GBIF Summary Paper:
The GBIF Integrated Publishing Toolkit (IPT)

DESCRIPTION

The GBIF Integrated Publishing Toolkit (IPT) is a software platform developed by the Global Biodiversity Information Facility (http://www.gbif.org/) to facilitate efficient biodiversity data publishing onto the Internet. This Java-based tool manages three types of data: taxon primary occurrence data, taxonomic checklists and resource metadata. Through the IPT, users will be able to publish data residing in local databases, upload existing files in comma-separated and tab-delimited formats, and access central services to make use of standardised controlled vocabularies and data quality reporting services. Extensions such as common names, species descriptive data or multiple identifications may be defined and shared amongst the IPT user community. The IPT also acts as a visualisation tool, as it includes a web application allowing for browsing and searching of published data, along with basic mapping services.
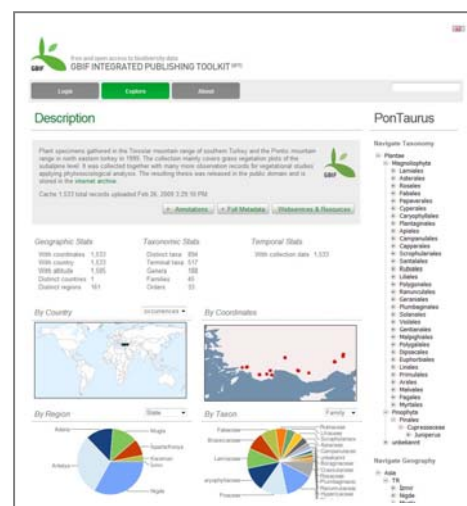
One of the main technical challenges of the GBIF decentralised architecture is to remove bottlenecks within data flow. The IPT offers interfaces to transfer complete resource archives efficiently in order to reduce the latency between data publication and discoverability through the GBIF indexes. Currently, much of the quality analysis is performed on the data after it is published onto the network. The IPT aims to provide data holders with the ability to easily run a growing number of quality checks against their data, before any data aggregators index the source. By decentralising the quality routines, it is expected that the data will become fit for a wider number of uses more quickly than possible with central quality analyses.

With an embedded database and web application, the IPT is an open platform to build upon, offering additional services and benefits both to the biodiversity data holders and users.

SPECIFICATIONS

GENERAL
- Open source (Apache 2.0 license) Java based customisable, multilingual web application.
- Connects and serves 3 types of data: taxon primary occurrence data, taxon checklists and general resource metadata. No manual data entry accepted (except for metadata).
- Manages multiple data sources (with individual metadata descriptions).
- Role-based user management, allowing for multiple data curators to share a common instance.
- Web interface with data browsing, full text search and individual pages for the three types of data, displaying features available for each type (i.e. density maps, charts, metadata).
- Several upload options: Relational Database Management System (RDBMS), tab files.
- Manages Universally Unique IDentifiers (UUID): recycles existing IDs or provides new ones, as appropriate.
- Manages technical metadata about the installation.
- Easy setup of Google Analytics™ for usage statistics.

TAXON PRIMARY OCCURRENCE DATA MANAGEMENT
- Data model with a core table using DarwinCore (DwC) terms and extension tables in a "star" schema.
- Data validation routines included via local files or central GBIF services for names, countries, basis-of-record, etc.

- Interfaces: URL based resolution, Open Geospatial Consortium (OGC) Web Feature Service (WFS) and Web Map Service (WMS), TAPIR *intermediate* search and inventory, simple REST web services and complete resource available for download as a DwC based archive file.

TAXONOMIC CHECKLIST

- Data model with a core table using DarwinCore (DwC) terms with extension tables in a "star" schema.
- Structure allows for modelling taxonomic hierarchy, synonymy and nomenclature notes, support for multi-lingual vernacular names and URLs associated with external sources for full taxonomic information.
- Interfaces: single TCS-RDF document, URL-based taxon resolution, complete resource available for download as a DwC based archive file.

RESOURCE METADATA

- Manages metadata information for every resource connected to the IPT (taxon primary occurrence datasets and names checklists) and any other external resource defined by the user but not published through the IPT.
- Data model based on a profile of the Ecological Metadata Language (EML) allowing for definition of taxonomic, geospatial and temporal scope, along with rights and citation, contact information and keywords.
- Interfaces: EML data file. RSS feed for most recently changed/created resources.

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 IPT public beta version release: 30 January 2009.
- GBIF IPT version 1.0 release: 31st March 2009.
- Software evaluation and bug-fixing May-October 2009
- New requirements gathering May-October 2009
- GBIF IPT version 1.1 release: October 2009.
- Review of new requirements October-November 2009
- GBIF IPT version 2.0 release: December 2009.

RESOURCES

1. **http://code.google.com/p/gbif-providertoolkit/**
   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
SAMY GAIJI
sgaiji@gbif.org

Systems Architect
TIM ROBERTSON
trobertson@gbif.org

Senior Developer
MARKUS DÖRING
mdoering@gbif.org

Training Officer
ALBERTO GONZÁLEZ-TALAVÁN
atalavan@gbif.org