ISWC 2013 Doctoral Consortium
‘Ontology Evolution for End-User Communities’

Peter J. Goodall and Peter Eklund

Centre for Digital Ecosystems
University of Wollongong,
peterg@acm.org, peklund@uow.edu.au

Abstract. This project supports application for a *practice-based*[4] PhD by Peter Goodall. The project will produce a system architecture and proof-of-concept laboratory implementation to model, instrument, prototype, and evaluate the effectiveness of several alternate system designs which are intended to enable small end-user communities to evolve specialized ontologies and annotations for entities important to them. Depending on available resources, the laboratory may also be used to study bridging-subset ontologies for interchange, federation and seeding of community ontologies. There will be a discussion of the strengths and weaknesses of various approaches tried by others, and a reflection on the usefulness of the system resulting from innovative work of this project.

Keywords: information systems, ontology, taxonomy, emergent systems, digital ecosystems, curation

1 Problem Description

Each living human community, small or large, cultural, scientific or commercial has its characteristic evolving shared nomenclature which enables discussion and action involving entities and activities important to that community. Further, communities that interact or exchange information with each-other require intersecting terminologies.

Ontology is the concept which encompasses terminology with its meanings and mechanisms. Ontology has philosophical and computational streams of interest to this project: *Philosophical Ontology* has been described as ‘that branch of metaphysics concerned with the nature or essence of being or existence’ [17]. Its modern child - *Computational Ontology* can be described as ‘a conceptualization of some universe of discourse, embodied as a declarative formal abstraction’[14].

This project is motivated by observation of a contradiction between the curatorial and cultural perspectives of an Australian Research Council Linkage Grant (now inactive) which I project-managed - known as ‘The Virtual Museum of the Pacific’[10,9] (no longer active).

The Cultural Collections section of our project partner - The Australian Museum, has over many decades iterated through developing or adopting controlled vocabularies for cataloging their Pacific Collection. They currently use a concise in-house taxonomy for very practical reasons.

While preparing for the launch of the project web-site, we spent time working with representatives of some Pacific communities. They were concerned that the collection taxonomy had little relationship with their own way of describing their artifacts within the collection.

Both the cultural and curatorial ontologies struggled to be effective in the context of the collection. Both the curatorial staff and cultural owners were experts in the conceptualization of their particular perspectives of the collection, and both need aid in ongoing development of terms to suit their both separate and overlapping needs.

My project goal is to research and develop systems for exploring ontologies computationally emergent from technically-informal community annotation.

2 Relevancy

This project has both social and commercial relevancy:

Social - many communities of commitment or interest [21,11] not equipped or resourced to create formal ontologies for curating their physical and electronic artifacts, even though their members are the primary domain experts of those communities.

Commercial - because nearly every modern large-scale electronic market or business, has a shortage of technical experts to formally classify objects and data for cataloging and recommendation.

Both of these application areas could greatly benefit from further development of categorization generated from community or customer based annotation.

3 Related Work

Eleanor Rosch and others, in a number of seminal papers [18,19] developed the notions of Basic Categories and Prototype Theory, prosecuting the idea that there are a number of fundamental levels of category that have a degree of coherence amongst end-users. This experimental work inspires some confidence in community labeling of objects being used to produce derivable category structure. More recently work on the Rational Model of Categorization and the Feature Induction Model examined by Sanborn [20] show promise for inferring categories from user tags.

The work of Cimiano and colleagues [6,7] provides a theoretical and practical resource for generating Formal Concept Lattices for adaption to work done in our *Virtual Museum of the Pacific* project and its Formal Concept Analysis-based navigation and tagging [8].

Mika [15] provides an extended ontology tripartite hypergraph model, splitting the hypergraph into three bipartite graphs associating actors with concepts, concepts with instances and actors with instances. These bipartite graphs are then folded using matrix operations to create various useful affiliation networks, whose properties are used to develop lightweight ontologies demonstrated via a number of case-studies.

A data-model for capture, archiving and processing of user tagging events requires some thought. A user labels an object while engaged in some cognitive context, the person chooses a symbol relevant to that context as a reminder of the concept from that context. Some investigation of processes and schools of Semiotics [5,16] leads to a semiotic triad augmented with a time-stamp and some flexible metadata.

The time-trace of these tags bears an interesting affinity with Dynamic Topic Modeling and Topic Hierarchies [2,3,12] - an approach for incremental discovery of hierarchic categories which appears worth investigating.

As noted in the Problem Description of this overview, development of category systems by domain experts who are not ontologists is discouraged by the complexity of traditional approaches. This complexity is reflected in the software tools readily available. Protege [13] is an example of a quality open-source ontology development system, many others are available in various states of development usage and repair [1] .

