Project no. 018340

Project acronym: EDIT

Project title: Toward the European Distributed Institute of Taxonomy

Instrument: Network of Excellence

Thematic Priority: Sub-Priority 1.1.6.3: “Global Change and Ecosystems”

**C6.5 Test manuscripts for XML data mining**

Due date of component: Month 15
Actual submission date: Month 15

Start date of project: 01/03/2006 duration: 5 years

Organisation name of lead contractor for this component: NHML

---

**Project co-funded by the European Commission within the Sixth Framework Programme (2002-2006)**

<table>
<thead>
<tr>
<th>Dissemination Level</th>
<th>PU</th>
<th>PP</th>
<th>RE</th>
<th>CO</th>
</tr>
</thead>
<tbody>
<tr>
<td>Public</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Restricted to other programme participants (including the Commission Services)</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Restricted to a group specified by the consortium (including the Commission Services)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Confidential, only for members of the consortium (including the Commission Services)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
GoldenGATE evaluation overview

Contact: Julius Welby (julius.welby@gmail.com).

Background

This report is the outcome of a 12 week contract (ending 8 June 2007) at the Natural History Museum to evaluate the GoldenGATE software package (GG), looking at its potential use as a means of translating existing taxonomic literature into one or more XML formats. The goal is to assess GG's suitability for converting documents into a marked-up format for addition to the XML repository which is part of the planned infrastructure of Work Package 6 of the EDIT project.

The full report is available at: http://www.editwebrevisions.info/node/60.

A number of people of people have provided assistance and input for this report, particularly Dave Roberts, Vince Smith, Simon Rycroft and Chris Lyal at the NHM, Guido Sautter and Terry Catapano of the GoldenGATE/TaxonX team, and Ben Clark of GATE. Many thanks to them and the other contributors.


Please note that the development of GoldenGATE is ongoing and new versions are released regularly. Technical issues identified in this report will no doubt be fixed or changed significantly in later versions.

Summary

- GoldenGATE is a promising application under active development.
- GoldenGATE is a good general XML mark-up tool, allowing rapid manual and semi-automatic annotation of text. It also offers a number of functions specific to the handling of taxonomic documents.
- The application is of modular design and can support a number of import and output formats.
- User defined functions are easily set up and can be set up to run sequentially as a pipeline (macro), making GoldenGATE highly customisable.
- The main technical limitations of the version evaluated are:
  - Text files larger than approx 200kB refresh very slowly in the editor. This appears to be caused by the poor performance of the Java Swing component used for the editor window. The current workaround is to split the input file into smaller pieces prior to marking-up.
  - Characters outside of the Western European character set are not handled properly. Support for a set of Unicode formats is in the process of being added and should appear in the next version.
  - A number of areas of direct interest to WP6, such the encoding of taxonomic relationships and bibliographic information, are not directly supported in the
evaluated version of TaxonX. Nor are the relevant tools available within GG. The addition of new external schemas is currently being implemented to address the mark-up related aspects many of these issues. In parallel, the future development of GG will presumably provide a user interface to add this mark-up.

- Some complex functions, such as the automatic recognition and annotation of taxon names, shift the user input away from the manual addition of annotations and instead require the user to check and, where necessary, correct the automatically added mark-up. The UI supporting this checking and correcting phase could be improved, and this would increase the usefulness of these functions.

- GoldenGATE supports the automatic retrieval of taxon IDs from the Hymenoptera Name Server. If this feature would be more generally useful if extended to cover a broader range of taxa.

- The documentation in the form of a user manual and help information are of good quality, but the rapid pace of development inevitably means that there are some areas in which the documentation lags behind the state of the tool.

- Documentation of the workflow suitable for the export of TaxonX to level 3 would be required if the output is to meet the requirements of the WP6 data warehouse. (At the time of this evaluation, only the level 1 workflow was available).

- In addition, a set of specifications of the intermediate formats created prior to export as TaxonX for specific levels of mark-up would be required for WP6 use.

- Like GoldenGATE, TaxonX is evolving rapidly.

- The schema takes a relatively minimalist approach when compared with more atomised formats such as taXMLit.

The combination of a schema and a customised XML editing tool is powerful, and has great potential. While GoldenGATE and taxonX are in rapid development, I would suggest that EDIT staff continue to liaise with the developers to remain aware of changes and provide feedback where appropriate.