EnTag: Enhancing Social Tagging for Discovery

Koraljka Golub
UKOLN, University of Bath
BA2 7AY
Bath, United Kingdom
+44 (0)1225 383691
k.golub@ukoln.ac.uk

Catherine Jones, Brian Matthews,
Bartłomiej Puzoń
STFC Rutherford Appleton Laboratory
OX11 0QX
Chilton, Didcot, United Kingdom
+44 (0)1235 446648
{catherine.jones,brian.matthews}@stfc.ac.uk

Jim Moon, Douglas Tudhope
Faculty of Advanced Technology,
University of Glamorgan
CF37 1DL
Pontypridd, United Kingdom
+44 (0)1443 482271
{jnjmoon,dstudhope @glam.ac.uk}

Marianne Lykke Nielsen
Royal School of Library and Information Science
Fredrik Bajers Vej 7K
9220 Aalborg Øst, Denmark
+45 (0)98 157922
mln@db.dk

ABSTRACT
The EnTag (Enhanced Tagging for Discovery) project investigated the effect on indexing and retrieval when using only social tagging versus when using social tagging in combination with suggestions from a controlled vocabulary. Two different contexts were explored: tagging by readers of a digital collection and tagging by authors in an institutional repository; also two different controlled vocabularies were examined, Dewey Decimal Classification and ACM Computing Classification Scheme. For each context a separate demonstrator was developed and a user study conducted. The results showed the importance of controlled vocabulary suggestions for both indexing and retrieval: to help produce ideas of tags to use, to make it easier to find focus for the tagging, as well as to ensure consistency and increase the number of access points in retrieval. The value and usefulness of the suggestions proved to be dependent on the quality of the suggestions, both in terms of conceptual relevance to the user and in appropriateness of the terminology. The participants themselves could also see the advantages of controlled vocabulary terms for retrieval if the terms used were from an authoritative source.

Categories and Subject Descriptors
H.3.1 [**INFORMATION STORAGE AND RETRIEVAL**]: Content Analysis and Indexing – *indexing methods, thesauruses*.

H.3.3 [**INFORMATION STORAGE AND RETRIEVAL**]: Information Search and Retrieval

H.3.7 [**INFORMATION STORAGE AND RETRIEVAL**]: Digital Libraries

General Terms
Design, Experimentation, Performance.

Keywords
Social tagging, folksonomies, subject indexing, controlled vocabularies, Dewey Decimal Classification, ACM Computing Classification Scheme, Intute, digital collection, institutional repository, retrieval.

1. INTRODUCTION
Controlled vocabularies such as classification schemes and thesauri have been used as tools for information discovery and retrieval in libraries and abstracting and indexing services, some for more than a century. They are designed to reduce the ambiguity of natural language when describing and retrieving documents, and to allow access via browsing and navigation. However, manual indexing and classification pose significant resource costs, especially when performed by trained indexers. Automated means and social tagging are often suggested as the two possible solutions. Both, however, have disadvantages and, depending on the purpose of use or context, require additional manual input.

Many of the existing social tagging applications have not been designed with information discovery and retrieval in mind. In this context people tend to use tags mainly to organize their documents [34]. The resulting folksonomies (collections of tags) lack even basic control of word forms such as spelling variants, synonyms and disambiguation of homonyms ([10], [31]). The need for controlled vocabularies in relation to folksonomies has been reported in the literature; at the same time, end-user natural language tags could cover aspects that are not available in a controlled vocabulary, especially when it comes to new concepts ([25], [29]).

EnTag (**En**hancing **Tag**ging for Discovery) [6] was a one-year UK's JISC-funded [18] project that investigated ways of enhancing social tagging via controlled vocabularies, with a view

to improving the quality of tags for increased information discovery and retrieval. Objectives were the following:

- Investigate indexing aspects when using only social tagging versus when using social tagging in combination with a controlled vocabulary;

- Investigate the above in two different contexts: tagging by readers and tagging by authors; and,

- Investigate the influence on retrieval of only social tagging versus social tagging in combination with a controlled vocabulary.

The remainder of this paper discusses related work (section 2), the methodology for the two user studies (section 3), results (section 4), and conclusions with recommendations for future steps (section 5).

2. RELATED WORK

The need for controlled vocabularies in relation to folksonomies has been reported in the literature. Weller [36] compares ontologies and folksonomies, suggesting that they are not to be seen as rivals but complementary to each other. Noruzi [24] provides seven arguments for why a folksonomy-based system should use a thesaurus, emphasizing that it is impossible to maintain consistency over time or across folksonomy users without a thesaurus.

Several social tagging applications and projects have recognized the problem of folksonomies and tried to implement solutions outside the area of traditional controlled vocabularies. For Connotea [3], a service for organizing references, an add-on tool has been developed which allows taggers to select terms from a controlled vocabulary [7]. ZigTag [37] is a social bookmarking service that provides disambiguating definitions of entered terms from which the user can choose when tagging or searching. Marchetti et al. [21] describe the SemKey system in which they use WordNet and Wikipedia to introduce unique concepts. They conclude that there are problems with both: WordNet contains parts of speech information and structured network of relations between them, but lacks data for proper names disambiguation; Wikipedia on the other hand has strong proper names coverage but lacks a structured set of relations between the concepts described. Faviki [8] is another social bookmarking tool that uses DBpedia, Wikipedia-based RDF data, as unique concepts. The Meaning Of A Tag (MOAT) project [22] goes along similar lines. Al-Khalifa [1] combined folksonomies and domain ontologies to automatically annotate educational documents. ZoneTag [23] is a mobile application for images which allows them to be uploaded to Flickr, at which point tags suggestions are provided. The suggestions are based on tags assigned to similar images and similar contexts such as location (derived using cell-tower or GPS) and user's network.

