Web Semantics: Science, Services and Agents on the World Wide Web (2007), doi:10.1016/j.websem.2007.11.011

Collective Knowledge Systems: Where the Social Web meets the Semantic Web*

Tom Gruber
TomGruber.org

Summary
What can happen if we combine the best ideas from the Social Web and Semantic Web? The Social Web is an ecosystem of participation, where value is created by the aggregation of many individual user contributions. The Semantic Web is an ecosystem of data, where value is created by the integration of structured data from many sources. What applications can best synthesize the strengths of these two approaches, to create a new level of value that is both rich with human participation and powered by well-structured information? This paper proposes a class of applications called *collective knowledge systems*, which unlock the "collective intelligence" of the Social Web with knowledge representation and reasoning techniques of the Semantic Web.

The Vision of Collective Intelligence

The Social Web is represented by a class of web sites and applications in which user participation is the primary driver of value. The architecture of such systems is well described by Tim O'Reilly [30] , who has fostered a community and media phenomenon around the banner of *Web 2.0* [48] . Headliners for the festival include Wikipedia, MySpace, YouTube, Flickr, Del.icio.us, Facebook, and Technorati. Discussions of the Social Web often use the phrase "collective intelligence" or "wisdom of crowds" to refer to the value created by the collective contributions of all these people writing articles for Wikipedia, sharing tagged photos on Flickr, sharing bookmarks on del.icio.us, or streaming their personal blogs into the open seas of the blogosphere. The excitement is understandable. The potential for knowledge sharing today is unmatched in history. Never before have so many creative and knowledgeable people been connected by such an efficient, universal network. The costs of gathering and computing over their contributions have come down to the point where new companies with very modest budgets provide innovative new services to millions of on-line participants. The result today is incredible breadth of information and diversity of perspective, and a culture of mass participation that sustains a fountain of publicly available content.

Collective intelligence is a grand vision, one to which I subscribe. However, I would call the current state of the Social Web something else: *collected intelligence*. That is, the value of these user contributions is in their being collected together and aggregated into community- or domain-specific sites: Flickr for photos, YouTube for videos, etc. I think it premature to apply term

[*] This paper is based on the author's keynote presentation given at the 5th International Semantic Web Conference, November 7, 2006, which was a call for unifying the Social and Semantic Webs. Since then, several research projects and workshops have published reports of research in this area (e.g., [18]). Please consider this paper as a conceptual framework with which to characterize such research, rather than a claim to specific results. See also related vision papers on the "Social Semantic Desktop" [10] and "Semantic Web 2.0" [15] .

collective intelligence to these systems because there is no emergence of truly new levels of understanding. From the Social Web collective we can learn which terms are popular for tagging photos or the buzz in the latest blog posts, and we can discover the latest new talent in video, photography, or op-ed. However, while popularity is one measure of quality, it is not a measure of veracity. Mass authoring is not the same thing as mass authority. Particularly in the presence of spam and other fraudulent sources in the mix, simply collecting the contributions of the masses does not lead to new levels of intelligence.[†]

Collective intelligence has been the goal of visionaries throughout the history of the Internet. Douglas Engelbart, who invented groupware, the mouse, and a form of hypertext designed for collective knowledge, wrote in 1963 of his career and project objective: "The grand challenge is to boost the collective IQ of organizations and of society" [12] . His Bootstrap Principle was about a *human-machine system* for simultaneously harvesting the collected knowledge for learning and evolving our technology for collective learning. In human-machine systems, both the human and machine contribute actively to the resulting intelligence, each doing what they do best. Other early pioneers of the human-machine model of collective intelligence include Norbert Wiener, the father of cybernetics, Buckminster Fuller, the consummate inventor and system thinker [13] , and Stewart Brand, creator of the first large virtual community on the Internet [43] . Tim Berners-Lee, the inventor of the World Wide Web, describes his vision of the Semantic Web in these terms: "The Semantic Web is not a separate Web but an extension of the current one, in which information is given well-defined meaning, better *enabling computers and people to work in cooperation*." [emphasis added] [3] .

I would suggest that collective intelligence be taken seriously as a scientific and societal goal, and that the Internet is our best shot at seeing it happen in our lifetimes. The key, as the visionaries have seen, is a synergy between human and machines. What kind of synergy? Clearly, there are different roles for people and machines. People are the producers and customers: they are the source of knowledge, and they have real world problems and interests. Machines are the enablers: they store and remember data, search and combine data, and draw mathematical and logical inferences. People learn by communicating with each other, and often create new knowledge in the context of conversation. The Internet makes it possible for machines to help people create more knowledge and learn from each other more effectively.

Some progress has been made in allowing machines to learn from people and data. Artificial Intelligence technology allows people to build "expert systems" that act competently as individual

[†] Intelligence is a larger topic than this paper can discuss, but it is reasonable to say that if the word has any meaning it includes at least some notion of competence. A system that shows collective intelligence, then, must be at least as competent as the individuals that comprise it. Jaron Lanier warns that to assume otherwise is to confuse collectivism with intelligence, with dangerous consequences [22] . In contrast, the markets described by Surowiecki in *The Wisdom of Crowds* [40] are, if measured by their competence, intelligent; under the right conditions, a market can outperform very intelligent individuals. However, markets with the requisite properties are not simple collections. They are orchestrated by formal rules and structured channels of communication and action.

experts, by embodying their problem solving knowledge in models and data. However, the knowledge acquisition bottleneck has limited the reach of these systems, because it takes a lot of work to get the knowledge into a form that machines can use to solve problems. Machine learning and text mining techniques can find structures and patterns in large data sets, and thereby help us make better use of our collected data. However, these techniques depend on their data for the power to reveal new insights. There is a symbiosis between data and machine. For example, the unprecedented acceleration of scientific progress in human genomics research is the result of *both* an enormous data gathering effort -- most notably the Human Genome Project -- and advances in computational biology.

With the rise of the Social Web, we now have millions of humans offering their knowledge online, which means that the information is stored, searchable, and easily shared. The challenge for the next generation of the Social and Semantic Webs is to find the right match between what is put online and methods for doing useful reasoning with the data. True collective intelligence can emerge if the data collected from all those people is aggregated and recombined to create new knowledge and new ways of learning that individual humans cannot do by themselves.

Collective Knowledge Systems

How can we characterize a system that can deliver on the opportunity of collective intelligence? Consider the class of *collective knowledge systems:* human-computer systems in which machines enable the collection and harvesting of large amounts of human-generated knowledge. A simple example is something I call the "FAQ-o-Sphere". As shown in Figure 1, it has three parts:

1. A social system, supported by computing and communication technology, which generates self-service problem solving discussions on the Internet. These are the product support forums, special interest mailing lists, and structured question-answer dialogs in which people pose problems and others reply with answers on the Internet.

