AntaBIF.BE
BIODIVERSITY.AQ

AntaBIF - technical document

June 2011

Draft V0.1

Background

The AntaBIF (Antarctic Biodiversity Information Facility), funded by the Belgian Science Policy Office (www.belspo.be), is building a dedicated Antarctic biodiversity data portal giving access to a distributed network of contributing database, according to the principles of the Global Biodiversity Information Facility (www.gbif.org), in the framework of the International Year of Biodiversity 2010.

This document gives an overview of the technical aspects and solutions of the AntaBIF webportal as of June 2011. Some of these aspects are still under refinement/testing phase. The data flows and standards are described as well as possibilities for users to contribute to the network.

This document will be sent in priority to the members of the AntaBIF ISC, and will be posted in the shared repository.

Table of Contents

Websites and sub-domains

A series of website and subdomains give the AntaBIF project some webpresence, while the data portal is under development. The main website, giving general information about the project is available on two urls: www.biodiversity.aq and www.AntaBIF.be. The list below describes the different subdomains as well as a short description of their scope:

http://www.biodiversity.aq or http://www.AntaBIF.be : shopwindow website, provides basic information on AntaBIF, as well as a dynamic link to the AntaBIF blog. Accesspoint to the latest news and resources available from the network. The website is currently maintained/developed by the ANTABIF project manager using iWeb 3.0.3, and hosted on the Belgian Biodiversity Platform ULB-VUB servers.

http://share.biodiversity.aq : Filesharing webspace, offers free and open access to original datasets, documents, R-code, GIS layers, reports and other relevant resources. The content is managed by the project manager and the science officer. Website is hosted on the Belgian Biodiversity Platform ULB-VUB servers.

http://afg.biodiversity.aq: Interactive Antarctic Field Guides (AFG). The AFG are built by aggregating data from three sources: the AFG database (life history, multimedia content), the Register of Antarctic Marine Species (taxonomy) and the Global Biodiversity Information Facility (occurence records) using RESTish webservices.

gcmd.gsfc.nasa.gov/KeywordSearch/Home.do?Portal=AntaBIF&MetadataType=0: AntaBIF metadata portal, hosted by NASA. The GCMD hosts the metadata backbone of AntaBIF. Webservices are used to publish and edit metadata records in DIF (Directory Interchange Format) format.

http://ogc.biodiversity.aq: AntaBIF Geoserver (Geoserver 2.0.1) instance, allows to serve geospatial layers in many formats in a dynamic way. It is an important component of the AntaBIF architecture, and is designed to manage geospatial data of different types. The Geoserver will give access to the occurrence data as welll as environmental layers. This data will be served using OGC webservices. The Geoserver instance is hosted on the Belgian Biodiversity Platform ULB-VUB servers.

http://www.princesselisabeth.be: domain name for a portal dedicated for the Princess Elisabeth context of the AntaBIF databases. Development Framework (Mahara) in test.

http://ipt.biodiversity.aq: AntaBIF Integrated Publishing Toolkit instance, allows to serve original data to GBIF. IPT technology allows to publish data to GBIF and alerts it when new datasets are available for trawling using an RSS-based data casting system. The IPT instance will also be used to publish new data on the AntaBIF data portal.

Technical architecture

The AntaBIF technical integrates the experience gained in the framework of SCAR-MarBIN and SCAR-MarBIN V2, building upon webservice-oriented, dynamic, technologies. Careful attention is also taken to make the best possible use of GBIF informatics tools (HIT and IPT). Using this approach, the networks will be interoperable with many entities, allowing the publication of the data in many different contexts, in a distributed fashion.

