Cover Sheet for Proposals	JISC *IE Programme*

Name of Call Area Bidding For (tick ONE only):

- [x] **Strand A1: Automated metadata generation and text mining**
- [ ] **Strand A2: Developing e-infrastructure to support research disciplines**
- [ ] **Strand A3: Repositories: start-up**
- [ ] **Strand A4: Repositories: rapid innovation**
- [ ] **Strand A5: Repositories: enhancement**
- [ ] **Strand A6: Preservation exemplars**
- [ ] **Strand B1: VRE Innovation: Tools and interoperability**
- [ ] **Strand B2: VRE Innovation: VRE Frameworks**
- [ ] **Strand B3: VRE Innovation: VRE National and Institutional Interoperability**

Name of Lead Institution:	UKOLN
Name of Proposed Project:	Evaluating Automated Subject Tools for Enhancing Retrieval (EASTER)
Name(s) of Project Partner(s):	UKOLN, University of Bath; University of Glamorgan; Intute (MIMAS, University of Manchester); City University London; Dagobert Soergel (University of Maryland) who will serve as a consulting expert Non-funded supporting partners: Royal School of Library and Information Science, Denmark; University College London; OCLC Office of Research, USA

Full Contact Details for Primary Contact:

Name: Dr. Koraljka Golub Position: Research Officer

Email: k.golub@ukoln.ac.uk

Tel: 01225 383619 Fax: 01225 386838

Address: UKOLN
University of Bath,
Bath, BA2 7AY

Length of Project:	**18 months**		
Project Start Date:	1 April 2009	**Project End Date:**	30 September 2010

Total Funding Requested from JISC:

Funding Broken Down over Financial Years (April - March):

April 09 – March 10	April 10 – March 11	April 11 – March 12

Total Institutional Contributions:

Outline Project Description:

The purpose of the project is to test and evaluate existing tools for automated subject metadata generation in order to better understand what is possible, what the limitations of current solutions are, and make subsequent recommendations for services employing subject metadata in the JISC community. The information centre to be chosen as a test-bed for this project will be Intute, a free online service providing access to quality-controlled, manually selected and catalogued Web resources for learning and research. We envision the project outputs would help further understand the value of subject metadata tools and their evaluation and identify opportunities that should be further exploited as part of the e-infrastructure for education and research.

The project is concerned with the creation and enrichment of subject metadata using existing automated tools. Subject metadata are most important in resource discovery, yet most expensive to produce manually. In addition, they are much more difficult to generate automatically especially in comparison to formal metadata such as file type, title, etc. Also, due to the high cost of evaluation, automated subject metadata tools are rarely tested in live environments of use. 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 subject metadata for increasing numbers of digital information at reasonable costs.

The project will examine existing tools in order to determine to what degree they can be integrated into (semi-)automated workflows. The tools for automated subject metadata generation will be tested in two contexts: by Intute cataloguers in the cataloguing workflow; and by end-users of Intute who search for information in Intute as part of their research, learning, and information management . The project will first develop the methodology for evaluating tools for automated subject metadata. The methodology will then be implemented in the above contexts. First, all tools will be evaluated for results using a created 'gold standard'. The best tool(s) for the purposes of Intute will be implemented into a demonstrator that will feed its results into the cataloguing workflow. This will be evaluated. Furthermore, a task-based end-user retrieval study will be conducted to 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).

I have looked at the example FOI form at Appendix B and included an FOI form in the attached bid (Tick Box)	**YES**	
I have read the Call and associated Terms and Conditions of Grant at Appendix D (Tick Box)	**YES**	

FOI Withheld Information Form

Nothing in this proposal need be withheld from disclosure under the Freedom of Information Act 2000.

