The Networked Semantic Desktop

,

Stefan Decker
DERI, NUIG, Ireland
stefan@deri.ie

Martin Frank
ISI, USC, USA
frank@isi.edu

,

Abstract

We present our vision of a new group collaboration infrastructure, the Networked Semantic Desktop, drawing from co-evolving research in the Semantic Web, Peer-to-Peer (P2P) Networks, and Online Social Networking.

1. Introduction

The Internet, electronic mail, and the Web have revolutionized the way we communicate and collaborate - their mass adoption is one of the major technological success stories of the 20th century. We now face a qualitatively different problem, information overload, that necessitates smarter and more fine-grained computer support for networked information, and that has to blend the boundaries between personal and group data, while simultaneously safeguarding privacy and establishing trust. In other words, the current computing infrastructure does not really support knowledge workers all that well: for example, sending a single file to a mailing list multiplies the cognitive processing effort of filtering and organizing this file times the number of recipients – leading to more and more of peoples' time going into information filtering and organization activities. Centralized collaborative infrastructures (like BSCW or Sharepoint) help to a certain extent, but the current application infrastructure does not let you interconnect separate data items, like the author of a document and her corresponding entry in your address book - much less let you share that interconnection with others.

Several new technology thrusts have now emerged which could dramatically impact how people interact and collaborate: The Semantic Web, P2P Computing, and Online Social Networking. This paper presents a vision of how the different thrusts will evolve to produce the Networked Semantic Desktop, which enables people and communities to directly collaborate with their peers while dramatically reducing the amount of time they spend filtering and filing information.

2. Usage Scenarios

We exemplify the impact of a Networked Semantic Desktop with two usage scenarios.

2.1. Surviving the Information Flood and Creating Knowledge in the Process

In our daily life, many of us get hundreds of emails, often with documents attached from the various different projects and communities we are involved in. These documents are always created within a context on the author's machine – but they are send out as if they had no context – they arrive without trusted metadata that would allow automatic processing and filing on the recipient's machine. The process has the following shortcomings:

- Apart from the folder structure the current Windows-style desktop and file system provides no support for organizing the information in the documents. This means the recipient has to cope with an insufficient support of current desktops systems for organizing the information – you may put them under the 2004-04 folder, or the Proposals folders, or the Semantic Web Research folder – but not all three.
- Since the metadata of the document has been lost when sending the email every recipient has to reinvent and recreate their own metadata, re-categorize the document and create the possible connections to other information. This is only of marginal societal value because most metadata has been created before.
- Even if metadata was integrated into the email, the author and the recipient of the document usually have different, personal classification schemes, and there is no way to selectively "open-source" them and align them with others'.

It is clearly possible to share and replicate documents as well as metadata via direct P2P connections. It should be possible for sub-communities to derive metadata in a distributed fashion via an implicit or explicit consensus process. Connections and relationships with other pieces of information could then be accessed by all members of the community – in this sense a distributed knowledge base is constructed around the work topics, documents, and information contained in these document. The information can then be viewed in multiple dimensions (year, type of document, content of document, and so on). Also, the community itself can be queried, similar to current Online Social Networks: members of the community can query to whom they are connected to and can share information with “friends of a friend” or can query for information – thus existing trust relationships between individuals can be leveraged to compute one’s trust in metadata.

2.2 Connecting to Trusted Colleagues

Many of the same ideas are applicable to the audience of this workshop - research is performed in scientific communities of interest in which members develop, review, publish and discuss each other work. Membership in a scientific community is based on interest and abilities: promising applicants are included by accepting their conference papers or inviting them on editorial boards. As a research field progresses these communities define their own language, which is often incomprehensible to outsiders and difficult to learn, yet necessary for efficiently classifying and communicating the topics of interest and for gaining an overview of the different types of research. This language evolves in a sluggish community process today - usually by members reading each others’ papers, by attending the same scientific events (e.g., workshops such as this one) and by then adopting each others’ terminology over time; the appearance of text books and dedicated overview papers consolidates a field’s vocabulary in the later stages. This evolutionary structuring of new scientific fields is largely invisible to non-scientists, both because journal publications can often not be accessed without paying hefty fees and because in any case it would take too much effort to derive a taxonomy of a scientific field by reading scholarly publications.