4 Research Questions

1. How to model simple hierarchies of community generated categories?
2. What is the relationship between community and classification?
3. Are time-based data-sets available which demonstrate category drift within a community over time?
4. How much information external to the tagging discourse is required for an ontology to emerge?
5. Is emergence as a computational model valid?
6. How to represent collections of tagging events in a way that is simple and effective, yet consistent with *relevant* semiotic theory?
7. How to model requirements for selecting suitable experimental datasets and feeds
8. How to access and transform datasets/feeds for ingestion into the system, and select appropriate tool-sets for this?
9. How to represent and derive ontologies from tagging event streams?
10. Which formal models of tags and ontologies are suitable for incremental and bulk computation?
11. Define a system architecture for processing, interaction, and visualization.

5 Hypotheses

These hypotheses are still very much under development.

1. Basic-level ontologies can be computationally emergent from a suitably annotated tagging event stream. The emergent ontologies will be judged useful by the annotating community.

2. Ontologists or curators working in the domain of a community's emergent ontologies (as derived above) will judge those ontologies to be useful to their work.
3. A proof of concept technical work-bench/tool-chain can be constructed from Free and Open Source Software (FOSS) and reasonably generic cloud computing resources, which can be used to effectively experiment with extraction of latent categories in tagging discourses.
4. Cohesiveness of a subject domain's community of interest corresponds to the coherence of the vocabulary used when tagging objects from that domain.
5. A cohesive community of interest tagging their domain of interest, will produce a stream of tagging events whose latent categorizations will converge into a set of basic and superordinate categories [19], which are recognized as useful by that community.
6. A tagging event data-set with a long-enough time-base will demonstrate temporal drift of its latent categories.
7. Scalable algorithms can be found to derive basic and superordinate categories incrementally from suitable annotation event streams.

6 Approach

1. Literature Survey - Research literature to understand
   - the fundamental relationships between semiotics, classification, community, crowd-sourcing, folksonomy.
   - computational topics - such as Formal Concept Analysis, ontology extraction from text, functional programming, analysis packages such as R, Incanter and Pandas.
2. Characterize and discover suitable data-sets. It is important to find, where possible, suitable existing data-sets for testing hypotheses before committing to the expense and difficulty of user-testing.
   - There are a number of ad-hoc web-sites indicating the existence of possibly suitable data-sets, many of these sites are out-of-date or informally curated.
   - Formal experimental data-set registries are being actively implemented. These also need characterization and selection to find examples suitable for this project.
   - Stream based data-feeds of tagging events are particularly desirable, especially if the same sources have archives of their data-feeds.
3. Identify end-user communities that have an interest in curation of their subject areas. Although very disappointing, strong advice that the difficulty of progressing an ethnographic project through ethics committee process has caused me to design experiments using contemporary western user domains. I am still expecting to use members of some communities of interest - where groups have overlapping domains but divergent interests.
4. Perform preliminary processing of data-sets to refine technical approach and use that to inform further development of my thesis. I expect to iterate through this cycle several times.

5. Once algorithms for deriving categorizations latent in the test data-streams are performing adequately it will be feasible to begin on user-interface development and visualization. I expect to base the display and tagging interface for end-users on the The Virtual Museum of the Pacific, and iteratively evolve it from that base.
6. Once an end-user user-interface is working under informal testing, finalization of user-testing strategy should be developed. Resources such as Amazon Mechanical Turk will probably be used for scalable user-testing.
7. Finalize the system architecture for the experimental work-bench. At this stage of the project I should have sufficient knowledge of the characteristics of data-sets, user-iteration and analysis tools to perform this task with some confidence. Initially I am strongly motivated to use FOSS where practicable, and to implement the system so that it can be scaled by moving from a Linux-based workstation onto a cloud platform.
8. Complete the thesis and package the system for archiving.

7 Reflections

I don't regard my possible success as a sign of failure of others, rather I hope to answer a variant requirement with my own synthesis. Most of the work done with Ontology development has been very formal, which excludes most busy domain-experts from their development.

I hope to produce some interesting results which are a step on the way to generating useful, evolvable user-ontologies, and that the 'workbench' of tools I compose will be useful to others in the field, or a good starting point for further development.

8 Evaluation plan

There are three facets of evaluation for this project:

1. Subject domain usefulness - If a community of end-users tags objects from their domain of interest, they should recognize the extracted categories and rate them as useful
2. Ontologist usefulness - A working ontologist or curator should find the extracted categories useful in themselves, and in a standard format that they can use of in their normal working environment. They should also recognize features in the project's developed workbench that they would like to see as an improvement in their professional working environment.
3. Workbench deployment effectiveness - The running and deployment of the work-bench should be accessible to a researcher in ontology extraction who has a reasonable ability to maintain and configure a Linux workstation.
4. Demonstrate that relationships between annotators can be convincingly deduced by agent affiliation analysis from the tripartite graph of agent, object, and annotation. Agents would need some verifiable, relevant relationship whose description is independent of the tagging dataset.

5. Given a reasonably large tagging event stream from a particular community, determine how well computationally emergent ontologies from that dataset are received by that community.
6. Use temporal sliding windows to generate subsets of a large tagging event stream as input to observe if temporal category drift is observable.
7. Demonstrate computation that generates emergent ontologies incrementally in an environment where the whole data-set cannot be held in system RAM.

