SemCards: A New Representation for Realizing the Semantic Web

Kristinn R. Thórisson[1,2] Nova Spivack[2] James M. Wissner[2]

[1]Center for Analysis & Design of Intelligent Agents
Reykjavik University, Kringlunni 1, 103 Reykjavik, Iceland

[2]Radar Networks, Inc.
410 Townsend Street, Suite 150, San Francisco, CA 94107 USA

thorisson@ru.is,{nova,jim}@radarnetworks.com

Abstract. The Semantic Web promises increased precision in automated information sorting, searching, organizing and summarizing. Realizing this requires significantly more reliable meta-information than is readily available for basic human-readable data types today. Relying solely on hand-crafted ontologies and annotation, or solely on artificial intelligence techniques, seems less likely for success than a combination of the two. How this is best done, however, is far from obvious. We propose an intermediate ontological representational level we call *SemCards* that combines ontological rigour with flexible user interface constructs. SemCards are machine- and human-readable entities that allow non-experts to create and use semantic content with ease, while empowering machines to better assist and participate in the process. We have implemented the SemCard technology on the Semantic Web site *Twine.com*, which to date has a growing 250k subscribers and over 2 million monthly unique visitors. SemCards allow users to quickly create semantically-grounded data that in turn acts as examples for automation processes, creating a positive iterative feedback loop of metadata creation between user and machine. The result is an increasingly larger, more accurate amount of metadata than with either approach alone. The SemCard provides a holistic solution to the Semantic Web, resulting in powerful management of the full lifecycle of data, including its creation, retrieval, classification, sorting and sharing. Here we present the key ideas behind SemCards and describe the initial implementation of the technology on *Twine.com*.

Key words: Semantic Web, Ontologies, Knowledge Management, User Interface, SemCards, Human-Machine Collaboration

1 Introduction

One thing we can all agree on: The world of information workers welcomes any improvement in information processing automation. Intelligent automated retrieval, manipulation, combination, presentation – and organized disposal – of information defines the speed of progress in much of today's high-technology work. Included in many people's vision of the Semantic Web is for machines to have better knowledge of the data they manipulate. This requires metadata – data about the data. Making machines smarter at tasks such as retrieving relevant information at relevant times automatically from the vast collection, even on the typical laptop hard drive, requires much more meta-information than is presently available for such data.

Accurate metadata can only be derived from an understanding of content; classifying photographs according to what they depict, for example, is best done by a recognition of the entities in them, lighting conditions, weather, film stock, lens type used, etc. Hand-crafting metadata for images, to continue with this example, will be an impossible undertaking, even if we limited the metadata to surface phenomena such as the basic objects included in the picture, as the number of photographs generated and shared by people is increasing exponentially. Powertools designed for manual metadata creation would only improve the situation incrementally, not exponentially, as needed.

Although text analysis has come quite a long way and is much further advanced than image analysis, artificial intelligence techniques for analyzing text and images have a long way to go to reliably decipher the complex content of such data. The falling price of computing power could help in this respect, as image analysis is resource-intensive. This will not be sufficient, however, as *general-purpose* image analysis (read: software with commmon sense) is needed to analyze and classify the full range of images produced by people based on content. Relying exclusively on automated creation of metadata seems thus equally doomed as complete reliance on hand-crafting. The question then becomes, What kind of collaborative framework will best address the building of the Semantic Web?

