

# Modeling Provenance for Semantic Desktop Applications

A. Marins, M. A. Casanova, A. Furtado, K. Breitman

Departamento de Informática – Pontifícia Universidade Católica do Rio de Janeiro  
Rua Marquês de S. Vicente, 225 – Rio de Janeiro, RJ – Brazil – CEP 22451-900

{amarins,casanova,furtado,karin}@inf.puc-rio.br

***Abstract:** As the volumes of digital resources grow exponentially, users face the threat of information overload. Almost everything we see, read, hear, write and measure is collected and made available via computational information systems (Carvalho et al. 2006). The problem is not so much finding information, but rather, developing computational solutions that help manage digital data in a meaningful way. In this paper we tackle this problem from the user's perspective. We explore Semantic Desktop applications, which combine ontologies, taxonomies, and metadata in general to enhance information management and help reduce the difficulty of locating data stored in personal computers. We argue that such applications would benefit if endowed with the ability to autonomously harvest provenance metadata and index content accordingly and propose a generic provenance model for this purpose.*

## 1 Introduction

In September 2006, Google search crawler dealt with 850 TB of raw data from the Web. Taking into account that the crawler has a compression rate of 11%, we are talking of 8.5 Peta Bytes of information available on line (Chang et al. 2006). Apart from radio and TV broadcasts, that are not fully digitalized yet, one can find/buy just about any information in digital format. The mailing list business has scaled up to the selling of credit card lists (over 3 billion numbers), criminal files (100 million), the name and address of every voting Mexican, ID and phone number of every Argentinean, as advertised by companies such as LexisNexis and Choice Point (Gaspari 2005).

Because the amount of available digital data is growing at such an exponential rate, it is becoming virtually impossible for human beings to manage the complexity and volumes of available information. The danger is to create “write only” databases, in which information is constantly stored, but impossible to be mined for meaningful purposes. This phenomenon, often referred to as information overload, poses a serious threat to the very usefulness of today's Web.

New abstractions are required, to help model and summarize massive data volumes to a more humane dimension. Researchers from industry, government, and academia are now exploring the possibility of creating a Semantic Web in which meaning is made explicit, allowing machines, as opposed of humans, to process and integrate data resources intelligently (Breitman et al. 2007).

In this paper we argue in favor of semantic desktops, which are applications that combine ontologies, taxonomies, and metadata in general to enhance information management and help reduce information overload in personal computer environments. The rest of this paper is organized as follows. In Section 2, we briefly introduce semantic desktop applications. In Section 3, we summarize the provenance concepts need throughout the paper. In Section 4, we introduce our provenance model. In Section 5, we outline a provenance-ready semantic desktop architecture. Finally, in Section 6, we present our conclusions.

## 2 Semantic Desktop Applications

The desktop metaphor helps individuals manage data stored in their personal computers. It is the single point of entry to the wide set of applications designed to access, manage and deliver personal or corporate digital objects. To provide users with the ability to organize their information resources in ways that suit their individual needs, while maintaining semantic interoperability with other applications researchers are investigating the development of semantic desktop applications (Sauermann et al. 2005, (Brunkhorst et al. 2006; Chirita et al. 2006, Quan et al. 2003). The set of methods, data structures, and tools that extend the traditional desktop metaphor, giving data a well-defined meaning, represent what we call the semantic desktop. Furthermore, a semantic desktop that enables the exchange of data across individual boundaries thereby improving online collaboration and helping organizing data created by a group of users will be referred to as a ‘social’ semantic desktop (Chernov et al. 2006).

In this paper we are interested in automatic ways to harvest the semantics of data items. In this light, we argue that relevant information can be obtained by tracing the origins of the items<sup>1</sup> themselves, i.e., capturing its provenance metadata. Typical provenance metadata includes relevant relationships with facts, people, publications, etc., as well as the interaction history of the data item. Today's desktop applications offer innumerable opportunities to capture useful provenance metadata. For instance, when an e-mail attachment is saved to the hard disk, information about the connection between that file and who sent can be stored as provenance metadata for the file. Also, when opening the browser through a link inside a message body from a trusted e-mail sender and downloading a file from this Web page, also provides a rich set of provenance metadata for the file: the fact that the saved file came from the Web page, and why the file was saved. These two simple examples illustrate why desktop applications should be instrumented to save and store provenance metadata, which will then be used by (semantic) desktop search applications to help users locate data stored in their personal computers.