Paradoxically, being an “insider scientist” can nevertheless be a lonely experience because communication with other community members by e.g. ping-pong journal publications is a slow and faceless process, lacking the spontaneity and friendship-building opportunities of face-to-face communication.

The Networked Semantic Desktop has the potential to accelerate scientific collaboration via a peer-to-peer end-user application for maintaining shared views of scientific fields, as well as to make these evolving views explicit and available to the public at no cost. Such a collaborative application is peer-based both in the scientific sense (“peer review”) and technical sense (“peer-to-peer technology”). Participants automatically become part of a global peer-to-peer network for scientific meta-data, and takes responsibility for a (proportional) fraction of the disk storage, bandwidth, and computing cycles to support it, probably based on a structured P2P network. This single global network could then support a large number of small scientific sub-communities, each of which revolves around jointly maintaining a shared view of a small sub-field. Maintaining this shared view is, of course, not typically an end to itself, but serves as the focal point that enables a scientific community to effectively exchange research papers, data sets, and ideas. Such close-knit communities may work on bottom-up taxonomies for a tiny new sub-field of Science, such as say Pteroylglutamic Acids, or may work on a top down categorization of say Liberal Arts as a whole, and each community can refer to concepts in other communities. Think of this type of Networked Semantic Desktop as collectively harnessing the power of millions of currently hand-scribbled categorizations of scientific sub-fields into an inter-linked, grass-roots, and world-wide view of Science.

For individual scientists to participate in this global scientific meta-data construction effort, there must be immediate benefits to joining, as well as assurances that certain things participants will naturally fear will not occur. The benefits are: By joining the global scientific meta-data network, (b1) one can access and query existing scientific terminology, (b2) one can mine queries by others to find out what the “hot” research topics are, and (b3) one can view others’ structuring of their fields - thus providing instant gratification to joining the global meta-data network. By joining a specific scientific sub-community as an active participant, over time, (b4) one can get to know others in one’s research area for friendship and scientific collaboration, and (b5) one can make a name for oneself by contributing to community taxonomies and the research itself. The assurances are: The peer-to-peer application must re-assure prospective participants that (a1) others will not claim their intellectual contributions as their own, (a2) they will always be in control of their personal view of the field, and they can always choose to cease collaborating with any individual or sub-community, (a3) they will not be politically dependent on any individual in some “gatekeeper” position in the global scientific meta-data network (a4) they will not be technically dependent on the people who own the technical community infrastructure, and (a5) their machine will not be intolerably slowed down by participating. While these requirements seem daunting, we have hope that with recent work on structured P2P networks for meta-data exchange as well as advances in the Resource Description Framework-based Semantic Web languages, such a global scientific collaboration network is within reach in the next five years - whereas just five years ago it would have seemed like a fantasy.

