Automatic Spatial Metadata Enrichment: Reducing Metadata Creation Burden through Spatial Folksonomies

Mohsen Kalantari[1], Hamed Olfat[2], Abbas Rajabifard[3]

Centre for SDIs and Land Administration
The University of Melbourne, Victoria 3010

[1] Research Fellow saeidks@unimelb.edu.au
[2] PhD student h.olfat@pgrad.unimelb.edu.au
[3] Associate Professor abbas.r@unimelb.edu.au

Abstract

Metadata plays a key role in facilitating access to up-to-date spatial information and contributes to the finding and delivering of high quality spatial information services to users. In particular, metadata is an important element in functioning and facilitating spatial data infrastructure (SDI) initiatives. With huge amount of spatial information being generated, a spatial application must be sufficiently flexible to extract and update spatial metadata automatically.

Automatic spatial metadata generation framework includes three fundamental but complementary streams; automatic creation, automatic update and automatic enrichment of spatial metadata. This paper explores the automatic metadata enrichment stream based on the tagging and folksonomy concepts. The paper argues how folksonomies help bringing the vocabulary of spatial data users into play and using them hand in hand with those sometimes mysterious terms supplied by experts in metadata records.

The paper then builds on the tagging and folksonomy concepts and proposes a conceptual model to employ them for spatial metadata enrichment. The paper finally discusses advantages and disadvantages of this approach against formal type of organizing spatial metadata.

Keywords: spatial data, metadata, automation, tagging, folksonomy

1. INTRODUCTION

Spatial Data Infrastructure are placed under pressure by a need for efficient and effective ways of indexing and organizing an increasing number of spatial datasets that are being added through both previously and recently created datasets. A well compiled metadata for datasets plays a critical role in searching and discovering spatial datasets. This is in particular essential for spatial data users in finding the correct datasets and for the spatial data systems such as geographical information systems and geo web services for interoperability.

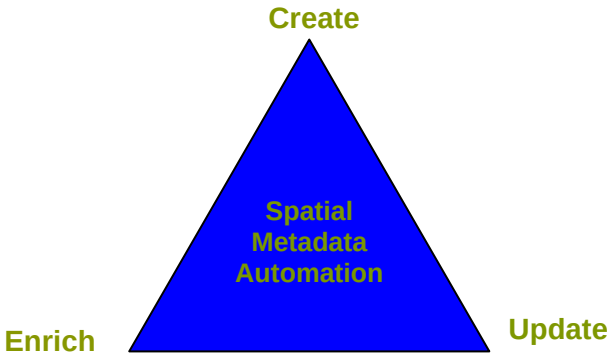
However authoring and compiling metadata for spatial datasets are often labor intensive and time consuming. Methods and approaches to overcome these issues are welcomed by the spatial industry. The concept of automatic spatial metadata generation research is rooted in automatic indexing, abstracting, and classification of spatial data content, which began with the need to organize increasing amounts of

spatial related data and the inability of manual methods to cope with huge amounts of spatial metadata (Rajabifard *et al.* 2009).

Today, automatic metadata generation should move beyond subject representation to encompass the production of author, title, date, format, spatial extension and many other types of metadata. In addition, thousands of spatial databases are now networked via the Internet, and information resources are frequently rendered in open and interoperable standards (e.g., eXtensible Markup Language or XML). These developments should enable automatic metadata generation systems to work on far larger spatial data directories.

Automatic spatial metadata generation research efforts are relatively new in the spatial field however a framework for spatial metadata automation which includes automatic creation, update and enrichment of spatial metadata is shown in Figure 1 (Kalantari *et al.* 2009). This framework defines three fundamental but complementary streams that can be employed for metadata automation as described below:

Figure 1: Spatial Metadata Automation Framework



1.1 Automatic update

Automatic spatial metadata update or synchronization is a process by which properties of a spatial dataset are read from the dataset and written into its spatial metadata. Some software vendors such as ESRI conceptualize metadata automation as synchronizing the metadata content when values in the spatial data change. In this type of model, for instance when a change occurs with a spatial data property such as its projection, the metadata will be updated with the new information. Olfat *et al* (2010) proposed a GML based approach to facilitate automatic metadata updates. In this approach GML is used as a medium for transferring metadata details from one file to another and monitor changes for automatic updating of the metadata content.

