Freedom and Restraint: Tags, Vocabularies and Ontologies.

K Faith Lawrence, dr m c schraefel
University of Southampton
kf03r@ecs.soton.ac.uk

Abstract

The benefit of metadata is widely recognized. However, the nature of that information and the method of production remains a topic of some debate. This division is most noticeable between those who believe in 'free tagging', and those who prefer the more formal construction of an ontology to define both the vocabulary of the domain and the relationships between the concepts within it. Looking at the community surrounding online amateur authors and the descriptive metadata they have developed over the last thirty years we consider what we can learn from a mature but amateur tagging community. This paper considers how these two systems might be used together to add the easy usability of free tagging to ontology descriptions and the conceptual richness of ontologies to free tags.

Keywords *— ontology, folksonomy, free tagging, indexing, case study*

1. Introduction

Free tagging, the addition of user determined keywords or phrases to an object, has become one of the new electronic trends and has driven the popularity of social bookmarking sites such as del.icio.us[1], flickr[2] and 43 Things[3]. At the other end of the spectrum formal ontologies provide a rich resource but require expert knowledge, not found in the average Internet users. Discussion on the relative values of the two methodologies, top-down vs bottom-up, has been widespread[12], [5], [9, etc]. One of the big questions has been how free tagging sites will mature. Will popularity and the influx of new users solidify the system or hopelessly distort the noise to signal ratio? While it is not possible to predict the eventual outcome, this paper considers an existing community which uses tags and has done for some time. By analysing that community we can gain some insight into how tagging communities might develop and where future development will be needed.

1.1. Case Study: Amateur Fiction

The group selected for consideration is amateur online writers and their readers. The Internet has opened up many opportunities in electronic publishing. While the commercial world is weighed down by formats and digital rights management, the amateur world has embraced the new means of distribution. A lot of attention has been given to blogs as representing a new paradigm in free information exchange. They are seen as both challenging traditional news sources and introducing new, more democratic, social aspects to online discourse. While a popular area of study, the 'Blogosphere', as it is called, is a relatively new phenomenon. Early blogs such as Justin Hall's[4] have been around since the mid 1990s but it did not truly gain momentum until the late 1990s and the early years of the present decade. If we compare this to shared fictional, rather than supposedly factual, community-based communication, we gain an interesting comparator.

There are two types of amateur fiction available online original fiction and fan fiction. Most of the work in this paper deals with the fan fiction community because they were the more organised and responsive of the two. Fan fiction, in its simplest form, is (amateur) fiction written about characters or set in a world that has been previously created by someone else. Both fan fiction and original fiction have been around since the oral tradition held sway and telling the difference between the two isn't always obvious. In its most current form, fan fiction is generally agreed to have been around since the 1960s with credit (or possibly blame) most often being given to Star Trek, and other popular shows of the era, and the related influence of the science fiction community. Fan fiction online dates back to UseNet and the early 1980s although it has been suggested that fan fiction archives were hosted on FTP servers as early as 1973. For a dispersed community communicating by post, the Internet offered many advantages such as speed and ease of use. The result was a massive expansion both of the community and, therefore, the amount of material being created and shared.

The practice of adding additional, non-bibliographic information to the material became common practice within the community during this partial but major transition from page to screen. While initially this consisted of warnings for subjects such as character death or sexual violence which could easily upset readers, the range of 'warned' for content expanded to a rough classification system. From authors adding lists of keywords or short phrases, many of which were developed and evolved within the community, a taxonomy of terms was developed. Since this shared vocabulary was created through a 'bottom=up' and democratic process it could be argued that it represents an early folksonomy, as the vocabularies derived from popular tags are often called. These community taxonomies are often shared as glossaries[5] but the

[1] http://del.icio.us/

[2] http://flickr.com/

[3] http://www.43things.com/

[4] http://www.links.net/

[5] The Fanfiction Glossary: http://www.subreality.com/glossary.htm

meaning and context of individual terms is kept alive through usage. As a result it mirrors tagging in both its approach to data description, and in its issues and drawbacks.

2. Community Categorisation & Tags

In his recent piece on the overrating of ontologies Clay Shirky makes two distinct arguments [12]. The first is for the superiority of free tagging over structured ontologies in certain circumstances and the second is for probabilistic categorisation. His second argument shows an unrequited conflagration of development methodology and useage but is outside the scope of this paper. Considering his first point, Shirky sets out the type of context in which he argues that free tagging is better than an ontology. With its requirements for a large corpus, amateur users and non-static characteristics it would be difficult to find a more accurate description on the amateur writing online community.