3. Components of the Networked Semantic Desktop

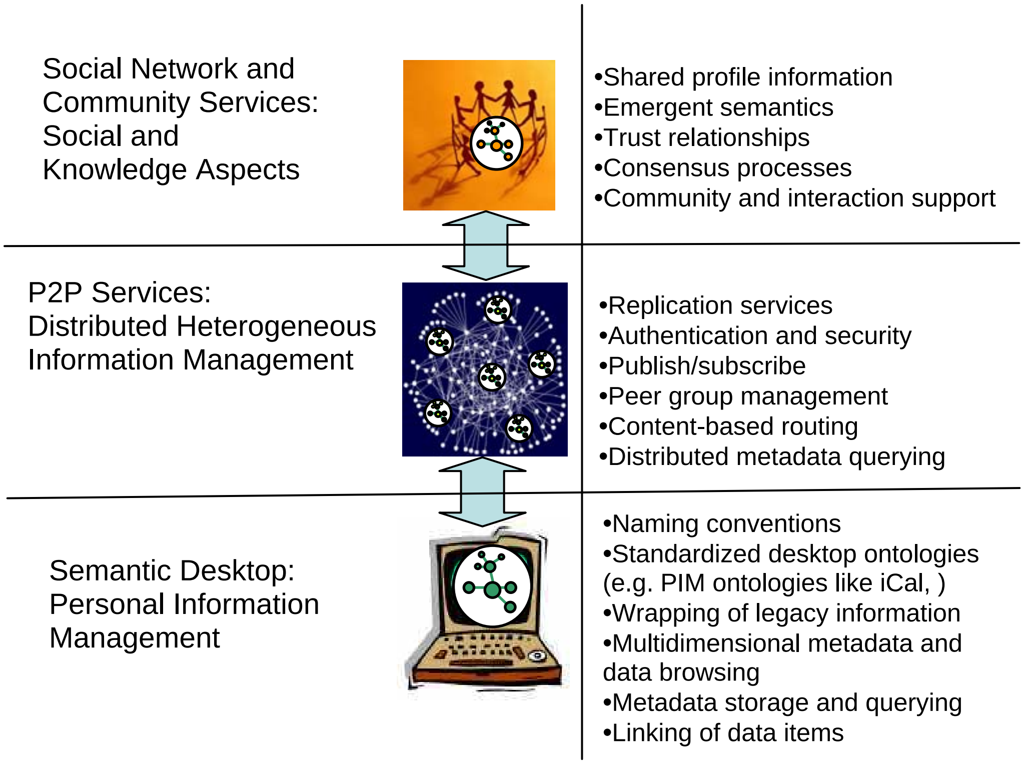

Figure 1 Component Architecture of the Networked Semantic Desktop

Figure 1 shows the highest-level architecture and connections between social networks and the P2P infrastructure that connects those social networks to individuals' desktops.

Traditional semantics, knowledge representation, and reasoning research is now interacting with other research areas, which not individually but together may have the explosive impact of the original Web:

1. *The Semantic Web effor*t (http://www.w3.org/sw) provides standards and technologies for the definition and exchange of metadata and ontologies. Available standard proposals provide ways to define the syntax (RDF) and semantics of metadata based on ontologies (OWL). Research covering data transfer, privacy and security issues is now also under development.
2. *Social Software* maps the social connections between different people into the technical infrastructure, as an example, Online Social Networking makes the relationship between individuals explicit and allows discovering previously unknown relationships. The most recent Social Networking Sites also help form new virtual communities around topics of interest and provide means to change and evolve these communities.
3. *P2P and Grid computing* develops technology to network large communities without centralized infrastructures, for data and computation sharing. P2P networks have technical benefits in terms of scalability and fault tolerance, but a main advantage compared to central sites is a political one: they allow to build communities without centralized nodes of control, much as the Internet grew as fast as it did because it was based on reciprocity – it avoided political debate as to who gets to own big, expensive central facilities. Recent research has provided initial ways of querying, exchanging and replicating data in P2P networks in a scalable way.

By projecting the trajectory of current trends we can (somewhat over-)simplify this picture by stating that next-generation Internet applications will support collaboration and information exchange in a P2P network, connecting online decentralized social networks, and enabling shared metadata creation and evolution by a consensus process, the result being he Networked Semantic Desktop. Figure 1 is our best attempt at the high-level architecture of these new applications.

The following sections describe the available technologies and technology convergences in more detail.

3.1 Desktop Technology and the Semantic Web

The Semantic Web delivers the basic representational infrastructure for metadata. Modern computer (even laptops) hold hard disks which carry in excess of 1 Mio files. Email and documents (and even information in emails and documents) created or downloaded from the Web increase the number of available files everyday.