Research aimed at integrating the worlds of folksonomies and controlled vocabularies has also been conducted. Smith [30] compared folksonomies in LibraryThing [20] and Library of Congress Subject Headings (LCSH) [19] and discussed the advantages and disadvantages of each. Trant and Bearman [35] compared end-user tagging of art against existing professional terms and found that the former increases access points. Hayman [12] argued for a combination of the best of the two worlds and described its application on an Australian collection of education

resources (Education Network Australia (edna)). While in edna controlled vocabularies are used for metadata creation and searching, a proof of concept was developed where users can also tag resources by choosing from an established taxonomy or by entering their own terms. Users' own terms are used later to feed back into the taxonomy to improve its quality. Smith [28] discusses end-user tagging and its potential for use in academic libraries, where the use of tags can be complementary to catalogue records, as given by examples of the universities of Pennsylvania and Huddersfield. The Library of Congress is collaborating with Flickr [9], in order to enhance bibliographic records for its images by end-user tags [26]. The HarvANA (Harvesting and Aggregating Networked Annotations) system enables authoritative metadata generated by traditional cataloguing methods to be merged with community annotations and tags ([11], [14]). The project employs an RDF model for representing the annotations/tags and OAI-PMH to harvest the annotations/tags from distributed community servers, which are then aggregated with the authoritative metadata in a centralized metadata store.

However, as known to the authors, no research has investigated the enhancement of social tagging with suggestions from a controlled vocabulary in a user trial with existing retrieval applications, comparing social tagging and enhanced tagging.

3. METHODOLOGY

The main focus of the project was to examine whether and, if so, how social tagging can be enhanced with suggestions from a controlled vocabulary. Two different contexts were explored: tagging by readers (Intute subject gateway [17]) and tagging by authors submitting papers to a repository (Science and Technology Facilities Council (STFC) repository [32]). For each context separate demonstrators were developed and user studies conducted. This allowed a general analysis of two different user communities, interfaces, digital collections and controlled vocabularies. The Dewey Decimal Classification (DDC) [4] was used in the former and ACM Computing Classification Scheme [33] in the latter.

3.1 Intute Study

3.1.1 Intute Demonstrator

The Intute subject gateway [17] is a UK-based database of hand selected and catalogued Web documents for education and research. For each document the catalogue record comprises title, description, controlled keywords, uncontrolled keywords, classification, type, URL, format, etc. (see [16] for detailed cataloguing policy). In the study, 11,042 Intute catalogue records were chosen, all from the area of politics. Records were classified by the DDC, IBSS (International Bibliography of the Social Sciences) thesaurus [15] and HASSET (Humanities and Social Science Electronic Thesaurus) [13]. Uncontrolled keywords mostly comprise names of countries.

The Intute demonstrator, developed at the University of Glamorgan, is a .NET application, employing Visual Studio, C# and ASP .NET. It operates over the extract of the Intute catalogue records held on the Glamorgan server. Original indexing and classification have been stripped out for purposes of the study. The URL to the web resource is provided so that the user can view the actual document.

In the Intute demonstrator, the controlled vocabulary from which suggestions are provided to one of the interfaces is the DDC

which also comprises mappings to Library of Congress Subject Headings (LCSH). DDC contains so-called class captions (names of classes) and relative index terms pointing to the captions (entry terms represented by class captions). Thus, the suggestions offered come from three sources: DDC class captions, DDC relative index terms, and LCSH terms. It was decided to retain the original DDC terms in suggestions for purposes of the study (see section 4.1.2).

The Intute demonstrator [4] comprises three major interfaces: searching, simple tagging, and enhanced tagging. Once a person logs in, he/she arrives to the searching interface (Figure 1). The searching interface provides the following features:

1) Main Tag Cloud: a tag cloud with tags linked to documents to which they were assigned. It is an alphabetical list of all tags in the demonstrator, with different font sizes relative to popularity. Filter By drop-down menu on top offers the My Tags option which presents the current taggers's tags only. By default everyone's tags are shown (Everyone's Tags).

2) Taggers: a cloud of names of taggers linked to documents they indexed.

3) A free-text search box, with an option to limit searching to tags, title and description fields.

The documents found are shown in the Results pane. They are automatically ranked according to the MySQL full-text natural language search. Title, description (from Intute) and existing tags are shown here. By clicking on the Title, the URL opens in a new window.

Figure 1. Intute demonstrator's searching interface

Once a document is selected from search results, clicking on the "Tag" button will return a tagging interface. Here title, URL, and description are displayed. Two tagging interfaces are provided: 1) the Simple Tagger, with tagging features common in popular social tagging applications; and 2) the Enhanced Tagger, with additional suggestions from the controlled vocabulary. (In the study the log-on screen provides the choice of the interfaces). Both tagging interfaces have the following options from which to select tags (Figure 2):

1) Main Tag Cloud.

2) Taggers: names of taggers linked to tags they have used, the latter listed in the All {Tagger Name}'s Tags pane; and,

3) My Tags For This Document.

By clicking on a selected tag, the tag will be shown in the text box. By pressing the Tag Document button the tag will be added

to the document as well as listed in the "My Tags for This Document" pane. A tag can also be typed in.

Enhanced Tagger (Figure 2) additionally provides suggestions from the controlled vocabulary, presented in three panes at the bottom of the screen. In the first pane to the left ("Automatically suggested matches, Find appropriate context(s)"), DDC classes are listed. They are automatically derived by a string-matching comparison of DDC vocabulary to a user-entered term from the text box above the panes upon clicking the Suggest button. Immediately after the user comes to the enhanced tagging page, initial suggestions are automatically generated by treating the document's title as if it had been entered as a tag.