2. A search engine that is good at finding questions and answers in this body of content. Google, for example, is very good at finding a message in a public forum in which someone has asked a question similar to one's query.[‡]

3. Intelligent users, who know how to formulate their problems in queries that the search engine can match to online question/answer pairs. In addition, users help the system learn when they provide intelligent feedback about which query/document pairs were effective at addressing their problems.

The FAQ-o-Sphere was not designed as a system, but as it exists today the FAQ-o-Sphere is an amazingly competent expert system. Type into Google the error message from a piece of buggy software, or the symptoms of a product malfunction, or even a common usability problem, and the

[‡] To be more precise, Google appears (to me, in 2007) to have reasonable *recall* (not necessarily good *precision*) when there is such a question/answer thread in a forum, and the question or answer message text matches the query text superficially (not at a conceptual level). The point is that, as part of the knowledge *system*, this level of matching works surprising well for many domains.

system will return surprisingly useful and insightful answers. The machine captures the collected experience of product users as they participate in a social process and the solutions created by helpful volunteers. The system enables human learning from peers at a scale that would not be possible without the machine components of the system.

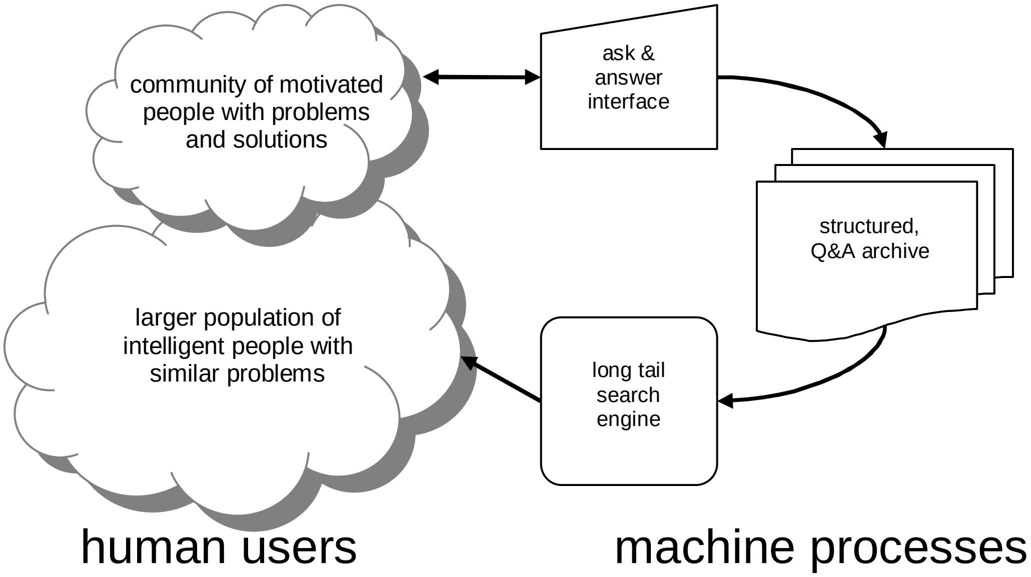
community of motivated people with problems and solutions
ask & answer interface
structured, Q&A archive
larger population of intelligent people with similar problems
long tail search engine
human users
machine processes

Figure 1. The FAQ-o-Sphere, an example of a collective knowledge system. It consists of (1) a community of contributors, participating in a social process (reward structure) augmented by computer mediated communication (dialog structure) and long-term memory (archive of conversation), (2) a search engine for retrieving information, and (3) intelligent users who actively query the system using strategies tuned to the content generation process and the search engine.

Other examples of collective knowledge systems include

- **Citizen Journalism** ([14]) - want to get the scoop on a story before the mainstream press, or get a diversity of opinions on a story? Search the blogosphere.
- **Product reviews** for consumer products. Want to buy a gadget, a piece of computer equipment, or digital camera? The best information is in user reviews, not the marketing literature.
- **Collaborative filtering** to recommend books and music. Like to read more books that are like a favorite book? Check Amazon's recommendations based on other customers' choices. Like a movie and want to see more like it? Get recommendations from Netflicks.

What do all these systems have in common that makes them so great? The key properties of collected knowledge systems are

- **User-Generated Content.** The bulk of the information is provided by humans participating in a social process. A traditional database or expert system, in contrast, gets the bulk of its information from a systematic data gathering or knowledge modeling process.
- **Human-machine synergy.** The combination of human and machine provides a capacity to provide useful information that could not be obtained otherwise. These systems provide

more domain coverage, diversity of perspective, and sheer volume of information than could be achieved by searching "official" literature or talking to experts.

- **Increasing returns with scale.** As more people contribute, the system gets more useful. The system of rewards that attracts contributors and the computation over their contributions is stable as the volume increases. In contrast, a text corpus and simple keyword search engine does not get more useful when the volume of content overwhelms the value of keywords to discriminate among documents. Similarly, if the reward system encourages fraud or fails to "bubble up" the best quality content, the system will get less useful as it grows.

If we wish to move from collected intelligence to collective intelligence, we would add a fourth property:

- **Emergent Knowledge.** The system enables computation and inference over the collected information, leading to answers, discoveries, or other results that are not found in the human contributions.

Emergent knowledge is not magic. Normal science works this way: scientists read the literature and talk with colleagues, synthesize new ideas, and "bubble up" the best work through the peer review processes. When a scientific discovery is made, the answer is not found by retrieving the right paper. In other words, science is not a collection of knowledge; it is a system for creating it. Technology can augment the discovery and creation of knowledge. For instance, some drug discovery approaches embody a system for learning from models and data that are extracted from published papers and associated datasets. By assembling large databases of known entities relevant to human biology, researchers can run computations that generate and test hypotheses about possible new therapeutic agents. For the vision of collective intelligence from the web, we need analogous technology for assembling large sources of human generated content in such a way that computations can discover and conclude new things. Enter the Semantic Web.

The Role of the Semantic Web

Technology has enabled the generation of collected knowledge systems by making it cheap and easy to:

- **Capture** - Cheap sensors, microprocessors, memory, fiber networks, and cellular telephony has meant that a lot of people have computers, smart mobile phones, digital cameras, and broadband. These things enable people to upload their digital lives and spend more time online.
- **Store** - Cheap disk storage, both in the home and on giant server farms, is removing the barriers for people to share a lot of information.
- **Distribute** - The Internet is an information superconductor connecting the planet.
- **Communicate** - asynchronous collaboration systems (email, wikis, blogs) overcome barriers of space and time and the number of parties in a conversation. We now can talk to anyone and learn from others' conversations without being there.

