Milestone: Review of options to use typographical information and other contextual clues

Partners: OU, KIT

Compiled by: David Morse (OU), Dauvit King (OU), Guido Sautter (KIT)

December 2011

WP7: Biodiversity literature access and data mining

Notes on M7.13 - Review of options to use typographical information and other contextual clues

Due

23:59 Sunday 31st July 2011

This Milestone Report provides an overview of progress toward using typographical information and other contextual clues for mark-up and data-mining of documents. As explained in the Milestone 7.11 report, our preferred option is to use the GoldenGATE software as the foundation for mark-up of taxonomic documents using XML. Therefore this document describes the functionality of the GoldenGATE software – in terms of its constituent parts rather than the interactive application.

Use of typographical information to support mark-up

The report documents the use of typographical information and other contextual clues (such as font changes, layout information and headings) within the context of GoldenGATE[1]. We intend to use this tool as the basis for a fully automated solution, rather than develop a new tool. GoldenGATE was developed at KIT, a ViBRANT partner organisation. It is a tool for semi-automated mark-up of taxonomic literature. As reported in Sautter *et al* (2009), while an experienced taxonomist can mark up materials more quickly when using GoldenGATE than when unaided, this remains a relatively slow process. The process will not scale to the needs of the ViBRANT project. However, GoldenGATE does form the basis for the development of an automated tool, and is thought to be suitable for enhancement through making increased use of typographical information. This assumption will be re-visited as ViBRANT progresses to confirm that it remains valid. If not, particularly if the latest developments in document analysis and information extraction prove difficult to integrate, then a new tool will be developed.

This Milestone report provides an underpinning for several milestones and deliverables that are forthcoming in the second year of the ViBRANT project for WP7. These milestones and deliverables are as follows:

ID	Short description	Month due
MS7.16	Mark-up modules delivering outline mark-up e.g. for article boundaries, treatment boundaries, headings and authors	18
MS7.17	Review of pilot mark up processes within the Scratchpad infrastructure	20
D7.2	Mark-up modules	24

1GoldenGATE can be downloaded from http://idaho.ipd.uka.de/GoldenGATE/.

In MS7.16 the team responsible for WP7 will be developing software to identify document structure elements such as article boundaries, treatment boundaries, headings and other significant document structures automatically. In other words, we will extend GoldenGATE's semi-automated approach to the identification of document structures in two ways.

1. We will extend the software so that it identifies a greater range and complexity of document structures than it does at present. This is particularly important for historic literature with its variety of styles.
2. We will enhance the software so that it identifies these document structures automatically, with a greater level of automation and degree of precision than GoldenGATE is able to achieve currently. This is important to achieve the scalability sought in ViBRANT.

The attached report on the existing GoldenGATE modules is a programmer-centric view of the problem because we are looking at the technical issues and their solution. This milestone is also on the development path towards MS7.16, MS7.17 and D7.2, which are themselves technical milestones and deliverables.

Finally, we think there is scope to publish a narrative review of this problem area and current solutions but this has not been done yet.

See attached report : *Report on Existing GoldenGATE Modules*

Reference

Sautter, G., Böhm, K., Agosti, D., and Klingenburg, C., (2009), 'Creating digital resources from legacy documents–an experience report from the biosystematics domain', In: *Proceedings of ESWC (European Semantic Web Conference) 2009*, Heraklion, Greece.