If the user clicks on one of the listed classes from the first pane, its narrower and broader classes are shown in the second pane ("Explore hierarchy around the selected context"), allowing interactive browsing of the hierarchical context. Simultaneously, in the third pane ("Select/edit relevant tags") a tag-cloud-like list of DDC captions, DDC relative index terms and LCSH mapped terms is presented as a source of suggestions from which the user may select a tag. Selecting a tag copies it to the text box, where it can be further edited; pressing the Tag Document button adds the tag to the document.

Figure 2. Intute demonstrator's enhanced tagging interface

What follows is an example of how a tag is chosen from suggestions in the enhanced interface. After searching for the term "slavery" in all fields, a list of documents is returned. The document chosen is "Slavery in New York", an online exhibition on history of slavery in New York State (http://www.slaveryinnewyork.org/). By clicking on the "Tag" button, the enhanced tagging interface opens. In the text box in the middle of the screen the tag "slavery" is entered and the "Suggest" button clicked, to get DDC-based suggestions. As shown in Figure 2, the first pane then displays 19 matching classes. Two of these belong to political science. The "End of Political Science" line means that the classes below it belong to different subject areas from DDC, which could be related to politics. Mousing over each caption will display its immediate superordinate class. The top found classes are the following:

Slavery and emancipation *(superordinate class: Political science)*

Emancipation *(superordinate class: Slavery)*

----- *End of political science* -----

Discriminatory practices and slavery *(belonging to Ethics of social relations)*

Extension of slavery *(belonging to Administration of Abraham Lincoln)*

Slavery *(belonging to Systems of labor in Economic institutions)*

American fiction-1830-1861, ... *(belonging to American literature)*

Etc.

Since the document is about the slave trade and emancipation movement, "Slavery and emancipation" is chosen, which results in the second pane containing "Slavery and emancipation" highlighted, surrounded by its broader and narrower classes, each hyperlinked to further broader and narrower classes. The third pane lists tagging suggestions, those being captions, relative index terms, and LCSH mappings for the term. Clicking on a suggestion will automatically have it appear in the text box. There one can further edit it before clicking on the Tag Document button.

For our document the tags chosen from the third pane are:

Antislavery movements - United States

Emancipation

Slave trade

Slavery

Thus, apart from the initial "slavery" tag, by using the DDC suggestions, three other tags are gained. Furthermore, unlike in common tagging applications, these terms are controlled.

3.1.2 Intute User Study
In order to compare simple and enhanced tagging an exploratory user study was conducted. It covered two main questions: 1) how a tag is chosen in the two different interfaces; and, 2) what are the retrieval implications.

The study took place remotely at participants' homes using their own computers, supported by communication via email. Initially, there were 61 participants who signed a participation consent form, and 54 who completed a pre-study questionnaire. Six participants took part in pilot testing. There were 42 participants who did at least some tagging but not everyone completed the whole study: 28 of them finished all tasks. The results reported here are based on these 28 participants' input.

The participants were politics students from British universities with one from the European University Institute. Based on a pre-study questionnaire, it was found out that they were experienced Web users. Half of them had used tagging applications before but conducted little tagging. Almost a third had some acquaintance with DDC; there were none with any acquaintance with LCSH or other controlled vocabularies. The majority have never used Intute.

The main data collection method was logging the participants' steps in the demonstrator. In order to contextualize and understand the results further, three groups of questionnaires were also used: 1) a pre-study questionnaire, which served to collect background information about the participants; 2) post-task questionnaires, to gather experience about each task as to familiarity with the topic, tagging in the task and interface features; and, 3) a post-study questionnaire, about the whole study and interfaces in general.

The procedure of the study comprised several steps. After signing the participation consent form and completing the pre-study questionnaire, the participants were given the training document through which they were to acquaint themselves with the demonstrator and tagging. Then they received the instructions document which described the steps they were supposed to take.

The steps included 4 tagging tasks, each comprising 15 documents. The tasks were on topics from politics; two were pre-defined (controlled) and two on a self-chosen topic (free). The controlled task for simple tagging was on the topic of "European integration", and for enhanced tagging on "peacekeeping". A hypothetical group project scenario was outlined as a rationale and motivation for the tagging activity. The scenario for the controlled enhanced tagging task follows:

"Imagine that as part of one of your courses, you are asked to write a four-page essay on the topic of European integration, as a joint project in groups of four. The essay should critically discuss existing theories about the creation of the European Union and its institutions. Your lecturer has instructed you to look for resources in the EnTag system. Since you will be working together with three other students, you should tag the documents you retrieve with tags that would be useful to you but would also enable other students to find those documents in EnTag and understand from your tags what the documents are about."

In each task they were first instructed to search for documents and then tag 15 of them. In controlled tasks they were asked to tag the top 15 retrieved documents; in free tasks they could choose any 15 that they found relevant. The analysis showed that 53 documents were tagged at least once in the controlled task, and 41 in the simple task (instead of 30 in each). This is because participants did not closely follow the 'top 15' instruction and because the URLs of some of the top 15 retrieved documents were temporarily unavailable. They were instructed to spend between 5 and 10 minutes on tagging each document, and to tag as many aspects and topics they thought appropriate for the task. They were also asked to open the URL, but did not need to follow further links. In case of long documents, they were instructed to focus on the abstract, introduction, conclusion, headings and table of contents. When tasks were being completed in the interface enhanced with suggestions from the controlled vocabulary, the instruction was to try to consider those suggestions if appropriate (they did not have to follow the suggestions).

The order of tasks was rotated in order to reduce the learning influence.

3.2 STFC Study
3.2.1 STFC Demonstrator
The Tagger interface is supplied in conjunction with the STFC's repository's ePubs [32] metadata editor so that tags can be entered for a specific publication by its authors. The figure below (Figure 3) shows a screen shot of a typical tagging screen.

