Project Acronym: EASTER
Version: draft v2
Contact: k.golub@ukoln.ac.uk
Date: 29 April 2009

JISC Project Plan

Overview of Project

1. Background

Subject metadata play a crucial role in resource discovery, but require the most resources to produce. Apart from helping to deal with scale and sustainability of processes, automated subject metadata can be used to enrich existing metadata and help establish more connections across and between resources, as well as to enhance metadata consistency. Moreover, automated subject metadata today find its use in a wide variety of applications, such as e-mail filtering, focused crawling and many others.

While automating the generation of any type of metadata is a big issue, automating subject metadata presents the hardest challenge. Research related to automated subject metadata can be found in a number of different areas (Polfreman et al. 2006), such as text categorization and document clustering (Sebastiani 2002; Jain et al. 1999), and can involve assigning controlled terms or extracting keyphrases (Toth 2002; Wu and Li 2008). How good the tools are, and how they compare against each other for different tasks or purposes, is largely unknown. This is mostly due to the fact that no comprehensive methodology for evaluating such tools exists.

There is a huge challenge facing UK HE digital collections, institutional repositories, and aggregators of institutional repository content, as to how to provide high quality metadata for increasing numbers of digital information at reasonable costs. While one can make an intuitive guess, one should strive for an objective estimate of the potential of existing automated tools.

2. Aims and Objectives

The project is concerned with the creation and enrichment of subject metadata using existing automated tools which will be tested with Intute in a live environment. Two processes and types of subject metadata will be explored:

1) The creation of subject metadata: using controlled terms from thesauri; and,
2) The enrichment of metadata records: with non-controlled subject keyphrases.

Automated subject metadata creation will be examined for different degrees of automation:

1) The possibility of entirely automating subject metadata creation; and,
2) The possibility of using existing tools for semi-automated subject metadata generation.

The tools for automated subject metadata generation will be tested in two contexts:

1) By Intute cataloguers in the cataloguing workflow; and,
2) By end-users of Intute who search for information in Intute as part of their research, learning, and information management processes.

3. Overall Approach

The project will first develop the methodology for evaluating tools for automated subject metadata, based on literature reviews or related evaluation methodologies, e.g., in the area of information retrieval. The methodology will then be implemented in the above contexts. First, all tools will be evaluated for results. Then, the best tool(s) will be implemented in a demonstrator, which will feed its results into the Intute cataloguing workflow. The demonstrator and integration will be evaluated, involving cataloguers study for Intute. Furthermore, a task-based end-user retrieval study will be conducted to determine whether relevant documents were successfully retrieved by automatically assigned terms, manually assigned terms or a combination thereof. Detailed use cases within the two predicted contexts will be further identified during the project, and in co-ordination with the JISC-

funded study of Automatic Metadata Generation: use case identification and tools/services prioritisation.

3.1 Tools

We will conduct an updated review of tools available and make final selection of at least four after preliminary analysis. So far we plan to use the ones listed below:

1) **Temis Categorizer** (http://www.temis.com/index.php?id=78&selt=1)
   Commercial. Assigns controlled vocabulary terms through text categorization and extracts keyphrases.
2) **KEA** (http://www.nzdl.org/Kea/)
   Free. Assigns controlled vocabulary terms through text categorization and extracts keyphrases.
3) **TextGarden** (http://kt.ijs.si/Dunja/textgarden/)
   Free. Assigns controlled vocabulary terms through text categorization and extracts keyphrases.
4) **TerMine** (http://www.nactem.ac.uk/software/termine/)
   Free. Extracts keyphrases.
5) **KnowLib's automated classifier** (http://www.it.lth.se/knowlib/auto.htm)
   Free. Assigns controlled vocabulary terms through string-matching.
6) **Scorpion** (http://www.oclc.org/research/software/scorpion/default.htm)
   Free. Assigns controlled vocabulary terms through string-matching.
7) **iVia project's libiViaClassification** (http://ivia.ucr.edu/manuals/stable/libiViaClassification/5.4.0/)
   Free. Assigns controlled vocabulary terms through text categorization.

Testing and reviewing whether the tools can ultimately be used will be necessary: since most of them are freely available, documentation is scarce and there are little guarantees with such software. Moreover, each tool will need to be examined for its compatibility with selected controlled vocabularies and datasets, taking into consideration the feasibility of any vocabulary format conversion. Combinations of tools in a pipeline will also be tested, if it would be recognized that they may complement each other.

