Knowledge Sharing and Reusability within the Public Sector: Security Challenges and Potential Solutions

Petros Belsis[1,2] Stefanos Gritzalis[2] Grammati Pantziou[1] Christos Skourlas[1]

[1]Department of Informatics, Technological Educational Institution of Athens

[2]Department of Information and Communication Systems Engineering

[1]Agiou Spyridonos St., Aigaleo 122 10 Greece

[2]Karlovassi Samos, 83200 Greece

{cskourlas, pantziou} @teiath.gr, {pbelsis, sgritz}@aegean.gr

Abstract: Knowledge reuse sharing can boost organizational performance, especially within the public sector. Organizations often fail to utilize existing knowledge when they attempt to solve similar problems; in other cases, in order to exchange information they need to establish time-consuming conventional communication knowledge exchanging procedures, involving many participants, which decrease seriously organizational response times. Deployment of cooperative Knowledge Management (KM) techniques is an interesting challenge towards this direction. In this paper we describe the challenges from both an information retrieval and security perspective towards the integration of KM repositories.

1. Introduction

Knowledge has always been an important asset for organizations. A big challenge is related with the possibility to merge shared distributed repositories between different organizations, therefore extending the possibilities for knowledge reuse within the public and health sector. Many challenges emerge towards the realization of this target. For example the various types of heterogeneity; another important aspect is related with security management. Towards the alleviation of the first problem, the use of ontology seems to be more prominent. We have utilised an ontology based approach, enabling the correspondence of semantics to multimedia files. For the creation and representation of the ontology, the RDF ontology framework has been utilised. Furthermore, for security management reasons, prior to file distribution we need to apply a flexible and scalable access control framework, which will be described in a following section. The interconnection and the integration (I&I) of operational, disparate Information and Knowledge Management Systems (INKOMES), which have been established to cover the needs of the same or separate enterprises, is a difficult problem, in general. We must stress that true integration takes more than the interconnection, which offers transparent access to heterogeneous information and knowledge management systems. It is also important, in the case of integration, to find "a common ontological basis for future component based systems" [Lenz et. al.].

Systems' interconnection and integration (I&I) has to be examined based on various topics and alternative strategies (of I&I). More precisely, a solution is based to the well - known methods and techniques of system analysis but also has to take under consideration various new concepts e.g. data and knowledge warehouse and mining. Thus, in such a case, the enterprise(s) must build some structure on top of the existing systems. That is the creation

of a “Data and Knowledge (virtual or real) warehouse”, where information, data and knowledge are copied into and accessed through a real or virtual repository, which is centralized or distributed.

Information and Text Retrieval techniques and new tools e.g. OLAP and Document Management tools have to play a key-role. As an example, various user categories (knowledge workers, domain experts, specialists, etc) use office automation software for writing documents, drawings etc. We have to collect and organize all these kinds of information and knowledge, which are usually not supported by ordinary Information Systems. Text and Document retrieval may offer the appropriate techniques and tools.

Interconnection standards also contribute to the effective solutions of the, I&I related, problems. Of course, there are (standardized) specifications related to the needs of specific domains.

Common examples from the Healthcare sector

It is common to see various distributed and separately operating (“isolated”) Hospital Information Systems and Laboratory Information systems. Especially, in the case of Hospital Information and Knowledge Management systems an example of interconnection standard could be the Health Level Seven (HL7) specification that refers to the application level ("Level Seven") of the International Standards Organization's (ISO) communications model for Open Systems Interconnection (OSI). In this framework, the electronic data exchange between applications can be seen as the need of intra-communicating applications to exchange sets of data.

Lenz, Blaster and Kuhn discuss advantages and disadvantages and propose that alternative strategies of integration must be evaluated in the case of Healthcare systems’ integration. Lenz and Kuhn summarize the state of the art in web technology and compare it with the needs of Hospital information systems’ integration.

