Tonkin, E. (2007). Signal and noise: Social construction and representation. 18th Annual ASIS SIG/CR Classification Research Workshop
doi: 10.7152/acro.v18i1.12873

Signal and noise: social construction and representation

E Tonkin, UKOLN
e.tonkin@ukoln.ac.uk

0. Abstract

This paper attempts to draw a quick sketch of some of the research that relates to the state of social tagging research today. The result is intended to be representative rather than exhaustive. The goal of indexing consistency is discussed and examined with respect to the specificities of differing indexing systems. The relation of indexing consistency with 'language-in-use' is discussed. We then proceed to take a look at a few examples of much older systems that relate closely with the lessons now being learned in social tagging today, in order to situate the present activity in its historical context – and examine a few approaches used for text-based search-and-retrieval and their relevance to tag corpora. To conclude, some distinctions between personal, social and global information management are discussed.

1. Introduction

> *"History is the only laboratory we have in which to test the consequences of thought."*
> *Etienne Gilson*

This paper is primarily about consolidation. Firstly, some prior work is discussed that seems relevant to the current debates on social tagging services, systems that collect, aggregate and manipulate free-text tags or annotations. Such tags may typically contain almost arbitrary strings, with a few system-specific limitations (on special characters, length, and much less frequently, on content).

In general, tag services do not promote a highly-structured approach to tagging, although examples of use of tag information to mediate more complex data types exist. One example of this is the geotagging phenomenon; another is system tag/tag prefix functionality available on systems such as del.icio.us (prefix:tag1 and prefix:tag2 may be queried by means of the system function system:has:prefix; views faceted by object type may be attained in a similar manner). With these exceptions in mind, the majority of tags do not share this simplicity of structure, intended task or use. Our central topic is the nature of a tagging corpus, the rules by which the production of free-text tags may be said to be governed, and the metrics by which we might analyze a collection (corpus) of user-generated tags.

1.1 Plain text, semantic and concept

Indexing consistency is a measure of agreement; inter-indexer consistency refers to the level of consistency between two different indexers' chosen terms, and intra-indexer consistency refers to the level of consistency between one indexer's chosen terms at different occasions - in other words, the extent to which two or more people agree in their description of a body of information (Sievert & Andrews, 1991). The literature exhibits a number of interesting characteristics, not least the point that there is a certain level of inconsistency about the importance of indexing consistency as a metric. Some report consistency as 'long considered essential for effective retrieval' (Olson & Wolfram, 2006), whilst Soergel (1994) points out that it is 'not important in itself'. Although it has been used as an indicator of indexing correctness, an index may be consistently incorrect.

Soergel's description of indexing consistency as a problematic measure echoes Cooper's 1969 findings that consistent indexing does not necessarily produce the best results. Calling inconsistency 'the rule rather than the exception', Cooper suggested that even where indexer-requester consistency was constant (that is, where the

requester of the resource consistently chooses the same term as the indexer), increased consistency 'will not necessarily result in an increase in retrieval effectiveness.' Other studies, however, such as Leonard (1975), concluded that 'there exists a direct, positive relationship between high interindexer consistency and high retrieval effectiveness'.

Naively, it would seem likely that this disagreement results from differing conditions of use, that perhaps retrieval effectiveness is susceptible to the context(s) of use of the index in question. In other words, it is possible that the use cases tested in some circumstances were amenable to indexing consistency as a measure, whilst others were not. What different use cases are there for an index? Why would we expect consistency to be more important in some than others?

One obvious possibility suggests itself: that inconsistency in indexing may reflect the consistency of the user group for which the information is to be indexed, the level of acceptance of the characterisation by which the information is to be indexed, and the purposes for which the information is to be indexed.

1.2 Language games

> *"Really, do not you think Udolpho the nicest book in the world?"*
> *"The nicest--by which I suppose you mean the neatest. That must depend upon the binding."*
> *"Henry," said Miss Tilney, "you are very impertinent. Miss Morland, he is treating you exactly as he does his sister. He is forever finding fault with me, for some incorrectness of language, and now he is taking the same liberty with you. The word 'nicest,' as you used it, did not suit him; and you had better change it as soon as you can, or we shall be overpowered with Johnson and Blair all the rest of the way."*
> *Northanger Abbey*, Jane Austen, 1818

There are many characterisations of language, for example, the model of language as a complex adaptive system (Steels, 2000). In this view, a community of language users is described as a complex adaptive system which collectively solves the problem of developing a shared communication system, by reaching an agreement on a repertoire of forms, meanings and form-meaning pairs – that is, a system such as the sounds used in speech, a set of shared conceptualisations, and a lexicon and grammar. For descriptive purposes at least, this is an extremely useful model, allowing us to build up a model of language evolution in terms of simple rules and characteristics of agents within a community.

