemented in the AQL queries.

Test deployment

- ArangoDB in cluster mode ⇒ allow sharding
- Deployed 2 VMs in INCD Openstack
- Each VM with 2 types of processes:
  - **Coordinators**: receives requests, distributes them to the DBServers, executes AQL queries and returns the result to the clients. The coordinator also exposes information about cluster health and cluster statistics.
  - **DBServers**: can both store sharded (and non-sharded) collections.
  - A database and a coordinator can live on the same server.
- And… learning the business :)