Semantic Cards, or *SemCards*, is a technology that links ontology creation, management and usage to the *user interface*. SemCards provide an intermediate ontological representational level that allows end-users to create rich semantic networks for their information sphere. The technology has three important features: (1) it does not require expertise users, (2) it is tolerant to end-user mistakes, and (3) it provides examples of metadata and semantic relationship links to the underlying machine intelligence. The best known way to create structured metadata is through the use of carefully constructed ontologies. The casual Internet user, however, is not initiated to invest a lot of time in understanding the intricacies of the kinds of advanced ontologies required for this purpose. An important difference between many prior efforts and ours is this separation, done in a way that is tolerant to manual input errors. One of the big problems with automation is low quality of results. While statistics may work reasonably in some cases as a solution to this, for any single user the "average" user is all too often too different on too many dimensions for such an approach to be useful. The SemCard intermediate layer encourages users to create metadata and semantic links, which provides underlying automation with highly specific, user-motivated examples. The net effect is an increase in the possible collaboration between the user and the machine. Semi-intelligent processes can be usefully employed without requiring significant or immediate leaps in AI research. Other important differences between our approach and prior work are an integrated ability to share data between individuals and groups of users over a network, with complex policy control over access and sharing, and the flexible use of SemCards to represent metadata for real-world objects and hypothetical constructs - as “library index cards for digital content, physical things and abstract ideas”.

We have built a network portal, Twine.com, for deploying the SemCard technology. Although current enterprise portals are capable of organizing group or team information, they are often inaccessible to the public or to individuals and are expensive and monolithic. Even less utilitarian and intelligent with respect to organizing information are the popular online search engines which are completely unstructured and typically organize information and data by relevance to keywords. From the users’ perspective what we have developed is a network portal where they can organize their own information for personal use, publish any of that information to any group – be it “emails” addressed to a single individual or photo albums shared with the world – and manage the information shared with them from others, whether it is documents, books, music, etc. Under the hood are powerful ontology-driven technologies for organizing all categories of data, including access management, relational (semantic) links and display policies, in a way that is relatively transparent to the user. The result is a system that offers improved automation and control over access management, information organization and display features.

Here we describe the ideas behind the approach and give a short overivew of a use-case on the semantic Web site Twine.com.

2 Related Work

The full vision of the Semantic Web will require significant amounts of metadata, some of which describes entities themselves, other which describes relationships between entities. Two camps can be seen proposing rather different approaches to this problem. One extreme claims that manual creation of metadata will never work as it is not only slow and error-prone, the level to which it would have to be done would go well beyond the patience of any average user – quite possibly all. To this camp the only real option is automation. The other camp points out that automation is even more error-prone than manual creation, as current efforts to automatic semantic annotation on massive scales produces only moderate results of between 80% and 90% correct, at the very best [1]. They claim that the remaining 10% will always be beyond reach because it requires significant amounts of human-level intelligence to be done correctly. Further, as argued by Etzioni and Gribble [2], metadata augmentation has quite possibly not been done by the general user population because they have seen no benefits in doing so. Lastly, this camp points to the massive amounts of tagging and data entry done on sites such as Wikipedia, Myspace and Facebook as a proof in point that end-users are quite willing to provide (some amount of) metadata. Giving them the right tools might change this. Applications that connect casual end-users with ontologically-driven content and processes are, nevertheless, virtually non-existent.

Many efforts have focused on building digital content management with a focus on the object. Of these, our technology bears perhaps the greatest resemblance to the *Buckets* of Maly et al. (1999) [3] which are "self-contained, intelligent, and aggregative ... objects that are capable of enforcing their own terms and conditions, negotiating access, and displaying their contents". Like SemCards, Buckets are fairly self-contained, with specifications for how they should be displayed. Buckets grew out of Kahn and Wilensky's (1995) [4] proposed infrastructure for digital information services. Key to their proposal was the notion of digital object, composed of essentially the two familar parts, data and metadata. The subsequent work on FEDORA [5] saw the creation of an open-source software framework for the "storage, management, and dissemination of complex objects and the relationships among them" [6]. Buckets represent a focus on storing content in digital libraries, most likely manipulated by experts. In contrast, SemCards aim at enabling casual end-users to create metadata. Buckets are targeted to machime manipulation; SemCards are aimed at machine *manipulation* as well, but more importantly at supporting *automatically generated* meta-information. SemCards also differ from Buckets in that they are especially designed to be sharable between multiple users over mixed-architecture networks.