A brief thought experiment is in order. This characterisation leaves us looking at the evolution of language as ongoing – like self-organising cellular automata, which are examples of complex dynamical systems, constantly shifting 'emergent' meta-structures are formed through a great number of local interactions. In the present day, many interactions are now mediated in a manner other than face-to-face communication. Many interactions are asymmetric, as in the case of input from broadcast media. Language – speech or the written word – is the primary medium through which these interactions are mediated. This process does not seem to reach a global equilibrium. Instead, there are many areas which seem to have reached a local equilibrium – perhaps an optimal solution for that localised environment during a given time period.

Deutscher (2005) describes a triad of motives for linguistic change in general: *economy*, *expressiveness* and *analogy.* That is, firstly there is a motivation to save effort – for example, in pronunciation of spoken words or phrases. Secondly, speakers' wishes to 'achieve greater effect for their utterances' leads to novel or emphatic restatements of existing terms (for example, rather than saying 'no', we are likely to say 'no, not at all'; presumably, in written language this may translate into the same sort of tendency that leads an AI researcher in 2007 to redefine their own work as 'data mining', towards more expressive speech). Thirdly, the motive of *analogy* is responsible for the very human urge to find regularity in a system that does not have a great deal of it. Dutch speakers, for example, would probably recognise this force in the recently published editions of *Het*

Groene Boekje, or 'The Little Green Book', the officially regularised orthographic and grammatical reference of the language that lays down rules as to the appearance of hyphens and appropriately regularizes spellings. Smith (2006) recognises a significant limitation to the rate of change in the need for a listener to accurately reconstruct the speaker's meaning – but the operative term here is *a listener*, rather than the general population of potential listeners.

The development of the components that form language represents one participant mechanism in an eternal experiment in iterative design. There are other elements in this game, such as the mechanisms that govern category learning (in infants and in adults), and phenomena such as categorical perception that alter our perception of stimuli such as speech and colour. When presented with a range of sounds between two phonemes, for example, we hear discontinuities related to the categories with which we capture sound. To take a specific example, the distinction in French between the 'u' sound in 'tu' and the 'ou' sound in 'vous' may be an example of a difficult distinction for an English speaker to hear- the two may appear to be the same sound. As category perception is learnt, and may be influenced via training (eg Goldstone 1994), this represents a further set of mechanisms contributing to the language development game. To quote the introduction to a recent journal issue devoted to the topic of language emergence (Ellis and Larsen-Freeman, 2006):

> 'There are many agencies and variables that underpin language phenomena[...]. Language is complex. Learners are complex. These variables interact over time in a nonlinear fashion, modulating and mediating each other, sometimes attenuating each other, sometimes amplifying each other in positive feedback relationships to the point where their combined weight exceeds the tipping point, which results in a change of state.'

Chalmers (2004) describes this interplay as follows:

> 'The individual and their prejudice are changed through the use of language, and the language changes through its use by individuals[...] This endless process of seeing the part in and through the whole is the *hermeneutic circle*.'

A final, crucial point in this comes from sociolinguistics, and is represented by the observation that, as speakers of a given language, we are generally able to converse fluently in several linguistic norms or 'modes of speech' (sometimes also dialects), to speak and listen as specialists in several fields.

1.3 Language, task and community

There exist many studies in both historical linguistics (the study of the changes that occur in and between languages over a long period of time) and in sociolinguistics (the study of the interrelation between language and society) that may be of use to us in understanding the way in which a 'semi-formal' system of symbols that is intended to be relatively static can evolve and change with time, the drivers underlying those changes and the characteristics that may be observable.

Several concepts could be of use in discussing “the way in which language is used within a community”. Of these, the first is the *speech community*, a term originating in sociolinguistics. There exist various definitions, but in general this term is applied in order to characterize a set of individuals – for example, when studying the variation in language use between two groups, each group is tentatively characterized as a speech community (the assertion is made that members of this group share a set of linguistic norms). This functional definition leads us to think of speech communities as reflections of physical communities – for example, rural Yorkshire farmers versus retailers in inner-city York.

However, it is also possible to look on the speech community as a community to which one claims membership by attaining that set of characteristics in speech. An SC appears wherever a shared linguistic norm emerges from

group behaviour; speech within that group henceforth attracts these characteristics. Thus, in a given day an individual may switch between speech communities on a number of occasions – when they say goodbye to the children and head out to the office, when they leave the desk for a chat with their hockey buddies in the canteen at lunch hour, and so forth. Potentially, the characteristics of their speech – or more likely, written language – may also change as they interact within different communities on the Internet.