1.1 An Outline of the General Problem and the Rationale for Systems' Interconnection and Integration

Data and Knowledge warehousing arose for three reasons: First, the need to provide single, clean, consistent source of operational and historical data for decision support purposes; second, the need to do so without impacting operational systems; and third the need to “communicate” with experts and to access tacit and explicit knowledge (case bases, rules, facts, etc).

The integration and the interconnection of Information and Knowledge Management Systems are usually decided for four reasons:

- It is impossible to improve the management and support decisions without having a common pool of "unified" data or a “transparent” access in heterogeneous sources of data. Even simple statistics and summary information are difficult to be extracted without an integrated system or this transparent access.
- It is obvious that, there is a serious need to include historical data, previous results, and rules (knowledge) into operational records for supporting the everyday work.
- It is very difficult to control, in a daily basis, the relationship of the operational records with the historical data, the previous results and tacit knowledge.

- There is always a (financial) need to decrease or eliminate the number of system bugs and failures, the transactions carried out, by mistake, the insecure use of the system etc. Only explicit and tacit knowledge may offer the appropriate background for this possibility.

In the following section the "I&I" problem is formulated and examined following three directions (aspects): Data and Knowledge Warehouse, electronic data exchange between applications, Document and Text Retrieval.

2. Problem Formulation and Discussion

2.1 The Data and Knowledge Warehouse

The data and knowledge warehouse aspect can be understood as the need for an integrated and time-varying collection of facts, "historical" cases, case specific reasoning (rules, heuristics), summary data derived from the operational information systems and is primarily used in strategic decision making. In other words, there is a need for a new data and knowledge base that stores historical, aggregated and summary information and also stores, at least, explicit knowledge.

A Data and Knowledge Warehouse (DKW) could be understood as a multidimensional structure. In a simplified approach, which can reduce the potential cost of implementation, DKW could be seen as a combination of a multidimensional database of stored data and an organizational memory including interesting documents and cases (types of explicit knowledge, in general) and offering text and information retrieval possibilities. Ontologies could also play an important role in this overall approach.

Each dimension of the multidimensional DKW could be structured as a "hierarchy" of dimension levels and every level could be an attribute associated with a domain of values Dom(I). Hence, the dimension can be formalized as a lattice. The Hasse diagrams (see Fig. 1) illustrate three lattices and the specified partial orders (e.g. sup $L = L_{ALL}$):

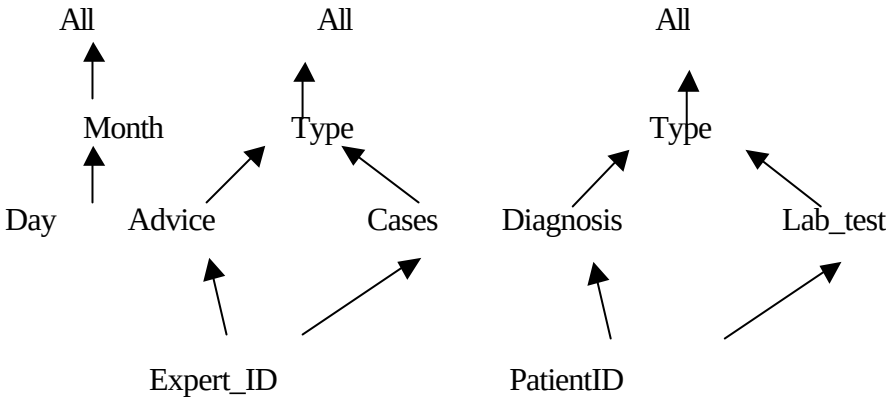

Fig. 1: Time lattice, Expert & Patient lattices.

A dimension scheme is a quadruple:

$$D=(\text{Dimension\_Name}, L, \leq, C),$$

where L is the set of dimension levels, (L, ≤) is a lattice (dimension hierarchy) and C is a (potentially empty) set of context dependencies.