The AntaBIF technical architecture is organized around the following components: data harvesting, data publishing, metadata, occurrence, taxonomy, genetics, search engine. Details are given below:

Overview

The technological ecosystem is largely based on the progress made in the framework of the SCAR-MarBIN V2 data portal. It is based on innovative, 100% Open Source solutions, and integrates the latest GBIF informatics components (Harvesting and Indexing Toolkit (HIT) and Integrated Publishing Toolkit (IPT)). Details are given in Table 1:

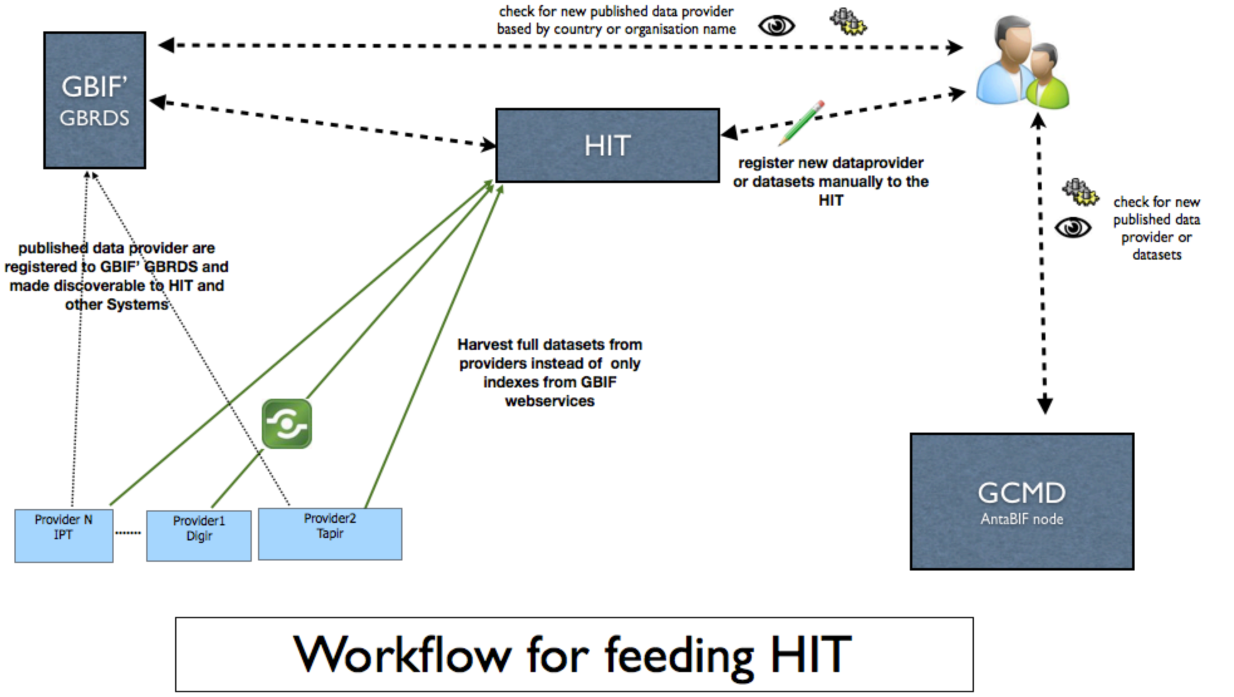
Table 1: description of technological solutions adapted for each component of the AntaBIF data system

Item	Technological solution
Design patterns	MVC (Model-View-Controller)/ORM (Object-Relational-Mapping
Framework	Rails (ActiveRecord) and YUI (User Interface Library)
Search engine	Full text (Elasticsearch-Lucene)
Language	Ruby
Database	PostGreSQL
GIS server	Geoserver
Spatial database	PostGIS
Mapping client	OpenLayers
Web services	RESTish (all resources)
Protocols	DIF (Data Interchange Format), DarwinCore, DarwinCore archive, Tapir
GBIF tools	HIT (Harvesting and Indexing Toolkit), IPT (Integrated Publishing Toolkit)
OS	FreeBSD
Hosting	BBPF (ULB/VUB joint IT Center)
Metadata systems	GCMD (Global Change Master Directory, mirrored), through API

Data harvesting component

To build the content of its databases, AntaBIF uses a customized version of GBIF's HIT (Harvesting Indexing Toolkit). The main changes brought to the original HIT is to use a geographic objects-supporting database solution (PostGreSQL/PostGIS) instead of the MySQL database, which is not spatially-enabled. Another customization to the HIT includes the possibility to create new datasets even if they haven't yet been registered in the Global Biodiversity Resources Discovery System (GBRDS). The data exchange protocols used by the custom version of the HIT include DiGIR and DarwinCore Archive (DWC-A). See Figure 1 for details.