There is a fifth role for technology, one which is just beginning to show its potential to create *collective* intelligence:

- **Create new value from the collected data**

I believe that creating value from data is the main role for the Semantic Web in collective knowledge systems. While there are plenty of ways to create value by aggregating user contributions today, there are few that go beyond summarizing or sorting the data. I see two major ways that Semantic Web technology can significantly change the game, moving us closer to emergent knowledge. First, one can add value to user data by adding structured data. That is, Semantic Web technologies can add structured data related to the content of the user contributions in a form that enables more powerful computation. Second, the standards and infrastructure of the Semantic Web can enable data sharing and computation *across* independent, heterogeneous social web applications. By combining structured and unstructured data, drawn from many sites across the Internet, Semantic Web technology could provide a substrate for the discovery of new knowledge that is not contained in any one source, and the solution of problems that were not anticipated by the creators of individual web sites.

Let us consider in more detail these ways that Semantic Web techniques can create value for collective knowledge systems.

Augmenting User-Contributed Data with Structured Data

The essential difference between the classic Web and the Semantic Web is that structured data is exposed in a structured way. For example, the classic Web might have a document that mentions a place, "Paris". The conventional way to find this document on the Web is to search for the term "Paris" in a search engine. Similarly, to find out more about the place one would plow through the search results on the term "Paris" and manually pick out the pages that seem to have something to do with the place. The heuristics employed by today's search engines for inferring what one means by the string "Paris" are biased by *popularity*, which means that one will encounter many pages about a celebrity heiress en route to the French capital.

The Semantic Web vision is to point to a representation of the entity, in this case a city, rather than its surface manifestation. Thus to find the city Paris, one would search for things known to be cities for entities whose names match "Paris", possibly limiting the results to cities of a certain size or in a particular country. Then one might look for information of the desired type about the city, such as maps, travel guides, restaurants, or famous people who lived in Paris during some period of history. The heuristics for searching the Semantic Web depend on conventions about how to represent things like cities (such as those specified in ontologies), and the availability of data which use these conventions. Such data is not available for most user contributions in the Social Web.

To move to the next level of collective knowledge systems, it would be nice to get the benefits of structured data from the systems that give rise to the Social Web. There are three basic approaches to this: expose data that is already in the databases used to generate HTML pages, extract the data retrospectively from user contributions, and capture the data as people share their information.

The first approach is to expose the structured data that already underlies the unstructured web pages. An obvious technique is for the site builder, who is generating unstructured web pages from a database, to expose the structured data in those pages using standard formats. For instance, social web sites could expose their links to users as FOAF data, which is a Semantic Web convention for representing personal contact information [27] . This, of course, requires the compliance of the site builder, which means that it will not likely happen without a business motivation that benefits the site.

There are several promising techniques for the second approach, to extract structured data from unstructured user contributions [2] [28] [39] . It is possible to do a reasonable job at identifying people, companies, and other entities with proper names, products, instances of relations you are interested in (e.g., person joining a company) [1] [7] , or instances of questions being asked [24] . There also techniques for pulling out candidates to use as classes and relations, although these are a bit noisier than the directed pattern matching algorithms [8] [23] [31] [32] [36] [38] [42] . In the context of this paper, what is interesting is that these techniques can be used to fold their results back into the data sources. That is, they can be used to *augment* the unstructured user data with structured data representing some of the entities and relationships mentioned in the text. For example, one could couple the structured data extracted from analyzing Wikipedia (such as that done by DBpedia [4]), into tools that allow users to add structured data while they are entering wiki pages (such as Semantic MediaWiki [46]). For instance, if a Wikipedia page mentions a book by its ISBN number, the link under the ISBN number could reference the book in structured databases of books and be used to call APIs for obtaining it ([5]). More sophisticated examples for extracting references to named entities and factual assertions can also be applied. It is important to note that all these techniques require open data access and APIs to have a real impact on the social web.

The third approach is to capture structured data on the way into the system. The straightforward technique is to give users tools for structuring their data, such as ways of adding structured fields and making class hierarchies. This is naïve for the social web, since the users in this space are not there to create databases; they are there to have fun, connect with other people, promote their ideas, and share their experiences. However, there is a way that makes sense for the user, the site builder, and the larger web. It is to provide services that give personal and social value to the individual in return for their using a tool that helps them add structure to their content. We call this technique *snap to grid*, and will illustrate it with an example in a later section. The term "snap to grid" refers to an interaction pattern that many of us use without consciously thinking about it. When you draw a shape on the screen in a drawing program such as PowerPoint, by default the system finds the nearest point in a discrete grid of locations to align your edges. That way, the next time you draw an object on that grid that *almost* lines up with the first, it will *precisely* line up in the data. Similarly, many text processing environments automatically correct your spelling as you type. They are snapping your text to the nearest word in the "grid" provided by the dictionary. Email programs do a similar "completion" on addresses typed into the address field. For the user of the drawing and text processing applications, snap to grid creates a more attractive and useful product; for the email user, the service is indispensable to achieving the core functionality of the product. The result is a mix of structured and unstructured data, which has far more value when aggregated into collections. For example, it is trivial to find all of the messages sent by the same person (or at least from the same address), since the machine validates the address when mail is sent and the namespace is a well-enforced standard.

Snap-to-grid assumes that there is a structure to data (such as constraints on its form or values), and helps users enter data within that structure. It is important to combine a snap-to-grid interface for soliciting structured data with motivations for providing this data. For example, there are tools for adding structure to Wikipedia ([46]), but they depend on voluntary compliance. An interesting approach is to combine data entry with a social system that structures *the behavior*. For example, Luis von Ahn [47] has created games in which people are rewarded for teaching the computer things such as what to label an image. The data structure is fairly simple: an entire image, or a well-defined region of the image, must be mapped to a word. The motivational structure of the game (try to match the label of other players) and the large number of players leads to quality of content.

Enabling Data Sharing and Computation Across Applications

The other major area where Semantic Web can help achieve the vision of collective intelligence is in the area of interoperability. If the world's knowledge is to be found on the Web, then we should be able to use it to answer questions, retrieve facts, solve problems, and explore possibilities. This is qualitatively different than searching for documents and reading them, even though text search engines are getting better at helping people do these things. Many major scientific discoveries and breakthroughs have involved recognizing the connections across domains or integrating insights from several sources. These are not associations of words; they are deep insights that involve the actual subject matter of these domains.