1.2 Automatic creation

While automatic update and synchronization is suitable for updating an existing metadata record, there is a need for other methods when there is no existing metadata associated with spatial data. Humans create metadata by writing descriptions of resources either in a structured or unstructured form. Computer applications can then extract certain information from a resource or its context. This may simply involve capturing information that is already available, such as the format of the file, or running an algorithm to determine the subject of a textual resource by counting keywords or by checking and analyzing pointers to the resource.

1.3 Automatic enrichment

The third aspect of metadata automation which will be further discussed in this article is automatic enrichment that involves improving content of the metadata through monitoring tags and keywords that are used by users for finding datasets. Creating metadata by monitoring user interaction is based on the Folksonomy concept, a concept that was first introduced by Thomas Vander Wal in 2004. He stated: “Folksonomy is the result of personal free tagging of information and objects for one's own retrieval. The tagging is done in a social environment (usually shared and open to others). Folksonomy is created from the act of tagging by the person consuming the information. The value in this external tagging is derived from people using their own vocabulary and adding explicit meaning, which may come from inferred understanding of the information/object. People are not so much categorizing, as providing a means to connect items (placing hooks) to provide their meaning in their own understanding.

Vander Wal adds, "Folksonomy is tagging that works. This is still a strong belief the three tenets of a folksonomy: 1) tag; 2) object being tagged; and 3) identity, are core to disambiguation of tag terms and provide for a rich understanding of the object being tagged.”

This paper introduces an automatic metadata enrichment stream based on tagging and folksonomies and also compares this stream against the other information indexing approaches. The paper then builds on the tagging concept and proposes a conceptual model to employ the tagging concept for spatial metadata enrichment. The paper finally discusses the advantages and disadvantages of this approach against other formal type of organizing spatial metadata.

2. FROM TAGGING TO FOLKSONOMIES

Traditionally metadata is created by dedicated professionals (Mathes 2004). Similarly in the spatial field, metadata experts create metadata, and this is the basis of most catalogues in SDIs. This often requires serious knowledge and background. The spatial industry in the metadata field has developed standards and schemes for cataloguing, categorization and classification of spatial data.

While professionally created metadata are often considered to be of high quality, it is costly in terms of time and effort to produce. This makes it very difficult to scale and keep up with the vast amounts of new spatial data being produced and updated, especially with new technologies like Global Navigation Satellite System, Satellite imagery, automatic map creation methods and in particular mediums like the World Wide Web.

An alternative is author created metadata. Original producers of the spatial data provide metadata along with their creations. The Dublin Core Metadata Initiative has been used with some success in this area (Greenberg *et al*. 2001). Author created metadata may help with the scalability problems in comparison to professional metadata, but both approaches share a basic problem: the intended and unintended eventual users of the spatial information are disconnected from the process.

There is a third approach that can be utilized to capture spatial information users’ notion of data and information to create metadata for the spatial information. Users will use their own language to describe the spatial information. This can go even deeper and they can express their comments not only about the data title but also on the other aspects of the metadata. In a sense this is a way of connecting users of spatial data to the process of creating spatial metadata. Sharing spatial datasets in

SDIs and allowing users to write notes about them, tag them and even share their notes with the other users opens another horizon for automation of spatial metadata and efficient using of them. This section provides a background review of tagging and folksonomy concepts to explore the potential of them in spatial metadata automation.

2.1 Tagging Concept

Returning to the history of the web technology, web browsers have allowed a user to “bookmark” a web site and organize these bookmarks into hierarchical file folders similar to the filing systems they use with paper files and the electronic files on their computer. As the web grew, filing bookmarks in one’s browser became unwieldy. Sites like Delicious.com allow users to save the URL for a web site; provide an annotation if desired; supply a number of tags that will help to retrieve it again; and group similar URLs together. These tags can be whatever or how many associations the user wishes to make with the URL. In this way, the URL can be associated with many concepts at once. This ability to use tags to bring out different aspects of a resource is a major advantage of tagging over formal systems of classification and taxonomies (controlled vocabularies)(Thomas et al. 2009) which can be adopted for spatial metadata automation.

