Hugo Liu & Pattie Maes (2007): Introduction to the Semantics of People & Culture (Editorial Preface), *International Journal on Semantic Web and Information Systems, Special Issue on Semantics of People and Culture* (Eds. H. Liu & P. Maes), 3(1), Hersey,PA: Idea Publishing Group.

Introduction to the Semantics of People & Culture

Hugo Liu, The Media Laboratory, MIT, USA
Pattie Maes, The Media Laboratory, MIT, USA

Welcome reader, we are excited to present to you four articles themed around the Semantics of People and Culture. Now that we've piqued your interest with a provocative special issue title, an explanation of what we mean by it is in order.

VISION

As you'll well appreciate, the original Semantic Web advocacy article put forth by Tim Berners-Lee *et al.* (2001) proved catalyst for a revolution of sorts in information systems science. By deploying age-old Artificial Intelligence tactics for representation and meta-reasoning to the World Wide Web, their manifesto envisioned a network of service-oriented entities that could access, negotiate, and interoperate all of the Web's treasures as elegantly as human agents can. And their vision was well justified, because the Web circa 2001 was still, by and large, dominated by unstructured free texts, the information potential of which could nary be unlocked by computers.

Six years later, we find ourselves with a very different Web. With delight and surprise, we find that web users and formal semantic services have successfully married. Blogs and news articles are mirrored by their RSS feeds. Online social network participants express themselves and their circle of friends, but their profiles are also semi-structured and their network of friends imply formal link structure that any scientist would eagerly compute. People rate consumer products, websites, and other people, jotting down some notes about each for their own edification; and automatically, their notes and ratings are correlated with the judgments of their peers into an impressive landscape of keyword tags and scores.

The integration of persons with web services has thus far been surprisingly seamless. And this second lowercase s, lowercase w, "semantic_web" revolution, unlike its predecessor, has evolved organically. When we compare the Semantic Web vision to the semantic_web reality, one might learn these key lessons— 1) be mindful to not over-architect information systems, because unlike a Field of Dreams, if *you build it, no one may come;* 2) everyday people won't formalize their content for fun, but they can be coaxed to annotate content with simple words and scores if they get immediate and tangible value from the process, such as making their content more searchable; 3) the implicit-but-shallow semantics generated by tens of millions of web users can be more powerful than the formal-and-deep semantics of a network of faceless web services, if only we focus on a layer of sense-making technology to mine and organize these implicit semantics, *cf.* (Sheth *et al.* 2005). In short, the surprise of the semantic_web is that, without much formalization, people can already generate computable and interoperable semantics themselves, thank-you-very-much.

Now, the punch line. Why is this issue focused on the semantics of people and culture? Because people and cultures (context communities) already have implicit semantic machinery that guide how they conceptualize, organize, and communicate. These machineries and their internal representations are not "ambiguous" (a scare word often brandished by formalists), but rather, they are contextually sophisticated, and framed by our native meta-languages like psychology and

perception. Nouns and verb phrases can express ideas and relationships more sophisticated than any RDF or OWL description ever could. If the Web 2.0 is any indication, people *are* the new machines and web services for the foreseeable future. But unlike an ordinary computer that is designed top-down, if we are to make full use of the information content produced by people, we have to discover the semantic specifications of people from the bottom up, through reverse engineering, and inevitably the organic process that is trial and error.

The Semantics of People and Culture, then, is what we should at this stage in the web's evolution, aim to research and uncover, be it from the point of view of a psychologist, a sociologist, an artificial intelligence researcher, an information systems designer, or a web scientist, who inherits skills from all of the above. The goal for the Semantics of People and Culture research effort is to uncover the latent semantic boundaries that frame human and cultural communication, so that we can design social media systems to better suit people's natural expressive tendencies. These advances in *semantical ergonomics* should in turn enable future iterations of the semantic_web to capture an ever-growing share of people's natural gift for semantic production.

POLITICS OF TAGGING

In February 2006, the Call for Papers (CFP) for this special issue was distributed. By October 2006, we had received a very favorable response to the Call with many more article proposals and submissions than we could hope to accommodate. Thanks to the support of the IJSWIS Editors, and the hard work of our theme issue's distinguished Program Committee through two rounds of peer review, we are able to deliver into your hands now, four high quality articles, which taken together, illuminate three major topics of interest within the Semantics of People and Culture research area.