Provenance metadata has been explored in a wide range of application areas. For example, the CIDOC *Conceptual Reference Model* (CRM) (CIDOC 2006), published by the International Committee for Documentation of the International Council of Museums (ICOM-CIDOC), provides definitions and a formal structure for describing

---

<sup>1</sup> We use the term data item in a very broad sense. It includes any multimedia item that can be referenced by a URI (Uniform Resource Identifier). The concept of an URI is fundamental for understanding the Semantic Web as a distributed, federated information space, because it provides an addressing scheme that is stable, distributed, and effective.

the concepts and relationships used in cultural heritage documentation. It has been accepted as a working draft by the ISO/TC46/SC4/WG9 in September 2000 and is currently in the final stages of becoming a standard, known as the ISO 21127:2006 (ISO 2006). A closer look at the CIDOC CRM reveals that the model is entirely based on the concept of provenance. Ram (2005) investigates the use of provenance concepts in the context of product design. The provenance model we introduce in this paper combines concepts from both of these references. We also draw provenance concepts from the PASOA Project (Groth et al. 2006).

Sauermann et al. (2005) and Breitman et al. (2007) survey semantic desktop projects. Gnowsiss (Sauermann and Schwarz 2004), Haystack (Quan et al. 2003) and Beagle++ (Brunkhorst et al. 2006; Chirita et al. 2006), which resulted from the European IST Project NEPOMUK, are examples of semantic desktop systems. They all explore Semantic Web technologies to store data semantics in a knowledge base, which is later on queried to help the user locate data. However, none of these systems explore provenance concepts to guide the design of the knowledge base.

The major contribution of this paper lies exactly in defining a provenance model, cast as a OWL ontology, that provides an uniform way to model knowledge bases that support semantic desktop applications.

### 3 Understanding Provenance

#### 3.1 Definition

Firstly we introduce the intuitive definition of provenance. Its etymology is rooted in the French verb '*provenir*', i.e., to come forth, originate. The Oxford English Dictionary provides the following definition for the word provenance:

**Definition 3.1.1:** (i) the fact of coming from some particular source or quarter; origin, derivation. (ii) the history or pedigree of a work of art, manuscript, rare book, etc.; concr., a record of the ultimate derivation and passage of an item through its various owners.

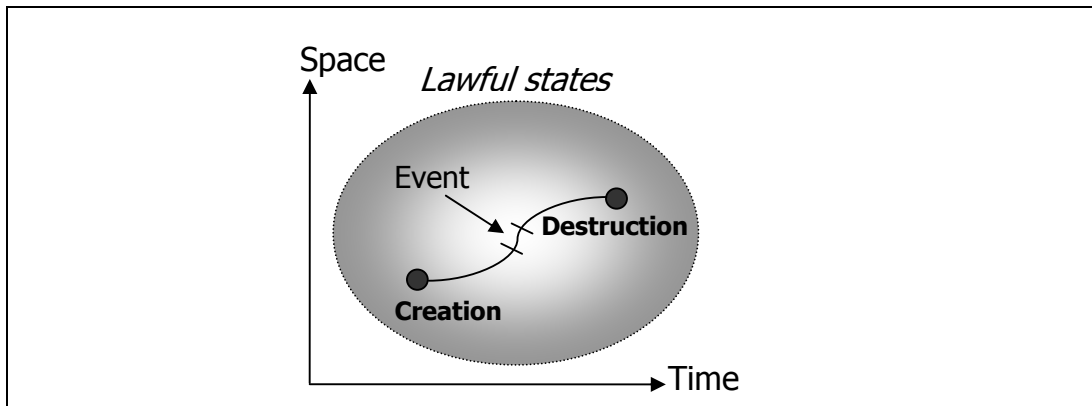
The Merriam-Webster Online Dictionary defines provenance as follows:

**Definition 3.1.2:** (i) the origin, source; (ii) the history of ownership of a valued object or work of art or literature.

Both definitions are compatible as they regard provenance as the derivation from a particular source to a specific state of an item. We further clarify the concept of provenance by investigating how to map History into provenance questions. According to Bunge (1977), History is based on evidence or documentation of events that occurred in the past, as illustrated in Figure 1. Central to Bunge's theory are the notions of event, space, time, action and agent. To facilitate their elicitation, the author suggests the use of the following questions:

- *Who* are the creators, publishers or contributors? Who is the sender? Who has permissions, and what are them? Is this entity trustable?
- *When* was the data created, accessed, modified? This question is linked to who has accessed or modified the data.

- *Where* was the data first reported or, complementarily, where is it stored, or where are the locations, if there are more than one copy.
- *how* has the data been derived or transformed. This question is related to which procedures or computations were applied to transform the data.
- *Which* applications, software configuration or tools' settings were in use when the data was created. This question is connected to environmental conditions.
- *What* is the data? Is it a creation, a transformation, a derivation, a management or a destruction?



**Figure 1. Bunge's view of history (Ram 2005)**

The questions serve to elicit the central provenance notions, according to Bunge (1977), i.e., event, space, time action and agent. The mappings are intuitive and unary, as depicted in Table 1.