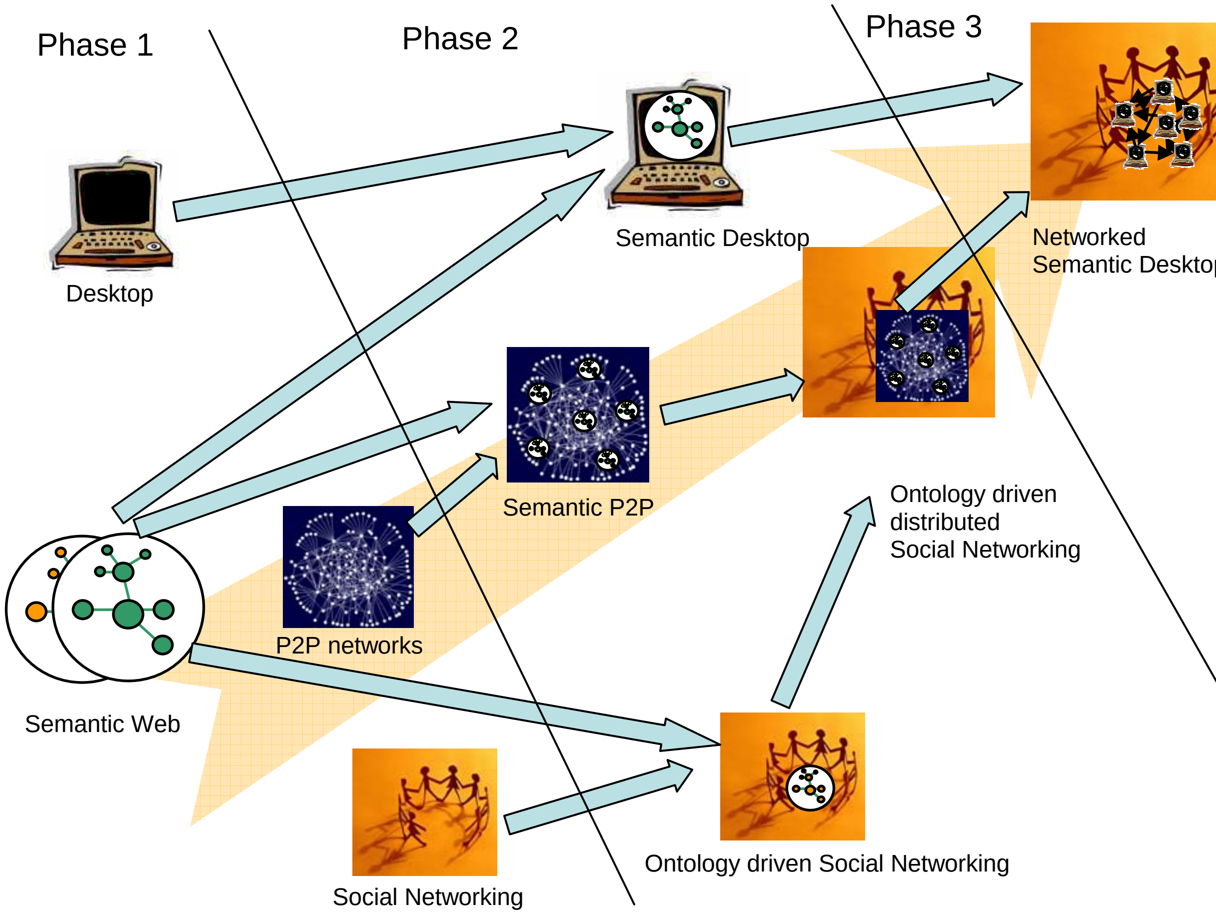

Figure 2 Phases towards the Networked Semantic Desktop

These information items can be treated as Semantic Web resources, which enable the usage of Semantic Web technologies to manage desktop data. Similar to Web resources, files, emails and other items can be identified using a Uniform Resource Identifier (URI). The metadata of these desktop items are represented as RDF graphs, which can then be used for browsing or searching by faceted metadata browsing techniques [17] and can again be shared with others.

This enables a co-evolution with current desktop technology, enabling the development of methods to create metadata and to interlink information on a on a local desktop computer (e.g. the author of a text document with the entry in the address database). Approaches like MITs Haystack [1], or European approaches like Gnowsis [2], or Fenfire [3] show that there is activity going on in this direction. However, these approaches concentrate on the higher levels of the desktop system like the user interface, or realize "add-on" components to handle metadata rather than being built upon meta-data from the ground up. For example, the file system in most operating systems does not have the capability to manage larger amount of heterogeneous metadata because of minimum file size issues. File system approaches like ReiserFS (see [4]) or Microsofts Longhorn raise the hope that current operating system can evolve to effectively support management of metadata on all levels – from the disk storage to the user interface.

3.2 P2P Systems

Starting initially with file sharing application like Gnutella, P2P systems have achieved an enormous amount of research attention, especially in the database community [14] [15]. [13] suggested to add metadata management to P2P networks. This idea has been taken up by approaches like Edutella [5][6][7][11], RDFPeers [8], P-Grid [121] and other projects [9][10]. Initial ideas of a collaboration infrastructure based on P2P technology has been realized in Groove which allows small groups to share a calendar, discussions, and files without a central server and see who is off- and on-line and to sent instant-message with them (but replicates all content to all group members which does not scale well with group size).

