An algebra and conceptual model for semantic tagging of collaborative digital libraries.

Epaminondas Kapetanios[1]
Markus Schaal[2]

[1] Harrow School of Computer Science, University of Westminster
[2] Department of Computer Engineering, Bilkent University

This is an electronic version of a paper presented at the Second Workshop on Foundations of Digital Libraries in conjunction with 11th European Conference on Research and Advanced Technologies on Digital Libraries (ECDL 2007), 20 Sep 2007, Budapest, Hungary. Available online at:

http://www.delos.info/index.php?option=com_content&task=view&id=597&Itemid=328

A Model and Algebra for Collaborative Semantic Annotation in Digital Libraries

Epaminondas Kapetanios[1] and Markus Schaal[2]

[1] University of Westminster, School of Computer Science
London, United Kingdom
e.kapetanios@wmin.ac.uk
[2] Bilkent University, Department of Computer Engineering
Ankara, Turkey
schaal@cs.bilkent.edu.tr

Abstract. Cost-effective semantic description and annotation of shared knowledge resources has always been of great importance for digital libraries and large scale information systems in general. With the emergence of the Social Web and Web 2.0 technologies, a more effective semantic description and annotation, e.g., *folksonomies*, of digital library contents is envisioned to take place in collaborative and personalised environments. However, there is a lack of foundation and mathematical rigour for coping with contextualised management and retrieval of semantic annotations throughout their evolution as well as diversity in users and user communities. In this paper, we propose an ontological foundation for semantic annotations of digital libraries in terms of *flexonomies*. The proposed theoretical model relies on a high dimensional space with algebraic operators for contextualised access of semantic tags and annotations. The set of the proposed algebraic operators, however, is an adaptation of the set theoretic operators *selection, projection, difference, intersection, union* in database theory. To this extent, the proposed model is meant to lay the ontological foundation for a Digital Library 2.0 project in terms of geometric spaces rather than logic (description) based formalisms as a more efficient and scalable solution to the semantic annotation problem in large scale.

Keywords Social Web, Collaborative Systems, Conceptual Modelling, Web 2.0, Digital Libraries, Semantic Annotation

1 Introduction

With the emergence of the Social Web and Web 2.0 technologies, semantic tagging and annotation of shared knowledge resources promises to become a more intuitive, user specific and scalable solution for the Semantic Web. *Folksonomies* [13] are prominent examples of semantically tagging shared knowledge repositories on the Web. This approach contrasts with the currently suggested Web ontology description languages such as RDF and OWL in that a user centric, light weight tagging and semantic annotation of knowledge sources is enabled. This puts the emphasis on intersubjective communication, argumentation and interpretation rather than formally describing common agreed upon conceptions of artefacts.

A prominent case study of an application of *folksonomies* and the Social Web to digital libraries is given in terms of a *bibsonomy* (http://www.bibsonomy.org). In bibsonomies, each bibliographic reference is enriched by multiple, possibly concatenated, keywords as semantic tags by each user. Those can be understood as initial and emerging conceptual structures, or concepts, to be further evolved. A *bibsonomy* can be understood as an ontology model, which spans a conceptual space over user specific clouds of semantic tags for bibliographic entries.

However, this way of semantically describing and sharing of knowledge sources also poses some major challenges. For instance, heterogeneous in meaning and, eventually, conflicting semantic annotations arise out of particular users or user communities in collaborative environments. Furthermore, evolutionary aspects of semantic annotations and concepts via improved understanding of annotated objects over time has not been an issue. There is also a lack of foundation and mathematical rigour for *folksonomies* in terms of contextualised management and retrieval of semantic annotations throughout their evolution as well as diversity in users and user communities.

