EASTER

Evaluating Automated Subject Tools for Enhancing Retrieval

Koraljka Golub, Vanda Broughton, George Buchanan,
Emlyn Everitt, Debra Hiom, Marianne Lykke Nielsen,
Dagobert Soergel, Douglas Tudhope

Background

- EASTER is an 18-month JISC project funded under the Information Environment Programme 2009-11.
- Started April this year and involves eight institutional partners
- Aim is to test and evaluate a range of current tools for automated subject metadata generation
- Anticipated outcomes:
  - better understanding of limitations and what possible
  - recommendations for services employing subject metadata in JISC community

Rationale

- EASTER investigates the creation and enrichment of subject metadata using existing automated tools.

- Subject metadata are the most important in resource discovery, yet most expensive to produce manually. In addition, they are more difficult to generate automatically compared to formal metadata such as file type, title, etc.

- Due to the high cost of evaluation, automated subject metadata tools are rarely tested in live environments of use.

- Challenge facing digital collections, institutional repositories, and aggregators of how to provide high quality subject metadata at reasonable costs.

Intute testbed

- Test-bed is Intute http://www.intute.ac.uk

- Tools for automated subject metadata generation
  will be tested in two contexts:
  Intute cataloguers in the cataloguing workflow;
  end-users of Intute who search for information

- Task-based end-user retrieval study will examine contribution of automatically assigned terms and manually assigned terms

Methodology

- A methodology for evaluating such tools is also an intended outcome/contribution

- The methodology includes creating an enhanced ‘gold standard’ test collection by careful manual cataloguing and expert review by cataloguers and users

Types of subject metadata

Two processes and types of subject metadata will be explored:

1) The creation of subject metadata: using controlled terms from thesauri

2) The enrichment of metadata records: with non-controlled subject keyphrases

Candidate Tools

Initial candidate tools (a subset will be selected after review)

1) Temis Categorizer (http://www.temis.com/index.php?id=78&selt=1)
2) KEA (http://www.nzdl.org/Kea/)
3) TextGarden (http://kt.ijs.si/Dunja/textgarden/)
4) TerMine (http://www.nactem.ac.uk/software/termine/)
5) KnowLib's automated classifier
   (http://www.it.lth.se/knowlib/auto.htm)
6) Scorpion
   (http://www.oclc.org/research/software/scorpion/default.htm)
7) iVia project's libiViaClassification
   (http://ivia.ucr.edu/manuals/stable/libiViaClassification/5.4.0/)

Contact

Project website

http://www.ukoln.ac.uk/projects/easter/

Project publications
http://www.ukoln.ac.uk/projects/easter/dissemination/

k.golub@ukoln.ac.uk
dstudhope@glam.ac.uk