Swales (1990) provided an alternative term, *discourse communities*, based on Nystrand's 1982 proposal. These share:

- common public goals
- mechanisms for intercommunication
- participatory mechanisms to provide members with feedback and information
- discourse expectations reflected in genres
- specialized terminology, and
- a critical mass of experts

DCs are 'spatially dispersed, formed around sociorhetorical functions, and mainly mediated by texts' (Prior, 2003). These were later revised to distinguish *place* DCs from *focus* DCs. *Place* DCs are 'local groups involved in some mutual project that brings about such things as shared lexis, regular communicative genres, and recognized, though not necessarily consensually accepted, senses of purpose and role'. *Focus* DCs are 'not defined by mutual engagement, but consist of individuals who co-participate in discursive practices with some purposeful focus even when they are separated by time, language, geography and so on'.

The description of DCs as stable entities has been criticised as 'warm fuzzy “discursive utopias”' by Harris (in Prior, 2003). The idea of a large, stable meta-structure containing a linguistically stable group of individuals can be discounted; Prior also quotes Marilyn Cooper as arguing for 'seeing discourse communities... as the products of continual hermeneutic work, as social phenomena where varied values and practices intersected, as ways of being in the world, not narrow intellectual commitments'. Studies have found that in academic and disciplinary writing, where one might expect to find a DC, one finds 'complex spaces, shot through with multiple discourses, practices, and identities'.

After the discourse community came Lave and Wenger's *communities of practice*, which describe 'participation in an activity system about which participants share understandings concerning what they are doing and what that means in their lives and for their communities.' (Lave and Wenger, 1991, pp. 97-98). Communities of practice involve *mutual engagement*, *an enterprise*, and *a shared repertoire*, and are 'somewhere between an interaction or series of casual interactions and larger categories like organizations, cultures or professions'.

Prior points out that DCs – and communities of practice, and activity systems – all share the characteristic that they are already *named social entities*. In effect, the existence of these communities is already assumed – we find it easy to believe in their existence because we are predisposed to do so. It is easy to fall into 'folk-linguistic, folk-sociological ideology'; therefore, it is as well to exercise some caution in this sort of analysis. Which if any of these terms – of these models – is most appropriate for our purposes?

1.4 Code switching

The modification between modes of speech alluded to earlier is rather similar to the phenomenon referred to as 'code-switching', a term which refers to switching between languages or dialects (say, between French and German) during the course of a conversation or other act. Code switching (CS) is often described, especially when discussing second language (L2) learners, as characteristic of linguistic deficiency – in fact it is often caricatured in this manner. It is important to recognise that this characterization, whilst in some instances accurate, may result from an assumption about the relative legitimacy of one variety of language over another. CS is used purposefully and fluently in many applications. Note that it is also valid to speak of register, dialect or

style switching or alternation; for the sake of the discussion, code-switching is here used according to a relatively vague definition encompassing all of the above as well as language or dialect alternation. A serious analysis of patterns of inconsistency in indexing would need to begin with an altogether more rigorous treatment.

Boztepe (2003) reviews various applications of CS. Fishman's Domain Analysis of CS describes patterns of use as relating to domain – interlocutor, place and topic relationships. Blom and Gumperz's study (ibid) describes situational switching, the definition of social rights and obligations, and metaphorical switching, the choice of topic. Three types of social constraints are described: *setting*, *social situation*, and *social event*. That is, physical environment, social situation and social event. Boztepe points out that Fishman's model of domain analysis does not talk about the reasons for this alternation, in the sense of what is gained from it for the speaker. What could be gained from CS?

One might argue that it results from conforming to the social norm. Boztepe quotes Myers-Scotton in arguing that CS is 'best explained by the optimal use of speaker's resources in their linguistic repertoires' – that 'the rewards of CS will be greater than those of maintaining a monolingual discourse pattern'. The relevance of speculation into cause is not great for our purposes, except inasmuch as it may inform speculation into the actual occurrence of CS or its subtler variants in the context of our problem – the ways in which tagging occurs.

As a final caveat, note that much work in this area relates primarily to spoken rather than written language. There is no reason to assume that internet-mediated communication would show very similar features to that seen in spoken discourse; some studies suggest that there are important distinctions between the forms (eg. Androutsopoulos, 2004). A general discussion of sociolinguistics and CMC can be found in the Journal of Sociolinguistics (Androutsopoulos, 2006).