**Table 1. Mapping between Bunge's notions and provenance questions**

Notion	Question
Event	What
Space	Where
Time	When
Action	How
Agent	Who, Which

To complement provenance semantics it is mandatory to capture decision rationale concepts: beliefs, desires and intentions. All of them are significant factors that affects decision making and can be represented using the Belief-Desire-Intention Model (Georgeff 1999). Beliefs represent knowledge of the world, desires are goals assigned to the agent and intentions are commitments by an agent to achieve particular goals. It is

crucial to trace the sequence of ideas and list of hypotheses related to the data, to capture Why provenance that explains why the data is being created.

To illustrate the question based elicitation process let's consider the development of a software package that provides dynamic analysis of vessels. Its purpose is to allow the dynamic simulation of the behavior of vessels subject to waves, winds and currents.

The development team considered adopting an ocean simulation API, which turned out to be unqualified for the project since the vendor of the API does not belong to a list of previously approved vendors. The API was developed outside of Brazil, and its provider released its first production version only 1 year ago. In this case the use of Who, Where and When questions would play a fundamental role in determining data quality and vendor reliability.

Suppose now that the virtual vessels designers' team discovered a new navigation design pattern. As the pattern was fairly new, it was not clear how to use it. At a given point the team realized that another group used that pattern in a different project (which had similar software requirements). The team referred to this project provenance record to find out how and why the pattern has been used and what were the lessons learned from this previous experience. It is clear that Who, hoW and Why questions would facilitate data reuse and sharing of this point.

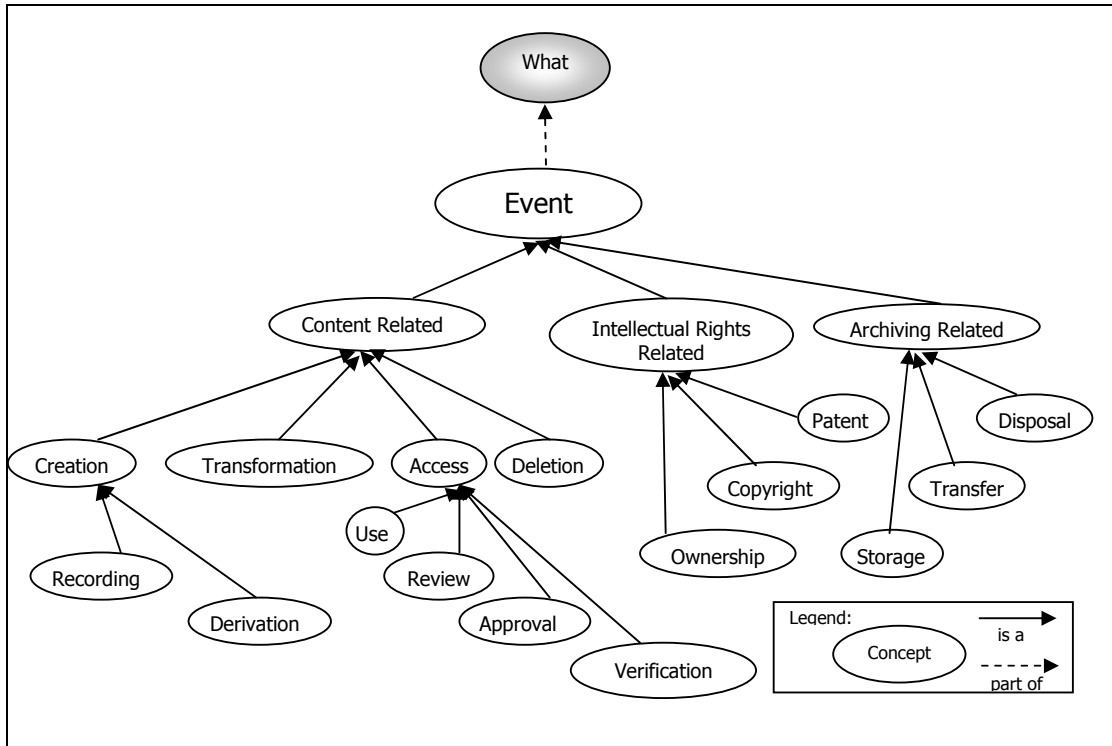
Now consider that two senior analysts were responsible for testing the new API. They both performed the same test on several use cases. The first analyst explored the API by simulating middle size waves over a long period, while the second computed it by recording results for giant waves over small periods. This scenario puts in evidence the necessity of recording the rationale of the processes, using the hoW question to capture both analysts' derivation procedures. In case that record went missing, there would be no way to compare the procedures.