Figure 1: diagram describing the workflow to feed AntaBIF's HIT instance.


check for new published data provider based by country or organisation name
GBIF'
GBRDS
HIT
register new dataprovider or datasets manually to the HIT
check for new published data provider or datasets
published data provider are registered to GBIF' GBRDS and made discoverable to HIT and other Systems
Harvest full datasets from providers instead of only indexes from GBIF webservices
GCMD
AntaBIF node
Provider N
IPT
Provider1
Digir
Provider2
Tapir
Workflow for feeding HIT

Data publishing component

AntaBIF will be using GBIF's IPT (Integrated Publishing Toolkit, IPT2) to publish original data to OBIS and GBIF. Original datasets (checklists or occurrence datasets) can be uploaded in the AntaBIF IPT instance, after it has been transformed to a DarwinCore Archive and that its consistency is checked using GBIF's DarwinCore Validator. The new data is then exposed using an RSS-based data casting system to GBIF's GBRDS which make it discoverable to other Systems as OBIS. For more details, see "Spatial data flow" section below.

Metadata component

AntaBIF metadata system will mirror the relevant Global Change Master Directory (GCMD) content, using an API (REST webservices). The GCMD data models which is based on DIF standard will be integrated in the AntaBIF database as a cache, linking existing datasets and new ones to the relevant metadata records.

Occurrences (Geospatial) component

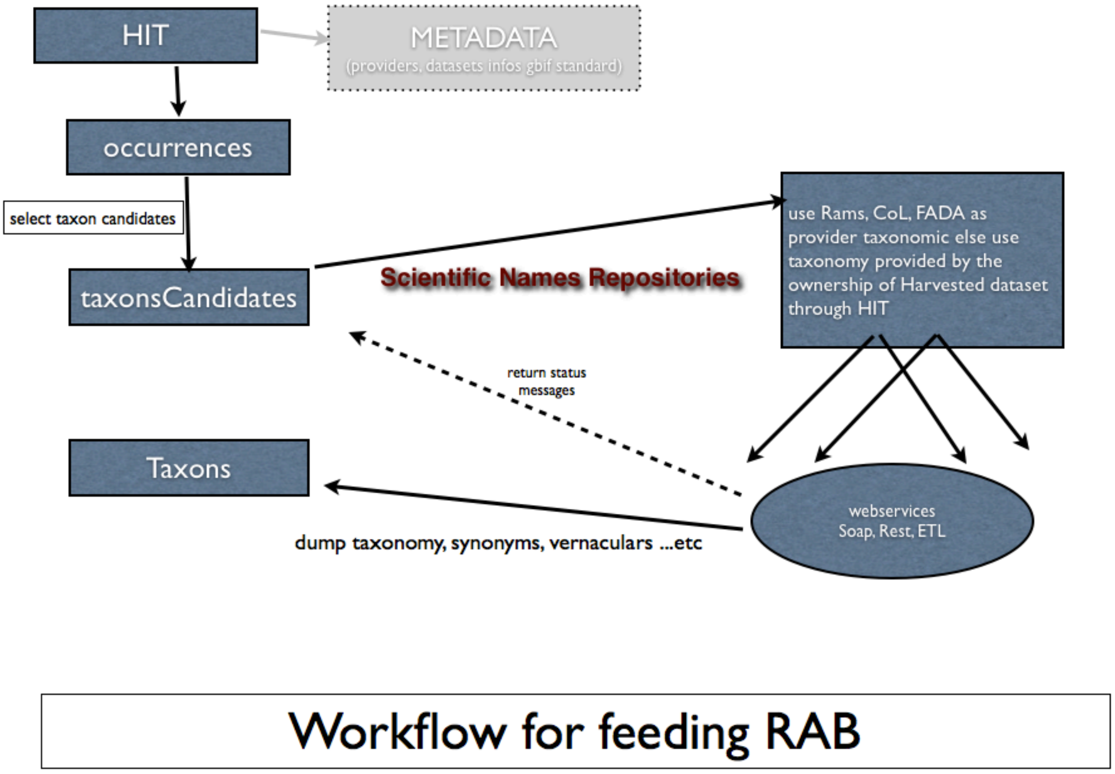
The AntaBIF geospatial component includes the use of PostGRES, PostGIS, Geoserver and OpenLayers. The DarwinCore standard is used for the occurrences model that maps harvested occurrences records.