Section / Paragraph No.	Relevant exemption from disclosure under FOI	Justification

1 Appropriateness and Fit to Programme Objectives and Overall Value to the JISC Community

1.1 General Scope

1. The purpose of the project is to **test and evaluate existing tools for automated subject metadata** generation in order to better understand what is possible, what the limitations of current solutions are, and to make subsequent recommendations for services employing subject metadata in the **JISC community**. The information centre to be chosen as a test-bed for this project is **Intute,** a free online service providing access to quality-controlled, manually selected and catalogued **Web resources for learning and research**. We envision the project outputs would help further understand **the value of subject metadata tools** and their **evaluation** and **identify opportunities** that should be further exploited as part of the e-infrastructure for education and research.

1.2 Rationale and Need

2. **Metadata** is a crucial yet expensive part of providing information in digital collections such as repositories and subject gateways. **Descriptive metadata**, in particular **subject metadata,** play a crucial role in resource discovery. Subject metadata describe the 'aboutness', i.e., the subject content of the resource, and the topics, issues, and purposes for which a resource is relevant. Subject metadata can be from a **controlled vocabulary** (e.g., classification schemes, thesauri, subject heading lists) or freely assigned terms such as **keyphrases**. In comparison to **free-text search**, there are many advantages to searching by **controlled subject metadata** (however generated), such as the following:

- Most relevant search terms are selected, and relevant search terms not explicitly mentioned in document may be added;
- Search terms are controlled, i.e., disambiguated so that there is no confusion between terms that look the same but have different meaning;
- Search terms can come from semantically structured vocabularies - hence documents can be found through searching for synonyms, narrower, broader, and even related terms that may not be present in the document itself (semantic query expansion).

3. While **subject metadata** play a crucial role in resource discovery, they **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 finds its use in a wide variety of applications, such as e-mail filtering, focused crawling and many others (see Sebastiani 2002, 6-9).

4. 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 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, especially for subject metadata (see Section 2.3).

5. 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. **Intute,** being the biggest UK information centre providing quality Web resources for learning and research, has a recognized need to deal with the scale and sustainability of cataloguing processes in which subject metadata demand the biggest effort. Unless automated methods are used, metadata enhancement cannot take place either. Similar data centres need to know to what degree existing tools for automated subject metadata can provide qualitative benefits to its cataloguers, and quantitative benefits, such as time and money saved. For its end-users, the quality of automatically produced terms either in place or for enhancement, also needs to be investigated. Finally, if a tool proves useful, they need to know the feasibility of adopting the tools, including issues of scale, skills, sustainability and costs.

1.3 Aims and Objectives

6. 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. Many argue that a combination of intellectual and automated methods is currently an optimal approach (e.g., Hagedorn 2001).

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.

7. 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, which is due in August 2009. For further details see Sections 2.1 through 2.4.

1.4 Benefits and Overall Value to the JISC Community

8. The project builds on JISC's past investment in exploring terminology services (Terminology Services and Technology Review, Terminology Registry Scoping Study, both by UKOLN and University of Glamorgan) and user-focused services like Intute. 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 will 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).

9. 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.

10. The project is expected to have both **immediate and long term impact** and the project outcomes are **ensured sustainability beyond the JISC funding period** (see section 4 Impact).

11. 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. The project Web site will seek to conform to **HTML standards** and **WCAG 2.0 guidelines**.

2 Quality of Proposal and Robustness of Workplan

2.1 Tools for Automated Subject Metadata Generation

2.1.1 Research behind Automated Subject Metadata Generation

12. Research related to **automated subject metadata generation** is spread around three major areas: text categorization, document clustering and string (pattern) matching (Golub 2006a). While the discussion below refers to textual documents, which will also be the target in this project, approaches to metadata extraction in the wider world of multimedia exist but are much harder to tackle.

13. In **document clustering**, both clusters (classes) into which documents are classified and, to a limited degree, relationships between them, are produced automatically. Labelling the clusters is a major research problem, with relationships between them, such as those of equivalence, related-term and hierarchical relationships, being even more difficult to automatically derive (Svenonius 2000, 168). In addition, "[a]utomatically-derived structures often result in heterogeneous criteria for category membership and can be difficult to understand" (Chen and Dumais 2000, 146). Also, clusters' labels, and the relationships between them, change as new documents are added to the collection; unstable class names and relationships are user-unfriendly in information retrieval systems, especially when used for subject browsing. Thus, tools for clustering will not be considered in this project.