Finally, imagine that a database analyst detected a serious performance issue in a persistence class, and confirmed that the error might have existed since October 2005. He or she would want to locate all data related to projects that have been developed using this class since then. If we are able to track *Which* and *When*, it would be possible to easily identify the projects that needed to be recalled. It is clear that a major issue in provenance harvesting is the order in which events take place. In the next section we explore provenance from a lifecycle angle.

### **3.2 Provenance Lifecycle**

We anchor provenance lifecycle on the event notion. The CIDOC CRM (CIDOC 2006) defines event as "something (*What*) that happens in space and time and brings about some change in the world". An event can involve people, objects in the world and ideas (concepts). In addition, an event occurs in a time frame, i.e, has a measurable duration; and occurs in space (Boeuf 2006).

The provenance lifecycle begins with the, often automated, capture of events from users/programs interactions, followed by monitoring and analysis of information from different sources in different formats. Figure 2 adapted from (Ram 2005), illustrates the information lifecycle of events in the context of new product design and development.



**Figure 2. Information Lifecycle Events**

The provenance lifecycle can be modeled as a four phase process: Create, Record, Query and Manage (Groth et al, 2006a). The first phase covers semantic discovery and creation.

The second phase focus on the storage of provenance metadata for future use. The long-term strategy for storing provenance metadata demands a component, called a *provenance store* (Groth et al, 2006a), that provides persistent storage, management and access of provenance metadata.

The third phase comprehends the querying of the provenance store by users or applications to obtain provenance metadata. At the most basic level, the result of a query is just provenance metadata. Advanced query facilities may return more sophisticate information about the data, derived from the provenance metadata.

Finally, the fourth phase covers management of the provenance store to handle: provenance archival, deletion and disposal, to maintain the metadata synchronized with the data. A provenance model should support all these four phases, as explored in the next section.

## 4 Provenance Model

In view of the discussion in the previous, we propose a provenance model based on the W7<sup>2</sup> model (Ram 2006) and on parts of the CIDOC CRM, and formalized as an OWL ontology.

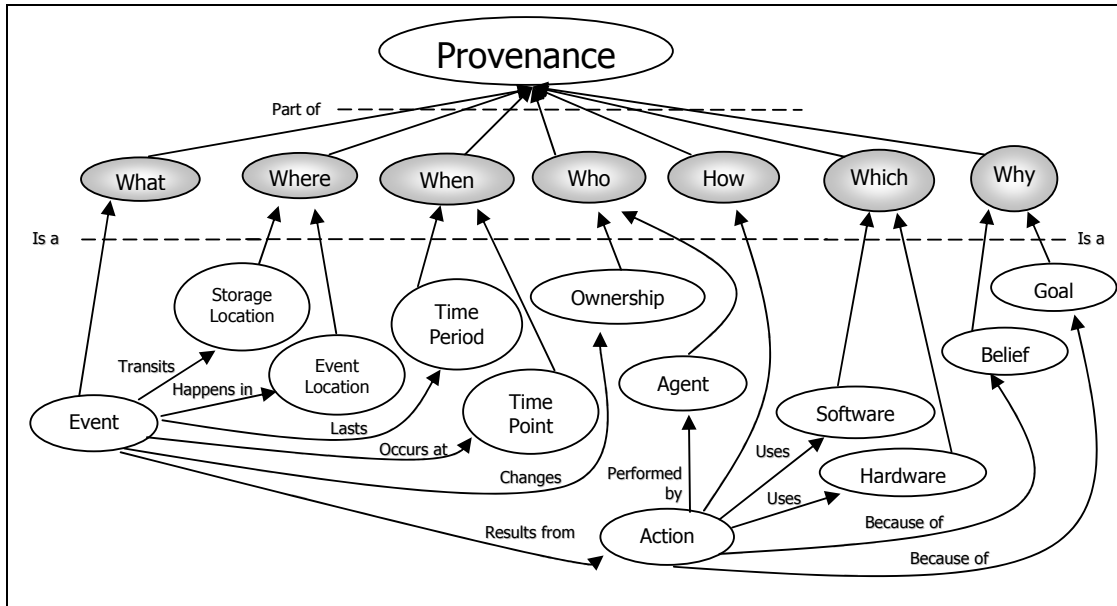


Figure 3. Provenance Model

For the sake of brevity, Figure 3 is restricted to the RDF graph of the ontology, where:

- classes are shown in ovals and object properties are shown as arrows;
- classes are stratified into three levels:
  - the top level contains just the Provenance class;
  - the mid level contains the classes that capture the W-Questions;
  - the bottom layer contains the classes that capture provenance details;
    - the Event, EventLocation, TimePeriod, TimePoint, Action and Agent classes are imported from the CIDOC CRM;
    - all other classes will typically be specialized to the specific environment (hardware, software and users).