The Semantic Web has the machinery to help address interoperability of data from multiple sources. Ontologies help specify common conceptualizations, independent of data model, so people can align their systems semantically by adopting vocabularies at their external boundaries without committing to a common data format internally [16] . RDF allows the encoding of structured data by reference to well-maintained namespaces. This does two things. It ties the data that is exposed and exchanged to the common vocabularies from the ontologies, which allows one to know *exactly which* class, attribute, or relation is being used in a statement. For example, it allows two or more applications to expose their references to cities as cities -- and not text strings. Secondly, the RDF format allows the entities mentioned to be identified unambiguously within a namespace. For example, if system A refers to its entry for Paris in 100 different pages, we know that it means the same city each time (the 100 references to Paris are explicitly tied to the same instance of Paris in a namespace). It does not mean that it refers to the same city as system B's reference to Paris. But when combined with common agreements about how to refer to a city versus a person, it enables a third, intermediary system to find the connections between the two systems' location references. This is potentially more powerful than matching the texts of documents from the two systems, particularly where precision is important.

In addition to meaningful integration of data from multiple sources, exposing structured data within a common framework of ontologies enables new technologies for distributed queries and data integration. Since each data source on the Semantic Web is potentially different in its internal data model, and there can be many ontologies with different meanings for terms, querying a large set of data sources is much more than issuing the same query to each. While this is a fundamentally difficult problem, the mechanisms of the Semantic Web make it possible to reason about the different ontological assumptions made by each system and adapt accordingly. A simple analogy from computer science illustrates how reasoning about data assumptions is important when operating on data from multiple sources. Let's say that each of two systems represented a table of numbers - one using 2 bytes and the other 8 bytes to store each number. Each number format places an upper bound on precision, by definition. When one does repeated computations over such numbers, the errors of precision are compounded. If a third system wanted to do a cross-application computation, such as combining the data from two scientific studies, then it would need to account for the different precisions when drawing conclusions. In particular, it would need to treat the data as if they all had the lower precision, or risk introducing spurious conclusions. Much more sophisticated analogs of this occur when two systems differ in the assumptions about how they represent entities and their properties, such as the spread of disease though the populations of cities (e.g., the city is a point location in one system and a polygon at some mapping resolution in another). The Social Web also has these problems. Even simple problems like determining whether two users with the same first and last names are the same person across applications involves this kind of analysis.

An example of how the Semantic Web community can help with this problem for the Social Web is the integration of tagging data. Tagging is the labeling of an entity (usually a web page or something with a URI) with words or phrases so one can remember them later and group them with related finds [44] . For example, photos on Flickr are tagged with labels that allow people to find photos with words and collaborate with other people interested in the same words. While many tagging sites make the data available over APIs, the meaning of the tagging data is completely unspecified. For some people, this is the point - let a thousand tags bloom. However, I am not talking about the meaning of the *labels*, which have all the problems of matching unconstrained natural language; the problem is that to interpret the tagging data itself one needs a great deal of knowledge about the conventions and software used at the tagging site. For example, sites differ by how they canonicalize labels (eliminating white space, folding case, stemming, synonym equivalence, etc.), how they identify users (implicitly or explicitly, and if so, with respect to which namespace and under what privacy rules), and how they identify tagged objects (there are multiple URLs one could use to see a picture on Flickr).

The TagCommons.org [41] is an open project applied to the problem of identifying ways to connect the world's tagging data in a *semantically meaningful way*. This involves a process of identifying use cases for sharing tag data, studying existing sources of tag data and applications for using it, defining a common conceptualization of all the distinctions needed to represent the data in these systems, specifying this conceptualization in one or more ontologies, showing how these ontologies map to data formats, and eventually offering a platform for tools that operate on the common understanding. The project's analysis to date has identified a range of use cases and existing tag systems, and is defining a common conceptualization using the collaboration processes of a wiki. For example, discussions have revealed assumptions that are implicit in existing tag data sources, such as the tagger identity (e.g., in blogs), the correlation among tags in time made together, tagging across natural languages, and the tagging of objects that are not web pages (e.g., photo identities assigned by cameras and mobile phones).

The TagCommons project is consciously a *mapping* effort rather than *homogenization* effort, and is focusing on clarifying the conceptual distinctions rather than the low level data models. Although its goal is to create an ontology that covers all the use cases and conceptual distinctions, a main use of this ontology will be to enable interesting mappings across existing systems, including those based on existing ontologies. The overlap and differences among tag data formats ([6] [33]), ontologies for tagging ([17] [26] [29]), general annotation [21] , and conceptual vocabularies [19] [20] [37] are being explored. The common conceptualization and ontology allows the community to give context-free definitions and neutral names for the distinctions that differ across systems, so that meta-tagging systems such as TagWatcher [45] can be designed to anticipate and reason about these differences when presenting a multi-site tag interface to people. In a sense, the TagCommons project is attempting to create a platform for interoperability of social web data on the Semantic Web that is akin to the "mash-up" ecology that is celebrated in Web 2.0.

Example: Collective Knowledge System for Travel

An example of how a system might apply some of these ideas is RealTravel. RealTravel is an example of "Web 2.0 for travel". It attracts travelers to share their experiences: sharing their itineraries, stories, photographs, where they stayed, what they did, and their recommendations for fellow travelers. Writers think of RealTravel as a great platform to share their experiences -- a blog site that caters to this domain. People who are planning travel use the site as a source of information to research their trip, augmenting the published travel guides and marketing literature from travel service providers. As illustrated in Figure 2, it has all of the properties of a collective knowledge system introduced earlier:

- **User generated content** - most of the content is from real people who are reporting on their experiences in the field.   It is a global travel guide written by thousands of authors.

- **Human-machine synergy** - travel planners can do the equivalent of asking many thousands of people their advice when deciding where to go, what to do, and where to stay. This augments what can be found in official travel guides or by consulting travel experts.

- **Increasing returns with scale** - as more people report on their travel experiences, the system can offer better coverage of more exotic locations and better depth on what to do and avoid. In addition, an editorial process and community feedback means that the "best" content for any travel destination in any category is always bubbled up for the reader.

- **Emergent knowledge** - the system offers recommendations for planning a trip that are based on unsupervised learning from the texts of travel blogs and a multidimensional match to structured data such as traveler demographics and declared interest.

To achieve these properties, RealTravel applies the principles discussed earlier in this paper for integrating structured and unstructured data and creating value by computing over user contributed data. In particular, let us look at three illustrations: snap to grid, contextual browsing, and learning from semistructured data.