Indeed, the current system in use by the community is fairly close to the free tagging system advocated by Shirky except that, with the exception of inside archives and LiveJournal tags, it is not currently machine readable. Community users are very much focused on human-readable metadata. This is hardly surprising because many of the community practices predate the Internet let alone the more recent developments in machine-readable information processing. However, beyond that fact, the community is made up of regular Internet users and advanced technical knowledge cannot be assumed. In a survey[8] run to gather user requirements just over half of the respondents were not comfortable with direct exposure to HTML and a further 30% were willing to work with raw HTML but lacked knowledge of scripts or other computer languages. 100%, however, were able to read.

When considering metadata used by the amateur writing community it is necessary to consider it in the context of its use as well as in the general context of metadata use. An initial investigation into community practice shows some idiosyncrasies that clearly illustrate the amateur nature of the community. There is a conceptual difference seen between categories and warnings, despite the fact that they are to all intents and purposes the same thing. The realisation that warnings could be used to find as well as avoid came early on but, despite that, the split between the two remains strong in many places. All the mailing lists except one required or advised that warnings should be stated and yet none asked for any indication of genre or other normally expected classification. Of the archives twelve out of the fifteen separated warnings from classifications. Of the remaining three one did not allow adult content which removed the necessity for all the most common warnings and one had no categories at all other than a list of the romantic pairings and the warnings.

A review of 7 mailing lists and 15 archives gives a overview of what type of metadata is commonly found. Both the mailing lists and archives were chosen at random. In the case of the mailing lists the criteria for selection was that they had an introductory text file to which we had access and which explicitly listed the expected metadata rather than relying on user knowledge of standard posting etiquette. In the case of the archives, it was decided to focus on the medium-sized fandom archives as these best represented the community values for their specific group. All but one of the archives was automatic, meaning that the authors selected the metadata from a list of preset options rather than being all free text entry. The list of categories was derived either from the story upload form (when accessible) or from the categories listed as available on the search page when full search available. Almost all archives carry some restrictions as to the type of fiction allowed to be uploaded. Because we were working with small to medium sized archives these restriction mostly focused on the universe that the stories were required to be set in and the romantic, or not, relationships that were allowed. For this reason those particular, and otherwise expected, metadata classifications were removed from the study because in most cases they were implicit and therefore their lack of presence was not indicative of anything but the restrictions on the archive.

2.1. LiveJournal Interest Tags

Tags can take many forms. Social bookmark sites such as Del.icio.us and Flickr have become the poster children for so-called 'free tagging'. They both use white space as a delimiter between tags which places restrictions on the format of the tag - that it must by a single keyword. This can be contrasted with systems such as LiveJournal's interest and content tags which are comma separated and can, therefore, be single world, multiple words or phrases. Where tags are restricted to keyword tags users typically find ways around the problem. The use of punctuation, especially hyphens and underscores, in common as is the simple amalgamation of the words into one[2]. It is an interesting question whether this requirement for user invention to circumvent technically imposed limitations adds to the synonym problem that is implicit in free tagging.

Shirky uses LiveJournal interest lists as an example of where this free-text tagging. Shirky's argument is that in his example "the terms actually encode different things, and the assertion that restricting vocabularies improves signal assumes that that [stat] there's no signal in the difference itself, and no value in protecting the user from too many matches [12]. For a system that is based on the concept of 'no experience necessary' there is an assumption being made by Shirky that the user is coming to the system with an expert knowledge of the domain they are tagging. It may be true that film people and movie people oc-

and Ye Olde Jolly Jolly Anal-Retentive General Fandom and Fanfiction Glossary: http://www.theparapet.net/fanfic/glossary.html are just two examples cupy two separate community spheres but unless the user already knows which of those two communities they wish to associate with that distinction is meaningless. Film people and movie people only want to not talk to each other if they self-identify as members of one of the two groups. The community aspect of social bookmarking sites is often raised as a benefit. It is interesting to consider this in comparison with the continued debate on community fragmentation that has become commonplace inside the fan fiction community, with LiveJournal frequently being represented as the main culprit. Dr Cherry [1], in her recent presentation at De Montfort University, considered the perceived advantages and disadvantages that were seen to come with this fragmentation. While the advantages are not insignificant there is a concern becoming more evidenced within the community that too much fragmentation and the increased insular nature of these subdivisions are detrimental to the community as a whole. In this context, therefore, there is a case to be made that forcing film and movie people to interact actually has social benefits beyond making it easier for the user to find what they are looking for.

