Semantic Web Techniques for Personalization of eGovernment Services

Fabio Grandi[1], Federica Mandreoli[2], Riccardo Martoglia[2], Enrico Ronchetti[2], Maria Rita Scalas[1], and Paolo Tiberio[2]

[1] Alma Mater Studiorum – Università di Bologna, Italy
Dipartimento di Elettronica, Informatica e Sistemistica,
{fgrandi, mrscalas}@deis.unibo.it
[2] Università di Modena e Reggio Emilia, Italy
Dipartimento di Ingegneria dell'Informazione, Modena,
{fmandreoli, rmartoglia, eronchetti, ptiberio}@unimo.it

Abstract. In this paper, we present the results of an ongoing research involving the design and implementation of systems supporting personalized access to multi-version resources in an eGovernment scenario. Personalization is supported by means of Semantic Web techniques and is based on an ontology-based profiling of users (citizens). Resources we consider are collections of norm documents in XML format but can also be generic Web pages and portals or eGovernment services. We introduce a reference infrastructure, describe the organization and present performance figures of a prototype system we have developed.

1 Introduction

The Semantic Web (SW), once it has been proposed as the next generation of the existing Web [4], has solicited a crop of scientific interest both from academic and industrial research and has gained strong momentum for the last five years. However, after years of intensive research and impressive scientific results, SW is still in search for killer applications and real-world use cases which could demonstrate, beyond any reasonable doubt, its added value as enabling technology.

On the other hand, the last years have also witnessed a very strong, worldwide institutional effort towards the implementation of **eGovernment** (eGov) support services, which constitute an enormous challenge for the deployment of semantics and the exploitation of domain knowledge in the design, construction and operation of Web information systems. As a matter of fact, whereas SW technologies can be an ideal platform to envisage a knowledge-based, user-centric, distributed and networked eGov, the eGov domain can provide an ideal testbed for existing SW research and for the development of software applications with "ontologies under the hood". In this context, the first call is for *interoperability*: manifold semantic differences have to be settled in order to provide seamless services to citizens, as the eGov domain provides for differences in the interpretation of laws, regulations, life events, administrative processes, service workflows, best-practices, to be taken into account within and across regions, nations and

J.F. Roddick et al. (Eds.): ER Workshops 2006, LNCS 4231, pp. 435–444, 2006.

continents (not to talk about the usage of many different languages). The second call is for *personalization*: the achievement of a high level of integration and involvement of the citizens in the eGov and eGovernance activities, the necessity to fairly deal with different categories of citizens (including disadvantaged ones, with a potential risk of increasing digital divide), the requirements to suppport flexible, user-friendly, precise, targeted and non-baffling services, all claim for the personalization of the services offered and information supplied.

Whereas most of the recent and ongoing research on the convergence between SW and eGov is on the interoperability side (see e.g. [2]), we move on the personalization side, which we consider a legitimate corollary of the solution of the interoperability problems. If interoperability, including semantic integration of systems, processes and of the exchanged information, is the basis for the realization of complex networked eGov services, semantics-aware personalization of the Public Administration (PA) activities that concern the citizens and of the online offered services is aimed at improving and optimizing the involvement of citizens in the eGovernance process. In particular, we consider ontology-based user-profiling and personalized access to online resources (internally available in multi-version format), which may range from guided browsing of PA informative Web sites and portals to selective querying collections of norm documents, and to enactment of customized Web services [15] implementing administrative processes. Notice that, although all these kinds of resources are already available in existing eGov Web information systems, personalization is either completely absent or at most "predefined" in the Web site structure/contents or service definition/workflow (for example, hardwired in eGov portals by human experts according to the *life events* metaphor [9]). Effective, automatic, flexible, on-demand, "intelligent" and, last but not least, efficient personalization facilities are lacking.

In this paper, we present the results of an ongoing research started in 2003 (see [11]). In the first part of this research we focused on the design and implementation of Web information systems for personalized access to norm repositories. Building upon previous work on temporal management of multi-version norm documents [6], we developed a platform for semantics-aware personalized access to the repository. Personalization is based on the maintenance of an ontology which classifies the citizens according to the limited applicability of norm provisions. Semantic information is then used to map the citizen's identity onto the applicable norms in the repository thanks to an intelligent and efficient retrieval system. The ongoing research concerns the application of our ontology-based personalization techniques to the choice and execution of eGov Web services. The research activity will be described in Section 2 whereas conclusions can be found in Section 3.