Snapping to the Grid of Travel Destinations

There are well over 2 million official places in the world that might be called towns, cities, states, countries, or the like. Many have multiple names with multiple spellings, particularly for travelers reporting on trips to foreign lands. There are many entities with the same names (there is a Paris in Texas, one in Ohio, and several in France). Their names are culture and context dependent ("New York" the state versus the city, whose official name is "New York" and not "New York City"). Many travel destinations are not in the official gazetteers of places ("Tuscany" is not a state in Italy and "Lake Tahoe" is not a city in California). Furthermore, destinations live in a lattice hierarchy (San Francisco is in the Bay Area as well as directly in California, which is in the United States) and a geospatial topology (New Jersey is near New York City, but in a different state).

Despite the messiness of destination data, it is the backbone of travel information sites. If you want to offer advice about where to stay in New York city, you need to group together all the information about NYC, and depending on the context (e.g., which airports serve the area) you may want to bring in data from neighboring cities. If you want to show the "best of California" you need to include the content from San Francisco and San Diego. Similarly, if you want to offer a tour of the best user photos from trips to Europe, you need to figure out which photos are from places that are located in Europe, without expecting users to geotag them.

The way RealTravel solves this is to elicit, at the time when users contribute their data, the place they are talking about. The exact UI technique is constantly evolving, but the principle is *snap-to-grid*. For example, when one starts to write about a leg of a trip, the system asks which country they went to (this is easy to get accurately). Then, for a particular blog entry, the system asks for the location within that country ("where in France did you go?"). It uses Web 2.0 UI techniques to rapidly offer a set of completions of the place name from a ranked list of candidates, in real time as they type. If they select a candidate, it confirms with a map showing the location of the place. If their destination is not in the database, the system allows them to create a new destination asking for its approximate location (what is it near), its hierarchical position (what state or country it is in), and its type (city, town, park, etc). Since this is a rare occurrence, and the completion list is biased by user data, most users experience their intended location quickly and move on.

As a result, every piece of user contributed data is precisely located in a structured grid of travel destinations, which can be maintained by a central staff and extended with user input. From this, the following services are enabled:

- Blogs are grouped together by destination and ranked by rating, so the reader can browse "best first" all of the content relevant to any travel destination.

- "Nearby" blogs, hotels, things to do, etc. can be computed. For example, to go to the famous Angkor Wat temples in Cambodia, one typically stays in the nearby city of Siem Reap. The traveler planning a trip to Cambodia may not know this.
- Aggregations such as the "best of California" and "things to do in France" can be computed by joining fact that a blog entry is located in a particular city with data that the city is in a state or country.
- The locations of photos can be inferred from the locations of the trips they are associated with, and also aggregated up the hierarchy.
- From named destinations the system can get geocoordinates, which enable the generation of custom route maps of travelers' journeys.
- Data associated with various levels of destinations can be mapped to external content sources and services that also index by destination (travel guides and targeted advertising)

Contextual Browsing: Combining tags, location, and rating data

In addition to destination data, which is objectively validated, there are subjective dimensions of user-contributed travel information which are defined by structured data. One is rating data, common in social web applications. Editors and users give feedback on the quality of content they read, which is typically a small integer rating that can be used to rank order content. The other is tags. As introduced earlier, tags are labels used to aid in organization and retrieval. The collection of tags for a site is called the folksonomy, which is useful data about collective interests.

RealTravel, like many Web 2.0 sites, combines these structured dimensions to order the unstructured content. For example, one can find all the travel blogs about diving, sorted by rating. In fact, the site combines all of the structured dimensions into a matrix, which offers the user a way to "pivot browse" along any dimension from any point in the matrix. For example, one can find all the blogs about Bali, then switch to the photos of Bali best-first, then look at the blog that contains a particular photo, then look for hotels in nearby destinations. One combination of dimensions is surprisingly useful: combining tags with destinations. For example, one can go to the page for Thailand and then see the most highly rated tags for blogs in Thailand, pick one ("beach") and then see all of the blogs in Thailand tagged with "beach". Since tags are a dynamic classification system, when combined with measure of utility this produces a kind of dynamic index for a virtual travel guide. For example, some tags such as "new year's 2006" were popular in some parts of the world but not used elsewhere. Similarly, blogs tagged with labels such as "natural wonder" were more common in places with natural wonders. This is no surprise, except this is an *emergent* source of insight about where in the world one might go to find natural wonders.

By using a snap-to-grid interface for entering tags, the system can offer easy paths to tags that represent useful domain concepts. For example, entering "art" or "arts" or "art museum" might all snap to the same general domain concept, "traveling to experience art". The dynamics of folksonomy evolution can be guided using these biases.

These structured dimensions to user contributions are also natural compliments to full-text search. For example, one could search for an idiosyncratic interest "Jim Morrison" and then drill down by location to discover that his grave is a tourist destination in Paris. Similarly, one can use tags as "facets" for *faceted search*, in which the user can dynamically add and modify multiple constraints to refine a query. For example, one can search for travel experiences anywhere in Europe that are tagged with "art" and "history" containing the text "cathedral". Since tagging is, in a sense, a user's vote that a label is a good term for retrieving an item, it makes sense to give these terms special treatment in the relevance rank of search results. And if tagging is elicited in a snap-to-grid manner as described above, selecting a tag-as-facet in a faceted search constrains the query with something more like a concept than a word. I consider this sort of semistructured, faceted search as a precursor to the dream of full-blown semantic search on the web.

Learning from Semistructured Data

While tags and the structured data dimensions are useful ways of organizing, browsing, and searching content, for true collective intelligence we would like to actually learn something from the collective that we could not discover by just searching and reading a lot of entries. RealTravel provides a recommendation engine that suggests the kind of computation that is possible when one combines the data from the social web with the computational tools developed for dealing with large, noisy, multidimensional structured data.

If one were to harvest the "wisdom of crowds" from fellow travelers, what would that look like? The first generation of user-contributed travel sites offers collected-intelligence style aggregation. For example, on TripAdvisor, a leading hotel review site, one can easily see the aggregated user rating for hotels in a city, and browse photos of hotels contributed by reviewers. More comprehensive sites such as Yahoo! Travel [49] include sample trip itineraries contributed by experts and members, and tie them nicely to destination data. What if one wants personalized advice on where to go on vacation? This is the sort of service that used to be offered by travel agents, who have been displaced by the online travel agencies. If you knew someone who was like you and traveled a lot, you could ask them. What if you could ask an entire population of people for their advice about where to go, but you need the advice tailored to your interests? That was the goal for the recommendation engine in RealTravel.