Rather than "movies" and "films" let us consider terms more relevant to the domain under discussion in this paper. Fig. 1 shows the number of individual journals (bottom-left of table) and communities (top-right of tables) on LiveJournal using the interest tags "creative writing", "writing", "fan fiction", "fanfiction", "fan fic", "fanfic" and "fic". The lists of individuals and communities were then compared to see how often the terms overlapped. While it could be argued that "creative writers" do not wish to talk to "writers" or fan writers and visa versa it would be hard to insist that the users of the different variations in spelling and abbreviation of the term fan fiction do not wish to communicate. The numbers were harvested from LiveJournal on July 4 2005 at approximately 5:30pm.

From table 1 we see the various ways of indicating an interest in fan fiction are more closely linked to each other than either is to creative writing or writing (or creative writing and writing are to each other), which is what one would hope to see. However the nearly twenty percent of individuals and communities that use both fan fiction" and fanfic" is still a small section of what is in fact one interest.

One possible way around this problem is to allow 'fuzzy' tagging with machine reasoning identifying collections of tags and thus deducing their relationships. In the example above the frequency of relative position between "fan fiction" and "fanfic" would allow them to be automatically linked. The disadvantage of this method is that either the author has to enter multiple similar tags so that the terms can be associated, or tagging by the general populace must be allowed. The former creates more work for the author, while the latter depends on the community allowing such cross-tagging. In some communities this would be accepted but that assumption is not one that can be made without investigation into a specific group's user practices. Even when allowed there is the problem of false associates when tags are just grouped on text string rather than conceptual meaning.

3. Not So Restricted Vocabulary

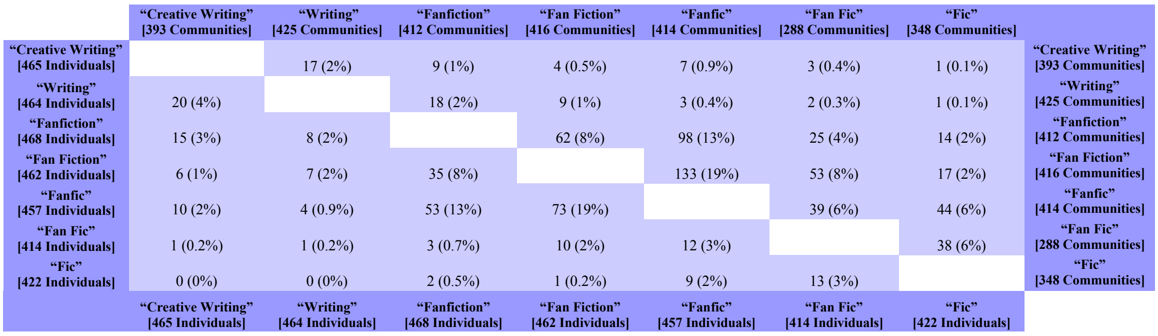
In total there were 136 tag concepts identified, space considerations prevents them being listed here, from the 15 archives. Just over 50% of the tags were used more than once but only approximately 10% were represented on more than half the the archives. This lack of inter-archive cohesion may represent the different community standards with the 10% representing the overarching globel standard. Or it may be due to lack of communication between the the different archives. As previously mentioned the majority of tags are human-readable only. Some archives output a RSS feed of stories as they are added but as yet this data has been used to feed update journals and individual use and has not been amalgamated into an general site. Due to this, there has been no community pressure for the individual archives to standardise so long as they represent the needs of the immediate community that they serve.

However the community archives act as both collection point for media items produced by the community and as focal nodes. Since they are created and maintained, almost exclusively, by members of the community the archives occupy an interesting position of being directly influenced by community expectations while, at the same time, reinforcing those behaviours by explicitly defining the information that the archive requires in addition to any media item. The mailing lists, eGroups and community journals occupy a similar place in the community hierarchy although, being totally free text entry, they allow more scope for individual expression. It could be argued that because these systems allow free text entry they are closer to free tagging systems than the archives which, while allowing some free text entry, for the most part give a choice of tags for the user to select. Given the choices embodied in the archives are drawn from the user experience with the free tags in the mailing lists and journals, what the list actually represents is the folksonomy that theory has suggested will develop from free tag use.