The ETL (Extract, Transform, Load) tools are developed incrementally for contexts and scopes extraction (taxonomy candidates, temporal data, geopatial data) .

Taxonomic backbone

The taxonomic backbone of AntaBIF is based on the occurrence model, generating a list of candidate names used to trigger taxonomic information retrieval. The AntaBIF taxonomic component (RAB) is designed as a pointer to other taxonomic data providers, including direct access to a minimal amount of locally stored data. The approach will be using a client consuming the Webservices for Rams, Col and FADA. Also a taxonomic candidates extractor tool from occurrences as well as client web-services for other taxonomic data providers will be used. See Figure 2 for details.

Figure 2. Diagram describing the workflow to feed AntaBIF's taxonomic component (RAB)


HIT
METADATA
(providers, datasets infos gbif standard)
occurrences
select taxon candidates
taxonsCandidates
Scientific Names Repositories
use Rams, CoL, FADA as provider taxonomic else use taxonomy provided by the ownership of Harvested dataset through HIT
return status messages
Taxons
webservices
Soap, Rest, ETL
dump taxonomy, synonyms, vernaculars ...etc
Workflow for feeding RAB

Genetics component

AntaBIF aims at integrating DNA-related information related to taxonomy. Webservices will be used as well to retrieve DNA sequences in Blast format from GenBank. This component will be suing client web-services to retrieve the DNA data form NCBI, establishing a relation to relevant RAB taxon as well as local DNA database for the taxa for which only DNA-based taxonomy is available from GenBank tools.

Search engines component

The AntaBIF search engine can search full text content in all AntaBIF databases, within any context (taxonomy, DNA, spatial, temporal, metadata).

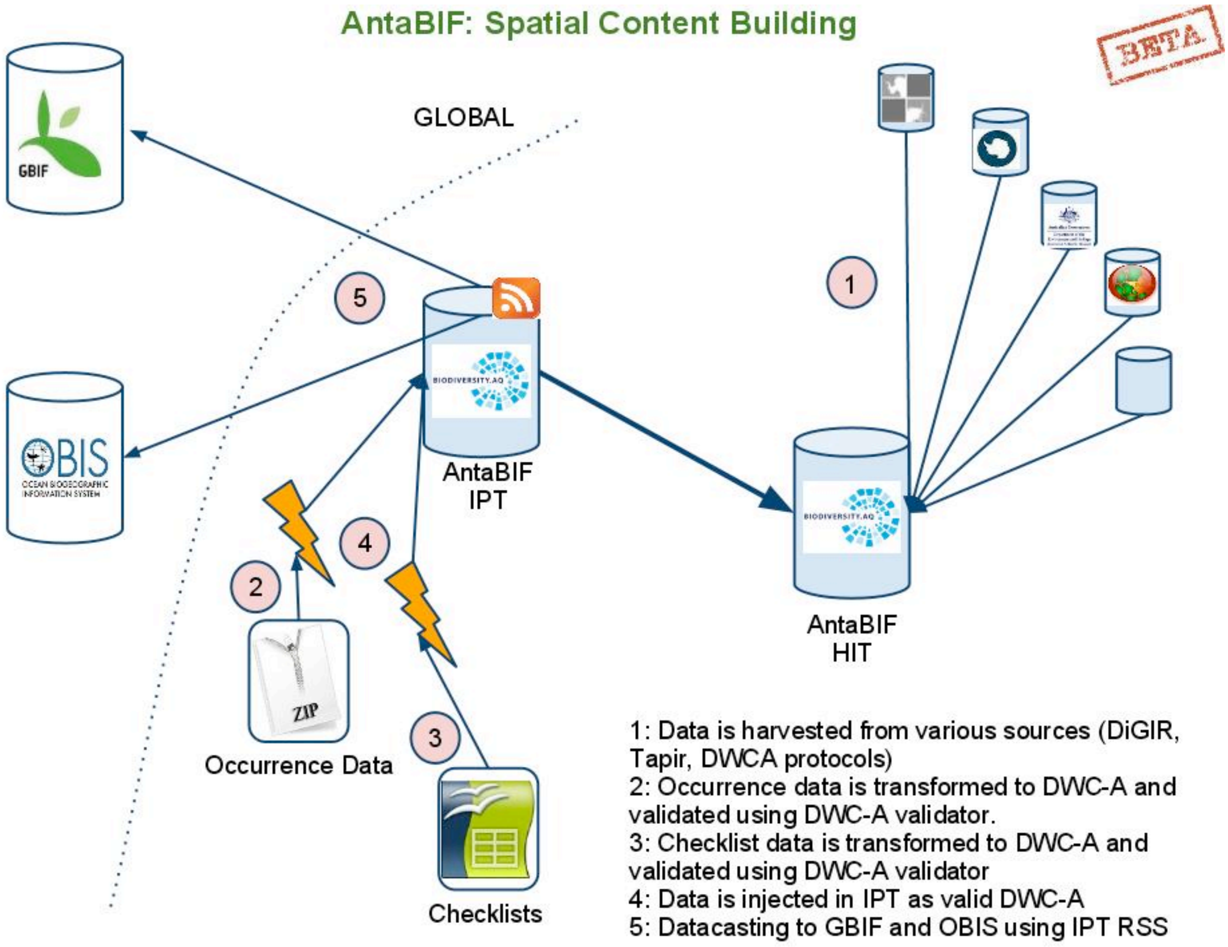
The component will be using an index server (Elasticsearch node). The Elastic search is a distributed Lucene-based search engine with sharding the index itself and using map/reduce to search on nodes. It has RESTful API with JSON standard as wire protocol.

