A FEDERATED TRIPLE STORE ARCHITECTURE FOR HEALTHCARE APPLICATIONS

Bruno Alves, Michael Schumacher and Fabian Cretton
Institute of Business Information Systems, University of Applied Sciences Western Switzerland, Sierre, Switzerland
{*bruno.alves, michael.schumacher, fabian.cretton*}*@hevs.ch*

Keywords: Interoperability; eHealth; Metadata, Semantic Web.

Abstract: Interoperability has the potential to improve care processes and decrease costs of the healthcare system. The advent of enterprise ICT solutions to replace costly and error-prone paper-based records did not fully convince practitioners, and many still prefer traditional methods for their simplicity and relative security. The Medicoordination project, which integrates several partners in healthcare on a regional scale in French speaking Switzerland, aims at designing eHealth solutions to the problems highlighted by reluctant practitioners. In a derivative project and through a complementary approach to the IHE XDS IT Profile, we designed, implemented and deployed a prototype of a semantic registry/repository for storing medical electronic records. We present herein the design and specification for a generic, interoperable registry/repository based on the technical requirements of the Swiss Health strategy. Although this paper presents an overview of the whole architecture, the focus will be on the registry, a federated semantic RDF store, managing metadata about medical documents. Our goals are the urbanization of information systems through SOA and ensure a level of interoperability between different actors.

1 INTRODUCTION

Interoperability is known as the ability for two or more systems to communicate together. It is a fundamental requirement in eHealth for enjoying the promised benefits of the adoption of electronic medical records (Brailer, 2005). Together with health information communication, interoperability can make data available where and when it is required. However, connecting systems is not enough to overcome the complexity and heterogeneity of modern healthcare infrastructures. Different systems must understand each other, they need a common knowledge.

Semantic web technologies provide the necessary level of intelligence to enable communication and knowledge sharing between disparate systems. Semantics encode the definition of each element of data, including its relationships with other elements. Semantic data convey a meaning, which is understandable by third parties sharing some common domain concept knowledge. Semantic interoperability is the key stone of information processing by computers (Della Valle et al., 2005a) and a new trend in healthcare informatics.

Multi-level interoperability in health information communication has the potential to improve the care processes and decrease costs of the health care system (Hillestad et al., 2005). To tackle the high potential of the domain of medical interoperability, but also respond to potential risks of data abuse, strategies for the interoperability exist in many countries (Lee et al., 2009)(Ruotsalainen et al., 2008), but also on European level (dec, 2008).

In this context, the Swiss Confederation also started an eHealth strategy late 2006 (OFCOM, 2007). Switzerland defined its strategy relatively late compared to its neighbors, because of its particularly fragmented health system. The Swiss eHealth strategy strives to create a clear outline for the next ten years in managing health data at various scales, and including participants from a large number of interest groups. This effort has lead to several concrete propositions for potential standards for data exchange and particularly an identification of partners in the system.

Based on these standards, we architecture a specification for a distributed semantic storage platform, called Medicoordination.

Medicoordination is a research project taking a complementary approach to the IHE Profile IT specifications. It describes a Service-Oriented Architecture (SOA), which can be used by different medical actors for sharing patient records. The goal is to provide a fully distributed storage solution for Patient Electronic Health Records (PEHR). The platform provides a federated metadata infrastructure allowing semantic descriptions of medical documents and a decentralized repository with versioning and access control mechanisms sharing strong security policies. The finality is to be able to exchange documents between actors working in different IT environments seamlessly and achieve a high-level of interoperability in completely opaque and heterogeneous environments.

Although Medicoordination is a wide project involving several components and modules, this paper concentrates on a specific point: the metadata system architecture or Metadata Service Layer (MSL). It consists in a federated RDF store, where metadata is stored as triples. This paper provides first an overview of the global architecture and proposes a view on the models, the derived architecture and a partially implemented prototype based on it.

2 RELATED WORK

The problem of the integration and exchange of distributed health data has already been thouroughly discussed in many articles, among which (Lenz et al., 2007)(McMurry et al., 2007)(Bergmann et al., 2007).