2 Personalized Access to eGovernment Resources

In the framework of eGov, a large number of online resources including PA portals, informative Websites, usable administrative services are progressively

being made available to citizens. In particular, collections of norm texts and legal information presented to citizens (e.g. stored in large repositories in XML format [10]) are being made available and becoming popular on the internet owing to big investments and efforts made by governments and administrations. Such portals or websites are usually equipped with a keyword-based search engine or contain indexes and predefined navigation paths for user guidance (e.g. following the life events approach).

The main objective of such activity has been the development of techniques allowing an effective and efficient personalized access to multi-version norm repositories. First of all, the fast dynamics involved in normative systems implies the coexistence of multiple *temporal versions* of the norm texts stored in a repository, since laws are continually subject to amendments and modifications (e.g. it is crucial to reconstruct the consolidated version of a norm as produced by the application of all the modifications it underwent so far). Moreover, another kind of versioning plays an important role, because some norms or some of their parts have or acquire a limited applicability. For example, a given norm defining tax treatment may contain some articles which are only applicable to particular classes of citizens: one article is applicable to unemployed persons, one article to self-employed persons, one article to public servants only and so on. Hence, a citizen accessing the repository may be interested in finding a *personalized version* of the norm, that is a version only containing articles which are applicable to his/her personal case. Notice that personalization avoids in some cases the user to have to go through a huge amount of irrelevant text to find out the relevant one and, thus, may help to make the search feasible. For instance, the annual budget law of a state, maybe composed of several hundreds of articles, may contain one article whose provisions have some consequences on the way research funds must be managed by universities (maybe without ever explicitly mentioning "university" in the text). One university professor may be interested in accessing the repository to retrieve the personalized version of the budget law, which will only contain the pertinent article, without having to go through the whole norm text, which would result in a very time-consuming and daunting activity.

Introducing the civic ontology. In general, in order to enhance the participation of the citizens to an eGovernance procedure of interest through the provision of personalization facilities, automatic and accurate positioning of them within the reference legal framework is needed. To this purpose, we propose to employ SW techniques and introduce an ontology, called *civic ontology*, which corresponds to a classification of citizens based on the distinctions introduced by subsequent norms which imply some limitation (total or partial) in their applicability. In the following, we refer to such norms as *founding acts*. Hence, in order to define a mapping between ontology classes and relevant norm parts, applicability annotations are added to the XML encoding of norms. More precisely, a semantic versioning dimension is introduced in the multi-version data model used for the representation and storage of the XML resources.

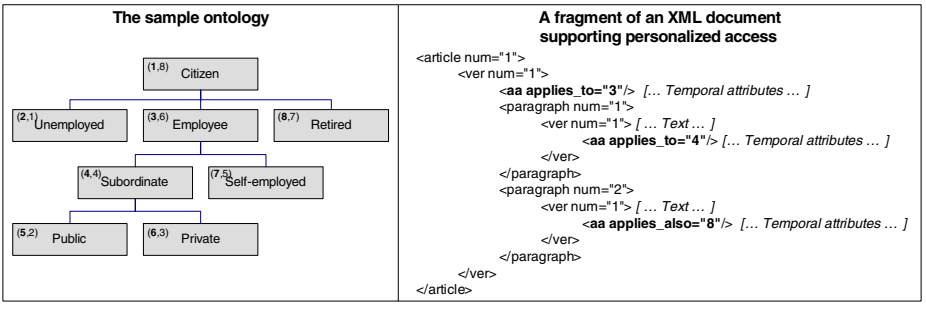

Fig. 1. An example of civic ontology, where each class has a name and is associated to a (pre,post) pair, and a fragment of a XML norm containing applicability annotations

For instance, the left part of Fig. 1 depicts a simple civic ontology built from a small corpus of norms ruling the status of citizens with respect to their work position. The right part shows a fragment of a multi-version XML norm text supporting personalized access with respect to this ontology, where the "*aa*" tag (**a**pplicability **a**nnotation) contains references to reference classes in the ontology.