In this section and the two to follow, we will introduce the four articles of the special issue to the reader, placing each in the context of the major topic of interest that it represents. The first topic of interest that we should like to suggest is the politics of tagging. As computer scientists are seduced by the power of all that is generic, it is tempting then to view community-produced annotations, or *tags,* simply as strings. But in actuality, web 2.0 users are getting much more information mileage out of tags than might be expected. Tags are not just words, they_are_also_concepts. Tags are enriched by, and can also play against, the social context that underlies them. Tags and bags-of-tags found in the wild on sites like del.icio.us and flickr can improvise syntactic behaviors like negation and coordination, or improvise opinions by collocating keywords with subjective language, e.g. 'tv', 'ads', 'funny'. Collectively, we can attribute these off-label uses of tags and tagging to what we term the *politics of tagging.* After all, tagging is an activity engaged in by humans maneuvering for self-amusement and community status in online social arenas. Of course it is political—politics being the communicative strategies devised to achieve social ends.

By recognizing that tagging is a political activity, information designers can begin to re-view the off-label usage of today's tagging systems as unimplemented features, rather than as bugs. In the "Ontology of Folksonomy: A Mash-up of Apples and Oranges," an article that we invited for this special issue, Tom Gruber (2007) of tomgruber.org and RealTravel.com narrates the short history of tagging with clarity, and also lays out a vision for the next generation of tagging systems. Gruber suggests that, not only has the purpose of ontology been misunderstood, but also, social media and its freeform sea of tags still stand to benefit from the structuration that ontology can provide. Inspired by the things that people are wanting to do with tags, such as collaborative filtering, voting out tag spammers, and making assertions, Gruber hammers out a wish list of design considerations for a next-generation tagging system that is being developed, called TagOntology. By Gruber's prognostications, the future of social media may yet lie in the ontology.

While an understanding of tag politics can further the design goal of more ergonomic semantic elicitation and knowledge sharing systems, another pressing research question is whether or not tags produced by annotation communities are any good, and more provocatively, are human-produced tags better than machine-extracted keywords? The latter question is the topic of Hend Al-Khalifa and Hugh Davis (2007) of the University of Southampton UK's article, entitled, "Exploring the Value of Folksonomies for Creating Semantic Metadata," which appears in this issue.

Al-Khalifa and Davis narrate a series of experiments that measured the relative semantic utility of del.icio.us tags and automatically extracted keywords against the gold standard of quality set by professional indexers. According to their findings, human evaluators found community-based tags to be more relevant descriptors of websites than were keywords derived using term extraction. Further decomposing the del.icio.us and extracted keywords into equivalence classes, the authors made the fruitful observation that del.icio.us tags were particularly concentrated on 'broader-terms' and (non-obviously) 'related-terms' than on 'same-terms' or 'synonyms'.

In short, perhaps community-produced tags are of higher semantic quality than machine-generated tags because 1) they do not dwell on what is obvious (i.e. 'same-terms', 'synonyms') and 2) they express concepts that are definitively related to a website, but through non-obvious relation types (i.e. 'related-terms'). We endorse further experiments into the semantic quality and genre of tags along the lines of Al-Khalifa and Davis' efforts.

SUBJECTIVITY

It would be challenging enough just to uncover the hidden preferences of people *qua* producers of semantic content, but to further complicate matters, different persons are governed by different semantic tendencies. Must we uncover and model them all? There are so many macroscopic tendencies that color or influence the tags we produce and the ratings we assess, such as tastes, personality, culture, ethics, and so on. In the face of so many subjective barriers that divide us, it is remarkable that we are also capable of tuning our expressions, e.g. tags, such that they could be intelligible by the community that we belong to, or even to that gold standard of human interoperability—objectivity.

While it is possible, even easy, for a hundred people to produce similar lists of relevant tags for a website such as nytimes.com, there are domains in which community annotations will not converge as quickly if at all—these are the subjective or political domains. The community forums in political websites, online dating websites, stock picking websites, and consumer product opinion sites are ready examples of argumentative domains.