However, the field is still in his infancy –there is an distributed efficient query model for distributed RDF sources, but no agreement how "semantics" (as in, inferencing within the network) should be incorporated into P2P networks. Topics like scalable infrastructures for service discovery in P2P networks, interoperation infrastructure for heterogeneous peers to translate information, emergent semantics and incremental learning, evolution of ontologies in an P2P environment, semantics-based routing, and semantics-based topologies for P2P networks are currently all open topics and actively discussed in the scientific community (e.g., [18][19]).

3.3 Online Social Networking

A relatively new field is Online Social Networking, which recently got a lot of attention by venture capitalists and internet users. Sites like linkedin.com, Friendster, and Orkut.com were able to attract millions of users and provide the infrastructure to

a) make relationships explicit, so persons can explore their personal network, and
b) make new connections and establish new relationships.

People are using social networking sites for personal and professional use, communications, new business developments and contacts, dating and virtual meetings. New communities can be established (like in Orkut.com or Tribes.net).

Individuals are highly motivated to sign up (for example, Orkut invitations have been auctioned off on Ebay) and to increase their visibility within a network, and to get as many people to join their network, driven by vanity (much like much of the original Web sites were).

While Online Social Networking (OSN) itself – as the current success of the existing sites shows – provides motivation for individuals to sign up, the current use possibilities of online Social Networking Sites are rather limited. The sheer availability of social connection information, however, opens up new collaboration possibilities, such as “Link Routing”: the routing of information based on the social connections between people. To exploit this kind of information the relationship information has to be made explicit and accessible. Approaches like the “Friend of a Friend” (FOAF) project [21] – actually predating the current Online Social Networking boom – provide this information in a machine accessible exploitable way.

Furthermore current Online Social Networking sites have another disadvantage:

Maintaining an Online Social Networking site requires a major investment. The main exploitable capital of the OSN sites is the user profile and relationship information – information that can be used e.g., for providing targeted advertisement (see e.g., the privacy statement of Orkut.com). Thus these sites are extremely unlikely to share this information openly or even with each other, which will seriously hamper the development of a next generation collaborative infrastructure. The obvious general solution is to build an Online Social Networking Infrastructure on top of a P2P system, with the following advantages:

- New applications can be easily added to the network
- Profile and user information remains the property of individual users – multiple OSN sites can crawl it, and the user is protected against losing here data if an OSN site shuts down.
- No large investment in a centralized site is necessary, since the cost for maintaining the overall network is shared among all users.

As a result it seems desirable to a) base Online Social Networks on Semantic Web technology, and b) exchange and deploy the social information in P2P infrastructures to resolve the control and ownership issue, resulting in an infrastructure similar to the one described in [20].

4. Development Phases for the Networked Semantic Desktop

We envision three research, development and deployment phases for the Networked Semantic Desktop (see figure 2). In the first phase, Semantic Web, P2P, and Social Networking technologies are developed, researched and partially deployed. In the second stage, we will see convergence: Semantic Web technology is deployed on the Desktop, resulting in the Semantic Desktop. Similarly, Semantic Web technology is incorporated in P2P networks and Social Networking. Once there is a reliable technology available for the technology convergences, the next phase can be tackled: the combination of the three fields Semantic Desktop, Semantic P2P and ontology-driven Social Networking into the Networked Semantic Desktop.

5. Conclusion

In this paper, we described our vision of how different, currently very active research fields will interact and co-evolve, resulting in a new internet-based group collaboration infrastructure we called the “Networked Semantic Desktop”, which will let individuals collaborate at a much finer-grained level as is possible with the Windows-style desktop today, and will result in dramatic time savings in filtering out marginal information and discovering vital information.