The screen is divided into four main areas:

1) At the top-centre, the title and abstract of the publication selected for tagging are displayed.

2) At the bottom-centre, a browse interface for the controlled vocabulary is shown. Initially top-level terms are shown. Clicking on a term will show its narrower and related terms. The current path to the top of the hierarchy is always shown as a 'breadcrumb' trail along the top of the hierarchy. Terms can be selected as tags by clicking on the "+" symbol to the left of each term. Apart from browsing the controlled vocabulary, one can also search it (the "Search thesauri" link).

3) To the left a tag cloud is displayed, with tags ranked in order of descending use frequency. They can be selected by either clicking on them (if they are free-text terms) or clicking on the spyglass symbol to their left (if they are from the controlled vocabulary). This will enter them into the "Add" term box to the right of the screen, where they can be accepted as a tag for the paper. The tag cloud as a default shows the terms used by the current author. "Show all" will display all authors' tags.

4) To the right, the current selected free-text (top) and controlled vocabulary terms (bottom) are shown. They can be deselected by clicking on the "—" sign to their left. In the centre of the panel, there is a free-text box, where the user can enter free-text terms. Multiple terms can be entered by separating them with commas.

Once a suitable selection of tags has been made using the tool, the user can accept them by clicking on the "Confirm" button at the bottom of the screen.
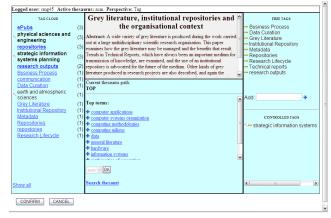

Figure 3. STFC demonstrator interface

The ACM Computing Classification Scheme was the chosen controlled vocabulary. Since its main purpose is to classify papers which are submitted to various ACM journals, it has a widespread awareness and authority within the computing community. It was first imported into SKOS [27] and then into the tool.

The STFC demonstrator is an Apache Cocoon application using combined Java and XML techniques, the underlying database being Oracle. It links dynamically to the ePubs institutional repository so that once the user performs a search in the repository, in a specially adapted edit mode there is an option to enter the Tagger system. Once the system is started, the title and abstract of the work are transferred together with any existing free tags.

3.2.2 STFC User Study

The STFC user study involved 10 participants working in the fields of computer science and information technology, who have deposited more than 10 papers within the repository. None of the participants identified themselves as frequent taggers. Two participants had some knowledge of the ACM Computing Classification Scheme before the tests, though several others were aware that it existed.

The sample of authors being by necessity small, this study sought to explore the purpose and ways of tagging from an author's perspective in a qualitative manner. The issues dealt with included the following: 1) what they feel the purpose of tags is and how they should be used; 2) whether they consider using controlled

vocabularies a worthwhile exercise over free tagging; and, 3) whether the user interface is intuitive and easy to use.

Data collection included automatic logging of tagging activities, observation and a semi-structured interview. The study comprised supervised sessions with individual participants. Each participant was given a worksheet with notes and guidance on the task. The task itself lasted approximately 20 minutes and involved using the tagging interface to select tags for a number of the author's papers. Initially the participants were free to choose tags from any of the three approaches offered (free text, controlled vocabulary, tag cloud). Later on the observer could suggest that the participant try another tagging approach in order to gauge their response to a method different to their first preference.

4. RESULTS

4.1 Intute

4.1.1 Number of Tags
As seen from Table 1, in total 7568 tags were assigned in both interfaces and in all the tasks. More tags were assigned in Simple Tagger than in Enhanced Tagger, which could be explained by the fact that the participants, as they reported, spent more time in Enhanced Tagger due to exploring the different controlled suggestions, instead of adding the first tag that came to their mind. In Enhanced Tagger more tagging activity took place when the topic was freely chosen.

In controlled tasks on average 41 tags per document were assigned (for the 94 documents that were tagged), and in free tasks 5 tags per document (for the 751 documents – many more because typically fewer users viewed a given document since the search was freely chosen). This is similar to findings of [2] where on average four tags were assigned. Each participant assigned on average 278 tags in total, a few more in Simple Tagger and a few more in the free task, probably due to faster tagging process in Simple Tagger and more motivation in the free task.

Table 1. Number of tags

	Simple	Enhanced	Total
Tags in total	4022	3546	7568
Controlled task	2025	1688	3713
Free task	1997	1858	3855
Tags per document (controlled)	avg. 49 (41 docs)	avg. 32 (53 docs)	avg. 41 (94 docs)
Tags per document (free)	avg. 5 (374 docs)	avg. 5 (377 docs)	avg. 5 (751 docs)
Tags per tagger (controlled)	avg. 72	avg. 63	total avg. 135
Tags per tagger (free)	avg. 74	avg. 69	total avg. 143

4.1.2 Choosing a Tag
As seen from Table 2, in simple tasks the majority of tags (90.9%) are assigned by typing them directly in (Typing Own Tag), as common in social tagging applications. In enhanced tasks 71.2% are typed in directly, while 16.9% come from DDC-based suggestions instead (Dewey Tag).

Table 2. Choosing a tag

Activity	Simple		Enhanced		Total	
Typing Own Tag	3656	90.9%	2525	71.2%	6181	81.7%
Main Tag Cloud	94	2.3%	88	2.5%	182	2.4%
Own Tag	0	0%	32	0.9%	32	0.4%
Certain Tagger's Tag	272	6.8%	303	8.5%	575	7.6%
Dewey Tag			598	16.9%	598	n/a
Total	4022		3546		7568	

The instructions for the enhanced tasks encouraged the participants to consider the suggestions if they thought them appropriate. While the figures should not therefore be considered a simple measure of popularity, some participants commented favourably on suggestions (see section 4.1.4 Post-Questionnaires). Selecting from another tagger's tags is a common feature in popular social tagging applications, and has been used in both of our systems, too – on average 7.6% of all tags assigned (Certain Tagger's Tag). Main tag cloud is yet another common feature, but has hardly been used at all (see section 5 Conclusions).