The separate representation layer provided by SemCards is a key difference between prior efforts and ours. They enable ontologically-driven constructs to be collaboratively built by ontology specialists, algorithms and end-users, encouraging them to provide examples to improve the automation.

3 SemCards: Semantic Objects for Collaborative Ontologically-Driven Information Management

SemCards form an intermediate separation layer between ontologies and the user interface. By isolating the stochastic nature of end-user activity from underlying semantic networks built with ontological rigour, two important goals are achieved. First, end-users are encouraged to create metadata for their content, as the input methods are familiar and straight-forward. SemCards provide an "isolation layer" that shields the deep ontology from being affected by end-user activity. This does not only help stabilize the system, it also helps the automation processes from having to deal with the "ground shifting from underneath". Second, the automation processes are provided with semantic net examples, created directly and continuously by end-users, that can be used to improve the automatic metadata creation. The net result of this is a significant improvement in automation quality and speed. The result is an improvement in the many tedious details of information management such as data sharing policy maintenance, indexing, sorting – in fact, the of the full data management lifecycle.

A single SemCard can be characterized as an intuitive user interface construct that bridges between a user and an underlying ontology that affords all the benefits of a Semantic Web such as automatic relationship discovery, sorting, data mining, semantic search, etc. Together many SemCards form semantic nets that are in every way the embodiment of what many have envisioned the Semantic Web to be. Instead of being complex, convoluted and non-intuitive as any machine-manipulatable ontology will appear to the uninitiated (c.f. [7]), SemCards provide a powerful and intuitive interface to a unified framework for managing information.

Every SemCard instance has a GUID, timestamps representing time of creation and related temporal aspects such as times of modification, as well as a set of policies. Its author is also represented, and any authors of modifications throughout the SemCard's lifetime. The SemCard's policies allow it to be displayed, shared, copied, etc. in prescribed ways, through the use of rules. In its simplest version a SemCard will appear as a form with fields or slots. A SemCard has one template and one or more instances, which corresponds roughly to the object-oriented programming concepts of object template/class, and object instance, respectively. Under the hood their slots are ontologically defined; however, the end user normally does not see this. To take an example, a SemCard for holding an e-mail message may look exactly like any interface to a regular email program. However, the slots reference an ontology that defines what kinds of data each slot can take, what type of information that is, etc. The e-mail SemCard, when created, will contain information about who authored which part of the content and when. Additionally, the author will not simply be a regular "from" but have a link to the SemCard representing the author of the email SemCard.

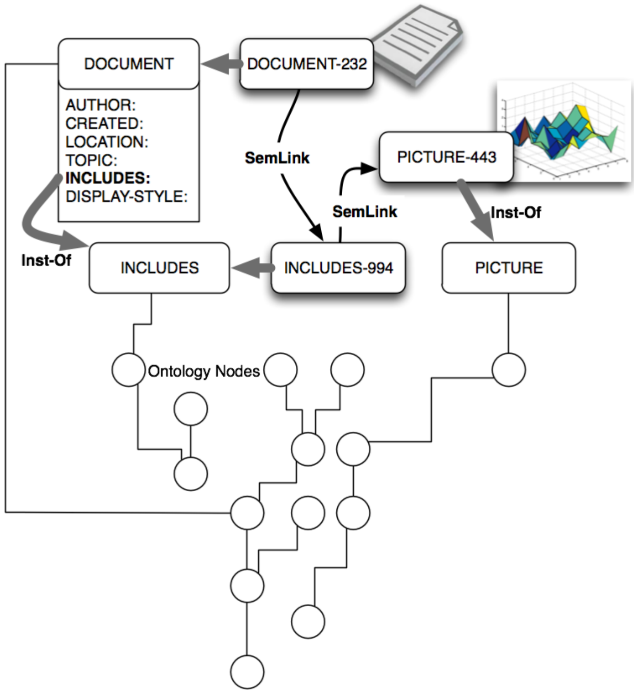

Fig. 1: Metadata for entities, digital or physical, is semantically defined by an underlying ontology that appears to the user as (networks of) SemCards.