In this paper and in response to these challenges, we envision and discuss a *flexonomy* as a collaborative semantic annotation reference model of digital library artefacts, e.g., images, manuscripts, etc. This approach keeps the flexibility and scalability of collaborative and user specific semantic tagging and annotations, however, on an mathematically and ontologically founded ground. To this extent, a model and algebra for *flexonomies* is discussed, which reflects the dynamics of semantic annotations as directed by their variation across time, sources and agents, e.g., semantic taggers, as well as by uncertainty and flexibility in definitions, labelling and networking of semantic tags and annotations. Semantic annotations in a *flexonomy* are bound with a particular context as posed by the dimensions, across which variation in their definition occurs.

The proposed theoretical model relies on a high dimensional conceptual space within which semantic tags and annotations are located according to their mappings into dimensional points. In contrast with other, usually vector space models for context modelling in information retrieval, it also relies on algebraic operators for contextualised access and manipulation of semantic tags and annotations. The proposed algebraic operators are an extension of the classical set theoretic operators *selection, projection, difference, intersection, union* towards embedding of n-dimensional subspaces into their definitions. To this extent, the proposed model is meant to lay the ontological foundation for a Digital Library 2.0 project in terms of geometric spaces rather than logic (description) based formalisms as a more efficient and scalable solution to the semantic annotation problem in large scale.

2 Related Work

Since concepts are the most basic units of thought, it is not surprising that they became important building blocks of suggested conceptual structures for knowledge representation. In particular, their appearance is prevailing in semantic networks [10], conceptual graphs, taxonomies, description logics and ontologies [2], which became a key issue with the emergence of the Semantic Web [4].

The predominant, logic based paradigm of formalising knowledge for the Semantic Web, however, concentrate mostly on mechanising subsumption reasoning within a common agreed upon Ontology, which has its limitations, especially when it comes to dealing with conceptual diversities, overlapping knowledge, versioning and conflicting views within an emerging Ontology or conceptual structure. Approaches to deal with diversities of Ontologies have become an issue for, e.g., agent technologies by introducing local concensus Ontologies rather than global ones. This is also driven by the fact that agents normally wish to start with individualised Ontologies and collaboratively develop a global, consensus Ontology [11]. Therefore, engineering and merging of domain specific Ontologies has also been addressed, to some extent, within the context of end user driven knowledge engineering too such as in [7].

Despite the fact that researchers already addressed context modelling [6] and contextualised ontologies via logic based formalisms such as the C-OWL [3] approach, an attempt to extend the OWL formalism in order to express context, there has also been attempts to restructure logic or seek alternative context representation schemes. For instance, Rudolf Will and his students formulated a mathematical theory in 1978 in terms of a *Formal Concept Analysis (FCA)* and its convergence with conceptual graphs. FCA has been introduced in order to provide some understanding of the term *concept* in terms of lattice theory. Since then, FCA has been considered not only within AI, but also in other computer science domains such as Software Engineering or Database Theory [9]. The convergence of FCA with computer science increased significantly by the series of International Conferences on Conceptual Structures (ICCS). An exemplary convergence with conceptual graphs, in particular, is given by [14].

An alternative to logic based context modelling has also been offered by the means of geometrical spaces [5], especially in the field of information retrieval [12]. With digital libraries as a predominant application in information retrieval, context modelling by using vector space bases has been addressed, for instance, in [8]. However, these modelling approaches are primarily targeting indexing issues for documents retrieval, whereas in our approach, we are discussing an algebra and operators for retrieval of contextualised and personalised views of collections of concepts within the proposed geometrical space. Furthermore, the definition of our high dimensional space allows the existence of dimensions at various granularity levels.

With the emergence of more pragmatical approaches for semantically annotating knowledge resources on the Web, *folksonomies* have been recently a response to the need for collaborative and flexible taxonomies [13]. This user centric approach to semantic tagging and classification of concepts, however, lacks a mathematical and ontological foundation. An attempt to formally describe personalised or user specific annotations has been offered by [1]. In our approach, however, we lay the foundation for a conceptual space where information access to the semantic annotations are enabled via a series of algebraic operators, which are bound with the dimensions of the space. To this extent, a more expressive query language for contextualised or personalised views of concepts or semantic annotations could be built upon these operators.