Table 3 shows where the participants looked for potential tags. In simple tagging the most frequent feature was others' tags, 72.9% in total (Certain Tagger's Name (18.9%) and Certain Tagger's Tag (54.0%)). Main Tag Cloud and Own Tag were less used (16.8% and 10.3% respectively). In enhanced tagging most frequent activities were the ones related to exploring DDC suggestions, 74.0% in total. Of other features, as in simple tagging it was others' tags that were most popular, 18.2% in total.

Table 3. Browsing for tags

Activity	Simple N = 614	Enhanced N = 2468
Main Tag Cloud Clicked	16.8%	5.1%
Own Tag Clicked	10.3%	2.7%
Certain Tagger's Name Clicked	18.9%	3.7%
Certain Tagger's Tag Clicked	54.0%	14.5%
Dewey Suggest Button Clicked		28.9%
Dewey TreeView Clicked (1st pane)		13.7%
Dewey Hierarchy Clicked (2nd pane)		3.2%
Dewey Suggested Tags Clicked (3rd pane)		28.2%
	100.0%	100.0%

Results showed that the 'raw DDC' is not necessarily appropriate to present online to end users. For example a number of DDC captions contain quite complex phrases, some of which are 'built terms' and require hierarchical context. Certain captions are unclear without displaying their coordinate, superordinate or subordinate classes. This can be due to captions' polysemic nature, brevity etc. In the demonstrator displaying superordinate classes has been implemented through mousing over the class. Suggestions to deal with problematic terms for a future study or demonstrator are being discussed in cooperation with OCLC. One

possible solution for captions that are unclear when on their own would be to pre-coordinate them with a term from the Relative Index to provide context. For example, the caption "Other conditions of employment" could be transformed into "Pensions and other conditions of employment" [Panzer, personal communication].

4.1.3 Retrieval Implications

4.1.3.1 Tags in Documents' Catalogue Records

End-user tags were compared against three Intute catalogue records fields: document's title, URL and (manually created) description. These are the fields the participants could see when tagging. Because most of the catalogue records originally had manually assigned controlled keywords, and more than half of them also uncontrolled keywords (both hidden to the study participants), end-user tags were compared against the manually assigned keywords as well. The comparison was case-insensitive while spelling variations were not considered, i.e., each spelling variation was counted as a different term.

The analysis showed that end-user tags, both from Simple and Enhanced Tagger, are found in 64% of tagged documents' title, URL or description. So, for 36% of the documents tags potentially present additional access points.

Table 4. Number of documents' records where tags found

	Simple		Enhanced		In Both	
Tags in Title	200	48.3%	174	41.8%	350	44.9%
Tags in URL	76	18.4%	72	17.3%	137	17.6%
Tags in Description	262	63.3%	232	55.8%	468	60.0%
Tagged Documents in Total	414		416		780	

As seen from Table 4, most tags are found in the description (it being the longest piece of the catalogue record), followed by title, and then by URL. Looking at the three fields, there are on average 5% less tags from Enhanced Tagger found in the catalogue records than there are from Simple Tagger. In addition, out of 143 documents to which tags from DDC suggestions were assigned, only in 17 documents were they found in title (12%), 2 in URL (1.4%), and 25 in description (18%). This suggests that, in retrieval, enhanced tagging could potentially provide more access points (that are not present in the document text) than simple tagging when used in combination with a full text search engine.

Table 5 shows the degree to which simple and enhanced tags overlap with pre-assigned Intute keywords. Both simple and enhanced tags show very low overlap. The top two rows show the degree to which tags are the same as controlled and uncontrolled keywords, 5.5% in total. Third and fourth row include the exact matches and also situations where tags are parts of keywords, in which case there is 7.5% overlap in total. This small overlap indicates that more than 90% of tags could serve as additional access points beyond the original Intute indexing.

Table 5. Tags versus pre-assigned (un)controlled keywords

	Simple		Enhanced		In Total	
Tags like controlled keywords	125	3.1%	119	3.4%	244	3.2%
Tags like uncontrolled keywords	60	1.5%	39	1.1%	99	1.3%
Tags parts of controlled keywords	251	6.2%	175	4.9%	426	5.6%
Tags parts of uncontrolled keywords	84	2.1%	60	1.7%	144	1.9%

Of the tags selected from DDC, only a few overlap with existing Intute controlled keywords (these came from DDC, IBSS and HASSET thesauri): out of 598 DDC tags, 12 are contained in (or the same as) controlled keywords. These are the following:

Controlled keyword	DDC tag
peace keeping	Peace
foreign policy	Foreign policy
civil-military relations	Civil-military relations
intelligence	Intelligence
local government	Local government
Democracy	Democracy
Diplomacy	Diplomacy
nationalism	Nationalism
rwanda	Rwanda
european union	Europe
socialism	Socialism
maps	Maps

4.1.3.2 Search Terms
Table 6 compares tags and pre-assigned keywords to search terms the participants used in their tasks.

Table 6. Search terms versus tags and keywords

	Simple N=98	Enhanced N=122
Nbr of tags in which search terms are found	249	254
Nbr of controlled keywords in which search terms are found	91	40
Nbr of uncontrolled keywords in which search terms are found	111	101

There were 98 search terms used in Simple Tagger, and 122 in Enhanced Tagger. In Simple Tagger, almost three times as many search terms are found in tags as in controlled keywords, and twice as many as in uncontrolled keywords. The difference is even bigger for Enhanced Tagger. The hypothesis that there is a greater probability of finding search terms in tags is an interesting avenue of future work. However, caution must be exercised since

the study context might naturally encourage the use of search terms in tags, since tagging was always preceded by a search.