At the current stage of the research, semantic information is mapped onto a *tree-like* civic ontology, that is based on a taxonomy of citizens induced by IS-A relationships. The tree-like civic ontology is sufficient to satisfy basic application requirements as to applicability constraints and personalization services, although more advanced application requirements may need a more sophisticated ontology definition. The adoption of tree-like ontologies allows us to exploit the pre-order and post-order properties of trees in order to enumerate the nodes and quickly check ancestor-descendant relationships between the classes. These codes are represented in the upper left part of the ontology classes in the Figure, in the form: (pre-order,post-order). For example, the class "Employee" has pre-order "3", which is also its identifier, whereas its post order is "6". The pre- and post-order information is then used to process queries in a very efficient way.

The article in the XML fragment on the right part of Fig. 1 is composed of two paragraphs and contains applicability annotations (tag *aa*). Notice that applicability is inherited by descendant nodes unless locally redefined. Hence, by means of redefinitions we can also introduce, for each part of a document, complex applicability properties including extensions or restrictions with respect to ancestors. For instance, the whole article in the Figure is applicable to civic class "3" (tag *applies_to*) and by default to all its descendants. However, its first paragraph is applicable to class "4", which is a restriction, whereas the second one is applicable to class "8" (tag *applies_also*), which is an extension. The representation of extensions and restrictions gives rise to high expressiveness and flexibility in an eGovernment scenario, where personalization requirements have to be met.

The reference infrastructure. In order to use the semantic versioning mechanism for personalization, we define the citizen's *digital identity* as the total

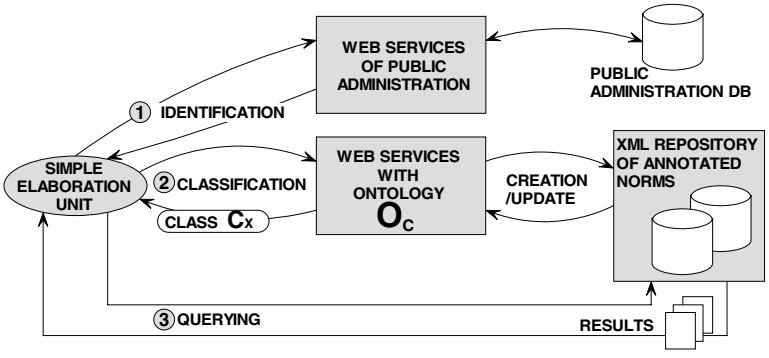

Fig. 2. The Complete Personalization Infrastructure

amount of information concerning him/her –necessary for the sake of classification with respect to the ontology– which is available online [14]. Such information must be retrievable in an automatic, secure and reliable way from the PA databases through suitable Web services (*identification services*). For instance, in order to see whether a citizen is married, a simple query concerning his/her marital status can be issued to registry databases. In this way, the classification of the citizen accessing the repository makes it possible to produce the most appropriate version of all and only norms which are applicable to his/her case.

Hence, the resulting complete infrastructure which is needed to perform all the required tasks is composed of various components that exchange information and cooperate to produce the final results as shown in Fig. 2. Firstly, in order to obtain a personalized access, a secure authentication is required for a citizen accessing the infrastructure. This is performed through a simple elaboration unit, also acting as user interface, which processes the citizen's requests and manages the results. Then, we can identify the following phases:

- the **identification phase** (step 1 in Fig. 2) consists of calls to identification services to reconstruct the digital identity of the authenticated user on-the-fly. In this phase the system collects pieces of information from all the involved PA web services and composes the identity of the citizen.
- the citizen **classification phase** (step 2 in Fig. 2) in which the classification service uses the collected digital identity to classify the citizen with respect to the civic ontology (O_C in Fig. 2), by means of an embedded reasoning service. In Fig. 2, the most specific class C_X has been assigned to the citizen;
- finally, in the **querying phase** (step 3 in Fig. 2) the citizen's query is executed on the multi-version XML repository, by accessing and reconstructing the appropriate version of all and only norms which are applicable to the class C_X.