SemCards are "dumb" in that they do not carry with them any executable code: We have completely separated the services operating on the SemCards from the SemCards themselves, leaving only a spec for the desired operations to be done on a SemCard in the SemCard itself. A SemCard can thus represent any digital item, like a png image or pdf document, *physical entities* such as a person, building, street, or a kitchen utensil, *as well as immaterial things* like ideas, mythologicical phenomena and intellectual creations.

Any type of digital object or information can be pointed to with a SemCard, e.g. a Web page, a product, a service offer, a data record in a database, a file or other media object, media streams, a link to a remote Web service, etc. Equally importantly, SemCards can represent relationships between SemCards, for example, that a person is the author of an idea. To fill out a SemCard instance, one or more slots are filled with values – these could be semantic links to other SemCards, typed entities or unclassified content. Each SemCard instance, its semantic dimensions and their values, can be stored as an XML (extensible Markup Language) object, using e.g. the RDF (Resource Description Framework) format [9].

Display rules dictate how the SemCard itself (as well as its target reference - the thing it represents) should be displayed to the user. These can describe, for example, its owner's preferences or the display device required. As SemCards carry with them their own display specifications their on-screen representation can be customized by their userss; the same SemCard can thus be displayed differently to two different users with different preferences. The rules can specify how metadata and slot values in the SemCard should be organized and what human-readable labels should be used for them, if any, as well as what aspects of the SemCard appear as interactive elements in the interface, and the results of specific interaction with those elements.

For viewing and manipulating SemCards we have developed both client-based editors in the spirit of Haystack [8] and Web-based interfaces. SemCards can be created in many ways; doing so manually from scratch involves selecting a SemCard template type, making an instance of it and customizing its slots using typed entities from an underlying ontology.

As SemCards isolate the user from the related ontologies, classificatory mistakes in their creation does not destroy the underlying ontologies. This results in a kind of graceful degradation; instead of breaking the system such mistakes only make the automated handling of information in the system slightly less accurate. The relationship between SemCards and the unerlying ontology can be likened to non-destructive editing for video: As the creation history (original data, i.e. ontologies) are not changed but rather represented in a separate intermediate layer, the edit history of any SemCard can be traced back and reverted, if need be, with no change to the underlying ontologies.

As explained, a SemCard can represent any digital item, like a png image or pdf document, as well as physical entities, even ideas. SemCards are also be used to represent relationships between SemCards; the type of relationship is then that SemCard's *type*. For example, A SemCard can also represent collections, for example a SemCard representing a group of friends would contain links to the SemCards representing the individuals of that social group.

Behind each SemCard is an ontology that defines the meaning of the SemCard slots, specifies valid values and relations between slots (see Figure 1). An ontology like FOAF (c.f. [10]) or the Dublin Core [11] can be used with SemCards, as each SemCard carries with it a reference to the ontology it is based on. Thus, networks of ontologies can be used with SemCards, whether they use a basic, simple and singleton ontology like the Dublin Core or are definded more deeply in e.g. foundational ontologies such as DOLCE, SUMO [12] [13], or OCHRE [14].

In our current implementation we have created a fairly extensive ontology for important digital data types including Web page, 2-D image, URL, text document, as well as for physical entities such as person, place and organization. This ontology will be made open-source, so as to encourage linking of other ontologies to it, extending its reach and improving its utility.