To achieve this function, the system processes every user contribution, looking at the text, tags, user profiles including demographics, and all of the other structured data in the system. It then does a clustering of the content to find synthetic dimensions. The exact algorithms are proprietary but the techniques result in the stable classification of documents and users into buckets. When travel researchers request a recommendation, they are asked for general constraints such as the area of the world they are considering, the length of their trip, their age, and family status. These factors are then used to filter the set of dimensions that might be relevant to them, and they are asked to rate their interests on these dimensions such as travel for adventure versus social experience. Based on their answers to these questions, the system then matches against all the known dimensions, including synthetic dimensions, how well their demographics match the blog authors, and the constraints such as destination and time to travel. It then rank-orders a list of recommended places to go, and offers a personalized list of the most relevant blogs to read to help choose a destination.

The results can be surprising and insightful. For example, when I first tried the recommendation engine for my own travel, it asked me questions that were relevant to my demographic profile (such as, whether I like wildlife viewing), and then based on my answers (such as the fact I like wildlife viewing, diving, and nature) recommended that I try the Galapagos Islands (first) and two other places I had already visited. As I browsed the stories and photos from people who have been to the Galapagos, I got a sense of why this would be a great place for me to visit. I had experienced the collective knowledge of thousands of travelers, applied to my personal needs.[§]

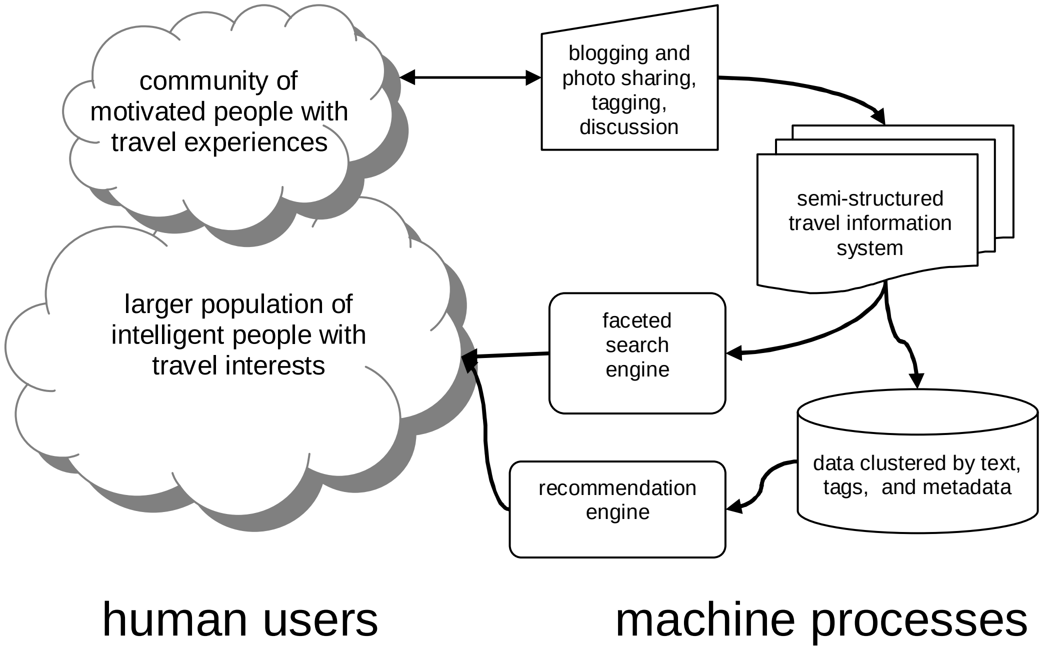
community of motivated people with travel experiences
blogging and photo sharing, tagging, discussion
semi-structured travel information system
larger population of intelligent people with travel interests
faceted search engine
data clustered by text, tags, and metadata
recommendation engine
human users
machine processes

Figure 2. A collective knowledge system for travel. RealTravel can be viewed as a human-computer system that consists of (1) a community of contributors, participating in a social process (sharing experiences) augmented by computer mediated communication (blogging, tagging, commenting) and long-term memory (travel information system, clustered by multidimensional analysis), (2) a faceted search engine and a recommendation engine, and (3) intelligent users who actively search for destination information and answer questions to get personalized travel advice.

Conclusions and Speculations

This paper argues that the Social Web and the Semantic Web should be combined, and that collective knowledge systems are the "killer applications" of this integration. The keys to getting the most from collective knowledge systems, toward true collective intelligence, are tightly integrating user-contributed content and machine-gathered data, and harvesting the knowledge from this combination of unstructured and structured information.

[§] Disclosure: the author was the CTO of RealTravel, and the recommendation engine was built by Sergei Lopatin. This anecdote is not an evaluation, and this paper is not claiming a new technology for machine learning or recommendation engines. It is, however, an example of what is possible by applying serious computing to harvest collective knowledge in a Social Web application.

The RealTravel experience is suggestive of what might be done in a collective knowledge system. It used social processes to attract the content providers, snap-to-grid to elicit structured data from users, pivot browsing and faceted search to aggregate the structured and unstructured data, and clustering technology to induce implicit underlying dimensions. To harvest the value of the collected knowledge, it combined personal user data such as demographic and travel preferences with all of the other explicit and implicit dimensions to power a travel recommendation engine.

Despite the emphasis on a grid of structure, however, the RealTravel application designers found no need for named entity extraction, taxonomic categorization, semantic search or other techniques associated with the Semantic Web. In keeping with the slogan "a little semantics goes a long way", most of the powerful inferences in the travel domain are along the basic dimensions of who, where, when, and why.[**] The commercial nature of the application also enforced a minimalism in architecture; that is, the types of structured data to be gathered and represented are driven by the computations required to help people research their travel. For example, while it is possible to recognize that a mention of "Jim Morrison" is probably referring to a person, it is not necessary to explicitly represent a fact that he is buried in Paris to know that his grave is a tourist attraction (the existence of a travel blog mentioning Jim's name and accurately located in Paris is enough). Nonetheless, the RealTravel application was built without the benefit of standard ontologies, data sources, and web services for personal identity, destinations, scheduling, and tagging. Were they available, these resources may have accelerated the development of functionality and the integration with other travel services. As the infrastructure of the Semantic Web becomes more available to the developers of Social Web applications, the next generation of collective knowledge systems may be designed from scratch to get the best of both worlds.

