

A Model for Document Processing in Semantic Desktop Systems

Ola Leifler

(Linköping University, Sweden
olale@ida.liu.se)

Henrik Eriksson

(Linköping University, Sweden
her@ida.liu.se)

Abstract: There is a significant gap between the services provided by dedicated information systems and general desktop systems for document communication and preparation. This situation is a serious knowledge-management problem, which often results in information loss, poor communication, and confusion among users. Semantic desktops promise to bring knowledge-based services to common desktop applications and, ultimately, to support knowledge management by adding advanced functionality to familiar computing environments. By custom tailoring these systems to different application domains, it is possible to provide dedicated services that assist users in combining document handling and communication with structured workflow processes and the services provided by dedicated systems. This paper presents a model for developing custom-tailored document processing for semantic-desktop systems. Our approach has been applied to the domain of military command and control, which is based on highly-structured document-driven processes.

Key Words: semantic desktop, document-driven processes, semantic documents, planning

Category: H.5.3, H.5.4, I.7.1, I.7.5, M.1, M.4

1 Introduction

Information handling and knowledge communication in dedicated application systems, such as accounting and planning systems, are typically designed to support users in achieving common organizational goals. Normally, such dedicated information systems assume well-defined information processes and workflows. Conversely, there are generic desktop applications that support users in a wide variety of tasks such as communication via e-mail, document preparation through word processors and information navigation through web browsers. Such general desktop systems do not make any commitments to a particular work process.

Unfortunately, there is a significant gap between dedicated information systems and common desktop applications for everyday work. Often, users must communicate outside dedicated information systems, for example when preparing and sending documents to one another. In case users have needs for communicating or storing information that cannot be managed within dedicated

information systems, users resort to generic desktop systems. In such situations, it is difficult to keep the activities in dedicated and generic systems synchronized. In document-driven activities such as planning and reporting, where document preparation and use are core activities, it is especially important to bridge the core process of using dedicated support systems with the document authoring task and to keep the document flow manageable.

Important elements of knowledge management are document preparation, communication, archival, and retrieval. Many knowledge-intensive activities are document-driven in the sense that they focus on document authoring and document use to support human tasks such as planning, decision making, and information dissemination. Today, users take advantage of computer-based office programs such as Microsoft office and OpenOffice to create and edit documents. However, these document-processing environments mainly provide support for text-based editing tasks rather than semantic services to support knowledge management.

Semantic desktops are inspired by the semantic web in the sense that the semantic desktop brings semantic-web technologies to the user's desktop [Cheyer et al., 2006, Sauermann et al., 2006, Dong and Halevy, 2005]. A principal idea behind the semantic desktop approach is to support working and reasoning with semantic entities that are normally scattered across several different resources. One of the advantages of semantic desktops is that they promote both formal and informal work processes and flexible information flows. Furthermore, these systems avoid the rigid structures sometimes imposed by traditional dedicated information systems. However, without in-depth knowledge of the application domain (i.e., specific information about the content, structure, and purpose of the documents), it is difficult to provide semantic services beyond general document indexing and search.

Our approach is to extend a pre-existing semantic desktop with domain-specific functionality that enables relevant document analysis based on both the document structure (e.g., outline, tables, and diagrams) and textual content (e.g., keywords, terms, and distinguishing phrases). The development of such semantic-desktop extensions includes modeling of the document workflow, definition of domain concepts in ontologies, and implementation of document-analysis components. The implementation of extensions to a semantic desktop can involve the addition of domain-specific document tracking, indexing and visualization. Naturally, it is necessary to precede this extension development with a traditional requirements analysis identifying the relevant services that the environment should provide.