Different preparation tasks need to be distinguished for the automated tools. Assigning controlled vocabulary terms requires processing of controlled vocabularies, at least for converting them to a format accepted by the target tool. Text categorization tools (e.g., TextGarden) require a set of training documents from which to 'learn'. This set would be developed as part of the 'gold standard' (see below). Furthermore, resources to be classified need to be processed into appropriate formats; and, certain sections of HTML need to be identified, again depending on tools as some tools already have the parsing included.

3.2 Data Collection

Project partner Intute will provide access to a selection of textual resources. The following areas with accompanying controlled vocabularies are envisioned:

1) Social Sciences, with IBSS and HASSET thesauri;
2) Health and Life Sciences, with the CABI thesaurus; and
3) Arts, with the AAT and the Getty Names Thesaurus.

For each of these areas, Intute will provide a selected number of documents for evaluation.

3.2 Evaluation Framework

We propose the following steps for producing the gold standard:

1. Start with a sample of documents that have already been subject-indexed;
2. Have each document subject-indexed again by two highly qualified cataloguers working in a user-centred mode;
3. Have at least some documents examined by a small focus group of three users who would discuss all angles from which the document should be discovered; and,
4. Once the tools have run, create a combined list of all the terms assigned (keeping track of where they come from) and get very knowledgeable cataloguers as well as users to comment on each term.

Once the “gold standard” is in place, different evaluation measures could be used. In addition, the average number of classes assigned to each document will be taken into account. Several other factors, such as the number of documents that are classified, whether the main concept is discovered should be also taken into consideration. Any failure analysis should be conducted, both for missed and for wrong descriptors. Any source of error needs to be traced; for example, it could derive from the thesaurus used by a tool rather than in the algorithm itself. Correct and incorrect descriptors should be analyzed, as affected by various factors: subject facet; explicitly present in the document versus inferred; level of exhaustivity - how a tool performs at different levels; and the subject domain of the document.

In order to evaluate automated metadata in live environments, an end-user retrieval test based on different use cases for supporting research, learning and the management and use of content will be conducted. A reasonably large collection that has been manually indexed will be run through automated tools as well. Then, users will conduct searches on assigned tasks. We will determine the contribution of automatically assigned terms and manually assigned terms, each alone and in combination, to retrieval success (retrieving relevant documents) and failure (missing relevant documents and retrieving irrelevant documents).

A demonstrator of an automated subject metadata system will be evaluated through an in-use observation. The observation will comprise of four elements: 1) a familiarisation tutorial; 2) an extended in-use study; 3) a manual metadata entry session; 4) a summative semi-structured interview. Sessions 2 and 3 will include short summative interviews for that session. We will carry out this observation with practicing cataloguers from Intute, using different subject areas. The study will determine the cataloguers’ assessments of the quality of the automated metadata created, identify usability issues for automated metadata extractors, and evaluate the impact of automated metadata on catalogue entry, in comparison to manual methods. Throughout, both qualitative and quantitative measures will be taken. The result will be a concrete understanding of the practical consequences of using automated metadata generation.

4. Project Outputs

The following deliverables are planned:

1) Evaluation methodology
2) ‘Gold standard’ data
3) Evaluation report of chosen tools
4) Intute workflow integration demonstrator
5) Intute workflow integration report
6) End-user retrieval study interface
7) End-user retrieval study report
8) Updated evaluation methodology
9) Final report
10) Dissemination in various forms

Apart from recommendations for the best subject tools, the project’s outcomes will include a proven methodology for evaluating automated tools as well as a report on the experience of implementing and testing the different free and commercial automated tools. These will benefit both the practitioners wanting to evaluate automated tools for their particular tasks and purposes, and researchers developing new tools. The resulting methodology framework will be of considerable interest to national and international researchers, as well as practitioners working in the field of automated metadata.

5. Project Outcomes

The project is envisaged to help understand and identify opportunities that should be further exploited as part of the e-infrastructure for education and research. While the results will be of great value to Intute, whose particular needs will be considered in the context of its own end-users and cataloguers, the outcomes will be highly relevant for all UK HE digital collections and JISC Information Environment Programme. The project would inform the community of the potential of the automated

tools within institutional and/or other service infrastructure environments. This will include issues of scale, skills, sustainability and costs. The results may be applicable to other parts of the metadata creation workflow and to different digital collections, such as repositories. Automatically created metadata could be used at various stages of the metadata creation workflow: 1) by an author creating original metadata at the time of deposit; 2) by a reader annotating (for colleagues/world or for recommendation for inclusion in a collection); and, 3) by a cataloguer. The results could also be applied to vocabulary-oriented metadata normalising and enhancement service, e.g. an aggregator harvesting relevant metadata, enhancing it and then offering harvesting of the improved metadata, as suggested in Tudhope, Koch, and Heery (2006).