Structured and unstructured, formal and informal -- these are not new dimensions. They are typically considered poles of a continuum. For example, it is illuminating to plot the range of knowledge representations along a two dimensional space, with this dimension of degree of structure on the x-axis (see Figure 3). Correlated with the degree of structure are the expressive power of the knowledge representation used and the cost to develop and maintain the knowledge base. On the y-axis is the value of the computational service offered, from simple retrieval to sophisticated reasoning. For most of the past 25 years, points in this space have been roughly correlated as well, suggesting there is no free lunch: to get better reasoning, you need to gather better data and represent it in a more complex way. I think the confluence of the social / economic wave of the Social Web with the maturing of the technology of the Semantic Web may present us with a turning point. Maybe we can get a breakthrough in computational value by applying our sophisticated reasoning to a mix of unstructured and unstructured data. We are beginning to see companies launching services under the banner of *Web 3.0* [25] that aim explicitly

[**] The essential dimensions of structured data for the travel domain seemed to be: personal identity (who is the travel writer and who is the travel researcher), geopolitical and geospatial location, time and sequence at the granularity of planning a trip, content quality ratings, and travel preferences (expressed as tags such as "wildlife viewing" and answers to questions such as "how important is meeting new people on your trip?"). These dimensions are the basis for aggregation and harvesting value from the unstructured text and images in the travel content. They would also be ideal ways to combine data across web sites using the machinery of the Semantic Web.

at collective intelligence. For instance, MetaWeb [35] is collecting a commons of integrated, structured data in a social web manner, and Radar Networks [25] is applying semantic web technologies to enrich the applications and data of the social web.

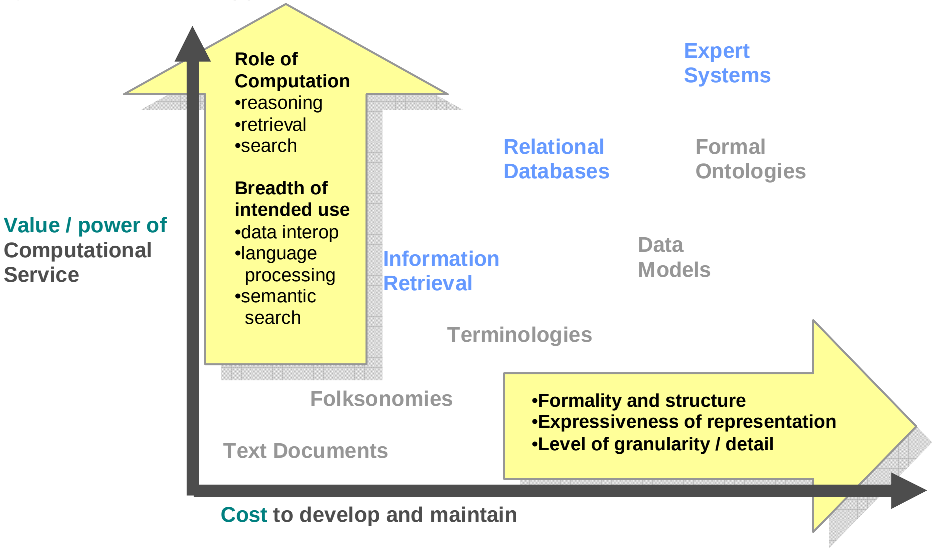

Figure 3. Value/cost tradeoff correlates with structure and depth of inference. For many information technology systems, the power of the computational service offered depends on the level of structure and completeness of the data. The integration of Social Web and Semantic Web may allow for a new synergy, lowering the cost of data and raising the computational value of gathering it.

Speculating a bit, I believe that we may be at the cusp of a change in the way we learn from each other on a global scale. To adapt Engelbart's ([12]) term, we may now be at the point where we are able to "bootstrap" our collective intelligence. The Human Genome project was a watershed event, showing the value of gathering and sharing data for the common good. The Web 2.0 ethos of participation can also apply to structured, scientific data (e.g., SDSS SkyServer project [34] for sharing astronomy data). The mass collaboration model of Wikipedia is being applied to deliberately capture knowledge for the common good, except in this case it will be organized by systematic taxonomy (e.g., the Encyclopedia of Life [11] for sharing knowledge about living things will be organized by species).

We will know we are crossing into the new learning paradigm when we see a qualitative change in the way people think of interacting on the web. Today, that interaction pattern treats the web as an information source: we learn by browsing, searching, and monitoring the web. Tomorrow, the web will be understood as an active human-computer system, and we will learn by *telling* it what we are interested in, *asking it* what we collectively know, and using it to *apply* our collective knowledge to address our collective needs.

References

[1] E. Agichtein and L. Gravano. Snowball: Extracting relations from large plaintext collections. In *Proceedings of the 5th ACM International Conference on Digital Libraries* (DL-00), pp. 85-94, San Antonio, Texas, 2000.

[2] S. Auer and J. Lehmann: What have Innsbruck and Leipzig in common? Extracting Semantics from Wiki Content. *4th European Semantic Web Conference* (ESWC 2007), June 3-7, 2007.

[3] T. Berners-Lee, J. Hendler, and O. Lassila. The Semantic Web, *Scientific American*, May 2001.

[4] C. Bizer et al.: DBpedia - Querying Wikipedia like a Database. Developers track presentation at the 16th international conference on World Wide Web, WWW 2007, Banff, Canada, May 8-12, 2007.

[5] C. Bizer, R. Cyganiak, and T. Gauß. The RDF Book Mashup: From Web APIs to a Web of Data, 3rd Workshop on Scripting for the Semantic Web, ESWC, Innsbruck, Austria, June 6, 2007. Available online as CEUR Workshop Proceedings, ISSN 1613-0073, online CEUR-WS.org/Vol-248/paper4.pdf.

[6] N. Borwankar. TagSchema: Database technology for tag based applications, blog at http://tagschema.com/blogs/tagschema/.

[7] M. Cafarella, D. Downey, S. Soderland, and O. Etzioni. KnowItNow: Fast, scalable information extraction from the Web. *In Proceedings of the Human Language Technology Conference* (HLT-EMNLP-05), pp. 563-570, Vancouver, Canada, 2005.

[8] M. J. Cafarella, D. Downey, S. Soderland, and F. M. Suchanek, G. Ifrim, and G. Weikum. Combining linguistic and statistical analysis to extract relations from web documents. In proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-2006), Philadelphia, USA, August 20-23, 2006.

[9] M. V. Copeland. Weaving the (Semantic) Web. Business 2.0., pp. 88-93, July 2007.

[10] S. Decker. The social semantic desktop: Next generation collaboration infrastructure. *Information Services & Use*, vol. 26, no. 2, pp. 139-144, 2006.