At the international level, large projects exist trying to solve the typical interoperability gap, which exists between hetegenous medical systems. The COOCON (Della Valle et al., 2005b) project aims at supporting health care professionals in reducing risks in their daily practices by building knowledge driven and dynamically adaptive networked communities within european healthcare systems. The ARTEMIS (Dogac et al., 2006) project proposes semantically enriched Web services in the healthcare domain in order to seamlessly connect medical institutions running heterogenous IT systems and exchange distributed medical data.

At the country level, most countries propose related initiatives. In Germany, for example, the bIT4health (better IT for better health) project described in (Blobel and Pharow, 2007) attempts to establish a telematics platform supporting seamless care combined with card enabled communication.

In our project Medicoordination, unlike many other presented here before, we focus on the distribution of the metadata, rather than the distribution of the documents. We attempt to model a Virtual Patient Record (VPR), described in (Records, 1997), supported by a distributed metadata architecture preserving the local ownership of the documents, while allowing patient consent on a per-document basis.

3 METHODS

The objective of the MediCoordination project is designing a decentralized management system for medical records that can be shared among regional care institutions. The system is composed of a registry, a repository, identity services and a coordination layer. It intends to address several integration problems with existing IT infrastructures and interoperability issues. This paper is focused on the registry component of the system. We describe herein the models resulting in a proposal of a technology-independent architecture and a partially implemented prototype of a federated metadata system.

The design of the system was constrained by recommendations made by the Swiss confederation eHealth coordination group. This organization provides recommendations on standards and technologies for Swiss pilot projects. It recently released a document on eHealth architecture guidelines (ehealthsuisse, 2009). These guidelines are grouped into three points of focus: security, distribution and information management and exchange. Security encompasses patient security, privacy, confidentiality, data protection and transparency. Distribution relates to decentralized structures, federalism, separation of roles, concerns about who owns the data and who can access it and integration of existing infrastructures and technology. Finally, data management and exchange relates to processes for data management.

Also, the metadata architecture is based on three models, which are derived from these points of focus. The distribution model specifies the distributed structure, which sustains the system; the data model describes how the metadata is specified and communicated; and finally, the security model describes the security mechanisms and how they are applied.

3.1 Distribution model

According to the recommendations, institutions should manage documents they generate. Since Switzerland is a highly fragmented country and distribution is required, a federated architecture is fore-

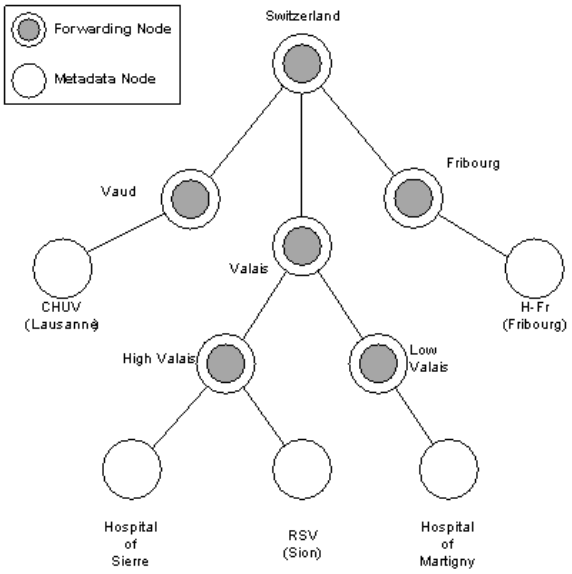

Figure 1: Storage architecture.

seen. Each node of the federation is a service accessible from the Web that can store metadata for the documents it issues. Nodes are connected in a hierarchical structure and distributed queries are performed globally on the network.