6 . References

[1] Dennis Quan, David Huynh, and David R. Karger: Haystack: A Platform for Authoring End User Semantic Web Applications. International Semantic Web Conference 2003: 738-753

[2] The Gnowsis project (visited April 2004): URL: http://www.gnowsis.com/wiki

[3] The FenFire project (visited April 2004): URL: http://www.nongnu.org/fenfire/

[4] Hans Reiser: Future Vision (visited April 2004): URL: http://www.namesys.com/whitepaper.html

[5] Mario T. Schlosser, Michael Sintek, Stefan Decker, and Wolfgang Nejdl: A Scalable and Ontology-Based P2P Infrastructure for Semantic Web Services. Peer-to-Peer Computing 2002: 104-111

[6] Mario Schlosser, Michael Sintek, Stefan Decker, and Wolfgang Nejdl: HyperCuP - Hypercubes, Ontologies and P2P Networks.Springer Lecture Notes on Computer Science, Vol. 2530 – Agents and Peer-to-Peer Systems

[7] Wolfgang Nejdl, Boris Wolf, Changtao Qu, Stefan Decker, Michael Sintek, Ambjörn Naeve, Mikael Nilsson, Matthias Palmér, and Tore Risch: EDUTELLA: a P2P networking infrastructure based on RDF. In 11th World Wide Web Conference, May 2002: 604-615

[8] Min Cai and Martin Frank. RDFPeers: A Scalable Distributed RDF Repository based on A Structured Peer-to- Peer Network, University of Southern California - Computer Science Technical Report 03-807, December 2003 (also accepted for the Semantic Web track of the 2004 World Wide Web conference -- this is the version of the paper as it was submitted).

[9] A. Castano, S. Ferrara, S. Montanelli, E. Pagani, G.P. Rossi (University of Milan): Ontology-Addressable Contents in P2P Networks. In: 1st Workshop on 1st Workshop on Semantics in Peer-to-Peer and Grid Computing at the Twelfth International World Wide Web Conference 20 May 2003, Budapest, Hungary. See: http://www.isi.edu/~stefan/SemPGRID/

[10] Arturo Crespo and Hector Garcia-Molina. Semantic Overlay Networks, Submitted for Publication, 2002. See: http://www-db.stanford.edu/peers/

[11] W. Nejdl, M. Wolpers, W. Siberski, C. Schmitz, M. Schlosser, I. Brunkhorst, and A. Lser. Super-peer-based routing and clustering strategies for RDF-based peer-to-peer networks. In 12th World Wide Web Conference, May 2003.

[12] Karl Aberer, Philippe Cudré-Mauroux, and Manfred Hauswirth: Start making sense: The Chatty Web approach for global semantic agreements. In: Journal of Web Semantics (Elsevier) 1(1), 2004.

[13] Rael Dornfest and Dan Brickley: The Power of Metadata. http://www.openp2p.com/pub/a/p2p/2001/01/18/metadata.html, excerpted from the book "Peer-to-Peer: Harnessing the Power of Disruptive Technologies (Jan. 2001).

[14] Mayank Bawa, Brian Cooper, Arturo Crespo, Neil Daswani, Prasanna Ganesan, Hector Garcia-Molina, Sepandar D. Kamvar, Sergio Marti, Mario T. Schlosser, Qi Sun, Patrick Vinograd, and Beverly Yang: Peer-to-peer research at Stanford. SIGMOD Record 32(3): 23-28 (2003)

[15] Groove: URL: http://www.groove.net/ (Visited April 2004).

[16] Alon Y. Halevy, Zachary G. Ives, Peter Mork, and Igor Tatarinov: Piazza: data management infrastructure for semantic web applications. WWW 2003: 556-567

[17] Yee, K-P., Swearingen, K., Li, K., and Hearst, M., Faceted Metadata for Image Search and Browsing, in the CHI 2003 Conference on Human Factors in Computing Systems.

[18] Karl Aberer, Stefan Decker, David De Roure, Carole Goble (eds.): 1st Workshop on Semantics in Peer-to-Peer and Grid Computing at the Twelfth International World Wide Web Conference 20 May 2003, Budapest, Hungay

[19] Karl Aberer, Stefan Decker, David De Roure, Carole Goble, Hongsuda Tangmunarunkit (eds.): The Second Workshop on Semantics in Peer-to-Peer and Grid Computing at the Thirteenth International World Wide Web Conference 18 May 2004, New York, USA.

[20] Ken Jordan, Jan Hauser, and Steven Foster: The Augmented Social Network: Building identity and trust into the next-generation Internet. In: First Monday, volume 8, number 8 (August 2003), URL: http://firstmonday.org/issues/issue8_8/jordan/index

[21] The Friend of a Friend (FOAF) project. URL: http://www.foaf-project.org/. Visited: April 2004.