According to Shirky (2005), traditional classification methods for information resources attempt to systematically organize knowledge by providing a single classification for a resource. For instance, in spatial information one can classify geo-coded features together with the addressing layer in a single classification as landmarks. Otherwise they can be separated into two different classification administrative and landmark layers.

Shrinky (2005) further argues that the free associations made by taggers are the only appropriate way to organize resources on systems as large and chaotic as the web for three reasons: classification fails to allow more than one place for an item; it is impossible to keep a classification system stable over time; and it is also impossible for an expert to truly predict how a user will search for something. This cannot be entirely relevant to SDIs but to an extent in spatial arena with an increasing number of spatial data layers, users and applications. First, it is difficult to stabilize a classification of information, secondly (and more importantly), it is impossible to predict how increasing number of interested users and applications might name or interpret a spatial data set.

The tagging approach provides more freedom for users, because when tagging, the user does not have to make a decision and restrict the resource to just one or two formal terms from a controlled keywords they may or may not be familiar with. Instead the users supply their own terms which are meaningful to them (Shirky 2005). For instance, spatial data users can select different tags to describe the same item. Items related to scale may be tagged "500", "1:500", "1/500". This flexibility allows users to classify their collections of items in a way that they find useful and users also have to decide whether each tagged item is actually relevant to what they're looking for.

Sinha (2005) performed a cognitive analysis of tagging, stating that it works well for users because it lowers the cognitive cost of making decisions on how to categorize a resource. It is easier for people to assign and remember their own terms for later retrieval. According to Mathes (2004) and Shirky (2005) tagging works by lowering the barrier to participation. Users do not have to be experts and learn complex rules and specialized vocabulary in order to tag as they do when applying a controlled vocabulary (Mathes 2004).

The ease and freedom with which one can apply tags explains why tagging is so popular and form folksonomies which will be discussed in the next section.

2.2 Folksonomies

Besides identifying different functions for tags, Golder and Huberman (2006) also describe collaborative tagging systems. Users tag primarily for themselves but the software makes it possible to see all the tags used for a resource so that a user may utilize tags from other users.

In this fashion, the folksonomy becomes a common vocabulary grown from the ground up. As the number of uses increases, each resource develops a “tag cloud” or a cluster of tags denoting popularity. Furthermore, the most popular resources are tagged the most frequently which, in turn, influences other users in their choice of tags.

Actually, the most popular tags for a resource turn out to be an accurate representation of the resource, exhibiting a power law distribution, with common descriptive tags being used in far greater proportion to the more varied or personally oriented tags (Golder and Huberman 2006). The system then will be operating on a folksonomy basis.

Users, tagging for themselves, collectively create useful sets of subject descriptors in the form of tags for the resources they are tagging and this user-added metadata can then be leveraged for information retrieval on a general as well as a personal level.

Having described tagging and folksonomies, the next section lays the ground work for metadata enrichment using these concepts towards automatic metadata creation.

3. AUTOMATIC SPATIAL METADATA ENRICHMENT

Considering spatial metadata’s small size compared to the data it describes, it is more easily shareable (ESRI 2002) and is considered as the surrogate of a spatial dataset which is referenced to its related spatial dataset. Hence, in a networked environment such spatial surrogates are discovered by the users seeking out required spatial datasets through catalogue systems, web services and the user interface. The user interface usually supports making a variety of queries (via basic and advanced searches) on spatial metadata records to retrieve the characteristics of the most appropriate datasets for end users.

These queries are generally based on the keywords or phrases used by the spatial data users. The keyword element is also one of the mandatory elements recommended by the ISO 19115 Metadata Standard which is embedded into each spatial metadata file and is defined as “commonly used word(s) or formalized word(s) or phrase(s) used to describe the subject”(ISO 2003).

In this regard, finding effective keywords to describe the spatial data sets is fundamental within any sharing platform. The right keyword for any spatial data set means the keyword which is consistent with the content of the data set and can reveal its essence and applications. In addition, a good keyword should be comprehensive and address the probable queries made by users from diverse categories. Moreover, a keyword should be popular meaning that most of the users agree on that keyword.