Besides regular metadata nodes, there also exist forwarding nodes, known as query hubs. These ”dummy” nodes forward the queries they receive and aggregate results according to the some environmental policies. Forwarding nodes cannot be used to write or read data directly, but they are involved in the organization of the network mesh. The federation of metadata nodes is built passively, i.e there is no discovery mechanism. A regular metadata node, such as RSV in Figure 1 must notify its existence to a query hub, or High Valais in this case. This construction allows defining security domains as a group of nodes sharing the same security policies. The concept of ”security domain” is explained in the next sections.

3.2 Data model

The current data model does not rely on semantics; however, we wanted to prepare the architecture for a future revision with semantic capabilities. Metadata in MSL basically consists in properties about the documents, such as file type, file identifier, author, date, etc. Since a large share of inferencing systems, among which Pellet and Racer, are based on RDF, document attributes in MSL are expressed using triples in the RDF format. A very simple ontology created with Protg was developed for the prototype.

3.3 Security model

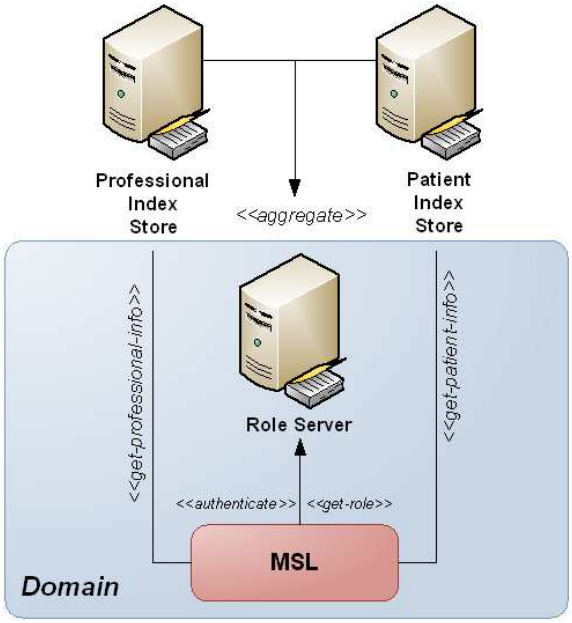

Figure 2: Security model.

The security model is broken into three sub-models: link-level security, authentication and authorization models.

Link-level security ensures the confidentiality and the integrity of the data by encrypting the communication channel. Certificate-based standards such as SSL/TSL 1.0 are typically used to encrypt the communication between endpoints. However, SSL does not prevent potential modifications of the SOAP messages between the application and the transport layer of the OSI Model. It is thus necessary to complement channel security with message-level protection mechanisms. Web Services Security (WSS) is a set of security policies based on XML, which provides primitives for encrypting, protecting and signing messages.

Besides link-level protection, it is important to ensure that only authenticated parties gain access to the resources. The difficulty of the MSL authentication model resides in the distributed nature of the service. Resources are spread all over a federated network and it is crucial to prevent access to users who are not authenticated to current node. The authentication model of the MSL relies on the concept of domains. A group of nodes sharing the same security policies (hospital campus, care services ...) is called a domain. Authentication is performed only once and users can directly access resources on nodes from the same domain, provided they have sufficient privileges.

Authenticating users do not ensure however that they are granted access to resources. For users to gain

privileges on them, it is necessary to define authorizations on documents. MSL uses a RBAC (Role-Based Access Control) model for controlling the access to the resources. Patients and medical staff are aggregated into roles, which are assigned specific privileges (read, write) on resources. Privileges are verified and validated by the repository in order to prevent or grant access to a category of users. Patient and medical staff information is stored into special databases, named Patient Index Store and Professional Index Store.

4 RESULTS

This section presents the global architecture, but focuses on the architecture of the MSL. It also gives an insight into a possible implementation.

4.1 Global architecture overview

Figure 3 presents the Medicoordination Healthcare Infrastructure or MHI. It is composed of the Metadata Service Layer (MSL - registry), the Storage Service Layer (StoSL - repository), Medicoordination Service Layer (MeSL - a coordinator service) and of Identity Services (IS - index stores for both patients and medical staff).