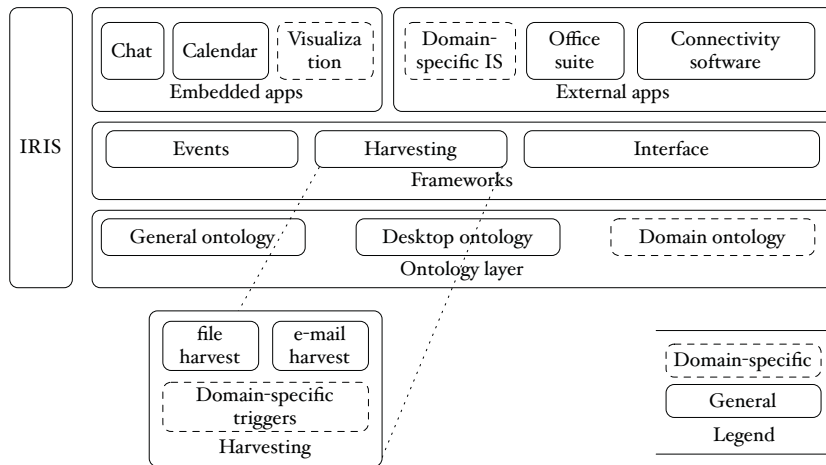


Figure 1: The IRIS semantic desktop model as a three-layered architecture with the ontology at the bottom, the frameworks used to interact with the ontology in the middle and applications at the top.

2 Background: Semantic desktop systems

Semantic Desktops [Cheyer et al., 2006, Sauermann et al., 2006] introduce technologies from research on the *semantic web* [Berners-Lee et al., 2001] to the computer desktop with the goal to empower users through improved information-management techniques. The semantic-desktop effort aims at managing information more intelligently through the use of powerful logical formalisms [Baader et al., 2003] for reasoning about semantic entities on the desktop.

Figure 1 presents an overview of the semantic-desktop model. It is based on the notion of an underlying formal ontology, which contains concepts and relationships that pertain to the use of a computer desktop environment. The concepts in the ontology describe interacting with applications, opening and editing files, reading e-mails etc. Also, it contains concepts that relate to semantic entities people refer to in their daily work, and which is what the semantic desktop is built to support. To give an example of such a semantic entity, we often refer to a *person*, although on a computer desktop, a person may be manifested in e-mails as the sender or recipient, in documents as the author, and in calendars as the participant in an event. All these application-specific references are identifiers of the same semantic object, however.

IRIS is a semantic-desktop environment [Cheyer et al., 2006] that contains functionality for harvesting references such as people, tasks and other common concepts from desktop resources. However, the flow of information between parts

of an organization may contain much richer semantics than that which can be described using such general concepts. Especially in highly-regulated activities, such as military command and control and medicine, additional information can be inferred from the context of work.

Like many other semantic-desktop environments, IRIS is an extensible framework where functionality can be added at all levels. To support the use of dedicated information systems along with generic desktop applications, the environment needs extensions primarily in the layers concerned with managing the ontology and ensuring that the semantics in the document flow is merged with information from dedicated information systems. The semantics of documents can be retrieved by injecting a context-aware IRIS plug-in that responds to events on the semantic desktop, such as the arrival and submission of documents via e-mail, and the creation and modification of documents. Together with a workflow model based on domain-specific documents, we propose to use semantic desktops as the foundation for new methods of merging and reasoning with information sources.

3 Document workflow modeling

The document flow within an organization is an important part of the knowledge-management process. Figure 2 illustrates the information flow between two levels of management. Both levels of management work in parallel to evaluate and plan activities, as well as monitor their progress and coordinate common resources. One way to characterize the interaction and workflow between these levels of management is through the documents that are created, modified, and accessed as part of the process. All traceable actions can be described as communication acts with specific senders and recipients within the organization, or with specific documents as attachments to e-mail communication. Moreover, for each specific domain, the contents of the documents communicated can be described in more detail, such as the character of the *situation brief* submitted by middle management. In the command-and-control domain, such situation briefs can contain descriptions of the current restrictions related to geographical, logistical, legal, and temporal conditions as well as medical restrictions. Also, the directive produced from upper management may contain a decision table with references to objective specifications, due dates for individual tasks, and allocated resources.

In domains where regulations or prevailing practice stipulate a well-defined structure for documents and communications, along with current trends towards standardized formats for documents such as OpenDocument and the proliferation of powerful toolkits for information extraction [Cunningham et al., 2002], a semantic desktop can be augmented with functionality that takes advantage of this structure and uses it to facilitate consistent and efficient use of information.