3 The Theoretical Foundation: Model and Algebra

In most modern mathematical formalisms, set theory provides the language in which mathematical objects are described. Complying with this tradition, the introduced model relies space models and set theory, however, with an extension towards *collections of concepts*. Given that *concepts* are always bound with an n-dimensional subspace to represent context, the algebraic operators is an introduction to what can be done with collections of concepts. To this extent, it might also help in precisely defining the still vague term *collection* in mathematics.

3.1 The Model

Given that *flexonomies* are meant to support light weight semantic annotation of knowledge sources and, in particular, digital libraries by relaxing many of the heavy weight ontology engineering principles, the following definitions aim at ontologically founding semantic tags and annotations as bound with flexibility and a context. Assuming that F is the set of potential concepts and C is the set of concepts represented in a *flexonomy*, i.e., things, relationships, instances, which semantically tag or annotate knowledge sources, we define the following.

Definition 1 *T is the function (total and one-to-one mapping) from F to* unique identifiers[3] *UUID's. Consequently, C_T, where $\{c \in C_T \subseteq C\}$, is defined as the set of* uniquely identified concepts*, i.e., those concepts having only a unique identifier, e.g., $C_T : \{100, 101, 102\}$.*

L is the function (partial and many-to-one mapping) from C_T to the set of labels[4]*. Consequently, C_L, where $\{c \in C_L \subseteq C\}$, is defined as a set of* labelled *concepts. This denotes that not all concepts or artefacts should be necessarily labelled and uniquely identified concepts might be assigned the same label, e.g.,*
$C_L : \{(100, Information), (102, Information)\}$

D is the function (partial and many-to-one mapping) from the union set $C_T \cup C_L$ to the set of descriptions[5]*. Consequently, C_D, where $\{c \in C_D \subseteq C\}$, is defined as a set of* described *concepts. This denotes that labelled or unlabelled concepts might be assigned a description. It also denotes that the same description might be assigned to more than one concept or artefact, e.g.,*
$C_D : \{(101, description1), (100, Information, description2)\}$.

Definition 2 *A collection of* **well-identified concepts** *C_I, where $\{c \in C_I \mid c \in C_D \wedge$ $\Pi_{UUID,Label}(c) \in C_L\}$, is defined as the set of concepts being members of C_D and their projection over* UUID *and* Label *is a member of C_L, i.e., they are composed of a unique identifier, a label and a description and, therefore, have been more semantically enriched. For instance, $C_I : \{(100, Information, description2)\}$.*

[3] A Universally Unique Identifier UUID has been suggested by the Open Software Foundation (OSF) as an identifier standard for software construction as part of a Distributed Computing Environment (DCE). It is meant to identify the same thing in different contexts.

[4] A label is a tag as a textual token to name entities. They are not unique identifiers.

[5] Short text to describe, annotate or disambiguate meaning of concept or artefact.

Definition 3 *R is a function (partial and many-to-many mapping) $\{R: C \rightarrow 2^C\}$ over the set of* flexonomy concepts. *R returns $C_R \subseteq C$ as a set of* **connected concepts** *and, therefore, more semantically enriched concepts. R denotes that not every concept needs to be connected or to be well identified in order to be connected.*

More specifically than R, N is a function (partial and many-to-many mapping) $\{N: C_I \rightarrow 2^{C_I}\}$ over the set of well identified concepts. *N returns $C_N \subseteq C$ as a set of* **well connected concepts**, *and, therefore, more semantically enriched concepts.*

The previous definitions implicitly indicate the dynamic and evolutionary aspects of semantic tags and annotations towards more semantically enhanced concepts, since they can be members of any of the previously defined subsets of C at some time point. In the following, we extend the definition of a *flexonomy* through the assignment of a concept to a particular context. For the sake of simplicity, we will refer to three dimensions $\{A, S, T\}$: A the set of tagging agents, e.g., users and user communities, S the set of sources to which a tag has been assigned, T the set of time points. These three dimensions are supposed to refer to *concept provenance* as a concept evolves throughout taggers, bibliographic references and time.