The search model is domain driven. Specific Ruby Scripts are developed to create, update, index objects, based on the database scopes/contexts. This approach allow AntaBIF to maintain the integrity of its data models and serve them Restfully.

Spatial Data flows

The data flow envisioned for AntaBIF occurence data is described in Figure 3: AntaBIF's HIT is capable of harvesting data from various sources using many different protocols, including DiGIR, TaPIR, IPT. AntaBIF has its own IPT instance (http://ipt.biodiversity.aq/), which pushes data to AntaBIF's HIT and to other information systems (OBIS and GBIF). Before it is uploaded in the IPT, new data is cleaned and checked for consistency using GBIF's DarwinCore validator.

Figure 3: diagram representing the content building for spatial data, and interaction with other information systems.


AntaBIF: Spatial Content Building
BETA
GLOBAL
GBIF
OBIS
OCEAN BIOGEOGRAPHIC INFORMATION SYSTEM
BIODIVERSITY.AQ
AntaBIF
IPT
AntaBIF
HIT
1
2
3
4
5
ZIP
Occurrence Data
Checklists
1: Data is harvested from various sources (DiGIR, Tapir, DWCA protocols)
2: Occurrence data is transformed to DWC-A and validated using DWC-A validator.
3: Checklist data is transformed to DWC-A and validated using DWC-A validator
4: Data is injected in IPT as valid DWC-A
5: Datacasting to GBIF and OBIS using IPT RSS

Standards

The standards adopted for the AntaBIF systems are described below, in function of the different data types.

Occurrence data (harvesting and publishing)

The Darwin Core standard has been used to mobilise the vast majority of specimen occurrence and observational records within the GBIF network. The Darwin Core standard was originally conceived to facilitate the discovery, retrieval, and integration of information about modern biological specimens, their spatio-temporal occurrence, and their supporting evidence housed in collections (physical or digital). The Darwin Core achieved this by defining a set of items in an ordered list, published in an XML document.