1.5 Hermeneutics and the index

A common reaction to the description of indexing as in any way a *situated*, *contextual* or *interpretive* process is to describe such suggestions as a variation on the theme of hermeneutical theories of indexing. Subjectivity in indexing is often acknowledged but infrequently explicitly handled, although implicitly handled on several levels. To quote Ellis and Larsen-Freeman (2006) once again:

> 'Variability pervades language production. But such demonstrations of pattern are too profound to allow us to relegate this to noise and random performance factors[...]There is systematicity despite persistent instability; however, the systematicity is to be seen in dynamic, contextualized patterns, not only in rule-governed behaviour.'

Hjørland (2002) suggests, in the domain of indexing academic works, that indexer disagreement may be systematic, in that it relates to different theoretical understandings – of the work, or of its relevance – and suggests as a remedy that, whilst it is implausible to index neutrally or without bias, subject analysis should support an explicitly specified set of goals and values.

Chalmers (2004) quotes Suchman's (1987) statement that: 'the *communicative* significance of a linguistic expression is always dependent on the circumstances of its use. A formal statement not of what the language means in relation to any context, but of what the language-user means in relation to some particular context, requires a description of the context or situation of the utterance itself. And every utterance's situation comprises an indefinite range of possibly relevant features.' He subsequently defines *hermeneutic theory* as 'based on accepting the effect of this indefinite, inevitable and infinitely detailed situational background'. The use of a symbol in human activity[1] carries the stamp of surrounding contextual information; 'assumptions, abbreviations,

1 'language-in-use', as the author's own simplistic mental shorthand reads, following Wittgenstein's assertion that 'the meaning of a word is its use in the language'

applicability, the people involved, the other information that they share as part of their current activity, their organizational structure and practices, and so on, endlessly.'

The question of interest to the analysis of social tagging is: what elements within this context are available for analysis; what elements may be expected to be represented? The idea that an index is tailored for a given 'discourse community' may be considered to be appropriate for use is not a new one. A formal classification is far from decontextualised; to quote Stephen Jay Gould on the purpose of classification: "[Classifications represent] theories about the basis of natural order, not dull categories compiled only to avoid chaos." Perhaps we might begin to answer our question by examining some previous developments in the area of tagging and seeing what may be learnt from the results.

1.6 Marshall's dimensions of annotation

Marshall defined annotation in the following manner:

> Annotation is a fundamental aspect of hypertext. Readers respond to hypertexts with commentary, make new connections and create new pathways, gather and interpret materials, and otherwise promote an accretion of both structure and content. [...] It has been construed in many ways: as link making, as path building, as commentary, as marking in or around existing text, as a decentering of authority, as a record of reading and interpretation, or as community memory.

Marshall's work (Marshall, 1998) provides an excellent framework through which to examine free-text tags as annotations. She characterises the forms that annotations take by reference to a set of dimensions, either innate to the annotation itself, to the annotation as it operates from the perspective of a reader, or to the annotation as it may be used to communicate with other audiences. In brief, the following continuums of characteristics were identified in her study:

Formal v. informal annotations: Structured metadata vs unstructured marginalia.
Explicit v. tacit annotations: Annotations designed for general readability differ a great deal from those designed purely as personal reminders; the example provided here is the ``cryptic marginal 'No!'"
Annotation as writing v. annotation as reading: Certain annotations may constitute a significant addition to the text, performing a role as part of a participatory process of working with a document.
Hyperextensive v. extensive v. intensive annotation: Annotations may refer/relate to broad links between documents, or to an in-depth engagement with a single text
Permanent v. transient annotations: Some annotations are of little long-term value, often by design. Others are of long-term value.
Published v. private: Certain annotations are expected to remain private; others do not, but were not designed for publishing. Others are explicitly designed for publishing.
Global v. institutional v. workgroup v. personal: Some annotation systems are personal in scope; at the far end of the continuum are systems with worldwide access.

Some elements of this description can be closely matched with the discussion of community- or audience-specific registers of language. Compare the description of a shared communicative environment in (Krauss & Fussell, 1991) to the dimension of scope, from global to personal. From the basis that communication requires a body of shared information ('common ground'), Krauss & Fussell describe how such an environment may be constructed from the perspective of each participant, and demonstrate via a number of studies that setting the intended audience of a message impacts on the formulation of the message in a number of ways, and hence may alter the effectiveness of communication.

2 Pointers and the distributed filesystem

Tonkin, E. (2007). Signal and noise: Social construction and representation. 18th Annual ASIS SIG/CR Classification Research Workshop
doi: 10.7152/acro.v18i1.12873

Having briefly looked at the forces that underly the use of language, it is worth applying this information to a little more history: specifically, the development of the filesystem. Usability concerns in this area have been extensively studied, and in somewhat more of a user-centric manner than is typically the case in surveys of indexing systems, so it may be considered a helpful analogy for our purposes.