If we are to make the fullest use of community annotation and knowledge sharing in these subjective situations, it's sometimes necessary to focus on local rather than global knowledge. Social information filtering (Shardanand & Maes 1995), or collaborative filtering as it is also known, is a proven technique for extracting knowledge from subjective domains. Consider a subjective domain such as film recommendation. Each user has annotated films on a one-to-five star rating scale. Since film preference is subject to personal tastes, it is right to be skeptical of global ratings on films. But rather, using collaborative filtering, we could identify a subset of the users whose tastes are closest to our own (i.e. k nearest neighbors), and rely just on the local rating calculated from those users. Hence, knowledge sharing becomes personalized. Numerous annotation communities have already deployed subjective metrics; for example, the DVD website Netflix.com displays not only a global one-to-five star rating for each film, but also a local rating based on the "average of raters like you."

While collaborative filtering creates a 'virtual locality', some subjective domains such as e-commerce merchant networks, are more sensitive to trust, and thus would prefer locality with more substantive basis. Paolo Massa and Paolo Avesani (2007) of

ITC-iRST, Italy, explore the adaptation of trust metrics to domains in which there are controversial users. In this issue's article entitled "Trust Metrics on Controversial Users: Balancing between Tyranny of the Majority and Echo Chambers," Massa and Avesani study the implications of controversial users in the Epinions.com product ratings community.

According to their account, "a controversial user is a user that is judged by other users in very diverse ways, for example, she is trusted or appreciated by many and is distrusted or negatively rated by many." Through computational experiments on the Epinions.com dataset, Massa and Avesani weigh the tradeoffs of employing global versus local (friend-of-a-friend) trust metrics. They advocate for a balance between overly-global metrics that discount valuable minority voices, and overly-local metrics which only tell a user what he or she wants to hear.

CULTURAL INHERITANCE

The quality and legibility of community-produced semantics are of varying degrees, and are greatly impacted by aforementioned factors such as tastes, personality, culture, etc. Subjective filtering techniques allow us the knowledge of those whom we already share a common background with; however, they do not allow us to access knowledge that is contextualized by a different culture or personality. A more radical approach for the investigation of the Semantics of People and Culture is to model the implicit semantics of cultures themselves, and then to use each "cultural module" to better interpret and qualify the knowledge produced by participants of that culture.

We might logically term this approach, 'cultural inheritance'. It is based on a conception of human semantics grounded in archetype/prototype theory. Consider that a person's semantics were represented by a programmatic object. The cultural context underlying the person, then, might be identified as the parent class from which the programmatic object subtypes. Each culture supplements its participants with defeasible knowledge and opinions, which nonetheless can be overwritten. We could envisage a multiple inheritance scheme in which a person actually subtypes many overlapping cultural modules—where 'culture' could be personality, religion, gender, personality, or any community prototype.

Possessing the cultural modules which underlie a person's tags or ratings could aid in interpreting these and other semantic productions. One trivial way to make use of cultural background is to renormalize our interpretation of a tag or numerical rating with the derived baseline. A more challenging use would be to translate knowledge—such as assertions and tag clouds—produced under the premise of one culture into another culture. That is the scenario which motivates the fourth and last paper in this special issue, entitled "GlobalMind: Automated Analysis of Cultural Contexts with Multicultural Common-sense Computing," by Hyemin Chung *et al.* (2007) of MIT. The authors report the engineering of GlobalMind—a website and database that elicits cultural common sense knowledge from users who speak different languages and hail from different countries.

Research on the representation and acquisition of *common sense* is one of the oldest topics of interest in Artificial Intelligence. Simply put, "common sense knowledge" is the collection of background world knowledge that a people share—all humans share a certain set of knowledge, as do members of a particular culture, age group, or community. These tiered and overlapping pools of knowledge can be thought of as coarse implementations of the cultural modules we speak of. One of the prevailing challenges of common sense is acquisition—when we communicate with other cultural participants, common sense facts are omitted because it is implied that both speaker and receiver already possess this information. Recent efforts in Artificial Intelligence have produced significant databases of common-sense knowledge on the order of almost a million facts, acquired either through manual programmer entry or through public contributors.