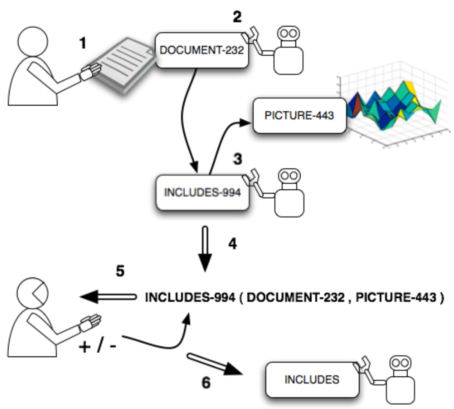

Fig. 2: The iterative nature of human-machine metadata annotation. (1) User creates digital document, (2) a SemCard instance is automatically created; the automation infers that a particular image is included in the document and (3) creates a SemCard for it and a SemCard of type *Includes* that links the two; (4) relationship between the SemCards now forms a triplet that the user can inspect (here shown in prepositional form, but is typically graphical); (5) user modifies the results (+/-) from which (6) the automation processes generalizes to improve own performance.

Although SemCards instances ideally derive from a SemCard template that is fully defined by one (or more) ontologies, the case could arise where a user wants to represent an entity for which no template exists. A user can create free-form slots and collect them into a new SemCard (that has no template). As long as the type of the SemCard – or at least one slot in it – has a connection to a known ontology (it will always have its author and date of creation), the automation mechanisms can use this information to base further automatic refinement of the SemCard instance, like linking it to (what are believed to be) related SemCards. Managing such automatic semantic links becomes akin to unstructured database managment; it will of course never be as good as that for fully-specified SemCards, but because these SemCards live in a rich network of other SemCards, this problem is not as large as it may seem.

4 Collaborative SemCard Creation by Man and Machine

An important feature of SemCards is that they record significant amounts of metadata about themselves, including their own genesis. This makes automatic creation of SemCards much more flexible as the automation process can make inferences about the quality of the SemCards (based on e.g. edit history). Because the same representational framework - SemCards - can be used for *all* data, including friend networks, author-entity relationships, object-owner, etc., inferencing can use the multiple SemCard relationship types (e.g. not only who created it but also who the creator's friends are) to decide how to perform automatic relationship creation, data-slot filling, automatic correction or deletion. Moreover, as the SemCard stores its edit history, including who/what made the edits, any such changes can be undone with relative ease. Since this history is stored as semantic information, it can be used to sort the SemCards according to their history. This makes managing SemCards over time much more flexible than if they were history-scarce, as e.g the losely-defined metadata of most data on people's hard drives. For example, caching, compressing or any other processe be made history-sensitive to a high level of detail.

As an example of collaborative automatic/ manual creation of SemCards, Nova, a SemCard end-user, finds a useful URL and creates a SemCard for it of type "bookmark for a Web-page". He makes personal comments on the Web page's contents by making a "Note" SemCard and linking it to the Webpage SemCard. Nova's automation processes, running on the SemCard hosting site, add two things: They fill the Webpage SemCard with machine-readable metadata from the Web page, and they also link these SemCards to a new SemCard that *it* created, containing further information mined from the Web site. Now Nova shares (a copy) of the SemCard with Jim (it gets saved in his SemCard space), who may add his own comments and links to related SemCards; the fact that the SemCard was shared with Jim by Nova is automatically recorded as part of the SemCard's metadata. Thus, events, data and metadata are created seamlessly and unobtrusively through collaborative paradigm.

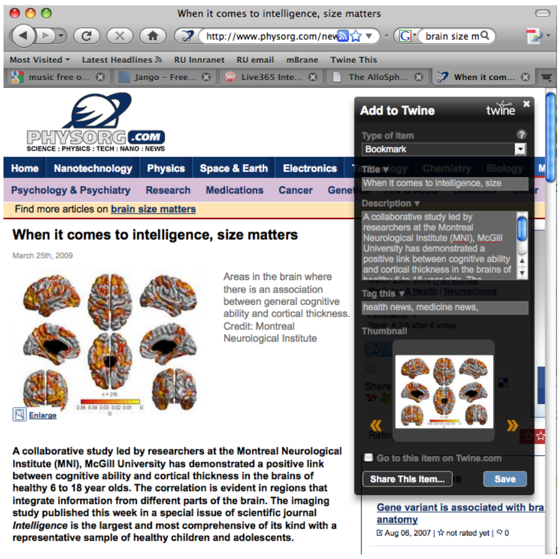

Fig. 3: The Twine bookmarklet popup enables Web surfers to create SemCards quickly. The system automatically fills in relevant information ("Title", "Description", "icon", etc.).