Years ago, Ward Cunningham described the wiki as 'the simplest online database that could possibly work'. An analogous description could be given for the tag: 'the simplest online filesystem metadata that could possibly work'. One might compare this with the way in which metadata is stored in a traditional filesystem, such as the FAT-16 system used by DOS.

The filesystem itself is generally designed principally with the machine in mind, with the primary aims of reading and writing speed, reliability – low probability of data loss, resilience against corruption of data resulting from crashes or hardware damage, and so forth. The mechanisms by which data is stored and retrieved typically make use of a machine-readable file identifier, which is unfortunately rather opaque to the user. For this reason, the additional layer for the filesystem designer to write is the user-interface layer – the way in which objects being managed are named, a human-readable attribute set (Tanenbaum & Woodhull, 1997). Users provide a certain amount of metadata about the file, such as a filename.

2.1 The classical filesystem

The file system designer thinks of a file as an array of blocks containing binary data, or rather, a file itself is a sequence of binary data that is stored across an array of blocks – the 'block' is the fundamental structure of a filesystem (a block device stores information in fixed-size blocks, each with its own address – each of which may be written to or read independently of the others). As files are typically discovered by looking through the contents of a directory, it is the directory structure (filesystem metadata, in general) that is actually read, since it is this structure that typically contains the metadata required to retrieve the file. The 'Standard Model' for a directory entry in a very simple filesystem might look something like the following, adapted from 'the MS/DOS filesystem', Tanenbaum & Woodhull (p420)

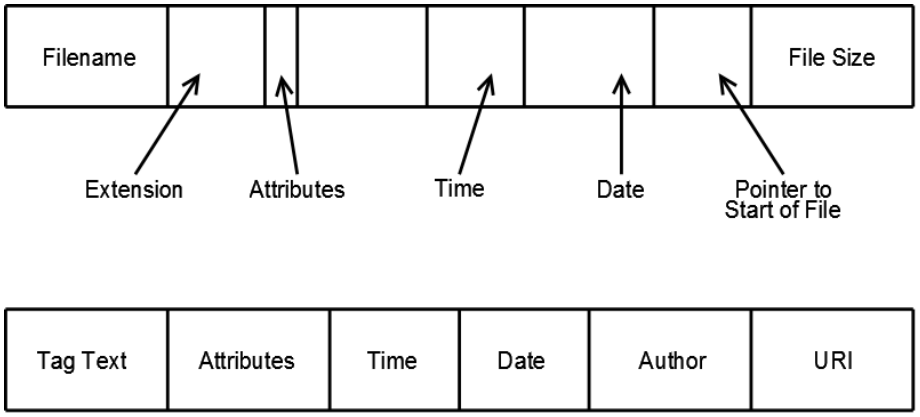

Fig. 1: 'Standard model' within filesystem and tag

A *pointer*, for programmers who work in low-level languages such as C, is a variable that contains a memory address – it does not contain the object itself, but it provides the address from which the object may be retrieved. In the case of a tag, the pointer is represented by a URI rather than a literal memory address. Time and date information are typically stored along with a tag, along with the tag itself – a plain-text string – in the place of the filename. The information stored in a basic filesystem and the data structure of a tag is not dissimilar in nature. More sophisticated filesystems provide more mechanisms for metadata storage - for example, many filesystems provide support for extended file attributes, enabling arbitrary attributes or attribute-value pairs to be

stored. Technically, the ability to create a relatively rich multidimensional single-user tag or metadata store is widely available, though it is seldom used.

2.2 The usable filesystem

A basic filesystem is simply a narrow tagging system on a limited data set, limited according to the filesystem's limitations. For example, this computer allows filenames of up to 255 characters, and virtually any character may be used in the filename (up to and including characters that are not easily readable, which is in itself problematic from the usability perspective). The traditional DOS filesystem permitted 11-character filenames but, by comparison to the ext3 filesystem, treats three of these characters as a file extension, forcing the 2-tuple convention of 'mnemonic word'.'filename extension'. The extension represents a short controlled vocabulary that could contain A-Z, 0-9, and several symbols. Symbols excluded, this suggests a limit of 36^3 possible file extensions from AAA to 999, of which a small subset are appropriate for use (pronounceable or memorable).

Conventionally, this very simple approach is supplemented by a filesystem hierarchy, by which files are organised into named sets, subsets and so forth to an arbitrary depth. This is typically monodimensional, in that a file can only inhabit one place in a filesystem hierarchy at any one time, with the exception of variations on the theme of the symbolic link, an OS-readable metadata record containing its own filename, but with a pointer that references and forwards certain operations toward the original file. In general, however, the filesystem hierarchy is widely described as difficult to use, in part as a result of the work of Barreau and Nardi (1995) (but see criticism in eg. Fertig et al (1996)).