4. Archive Folksonomy Analysis

As discussed above there is only a conceptual difference between classification given to categorize an item and one given to warn for content. Therefore when the various descriptive tags were collected from the sample archives the two types were combined. Obvious synonyms were identified (humour-humor, expended/non-expanded acronyms etc) and grouped together to give a list of tag concepts, the total number of times that the tag was available across the fifteen archives and the number of variations of that tag. The tags were characterised by:

- Source of the terminology



	“Creative Writing” [393 Communities]	“Writing” [425 Communities]	“Fanfiction” [412 Communities]	“Fan Fiction” [416 Communities]	“Fanfic” [414 Communities]	“Fan Fic” [288 Communities]	“Fic” [348 Communities]	
“Creative Writing” [465 Individuals]		17 (2%)	9 (1%)	4 (0.5%)	7 (0.9%)	3 (0.4%)	1 (0.1%)	“Creative Writing” [393 Communities]
“Writing” [464 Individuals]	20 (4%)		18 (2%)	9 (1%)	3 (0.4%)	2 (0.3%)	1 (0.1%)	“Writing” [425 Communities]
“Fanfiction” [468 Individuals]	15 (3%)	8 (2%)		62 (8%)	98 (13%)	25 (4%)	14 (2%)	“Fanfiction” [412 Communities]
“Fan Fiction” [462 Individuals]	6 (1%)	7 (2%)	35 (8%)		133 (19%)	53 (8%)	17 (2%)	“Fan Fiction” [416 Communities]
“Fanfic” [457 Individuals]	10 (2%)	4 (0.9%)	53 (13%)	73 (19%)		39 (6%)	44 (6%)	“Fanfic” [414 Communities]
“Fan Fic” [414 Individuals]	1 (0.2%)	1 (0.2%)	3 (0.7%)	10 (2%)	12 (3%)		38 (6%)	“Fan Fic” [288 Communities]
“Fic” [422 Individuals]	0 (0%)	0 (0%)	2 (0.5%)	1 (0.2%)	9 (2%)	13 (3%)		“Fic” [348 Communities]
	“Creative Writing” [465 Individuals]	“Writing” [464 Individuals]	“Fanfiction” [468 Individuals]	“Fan Fiction” [462 Individuals]	“Fanfic” [457 Individuals]	“Fan Fic” [414 Individuals]	“Fic” [422 Individuals]	

Figure 1.Community Involvement of Respondents

- Class of thing described
- Content range described with the PICS categories [10] used for guidance
- Complication of the tag

The breakdown can be seen in Table 1.

A chi-squared test was used to determine whether there was a statistically significant relationship between these groupings and the numbers of variations found for a term.

4.1. Results

A statistically significant relationship was found between the number of synonyms and the popularity of the term (p<0.000), the complication of the term (p<0.000) and the source of the terminology (p=0.035).

No significant relationship was found between the number of variations and the type of thing being described or the category of classification.

4.2. Discussion

It seems logical that the more parts the tag is comprised of, the more variations it may have. Observation of keyword tagging sites suggests that this is not reduced by limiting tags to a single word since users then struggle to convert the longer phrase they would otherwise have used. Synonym issues are a recognised feature of free tagging. Further analysis is needed to discover any measurable difference between keyword and phrase tagging.

As with the length of the tag, the relationship between concept popularity and variation is a logical one. The more people wish to express an idea, the more opportunity there is for divergence, whether accidental, contextual or intentional. The suggestion has been made that over time a natural consolidation of terms will occur especially around the more popular topics [2] and suggestions has been made for ways of supporting this process[11].

The fact that there was a statistically significant relationship between the derivation of the vocabulary and the number of variants raises a number of issues. While the two terms that carried the most variations (7 and 9) were both external terms being reused by the community it should be noted that one was an acronym and the other represents a concept which is currently under debate within the community. Given that, other than these two cases the most variations from borrowed vocabularies, literary or otherwise, is 3, there is a strong argument that they can be seen as special cases. If we disregard these two cases as outliers then it is clear that the terminology that evolves with the community has significantly more variation than that from outside.

The majority of the adopted terms are literary, legal, calendric or carnal with only a few not coming from a well defined and organised vocabulary. While the tags were freely added the difference that can be seen between the borrowed tags and the evolved tags may well be because at some point the borrowed tags were from a formally classified and defined categorisation. While statistically significant it could be argued that the practical difference is low and given the trend seems to be towards the condensing of terms perhaps time will continue to reduce this gap. Only time will reveal at what point the trend levels off but it should be noted in discussions that even though free tagging takes pride in its bottom-up approach it is drawing directly from existent imposed formalisations.

