GBIF

GBIF Summary Paper:
The GBIF Harvesting and Indexing Toolkit (HIT)

DESCRIPTION

The GBIF Harvesting and Indexing Toolkit (HIT) is a software platform developed by the Global Biodiversity Information Facility (http://www.gbif.org/) to manage biodiversity data harvesting and quickly build indexes of the harvested data.

This Java-based tool can harvest data from data publishers publishing their data through three protocols: DiGIR (http://digir.net/), BioCASe (http://www.biocase.org/products/protocols/), and TAPIR (http://www.tdwg.org/activities/tapir/). It can also harvest data directly from a single dump, created in accordance with the new Darwin Core Archive standard (http://rs.tdwg.org/dwc/terms/guides/text/index.htm). Built in such a way that it is fully extensible, the HIT allows support for additional protocols to be easily incorporated at a later time.

Because the various protocols differ from one another considerably, harvesting data from a heterogeneous set of data publishers can be quite complicated. The HIT's interface is effective at hiding these differences from the user, so that the same set of operations is presented to the user regardless of which underlying protocol is being used. By grouping data publishers together in one place, regardless of their protocol, the HIT provides a convenient mechanism whereby they can all be managed. Nodes needing/wanting to mobilise data from several data publishers, for example, will find this increasingly beneficial as their list of publishers continues to grow.

Another challenge encountered when harvesting is that it is extremely time-intensive for large datasets and that it requires uninterrupted communication with the provider. The HIT bypasses these problems by allowing harvesting operations to be automatically scheduled, and allowing them to restart from the point where they fail instead of beginning anew. Moreover, in most cases the HIT is capable of identifying the reason why an operation fails, responding with this information to the user. Whether it is a connectivity problem, or an issue with the way that the provider has structured their data, the user can quickly determine which course of action must be taken to enable harvesting to continue.

Once harvesting has been completed, the HIT displays relevant statistics such as the number of records successfully harvested and the number dropped. The user can then compare these numbers with the total number of records that were expected in order to determine which data publishers/datasets require more attention and which ones can continue to be harvested in an automated way. Ultimately the HIT aims to allow for a situation where all harvesting becomes a fully automated procedure, saving a great deal of time.

One of the main technical challenges of the GBIF decentralised architecture is to allow Nodes to aggregate data themselves which can then be made available back to GBIF. The HIT surmounts this challenge by allowing the harvested data to be synchronised with a relational database, and for an index of the data to be built. Once this has been done, it would be possible to build a regional or thematic portal on top of the database, providing the community with an additional means of accessing the data.

GBIF Summary Paper: The GBIF Harvesting and Indexing Toolkit

Entirely open source and extensible, the HIT is intended to grow as new protocols emerge. It is expected that the benefits to both the biodiversity data holders and users are significant, making the process of data mobilisation an easier one that people are keener to engage in.

SPECIFICATIONS

GENERAL
- Open source (Apache 2.0 license) Java based, customisable multilingual web application
- Synchronisation with the GBIF UDDI registry
- Harvesting of 3 types of protocols: DiGIR, BioCASe, and TAPIR; extensible to others.
- Harvesting from the Darwin Core Archive format
- Internationalised output log messaging.
- Role-based user management, allowing for multiple data harvesters to share a common instance.
- An interface displaying the complete list of data publishers and their datasets, filterable by provider name, dataset name, and country name, displaying statistics, etc.
- An interface displaying the complete list of operations currently scheduled.
- An interface displaying the output log messages, filterable by provider or dataset.
- Allows the in-browser viewing of each individual xml request or response sent as part of the various operations.
- Automatic scheduling of operations.
- Synchronisation with one or more external databases.
- Generation of an index on the harvested data

REQUIREMENTS

Pre-requisites for installing the software:
- Web server with a Servlet container (such as Tomcat or Jetty) connected to the Internet.
- Java version 5.

SCHEDULE

- GBIF HIT internal alpha version release: 30 April 2009.
- GBIF HIT semi-public alpha release 31 May 2009.
- GBIF HIT public 1.0 beta release: 15 August 2009.
- Software evaluation and bug-fixing September-December 2009
- GBIF HIT version 1.1 release: December, 2009.

RESOURCES

1. **http://code.google.com/p/gbif-indexingtoolkit/**
   Project Home Site: documentation, downloads, source code, bug reporting, etc.
2. **http://www.gbif.org/**
   GBIF communications portal.
3. **http://www2.gbif.org/WP2009-10.pdf**
   GBIF Work Programme 2009-2010

GBIF CONTACTS

Head of Informatics	**Systems Architect**
SAMY GAIJI	TIM ROBERTSON
sgaiji@gbif.org	trobertson@gbif.org
Developer	**Training Officer**
Kyle Braak	ALBERTO GONZÁLEZ-TALAVÁN
kbraak@gbif.org	atalavan@gbif.org

GBIF Summary Paper: The GBIF Harvesting and Indexing Toolkit