The storage service handles requests for reading and writing medical documents. The metadata associated to these documents is stored in a metadata node (from the issuer of the document), while the document contents are stored on the storage node. The role of the storage service is to manage health resources, organise documents in a patient medical record, maintain revisions of the files and perform sporadic audits.

The Medicoordination coordinator service represents the glue between all the components of the infrastructure. It is responsible for coordinating the registration and storage of the medical documents, while verifying the authorizations of the parties involved in the process (practitioners, medical doctors and patients). The Identity Services encompass Professional Index Stores, Patient Index Stores and Role Servers. A Role Server aggregates identities under a common denomination with common rights on a particular set of documents. The access rights for the roles are specified in the metadata for the moment and must be verified by the storage service. Besides that, Identity Services also provide authentication mechanisms, which can be endorsed by the Role Server itself. An authentication service is responsible for issuing tokens for the local domain and for the verification of foreign tokens issued by another domain.

4.2 MSL overview

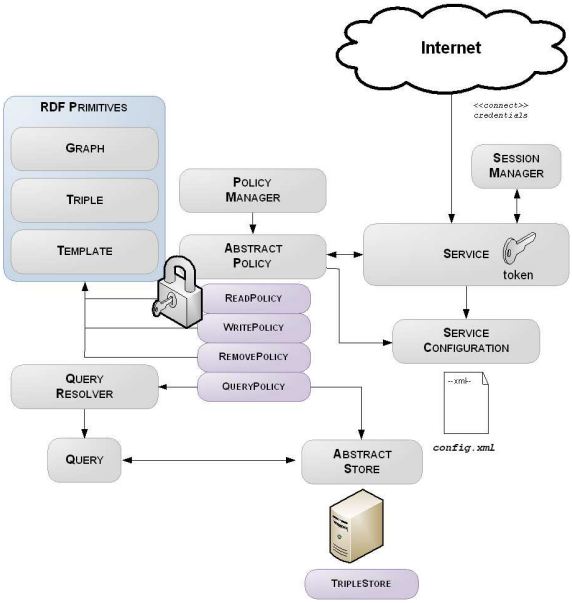

Figure 4: MSL architecture overview.

The Metadata Service Layer is a federated metadata system, which goal is to store information about fragments (i.e: documents composing a medical record), so that the information care professionals need is immediately available anywhere within the country.

The architecture is normally intended to be independent of any technology; however, we opted for communication using the Simple Object Access Protocol (SOAP 1.1 and 1.2) standard. This choice is primarily related to the excellent support for secure communications using the Web Services Security (WSS) specification.

Figure 4 presents the global architecture of the MSL. Each node is a Web Service endpoint provides four primitives: read(), write(), remove() and query(). Primitives work with graphs (set of triples about the same subject) and it is the role of the coordinator service to break up the initial metadata into triples. Metadata is stored in a database of triples (triple store) and is protected by a policy mechanism. The service is protected and requires authentication.

4.3 Representing information about documents

Meta-information or metadata about documents is related to the properties of the files.