14. **Text categorization** (machine learning) is the most widespread approach to automated classification of text. Here characteristics of subject classes, into which documents are to be classified, are learnt from documents with manually assigned classes. However, manually classified documents are often unavailable in many subject

areas, for different document types or for different user groups. Also, text categorization algorithms only perform well on new documents if they are similar enough to the training documents. Traditionally, research in text categorization seems to be focused on improving algorithm performance, and experiments are conducted under laboratory-like conditions.

15. In **controlled-vocabulary based string (pattern) matching**, matching is conducted between a controlled vocabulary and the text of documents to be classified. A major advantage of this approach is that it does not require training documents, while still maintaining a pre-defined structure. If using a well-developed classification scheme, it will also be suitable for subject browsing in information retrieval systems. Apart from improved information retrieval, another motivation to apply controlled vocabularies in automated classification is to re-use the intellectual effort that has gone into creating such a controlled vocabulary (see also Svenonius 1997).

2.1.2 Tools Selected

16. The tools to be tested will be selected largely on the basis of JISC reports such as MetaTools project reports (Polfreman and Rajbhandari 2008), the JORUM report on automated metadata (Baird 2006), the repositories consistency report (Charlesworth 2008), and a US AMeGA report (Greenberg, Spurgin, and Crystal 2005). The main criteria for selection were the following: 1) free or free to evaluate, 2) functionality for controlled vocabulary assignment or subject keyphrase extraction, and, 3) functionality for the English language resources.

17. We intend to use at least four of the tools listed below. Testing and reviewing whether they 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.

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.

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 Section 2.3.2.1). 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.

18. Combinations of tools will also be tested, if it is recognized that they may complement each other. For example, as shown in Golub et al. (2006c), a string-matching algorithm may yield high precision, and a text categorization one may yield high recall. Combining them could provide an optimal solution.

2.2 Content (Data Collection)

19. 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.

2.3 Evaluating Automated Subject Metadata

2.3.1 Current Approaches and Challenges

20. Although there is a lot of research reporting on different approaches to automated metadata generation (e.g., Yang 1999; Yilmazel et al. 2004), the evaluation of those approaches and evaluation methodology is scarce. One approach is testing the **quality of retrieval** based on the assigned metadata terms. But retrieval testing is fraught with problems; the results depend on many factors, so retrieval testing cannot isolate the quality of the metadata nor can it shed light on the question of how automatic creation of metadata can be integrated into a workflow. Another approach is to measure **indexing quality** directly. One method of doing so is to compare automatically assigned metadata terms against existing human-assigned terms or classes of the document collection used (as a 'gold standard'), but this method also has problems, as discussed below. In most cases

measures inspired by information retrieval measures are used. Effectiveness, the degree to which correct classification decisions have been made, is often measured as precision (correct positives/predicted positives) and recall (correct positives/actual positives), and F1 which combines the two. These three measures were used in MetaTools for title and keywords, although they had initially considered using a number of other intrinsic and extrinsic measures as well (Polfreman and Rajbhandari 2008).

21. In order to develop measures for automated tools, the literature on **indexing quality** will be reviewed. According to Lancaster (2003, 83-99), an indexing "failure" could occur in the conceptual analysis phase of indexing, where a topic of user interest is not recognized or is misinterpreted, and in the translation phase, where the term assigned is not the most specific one or is inappropriate. When indexing, people make errors such as those related to exhaustivity policy (too many or too few subjects become assigned), specificity of indexing (which usually means that the assigned subject is not the most specific one available); they may omit important subjects, or assign an obviously incorrect subject. Soergel (1994) reviews indexing characteristics such as exhaustivity, correctness, and consistency in the light of their influence on retrieval. In addition, it has been reported that different people, whether users or professional subject indexers, would assign different subjects to the same document. Studies on inter- and intra-indexer consistency report generally low indexer consistency (Olson and Boll, 2001, p. 99-101). Markey (1984) reviewed 57 indexer consistency studies and reported that consistency levels ranged from 4% to 84%, with only 18 studies showing over 50% consistency. As this analysis shows, when interpreting consistency as a quality measure, one must consider the context.