Generally speaking, we define the context as modelled by the n-dimensional space R^n with $\{D_i, i = 1, ...n\}$ the set of dimensions and $V \subseteq \{D_1 \dots D_n\}$ a vector space defined within R^n, since not all dimensions in R^n can be scaled, e.g., agents(users), sources, etc. D_i is considered as the power set (groupings are possible) of discrete points indicating the concept provenance with respect to parameters such as origin, timeliness, reference to source, etc., including the empty set $\emptyset$. This denotes that a particular concept might have been assigned a particular dimension, however, it might not have been assigned any discrete point on that dimension, e.g., to some user or source, which is unknown.

Definition 4 *A flexonomy C with respect to its context is defined as a collection of contextualised concepts $\{\{C_{v_1}\}, \{C_{v_2}\}, ..., \{C_{v_k}\}\}$, $v_i \in M^n = \{(P_1(D_1) \times ... \times P_n(D_m)) \cup \emptyset\}$, $i = \{1, ..., k\}$, $M^n \subseteq R^n$, $m \leq n$.*

Definition 4 denotes flexibility not only in terms of definitions of concepts in C but also in terms of their mappings to a context. In other words, a concept does not need to be assigned all dimensions of R^n and can be assigned to zero or any dimensional points including groupings, e.g., groups of users. Assigning the empty set $\emptyset$, i.e., the 0-dimensional (vector) space to a concept denotes its dimensionless validity and independence of any context.

3.2 Basic Algebraic Operators

In the following, the algebraic operators *C-Selection, C-Projection, C-Union, C-Difference, C-Intersection* are defined in order to lay the foundations for an algebra and other compound operators to access and manipulate a *flexonomy*. All operators of the algebra are subject to constraints as posed on both values of concepts in the flexonomy and on R^n or any subspace denoted by M^n. Constraints are expressed as compound predicates, which are connected via the logic operators AND, OR, NOT. For the sake

of understanding of the following definitions, we need to recall that according to definitions in 3.1, similar labels or annotations of semantic tags are allowed, however, only over different contexts. Moreover, it is worth noting that, in contrast with classical set theory, equality or difference among values alone does not suffice as a criterion for these operators. The subspace M^n representing a particular context is always taken into consideration.

Definition 5 **C-Selection** *operator selects a collection of concepts C_σ within the flexonomy C according to conditions on ontological values of concepts, such as labels, descriptions, relationships, as expressed by a compound predicate P_k, $k \in N$, independent of any dimensions M^n including the zero dimension.*

- **Input:** *The collection of all concepts C in the flexible ontology and condition P_k.*
- **Output:** *A collection of concepts C_σ defined as $\{c_{v_i} \in C_\sigma \mid P_k(c_{v_i})\}$ denoting that they satisfy condition P_k. It holds that M is constructed from C_σ, i.e., $C_\sigma \rightarrow M^n$.*
- **Mathematical Notation:** $\Sigma_{P_k}(C) = C_\sigma$
- **Example:** $\Sigma_{P_1:label=" \star collaborative \star "}(C)$

Definition 6 **C-Projection** *operator returns a collection of concepts C_π within C according to conditions as posed by a predicate $M^n_{P_k}$ in terms of constraints on both dimensions M^n and dimensional points P_k. If M^n is the empty set, then it is the zero dimension that counts meaning that only common agreed concepts are returned. If P_k is left empty, then only concepts, which are not assigned any dimensional points on a particular dimension, are returned.*

- **Input:** *The collection of all concepts C in the flexible ontology and $M^n_{P_k}$*
- **Output:** *A collection of concepts C_π defined as $\{c_{v_i} \in C_\pi \mid M^n_{P_k}(c_{v_i})\}$*
- **Mathematical Notation:** $\Pi_{M^n_{P_k}}(C) = C_\pi$
- **Example:** $\Pi_{User_{(=David)} \wedge Time_{(>11/2005) \wedge \leq (11/2003)}}(C)$

