Using the Harvesting and Indexing Toolkit (HIT) in Special Interest Networks (SINs)

Kelbert, P., Droege, G., Holetschek, J. & Güntsch, A.

Botanic Garden and Botanical Museum Berlin-Dahlem
Freie Universität Berlin
• Harvesting and Indexing Toolkit
  – Developed by GBIF
    • Several adaptations done by different institutions -> different versions exist
  – Users: several GBIF national nodes, GGBN, OpenUp!, BGBM etc.
  – Can handle ABCD, DwC, DwC-A
  – Java & MySQL
Background

• BGBM loves SINs
  – Technical node of Global Genome Biodiversity Network (GGBN)
  – Technical support for BioCASe/ABCD
  – Hosting several other SIN portals (national and international)
  – BiNHum (Biodiversity Network for the Humboldt-Ring, funded by DFG)
    • 6 museums/research institutions in Germany, one shared BiNHum portal
→ use HIT with several extensions to handle complexity of SINs
Principal Harvesting Workflow - HIT

- Provider / Dataset
- (Correction)
- Registration And Harvesting
- Feedback to the provider
- Quality tests / Data cleaning
Principal Harvesting Workflow - HIT

Provider / Dataset

Supports: ABCD 2.06, DwC-A, DwC
Extended for:
- ABCDEFG
- ABCD 2.1
- ABCD archives
- GGBN Data Standard (ABCD and DwC-A)*

(Correction)

Feedback to the provider
Quality tests / Data cleaning

*presentation in S03, today 11-12.30
Principal Harvesting Workflow - HIT

BioDatasource List

An overview of all BioDatasources managed locally, divided into 2 categories: metadata update and operators. Metadata update gather information about the number of resources behind a given access point each one. The operators are then used to manage and perform actions against that individual resource located at the given access point.


List: [ ] All [ ] Recently Added [ ] Deleted [ ] Metadata Updaters: [ ] DiGiR [ ] BioCASE [ ] TAIPIR [ ] Dwc Archive [ ] Operators: [ ] DiGiR [ ] BioCASE [ ] TAIPIR [ ] Dwc Archive

Select: [ ] All [ ] None

Datasources

<table>
<thead>
<tr>
<th>Data source</th>
<th>Provider</th>
<th>URL</th>
<th>Target</th>
<th>Harvesting Start</th>
<th>Last Inventory</th>
<th>Last Harvested</th>
</tr>
</thead>
<tbody>
<tr>
<td>BoBo</td>
<td>BGBM</td>
<td>http://</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GBIF - D Protoreza - GBIF - D Protezr data collection</td>
<td>BGBM</td>
<td>http://</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GBIF - D Protoreza</td>
<td>BGBM</td>
<td>http://</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SMNS Herpetology I - BihU Herpetology I</td>
<td>SMNS</td>
<td>http://</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SMNS Herpetology I</td>
<td>SMNS</td>
<td>http://</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Herbarium all associated BGBM - Herbarium Berolinense</td>
<td>BGBM</td>
<td>http://</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AlpTerrafy</td>
<td>BGBM</td>
<td>http://</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AlpTerrafy - AlpTerra Types</td>
<td>BGBM</td>
<td>http://</td>
<td>7865</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AlpTerrafy - AlpTerra Types</td>
<td>BGBM</td>
<td>http://</td>
<td>7865</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AlpTerrafy</td>
<td>BGBM</td>
<td>http://</td>
<td>23</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AlpTerrafy</td>
<td>BGBM</td>
<td>http://</td>
<td>23</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AlpTerra - AlpTerra</td>
<td>BGBM</td>
<td>http://</td>
<td>3099</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Herbarium all associated BGBM</td>
<td>BGBM</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Displaying 1 - 100 of 120
Principal Harvesting Workflow - HIT

Extended for:
- Associations between records (ABCD, DwC-A)
- Multiple identifications per record
- Multiple multimedia urls per record
- Measurement Or Fact
- Harvesting of user-defined filter or list of records
- Storage in (extended) MySQL database

Registration And Harvesting

Feedback to the provider

Quality tests / Data cleaning
### Principal Harvesting Workflow - HIT

#### BioDatasource List

An overview of all BioDatasources managed locally, divided into 2 categories: metadata updaters and operators. Metadata updaters gather information about the number of resources behind a given access point for each one. The operators are then used to manage and perform actions against that individual resource located at the given access point.

Blabla

<table>
<thead>
<tr>
<th>Data provide</th>
<th>search</th>
</tr>
</thead>
</table>

**List:**
- All
- Recently Added
- Deleted
- Metadata Updaters: All, DiGIR, BioCASE, TAPIR, Dwc Archive
- Operators: All, DiGIR, BioCASE, TAPIR, Dwc Archive

**Select:**
- All
- None

<table>
<thead>
<tr>
<th>Datasources</th>
<th>Associated from</th>
<th>Provider</th>
<th>URL</th>
<th>Target</th>
<th>Harvest</th>
<th>Harvesting Start</th>
<th>Last Inventory</th>
<th>Last Inventory Process</th>
<th>Last Harvested</th>
</tr>
</thead>
<tbody>
<tr>
<td>BSMeryscol - The Erysiphales Collection at the Botanic Gardens of the University of Amsterdam (BM)</td>
<td>BSMeryscol</td>
<td>SNSB</td>
<td>http://...</td>
<td>11007</td>
<td>11003</td>
<td>2014-07-21 11:45:34</td>
<td>2014-07-21</td>
<td>2014-07-21</td>
<td></td>
</tr>
</tbody>
</table>