The Dimension schemes, in our example, could be the following:

D_1=(Time, L_1, $\leq_1$, ∅), D_2=(Expert, L_2, $\leq_2$, ∅), D_2=(Patient, L_3, $\leq_3$, ∅).

The atomic information units of a DKW are given by facts (cases etc.). Hence, a fact is a point of the multidimensional space. A measure can be assigned to every fact.

Let us consider the formation of a fact scheme F as a quadruple (*Lechtenborger J.*, 2003): F=(Fact_Name, D, (M, FD_M), S), where D is a set of dimension schemes, (M, FD_M) is a measure scheme, and S is a set of summarization constraints. To clarify the concept of the measure scheme we can add that M is a set of attributes, which are called measures, and FD is a (potentially empty) set of functional dependencies (FDs) of the form

$$\{m_1 \ \ldots m_n \rightarrow m \mid m_1 \ldots m_n, m \in M, 1 \leq n\}$$

specifying a derivation order on M. Intuitively, a measure scheme (M, FD_M) specifies how measures can be computed from each other and a fact schema can be seen as a multidimensional representation of a certain universe of attributes (*Lechtenborger J., 2003)*. Hence, a portion of the fact schema, in our example, is the following:

F=(Facts, D, M, S), where D={D_1, D_2, D_3}

(M, FDs)=({Balance, BalanceClass}, {Balance → BalanceClass}), and S is a set of summarization constraints.

Table 1 shows a short simplified sample of the multidimensional DKM including data related to patients' examinations and experts' comments on specific cases

Table 1. A portion of the DKW

DD	MM	caseID	Diagnosis (eg use of ICD-10)	Lab_Test	Type	Balance	Balance Class
01	01	1	….	T3	P	1000	…
01	01	2	….	T2	P	800	…
02	01	3	….	T3	P	1000	…
02	01	4	….	T2	C	-800	…
…	….	…	…	…		…	…

DD: day, MM: month, P: Patient covered by a private scheme of social security

C: Patient covered by a company scheme (agreement)

DD	MM	Diagnosis (eg use of ICD-10)	CaseID	Lab test	Short Description of Results	Filename of detailed results	Case description and advices	Expert Id
01	01	D1	2	T2	Aaaaaaa	Test1.pdf	… written by …	…
01	01	D1	4	T2	Bbbbbbb	Test12.tiff	…	
02	01	D10	1	T3			…	

02	01	D10	3	T3	Ddddddd	Test30.pdf	…	

Lechtenborger and Vossen (2003) stress that the warehouse design is a non-trivial problem. They present a sequence of multidimensional normal forms and discuss how these forms allow reasoning about the quality of conceptual data warehouse schemata. Lu and Lowenthal (2003) examine strategic arrangements of fact data in order to answer analytical queries, efficiently, and improve query performance.

An emphasis must be given to the fact that the design, construction, and implementation of the DKW is an important and challenging consideration that should not be underestimated.

2.2 The Electronic Data Exchange between Applications.

We need a basis for solving the interconnection problem. Such a basis must provide standards for the exchange, management and integration of data that support decisions and office workers' support and the management, delivery and evaluation of services. Specifically, to create flexible, cost effective approaches, standards, guidelines, methodologies, and related services for interoperability between information systems. As an example, we can use a messaging standard that enables disparate applications to exchange key sets of operational data and knowledge and supports such functions as security checks, persons' identification, availability checks, exchange mechanism negotiations and, most importantly, data and knowledge exchange structuring. It must be designed not only to support a centralized case based system but also to serve a distributed environment where data and knowledge resides in enterprise / departmental systems.

2.3 Document Storage and Retrieval

The similarity of a document against a submitted query has been a field of continuing research for more than 20 years. In the popular vector space model a data set of n unique terms is specified, called the index terms (or keywords or uncontrolled terms or key phrases) of the document collection, and every document can be represented by a vector,

$$(T_1, T_2, \ldots, T_n)$$

where T_i=1, if the index term i is present in the document, and 0 otherwise.