The project is expected to have both immediate and long-term impact. Intute is a free online service providing access to the very best Web resources for learning, education and research in HE. For new innovative research to flourish, ease of access to and use of information services such as the ones provided by Intute are required. This project will examine to what degree and how information centres such as Intute could provide more resources at a faster rate and whether and how they could enhance subject access to information by addressing the hardest and most important metadata processes. Gaining this knowledge represents immediate impact. If proven successful, the tools evaluated in the project could be implemented and as such would enhance the cataloguer's efficiency, provide more metadata that will also be more consistent (medium-term impact), which would make the discovery of relevant information simpler for the wider academic community, thus encouraging the development of new science and improved learning (long-term impact). In addition, apart from helping to deal with scale, sustainability, and enrichment of metadata, automated subject metadata could be used in a wide variety of other applications, such as e-mail filtering, focused crawling and many others. Thus, the knowledge gained about the subject metadata tools, as well as their evaluation methodology, will be valuable to a range of practitioners and researchers in these areas.

The collaborative proposal addresses Welsh priorities concerning promotion of research capability and collaboration. Its outcomes will support the development of e- and distance learning/research through the enhancements to Intute capabilities for the improvement of metadata availability, with implied guidance for other digital collections.

6. Stakeholder Analysis

Stakeholder	Interest / stake	Importance
Digital librarians and other information professionals	Ability to reduce indexing costs and improve retrieval.	high
End users	Improved subject access to information.	medium
Discovery and delivery service providers	Improved ways to reduce indexing costs and improve retrieval. Guidance on how to build systems to support the above.	high
Information scientists	New findings related to subject access to information.	medium

7. Risk Analysis

Risk	Probability (1-5)	Severity (1-5)	Score (P x S)	Action to Prevent/Manage Risk
Staffing	1	4	4	Existing staff will work on the project. Multiple staff at each site have the expertise and skills required.
Organisational	2	3	6	Experience from previous projects is in place. Involve partner representatives in project meetings. Consortium agreement.
Technical	2	3	6	Establish contact and co-ordinate efforts with software producers early in the process. Scope Intute requirements.
External suppliers	2	4	8	Establish contact and co-ordinate efforts with software producers early in the process.
Legal	1	4	4	Regulate possible issues with a consortium agreement.

8. Standards

Name of standard or specification	Version	Notes
HTML standards and guidelines		For the project Web site

9. Technical Development

The final demonstrator will be built from existing software. The novel parts (i.e. automated metadata extractor software) are themselves technically proven, but not practically evaluated. Following the project, a public, open-source version will be made available to the professional practicing library community with documentation. Overall, QA policies and procedures will be developed based on the recommendations of the JISC-funded QA Focus project. Temis commercial interests will be protected as agreed with them.

10. Intellectual Property Rights

The project will comply with the terms of the JISC Funding Agreement. The IPR of material generated as part of the project will remain with the respective creators. All outputs, including documentation and code, created during the fulfilment of this project will be disseminated to the wider HE community with the expectation that it will be made freely available under an appropriate open source or creative commons license as appropriate. Temis commercial interests will be protected as agreed with them.

Project Resources

11. Project Partners

UKOLN, University of Bath
Role: project management, evaluation methodology development, end-user study, dissemination (WP **1**,**2**,3,4,**5**,**6**)
Main contact: Koraljka Golub

University of Glamorgan
Role: tools implementation and evaluation, end-user study (WP 1,2,**3**,4,**5**,6)
Main contact: Douglas Tudhope

Intute, MIMAS – University of Manchester
Role: provision of data, end-users and cataloguers (WP 1,2,**3**,**4**,**5**,6)
Main contact: Debra Hiom

City University London
Role: demonstrator workflow integration (WP 1,2,3,**4**,5,6)
Main contact: George Buchanan

Consulting expert: Dagobert Soergel
Role: expert consultant in the area of metadata and evaluation (WP 1,**2**,**3**,4,**5**,**6**)

Royal School of Library and Information Science, Denmark
Role: non-funded expert researcher in the area of knowledge organization systems and user-based evaluation (WP 1,**2**,**3**,4,**5**,**6**)
Main contact: Marianne Lykke Nielsen

University College London
Role: non-funded expert researcher in the area of knowledge organisation systems (WP 1,**2**,3,4,**5**,6)
Main contact: Vanda Broughton

OCLC Office of Research, USA
Role: non-funded support provider for Scorpion (WP 1,2,**3,4,**5,6)
Main contact: Diane Vizine-Goetz

The consortium agreement will be signed within three months of the start of the project.