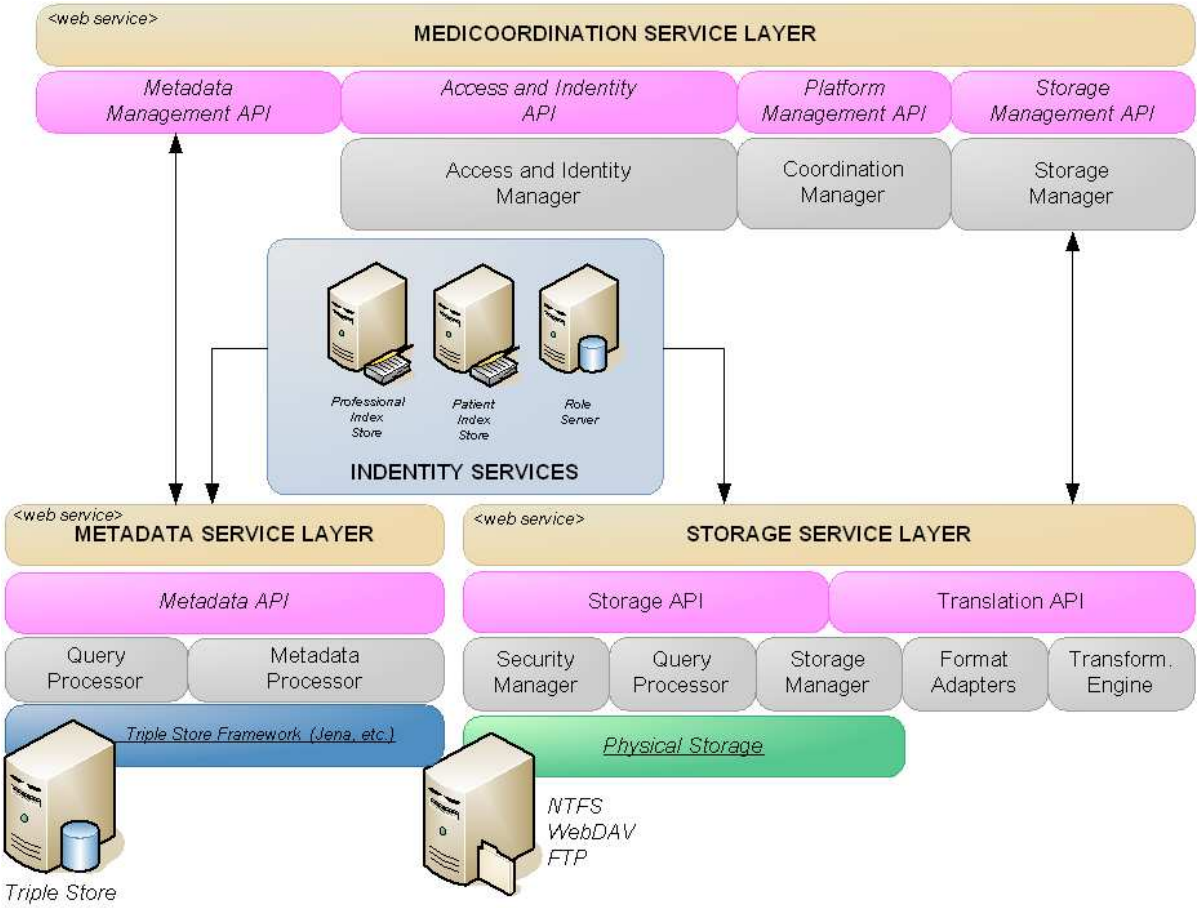

Figure 3: Medicoordination Healthcare Infrastructure

4.3.1 Representing metadata as triples

Metadata can be represented as a set of RDF triples referring to the same subject. RDF is a graph model language originally used as a metadata data model. RDF is not tied to a particular data format, but is often associated with XML and is called RDF/XML. RDF consists in graphs of triples referring to a same subject and allows expressing simple predicates (subject-property-object). For example, expressing that a particular fragment contains content about allergies, would yield the following triples (in N-Triples notation):

```
<http://medicoordination.ch/frag/1a23dd12>
<http://medicoordination.ch/onto/is-about>
'Allergies' .
```

The data model of the MSL does not impose any particular metadata requirements. However, there exist metadata definition initiatives for describing medical documents, such as (Malet et al., 1999). The metadata system described here does not strictly require inferring new triples when adding new data. In the prototype MSL, metadata was limited to the use of file properties and attributes like date, author, recipient, etc. Since those metadatas are just properties of a single object and thus have no relationship, they do not need inferencing. This behaviour is intended for this first specification of the architecture. Because documents are spread all over the network, metadatas on two different nodes can relate to each other and thus, could require a distributed inference. Distributed reasoning patterns are very difficult to implementand are still a field of active research (Schenk and Petrák, 2008)(Fang et al., 2008).

4.3.2 Storing metadata

Metadata is stored in a database of RDF triples, which is called an RDF store. The MSL provides an interface abstracting their behaviour: the AbstractStore. The implementation of new types of stores can be supported by any library supporting RDF, such as Jena or Virtuoso. Stores are loaded at run-time either programmatically or from a configuration file.