A similar criticism is levelled by Woods (1997) toward hierarchical library classification systems such as the Dewey decimal classification system or the Library of Congress classification system, raising the point that a 'concept' may be placed in 'exactly one place in the hierarchy', and that the choice between the several plausible parent categories that may exist is essentially arbitrary. Indeed, Woods' discussion of the various difficulties of an hierarchical index is quite applicable to the filesystem. O'Toole and Gifford (1992) made a similar point, suggesting that file metadata should describe the content of a file, as well as the location where it may be found.

Some findings regarding user experience with modern desktop systems, summarised in Ravaiso et al (2004):

- Systematic separation of files, emails and bookmarks was seen by users as inconvenient
- The filename and path were not of significant use in resource discovery
- Background context behind the creation of a document played a part in storing and accessing that file, because of relevant background knowledge or due to the use of contextually relevant metadata.
- Users were reluctant to use search functionality, preferring to search manually
- Manual classification 'required considerable effort' - users thought that this was a task that their computer could address, although they also wanted to be able to adjust the result
- Users felt the need for a 'comfortable, flexible, easy-to-use method for adding notes and remarks to documents' - a form of free-text annotation, perhaps?

Additionally, the same user may at different times take several viewpoints on data: task-oriented, context-oriented and content-oriented. Any specific tool tends to support only one of these.

2.3 Tagging in the filesystem

Many research efforts have focused on filesystem improvement. MIT's Semantic Filesystem project (Gifford et al, 1991) had outlined a method by which attributes could be automatically extracted from files, and could then be used as a basis for separating content into 'virtual directories' - in this view, files are collected into virtual directories according to automatically extracted features. Various further alternatives were proposed, such as the 'pile' metaphor for the sorting of documents (Rose et al, 1993).

This project was designed to replace the hierarchical structure with a document management system that made use of 'meaningful, user-level document attributes, such as “Word file”, “published paper”, “shared with Jim” or “Currently in progress”'. In Placeless Documents, a file could inhabit multiple places in the filesystem and many views could be taken very quickly and simply on large sets of files, faceted according to the task currently in hand. This was designed to expose what Dourish et al (2000) referred to as the *multivalent* nature of documents – that a given document may 'represent different things to different people'.

Later refinements on this system took advantage of the Placeless Documents' structure of annotated pointers in order to create a structure referred to as *abstract documents*, which held no content but were simply representative metadata records that could be used as collections of metadata properties, enabling the creation of category types, representing a set of values – this is not dissimilar to the 'tag bundle' concept, that permits a sort of thesaurus of synonyms to be developed by grouping several related terms under a parent category. More traditional attributes were also held and inherited across this structure, such as access control attributes.

Dourish et al (2000) describe some of the limitations of the original project -specifically, the fact that the key-value pairs were untyped, and that there was no facility for organising document property values according to a hierarchy or other structure. A central issue for this project was the high cost of metadata authoring in user time and effort; hence, a central interest for this project is the question of the extent to which metadata can be retrieved automatically. An inadequate quantity of metadata results in a poor user experience.

The metadata-rich filesystem continues to evolve, with content-based indexing systems now available on several operating systems. BeOS had a sophisticated model with a base set of file attributes, extended according to the type of the file, and indexed access. Many common Linux/Unix filesystems permit arbitrary text attributes to be added to files (using the 'extended attributes' system); the Reiser4 filesystem provides hooks for metadata-aware plugins, and Apple's 'Spotlight' feature allows searches based on various metadata attributes. Unfortunately, while filesystems are growing smarter, the question of interoperable transfer of metadata between filesystems has not yet been solved.

2.4 Looking a little closer

> *Granny peered closer. 'What's the curly thing?' she said.*
> *'Oh, that's the Adjustable Device for Winning Ontological Arguments,' said Shawn.*
> *'And this?'*
> *'That's the Tool for Extracting the Essential Truth from a Given Statement,' said Shawn.*
> *-- 'Carpe Jugulum', Terry Pratchett*

No matter the scale of the development, information management has a tendency to trip over a familiar set of limitations. This is by no means a revelation – at the risk of stating the obvious, functional systems are generally engineered according to design strategies that have proven successful in the past, and therefore operate in a manner that minimises most of these limitations. Whether this is explicitly recognised is a matter of little consequence. One intriguing possibility arises; whilst a tag corpus is developed under far less structure and guidance than – for example – a DC metadata record, and could therefore be expected to demonstrate less regular structure or content, it is nonetheless possible that strategies developed for manipulating a tag corpus could be applied to improve the effectiveness and power of search across 'semi-formal' metadata records.