In order to supply the desired services, the digital identity is modelled and represented within the system in a form such that it can be translated into the same language used for the ontology. In this way, during the classification

```
FOR    $a IN norm
WHERE  textConstr ($a//paragraph//text(), 'health AND care')
AND    tempConstr ('vTime OVERLAPS PERIOD('2002-01-01','2004-12-31')')
AND    applConstr ('class_7')
RETURN $a
```

Fig. 3. An XQuery-equivalent executable query

procedure, the matching between the civic ontology classes and the citizen's digital identity can be reduced to a standard reasoning task (see [3,8]).

Furthermore, the civic ontology used in steps 2 and 3 requires to be created and constantly maintained: each time a new founding act is enforced, the execution of a **creation/update phase** is needed. Notice that this process is a delicate task which needs advice by human experts and "official validation" of the outcomes and, thus, it can only partially be automated.

The resources of interests (e.g. norm documents) are stored in the XML repositories in a compact format according to a multi-version data model supporting temporal and semantic versioning (details can be found in [7]). Notice that temporal and limited applicability aspects may also interplay in the production and management of versions. However, since temporal and semantic versioning are treated in an orthogonal way in our model, also complex situations can be easily captured.

The queries that can be supported can contain four types of completely orthogonal constraints (temporal, structural, textual and applicability) allowing us to specify very specific searches in the XML norm repository. Let us focus first on the applicability constraint. Consider again the ontology and norm fragment in Fig. 1 and let John Smith be a "self-employed" citizen (i.e. belonging to class "7") retrieving the norm: hence, the sample article will be selected as pertinent, but only the second paragraph will be actually presented as applicable. Furthermore, the applicability constraint can be combined with the other three ones in order to fully support a multi-dimensional selection. For instance, John Smith could be interested in all the norms ...

- which contain paragraphs (*structural constraint*) ...
- dealing with health care (*textual constraint*), ...
- which were in force between 2002 and 2004 (*temporal constraint*), ...
- which are applicable to his personal case (*applicability constraint*).

Such a query can be issued to our system using the standard XQuery FLWR syntax [16] as in Fig. 3, where `textConstr`, `tempConstr`, and `applConstr` are suitable functions allowing the specification of the textual, temporal and applicability constraints, respectively (the structural constraint is implicit in the XPath expressions used in the XQuery statement). Notice that the temporal constraints can involve several time dimensions (see [6]), allowing high flexibility in satisfying the information needs of users in the eGovernment scenario. In particular, it is possible to extract consolidated current versions from the

multi-version repository, or to access past versions of particular norm texts, all consistently reconstructed by the system on the basis of the user's specifications and personalized on the basis of his/her identity.

Implementation and performance evaluation. In order to test the efficacy of the proposed approach, we built a prototype system supporting our data model. The system is based on a Multi-version XML Query Processor designed on purpose, which is able to manage the XML data repository and to support all the temporal, structural, textual and applicability query features in a single component. In addition to the introduction of the semantic versioning dimension, the system represents a complete redesign and extension of a previous system supporting temporal versioning described in [6], which we had built on top of a commercial DBMS with XML storage and query support. Details of the migration and a comparison between the two systems can be found in [7,12].

The prototype is implemented in Java JDK 1.5 and exploits ad-hoc data structures (relying on embedded "light" DBMS libraries) and algorithms which allow users to store and reconstruct on-the-fly the XML norm versions satisfying the four types of constraints. Such a component stores the XML norms in a partitioned way, which is used, during query answering, in order to efficiently perform structural-join algorithms [1] specifically adapted and tuned for the temporal/semantic multi-version context. Textual constraints are handled by means of an inverted index. Owing to the properties of the adopted pre- and post-order encoding of the civic ontology classes, the system is able to very efficiently deal with applicability constraints during query processing by means of simple comparisons involving such encodings and semantic annotations.

The experiments have been effected on a P4 2.5Ghz Windows XP workstation, equipped with 512MB RAM and a RAID0 cluster of two 80GB EIDE disks with NT file system (NTFS). We performed the tests on three XML document sets of increasing size: collection C1 (5,000 XML normative text documents), C2 (10,000 documents) and C3 (20,000 documents). We will describe in detail only the results obtained on the collection C1, then we will briefly describe the scalability performance shown on the other two collections. The total size of the collections is 120MB, 240MB, and 480MB, respectively.