As their authorship is automatically recorded in the SemCards, this can be later used to e.g. exclude all SemCards created by particular automation processes, should this be desired. Proactive automatic mining of a user's SemCards can reveal implicit relationships that the system can automatically make explicit, facilitating faster future retrieval through particular relationship chains in the resulting relationship graphs.

A positive feedback loop of iterative improvement on the network is created through the collaboration between a user and his/her data (2): When the initial manual data entry reaches a critical point the automation starts to provide significant enhancements to the user. Increased manual input, especially in the form of additions to automatically generated semantic links, allows the automation system to make inferences about the quality of the data entry, not just for a single user but for many. This enables it to improve the accuracy of its own automation, and suggestions to users about related data will be more relevant and targeted.

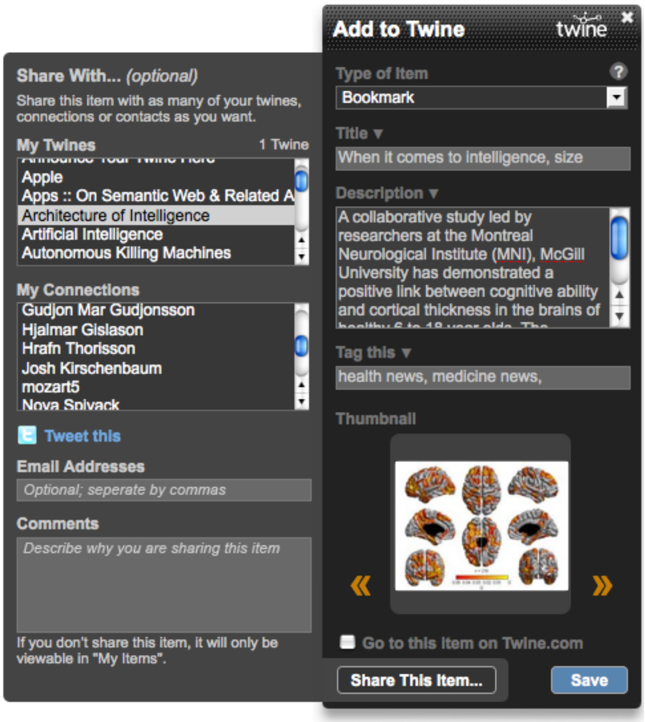

Fig. 4: Upon creation of the bookmark, SemCard users can choose to share it (left side of popup) with users via Twines they have created or subscribed to ("My Twines"), or directly with people they have connected with ("My Connections").

As a users's SemCard database grows user-customized automation becomes more relevant; in the long run, as the benefits of automation become increasingly obvious to each user, people will see the benefits of providing a bit of extra meta-information when they create e.g. a word processing document or an image. This will trigger a positive upward spiral where increased use of automation will motivate users to add more pieces of metadata, which will in turn enable better automation.

5 Deployment on Twine.com

We have started to implement the SemCard technology and deployed it on the *Twine.com*, an online Semantic Web portal where people can create accounts and use a SemCard-enabled system to manage their online activities and information, including bookmarks, digital files, sharing policies, and more.