The extraction of meaning from a piece of written language is, unfortunately, an extremely difficult problem; Sahlgren (2006) refers to the concept of 'meaning' as a 'holy grail' in the study of language. Not only is there a great deal of uncertainty as to exactly what the term means, there is even uncertainty as to whether the concept is a useful one (should it in fact exist). Fortunately for us, we do not have to retrace the steps of Wittgenstein any

further; instead, it is possible to engineer the problem such that the question does not arise, by precising that we are not interested in extracting 'meaning'; we are simply interested in modelling what one might by analogy to meteorology call 'isobars' of meaning. Extracting essential truth is not our motive; rather, extracting some idea of equivalence or similarity between term uses will suffice. Our principal interests are, for example, in granting a system with the ability to reliably guess as to whether two users who make use of a given string are referencing the same concept - or whether two users who make use of different strings are nonetheless attempting to reference the same concept – or whether two terms have similar meanings.

A model that granted us with an *inexpensive*, moderately reliable viewpoint on these questions would represent an invaluable resource. One such, the word space model, is reviewed in Sahlgren (2006). The word space model approach in general uses string co-occurrence to construct high-dimensional semantic vector spaces, with words represented as context vectors for computing semantic similarity. String co-occurrence models based around graph network analysis are simple to apply in tag analysis, and frequently appear, inherited largely from techniques applied in social network analysis (SNA). If language is designed around social lines, accurate social network information – or, in line with the intent to stick with inexpensive models, a cheap form of developing a mutually dependent set of models containing best fits at both social networking and semantic space models – could perhaps benefit in terms of increased accuracy.

Alternative approaches towards tag corpus analysis also abound. For example, the Kinds of Tags project (Baptista et al, 2007) describes an approach towards annotation of a tag set via Dublin Core attributes, with the intention that this annotation will support a number of aims, including investigation into the following set of questions:

- Into which DC elements can tags be mapped?
- What is the relative weight of each of the DC elements?
- What other elements come up from the analysis of the tags?
- Do tags correspond to atomic values?

An interesting side-effect of this kind of investigation is the exposure to differing theoretical backgrounds and research aims. DC metadata terms might be chosen as a starting point for such an annotation for a variety of reasons: for example, they are well-known and many participants have experience in their use; were they to prove a good fit for a tag set, an automated process for guessing at the corresponding DC term for a tag could benefit clear interface design; DC terms were developed over a decade to represent a 'best-fit' for use in resource description. Early indications are that several classes of tag exist that are not readily described via the set of DC terms.

A peculiarity of data mining as an approach is that it is often easy to find apparent structure in a given dataset. It is not, however, easy to demonstrate that this data is likely to appear in a consistent, reproducible manner – and it is often very difficult to reach an understanding of what this structure may actually signify or where it originates. What types of structure may be expected to exist in a corpus of contributed tags? What relations may exist between terms? If we were to demonstrate the accuracy of an assertion that tag use is generally declarative (ie. expressing a belief or statement), are the relations between terms similar to those found in other linguistic constructions? Perhaps tags are instead often applied according to Marshall's description, as part of a participatory process of working with a document. The tagging system on slashdot.org comes to mind, in which popular tags appear just below an article, meaning that tags are typically used for brief reactions heavy in cultural references – in the Slashdot vernacular – such as '!news', 'dupe', 'itslifejimbutnotasweknowit' and 'fud'. It is unlikely that the lessons learned in examining a single corpus apply seamlessly or in their entirety to another, and so it is most appropriate to see this approach as informing the development of a set of tools rather than a set of solutions or any form of stable or globally applicable heuristics.

3. Conclusion

As mentioned at the beginning of this paper, limitations in indexing consistency are often considered to be a question of the ongoing improvement of subject headings, and education and training for indexers. There is nothing inappropriate about this approach, assuming an appropriate context of use. Because tags are like little snatches of natural language, they fascinate some for the promise that they seem to offer of investigating the way in which people use language. Since social tagging systems are simple to use, appealing, can be created by novice users and are therefore of low cost, tagging is of interest to others for very pragmatic purposes, often in conjunction with other (more formal) systems.

It is all too easy to see this span of interests as representative of opposing views, rather than as a spectrum. This is a false dichotomy. Rather, the discontinuities result from differing contexts of use. With a population of well-trained indexers, the resulting terms are on the formal side of Marshall's dimensions of annotation, with an appropriately chosen audience and so forth. A heterogeneous population of individuals with little in common undertaking a wide range of tasks across a long period of time, on the other hand, generates a wide range of terms with a great deal more 'noise' – dissimilar contexts of communication.

