INFORMATION SOCIETY TECHNOLOGIES
(IST)
PROGRAMME

Project IST-2001-33562 **MoWGLI**

D6.c Validation 3: Journal interface.

Author: Romeo Anghelache

Project Acronym: **MoWGLI**
Project full title: Mathematics On the Web: Get it by Logic and Interfaces

Proposal/Contract no.: IST-2001-33562 **MoWGLI**

1 Introduction

This validation prototype is a capability test of **Hermes** as a conversion and authoring tool of documents containing mathematical expressions, and it required a trial conversion of all the articles in Living Reviews in Relativity, which is an international, peer-reviewed, open-access electronic publisher of review articles in gravitational physics.

We describe here the current status of the **Hermes** tool, the results of the "Journal interface" validation prototype and the external (to the MoWGLI project) collaborations which started around the **Hermes** tool meanwhile.

This document ends with a few general conclusions we derived from these efforts, concerning the semantic needs of authoring, archiving and publishing of scientific documents.

2 Current status of Hermes

Hermes was originally intended to provide a way to convert mathematical expressions in Content-MathML. Now, at the validation prototype stage, it is a semantic oriented, full document converter and authoring tool.

The current implementation of **Hermes** has the following components:

- a set of semantic helper macros, for LaTeX, AMS-LaTeX and AMS-TeX (the dlt.tex, dalt.tex and da.tex files available in the source distribution). These macros enable the author to add semantic layers to his documents, that is, authoring with, or constructing, his own semantic vocabulary, such as MathML-content, or, with minimal manual intervention, making his document renderable in a web browser using XML and MathML-presentation.

- a scanner, written in **flex**, which extracts from the resulting 'semantic **dvi**' file the tokens seeded by the macro collection above and sends them to the parser below (the 'hermes.l' file in the **Hermes** distribution);

- a parser, written in **bison**, which is a grammar that performs a semantic action when a structured set of tokens is recognized (the 'hermes.y' file in the **Hermes** distribution); the semantic action is the creation of parts of the **XML** output; the parser and the scanner compile into a 'semantic **dvi**' translator called 'the **Hermes** translator'.

Using the dvi for conversion is a design decision which has been taken to minimize the complexity which structural dependencies of various packages and their use of the TEX achitecture may add to the process of parsing directly the sources.

The current implementation of **Hermes** converts mathematical expressions authored in LATEX, AMS-LATEX and AMS-TEX as follows:

- arbitrary expressions/symbols are encoded first in Presentation-MathML (the presentational semantic layer, understood by mathematical expressions rendering engines).

  This level of conversion requires inclusion of a single macro in the originally authored source.

  Hermes expects all the oriented delimiters (parenthesis, brackets ...) in the mathematical regions to be balanced, in order to generate well formed expressions, otherwise it dies verbosely.

  The TEX primitives which imply source backparsing (e.g.\over) are not interpreted correctly by **Hermes**, therefore a minimal clean-up is required before attempting the conversion. The clean-up process may be designed to avoid human intervention.

- expressions which have a clear meaning (are authored using the **Hermes** content macros) are wrapped further in Content-MathML

- expressions which use a home-grown semantic vocabulary can generate a corresponding XML vocabulary (the semantics has to be added to the **Hermes** grammar, in **bison** notation, as a source module);

The current implementation of **Hermes** converts the document structure as follows:

- recovers typical metadata: title, author, creation date... (the metadata model is currently in expansion, following the Living Reviews metadata needs: structured bibliography, publication status, etc.);

- preserves the internal references: citations and equations references (a more semantic approach, which is under construction, will enable the preservation and validation of all the internal and external references: references to sections, pages, files, URLs)

- saves presentational hints in the output XML document

- saves the generic structure of the document (sections, paragraphs)

3 Changes in architecture

No major structural changes were necessary in the software architecture since the extended prototype version.

A lot of improvements have been made, all of them beyond the original MoWGLI requirements: recognition of more document metadata, a larger number of TEX fonts are mapped to Unicode, two supplementary macro sets have been added (for $\mathcal{AMS}$-TEX and $\mathcal{AMS}$-LATEX) to support the transformation of these, major packages specific, mathematical regions (structures like multline, gathered, split, etc.), used at authoring time, into the corresponding MathML macro-structures.