4.3. Extending Tags

One alternative being investigated is the combination of free tagging and ontologies by separating out ontology and vocabulary. When individuals add tags to documents they tend to use an individual or personal vocabulary. A user’s personal vocabulary rarely changes between documents and users within contexts often share vocabularies at a practical level. These facts suggest that it would be possible to produce a small number of vocabulary lists mapping to an ontology which would satisfy a majority of users.

By giving users the option to personalise the map between the vocabulary they use and the most popular definitions of the ontology’s concepts it is possible to hide the complexity of the ontology from the casual user yet retain the richness of the description. The disadvantage is that this requires the design of these initial lists through analysis of the community

Table 1.Archive Categories

Synonyms	Count	Vocabulary	Count	Class	Count	Content	Count
1	100	Developed	51	Genre	32	Not Content	48
2	18	Literature	28	Content	71	Sex	28
3	12	Other	57	Genre/Content	15	Violence	16
4	1	Complication	Count	Document Type	10	Harmful	2
5	3	Phrase	48	Source	5	Hate	1
7	1	Keyword	79	Language	1	Language	1
9	1	Acronym	9	Fandom Dependent	2	Other	40

- especially those communities which do not include ontology experts.

Rel-tags[13], tags which include a hyperlink, provide an extension of the tag system. These tags allow users to mark material which is outside the immediate domain of a free tagging site. The visible string is independent from the URI component which acts as the actual tag (See Listing 1). The URI specified in the tag is required by the specification to point to a valid 'tag space' which it defines as "a place that collates or defines tags...where the last segment of the path of the URL is the tag"[13].

While this provides a way to extend tagging beyond the social bookmarking sites it is still limited to the tagging mindset. The displayed, user readable tag is not processed in any way, only the final section of the URI and any query parameters or additional fragments attached to the tag are ignored. This imposed limitation may define the scope for the rel-tag but it does not define the scope of the possibilities.

The concept of the triple as defined within RDF[4] requires a subject, predicate and object. If we consider a rel-tag in this space then the following mapping is possible:

Table 2.Tags to Triples

Triple Component	Tag Component
Subject	Document being tagged
Predicate	URI defined by tag
Object	String value of tag

Given the necessity of processing the Document Object Model to retrieve the tag URI it is logical to consider what other information could be parsed for processing at the same time. If we regard the URI as a link to the definition of the tag rather than a string to be processed then it becomes possible to link directly with more structured and exact definitions. For example the URI could reference an ontology definition which could provide the additional contextual information that a flat, string tag lacks. For example it could link to something as simple as the RDF WordNet definitions[6] which would allow the user to specify which meaning of a word they were intending. The concern with this system would be the effort required on the part of the user to add in the correct link and possibly have to create the ontological definition to which they intend to associate the tag. This brings us back to the idea of a shared vocabulary.

If we agree with the idea that term usage is context based then within a context, for example an online community, there is a shared understanding. This the reverse of the free tagging concept that community will be created by finding people who have a shared worldview as seen by their terminology usage. Given the existence of many online communities such as the one discussed in this paper it is not inconceivable that, at this stage of free tagging community coalescence, the community pre-exists the tagging. It would only require one member of the community to create a taxonomy of definitions for the less technical members of the community to be able to make use of the system. It would even be possibly to create one or more distributable lists of vocabulary/mapping which could underpin an automatic completion system so that the user would not need to see any of the tag beyond the human readable text. Pick-and-mix systems could easily allow users to create a vocabulary system for themselves choosing which definitions of words they wished to use, perhaps with a simple to understand disambiguation system such as Wikipedia uses. Power-users could create new definitions for their own and other's use - a social concept marking system as opposed to a bookmarking one.

This is not about simply providing a link to a definition which can be used by a human to understand what is meant, but a way of linking to machine processable data. (See Figure 2).

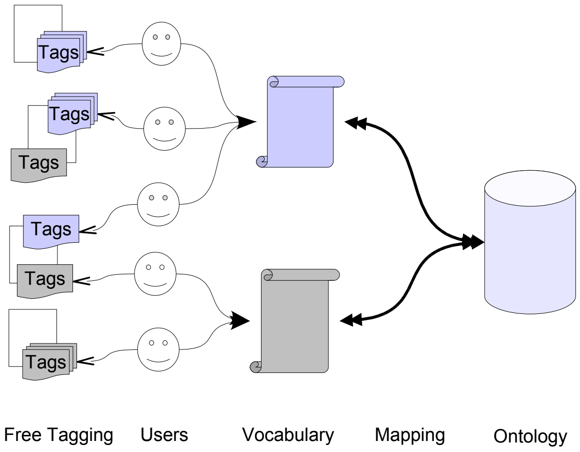


Figure 2.From Vocabulary to Ontology

By linking the vocabulary to the more complex structures it becomes possible to carry out processing that