[11] The Encyclopedia of Life. http://www.eol.org/

[12] D. C. Engelbart. A Conceptual Framework for the Augmentation of Man's Intellect. *Vistas in Information Handling*, Howerton and Weeks (eds), Washington, D.C.: Spartan Books, pp. 1-29. Republished in Greif, I. (ed) 1988. *Computer Supported Cooperative Work: A Book of Readings,* San Mateo, CA: Morgan Kaufmann Publishers, Inc., pp. 35-65. The original technical report (1963) is available as http://www.bootstrap.org/augdocs/friedewald030402/augmentinghumanintellect/ahi62index.html.

[13] R. B. Fuller and E. J. Applewhite. *Synergetics*. New York: Macmillan, 1975.

[14] D. Gillmor. *We the Media*. Sebastopol, California: O'Reilly Media, 2004. Also available online as http://www.authorama.com/book/we-the-media.html .

[15] Mark Greaves, Semantic Web 2.0, IEEE Intelligent Systems, vol. 22, no. 2, pp. 94-96, Mar/Apr, 2007.

[16] T. R. Gruber. A Translation Approach to Portable Ontology Specifications. *Knowledge Acquisition*, **5**(2), 199-220, 1993. See also http://www-ksl.stanford.edu/kst/what-is-an-ontology.html

[17] T. R. Gruber. Ontology of Folksonomy: A Mash-up of Apples and Oranges, *International Journal on Semantic Web and Information Systems*, Vol. 3, Issue 1, 2007. Based on paper presented at MTSR’05, November 2005.

[18] A. Hotho and B. Hoser, Eds. Bridging the Gap between Semantic Web and Web 2.0. Workshop located at the ESWC, Innsbruck, Austria, June 6, 2007. http://www.kde.cs.uni-kassel.de/ws/eswc2007/

[19] H. L. Kim, J. G. Breslin, S. H. Hwang, and H. GF. Kim. Social Semantic Cloud Of Tags: Let's Share Tags, http://scot-project.org/.

[20] T. Knerr. Tagging Ontology- Towards a Common Ontology for Folksonomies, 2006, available at: http://code.google.com/p/tagont/

[21] M. Koivunen. Annotea and Semantic Web Supported Collaboration, http://www.annotea.org/. UserSWeb workshop at the *2nd European Semantic Web Conference* (ESWC 2005).

[22] J. Lanier. Digital Maoism: The Hazards of the New Online Collectivism. *The Edge*, May 30, 2006. http://www.edge.org/3rd_culture/lanier06/lanier06_index.html

[23] D. Lin and P. Pantel. Concept discovery from text. In *Proceedings of the 19th International Conference on Computational Linguistics* (COLING-02), Taipei, Taiwan, pp. 1-7, 2002.

[24] L. Lita and J. Carbonell. Instance-based question answering: A data driven approach. In Proceedings *of the Conference on Empirical Methods in Natural Language Processing* (EMNLP-04), pp. 396-403, Barcelona, Spain, 2004.

[25] J. Markoff. Entrepreneurs See a Web Guided by Common Sense. *New York Times*, November 12, 2006.

[26] P. Mika. Ontologies Are Us: A Unified Model of Social Networks and Semantics, in Proceedings of the *International Semantic Web Conference 2005* (ISWC 2005), pp. 522–536, 2005.

[27] D. Miller and D. Brickley. Friend of a Friend Project. http://www.foaf-project.org/ and http://xmlns.com/foaf/0.1/. 2000.

[28] R. Mooney and R. Bunescu. Mining knowledge from text using information extraction. *SIGKDD Explorations*, 7(1):3-10, 2005.

[29] R. Newman. Tag Ontology design, available at: http://www.holygoat.co.uk/projects/tags/, 2005.

[30] T. O'Reilly. What Is Web 2.0: Design Patterns and Business Models for the Next Generation of Software, September 30, 2005. http://www.oreillynet.com/pub/a/oreilly/tim/news/2005/09/30/what-is-web-20.html.

[31] P. Pantel and D. Ravichandra. Automatically labeling semantic classes. In *Proceedings of the 2004 Human Language Technology Conference* (HLT-NAA Cl-04), pp. 321-328, Boston, Massachusetts, 2004.

[32] P. Pantel and M. Pennacchiotti. Espresso: Leveraging generic patterns for automatically harvesting semantic relations. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics* (COLING-ACL-06), pp. 113-120, Sydney, Australia, 2006.

[33] The rel-tag microformat. http://microformats.org/wiki/rel-tag.

[34] SDSS SkyServer, http://cas.sdss.org/dr6/en/.

[35] Sharing what matters. *The Economist*, June 7, 2007.

[36] Y. Shinyama and S. Sekine. Preemptive information extraction using unrestricted relation discovery. In *Proceedings of the 2006 Human Language Technology Conference* (HLT-NAACL-06), pages 204-311, New York, USA, 2006.

[37] L. Specia and E. Motta. Integrating Folksonomies with the Semantic Web, Proceedings of the *4th European Semantic Web Conference* (ESWC 2007), 2007.

[38] R. Snow, D. Jurafsky, and A. Y. Ng. Semantic taxonomy induction from heterogeneous evidence. In ACL, 2006.

[39] F. M. Suchanek, G. Kasneci, G. Weikum: Yago: A Core of Semantic Knowledge - Unifying WordNet and Wikipedia. In *Proceedings of the 16th international conference on World Wide Web*, WWW 2006, Banff, Canada, May 8-12, 2007.

[40] J. Surowiecki, *The Wisdom of Crowds*, Random House, 2004.

[41] TagCommons project. http://tagcommons.org.

[42] K. Tokunaga, J. Kazama, and K. Torisawa. Automatic discovery of attribute words from Web documents. In *Proceedings of the 2nd International Joint Conference on Natural Language Processing* (IJCNLP-05), pages 106-118, Jeju Island, Korea, 2005.

[43] F. Turner. From Counterculture to Cyberculture: Stewart Brand, the Whole Earth Network, and the Rise of Digital Utopianism. University Of Chicago Press, 2006.

[44] T. Vander Wal, (2004). Folksonomy. http://vanderwal.net/folksonomy.html.

[45] T. Vander Wal. TagWatcher, http://www.tagwatcher.info/blog/.

[46] M. Völkel, M. Krötzsch, D. Vrandecic, H. Haller, R. Studer. Semantic Wikipedia. In *Proceedings of the 16th international conference on World Wide Web*, WWW 2006, Edinburgh, Scotland, May 23-26, 2006. May 2006.

[47] L. von Ahn. Games With a Purpose, IEEE Computer Magazine, June 2006. Pages 96-98.

[48] Web 2.0 Conferences. http://www.web2summit.com/

[49] Yahoo! Travel. http://travel.yahoo.com/