At the time of this writing there are well over 4 million SemCards on Twine.com. Twine.com has around 250 thousand registered users and over 2 million unique visits montly. The rate of new SemCard creation has grown to 3K per day, by an estimated 10% of the users. So far, users rarely correct the automatically-generated SemCards, but a subset of the users add extensive additional information to them. We are currently in the process of scaling up the system as well as developing increasingly sophisticated automation and mining processes to provide the user with related items, related users and recommended items. Users' behavior with regards to these items is used for continuous improvement of these processes.

Fig. 5: Automatic processing of a bookmarked Web page can detect places, people organizations and various named entities ("tags"). The user can then modify these by deleting (clicking on the [x]) and adding new ones. Two snapshots of the same drop-down box are shown, the left image showing its "other tags" that were auto-recognized, and the right side "organizations", as well as one user-added "place" ("Montreal").

We will now give an actual example of making a SemCard for a Web page, a short article on the Physorg.com Web site.[1] As can be seen in Figure 3, when a user comes to a Web page of interest they can click on the bookmarklet "Twine This", which brings up a simple menu with a few information fields. Parts of the SemCard slots have been filled in; the user can choose to edit these, overwrite them with her own information or to leave them as-is. When the user clicks "save" a SemCard for this Web page is created in their Twine account. The user can choose to share this item with users and/or *twines* (see Figure 4) – a twine is a SemCard that can be described as "a blog with controlled access permissions" – in other words, a SemCard for a set of SemCards with particular visual presentation and adjustable viewing permissions. The twine SemCard shows the dynamic properties of SemCards for specifying dynamic processes, e.g. calling on services from mining, inferencing, etc.

When the "bookmark" SemCard is saved, using the "Save" button on the lower right on the bookmarklet popup, the SemCard is stored on Twine.com. Any sharing selection that the user had made during the creation will make the bookmark SemCard available to the users who have permissions to read

[1] http://www.physorg.com/news157210821.html

those Twines; for example, sharing it with the twine *Architecture of Intelligence* (Figure 4) will enable everyone who has been invited to subscribe to this twine to see it. In their home page on Twine.com this SemCard will now additionally bring forth a lot of information, including auto-tagging (recognized entities, relationships, etc.).

As seen in Figure 5 a cross next to auto-generated tags allow the user to delete the ones that they don't agree with. Further related information is automatically pulled forward, sorted into "places", "people", "organizations", "other tags" and "types of items": The last one is interesting as it is a unique feature of semantic Webs – here one can find related SemCards of type "video", for example, or "product". Vector space representations are used to profile users and their semantic networks and subsequently select related items from other semantic nets. Using a (semantic) drill-down search mechanism a user can keep refining a search for a SemCard, by selecting any combination of type, tags, author, etc. (Figure 6). During such drill-downs, suggestions by the automation of related material become increasingly better.

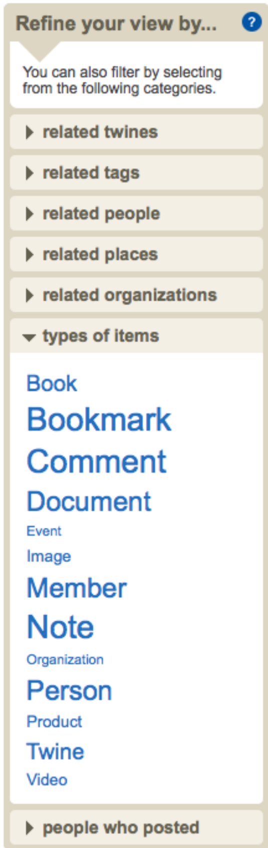

Fig. 6: Semantic search box ion *Twine.com*.

6 Conclusions and Future Work

SemCards are a powerful representation scheme for enabling collaborative human-machine development of Semantic Web information. The technology achieves this by separating hard-core ontologies from the end-user, mediating these via graphical information structures, represented under the hood using RDF and OWL but supplying their own visual representation schemas for on-screen viewing. The SemCard framework allows better sharing, storing, annotating, enhancing and expanding semantic networks, creating true knowledge networks through a collaboration between people and artificial intelligence programs.