References

1. Bergman, M.: The sweet compendium of ontology building tools (Jan 2010), `http://www.mkbergman.com/862/the-sweet-compendium-of-ontology-building-tools/`
2. Blei, D.M., Lafferty, J.D.: Dynamic topic models. In: Proceedings of the 23rd international conference on Machine learning. p. 113120 (2006), `http://dl.acm.org/citation.cfm?id=1143859`
3. Blei, D.M., Griffiths, T.L., Jordan, M.I., Tenenbaum, J.B.: Hierarchical topic models and the nested chinese restaurant process. In: Advances in Neural Information Processing Systems. p. 2003. MIT Press (2004)
4. Candy, L.: Practice based research: A guide. CCS Report: 2006-V1. 0 November p. 19 (2006)
5. Chandler, D.: Semiotics the Basics. Taylor & Francis, Hoboken (2007), `http://public.eblib.com/EBLPublic/PublicView.do?ptiID=308502`
6. Cimiano, P., Staab, S., Tane, J.: Automatic acquisition of taxonomies from text: FCA meets NLP. In: International Workshop & Tutorial on Adaptive Text Extraction and Mining held in conjunction with the 14th European Conference on Machine Learning and the 7th European Conference on Principles and Practice of. p. 10 (2003)
7. Cimiano, P.: Ontology Learning and Population from Text: Algorithms, Evaluation and Applications. Springer-Verlag New York, Inc., Secaucus, NJ, USA (2006)
8. Eklund, P., Goodall, P., Wray, T.: Information retrieval and social tagging for digital libraries using formal concept analysis. In: Computing and Communication Technologies, Research, Innovation, and Vision for the Future (RIVF), 2010 IEEE RIVF International Conference on. p. 16 (2013)
9. Eklund, P., Goodall, P., Wray, T., Bunt, B., Lawson, A., Christidis, L., Daniels, V., Van Ollfen, M.: Designing the digital ecosystem of the virtual museum of the pacific. In: 3rd IEEE International Conference on Digital Ecosystems and Technologies. IEEE Press (2009)
10. Eklund, P., Goodall, P.J., Wray, T., Daniel, V., Van Olffen, M.: Folksonomy with practical taxonomy, a design for social metadata of the virtual museum of the pacific. In: Proceedings of the 6th International Conference on Information Technology and Applications. p. 112117 (2009)
11. Fischer, G.: Communities of interest: Learning through the interaction of multiple knowledge systems. In: Proceedings of the 24th IRIS Conference. p. 114 (2001)
12. Fu, W.T.: The microstructures of social tagging: a rational model. In: Proceedings of the 2008 ACM conference on Computer supported cooperative work. p. 229238 (2008), `http://dl.acm.org/citation.cfm?id=1460600`

13. Gennari, J.H., Musen, M.A., Fergerson, R.W., Grosso, W.E., Crubzy, M., Eriksson, H., Noy, N.F., Tu, S.W.: The evolution of protg: an environment for knowledge-based systems development. International Journal of Human-computer studies 58(1), 89123 (2003), http://www.sciencedirect.com/science/article/pii/S1071581902001271
14. Gruber, T.R., et al.: A translation approach to portable ontology specifications. Knowledge acquisition 5, 199199 (1993)
15. Mika, P.: Ontologies are us: A unified model of social networks and semantics. Web Semantics: Science, Services and Agents on the World Wide Web 5(1), 5–15 (Mar 2007), http://www.sciencedirect.com/science/article/B758F-4MYF67P-1/2/56984a3ddf4632bb98b722551cdb1151
16. Nth, W.: Handbook of semiotics. Indiana University Press (1995), http://books.google.com/books?hl=en&lr=&id=rHA4KQcPeNgC&oi=fnd&pg=PR9&dq=Handbook+of+Semiotics&ots=ddo1tWkS4f&sig=9NNalr_6O7bREpvCPhwQXB9lBn4
17. Oxford English Dictionary: "ontology, n.". (2004), http://www.oed.com/view/Entry/131551?redirectedFrom=Ontology
18. Rosch, E., Mervis, C., Gray, W., Johnson, D., Boyes-Braem, P.: Basic objects in natural categories. Cognitive psychology 8(3), 382439 (1976)
19. Rosch, E.: Principles of categorization. Concepts: core readings p. 189206 (1999), http://books.google.com/books?hl=en&lr=&id=sj1gczQ-7K8C&oi=fnd&pg=PA189&dq=%0D%0A1%0D%0APrinciples+of+Categorization%0D%0AEleanor+Rosch,+1978+&ots=NoqbGizR2u&sig=0WEAZVboo9yC0P8tnryxMC7DT3M
20. Sanborn, A.N., Griffiths, T.L., Navarro, D.J.: A more rational model of categorization. In: Proceedings of the 28th annual conference of the cognitive science society. p. 726731 (2006), http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.163.3800&rep=rep1&type=pdf
21. Wenger, E.: Communities of Practice: Learning, Meaning, and Identity. Cambridge University Press (Sep 1999)