The new form of metadata that are created by users, the Folksonomy introduced earlier, can facilitate the generation of good keywords for any sharable resources. Folksonomic metadata consists of words that users generate and attach to content, which are well known as tags (Alexander 2006).

"Geo-tags" might be the good example of tags as the keywords linked to a concrete position (Heuer and Dupke 2007). Geo-tagging allows for easily combining attributive information with spatial location. People share their geo-tagged contents on platforms like Flickr or the Google Earth Community.

However, the tagging concept is new in the spatial arena and can be considered as one of the potential ways to enrich the spatial metadata content. As a result of this, the automatic spatial metadata enrichment, as one the main streamlines of spatial metadata automation (Kalantari *et al.* 2009) involves improving the content of metadata through monitoring the popular searched keywords and tags used for finding the spatial data sets to their related metadata records.

Generating this kind of spatial metadata can help describe a data set and allow it to be found again by browsing or searching easily and quickly. These tags will be chosen systematically or informally and personally by the spatial data publisher or by its users, depending on their use. On a spatial data directory where many users are allowed to tag much spatial data, this collection of tags can become a spatial folksonomy, that is, a method that can collaboratively create and manage metadata to annotate and categorize spatial data.

It is arguable that tagging and folksonomies are only workable in large scale information resources such as broader WWW with huge number of users rather than relatively small scale networks such as SDIs. Addressing this argument, the next section proposes a mechanism to directly and indirectly involve the users in the tagging process and capture their knowledge in the system to create a spatial folksonomy for SDIs. The folksonomy then created can be used to enrich the metadata content of spatial datasets.

4. CONCEPTUAL DESIGN FOR AUTOMATIC METADATA ENRICHMENT

As discussed above, tagging and folksonomies can be employed to help automate metadata by enriching the content. In this space, there are two complementary approaches for metadata enrichment; system and user oriented approaches.

Consequently, to implement the automatic spatial metadata enrichment concept based on tagging, two models have been designed. The first model is indirect tag generation for spatial data sets based on system oriented approach and the second model is direct tag generation by engaging the spatial data users in tagging process. These streams are discussed below:

4.1 Indirect automatic enrichment model

The system oriented design is concentrated on monitoring tags that are employed by spatial data users, analyzing them and then employing the tags to enrich the content of metadata.

The indirect automatic enrichment model is streamlined in three stages (Figures 2, 3, 4):

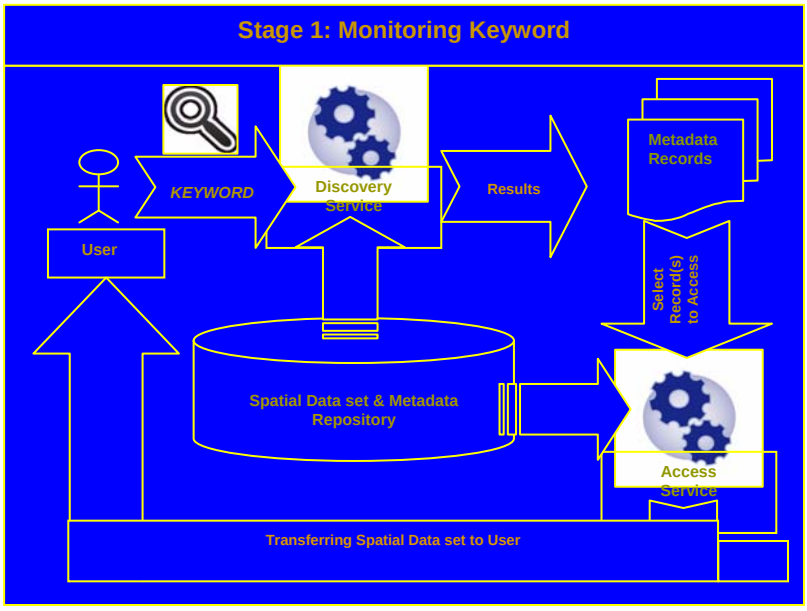
- Monitoring keyword

- Recording keyword
- Assigning keyword

Stage 1: Monitoring keyword

The data catalogue systems consisting of data and metadata repositories (distributed or centralized) typically provide the users with the services to discover and access the data. As mentioned earlier, through these systems one of the common ways of data discovery is to query the metadata records via searching using keywords. Then, the discovery service will search and retrieve all the corresponding metadata records with that keyword. The users will be able to view the results by opening the metadata files and deciding on which data is more suitable for their needs. Finally, through the access service they will be able to access the required data or be aware of the access policies and rules.