A query is a document and can be represented in the same manner. The document and query vectors can be envisioned as an n-dimensional vector space. A vector matching operation, based on the cosine correlation used to measure the cosine of the angle between vectors can be used to compute the similarity. Hence, the following equation (Karanikolas and Skourlas, 2002) gives a well-known method to measure the similarity of document D_i against query Q:

$$S(D_i, Q) = \frac{\sum_{j=1}^{n} q_j t_{ij}}{\sqrt{\sum_{j=1}^{n} q_j^2 \cdot \sum_{j-1}^{n} t_{ij}^2}} = \frac{\sum_{j=1}^{n} q_j t_{ij}}{L_Q \cdot L_{D_i}}$$

where n is the number of index terms used in the collection, t_{ij} is the weight of term j in document D_i and q_j is the weight of term j in the query. The following two equations can be used to measure the terms t_{ij} and q_j:

$$t_{ij} = 0.5 + 0.5 \cdot \frac{F_{ij}}{\max F_i}$$

$$q_j = \log_2 \left(\frac{N}{DOCFREQ_j} \right)$$

where F_{ij} is the frequency of term j in document D_i, $maxF_i$ is the maximum frequency of the terms in document D_i, N is the number of documents in the collection and $DOCFREQ_j$ is the number of documents that include the index term j.

2.4 System Architecture and Implementation

Our architecture is materializing a distributed organizational memory. An Organizational Memory (OM) comprises a variety of information sources where information elements of all kinds, structures, contents and media types are available. In addition, a distributed OM utilizes knowledge from interconnected domains, representing knowledge assets in a location independent form. Several instances of an organizational memory are established in different organizational domains and are stored on local nodes. We provide a brief description of its core capabilities, in order to emphasize to its extensions that provide flexible authorization among distributed knowledge nodes.

The organizational memory module supports storage and retrieval for semi-structured documents with multilingual support. Organizational experience is being codified by subject in semi-structured documents, which consist of raw text, brief abstract description and keywords in order to facilitate retrieval. The system also attempts to provide support for tacit knowledge exploitation through its capability to interconnect users among them in order to share experiences when facing a specific type of problem. We also provide support for retrieving several types of files, such as images or multimedia files. All the repositories are implemented using Oracle while Java is used for interface implementation.

This implementation scheme is replicated on different nodes and it is supposed that different domains would like to contribute their knowledge potential. Our aim is to ensure that only authorized personnel among the two domains will have access to the knowledge sources. This situation is typical in e-Government environments, where all cooperating agencies need to share access to each other's data for a common purpose. In relation to these issues, in our research, we consider two main problems:

1. First, the problem of knowledge discovery upon different domains and second
2. how a user from one domain can be authorized to access resources from another domain and how this procedure can happen transparently and securely, which

means by minimizing the effort on the user's behalf and at the same time without exposing the knowledge assets to unauthorized disclosure or modification.

As far as it concerns to the first problem, the role of ontology is crucial. Each domain maintains its own domain ontology. We also introduce a central ontology repository, accessible from all the domains for retrieving domain ontologies. Ontologies define the concepts for each domain and their properties and enable semantically enabled description and querying over the knowledge assets. In order to enable transparent knowledge assets identification we utilize software agents that act on the user's behalf and query the distributed domains. Ontologies play also a crucial role in facilitating communication between software agents as they enable standardization of terminology in agent communication messages.
In relation to the second problem, we adopt a security policy based approach and we introduce a software agent that handles the necessary negotiations in order to authorize a user to gain access over a specific asset. The security policy defines the roles that deserve access to a specific asset, and the agent carries the user credentials which enable user identification, and accordingly by a policy interpretation the user is authorized or not to retrieve the remote asset.