Our model is based on a small number of unifying concepts and can be a good generalization for other models. It is not directly tied to any standards, technologies or other concrete implementation details and it does seek to provide a common semantics that can be used unambiguously across and between different implementations.

<sup>2</sup> We consider hoW as one of the Ws. We can interpret W7 model as a superset of the popularly known 5W and 1H.

(Barbosa et al 2007) argue that analogy mappings facilitate conceptual modeling by allowing the designer to reinterpret fragments of familiar conceptual models into other contexts. A good example of conceptual modeling by analogy and metaphor (Barbosa et al 2007) is the mapping between Design Rationale and Provenance context. Design rationales are usually adopted to capture representation schemas based on arguing because they offer an infrastructure to identify what decisions were made and related reasons. Kuaba's elements presented in Table 2 represent the people involved in a design activity and their respective roles as well as methods, arguments and justifications tied to the decisions and solution ideas (Medeiros 2006).

**Table 2. Analogy between Kuaba's elements and provenance questions**

Element	Question
Decision	What
Duration	When
Activity	How
Method	Which
Person	Who
Argument/Justification	Why
-	Where <sup>3</sup>

Now, querying the provenance model means querying the provenance of one or more objects, which implies that a provenance query will typically have two parts (Miles 2006):

- *query data handle*, that identifies the object(s) of interest;
- *relationship target filter*, that scopes the query results, restricting it to a manageable amount of information returned.

Figure 4 depicts a 360-degree of possible starting points for the target filters. The specific query language will naturally depend on the knowledge base management system adopted.

Suppose the user wants to find the map of a sightseeing action performed at Rodrigo de Freitas Lagoon in Rio de Janeiro and do not have a specific target filter value. Starting from a 360-degree metaphor it is possible to search for this information by fixing other available "axis" like time, space or object. First, use the time frame for the specific trip that returns relevant files. Then filter results and identify other objects available at this time frame. Or fix action, for instance sightseeing, and discover all other relations available. This is similar to the way we recall information that we do not

---

<sup>3</sup> There is no explicit direct Kuaba's element that maps to *Where* provenance question.

remember by exploring all other provenance questions. To accomplish this, the query results needs to show the relevant associations for each object returned.