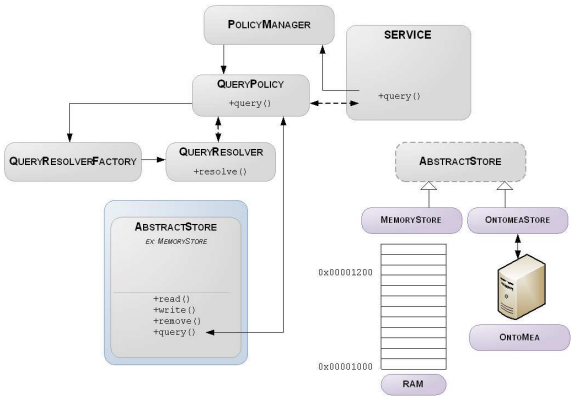

Figure 5: MSL storage architecture overview.

4.3.3 Managing access to metadata

The AbstractStore provides almost the same interface the Web service provides, except that its access is restricted by the use of policies. Policies, which are illustrated in Figure 6, represent a barrier between the users and the store. Policies allow controlling the data flow between the user and the resource. They can be used to verify access authorizations, implement security mechanisms or transform the input/output data. There is only one active policy for each service primitive (read, write, remove and query) and it is defined in a configuration file, but can be changed programmatically at run-time.

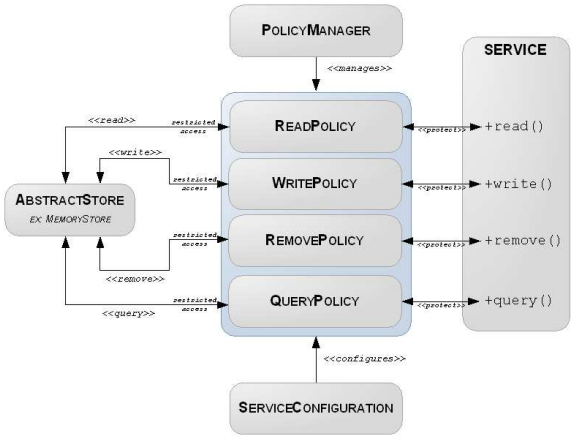

Figure 6: Policy mechanism overview.

When a service call is triggered, the PolicyManager class finds the active Policy for that method and invokes it. The Policy class then accesses the AbstractStore, but can perform some transformations or access control of the data it receives. This mechanism is particularly useful in the context of distributed queries.

4.3.4 Federating nodes

Although a MSL node is intended to work in a federated network topology, it provides no mechanisms for constructing the federation. This limitation allows on the other hand reusing the service in other environments.

Regular and forwarding nodes must be registered in federated structure supported by an external framework. The prototype uses WSDIR (Schumacher et al., 2007), a federated directory system which allows registration and discovery of semantic web services to build the federation. Each service is registered in a node of the WSDIR service and it is then used to find all metadata nodes registered in the federation children nodes.

4.3.5 Handling queries in a federation - from query resolution to execution

Metadata nodes do not have any knowledge of the other nodes. The federation access mechanisms are handled at the policy level. In this particular case, only queries need distribution, because reading, writing and removing is done locally. Federated queries are handled at the policy level. The prototype application implements a FederatedQueryPolicy that first performs the query locally and then forwards it to the other MSL nodes (children). Finally, it aggregates the results and forwards it to the calling node until the user gets a response.

4.4 Security considerations

The MSL uses a policy-based design in order to protect access to the underlying service RDF store. There exist a policy class for each service primitive (read, write, remove and query), preventing unauthorised inbound and outbound accesses to the resources.

Policies play a double role in the design of the architecture. First, they allow delaying the security decisions to a latter implementation phase, allowing thus concentrating on other programming aspects. Policies are defined at run-time in a configuration file. Second, policies allow applying protection mechanisms (resource observation) at the lowest levels (access level).

