ALETHEIA – TOWARDS A DISTRIBUTED ARCHITECTURE FOR SEMANTIC FEDERATION OF COMPREHENSIVE PRODUCT INFORMATION

Matthias Wauer, Daniel Schuster, Alexander Schill
Technische Universität Dresden, Faculty of Computer Science, Institute of Systems Architecture
Helmholtzstraße 10, 01062 Dresden, Germany
{matthias.wauer, daniel.schuster, alexander.schill}@tu-dresden.de

Johannes Meinecke, Thomas Janke
SAP Research CEC Dresden, Germany
{johannes.meinecke, thomas.janke}@sap.com

ABSTRACT

Having access to relevant product information is a key requirement for improving the efficiency of a business. This paper presents our approach towards an architecture federating information from heterogeneous sources. We explain key design decisions and important functions of the identified components. The approach is determined by requirements from several industry partners and evaluated at an early stage with a first implementation.

KEYWORDS

Information Retrieval, Information Federation, Semantic Web, Information Extraction.

1. INTRODUCTION

Requirements for the management of product information are changing rapidly. The information that arises in the course of a product lifecycle is expected to grow significantly, due to an extended focus on the early and late phases that are considered more and more important (Burkett 2006, Lewis 2005). Furthermore, a large and not yet utilized part of that data is unstructured. Existing information systems are not capable of fully capturing and utilizing this information (Blumberg and Atre 2003).

The Aletheia project (Aletheia 2009) is targeted at semantic federation of heterogeneous product information. It aims to solve this issue by combining information extraction techniques with Semantic Web concepts (Berners-Lee et al. 2001). Information is maintained by different institutions and represented in various data formats and types, which can be classified as structured, semi-structured and unstructured information. By describing and managing this information semantically, Aletheia enables information exchange between different organizational domains. Industry partners provide application scenarios along the product lifecycle, including domains such as automotive, e-commerce, logistics, and industrial maintenance.

The primary research challenge that will be covered by this paper is how to effectively integrate a large number of decentralized information sources across those administrative boundaries. This paper presents our approach towards a service-oriented system architecture (Section 2) that enables efficient queries on distributed heterogeneous product information. The concept of so-called Aletheia Service Hubs is introduced, federating information supplied by information providers using Web services.

We already implemented core parts of the architecture (Section 3) while there still remain many open research questions. This article thus should foster discussion on how these questions can be solved. A short survey of related work (Section 4) shows that there are comparable approaches regarding the Semantic search part of the system. But integrating semi-structured and unstructured information sources and crossing organizational boundaries are quite new challenges we are facing in Aletheia. The architecture presented here is a first step towards the goal of semantic federation of comprehensive product information.

2. ARCHITECTURE

The architectural decisions for the Aletheia project are based on several major assumptions, which resulted from analysing the requirements of the five industry partners involved. Federating heterogeneous information requires a semantic repository in order to align and map between the different sources. The potentially large amount of information suggests that a central node closely connected to this repository is a sensible decision regarding the efficiency and also tends to simplify the system design.

There are many potential applications for the Aletheia system. Therefore, a variety of front-end applications must be supported, ranging from easy to use enterprise search engines for product support to personal information management solutions for product design to applications that enable the invocation of management tasks based on related information. Extensibility of central components hence is a necessity.

In order to adapt to environmental changes, the Aletheia system must be extensible with regard to the connected information sources as well. Thus, general classes of information sources need to be identified in order to provide an interface to them. They allow a simple inclusion of other information sources at runtime.

Finally, the system architecture must ensure that an individual domain, such as a company's department, can deploy and operate an Aletheia system on its own while several central nodes (Aletheia Service Hubs) may interact with each other, enabling collaboration between and use of information from different domains. The resulting generic architecture is visualized on a high level in Figure 1.