4.1.3.3 Simple vs. Enhanced Tags
A full qualitative analysis of the differences between tags derived from Simple and Enhanced Taggers is under way. A suggestive example is presented to illustrate one of the main lines of inquiry. The example is a document where in Simple Tagger at the end of the study there was only one tag assigned by six different users: "afghanistan". In Enhanced Tagger the tags assigned were the following (the number after comma indicates the number of times the tag was assigned, and the ones in bold are DDC tags):

> Afganistan, 2
>
> **Afghan War, 2001-, 1**
>
> afghanisan and liberia, 1
>
> **Afghanistan, 5**
>
> **Afghanistan - 20th century, 1**
>
> Afghanistan - Politics and government, 1
>
> **Afghanistan - Politics and government - 1973-, 1**
>
> afghanistan and liberia, 1
>
> Afghanistan and Liberia - Civil rights, 1
>
> afghanistan casestudy, 1

This illustrates that the suggestions from Enhanced Tagger can describe the document by more facets: time (20th century), specific event (Afghan war), politics and government.

4.1.4 Post-Questionnaires
After each task, the participants completed a post-task questionnaire. In both Simple and Enhanced Tagger, the majority were on average familiar with the topic of the task, they found it easy to choose tags, were satisfied with tags assigned, and were certain that they assigned the tags correctly.

Of the tagging-support features, they found the following ones helpful:

- Listing of own tags;
- DDC disambiguation tree (first panel on the left bottom part of the screen); and,
- DDC/LCSH suggestions.

The ones they did not find as useful were:

- Main tag cloud;
- Clickable names of others; and,
- DDC hierarchical browsing for narrower or broader classes (bottom middle pane).

At the very end, the participants filled in a final questionnaire about the whole study. The majority enjoyed the study and believed that a similar system would be useful in real life. They thought it extremely or very easy to learn and use Simple Tagger and somewhat or very easy to learn and use Enhanced Tagger. Of the comments about Simple Tagger, they liked freedom to have own choices, its simplicity and speed. They did not like the fact there were no suggestions and the need to scroll in the main tag cloud. About Enhanced Tagger they liked suggestions when they thought them useful; they found them time-saving and valuable in particular when they were unsure of which tags to assign; and, they considered them important for consistency. One participant preferred simple tagging when indexing a document on a familiar

topic and enhanced tagging when a document is on an unfamiliar topic, suggesting how both are useful in different contexts. What they did not like about Enhanced Tagger were sometimes irrelevant suggestions, there were several comments on the cluttered interface and the number of interaction steps required to tag a suggestion. One or two experienced excessive scrolling when employing browser platforms the system was not optimised for. A few suggested that a streamlined version of Enhanced Tagger would be useful, with easier interaction and more selective focused suggestions.

4.2 STFC

In the study on average 6 tags per item were assigned and 67% of the total number of tags assigned were free terms. The participants could overall understand the importance of tagging for information retrieval purposes. There was a general pervading sentiment amongst them that choosing terms from a controlled vocabulary was a "good thing" and in fact better than own terms.

Most subjects claimed that they would be willing to use a tool similar to the one provided, albeit with some reservations about how this could be realised in practice, and proposed suggestions on the interface, and additional automation. The tag cloud was not a success: most did not use it, and those that considered it found it confusing or overloaded with options.

The suitability of the proposed controlled vocabulary was also an issue with most subjects. While they recognised that this was a well-known vocabulary used in computer-science publishing, they questioned its usefulness to accurately index their work for retrieval. Several wanted choices of vocabularies.

Most depositors had a strong preference for the way they interact with the system. The findings suggest three main groups:

1) Free text taggers: they enter many free-text terms, and don't care about the use of the controlled vocabulary.

2) Thesaurus browsers: they systematically browse the hierarchy of the controlled vocabulary, and only enter free-text term when the controlled vocabulary does not have a term they are comfortable with.

3) Thesaurus searchers: they prefer to interact with the controlled vocabulary via the search tool, then move to browsing and only enter free-text term when the controlled vocabulary does not have a term they are comfortable with.

4.3 Comparison of Intute and STFC Results

The subject groups of searchers/readers (Intute) and authors/depositors (STFC) clearly have different roles (although they can overlap). Also regular depositors tend to be more mature than the searchers, due to more experience within a discipline being required before authoring papers. Nevertheless, a number of similarities between the Intute and STFC users could be identified. In both contexts the following was the case:

- Users appreciated the benefits of consistency and vocabulary control and were potentially willing to engage with the tagging system;

- There was evidence of support for automated suggestions if they are appropriate and relevant;

- The quality and appropriateness of the controlled vocabulary proved to be important;

- The main tag cloud proved problematic to use effectively; and,

- The user interface proved important along with the visual presentation and interaction sequence.

5. CONCLUSIONS

The main focus of the project was whether and, if so, how social tagging can be enhanced with suggestions from a controlled vocabulary. Two different contexts were explored: tagging by readers (Intute) and tagging by authors submitting papers to a repository (STFC). For each context separate demonstrators were developed and user studies conducted. This allowed a general comparison of two different user communities, interfaces, digital collections and controlled vocabularies. The Intute study in particular focused on comparison between simple tagging and tagging enhanced by suggestions from a controlled vocabulary, the latter being DDC. The STFC study investigated from an author's perspective in a qualitative manner the purpose and ways of tagging, one way involving a controlled vocabulary (ACM Computing Classification Scheme).