Experiments were conducted by submitting queries of five different types (Q1-Q5). Table 1 presents the features of the test queries and the query execution time for each of them. All the queries require structural support (St constraint); types Q1 and Q2 also involve textual search by keywords (Tx constraint), with different selectivities; type Q3 contains temporal conditions (Tm constraint) on three time dimensions: transaction, valid and publication time; types Q4 and Q5 mix the previous ones since they involve both keywords and temporal conditions. For each query type, we also present a personalized access variant involving an additional applicability constraint (Ap constraint), denoted as Qx-A in the first column of Table 1. "XML-Native" denotes the system described in this paper, whereas "DOM-based" represents our previous prototype only supporting temporal versioning.

Table 1. Features of the test queries and query execution time (time in milliseconds, collection C1)

Query	Constraints				Selectivity	Performance (msec)	
	Tm	St	Tx	Ap		DOM-based	XML-Native
Q1	-	√	√	-	0.6%	2891	1046
Q2	-	√	√	-	4.02%	43240	2970
Q3	√	√	-	-	2.9%	47638	6523
Q4	√	√	√	-	0.68%	2151	1015
Q5	√	√	√	-	1.46%	3130	2550
Q1-A	-	√	√	√	0.23%	n/a	1095
Q2-A	-	√	√	√	1.65%	n/a	3004
Q3-A	√	√	-	√	1.3%	n/a	6760
Q4-A	√	√	√	√	0.31%	n/a	1020
Q5-A	√	√	√	√	0.77%	n/a	2602

Let us first focus on queries without personalized access. Our approach shows a good efficiency in every context, providing a short response time (including query analysis, retrieval of the qualifying norm parts and reconstruction of the result) of approximately one or two seconds for most of the queries. Notice that the selectivity of the query predicates does not impair performances (as it happened to the "DOM-based" approach), even when large amounts of documents containing some (typically small) relevant portions have to be retrieved, as it happens for queries Q2 and Q3. Our new system is able to deliver a fast and reliable performance in all cases, since it practically avoids the retrieval of useless document parts. Furthermore, for the same reasons, the main memory requirements of the new system are very small, less than 5% w.r.t. "DOM-based" approach.

The time needed to answer the personalized access versions of the Q1–Q5 queries is approximately 0.5-1% more than for the original versions. Moreover, since the applicability annotations of each part of an XML document are stored as simple integers, the size of the tuples with applicability annotations is practically unchanged (only a 3-4% storage space overhead is required with respect to documents without semantic versioning), even with quite complex annotations involving several applicability extensions and restrictions.

Finally, we ran the same queries of the previous tests on the larger collections and saw that the computing time always grows sub-linearly with the number of documents. For instance, query Q1 executed on the 10,000 documents of collection C2 (which is as double as C1) took 1,366 msec (i.e. the system was only 30% slower); similarly, on the 20,000 documents of collection C3, the average response time was 1,741 msec (i.e. the system was less than 30% slower than with C2). Also with the other queries the measured trend was the same, thus showing the good scalability of the system in every type of query context.

Current extensions. In our current research work, we are extending our ontology-based personalization approach to the definition and management of multi-version eGov Web services. For instance, ontology-based personalization

has been used in the field of eLearning services [5,13] and we are experimenting the adoption of similar techniques for the eGov application domain. In particular, we are applying our semantic versioning techniques also to solve this problem.

In particular, the application scenario and the reference infrastructure remain the same as in Fig. 2, along with the ontology management module and the identification and classification services. However, the XML repository and the query engine are being extended to also deal with the data required for the definition of multi-version Web services (including the specific data possibly required by a Workflow Management System to enact and support the execution of a workflow instance underlying the personalized eGov service delivered). Once the citizen has been classified with respect to the ontology during the identification phase, the semantic information is then used during the query phase to extract the data needed to build the personalized version of the requested Web service.

For instance, the eGov service for the "change of address" procedure may be different for public servants with respect to other categories of workers (e.g. public servants may have to declare that their new residence is within a fixed distance from their workplace, if required by law; hence a specific subservice for this task has to be included in the selected Web service). Hence, if the user has been classified as public servant, the query engine must retrieve a personalized version of the Web service for the "change of address" process for the citizen's case.