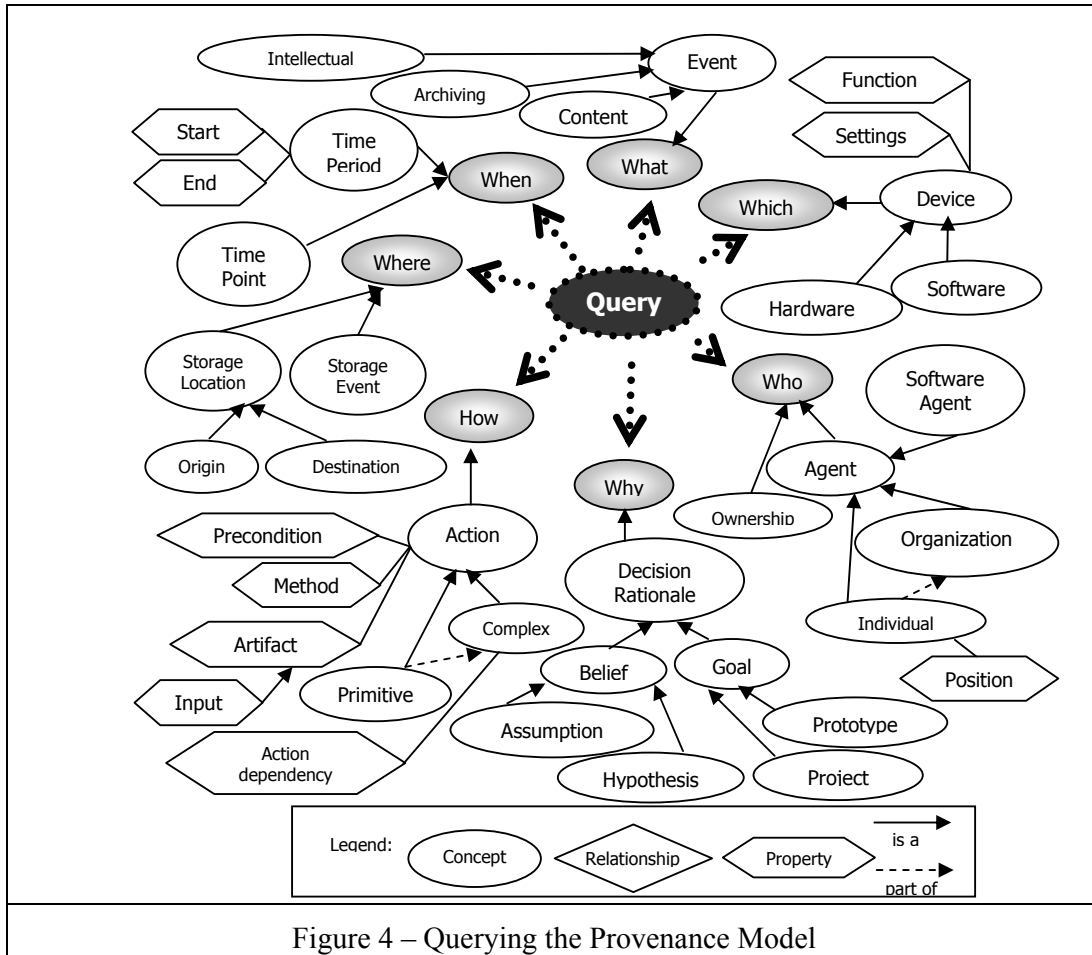
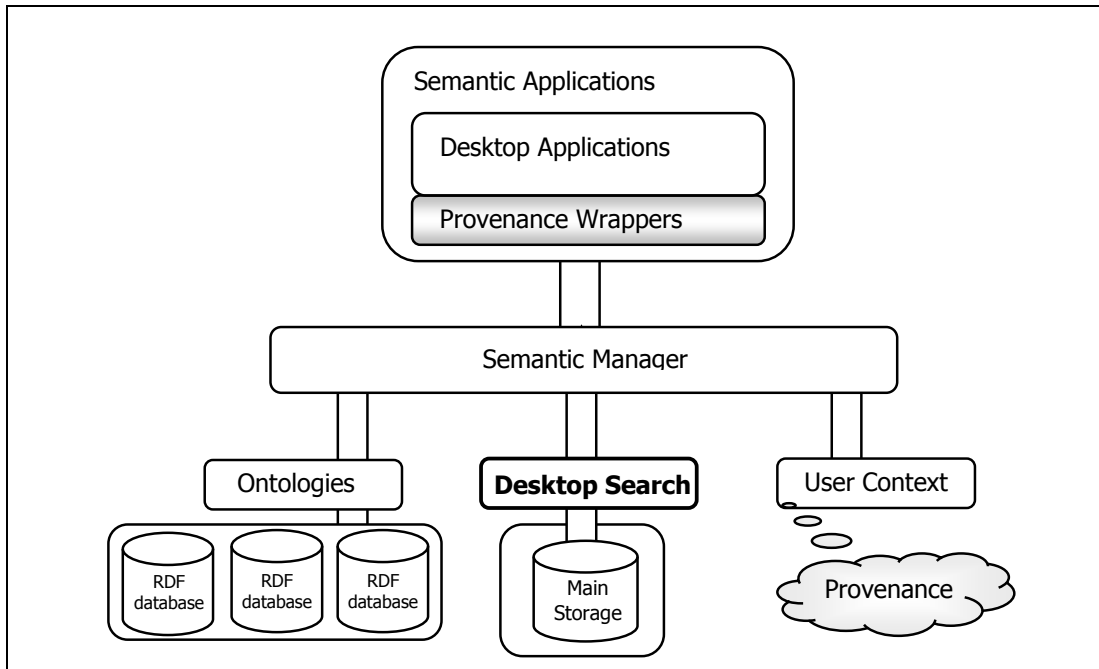


Figure 4 – Querying the Provenance Model

## 5 Provenance in the Semantic Desktop

Ideally, provenance metadata should be harvested from all desktop applications in order to give meaning to stored data. Desktop search is the ground technology that indexes the different types of data stored locally. Figure 5 shows a simplified provenance-ready semantic desktop framework, adapted from (Sauermaun 2005).



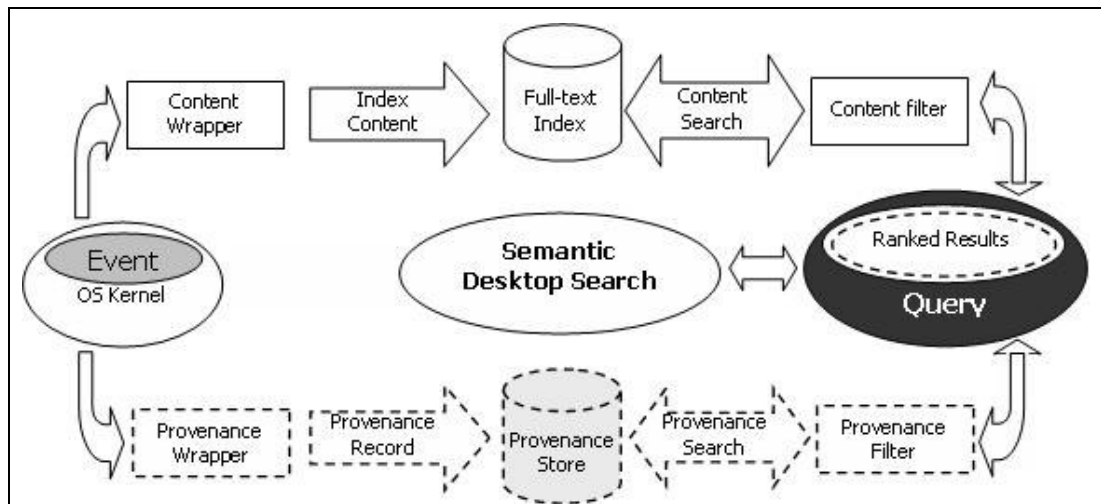
**Figure 5. Simplified Semantic Desktop Architecture**