However, protecting the access to resource is not sufficient to prevent data stealing and data confidentiality problems. The MSL architecture specification defines mechanisms to ensure the security of the communication medium, confidentiality and integrity of the messages as well as resource access protection through authentication and access control.

4.4.1 Securing the communication channels

The MSL communicates with the user, with other nodes and the authentication server. It is important to SSL-encrypt the communication channel between all parties. Encrypting the communication between the node and the authenticator service is important in order to protected the token. Encrypting the communication between the user and the node is necessary to prevent data confidentiality loss and document stealth. Encrypting the communication between two nodes is important in order to avoid eavesdropping on the metadata. The prototype application uses SSL with certificates on the service.

4.4.2 Securing the inter-component messages

As said previously, SSL only encrypts the TCP/IP message payload from the transport layer of the OSI Model to the transport layer of the remote host. It is thus possible to forge a new message or modify it (XML). WSS primitives allow to encrypt and sign (XML Signature) messages. The prototype application does not currently implement SAML tokens, but uses custom tokens, which are passed to the service methods. Custom tokens are validated by the service itself.

4.4.3 Securing the access to resources

Access to the resources uses a RBAC access controlling scheme. Users are associated roles on the Role Server and are assigned specific access rights. These rights are defined in the metadata as read_id and write_id. Each of these properties accepts a list of roles as a value, specifying roles able to read or write. Describing access lists in the metadata makes it easy for authenticated and authorized patients or professionals to change access rights when needed, but also describing access rights on a document level.

4.4.4 Authenticating users

The federated network is distributed into domains, which are clusters of nodes sharing the same security policies. Inside a domain, there is one single authenticator service, which can be a Kerberos server, or anything else that allows single sign-one. However, supporting security domains supposes the authentication server to be able issuing security tokens. The Security Assertion Markup Language (SAML) enables ”portable trust” by supporting the assertion of authentication of single principals between different domains and is thus the recommended specification for authentication across multiple domains. The prototype application currently has no support for external authentication. Instead it authenticates users themselves on the local level. So, the only way to support domain authentication is through a shared database of credentials.

5 DISCUSSION

The Medicoordination architecture has similarities with IHE XDS in the sense that it is constituted of a registry, a document repository and of affinity domains. However, Medicoordination supports RDF metadata in a federated scheme. The advantage of a federated triple store is to provide a high level of data ”semantization” and allowing each health care actor to manage its own documents while providing a decentralized storage. Very advanced searches are then possible on the documents. It is feasible to return a list of accessible documents for each patient which has a particular disease with specific symptoms. The benefits of Medicoordination are to give a specification particularly well adapted for the Switzerland. Since each canton has its own policies on terms of healthcare, it is convenient to have a solution with provides decentralized storage and meta-information while giving the full control of data to local authorities. For example, a hospital, which is also a Medicoordination metadata node, can manage and control information about all documents produced by itself, but can also share them with the other institutes. They are part of a global federation across the country, but are local to a healthcare institute. Furthermore, the Medicoordination architecture does not impose the choice of a specific infrastructure or implementation. It only gives some guidelines about how the specific parts should work together. Medicoordination is also intended to adapt to existing infrastructures. It only requires thin and small clients to make the bridge between the healthcare information technologies used by the care institutions and the platform. A part of the existing application already supports such extensions through plugins. In order to validate our results, a full prototype has still to be implemented.

6 CONCLUSIONS

This paper introduced an architecture to be used in situations where the heterogeneity of systems prevents classic interoperability solutions to work. We did not dig into low-level concepts to remain independent of any architecture. The implementation of

systems based on Medicoordination necessitates careful thought on how to get different parts working together. Medicoordination, as a research project is intended to give some guidelines about a possible architecture for electronic healthcare infrastructure cooperation, which empowers each healthcare actor to manage its own data, while providing a flexible platform, which adapts to existing standards and infrastructures. Future development of the Medicoordination architecture may involve modifications to make it partially compatible with IHE XDS Affinity Domains and XDS repositories. This is a necessary development since the use of IHE profiles seems to be in the focus of the Swiss Confederation eHealth Strategy. Future work will also include a more detailed specification based on new communication standards linked to semantic web activities.