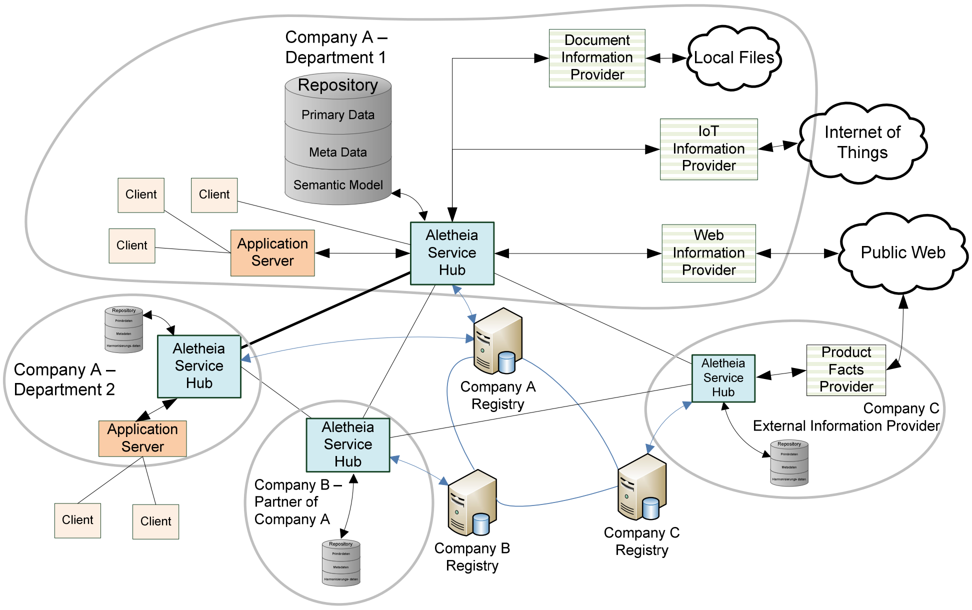

Figure 1. Aletheia architecture

2.1 Aletheia Service Hub

The Aletheia Service Hub is the main functional unit of the system. It provides an environment for different services, e.g. product search or annotations, that may be deployed inside the Service Hub and may be connected to corresponding services at remote Service Hubs. Beyond the concentration of functionality, the Service Hub also acts as a proxy for the applications to hide the complexity of the heterogeneous information source access. By deploying this proxy, we can enforce access control regarding the use of internal or external information sources within an administrative domain. Those sources may be unstructured (such as documents from file shares), semi-structured (like web pages from the Web), or structured (e.g. databases).

The main task of the Service Hub is to manage the semantic repository used to integrate primary data and meta data such as the time of last access, type of information, or reliability of information, gathered from different information providers. These may update information regularly at the Service Hub using a publish-subscribe service, e.g. if an RFID tag was scanned on the Internet of Things (IoT). Information can also be obtained at query time from information providers. The repository contains the semantic model, described by ontologies which are the central tool for semantic federation and semantic search. Product ontologies are used to unify facts gathered from the different sources and enable semantic processing of all this information.

Different Aletheia Service Hubs can be connected to each other. This allows one domain to access information sources of other domains, based on explicitly defined terms. For instance, policies can describe the confidentiality of shared information.

2.2 Information Provider

The Aletheia project not only federates information from structured sources, but from unstructured and semi-structured sources as well. The heterogeneous nature of this data requires a layer of information providers that encapsulate the functionality of accessing and preprocessing information sources (Wiederhold, 1992).

Both push and pull access is required, and there are data sources that either provide certain semantically described facts or fuzzy information, like entities extracted from natural language documents. Likewise, the repository not only consists of semantic information, but also features containers for fuzzy information and exclusively syntactic data such as a full text index. The generic functionality, e.g. routing requests to appropriate information sources and performing ontology matching, is part of this data integration layer.

The actual data providers range from local text files, office files on a file server, scraped Web sites and XML Web services to relational databases.

2.3 Distributed Registry

Each Aletheia Service Hub provides a registry. This component provides endpoint and configuration information for external applications and internal Aletheia Service Hub components. The registry also covers information on the data sources. We are currently evaluating different registry types w.r.t. their suitability.

3. IMPLEMENTATION AND EXAMPLE APPLICATION

Figure 2 shows a screenshot and the technical architecture diagram from an early implementation of the Aletheia prototype, using a single Service Hub as the backend with an integrated registry. The depicted example application provides faceted browsing over information on digital cameras. It takes advantage of the semantic and data federation capabilities of a first implemented Aletheia Service Hub.