Figure 2: Stage 1/ The Automatic metadata automatic enrichment flow

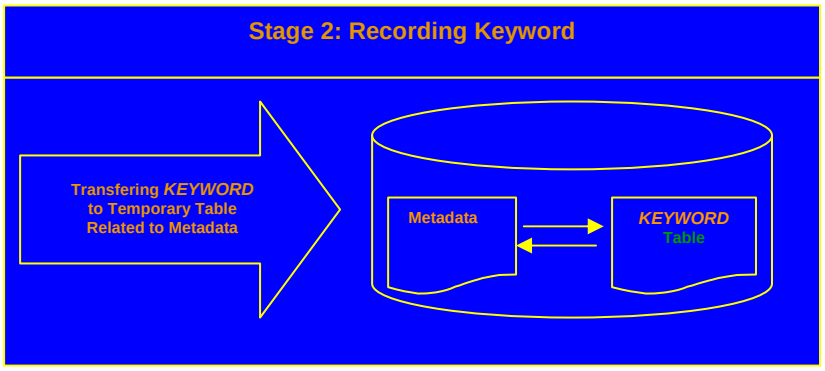


The main aim of Stage 1 is to identify the keywords used by the users which allow them to utilize the search process. To do so, any used keyword is monitored during the process of data discovery and access. This stage is based on the assumption that any keyword directing the user to the final required data may be used in the future by the same user or other users; thus it should be recorded by the system.

Stage 2: Recording keyword

In this stage, any keyword relevant to any spatial dataset which has been identified in Stage 1 would be recorded in a temporary database. This database is related to the corresponding metadata records through data fields such as “Metadata ID”, “Spatial Dataset Name”, “Keyword” , “Number of Repetition” etc. The main aim of this stage is to recognize how many times any keyword is used in the discovery process for the same spatial data set. Increasing use of the same keyword illustrates its popularity.

Figure 3: Stage 2/ The Automatic metadata automatic enrichment flow

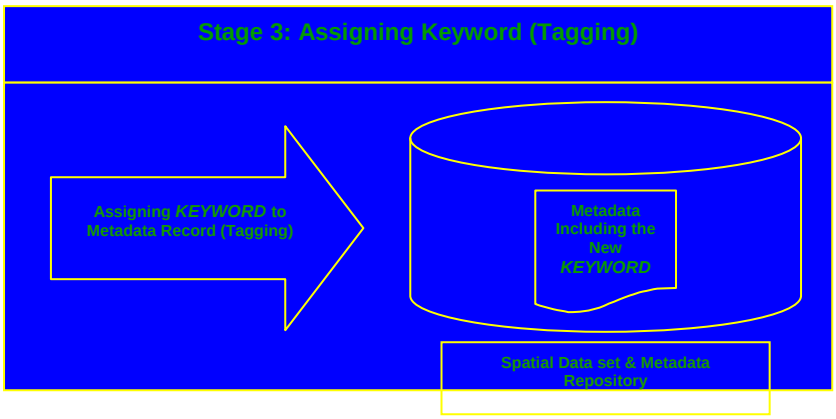


Stage 3: Assigning keyword

Among the keywords recorded in Stage 2, any of them which have a specific value of repetition will be assigned to its spatial metadata file and stored in its keywords field. Indeed, the new keyword will be added to the spatial data set.

Through indirect enrichment process, the popular keywords used for finding spatial data sets are identified and shared between users. In addition, this process will facilitate the spatial data discovery within data catalogue systems. This process enriches the metadata records related to spatial data sets through applying and refining the appropriate keywords.