Semantic Manager will interfaces between provenance wrappers, user context and available domain ontologies to harvest and store provenance information. Then, desktop search component will enable provenance query desktop content and consolidate and rank results using provenance semantics. Semantic applications are desktop applications enriched with provenance capturing and search capabilities.

Several releases of desktop search applications, such as Google Desktop and Beagle, recently emerged to suppress the difficulty of accessing data stored in our computers. Yet they are not provenance-ready. Hence, this work suggests that the metadata storage module, as in Beagle<sup>++</sup> architecture (Brunkhorst et al. 2006), should be re-engineered as a provenance store module, designed according to the provenance model introduced in Section 4. Figure 6 illustrates the provenance-ready process, pointing out the two majors desktop search related flows inspired in the Beagle<sup>++</sup> architecture.

In the middle we have the semantic desktop search component that should be available and integrate with all desktop applications interface. Also, Full-text and Provenance Index components are physically separated because Provenance metadata has different management demands from Content data. Thus, requires specific archiving and disposal actions.

Whenever a new event occurs in the operating system, like a new email saving operation into the main desktop storage, content and provenance wrappers are responsible to index content and store metadata for future search, covering the left one direction flow in Figure 6. The right bi-directional flow represents the user query action defined by content and provenance filters and the new semantic ranked results.



**Figure 6.- Provenance Ready Desktop Search Architecture based on Beagle<sup>++</sup>**

## 6 Conclusions

In this paper we tackled the problem of dealing with massive volumes of distributed multimedia data from the single user's perspective. We briefly described Semantic Desktop applications, which explore emerging Semantic Web technologies to provide a computational solution to what is known as the "information overload problem". In particular, we discussed the use of provenance metadata as a means to improve context-based data recovery and analysis in personal computer environments. A provenance model was briefly described in Section 4, after a discussion about the nature of provenance, in Section 3. Semantic desktop applications, in general, would benefit from the suggested model thereby becoming provenance-ready applications.

The proposed provenance model is being extended to cover questions related to granularity, scalability and security (Braun et al, 2006), which influence the design of the provenance store model. With the overwhelming volume of information made available today through the Web, validity and intellectual property become, more than ever, most pressing issues. They are both listed among the problems to be faced by the Computer Science community, at the end of section 2 of the document (Carvalho et al. 2006). And they raise questions related to provenance: "Can we trust in the validity of a given piece of information? Where does it come from?"

## 7 References

- Boeuf, P. Le (2006) Using an ontology-driven system to integrate museum information and library information. In: Proc. Symposium on Digital Semantic Content across Cultures, Paris, the Louvre, 4-5 May 2006. Available at: <http://www.seco.tkk.fi/events/2006/2006-05-04-websemantique/presentations/articles/>