22. In conclusion, existing metadata records cannot be used as a **gold standard**. For example, the classes assigned by algorithms, but not human-assigned, might be wrong; alternatively, they might also be right but omitted during human indexing by mistake. Also, as mentioned earlier, subject metadata creation involves determining subject terms or classes under which a document should be found; this goes beyond simply capturing what the document is about to what the document could be used for; text categorization algorithms might find such terms, given a good training set, but human indexers who are not well trained might miss them.

2.3.2 The Framework Proposal

2.3.2.1 Gold Standard

23. As illustrated by the above discussion, producing a **gold standard** is hard. We propose the following:

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.

24. Once the "gold standard" is in place, different evaluation measures could be used (see the MetaTools project, Polfreman and Rajbhandari 2008). 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.

2.3.2.2 Retrieval Test on Use Cases

25. 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).

2.4 Workflow Integration Demonstrator

26. 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.

2.5 Project Deliverables and Timetable

Major Deliverables	Month	Lead Effort
Project plan	1	**UKOLN** + partners
Evaluation methodology	7	**UKOLN** + partners
'Gold standard' data	10	**UKOLN** + partners
Evaluation report of chosen tools	13	**GLAM** + partners
Intute workflow integration demonstrator	14	**City** + partners
Intute workflow integration report	16	**City** + partners
End-user retrieval study interface	14	**UKOLN** + partners
End-user retrieval study report	17	**UKOLN** + partners
Updated evaluation methodology	18	**UKOLN** + partners
Final report	18	**UKOLN** + partners
Dissemination in various forms	3-18	**UKOLN** + partners

2.5.1 Workpackages

Work Package 1	**Project Management**
Lead:	UKOLN
Start:	Month 1
End:	Month 18
Description:	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, conference calls and informal methods. Project staff will work in partnership with members of relevant JISC Development teams, provide progress reports as required and participate in programme evaluation activities. All partners will contribute to reports.
Deliverables:	
1.1	Project Plan (month 1)
1.2	Project Web Page on JISC Web Site (month 1)
1.3	Web Site at UKOLN (month 3)
1.4	Consortium Agreement (month 3)
1.5	Project Plan (month 1)
1.6	Progress Reports (months 6 and 12)
1.7	Final Report (draft month 17, final month 18)
1.8	Completion Report, including financial statement (month 18)
Work Package 2	**Evaluation methodology development**
Lead:	UKOLN
Start:	Month 1
End:	Month 18
Description:	UKOLN and consulting expert will take the lead to review related literature and develop evaluation framework, with RSLIS, UCL and Intute providing input from their different research and practice perspectives.
Deliverables:	
2.1	Evaluation methodology (month 7)
2.2	Updated evaluation methodology (month 18)
Work Package 3	**Subject metadata evaluation**
Lead:	University of Glamorgan
Start:	Month 1
End:	Month 13
Description:	Intute will provide the data collection. University of Glamorgan and UKOLN will plan and conduct evaluation, with feedback from consulting expert, OCLC and RSLIS. They will produce a 'gold standard' data for evaluation, which will involve Intute cataloguers and a focus group of end-users. University of Glamorgan will prepare controlled vocabularies, install the tools, harvest Web pages (from Intute bibliographic records) to index, train and test the tools. University of Glamorgan and UKOLN will write an evaluation report.
Deliverables:	
3.1	'Gold standard' data (month 10)
3.2	Evaluation report of chosen tools (month 13)
Work Package 4	**Implementing most optimal tool(s) as part of a workflow**
Lead:	City University
Start:	Month 13
End:	Month 16