Figure 4: Stage 3/ The Automatic metadata automatic enrichment flow



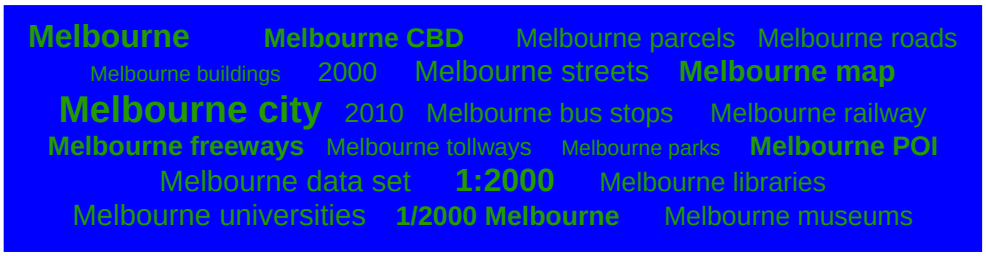
4.2 Direct metadata enrichment model

Contrary to the indirect metadata enrichment model based on the system oriented approach, in the direct model to enrich the spatial metadata content a tagging process by the users is also considered. Through this process, the spatial data users tag a data set with words they feel best describe what it is about. Accordingly, they will be involved in the enrichment process and the knowledge of the users about spatial data sets will be shared. Moreover, the tagging process will help the users easily and quickly find their tagged data sets again within the spatial data catalogue.

The users will tag the data sets according to their awareness of data and also the intention of using that data which is usually related to their circumstances. These tags are visualized by “Tag Cloud” in the spatial folksonomy. Within the tag cloud, the

tags which are used more frequently by the users will be highlighted and shown in a bold format (Figure 5). For instance, the tagging process for 1:2000 map of Melbourne city can be imagined here. The users of this map can be Melbourne citizens, tourists, students, or decision makers from different organizations.

Figure 5: The Tag Cloud for Melbourne Map 1:2000



The users are also able to agree and disagree with the existing tags in the tag cloud by clicking or not clicking on the tags. Having designed the model for metadata data enrichment through folksonomies the next section discusses the efficiency of a system based on folksonomies.

5. ENRICH OR NOT TO ENRICH BY TAGS

A spatial metadata enrichment system based on tagging and folksonomies will benefit both spatial data publishers and users in terms of facilitating the data discovery process, involving users in metadata creation and enrichment, making the data catalogue systems more user-friendly, and sharing the users' knowledge about spatial data sets.

A disadvantage of a spatial tagging system could be where there is no information about the meaning or semantics of each tag. For example, the spatial tag "Melbourne" might refer to the Central Business District of Metropolitan Melbourne or Metropolitan Melbourne itself and this lack of semantic distinction can lead to inappropriate connections between spatial data.

All tags have this problem of ambiguity. Different users may use the same tag to mean different things because they are applying it in different contexts to different resources. For instance, one user may assign the tag "Property" to a resource about generic cadastral layers while another user may use the same tag to refer only to parcel layers.

Tags can be applied at different levels of specificity by different users (or even by the same user at different times). Besides different terms may be used for the same concept (again by different users or by the same user – users will not necessarily be consistent) (Hayman 2007).

Typically, no information about the meaning of a tag is provided although the indirect metadata enrichment model, proposed earlier, makes sure the user is tagging in defined framework. The user will be provided with a number of suggestions, however still will remain independent deciding on labelling a data set.

However, larger-scale spatial folksonomies can address some of the problems of tagging, as users of spatial tagging systems tend to notice the current use of "tag terms" within these systems, and thus use existing tags in order to easily form connections to related items. In this way, spatial folksonomies collectively develop a

partial set of metadata standards through ongoing involvement of non-expert spatial users (Kalantari *et al.* 2009).

6. CONCLUSION

With increasing amount of spatial information being generated, SDIs must sufficiently manage updating spatial metadata automatically. This paper built on folksonomy and tagging concepts and proposed a solution for metadata automation by enriching the content of metadata through monitoring keywords used and tags allocated by users. A brief introduction to tagging and its evolution towards folksonomies in the first part of the paper outlined the weaknesses of expert generated keywords which are the strengths of folksonomies.