The Intute study showed that most tags are added by typing them in directly, as is common in social tagging applications; of the other features used, the most frequent one is DDC suggestions, and another tagger's cloud. That the participants appreciate the DDC suggestions was also seen from their comments. In the STFC study some participants prefer typing their own tags in, while others use the ACM Computing Classification Scheme suggestions. They overall see that choosing terms from a controlled vocabulary is important for consistency in retrieval.

The main tag cloud, common in social tagging applications, is little used in both studies. Part of the problem is that there are too many options from which to choose and excessive scrolling due to the size of the cloud. This could be reduced by improved search/browsing and personalised ranking, clustering or filtering. However, it is not clear that displaying all tags is useful for retrieval purposes with large collections. In future work, we will investigate further single-document, all-user tag clouds, as well as tag clouds based on communities of friends/colleagues.

The Intute study also showed that in enhanced tagging, while users appreciated the 'direct' suggestions (third pane) and made some use of the disambiguation interface element (first pane), they did not browse the Dewey hierarchy very much (middle pane). Further work is needed to explore when browsing functionality is desirable in this context.

In the Intute study both simple and enhanced tagging provided additional entry points beyond the original subject indexing, as well as compared to free text search engines. There was some evidence that, in particular, controlled suggestions provided additional access points beyond the literal text. That tagging contributes has also been found in earlier research ([35]).

User experience and task completion showed that both demonstrators were usable with little prior training. However comments showed that the Intute interface, particularly in the Enhanced Tagger was experienced as complex. By design the interface was cluttered because a variety of tagging features was to be tested. An operational system should have a simpler, less cluttered user interface, focusing on the key functionality and with user interaction streamlined.

Most participants from both studies claimed that they would be willing to use similar tools in real life. Because the results also showed the importance of controlled vocabulary suggestions, both

at the time of tagging (indexing) and retrieval, it is recommended that social tagging be allowed in repositories and other digital collections, supported by controlled vocabularies. However, it is important to further analyze, experiment and pilot test tools derivative from both Intute and STFC demonstrators. Further studies would involve the following:

1) Qualitatively analyze user tags, for example, whether they are different from those assigned by librarians in the Intute database, what topical facets are represented in the tags etc.

2) Further investigate user tagging styles and motivation for tagging. For users as searchers/readers, the rationale is less straightforward than with authors. In some situations, where a user is part of a natural community engaged in a joint endeavour (as in the scenario suggested in the Intute study), tagging content will serve for mutual benefit. In other cases, users may be motivated to act as good (informed) citizens and tag based on their desire to share expertise or enthusiasm. One aspect of this could be a consideration of the different types of tagging activity. For example, tags might express the genre or utility of a document for a user's purposes. To the extent that others share the same perspective, non-subject based tags might serve as useful access points for others, in addition to their potential in personalising access to a collection.

3) Conduct a retrieval test with appropriate performance measures of retrieval effectiveness. This is complex due to the interconnectedness of indexing, searching, motivation and evaluation.

4) Conduct extended (longitudinal) evaluation in live settings to gain information of such systems in real life.

5) Investigate different types of vocabularies and different domains. A different type of vocabulary, such as a thesaurus, may have different effects as a basis of tagging suggestions.

Furthermore, a number of enhancements to the functionality of the tools were proposed. There should be greater quality of automatic suggestions: they should be user-oriented as regards terminology, level of specificity, perspective and currency. The Intute demonstrator implemented a very crude form of automatic classification, in that the title of a document selected for tagging was fed through to the DDC matching system. The top ranked match yielded suggestions which automatically appeared in the suggested tag cloud, which worked remarkably well in some cases, but not in others. A more sophisticated automatic classification system would add considerable value to the quality of suggestions.

An auto-completion feature could also be provided which could include interactive term disambiguation and visualization. Some participants wanted a more graphical/visual representation of the controlled vocabulary, others asked for more information about the semantic meaning of a term in the controlled vocabulary and even other people's tags. Also, one could further study whether it is useful to structure the suggestions in a faceted check list of controlled terms.

Broadly speaking, this study has investigated the seeding of folksonomic tag clouds with controlled vocabulary elements. The resulting hybrid shows potential to increase access points to collections of information. Preliminary analysis suggests that vocabulary-based suggestions have potential to prompt taggers to escape the literal text, provided the suggestions are relevant and user oriented. This requires further analysis and work, including personalised filtering of suggestions and selection of appropriate

terminology for a concept, the expansion and adaptation of existing controlled vocabularies with social tagging data and the investigation of user-oriented methodologies for the development and construction of controlled vocabularies. If the hypothesis is borne out then vocabulary-based suggestions may potentially both encourage the description of resources by more topics (increased exhaustivity) and may also afford the capability of describing resources at a higher level of generalisation (the activity of classification). In other words, vocabulary-based suggestions might encourage both indexing and classification activities.

6. ACKNOWLEDGMENTS

7. REFERENCES

[1] Al-Khalifa, H. S. 2008. Automatic document-level semantic metadata annotation using folksonomies and domain ontologies. *ACM SIGWEB Newsletter*, Vol. 2008, Autumn (September 2008).

[2] Angus, E., Thelwall, M., and Stuart, D. 2008. General patterns of tag usage among university groups in Flickr. *Online Information Review*, 32 (1), 89-101.

[3] Connotea. 2008. http://www.connotea.org/.