We have implemented the SemCard technology on the Semantic Web site *Twine.com*, which has over 2 million monthly unique visitors. In close collaboration with automation processes, these users have created over 4 million SemCards to date. Our results so far show that SemCards can support all of the features described in this paper for hundreds of thousands of users and we have good reason to believe that the technology will scale well. Further, many more features are envisioned, relating to searching, querying and mining. While these have not been described here they have been prototyped in our labs and will be unveiled on Twine.com and in other Radar Networks applications and services in the near future.

Acknowledgments. The work was supported by Radar Networks, Inc., by Reykjavik University, Iceland, and by the School of Computer Science, Reykjavik University. The authors would like to thank the Twine team for the implementation of the SemCard technology on Twine.com, as well as the funders of Radar, Vulcan Ventures Inc., Leapfrog Ventures Inc. and the angel investors.

References

1. Dill, S., Eiron, N., Gibson, D., Gruhl, D., Guha, R., Jhingran, A., Kanungo, T., Rajagopalan, S., Tomkins, A., Tomlin, J.A., Zien, J.Y.: SemTag and Seeker: Bootstrapping the Semantic seb via automated semantic annotation. In: Proceedings of the World Wide Web Conference. (2003) DOI: 10.1145/775152.775178.
2. Etzioni, O., Gribble, S.: An evolutionary approach to the Semantic Web. In: Poster presentation at the First International Semantic Web Conference. (2002)
3. Maly, K., Nelson, M.L., Zubair, M.: Smart objects, dumb archives: A user-centric, layered digital library framework. In: D-Lib Magazine. Volume 5. (1999)
4. Kahn, R., Wilensky, R.: A framework for distributed digital object services. In: International Journal on Digital Libraries. Volume 6. (1995) 115–123
5. Payette, S., Lagoze, C.: Flexible and extensible digital object and repository architecture (FEDORA). In: Lecture Notes In Computer Science. Volume 1513. (1998) 41–59
6. Lagoze, C., Payette, S., Shin, E., Wilper, C.: Fedora: An architecture for complex objects and their relationships. In: Journal of Digital Libraries - Special Issue on Complex Objects. (2005) DOI: arXiv:cs/0501012v6.
7. Drummond, N., Jupp, S., Moulton, G., Stevens, R.: A practical guide to building OWL ontologies using the Protégé 4 and CO-ODE tools, edition 1.2. (2009)
8. Huynh, D., Karger, D.R., Quan, D.: Haystack: A platform for creating, organizing and visualizing information using RDF. In: Semantic Web Workshop, WWW2002. (May 2002)
9. Decker, S., Melnik, S., Harmelen, F.V., Fensel, D., Klein, M., Erdmann, M., Horrocks, I.: Knowledge networking in the Semantic Web: The roles of XML and RDF (2000)
10. Ding, L., Zhou, L., Finin, T., Joshi, A.: How the Semantic Web is being used: An analysis of FOAF documents. In: Proceedings of the 38th International Conference on System Sciences. (2005) p. 313.3 DOI: 10.1109/HICSS.2005.299.
11. Sutton, S.A., Mason, J.: The Dublin Core and metadata for educational resources. In: International Conference on Dublin Core and Metadata Applications. (2001) 25–31
12. Varma, V.: Building large scale ontology networks. In: Language Engineering Conference (LEC'02), Hyderabad, India (2002) p. 121
13. Hong, J.F., Li, X.B., Huang, C.R.: Ontology-based predication of compound relations: A study based on SUMO. In: Proceedings of PACLIC18, Waseda University, Japan (2004) 151–160
14. Schneider, L.: Designing foundational ontologies: The object-centered highlevel reference ontology OCHRE as a case study. In: Conceptual Modeling - 22nd International Conference on Conceptual Modeling, I.-Y., Springer (2003) 91–104