The world-wide experiment that is social tagging may be worth watching for another reason, too; having acknowledged that many of the intriguing questions surrounding the processes of indexing and metadata generation have been sidestepped by careful system design and engineering, it is interesting to examine a system in which these concerns have not been dealt with and their impact minimised during the design phase. The results of such investigations may even be of some interest for future designers of other classes of indexing systems.

Social tagging is an interesting topic, not despite its historical underpinnings but because of them. Many discussion points that have been raised, examined and dismissed in the past are brought back to our attention by the mechanisms underlying social tagging. For myself, this is primarily related to the ungoverned nature of this particular environment, the heterogeneous nature of the user populations and the wide range of use cases for which social tagging systems are used. The efficacy of such systems as indexes is almost a detail (though an important one), since there are so many modes in which they may be used – for example, search, browsing, annotation in the many senses of the term highlighted by (Marshall, 1998) and supporting information collection and organisation in a task-oriented manner.

The development and evaluation of retrieval methods based around tag corpora is an important and well-represented research area, which for some types of tagging system may benefit from the application of techniques originating from areas other than indexing, particularly those which depend upon more complex or general-purpose models of language use.

Bibliography

Androutsopoulos, J. (2004). Language contact on the Internet: code-switching, language crossing and code choice.

Androutsopoulos, J. (2006). Sociolinguistics and computer-mediated communication. Journal of Sociolinguistics 10 (4), 419–438.

Baptista, A. A.; Tonkin, E. L.; Resmini, A.; van Hooland, S.; Pinheiro, S.; Mendéz, E.; Nevile, L. (2007). Kinds Of Tags: progress report for the DC-Social tagging community. Retrieved August, 2007 from http://hdl.handle.net/1822/6881

Barreau, DK and Nardi, B. (1995). Finding and reminding: File organization from the desktop. ACM SIGCHI

Bulletin, 27 (3), 39-43.

Boztepe, (2003) Issues in Code-Switching: Competing Theories and Models. Working Papers in TESOL & Applied Linguistics, Vol 3, No 2 (2003)

Chalmers, M. (2004). "Hermeneutics, information and representation," European Journal of Information Systems (13:3), p 210

Cooper, W. S. (1969). Is Inter-indexer consistency a hobgoblin? American Documentation, 20, 268-278.

Deutscher, G. (2005). The Unfolding Of Language: The Evolution Of Mankind's Greatest Invention. Metropolitan Books. ISBN: 978-0805079074

Dourish, P.; Edwards, W. K.; LaMarca, A.; Salisbury, M. Presto: an experimental architecture for fluid interactive document spaces. ACM Transactions on Computer-Human Interaction. 1999 June; 6 (2):133-161.

Dourish, P., Edwards, W. K., LaMarca, A., Lamping, J., Petersen, K., Salisbury, M., Terry, D. B., & Thornton, J. (2000). Extending Document Management Systems with User-Specific Active Properties. ACM Transaction on Information Systems, 18(2), 140-170.

Ellis, N. C. & Larsen-Freeman, D. (2006). Language Emergence: Implications for Applied Linguistics. Introduction to the Special Issue. Applied Linguistics, 27(4), 558-589.

Fertig, S., Freeman, E. and Gelernter, D. (1996). "Finding and reminding" reconsidered. ACM SIGCHI Bulletin, 28 (1), 66-69.

Fishman, Joshua A. 1972. Domains and the relationship between micro- and macrosociolinguistics. In: J. Gumperz and D. Hymes, eds. Directions in sociolinguistics. The ethnography of speaking. New York: Holt, Rinehart and Winston, 407-434

Gifford, D. K.; Jouvelot. P.; Sheldon, M. A.; O'Toole, J. W. (1991) Semantic file systems, Proceedings of the thirteenth ACM symposium on Operating systems principles, p.16-25, October 13-16, Pacific Grove, California, United States

Goldstone, R. L. (1994). Influences of categorization on perceptual discrimination. Journal of Experimental Psychology General 123: 178-200.

Hjørland, B. (2002). Epistemology and the Socio-Cognitive Perspective in Information Science. Journal of the American Society for Information Science and Technology, 53(4), 257-270

Krauss, R. M. & Fussell, S. R. (1991). Constructing Shared Communicative Environments. In: Perspectives on Socially Shared Cognition, ed. Lauren B. Resnick, John M. Levine & Stephanie D. Teasley. American Psychological Association.

Lave, J., & Wenger, E. (1991). Situated learning: Legitimate peripheral participation. Cambridge, England: Cambridge University Press.

Leonard, L. E. (1975). Inter-indexer consistency and retrieval effectiveness: Measurement of relationships. Unpublished dissertation, University of Illinois, Urbana-Champaign.