[4] Dewey Services: Overview. http://www.oclc.org/dewey/overview/default.htm

[5] EnTag Intute demonstrator. http://reswin1.isd.glam.ac.uk/intute_test2/(S(gfpkqo55sh342i4 5zenuz545))/loginpage2.aspx

[6] EnTag project website. http://www.ukoln.ac.uk/projects/enhanced-tagging/

[7] Entity Describer. 2007. http://www.connotea.org/wiki/EntityDescriber

[8] Faviki. http://faviki.com/

[9] Flickr. 2008. http://www.flickr.com/

[10] Guy, M., and Tonkin, E. 2006. Folksonomies: Tidying up tags? *D-Lib Magazine*, January 2006. Available at http://www.dlib.org/dlib/january06/guy/01guy.html

[11] HarvANA: Harvesting and Aggregating Networked Annotations. http://www.itee.uq.edu.au/~eresearch/projects/harvana/

[12] Hayman, S. 2007. Folksonomies and tagging: New developments in social bookmarking. In Ark Group Conference: Developing and Improving Classification Schemes 27-29 June, Rydges World Square, Sydney. Available at http://www.educationau.edu.au/jahia/webdav/site/myjahiasite/s hared/papers/arkhayman.pdf

[13] Humanities and Social Science Electronic Thesaurus (HASSET). http://www.data-archive.ac.uk/search/hassetAbout.asp

[14] Hunter, J., Khan, I., and Gerber, A. 2008. HarVANA - Harvesting Community Tags to Enrich Collection Metadata. In

Joint Conference on Digital Libraries, Pittsburgh, PA, USA, June 16 - 20, 2008. http://www.itee.uq.edu.au/~eresearch/papers/2008/Hunter_JCDL2008.pdf

[15] IBSS thesaurus. http://www.lse.ac.uk/collections/IBSS/about/thesaurus.htm

[16] Intute – Policy. http://www.intute.ac.uk/policy.html

[17] Intute: Best of the Web. http://www.intute.ac.uk/

[18] JISC: Joint Information Systems Committee. http://www.jisc.ac.uk/

[19] Library of Congress Subject Headings. 2007. 30th ed. http://www.loc.gov/cds/lcsh.html#lcsh20

[20] LibraryThing. 2008. http://www.librarything.com/

[21] Marchetti, A., Tesconi, M., Ronzano, F., Rosella, M., and, Minutoli, S. 2007. SemKey: A semantic collaborative tagging system. In Tagging and Metadata for Social Information Organization, 16th International World Wide web Conference May 8-12, Banff, Aberta, Canada. Available at http://www2007.org/workshops/paper_45.pdf

[22] Meaning Of A Tag. http://moat-project.org/

[23] Naaman, M., and Nair, R. 2008. ZoneTag's collaborative tag suggestions: What is this person doing in my phone? *IEEE Multimedia*, vol.15 (3), pp. 34-40, July-Sept. http://infolab.stanford.edu/~mor/research/ZoneTagIEEEMM.pdf

[24] Noruzi, A. 2007. Folksonomies: Why do we need controlled vocabulary? Editorial. *Webology*, Vol. 4, No. 2. Available at http://www.webology.ir/2007/v4n2/editorial12.html

[25] Qin, J. 2008. Folksonomies and taxonomies: Where the two can meet. In New Dimensions in Knowledge Organization Systems, Washington DC, September 11. Available at http://nkos.slis.kent.edu/2008workshop/JianQin.pdf

[26] Raymond, M. 2008. My friend Flickr: A match made in photo heaven. Library of Congress Blog. Available at http://www.loc.gov/blog/?p=233

[27] SKOS Simple Knowledge Organization System - Home Page. http://www.w3.org/2004/02/skos/

[28] Smith, C. 2008. Adoption of social tagging in academic libraries: A leap of faith or a descent into the abyss? CILIP Cataloguing and Indexing Group Conference, The University of Strathclyde, Glasgow, 3 – 5 September. Available at http://www.ukoln.ac.uk/events/CIG/2008/conf-glasgow/presentations/ppt-2000-html/smith_files/v3_document.html

[29] Smith, G. 2008. Tagging: Emerging trends. The *Bulletin of the American Society for Information Science and Technology,* August/September, Vol. 34, No. 6. Available at http://www.asis.org/Bulletin/Aug-08/AugSep08_Smith.html

[30] Smith, T. 2007. Cataloging and you: Measuring the efficacy of a folksonomy for subject analysis . In Lussky, Joan, Eds. Proceedings of the 18th Workshop of the American Society for Information Science and Technology Special Interest Group in Classification Research, Milwaukee, Wisconsin. Available at http://dlist.sir.arizona.edu/2061/

[31] Spiteri, L. F. 2007. Structure and form of folksonomy tags: The road to the public library catalogue. *Webology*, Volume 4, Number 2, June, 2007. Available at http://www.webology.ir/2007/v4n2/a41.html

[32] STFC ePublication Archive. 2008. http://epubs.cclrc.ac.uk/

[33] The 1998 ACM Computing Classification System. http://www.acm.org/about/class/1998

[34] Tonkin, E., Corrado, E. M., Moulasion, H. L., Kipp, M. E. I., Resmini, A., Pfeiffer, H. D., and Zhang, Q. 2008. Collaborative and social tagging networks. *Ariadne* 54. Available at http://www.ariadne.ac.uk/issue54/tonkin-et-al/

[35] Trant, J., and Bearman, D. 2008. Public and professional vocabularies: Comparing user tagging with museum documents and documentation. The 7th European Networked Knowledge Organization Systems (NKOS) Workshop, 12th ECDL Conference, Aarhus, Denmark, Friday September 19th. Available at http://www.comp.glam.ac.uk/pages/research/hypermedia/nkos/nkos2008/presentations/steveNKOS08-trantBearman.pdf

[36] Weller, K. 2007. Folksonomies and ontologies: Two new players in indexing and knowledge organization. In Online Information Conference Proceedings, London 2007, 108-115.

[37] ZigTag. http://zigtag.com/