12. Project Management

Project management and partner co-ordination will be provided by UKOLN and will be achieved by an initial project start-up meeting, a mid-term meeting and a closure meeting. Communication between partners will be supported by email-based discussions and further telephone meetings. Project reports will be supplied and co-ordinated by the UKOLN. The project manager will spend 10% on the management.

Project team

UKOLN		
Michael Day	Project Director	m.day@ukoln.ac.uk UKOLN University of Bath, Bath, BA2 7AY tel: +44 (0) 1225 383923 fax: +44 (0) 1225 386838
Koraljka Golub	Project Manager/Research Officer	k.golub@ukoln.ac.uk UKOLN University of Bath, Bath, BA2 7AY tel: +44 (0) 1225 383619 fax: +44 (0) 1225 386838
Sarah Hext	Project Administrator	s.hext@ukoln.ac.uk UKOLN University of Bath, Bath, BA2 7AY tel: +44 (0) 1225 383618 fax: +44 (0) 1225 386838
University of Glamorgan		
Douglas Tudhope	Glamorgan demonstrator leader	dstudhope@glam.ac.uk School of Computing, University of Glamorgan, Pontypridd, CF37 1DL tel: +44 (0) 1443 482271 fax: +44 (0) 1443 482715
Emlyn Everitt	Software developer	eeveritt@glam.ac.uk School of Computing, University of Glamorgan, Pontypridd, CF37 1DL tel: +44 (0) 1443 482202 fax: +44 (0) 1443 482715
Intute		
Debra Hiom	Intute data, cataloguers and users coordinator	d.hiom@bristol.ac.uk Institute for Learning and Research Technology, University of Bristol 8-10 Berkeley Square Bristol, BS8 1HH tel: +44 (0) 117 928 7117 fax: +44 (0) 117 928 7112
City University London		
George Buchanan	Demonstrator workflow integration	George.Buchanan.1@city.ac.uk School of Informatics, City University Northampton Square London, EC1V 0HB tel: +44 (0) 20 7040 8469 fax: +44 (0) 20 7040 8859

Consulting expert		
Dagobert Soergel	Expert consultant	dsoergel@umd.edu College of Information Studies, University of Maryland College Park, MD 20742-4345, USA tel: 301-405-2037 fax: 301-314-9145
Danish Royal School of Library and Information Science		
Marianne Lykke Nielsen	Expert researcher	mln@db.dk RSLIS Fredrik Bajers Vej 7K 9220 Aalborg Øst, Denmark tel.: +45 98 15 79 22 fax: +45 32 84 02 01
University College London		
Vanda Broughton	Expert researcher	v.broughton@ucl.ac.uk Department of Information Studies, University College London Gower Street London WC1E 6BT tel: +44 (0) 20 7679 2291 fax: +44 (0) 20 7383 0557
OCLC		
Diane Vizine-Goetz	Supporting officer for DDC	vizine@oclc.org OCLC 6565 Kilgour Place Dublin, Ohio 43017-3395, USA tel: +1 614 764 6084 fax: +1 614 764 6096

13. Programme Support

Invitations to events on subject access to information.

Invitations to events on automation.

General alerts on other JISC projects and reports which are particularly relevant to EASTER.

14. Budget

See Appendix B.

Detailed Project Planning

15. Workpackages

See Appendix 2.