Description:	Intute will provide input on workflow. City University will implement a demonstrator integrating the most optimal tool into Intute cataloguer's workflow. Intute will provide feedback.
Deliverables:	
4.1	Demonstrator (month 14)
4.2	Evaluation report of the demonstrator and workflow integration (month 16)
Work Package 5	**End-user study**
Lead:	UKOLN
Start:	Month 12
End:	Month 17
Description:	Under guidance from consulting expert and RSLIS, UKOLN and University of Glamorgan will plan and conduct evaluation, which will involve generating automated terms, setting up a user interface, conducting the study, and reporting the results. Intute will provide the data collection and help find end-users.
Deliverables:	
4.1	End-user retrieval study interface (month 14)
4.2	End-user retrieval study report (month 17)
Work Package 6	**Dissemination**
Lead:	UKOLN
Start:	Month 3
End:	Month 18
Description:	The project will work with the JISC programme to disseminate results in a timely fashion. The project Web site will be kept up to date with progress on deliverables. Project partners will participate in JISC programme events, and will disseminate through conferences. All partners will participate.

2.5.2 Risks

Risk	Level	Likelihood	Contingency
Recruitment difficulties	Medium	Low	Existing staff will work on the project.
Loss of a team member	High	Low	Multiple staff at each site have the expertise and skills required.
Lack of engagement of partners	High	Low	Involve partner representatives in project meetings.
Project is over-ambitious in scope and/or over-runs	Low	MediumLow	Agree scope with JISC by means of project plan.
Problem with implementing a software	Low	MediumLow	Establish contact and co-ordinate efforts with software producers early in the process.
Problems with integrating tools with Intute workflows	Medium	MediumLow	Scope Intute requirements.

2.5.3 Intellectual property

27. 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 generated as part of the 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 licence as appropriate. Outputs will be made publicly available in a timely manner to ensure current information about the project is available throughout its life. We will respect the license model of all third party software used during the project.

3 Engagement with the Community

28. A close cooperation with Intute and its cataloguers as a representative of that group will be maintained throughout the project. Feedback involving the cataloguers has been planned as part of establishing the gold standard and workflow integration. Target subject areas and tools have been selected in coordination with their needs. End-users will also be involved as part of establishing a 'gold standard' and retrieval evaluation. Detailed use cases 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, which is due in August 2009.

29. The proposed consortium combines prominent research/development groups with an outstanding track record for dissemination and national and international contacts within the community. The mix of partners with different practice and research backgrounds (information service, OCLC, library science, computer science, HCI) brings different communities of use to play an active role in gathering requirements and providing input.

30. The project's major purpose is improvement of digital collection services such as those of Intute. Further testing of the applicability of project outputs within other domains (e.g. repositories) is desired, as well as the need to maintain wider communication with these. The project will thus track and collaborate as appropriate with related

projects, such as Deposit Plait, Enhanced Ingest to Digital e-research Repositories (EIDer), and Automated Archiving for an Institutional Repository (AIR).

31. The project will work with the JISC programme to disseminate results in a timely fashion. The project Web site will be kept up to date with progress on deliverables. Project partners will participate in JISC programme events, and will disseminate through conferences. One of the principle interests of partners from research institutions is to have strong publications, so journal articles will be another target. Through the project Web site and subsequent conference and journal publications, the outputs of the project will be made available beyond the JISC funding period. The final stages of the project will consider facilitating long-term access to the demonstrator and research outputs (see Section 4).

4 Impact

32. While the project will build on previous related JISC projects such as MetaTools and Automatic Metadata Generation for Resource Discovery, it is crucially innovative in many ways, especially because it develops concrete workflows for Intute, development of evaluation methodology, and comprehensive testing involving people on both ends of the spectrum: end-users as well as cataloguers. The project is envisaged to help further understand and identify opportunities that should be further exploited as part of the e-infrastructure for education and research. Recommendations will be drawn to address similar problems at a variety of digital information service levels, ranging from institutional, data centres, regional and national. The results will complement those of other metadata generation and repositories projects such as the Deposit Plait, EIDer, and AIR. In addition, the project will work together with a complementary project proposal, VIM (Value for money In automatic Metadata generation). VIM will focus on the role and cost of bibliographic and descriptive metadata in enhancing the search and retrieval experience of students and academics in UK Higher Education in order to identify the best value for money for information services in metadata generation.

