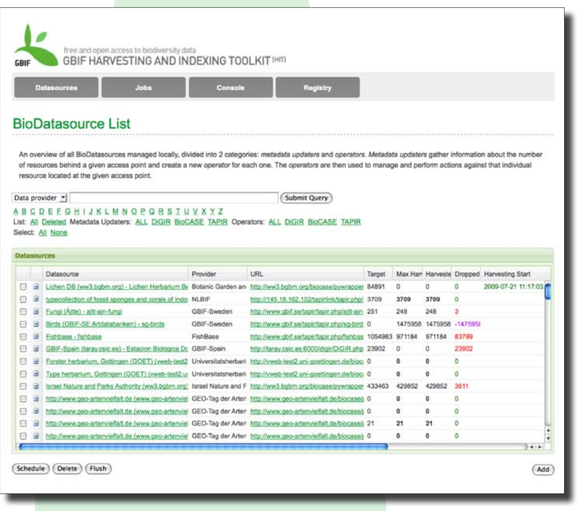
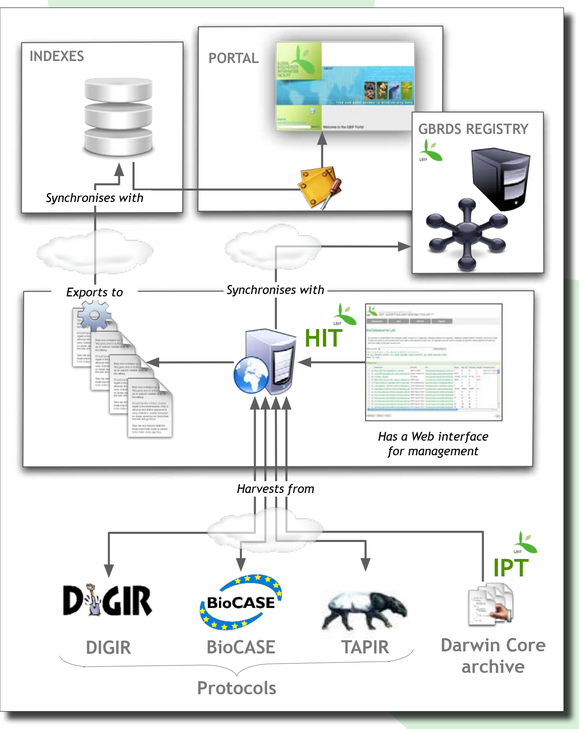
The GBIF Harvesting and Indexing Toolkit (HIT)



A screenshot of a HIT instance installed in GBIF showing the user-friendly dashboard from which to manage all harvesting activities



An overview of the data flows from data publishers to the raw indexes using the HIT

Resources

http://code.google.com/p/gbif-indexingtoolkit/

Source for GBIF HIT documentation, downloads, source code, bug reporting, etc.

Overall Description

The Global Biodiversity Information Facility (GBIF) aims to be the preferred gateway, worldwide, to a comprehensive, distributed array of primary species-occurrence data. In moving 'towards full operation' of a fully distributed network architecture, the key focus in portal design was to enable customisation by GBIF Participant Nodes to their local needs, through appropriate and user-friendly tools. In 2009, particular focus was given on simplifying the process of publishing data as well as to improve the frequency of data indexing. The GBIF Harvesting and Indexing Toolkit (HIT) is a software platform developed by GBIF (http://www.gbif.org/) to manage biodiversity data harvesting and quickly build indexes of the harvested data.

Overview

The HIT is capable of harvesting data from data publishers exposing their data through three protocols: Distributed Generic Information Retrieval (DiGIR)[1], Biological Collection Access Service (BioCASe)[2], and TDWG Access Protocol for Information Retrieval (TAPIR)[3]. It can also harvest data directly from a single export, created in accordance with the new Darwin Core terms[4] as a dump in Archive format[5] using the Integrated Publishing Toolkit (IPT)[6]. By accessing all data publishers through a single tool, regardless of the protocol used, the HIT provides a convenient mechanism to coordinate and manage indexing operations and scheduling. Anybody wanting to mobilise data from several data publishers will find this increasingly beneficial as their list of publishers continues to grow.

Features

HIT is an open source (Apache 2.0 license) Java based, customisable, multilingual web application that:

- Synchronises with the GBIF Universal Description, Discovery and Integration UDDI registry
- Harvests three types of protocols: DiGIR, BioCASe, and TAPIR; extensible to others
- Harvests from the Darwin Core Archive format
- Tracks activity with output log messages, filterable by provider or dataset
- Displays the complete list of data publishers and their datasets, filterable by provider name, dataset name, and country name, displaying statistics, etc.
- Displays the complete list of operations currently scheduled
- Allows the in-browser viewing of each individual XML request or response sent as part of the various operations
- Synchronises with one or more external databases
- Generates an index of the harvested data

1 http://digir.net/
2 http://www.biocase.org/products/protocols/
3 http://www.tdwg.org/activities/tapir/
4 http://rs.tdwg.org/dwc/index.htm
5 http://rs.tdwg.org/dwc/terms/guides/text/index.htm
6 http://code.google.com/p/gbif-providertoolkit/

Kyle Braak (kbraak@gbif.org)
Andrea Hahn (ahahn@gbif.org)
Samy Gaiji (sgaiji@gbif.org)
Tim Robertson (trobertson@gbif.org)
Markus Döring (mdoering@gbif.org)

Global Biodiversity Information Facility (GBIF)
Universitetsparken 15
2100 Copenhagen
Denmark

GBIF
www.gbif.org
...free and open access to biodiversity data

About GBIF

GBIF makes digital biodiversity **data openly and freely available** on the Internet for everyone, and endorses both open source software and open data access.
www.gbif.org

GBIF provides **scientific biodiversity data** for decision-making, research endeavours and public use.
data.gbif.org

GBIF is a network of **data publishers who retain ownership and control** of the data they share. Linked datasets provide a more robust representation of biodiversity than any single dataset.

GBIF provides access to primary biodiversity data held in institutions in developed and developing countries. **Data shared through GBIF are repatriated data**.

GBIF is a **dynamic, growing partnership** of countries, organisations, institutions and individuals working together to mobilise scientific biodiversity data.

GBIF invites you to download species occurrence data freely and openly from **http://data.gbif.org**

GBIF invites you to **join** the GBIF network **and share** your biodiversity data, as well as participate in developing **new tools and services.**

2010 International Year of Biodiversity

DESIGNGRAFIK.DK / PHOTO: PhotoDisc, Jon Sullivan, Derek Roberston, Else Østergaard Andersen, Anne Mette Nielsen / PRINT: PosterPrint