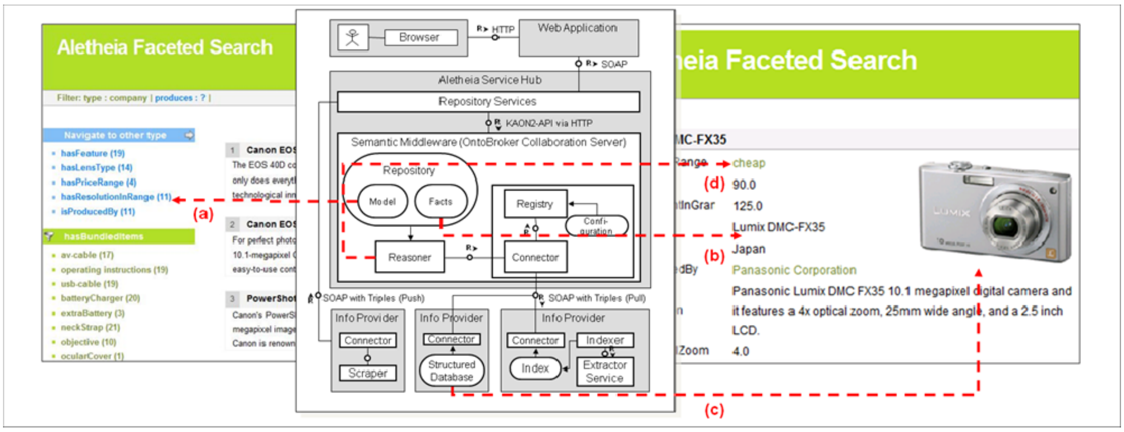

Figure 2. Aletheia prototype with example application

Navigation facets are automatically generated from the semantic model of the repository (a). The displayed product information originates from centrally stored facts entered into the repository by push information providers (b), from facts of an external pull-information provider with up-to-date vendor information (c) as well as from facts derived dynamically by a reasoner, as e.g. automatic price class categorization (d).

4. RELATED WORK

SemaPlorer (Schenk et al. 2008) is a similar approach to managing large sets of data, using a federated storage infrastructure for semantic data based on NetworkedGraphs. However, it is limited to semantic data described as RDF, whereas Aletheia aims for federating more heterogeneous data sources.

The NeOn project (Waterfeld et al. 2008) also provides an architecture for semantic integration, focusing on toolkits for ontology engineering, while ignoring semi- and unstructured information in the processing model.

Fedseeko (Walther et al. 2009) semantically federates product information, providing a complete view on specifications and facts. It also includes semi-structured information, but lacks an appropriate repository. Hence, the approach does not scale to a large user base.

Ensuring access control in knowledge federations (Evdokimov et al, 2009) is another aspect currently examined in Aletheia.

5. CONCLUSIONS

The presented architecture allows users to find relevant product information from a variety of sources. Furthermore, it extends the scope of related work by processing unstructured and semi-structured information in order to improve their accessibility.

Our goal is to enable collaboration between domains. Therefore, we consider replicating public information to other registries by implementing, e.g., distributed hash tables. We believe this research is valuable to the product information management and federated information system community.

ACKNOWLEDGEMENT

This work was partly funded by the German Ministry of Education and Research under the research grant number 01IA08001. We thank the partners of the Aletheia project for the fruitful discussions about the architectural approach presented in this paper.

REFERENCES

Aletheia, 2009. Aletheia project, http://www.aletheia-projekt.de

Berners-Lee, T., Hendler, J. and Lassila, O., 2001. The Semantic Web. *Scientific American,* Vol. 284, No. 5, pp. 34-43.

Blumberg, R. and Atre, S., 2003. The Problem with Unstructured Data. *DM Review Magazine*., February 2003, pp.42-46.

Burkett, M., 2006. *Validating Design Decisions Is Driving PLM Market Growth. AMR Technical Report.* AMR Research, Boston, USA.

Evdokimov, S., Fabian, B., Kunz, S., 2009. *Challenges for access control in knowledge federations*. International Conference on Knowledge Management and Information Sharing (KMIS 2009), Madeira, Portugal.

Lewis, P., 2005. *Business 2010: Manufacturing – Embracing the challenge of change*. Economist Intelligence Unit, London, UK.

Schenk, S., Saathoff, C., Baumesberger, A., Jochum, F., Kleinen, A., Staab, S., Scherp, A., 2008. SemaPlorer – Interactive Semantic Exploration of Data and Media based on a Federated Cloud Infrastructure. *In Billion Triple Challenge, International Semantic Web Conference*, Karlsruhe, Germany.

Walther, M., Schuster, D. and Schill, A., 2009. Federated Product Search with Information Enrichment Using Heterogeneous Sources. *Proceedings of Business Information Systems 2009*, Poznàn, Poland, pp.73-84.

Waterfeld, W., Erdmann, M., Schweitzer, T., Haase, P., 2008. D6.9.1 Specification of NeOn architecture and API V2. *Public Deliverable of the NeOn project.*

Wiederhold, G., 1992. *Mediators in the Architecture of Future Information Systems*. IEEE Computer, March 1992.