33. In order to reduce its processing costs and at the same time improve access to information for JISC HE end-users, Intute as a representative of digital collection providers has recognized the need to evaluate to what degree automated tools can be used in its processes – this project will address this need. Both cataloguers and end-users will be involved and provide feedback. **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. For further stakeholder details, see first paragraph in Section 3.

34. 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 (see also Section 1.4 on other examples of benefits). 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.

35. 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.

36. The project outcomes are **ensured sustainability beyond the JISC funding period**. Since the proposal has been developed in the context of Intute itself, it is envisioned that its outputs could easily be embedded in Intute, if shown to be beneficial. The actual demonstrator will be available for Intute to apply/adapt/re-engineer. A public, open-source version will be made available to the professional practicing library community with documentation. This will enable other digital collections and services to explore the demonstrator and to embed it in their production services. In addition, all reports on different tools, the evaluation framework and other project materials will be made freely available via project Web sites and repositories as appropriate, such as the JISC IE repository and JorumOpen. In addition, the project would engage with the National Center for Text Mining (NaCTeM) to explore ways of offering continued access to project outcomes.

37. A self-evaluation of the project will be a significant element of the Final Report, focussing particularly on the degree to which the planned deliverables have been accomplished, and feedback at national and international events, as well as tracking and collaborating with similar projects (see first paragraph in this Section).

5 Budget

6 Previous Experience of the Project Team

38. This bid is led by UKOLN at the University of Bath. Project management will come from UKOLN and be carried out by Koraljka Golub, with leadership and direction from Michael Day as UKOLN Research team leader. Work on this project will be informed by UKOLN's involvement in activities such as the Intute Repository Search (IRS) project, the Repositories Support Project (RSP), and the Repositories Research Team (RRT). The project also complements the FixRep project proposal that is being submitted by a UKOLN-led consortium to this call. FixRep will focus on using text analysis and information extraction techniques to evaluate the potential of generating formal metadata types – typically intrinsic metadata such as document titles, creator names or format information – within real-world workflows. In addition, UKOLN have proven expertise in management of successful projects. As outlined below, the partnership of the consortium provides a strong basis for successful outcomes, combining teams with proven expertise in relevant areas, and previous co-operation on similar projects.

UKOLN, University of Bath

Koraljka Golub, Research Officer

Role: project management, evaluation methodology development, dissemination (WP **1**,**2**,3,4,**5**,**6**)

Koraljka Golub has worked as a project manager and researcher on projects concerning metadata and interoperability, including EnTag - Enhanced Social Tagging - and TRSS – the Terminology Registry Scoping Study (funded by the JISC). She completed her PhD on the topic of automated subject classification with emphasis on evaluation using a triangulation of methods involving a user study, and has published a dozen publications in the area. In 2008, together with Emma Tonkin of UKOLN, she wrote a book chapter on Technologies for metadata extraction.

University of Glamorgan

Douglas Tudhope, Professor

Role: Glamorgan effort leader, guidance on tools evaluation and end-user study (WP 1,2,**3**,4,**5**,6)

Douglas Tudhope was Principle Investigator in the UK EPSRC Research Council funded FACET project, in collaboration with the Science Museum, investigating thesaurus-based query expansion. He is currently Principle Investigator on the AHRC funded STAR project (Semantic Tools for Archaeological Resources), in collaboration with English Heritage. Together with UKOLN he has worked on previous JISC projects, EnTag - Enhanced Social Tagging – and TRSS – the Terminology Registry Scoping Study. He is Editor of the journal New Review of Hypermedia and Multimedia and acting Theme Editor, Information Discovery, Journal of Digital Information (JoDI). He is a member of the network on Networked Knowledge Organization Systems (NKOS) and has co-organised 10 NKOS Workshops at the European Conference on Digital Libraries, Joint Conference on Digital Libraries and Dublin Core Metadata Initiative Conference, chairing three of them. He has published over 60 refereed publications.