The Darwin Core today is broader in scope. It aims to provide a stable, standard reference for sharing information on biological diversity. As a glossary of terms, the Darwin Core provides stable semantic definitions with the goal of being maximally reusable in a variety of contexts. This means that Darwin Core may still be used in the same way it has historically been used, but may also serve as the basis for building more complex exchange formats, while still ensuring interoperability through a common set of terms. One such exchange format is defined in the GBIF Integrated Publishing Toolkit, which allows for the definition of multiple extensions to a core 'taxon occurrence' or species entity. These extensions provide a means of serving multiple identifications to a specimen, multiple images of a specimen or multiple common names to a taxon concept. This is now possible due to the broadening of scope of the Darwin Core and a redefinition of its structure into a reusable glossary of terms.

The preferred format for publishing data to the AntaBIF network is the Darwin Core Archive (DwC-A), which is essentially a set of text (e.g., TAB or CSV) files with a simple descriptor to inform others how your files are organized. The format is defined in the Darwin Core text guidelines.

The central idea of this archive is that its data files are logically arranged in a star-like manner, with one core data file surrounded by any number of 'extensions'. Each extension record (or 'extension file row') points to a record in the core file; in this way, many extension records can exist for each single core record.

Metadata

The ISO 19115/TC211 geospatial metadata standard was adopted June 2004. Required elements and appropriate modifications were approved by the CEOS IDN Interoperability group and incorporated into the DIF to achieve full ISO compatibility.

The DIF does not compete with other metadata standards. It is simply the "container" for the metadata elements that are maintained in the IDN database, where validation for mandatory fields, keywords, personnel, etc. takes place.

The DIF is used to create directory entries which describe a group of data. A DIF consists of a collection of fields which detail specific information about the data. Eight fields are required in the DIF; the others expand upon and clarify the information. Some of the fields are text fields, others require the use of controlled keywords (sometimes known as "valids").

The DIF allows users of data to understand the contents of a data set and contains those fields which are necessary for users to decide whether a particular data set would be useful for their needs.

The DIFGuide document provides information about each field of the DIF, including its syntax, specifications, recommendations, and examples. Several example DIFs are also provided.

Geospatial data

OGC is the Open Geospatial Consortium, the leading standards organization in the Geospatial arena. Somewhat like the W3C, but focused on geospatial applications. The Web Map Server (WMS) specification defines a standard protocol for generating cartographic maps over the web. The Web Feature Server (WFS) defines a standard protocol for querying and retrieving vector features over the web. WMS and WFS are two of the most prominent OGC specifications. GeoServer is the reference implementation of the WFS spec, and also fully implements the WMS spec.

Contributing to AntaBIF

Scientists, Institutes or organizations wishing to publish their data on AntaBIF networks have the possibility to use a range of existing tools, ensuring the data is interoperable with other biodiversity information networks, and adapted to the level of IT competence/infrastructure available to the data provider.

IPT2

The vision of AntaBIF is to provide its data providers with tools which help shift the network architecture towards a distributed infrastructure. An essential tool to reach this logical infrastructure is GBIF's IPT. GBIF developed the IPT as a software platform to facilitate the efficient publishing of biodiversity data on the Internet, using the GBIF network. The IPT is the recommended solution by the AntaBIF network for publishers who can meet the technical requirements to install an IPT instance.

Template Spreadsheet (direct contribution) and Data cleaning tools

Data providers have the possibility to email spreadsheets to AntaBIF, using spreadsheet templates (.xlsx, .ods formats). The spreadsheets are similar to the GBIF template spreadsheets, listed below:

Metadata - Users can fill in this template to describe a database or other data resource. [XLSX format]

Species Occurrence - Users can fill this template to record or store basic species collections or observational data. [XLSX format]

The spreasheet can be processed using the GBIF spreadsheet processor tool, and cleaned using the GBIF DarwinCore Archive validator.

DiGIR

While the IPT is the recommended way of publishing data through the AntaBIF network, there are other publishing tools available from the wider community. Distributed Generic Information Retrieval (DiGIR) is a legacy tool, but still largely used in the AntaBIF community (eg by SCAR-MarBIN, AADC, BAS, Pangaea,...). Thus, DiGIR will continue to be harvested and indexed by the AntaBIF core infrastructure.