However, techniques very similar to the ones adopted for semantic annotation of XML documents and for efficient querying of the multi-version repository can be used also in such a case. Preliminary experiments are under way.

3 Conclusions

In this paper, we presented the results of a still ongoing research activity we are carrying out in the context of a national research project in order to support efficient and personalized access to multi-version resources in an eGovernment scenario. We defined a data model supporting ontology-based personalized access to XML documents, built a prototype system implementing the data model and evaluated its performance through some exploratory experiments. The results we obtained are very encouraging as to query response time, storage requirements and system scalability figures. Current efforts are focused on the extension of our approach to the support of ontology-based personalization of eGov services, as outlined at the end of the previous section.

In the future, we will strengthen the proposed approach, in particular by considering more advanced application requirements leading to a more sophisticated (e.g. graph-based) ontology definition, and by completing the required technological infrastructure with the specification and implementation of the remaining auxiliary services, including advanced reasoning services for management of the ontology. Further work will also include the assessment of our developed systems in a concrete working environment, with real users and in the presence of a large repository of real legal documents. In particular, a civic ontology based on a corpus of real norms (concerning infancy schools) is currently under development.

References

1. S. Al-Khalifa, H.V. Jagadish, J. M. Patel, Y. Wu, N. Koudas, and D. Srivastava. Structural joins: A primitive for efficient XML query pattern matching. In *Proc. of 18th ICDE Conf.*, San Jose, CA, 2002.
2. A. Abecker, A. Sheth, G. Mentzas, and L. Stojanovic (Eds.). *Semantic Web Meets eGovernment – Papers form the AAAI Spring Symposium.* AAAI Press, Menlo Park, CA, 2006.
3. F. Baader, I. Horrocks, and U. Sattler. Description Logics for the Semantic web. *Künstliche Intelligenz*, 16(4):57–59, 2002.
4. T. Berners-Lee, J. Hendler, O. Lassila, ' The Semantic Web, *Scientific American* 284(5):34-43, 2001.
5. R. Denaux, D. Dimitrova, and L. Aroyo. Interactive Ontology-based user modeling for personalized learning content management. In *Proc. of AH'04 Semantic Web for E-Learning Workshop.*, Eindhoven, The Netherlands, 2004.
6. F. Grandi, F. Mandreoli, and P. Tiberio. Temporal modelling and management of normative documents in XML format. *Data & Knowledge Engineering*, 54(3):327–354, 2005.
7. F. Grandi, F. Mandreoli, R. Martoglia, E. Ronchetti, M.R. Scalas, and P. Tiberio. Personalized access to multi-version norm texts in an e-government scenario. In *Proc. of 4th EGOV Conf.*, LNCS No. 3591, Copenhagen, Denmark, 2005.
8. I. Horrocks, and P.F. Patel-Schneider. Reducing OWL entailment to Description Logic satisfiability. In *Proc. of ISWC 2003*, Sanibel Island, FL, 2003.
9. The Italian e-Government portal. http://www.italia.gov.it/.
10. The "norme in rete" (norms on the net) home page. http://www.normeinrete.it.
11. The "Semantic web techniques for the management of digital identity and the access to norms" PRIN Project Home Page. http://www.cirsfid.unibo.it/eGov03/.
12. F. Mandreoli, R. Martoglia, F. Grandi, and M.R. Scalas. Efficient management of multi-version XML documents for e-Government applications. In *Proc. of 1st WEBIST Conf.*, Miami, FL, 2005.
13. L. Razmerita, and G. Gouarderes. Ontology-based user modeling for personalization of grid learning services. In *Proc. ITS'04 GRID Learning Services Workshop*, Maceio, Brazil, 2004.
14. S. Rodotà. Introduction to the "One world, one privacy" session. In *Proc. of 23rd Data Protection Commissioners Conf.*, Paris, France, http://www.paris-conference-2001.org/eng/contribution/rodota_contrib.pdf, 2001.
15. Web services activity. W3C Consortium, http://www.w3.org/2000/xp/Group/.
16. XML Query language. W3C Consortium, http://www.w3.org/XML/Query.