Emlyn Everitt, Senior Lecturer

Role: tools implementation and evaluation, technical support and development (WP 1,2,**3**,4,**5**,6)

Emlyn Everitt has 16 years experience working in both academia and industry (BT, Barclays, MoD, Syntegra, Microsoft, Cisco), working in the fields of information retrieval and retrieval analysis, and has been involved in a wide variety of high profile projects such as the NHS patient record system and the international inter-bank lending system.

Intute, MIMAS – University of Manchester

Debra Hiom, IntuteTechnical Manager

Role: provision of data, end-users and cataloguers (WP 1,2,**3**,**4**,**5**,6)

Debra Hiom has a first degree in Humanities and an MSc in Information Management. She has been involved in Internet research since 1992, with special interests in the area of networked resource discovery and digital libraries. As technical manager for Intute Debra has a keen interest in automated methods of metadata generation. She teaches on the MSc Course on Information and Library Management at the University of the West of England and has written extensively about the Internet, including the publications Online Information Services in the Social Sciences and the Library and Information Professionals Internet Companion. Debra also contributed to the EnTag - Enhanced Social Tagging project.

City University London

George Buchanan, Senior Lecturer

Role: demonstrator workflow integration (WP 1,2,3,**4**,5,6)

George Buchanan is currently a lecturer at Swansea University, moving to City University in March 2009. He is also a visiting academic at the UCL Interaction Centre, and has worked on the User Centred Interactive Search project. One of his main research areas includes information seeking in digital libraries, for which he has won several best paper awards at international conferences. His previous work has included extensive work on the open-source Greenstone Digital Library software, and extensions and adaptations of the DSpace institutional repository system.

Consulting expert

Dagobert Soergel, Professor at University of Maryland

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

Dagobert Soergel is an internationally renowned expert in the field of knowledge organization, information retrieval, information technology and software evaluation. His experience and knowledge is crucial to this project especially in relation to developing evaluation methodology. He works as a Professor at the College of Information Studies,

University of Maryland. His courses include Information Structure, Construction and Maintenance of Index Languages and Thesauri, Database Design, and Principles of Software Evaluation. His research interests cover information storage and retrieval, development of indexing languages, and computer applications. He has published 7 books and more than 130 journal and conference papers and presentations. He led the development of the Alcohol and Other Drug Thesaurus and was involved in the CliMB project. He served as a reviewer of the DELOS Network of Excellence on Digital Libraries) for the European Commission.

Royal School of Library and Information Science, Denmark
Marianne Lykke Nielsen, Associate Professor
Role: expert researcher in the area of knowledge organization systems and user-based evaluation (WP 1,**2**,**3**,4,**5**,**6**)
Marianne Lykke Nielsen has agreed to collaborate as a **non-funded supporting partner** (expenses only), with particular regard to planning, implementing and analysing evaluation. She also made a crucial contribution to the EnTag - Enhanced Social Tagging project. She lectures and researches on the design and evaluation of systems for knowledge organisation. She is co-PI in the US NSF Pathway Project that investigates user-centred indexing methods based on semantic components of documents. She is collaborating on evaluation and user aspects of the Glamorgan STAR project, co-edited the 2006 NRHM special issue on KOS and co-organised NKOS workshops in 2004, 2005, 2006, 2007, and 2008.

University College London
Vanda Broughton, Senior Lecturer
Role: expert researcher in the area of knowledge organisation systems (WP 1,**2**,3,4,**5**,6)
Vanda Broughton has agreed to collaborate as a **non-funded supporting partner** (expenses only) as advisor. She is the author of a number of books and papers on faceted classification and controlled vocabularies in digital environments. Her current research work focuses on the development of a general theory of facet analysis and the automatic generation of thesaural and systematic structures from core terminologies. She is Joint Editor of BC2, Associate Editor of the UDC, a member of the UK Classification Research Group, sometime member of the IFLA Committee on Classification & Indexing, and the Chair of the ISKO UK. Especially relevant to this project is her experience as a collaborator on the JISC Metadata Generation for Resource Discovery project.