Definition 7 **C-Intersection** *operator returns a collection of concepts C_τ, which are shared between two arbitrary collections of concepts C' and C'' in the flexonomy C and across different contexts M^n_P, with P pointing at the dimensional point(s) on the dimensions M under consideration.*

- **Input:** *Two arbitrary collections C' and C'' as restricted by some subspaces K^n_P and L^n_P, respectively.*
- **Output:** *A collection of concepts C_τ defined as $\{c_{k_i} \in C' \wedge c_{l_i} \in C''\}$, where it holds that c_{k_i} and c_{l_i} have the same unique identifier and $k_i \neq l_i, k_i \in K^n_P, l_i \in L^n_P$.*
- **Mathematical Notation:** $\top(C', C'') = C_\tau$
- **Example:** $\top(C' : ((100, Info)_{David}, (101, Database)_{Chris}, (102, Information)_{David}),$ $C'' : ((102, Information)_{Chris}, (103, Database)_{David}))$ $= C_\tau : (102, Information)_{(David,Chris)}$. $Database_{David}$ *and* $Database_{Chris}$ *do not qualify, since they are not sharing the same meaning, i.e., different unique identifiers 101 and 103, whereas* $Info_{David}$ *and* $Database_{David}$ *do not qualify either, since they are assigned the same context, i.e., user* David.

Definition 8 **C-Difference** *operator returns a collection of concepts C_Δ, which are NOT shared between two arbitrary collections of concepts C' and C'' in the flexonomy C and across different contexts M^n_P, with P pointing at the dimensional point(s) on the dimensions M under consideration.*

- **Input:** *Two arbitrary collections C' and C'' as restricted by some subspaces K_P^n and L_P^n, respectively.*
- **Output:** *A collection of concepts C_Δ defined as $\{c_{k_i} \in C', c_{l_i} \in C''\}$ such that $k_i \neq l_i, k_i \in K_P^n, l_i \in L_P^n$*
- **Mathematical Notation:** $\Delta(C', C'') = C_\Delta$
- **Example:** $\Delta(C' : ((100, Info)_{David}, (101, Database)_{Chris}, (102, Information)_{David}), C'' : ((102, Information)_{Chris}, (103, Database)_{David})) = C_\Delta : ((100, Info)_{David}, (101, Database)_{Chris})$*, where $Information_{David}$ does not qualify, since it is shared by user* Chris*, i.e., same unique identifier 102, whereas $Database_{Chris}$ does qualify, since it is not shared by both users despite similar labels.*

Definition 9 **C-Union** *operator returns a collection of concepts C_Γ, which is the union of two arbitrary collections of concepts C' and C'' in the flexonomy C by also merging or extending their different contexts M_P^n, with P pointing at the dimensional point(s) on the dimensions M under consideration.*

- **Input:** *Two arbitrary collections C' and C'' as restricted by some subspaces K_P^n and L_P^n, respectively.*
- **Output:** *A collection of concepts C_Γ defined as $\{c_{v_i} \in C' \vee c_{v_i} \in C''\}$, such that $v_i = k_i \cup l_i, k_i \in K_P^n, l_i \in L_P^n$.*
- **Mathematical Notation:** $\Gamma(C', C'') = C_\Gamma$
- **Example:** $\Gamma(C' : ((100, Info)_{David}, (101, Database)_{Chris}, (102, Information)_{David}, (103, Database)_{David}), C'' : ((102, Information)_{Chris}, (103, Database)_{David})) = C_\Gamma : ((100, Info)_{David}, (101, Database)_{Chris}, (102, Information)_{(David,Chris)}, (103, Database)_{David})$*, where $Database_{David}$ appears only once in C_Γ, since it holds that $k_i : \{David\} \cup l_i : \{David\} = v_i : \{David\}$*