3.1 Access Control Solutions

Managing the resources of a framework is a big challenge that requires a lot of effort on both the design as well as the implementation of countermeasures. Security policies are adopted to a high extent towards this direction. A policy can be considered to consist of a set of authoritative statements that determine the set of acceptable options in future selection processes. Relative to security, a policy can determine the set of acceptable actions, prohibitions and rights that are defined within the borders of an organization. A part of a security policy is determining the access control rights for each individual. Several challenges arise on this field, due to the very large number of subjects (resources) that need to be administered and due to the very large number of users. The Role Based Access Control (RBAC) (Shandhu) model seems to be dominant and widely accepted both in academic as commercial environments. The main principle of RBAC is related with the fact that usually users with similar roles, need to be accredited for the same actions, and need to have the same access rights. By classifying users to roles and accordingly by relating individuals with a role, the security management is simplified dramatically. For example, each time somebody enters the organization, we simply classify her to one of the predefined roles. Accordingly, when somebody leaves the organization, we do not need to manually withdraw all the access rights for every resource she was assigned to have access rights.

Security policies, provide a flexible means to automate the security management procedures as well as to enable the enforcement of access control decisions on distributed systems. Security policies can be codified in several special purpose languages, some of which provide codification in XML format, which makes them preferable, as they provide support for various platforms, and also makes them highly interoperable. The use of policies can simplify the management of distributed systems, which contain a large number of objects which often span across organizational boundaries. A more challenging option arises when it

comes to adapting to this framework resources from different domains which cooperate on the grounds of a common basis.

3.2 Access Control over Distributed Environments

We adopt the XACML (XACML) policy management framework in our application. XACML is standardized and allows extensions in order to become applicable to several types of networked environments, such as those incorporating Web-services. An overview of the XACML operational model is provided in the following: Among the key concepts we can distinguish those of the Policy Enforcement Point (PEP) and Policy Decision Point (PDP). Now the overall philosophy of XACML can be described in the following: First, the administrator is responsible for editing the security policy and encoding it in the appropriate format. Accordingly, she makes it available to the PEP. When a request is made, it is directed to the PEP. The PEP is requesting additive context related information, through another module, the context handler, responsible for constructing the messages in XACML format and collecting additive information, such as subject, action, resource and environment related attributes. Then, this XACML message is transmitted to the PDP which decides upon providing authorization. Accordingly, the PDP returns the response to the context handler in XACML native format and at the end the message is directed to the PEP, which fulfils its obligations, authorizing or not the requester to perform the requested action over the resource.

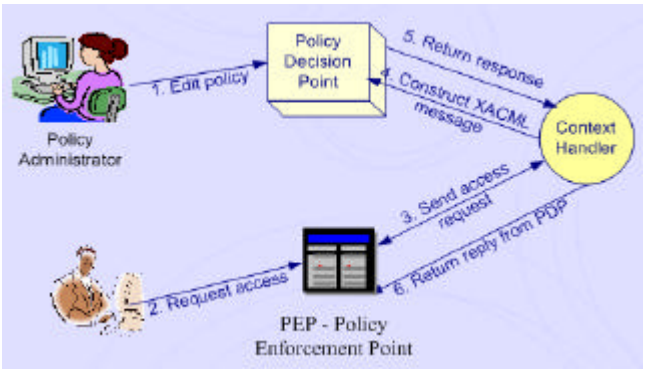

Fig 2: The XACML framework - overview

3.3 System Use Case Example Scenario

Imagine the following situation: a citizen visits a ministry asking for some documents (ex. Working permission). In order the civil servant to issue the requested permission he/she requests from the user to bring some documents from another department of the ministry or another ministry. In the second case, there is no way from users from other ministries to access resources from a different domain. Our work focuses exactly on the following. It enables through role correspondence a user from one ministry to be assigned a corresponding role on the remote ministry, which is pre-settled by the administrators of both domains, and

therefore to enable the joined management of an organization. Therefore, the processes are automated and simplified and there is no need to establish traditional means of contact such as physical presence.