OCLC Office of Research, USA
Diane Vizine-Goetz, Senior Research Scientist
Role: support provider for Scorpion (WP 1,2,**3**,**4**,5,6)
Diane Vizine-Goetz has agreed to collaborate as a **non-funded supporting partner**, with particular regard to Scorpion implementation and evaluation. She is lead researcher on the Terminology Services research project and is a member of the OCLC team conducting research involving the Functional Requirements for Bibliographic Records (FRBR) model. She joined OCLC in 1983 and has conducted research on the development of classifier-assistance tools and the application and use of the Library of Congress Subject Headings in online systems.

7 References

Baird, K. (2006). Automated Metadata. www.jorum.ac.uk/docs/pdf/automated_metadata_report.pdf

Charlesworth, A. (2008). Feasibility study into approaches to improve the consistency with which repositories share material (http://ie-repository.jisc.ac.uk/256/1/jisc-clax-final-report-repocon.pdf).

Chen, H., and Dumais, S.T. (2000), "Bringing order to the Web: automatically categorizing search results", *Proceedings of the ACM International Conference on Human Factors in Computing Systems, Den Haag*, pp. 145-152.

Golub, K. (2006a). Automated subject classification of textual Web documents. Journal of Documentation, 62,3, pp.350-371.

Golub, K., Ardö, A., Mladenić, D., and Grobelnik, M. (2006)c, "Comparing and combining two approaches to automated subject classification of text", *Proceedings of 10th European Conference on Research and Advanced Technology for Digital Libraries, Alicante, Spain, 17-22 September*, pp. 467-470.

Greenberg, J.; Spurgin, K.; and Crystal, A. (2005). Final Report for the AMeGA (Automatic Metadata Generation Applications) Project. Available at http://www.loc.gov/catdir/bibcontrol/lc_amega_final_report.pdf

Hagedorn, K. (2001). Extracting Value from Automated Classification Tools: The Role of Manual Involvement and Controlled Vocabularies. ACIA White Paper. http://argus-acia.com/white_papers/classification.html

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.

Lancaster, F.W. (2003), Indexing and abstracting in theory and practice (3rd ed.). London: Facet.

Markey, K. (1984). Interindexer consistency tests: A literature review and report of a test of consistency in indexing visual materials. Library and Information Science Research, 6, 155-77.

Olson, H. A., and Boll, J. J. (2001). Subject analysis in online catalogs (2nd ed.). Englewood, CO: Libraries Unlimited.

Polfreman, M., and Rajbhandari, S. (2008). MetaTools - Investigating Metadata Generation Tools - Final Report. http://ie-repository.jisc.ac.uk/258/

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.

Soergel, D. (1994). Indexing and Retrieval Performance: The Logical Evidence. JASIS 45(8): 589-599.

Svenonius, E. (1997), "Definitional approaches in the design of classification and thesauri and their implications for retrieval and for automatic classification", *Proceedings of the Sixth International Study Conference on Classification Research,* pp. 12-16.

Svenonius, E. (2000), *The intellectual foundations of information organization,* MIT Press, Cambridge, MA.
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.
Yang, Y. (1999). An evaluation of statistical approaches to text categorization. Journal of Information Retrieval, 1(1/2), 67-88.
Yilmazel, O., Finneran, C. M., and Liddy, E. D. (2004): Metaextract: an NLP system to automatically assign metadata. In: JCDL04: Proceedings of the 4th ACM/IEEE-CS Joint Conference on Digital Libraries 2004. pp. 241-242.

Appendix: Supporting Letters

Letters of support from the following institutions are attached:

1. UKOLN, University of Bath;
2. University of Glamorgan;
3. Intute (MIMAS, University of Manchester);
4. City University London;
5. Dagobert Soergel (University of Maryland);
6. Royal School of Library and Information Science, Denmark;
7. University College London;
8. OCLC Office of Research, USA;
9. National Centre for Text Mining (NaCTeM)