4 Conclusions and further work

We presented a model and algebra for *flexonomy* as a mathematical and ontological foundation for organising and sharing contextualised and personalised semantic tagging and annotation in digital libraries. The context model is based on a high-dimensional space R^n and on an algebra as an extension of set theoretic operators towards embedding $M^n \subseteq R^n$ subspaces into these operators. More advanced and specific operators can be further defined as compound operators. For instance, the $C - Restriction$ can be defined as a compound operator $\Pi(C_\sigma)$ or $\Sigma(C_\pi)$ for a more selective focussing across values and dimensions of semantic annotations. The algebra is meant to enable the highlighting of differences as well as commonalities of perceptions of artefacts in digital libraries across any arbitrary dimensions, e.g., users, time, sources, by allowing flexible, i.e., incomplete or vague, and user centric semantic annotations. Given also that $V \subseteq R^n$ is a vector space defined over R^n, we are looking forward to defining of operators for semantic distance, similarity and merging. We also plan to specify a query language and implement a prototype for a *flexonomy* as a collaborative environment for semantic tagging of shared bibliographic entries such as those in *http://www.bibsonomy.org*.

References

1. M. Agosti, N. Ferro, and N. Orio. Graph based Automatic Suggestion of Relationships among Images of Illuminated Manuscripts. In *Proceedings of the 2006 ACM Symposium on Applied Computing (SAC)*, pages 1063–1067, 2006.
2. Franz Baader, Diego Calvanese, Deborah McGuinness, Daniele Nardi, and Peter Patel-Schneider, editors. *The Description Logic Handbook: Theory, Implementation and Applications*. Cambridge University Press, 2003.
3. P. Bouquet, F. Giunchiglia, F. van Harmelen, L. Serafini, and H. Stuckenschmidt. C-OWL: Contextualizing Ontologies. In D. Fensel, K. Sycara, and J. Mylopoulos, editors, *Proceedings of the International Semantic Web Conference ISWC 03*, volume 2870 of *Lecture Notes in Computer Science*, pages 164–179. Springer Verlag, 2003.
4. D. Fensel, J. Hendler, H. Lieberman, and W. Wahlster, editors. *Spinning the Semantic Web*. MIT Press, 2003.
5. Peter Gärdenfors. *Conceptual Spaces: The Geometry of Thought*. The MIT Press, Cambridge, Massachusetts, 2000.
6. D. Lenat. *The Dimensions of Context-Space*. Cycorp, 1998. http://www.cyc.com.
7. D. L. McGuiness, R. Fikes, J. Rice, and S. Wilder. An Environment for Merging and Testing Large Ontologies. In *Proc. 17th Intl. Conf. on Principles of Knowledge Representation and Reasoning (KR'2000)*, pages 483–493, Colorado,USA, April 2000.
8. M. Melucci. Context Modelling and Discovery using Vector Space Bases. In *Proceedings of the International Conference on Information and Knowledge Management (CIKM'05)*, pages 808–815, Bremen, Germany, 2005.
9. I. Schmitt and G. Saake. Merging Inheritance Hierarchies for Database Integration. In *Proc. 3rd IFCIS Intl. Conf. on Cooperative Information Systems*, pages 122–131, NY,USA, August 1998.
10. John Sowa, editor. *Principles of Semantic Networks*. Morgan Kaufmann Publishers Inc., 1991.
11. L. Stephens and M. N. Huhns. Concensus Ontologies Reconciling the Semantics of Web Pages and Agents. *IEEE Internet Computing*, pages 92–95, September 2001.
12. C. J. van Rijsbergen. *The Geometry of Information Retrieval*. Cambridge University Press, 2004.
13. Thomas Vander Wal. Folksonomy. `http://vanderwal.net/folksonomy.html`, 2004.
14. R. Wille. Conceptual Graphs and Formal Concept Analysis. In D. Lukose, H. Delugach, M. Keeler, L. Searle, and J. F. Sowa, editors, *Conceptual Structures: Fulfilling Peirce's Dream*, volume 1257 of *Lecture Notes in Artificial Intelligence*, pages 290–303. Springer, Heidelberg, 1997.