REFERENCES

(2008). Commission Recommendation on cross-border interoperability of electronic health record systems (notified under document number C(2008) 3282). *Official Journal of the European Union*.

Bergmann, J., Bott, O., Pretschner, D., and Haux, R. (2007). An e-consent-based shared EHR system architecture for integrated healthcare networks. *International journal of medical informatics*, 76(2-3):130–136.

Blobel, B. and Pharow, P. (2007). A model driven approach for the german health telematics architectural framework and security infrastructure. *International Journal of Medical Informatics*, 76(2-3):169 – 175. Connecting Medical Informatics and Bio-Informatics - MIE 2005.

Brailer, D. (2005). Interoperability: the key to the future health care system. *Health Affairs*, 24(1):19–21.

Della Valle, E., Cerizza, D., Bicer, V., Kabak, Y., Laleci, G., and Lausen, H. (2005a). The need for semantic web service in the eHealth. In *W3C workshop on Frameworks for Semantics in Web Services*.

Della Valle, E., Gadda, L., and Perdoni, V. (2005b). COCOON: Building knowledge driven and dynamically networked communities within european healthcare systems. Presented at Med-e-tel 2005 Conference in Luxembourg.

Dogac, A., Laleci, G., Kirbas, S., Kabak, Y., Sinir, S., Yildiz, A., and Gurcan, Y. (2006). Artemis: deploying semantically enriched web services in the healthcare domain. *Information Systems*, 31(4-5):321–339.

ehealthsuisse (2009). Cybersanté Suisse: Normes et architecture, Premieres recommandations.

Fang, Q., Zhao, Y., Yang, G., and Zheng, W. (2008). Scalable distributed ontology reasoning using DHT-based partitioning. In *Proceedings of the 3rd Asian Semantic Web Conference on The Semantic Web*, pages 91–105. Springer.

Hillestad, R., Bigelow, J., Bower, A., Girosi, F., Meili, R., Scoville, R., and Taylor, R. (2005). Can electronic medical record systems transform health care? Potential health benefits, savings, and costs. *Health Affairs*, 24(5):1103.

Lee, M., Min, S., Shin, H., Lee, B., and Kim, J. (2009). The e-Health Landscape: Current Status and Future Prospects in Korea. *TELEMEDICINE and e-HEALTH*, 15(4):362–369.

Lenz, R., Beyer, M., and Kuhn, K. A. (2007). Semantic integration in healthcare networks. *International Journal of Medical Informatics*, 76(2-3):201 – 207. Connecting Medical Informatics and Bio-Informatics - MIE 2005.

Malet, G., Munoz, F., Appleyard, R., and Hersh, W. (1999). A model for enhancing Internet medical document retrieval with" medical core metadata". *Journal of the American Medical Informatics Association*, 6(2):163.

McMurry, A., Gilbert, C., Reis, B., Chueh, H., Kohane, I., and Mandl, K. (2007). A self-scaling, distributed information architecture for public health, research, and clinical care. *Journal of the American Medical Informatics Association*, 14(4):527.

OFCOM (2007). Stratégie Cybersanté (eHealth) Suisse .

Records, P. (1997). Virtual Patient Records. *COMMUNICATIONS OF THE ACM*, 40(8):111.

Ruotsalainen, P., Iivari, A., and Doupi, P. (2008). Finland's strategy and implementation of citizens' access to health information. *Studies in health technology and informatics*, 137:379.

Schenk, S. and Petrák, J. (2008). Sesame RDF Repository Extensions for Remote Querying. In *ZNALOSTI Conf.*

Schumacher, M., Pelt, T. v., Constantinescu, I., and Faltings, B. (2007). Wsdir: A federated directory system of semantic web services. In *WETICE '07: Proceedings of the 16th IEEE International Workshops on Enabling Technologies: Infrastructure for Collaborative Enterprises*, pages 98–103, Washington, DC, USA. IEEE Computer Society.