The gradual semantic annotation and document structure continue to belong to the 'beta' stage as they evolve slowly according to the new users' needs.

The content macros (responsible for creating MathML-content) in the distribution are unchanged, user interest seems to be low in this direction, mainly because the benefits are not obvious and the lazy approach of converting an already authored paper is less tedious than using a new vocabulary at authoring time. We expect that this interest will grow, slowly, once the benefits of expressing mathematics in XML become more popular.

4 Living Reviews conversion

All the currently published articles in the Living Reviews in Relativity collection have been converted to XML+MathML. They served also as a testbed for constructing the **Hermes** macros corresponding to the major LATEX packages used.

The conversion process involved the following steps:

1. validate the original sources: compiling the original sources into **dvi**
2. validate the sources with the insertion of the corresponding **Hermes** macro collection: compiling the modified sources into semantic **dvi**;

   this is the step where **Hermes** detects unbalanced mathematical expressions and author intervention is needed to fix the sources
3. parse the semantic **dvi** with **Hermes** and output the result into an **XML** file, encoded in Unicode UTF-8, which contains enough semantic information for making it archivable in a library.

4. convert the library **XML** file into a media dependent **XML** file (currently, the **Hermes** distributon comes with an **XSLT** stylesheet which converts a typical scientific article into an XHTML+**MathML** renderable on screen and printable in a pdf); the quality of the rendering depends on the fonts and the implementation of the MathML rendering engine.

The overall result of the sample conversion can be reached and browsed through the **Hermes** website, hosted by Max-Planck-Institut für Gravitationsphysik (Albert-Einstein Institut), Golm, Germany.
The **Hermes** website address is: http://www.aei.mpg.de/hermes .

5 External collaborations

Hermes development has passed beyond the requirements of the MoWGLI project (converting LaTeX into Math-ML): it continues to grow towards a fully endowed semantic authoring and publishing tool guided by the feedback of the current actual users.

One of the "early adopters" of **Hermes** is Zentralblatt für Mathematik (Berlin-Karlsruhe, Germany), which decided to use **Hermes** for displaying answers to user queries from their reviews database in XML+MathML. The author of this document has tested the conversion of 3 volumes of Zentralblatt of 90,000 records (abstracts) each, covering all the domains in mathematics typically handled by Zentralblatt, and developed, along the way, the **Hermes** macro collection for AMS-TeX.

Another independent group, led by Prof. Antal M. IVÁNYI, Faculty of Informatics of Loránd Eötvös University, Budapest, Hungary, is already using **Hermes** to create a small number of fundamental books in computer science, from LaTeX in XML, in a government funded project.

Yet another independent group, led by Prof. Günter Törner, from Duisburg University, Germany, in a project for archiving TeX documents in the mathematical domain, called TeXDocC, plans to use **Hermes** and move the emphasis on **archiving XML** documents and **storing** the TeX sources, instead of archiving the documents in the original TeX format.

Hermes will also be used, during this year, after the end of the **MoWGLI** project, in converting some of the Einstein's original papers into XML+MathML, in a collaboration with Max Planck Institute for History of Science, Berlin, Germany.

6 General conclusions

At the end of the **MoWGLI** project we can formulate some potentially useful statements:

- the "print on paper" oriented design of the TEX architecture proved unsurprisingly to be a hindrance to any practical attempt of using TEX for building semantic documents in a user (or archiver) friendly way;
- the process of scholarly communication has to be understood as a 3 steps process: authoring, archiving, publishing; this perspective hints which are the semantic priorities, and which of them are appropriate for which group;
- the presentational and administrative metadata layers seem to be the most atractive part for the currently interested **Hermes** users, but all the three categories above need to be addressed by a modern semantic authoring tool.

Note: **Hermes** is free software, covered by GPL, and will continue to be accessible online from its host institution: Albert-Einstein Institut, Golm, Germany, at the address: http://www.aei.mpg.de/hermes . Its development continues past the end of the **MoWGLI** project.