[6] http://xmlns.com/2001/08/wordnet/

Listing 1.Rel Tag Composition

```
<a href=‘‘http://any.site.tag.space/[myTag]’’ rel=‘‘tag’’>[displayedTag]</a>
```

requires those structures and their defined relationships. This does restrict the vocabulary available, unless the user is in a position to create their own formal definition and mapping, however it works within the confines of a shared community understand which also underpins the usability of free tagging. With resources such as WordNet already available a lot of the necessary structures are already available for use. For communities such as the fan fiction community which already have large human readable glossaries of community terms it is a small step to link to these and the next logical step to create a machine readable version[7].

5. Conclusion

The fan fiction community represents one which has moved from free tagging metadata to using the folksonomies created from those tags. The lack of machine readable metadata may have slowed the progression but the extended length of time which the community has been using the system more than compensates for any potential slowness. We can see how a mature system may behave with the tags feeding the taxonomy which in turn informs the community as to what tags are expected and appropriate vocabulary. We also see that where there is no formal structure underlining the vocabulary there is a greater variation in the exact tag string used. While a number of factors affect this, the underlining community agreement on vocabulary and meaning, whether the community using the term or the one for which it was originally formalised, has a strong effect on the divergence of terms.

While wanting to keep the system that the community knows and likes, we have also laid out a way in which the usability may be retained but, with the shift to machine readable metadata, the full advantages and capabilities may be exposed. While this case has been made with the needs of a specific community in mind the implications are much broader. We are in the process of introducing these changes to the fan fiction community and hope that the lessons taken from this can be used to inform other communities and tagging systems.

6. References

[1] B. Cherry. ‘so we said screw it and wrote it ourselves’: Creative ownership, gendered narratives and slash fiction in the on-line start wars fan communities. Presented at the Slash Fiction Study Day, Cultural eXchanges, De Montfort University, March 2006.

[2] M. Guy and E. Tonkin. Folksonomies: Tidying up tags? *D-Lib Magazine*, 12(1), January 2006.

[3] M. O. Jewell, K. F. Lawrence, A. Prugel-Bennett, and m. c. schraefel. Annotation of multimedia using ontomedia. Submitted to First International Workshop on Semantic Web Annotations for Multimedia (SWAMM 2006), Edinburgh, United Kingdom., 2006.

[4] G. Klyne and J. J. Carroll. Resource description framework (rdf): Concepts and abstract syntax. W3c recommendation, W3C, February 2004.

[5] E. Kroski. The hive mind: Folksonomies and user-based tagging. InfoTangle Blog, December 2005.

[6] K. F. Lawrence, M. O. Jewell, M. M. Tuffield, A. Prugel-Bennett, D. E. Millard, M. S. Nixon, m. c. schraefel, and N. R. Shadbolt. Ontomedia - creating an ontology for marking up the contents of heterogeneous media. In *Ontology Patterns for the Semantic Web ISWC-05 Workshop*, November 2005.

[7] K. F. Lawrence and m. c. schraefel. Ontomedia - creating an ontology for marking up the contents of fiction and other media. In *Proceedings of 1st AKT Doctoral Colloquium*. AKT, June 2005.

[8] K. F. Lawrence and m. c. schraefel. Web based semantic communities who, how and why we might want them in the first place. In *Proceedings of Web Based Communities 2006*, San Sebastian, Spain, Feb 2006. IADAS.

[9] P. Mika. Ontologies are us: A unied model of social networks and semantics. In *Proceedings of the 4th International Semantic Web Conference*, Galway, Ireland, 2005.

[10] J. Miller, P. Resnick, and D. Singer. Rating services and rating systems (and their machine readable descriptions). W3c recommendation, World Wide Web Consortium, October 1996.

[11] L. Pind. Folksonomies: How we can improve the tags. Blog, January 2005.

[12] C. Shirky. Ontology is overrated: Categories, links, and tags. Clay Shirky's Writings About the Internet Website, 2005.

[13] K. M. Tantek Çelik, Derek Powazek. rel-tag: Draft specification. Draft specification, Microformats, 2005.

[7] This is currently being done using the OntoMedia ontology[7], [6] with the first implementations just becoming public[3] (http://ontomedia.ecs.soton.ac.uk/rdf/)