Through the conceptual design of automatic metadata enrichment process the paper illustrated there is a fundamental difference between browsing and finding in spatial data discovery. Browsing tags in folksonomies is valuable for unanticipated discoveries of related datasets. That is a much different task than searching for every data layer in an SDI area using a specific term.

Folksonomies directly reflect the vocabulary of users instead of using the sometimes mysterious terms supplied by experts. Keywords in metadata records over time become rigid, out of date, and distant from the every day language of the growing number of users.

However it should be emphasized that folksonomies and controlled keywords have different strengths which can complement each other since the strengths of one are the weaknesses of the other. Keywords and standards enable uniform access and interoperability. Folksonomies on the other hand, brings user language, perspective, expertise, and eventually will lead towards more user oriented metadata. This article is based an ongoing research in the area of metadata automation and recommends combining folksonomies with controlled keywords in SDIs, to create richer metadata.

ACKNOWLEDGEMENT

The authors wish to acknowledge the support of the Australian Research Council – Linkage Fund and the members of the Centre for Spatial Data Infrastructures and Land Administration at the Department of Geomatics, the University of Melbourne, in the preparation of this paper and the associated research. However, the views expressed in the paper are those of the authors and do not necessarily reflect those of these bodies.

REFERENCES

Alexander, B. (2006). "Web 2.0: A New Wave of Innovation for Teaching and Learning?" *Educause Review* 41(2).

ESRI (2002) "Metadata and GIS." An ESRI ® White Paper,

Golder, S. A. and Huberman, B. A. (2006). "Usage patterns of collaborative tagging systems " *Journal of Information Science* 32(2): 198-208.

Greenberg, J., Maria Cristina Pattuelli, Parsia, B. and Robertson., W. D. (2001). "Author-generated Dublin Core Metadata for Web Resources: A Baseline Study in an Organization." *Digital Information* 2(2): 78.

Hayman, S. (2007). Folksonomies and Tagging: New Developments in Social Bookmarking. Ark Group Conference: Developing and Improving Classification Schemes. Sydney, Rydges World Square.

Heuer, J. T. and Dupke, S. (2007). "Towards a Spatial Search Engine Using Geotags". *GI-Days 2007 - Young Researchers Forum*.

ISO (2003). *Geographic Information - Metadata (ISO 19115:2003)*.

Kalantari, M., Rajabifard, A. and Olfat, H. (2009). "Spatial metadata automation". *Surveying and Spatial Sciences Institute Biennial International Conference*, Adelaide, South Australia.

Mathes, A. (2004). "Folksonomies - Cooperative Classification and Communication Through Shared Metadata." from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.135.1000&rep=rep1&type=pdf.

Olfat, H., Rajabifard, A. and Kalantari, M. (2010). Automatic Spatial Metadata Update: a New Approach. FIG 2010 Congress. Sydney, Australia.

Rajabifard, A., Kalantari, M. and Binns, A. (2009). SDI and Metadata Entry and Updating Toolsin. B. van Loenen, J. W. J. Besemer and J. A. Zevenbergen (Eds). *SDI Convergence. Research, Emerging Trends, and Critical Assessment* Delft:pp 121-136.

Shirky, C. (2005). "Ontology is overrated: categories, links, and tags." Clay Shirky's Writings about the Internet Retrieved 24 Feb 2010, from www.shirky.com/writings/ontology_overrated.html

Sinha, R. (2005). "A cognitive analysis of tagging." Retrieved 3 March 2010, from http://rashmisinha.com/2005/09/27/a-cognitive-analysis-of-tagging/.

Thomas, M., Caudle, D. M. and Schmitz, C. M. (2009). "To tag or not to tag?" *Library Hi Tech* 27(3): 411-434.

Author/s:
Kalantari, Mohsen; OLFAT, HAMED; RAJABIFARD, ABBAS

Title:
Automatic spatial metadata enrichment: reducing metadata creation burden through spatial folksonomies

Date:
2010

Citation:
Kalantari, M., Olfat, H., & Rajabifard, A. (2010). Automatic spatial metadata enrichment: reducing metadata creation burden through spatial folksonomies. In GSDI 12 World Conference: Realising Spatially Enabled Societies, Singapore.

Publication Status:
Published