**View Logs, Name-ranges, Rolling log**

**Displaying 1 - 9 of 9**
New:
- Original values are kept in the database
- Cleaned values are stored in extra tables
  - Geography, Coordinates (Gisgraphy, Geonames)
  - Country translation
  - Coordinates validity
  - ISO-code vs. Country
  - ISO/Country vs. Coordinates
  - Waterbodies extraction from locality/gatheringarea/country
- Name parsing (GBIF parser plus further algorithms)
- Multimedia URL validity
- Visualisation

Feedback to the provider

Quality tests / Data cleaning
Principal Harvesting Workflow - HIT

Results from the quality tests

Choose a datasource:

Show 10 entries

<table>
<thead>
<tr>
<th>Datasource</th>
<th>Provider</th>
</tr>
</thead>
<tbody>
<tr>
<td>Algata - Algata</td>
<td>BGBM</td>
</tr>
<tr>
<td>AlgaterraTypen - Algaterra Types</td>
<td>BGBM</td>
</tr>
<tr>
<td>Algaterravideos</td>
<td>BGBM</td>
</tr>
<tr>
<td>Animal Sound Archive - Animal Sound Archive</td>
<td>MN</td>
</tr>
<tr>
<td>BIOTA Southern Africa - BIOTA Southern Africa - The Collection of Lichens at the Botanische Staatsammlung München</td>
<td>SNSB</td>
</tr>
<tr>
<td>BoBo - BoBO - Botanic Garden and Botanical Museum Berlin-Dahlem Observations</td>
<td>BGBM</td>
</tr>
<tr>
<td>BSMeryscoll - The Erysiphales Collection at the Botanische Staatsammlung München</td>
<td>SNSB</td>
</tr>
<tr>
<td>BSMungicoll - The Fungal Collection at the Botanische Staatsammlung München</td>
<td>SNSB</td>
</tr>
<tr>
<td>BSMgrossebrcoll - The Fungal Collection of Helga Große-Brauckmann at the Botanische Staatsammlung München</td>
<td>SNSB</td>
</tr>
</tbody>
</table>

Start/Run Quality tests:
- Run all tests

Quality tests:
- See country quality
- See coordinates quality
- See date quality
- See coordinates vs. country
- See association statistics

Quality tests - export:
- Export country quality
- Export country vs. isocode quality
- Export coordinates quality
- Export date quality
- Export coordinates vs. country
- Export association statistics
- Export missing units
- Export empty multimedia
- Export names statistics
- Export all test results
Results from the quality tests

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Test</th>
<th>Result</th>
<th>Amount of occurrences concerned</th>
</tr>
</thead>
<tbody>
<tr>
<td>Animal Sound Archive - Animal Sound Archive</td>
<td>Coordinates validity</td>
<td>FAILED</td>
<td>15</td>
</tr>
<tr>
<td>SMNSStuttgartHerbar - Staatliches Museum für Naturkunde Stuttgart, Herbarium</td>
<td>Coordinates validity latitude</td>
<td>FAILED</td>
<td>161</td>
</tr>
<tr>
<td>Herbar all associated BGBM - Herbarium Berolinense</td>
<td>Coordinates validity latitude</td>
<td>FAILED</td>
<td>28</td>
</tr>
<tr>
<td>BSMichfungicoll - The Collection of Lichenicolous Fungi at the Botanische Staatsammlung München</td>
<td>Coordinates validity latitude</td>
<td>FAILED</td>
<td>4</td>
</tr>
<tr>
<td>Lichen - Lichen Herbarium Berlin</td>
<td>Coordinates validity longitude</td>
<td>FAILED</td>
<td>7</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Original value</th>
<th>Supplemental info</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>10.2,-183.9167</td>
<td>Coordinates validity longitude</td>
<td></td>
</tr>
<tr>
<td>10.4333,-184.05</td>
<td>Coordinates validity longitude</td>
<td></td>
</tr>
<tr>
<td>38.1607,-640.1667</td>
<td>Coordinates validity longitude</td>
<td></td>
</tr>
<tr>
<td>9.9833,-183.7167</td>
<td>Coordinates validity longitude</td>
<td></td>
</tr>
</tbody>
</table>
Principal Harvesting Workflow - HIT

- Generation of CSV files:
  - One file per quality test:
    - original value
    - cleaned value
    - log/explanation
    - concerned UnitIDs

<table>
<thead>
<tr>
<th>Test/Tested on</th>
<th>Value before</th>
<th>Value after</th>
<th>Suggestion or log</th>
<th>UnitIDs</th>
</tr>
</thead>
<tbody>
<tr>
<td>isocountrycode</td>
<td>Afghanistan-AFG</td>
<td>Afghanistan-AF</td>
<td>isocode replaced AFG by AF</td>
<td>171646 --- 171648 ---</td>
</tr>
<tr>
<td>isocountrycode</td>
<td>Armenia-ARM</td>
<td>Armenia-AM</td>
<td>isocode replaced ARM by AM</td>
<td>107684</td>
</tr>
</tbody>
</table>

Feedback to the provider

Quality tests / Data cleaning
Data enrichment for HIT done by ZFMK/BGBM

• Data enrichment implemented
  – Red List (csv list)
  – Common Names (web service NHM Vienna)

Coming soon:
  – GBIF Checklist bank (web service GBIF)
  – GGBN records (web service GGBN)
Harvesting MySQL

- HIT

SOLR indexing

- To increase performance; optional

Portals

- BiNHum
- GGBN (new portal release 11/2015)
- Virtual Herbarium Germany (migration planned)
- Algae & Protists (migration planned)
- BGBM (migration planned)
- ... etc. ...
Conclusion

Source Code available at:
http://ww2.biocase.org/svn/synthesys/trunk/BinHum/

Paper about extended HIT is work in progress