- Braun, U.; Garfinkel, S.; Holland, A. D.; Muniswamy-Reddy, K.; Seltzer, I. M. (2006) Issues in Automatic Provenance Collection. In: Proc. International Provenance and Annotation Workshop (IPAW'06), Chicago, Illinois, USA May 3-5, 2006.
- Breitman, K.; Casanova, M.A.; Truszkowski, W. (2007) *Semantic Web: Concepts, Technologies and Applications*. Springer. ISBN 1-84628-581-X. Springer-Verlag Heidelberg, 2007.
- Brunkhorst, I.; Chirita, A.; Costache, S.; Gaugaz, J.; Ioannou, E.; Iofciu, T.; Minack, E.; Nejd, W.; Paiu, R. (2006) The Beagle++ Toolbox: Towards an Extendable Desktop Search Architecture. Distributed Systems Institute – Knowledge Based Systems. Technical report, May 2006.
- Bunge, M. (1977) *Treatise on Basic Philosophy: Vol. 3: Ontology I: The Furniture of the World*. Reidel, Boston, MA.
- Carvalho, A.; Brayner, A.; Loureiro, A.; Furtado, A.; von Staa, A.; Lucena, C. J.; Souza, C.; Medeiros, C. M.; Lucchesi, C.; Silva, E.; Wagner, F.; Simon, I.; Wainer, J.; Maldonado, J. C.; Oliveira, J.; Ribeiro, L.; Velho, L.; Gonçalves, L.; Mattoso, M.; Ziviani, N.; Navaux, P.; Torres, R.; Almeida, V. A.; Meira, W.; Kohayakawa, Y. (2006) – “Grand Challenges in Computer Science Research in Brazil – 2006 – 2016” - Workshop Report – Available at <http://www.sbc.org.br>
- Chang, F.; Dean, J.; Ghemawat, S.; Hsieh, W.; Wallach, D.; Burrows, M.; Chandra, T.; Fikes, A.; Gruber, R. (2006) - "Bigtable: A Distributed Storage System for Structured Data" - Proceedings of OSDI'06: Seventh Symposium on Operating System Design and Implementation, Seattle, WA.
- Chernov, S.; Iofciu, T.; Nejd, W.; Zhou, X. (2006) The need for Semantic Desktop Dataset. L3S Research Center and University of Hannover, Germany.
- Chirita, P-A.; Costache, S.; Nejd, W.; Paiu, R. (2006) Beagle++: Semantically Enhanced Searching and Ranking on the Desktop. In Proc. 3rd European Semantic Web Conference.
- Gaspari, E. (2005)– “Vem aí o Estado policial-informático” – Jornal O Globo – April 17, 2005 edition.
- Georgeff, M.; Pell, B.; Pollack, M.; Tambe, M.; Wooldridge, M. (1999) The Belief-Desire-Intention Model of Agency. In Proceedings of the 5th International Workshop on Intelligent Agents.
- Groth, P.; Jiang, S.; Miles, S.; Munroe, S.; Tan V.; Tsasakou, S.; Moreau, L. (2006a) D3.1.1: An Architecture for Provenance Systems. Technical report, University of Southampton, February 2006. Available at: <http://eprints.ecs.soton.ac.uk/12023/>
- Groth, P.; Miles, S.; Munroe, S. (2006b) Principles of High Quality Documentation for Provenance: A Philosophical Discussion. In: Proc. International Provenance and Annotation Workshop (IPAW'06), Chicago, Illinois, USA May 3-5, 2006.
- ISO (2006) Information and documentation - A reference ontology for the interchange of cultural heritage information. Draft International Standard ISO 21127:2006. International Organization for Standardization. Available at: <http://www.niso.org/international/SC4/n500.pdf>
- Medeiros, A (2006) Kuaba: Uma Abordagem para Representação de Design Rationale para o Reuso de Designs baseados em Modelo. Rio de Janeiro, 2006. 149p. PhD Thesis – Departamento de Informática, Pontifícia Universidade Católica do Rio de Janeiro.
- Mills, D. (2006) Semantic Wave 2006 Part-1: Executive Guide to Billion Dollar Markets. Special report from Project 10X.

- Miles, S. (2006) Electronically Querying for the Provenance of Entities. In: Proc. International Provenance and Annotation Workshop (IPAW'06), Chicago, Illinois, USA May 3-5, 2006.
- Moreau, L. (2006) Usage of Provenance – A Tour of Babel. In: Proc. International Provenance and Annotation Workshop (IPAW'06), Chicago, Illinois, USA May 3-5, 2006.
- Quan, D.; Huynh, D.; Karger, D.R. (2003) Haystack: A platform for authoring end user semantic web applications. In: International Semantic Web Conference 2003.
- Ram, S (2005) Investigating Data Provenance in the Context of New Product Design and Development. In: The 2005 National Conference on Digital Government Research (dgo 2005) Available at:  
[www.digitalgovernment.org/library/library/dgo2005/digarch/ram.ppt](http://www.digitalgovernment.org/library/library/dgo2005/digarch/ram.ppt)
- Ram, S (2006) Provenance Project Summary. Available at:  
[kartik.eller.arizona.edu/Provenance\\_Project\\_summary.doc](http://kartik.eller.arizona.edu/Provenance_Project_summary.doc)
- Sauermann, L.; Schwarz, S. (2004) Introducing the Gnowsis Semantic Desktop. In: Proceedings of the International Semantic Web Conference 2004.
- Sauermann, L. (2005) The Semantic Desktop - a Basis for Personal Knowledge. In: Proceedings of I-KNOW '05. Graz, Austria, June 29 - July 1, 2005.
- Sauermann, L.; Bernardi, A.; Dengel, A. (2005) Overview and Outlook on the Semantic Desktop and Outlook on the Semantic Desktop. In: Proceedings of the 1st Workshop on The Semantic Desktop. 4th International Semantic Web Conference (Galway, Ireland), 2005, S. Decker, J. Park, D. Quan, and L. Sauermann (Eds.).