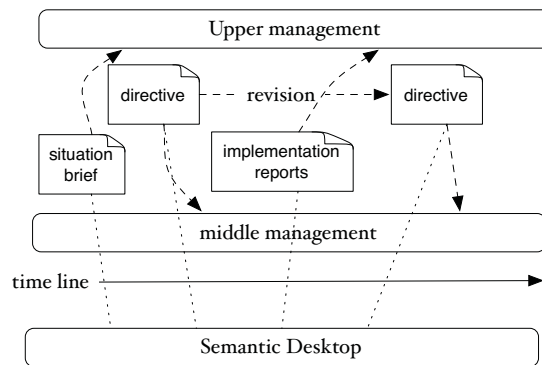


Figure 2: Document flow as part of the collaboration between two levels of management. The documents involved in this communication carry auxiliary information, such as specific file names, structured content, format, author, sender, and recipients. This type of information can assist the information-extraction process.

4 Domain-specific semantic desktop information management

Recent standards in document formats (e.g., the OpenDocument and Office Open XML formats) along with advances in natural-language processing simplify the inspection of generic desktop documents. Today, it is a relatively straightforward task to recognize the structure of a document and use that structure to harvest information on its contents. In addition, natural-language-processing tools can classify words and phrases as being references to specific semantic entities, such as people and locations. Together, these advances corroborate the case for the semantic desktop as a viable support tool for knowledge management.

The use of domain-specific information can support users with reconciliation of resources, efficient information navigation, and improved communication in the following manner:

- *Reconciliation of resources* can be performed using the mechanisms already in place in semantic desktops for reifying information [Dong and Halevy, 2005], with the only addition that the source of the information must originate from the dedicated information system as well as text documents. The IRIS environment in particular is designed to act as an open platform that merges semantic references from other applications and other computers through the use of web services.
- *Efficient information navigation* can be supported by IRIS through generic views that present information about the objects on the desktop, such as

people and tasks. However, because IRIS is an open platform, it is possible to add views for displaying and navigating among references to locations, resources, teams, and so on. Furthermore, developers can connect these views to the ontology via specific events that inform each view of updates to the specific kind of semantic entities that the view manages.

- *Improved communication* can also be supported by the semantic desktop by using the shared knowledge base. The knowledge base is set up to be shared over a network, along with an event-driven architecture that provides users with notifications of changes or additions to specific categories of semantic entities.

5 Implementation for command and control

To demonstrate its viability, we have successfully used our approach in a military command-and-control scenario, where we harvest domain-specific location references from standard OpenOffice documents and provide navigation among those references via a map-based user interface that supports information navigation. As another example, we have identified documents with a specific structure and purpose¹, and extracted domain-specific references to tasks in those documents, based on the structure of the document and the existence of a tabular definition of tasks and responsibilities, to reason about temporal dependencies between tasks. Furthermore, we are currently employing our approach to facilitate communication analysis by modeling the transactions between members of staff.

6 Discussion

The use of semantic-desktop environments promotes interoperability across organizations because most of the information interchange is based on e-mail and standard document formats rather than dedicated protocols. This design enables tracking and analysis of documents from outside sources and collaboration partners. In a way, the documents act as user interfaces to the tracking and analysis programs because they affect the actions of these systems.

In this work, we have addressed the analysis of standard documents. However, we believe that it is possible to augment documents with additional metadata to achieve *semantic documents* [Eriksson and Bång, 2006, Eriksson, 2007]. Such semantic document can contain ontologies with concepts that are linked to words, sentences, paragraphs, and other parts of the document. It is possible to add this information to standard document formats, such as PDF, to allow analysis programs to access metadata directly without extracting them from the document text. For example, by representing the content of decision tables in the

¹ to synchronize the use of resources over time during a mission

documents as ontology classes and instances, a semantic-desktop environment could compare documents directly by performing ontology matching. In other words, semantic documents have the potential of retaining semantic information available at the time of document preparation.