16. Evaluation Plan

Timing	Factor to Evaluate	Questions to Address	Method(s)	Measure of Success
Month 7	Readiness of the evaluation methodology	Is the methodology ready to be used in the study?	Pilot testing	Pilot testing shows that the methodology is ready
Month 10	Producing the 'gold standard' data	Is the 'gold standard' data ready to be used	Pilot testing	Pilot testing shows that the data is well

		in the study?		designed and ready
Month 14	Intute workflow integration demonstrator	Is the demonstrator ready to be used in the study?	Pilot testing	Pilot testing shows that the demonstrator is ready
Month 14	End-user retrieval study interface	Is the interface ready to be used in the user study?	Pilot testing	Pilot testing shows that the interface is ready
Months 7-14	Designing user and cataloguer studies, and 'gold standard' data collection study	Is the study well designed?	Pilot testing	Pilot testing shows that there the study is appropriate and well designed
Months 7-17	User and cataloguer studies, and 'gold standard' data collection study	As included in the study	Questionnaires and data logging	All data are collected and properly stored
Months 13-18	Reports	Are issues important to stakeholders addressed?	Check with stakeholders through personal contact	Production of report that represents interests of stakeholders

17. Quality Plan

Output	**WP2: Evaluation methodology development**				
Timing	**Quality criteria**	**QA method(s)**	**Evidence of compliance**	**Quality responsibilities**	**Quality tools (if applicable)**
Months 1 to 18	Appropriate sampling of the literature and use cases.	Feedback from other partners and colleagues.	Positive feedback.	Koraljka Golub, Dagobert Soergel	
Output	**WP3: Subject metadata evaluation**				
Timing	**Quality criteria**	**QA method(s)**	**Evidence of compliance**	**Quality responsibilities**	**Quality tools (if applicable)**
Months 1 to 13	Scientific Appropriateness	Testing		Douglas Tudhope, Koraljka Golub	
Output	**WP4: Implementing best tool(s)**				
Timing	**Quality criteria**	**QA method(s)**	**Evidence of compliance**	**Quality responsibilities**	**Quality tools (if applicable)**
Months 13 to 16	Scientific Appropriateness	Using established study methods and sampling	Pilot testing	George Buchanan	
Output	**WP5: End-user study**				
Timing	**Quality criteria**	**QA method(s)**	**Evidence of compliance**	**Quality responsibilities**	**Quality tools (if applicable)**
Months 12 to 17	Scientific Appropriateness	Using established study methods and sampling	Pilot testing	Koraljka Golub, Douglas Tudhope	
Output	**WP6: Dissemination**				
Timing	**Quality criteria**	**QA method(s)**	**Evidence of compliance**	**Quality responsibilities**	**Quality tools (if applicable)**
Months 3 to 18	Scientific Appropriateness and Stakeholders Interests Covered	Frequent discussions among partners, JISC, and other colleagues	Successful completion of external peer review	All main contacts	

Feedback and peer review from project partners, people at events and JISC throughout the project.

18. Dissemination Plan

Timing	Dissemination Activity	Audience	Purpose	Key Message
End of project	Report on the tools, integration and influence on retrieval	Repositories, digital collections	To inform beneficial developments and motivate buy-in	Whether automated subject tools can be useful
Throughout project and afterwards	Presentations at conferences and other events	Information services providers, researchers	To foster further collaborations and ensure buy-in	
Throughout project	Web site	All of above	All of above, enable access to demonstrator	

19. Exit and Sustainability Plans

Project Outputs	Action for Take-up & Embedding	Action for Exit
Knowledge on automated subject metadata tools and evaluation methodology	Further dissemination	Further research in other contexts

Project Outputs	Why Sustainable	Scenarios for Taking Forward	Issues to Address
Demonstrator and software tools		Investigate stakeholder's interest	Seek further funding

References

Jain, A.K., Murty, M.N., and Flynn, P.J. (1999), "Data clustering: a review", *ACM Computing Surveys*, Vol. 31 No. 3, pp. 264-323.
Polfreman, M., Broughton, V., and Wilson, A. (2006). Metadata Generation for Resource Discovery: Final Draft. V2.
Sebastiani, F. (2002), "Machine learning in automated text categorization", *ACM Computing Surveys*, Vol. 34 No. 1, pp. 1-47.
Toth, E. (2002), "Innovative solutions in automatic classification: a brief summary", *Libri*, Vol. 25 No. 1, pp. 48-53.
Tudhope, D.; Koch, T.; Heery, R. (2006). Terminology Services and Technology: JISC state of the art review. http://www.jisc.ac.uk/Terminology_Services_and_Technology_Review_Sep_06
Wu, Y. B., &, Li, Q. (2008). Document keyphrases as subject metadata: Incorporating document key concepts in search results. Information Retrieval, 11, 229-249.

Appendixes

Appendix A. Project Budget

Appendix B. Workpackages

See separate document.