The same challenge stands for interconnecting hospitals, where a doctor may seek information for one of his patients from another hospital, so that he can deliver in timely manner accurate important information about one of his patients situation.

4. Conclusions – Further Work

The interconnection and integration problem of disperse and separately operated systems can be solved using technically complex or more simple approaches. However, the use of a new "integrated" system (or interconnected systems) is controversial and much work has to be done for the desirable improvement of cooperation between the separate systems:

1. There is a need for an essential involvement of the management to influence the final acceptance and use of the integrated system.
2. "The more automation is established, the less deviations are allowed". New tasks, the way of doing things, etc., have to be clear. Advantages have to be analyzed and discussed, in depth.
3. "People have a reluctance to change their working habits". We must support them.
4. Specific user categories e.g. Nursing personnel in a Hospital have difficulties in using the new “system”.
5. Proprietary interconnection demands the solutions of the same problem each time we want to interconnect a new application e.g. to add a new Hospital or Laboratory in the “integrated” Health Care environment. There is a need for a general solution of the interconnection and integration problem.
6. There is a need for extracting and classifying information, (semi) automatically. As an example patient discharge letters form a potential source of information for extracting (semi) automatically the ICD codes related with the diagnosis. Data and Text mining can offer the appropriate techniques (Karanikolas and Skourlas 2002). Fuzzy sets can also be used for the mining of useful information (Hong et al., 2003).

Acknowledgments

This work is co-funded by the European Social Fund and National Resources – (EPEAEK-II)-ARXIMHDHS.

References

Davenport T., S. Volpel (2001). *“The rise of knowledge towards attention management”, Journal of knowledge management, vol. 5, No 3, 2001, pp 212-221.*

Chalaris I., Belsis P. (2003a) *From IT-Security & Quality Management Systems to IT-Governance: Trends and support using business modeling tools”, Proceedings The 9th EATA Netties Conference, Cyprus, October 2003, pp. 207-219*

Kim Y-G, S-H Yu, J-H Lee (2003). *"Knowledge strategy planning: methodology and case, Expert systems with applications, vol. 24, pp 295-307.*

King W., Marks P., McCoy S., (2002). *"The most important issues in Knowledge Management", Communications of the ACM, Sept. 2002, vol.45, No. 9*

Ravi Sandhu, David Ferraiolo, and Richard Kuhn (2000). *The NIST model for role-based access control: towards a unified standard. In Proceedings of the Fifth ACM Workshop on Role-Based Access Control (RBAC'00), pages 47–63, 2000.*

(XACML) *Extensible Markup Language Specification (XML), http://www.w3.org/XML/*

Lenz R, Blaser R, Kuhn KA (1999)*, Hospital information systems: changes and obstacles on the way to integration, Stud. Health Technol. Inform. 1999, No 68, pp. 25-30.*

Lenz R, Kuhn KA. (2001) *Intranet meets hospital information systems: the solution to the integration problem?, Methods Inf. Med., 2001, No 40, pp 99-105.*

(HL7)*http://www.hl7.org/*

Lechtenborger J., Vossen G. (2003)*, Multidimensional normal forms for data warehouse design, Information systems, No 28, 2003, pp 415-434.*

Lu X., Lowenthal F. (2003)*, Arranging fact table records in a data warehouse to improve query performance, Computers and Operations Research, 2003.*

Karanikolas N., Skourlas C. (2000),*Computed Assisted Information Resources Navigation. Medical Informatics and the Internet in Medicine, volume 25, No 2, 2000.*

Karanikolas N., Skourlas C. (2002)*, Automatic Diagnosis Classification of patient discharge letters. MIE 2002: XVIIth International Congress of the European Federation for Medical Informatics, August 25-29, 2002, Budapest, Hungary.*

Hong Tzung-Pei et. al. (2003)*, Fuzzy data mining for interesting generalized association rules, Fuzzy sets and systems, 2003*