A long-term goal for domain-specific semantic-desktop approach is to make it available to a broad range of applications by lowering the effort to custom tailor the environment. In the area of knowledge acquisition, researchers have developed metatools for instantiating domain-specific knowledge-acquisition tools from high-level descriptions [Gennari et al., 2002]. A similar approach could potentially be used to create domain-specific extensions for semantic-desktop environments. For example, we believe that a meta-level tool could generate a domain-specific extension from a combination of domain ontologies and document templates. Naturally, there are many design alternatives for making domain-specific commitments in future generalized semantic-desktop environments and for developers to specify the domain-specific aspects.

Our evaluation of semantic desktops as a viable approach for supporting document-driven staff work has currently been conducted through technical implementation work, based on analyses from participating in and observing actual staff work. The results of our work indicates that semantic desktops can be used successfully to extract and reason about information critical to commanders and represented as parts of the contents or the structure of documents. In military staff exercises, we have identified issues related to the management of temporal and spatial information in documents that was a major cause of concern. Since effective sense making is crucial to a staff, and since that process implies primarily interpreting information jointly, means for organizing the desktop content according to semantic entities instead of the physical structure of documents would make valuable contributions to the work environment of commanders.

7 Conclusion

Semantic desktop systems have the potential to enhance document-driven knowledge-management processes through more effective management of semantic entities in documents and communications. One of the main advantages of the semantic-desktop approach is that it supports the users in their daily activities without introducing traditional application systems that require separate streams of information handling. To maximize the benefit of semantic desktops, however, we believe that it is necessary to adapt these systems to the knowledge-management environment of the users and the tasks that the users are performing daily.

A remaining challenge for domain-specific semantic desktops is to find ways to customize the system to the documents and work patterns particular to the

application area. Specifically, it is important to streamline the adaptation of document analysis for new document types.

Acknowledgments

This work was supported by the Swedish National Defence College.

References

- [Baader et al., 2003] [Baader et al., 2003] Baader, F., Calvanese, D., McGuinness, D., Nardi, D., and Patel-Schneider, P., editors (2003). *The Description Logic Handbook*. Cambridge University Press.
- [Berners-Lee et al., 2001] [Berners-Lee et al., 2001] Berners-Lee, T., Hendler, J., and Lassila, O. (2001). The semantic web. *Scientific American*, 284(5):28–37.
- [Cheyer et al., 2006] [Cheyer et al., 2006] Cheyer, A., Park, J., and Giuli, R. (2006). IRIS: Integrate. relate. infer. share. In *Proceedings of the Semantic Desktop and Social Semantic Collaboration Workshop (SemDesk-2006)*.
- [Cunningham et al., 2002] [Cunningham et al., 2002] Cunningham, H., Maynard, D., Bontcheva, K., and Tablan, V. (2002). GATE: A framework and graphical development environment for robust NLP tools and applications. In *Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics*.
- [Dong and Halevy, 2005] [Dong and Halevy, 2005] Dong, X. L. and Halevy, A. (2005). A platform for personal information management and integration. In *Proceedings of the Second Biennial Conference on Innovative Data Systems Research*.
- [Eriksson, 2007] [Eriksson, 2007] Eriksson, H. (2007). The semantic document approach to combining documents and ontologies. *International Journal of Human-Computer Studies*, 65(7):624–639.
- [Eriksson and Bång, 2006] [Eriksson and Bång, 2006] Eriksson, H. and Bång, M. (2006). Towards document repositories based on semantic documents. In *Proceedings of the Sixth Conference on Knowledge Management*, pages 313–320, Graz, Austria.
- [Gennari et al., 2002] [Gennari et al., 2002] Gennari, J. H., Musen, M. A., Ferguson, R. W., Grosso, W. E., Crubézy, M., Eriksson, H., Noy, N. F., and Tu, S. W. (2002). The evolution of Protégé: an environment for knowledge-based systems development. *International Journal of Human-Computer Studies*, 58:89–123.
- [Sauermaann et al., 2006] [Sauermaann et al., 2006] Sauermaann, L., Grimnes, G. A., Kiesel, M., Fluit, C., Maus, H., Heim, D., Nadeem, D., Horak, B., and Dengel, A. (2006). Semantic Desktop 2.0: The Gnows Experience. In Cruz, I., editor, *Proceedings of the International Semantic Web Conference (ISWC 2006)*, volume 4273 of *Lecture Notes in Computer Science*, pages 887–900. Springer Verlag.