Chung *et al.*'s GlobalMind cultural knowledge database builds on one of the largest machine-useable common-sense databases, ConceptNet (Liu & Singh 2004), which represents knowledge in a semi-formal manner, as a semantic network of concept nodes interlinked by twenty different semantic relations (e.g. EffectOf("eat burger","feel full")). Practical and approximate inferences can be made by spreading activation across the network. GlobalMind takes the view that ConceptNet's knowledge is actually local to Anglo-American culture since that is the constitution of its contributors. GlobalMind then extends elicitation of cultural common sense to other languages and cultures, such as Korean, Japanese, Chinese, Spanish and Finnish. These various semantic networks are partially aligned with one another by means of bicultural mappings, thus affording promising methods for cultural reasoning such as difference finding and cross-cultural translation of knowledge.

CONCLUSION

As we focus on the future of the human-grown semantic web, it would be wise to remember the lessons of our recent past. Even when mountains of technical standards were complete, web 2.0 and social media did not really take off until the activities of semantic production—such as tagging, rating, and associating—became easy, transparent, and rewarding enough to sustain organic growth of participation. Looking forward, the next advancements upon today's tagging and rating systems would be wise to put semantical ergonomics first—intentionally designing systems to leverage the implicit semantic machinery that guides how people naturally want to conceptualize, organize, and communicate.

We envision the Semantics of People and Culture as the research agenda that will support new innovations in the human-grown semantic web. We have presented a roadmap of three promising areas of research—politics of tagging, subjectivity, and cultural inheritance.

Tagging and rating communities are political—people naturally want to use tags and ratings for certain purposes and in idiosyncratic ways. Then we should systematically measure the semantic value unique to community-produced annotations (Al-Khalifa & Davis 2007), as well as base future designs of tagging and rating systems upon ethnographies of real-world usage (Gruber 2007).

As tags and ratings are contributed by people, they will not always be objective. Subjectivizing factors such as tastes, personality, and culture can and do color human semantic productions. In many subjective domains, such as film recommendation, dating, or consumer products, 'truth' can sometimes be quite local to particular subjective segments. To make the full use of subjective semantics, we should develop techniques to support the localization of meaning, for example, by social information filtering, or through local trust metrics and the detection of controversy (Massa & Avesani 2007).

Ultimately, it could make sense to model the systematic subjective factors directly. Tastes, personalities, ethnicities, genders, subcultures, beliefs, etc. do have some systematic component. We can conceive of any individual's subjectivity as based in the multiple inheritance of various cultural prototypes, just as a programmatic object can inherit default properties and behaviors from various base classes. By capturing the shared common sense and sensibilities that define each subjective factor within 'cultural modules', it could afford improvisational manipulations of our tags and ratings corpora—such as normalizing away subjectivity, or translating tags, assertions, and ratings from one cultural context to another. Cultural modeling may sound exotic, but really, they aren't so far off. Engineering efforts to carefully capture knowledge that explicates culture (Chung *et al.* 2007), tastes, sentiment, and the like are gaining momentum and traction.

The human-grown semantic web has already proved its great potential. By researching semantic technologies to exploit real-world system usage, to cope with subjectivity, and to enhance interpretation using cultural context, we can create smarter

systems to harness and enhance humans' intrinsic semantic productivity.

REFERENCES

Al-Khalifa, H., & Davis, H. (2007). Exploring the Value of Folksonomies for Creating Semantic Metadata. *This Issue.*

Berners-Lee, T., Hendler, J., & Lassila, O. (2001). The Semantic Web, in *Scientific American*, May 2001.

Chung, H., Lieberman, H., & Bender, W. (2007). GlobalMind: Automated Analysis of Cultural Contexts with Multicultural Common-sense Computing. *This Issue.*

Gruber, T. (2007). Ontology of Folksonomy: A Mash-up of Apples and Oranges. *This Issue.*

Liu, H., & Singh, P. (2004). ConceptNet: A Practical Commonsense Reasoning Toolkit. *BT Technology Journal, 22(4),* 211-226.

Massa, P., & Avesani, P. (2007). Trust Metrics on Controversial Users: Balancing between Tyranny of the Majority and Echo Chambers. *This Issue.*

Shardanand, U. & Maes, P. (1995). Social Information Filtering: Algorithms for Automating 'Word of Mouth'. *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems* (pp. 210-217).

Sheth, A., Ramakrishnan, C., & Thomas, C. (2005). Semantics for the Semantic Web: The Implicit, the Formal and the Powerful. *International Journal on Semantic Web and Information Systems, 1(1),* 1-18.