Prepared for the ICT for Government and Public Services Unit
DG Information Society and Media
European Commission

Bringing Together and Accelerating eGovernment Research in the EU

Information Integration
Report

Report Date
March 2009

Authors:
I. Kotsiopoulos, P. Rentzepopoulos

DG Information Society and Media
European Commission

Ch. de Charleroi, 123A
B-1060 Brussels

Prepared for the ICT for Government and Public Services Unit

DG Information Society and Media

European Commission

Executive summary

This report, compiled under the framework of the study ***"Bringing Together and Accelerating eGovernment Research in the EU"*** examines the status of information integration as applied to eGovernment research.

The first chapter, although introductory, has an important mission: to show why information integration in eGovernment today mainly relates to semantic integration and, at the same time, to align the reader with the most important trends and problems faced by the international research scene.

The second chapter emphasises on eGovernment applications, where ontologies and ontology integration in web services environments play a dominant role. Various examples of ontologies and ontology creation techniques and methods are given spanning applications from local government to multi-lingual support to legal knowledge representation and management.

The third chapter explores the international dimension of information integration through references to national and cross-border initiatives and related research activities.

The report concludes that information integration in the public sector requires agreement both at policy-making level and at administrative and technical level, given the vast size of the terminological system and the entities (communal, national, regional) involved.

ICT for Government and Public Services

DG Information Society and Media
European Commission

EUROPEAN DYNAMICS

Ch. de Charleroi, 123A
B-1060 Brussels

Prepared for the ICT for Government and Public Services Unit

DG Information Society and Media

European Commission

Table of Contents

Executive summary **2**

1 Introduction **6**

1.1 The problem of information integration 6
1.2 Structural data integration: a taxonomy 6
1.3 Semantic integration 9
1.3.1 Ontologies in data integration 10
1.4 Semantic integration - the industry view 13
1.5 Information integration: current research trends 15

2 Information integration in eGovernment **18**

2.1 Ontology development 19
2.1.1 Ontology maintenance and design examples 20
2.1.2 Implementation of ontologies I 24
2.1.3 Implementation of ontologies II 26
2.1.4 More ontology examples 26
2.2 Ontologies for knowledge management 30
2.2.1 Ontologies for legal knowledge 31
2.3 Data mining and ontologies for fraud detection 34

3 International Dimension **37**

3.1 The global perspective 37
3.1.1 Introduction 37
3.1.2 Scope and direction 37
3.2 International research topics in information integration 38
3.2.1.1 Technical aspects 38
3.2.1.2 Semantic aspects 39
3.3 Activities 40
3.3.1 International Cooperation 40
3.3.1.1 UN 40
3.3.1.2 OECD 41
3.3.1.3 EICTA 42
3.3.1.4 Joint initiatives 42
3.3.2 Non-EU national initiatives 43
3.3.2.1 USA 44
3.3.2.2 Australia 48
3.3.2.3 Korea 48
3.3.2.4 China 49
3.3.2.5 South-east Asia 49
3.3.2.6 Japan 50
3.4 Other resources 51

4 Conclusions **53**

ANNEX: eGovernment readiness index Top 35 (2008) **54**

List of figures

Figure 1. General integration approaches at different architectural levels (Ziegler and Dittrich) 7
Figure 2. ONTOGOV ontologies used for modelling eGovernment services 21
Figure 3. Overview of FIT ontologies 23
Figure 4. iWebCare: Generic fraud ontology 36
Figure 5. Adoption of Information Sharing 49
Figure 6: Types of innovation and diffusion in Japan 50

List of tables

Table 1. SemanticGov: GEA to WSMO concept mapping 28
Table 2: EGovernment readiness index (UN 2008) 55

1 Introduction

In this chapter we survey the field of information integration and the challenges it poses in the internet era. Our emphasis is on the main directions and results of recent research and the implications they have on eGovernment.

1.1 The problem of information integration

Information integration or, in a sometimes narrower sense, “data integration” has been with us ever since the early database systems. Although advances in theory (database and information-related mathematical theories and algorithms) and practice (processing power, software environments and the Web) have enabled considerable progress, the explosion in quantities of available data of heterogeneous origin, which ICTs themselves have “encouraged” and “attracted” has kept the problem alive to this day.

Data sources range from fields such as science (for example various measuring sensors) and civil matters (documents concerning law, tax, discussion etc.) to business and technology. These sources produce data of not only incompatible format but also of incompatible representation. An example of the latter can be sets of measurements taken at different time intervals and under differing accuracy assumptions. To these, one can add the multiplicity and geographic distribution of storage media, which operate under differing data models, formats and platforms.

All of the above point to the fact that we are still faced with a major integration challenge. As can easily be manifested in the field of our interest, eGovernment, it is not only becoming difficult to find what one wants but equally so to be able to combine it with what one needs. The aim of effective integration, namely having the right information at the right place at the right time still remains largely unattained.

XML has been a great boost towards solving data integration problems, especially when referring to partially structured documents. Although it can only define syntax, its use of metadata to describe structure allows flexible coding and display of data. Solution of integration problems at the syntactic level only, however, entails schema (data model) integration, which, in turn, has to overcome variations of structure, quality and consistency of data and metadata. These difficulties, which are inherent in most documents, severely limit the applicability of syntactic-level-only integration and call for semantic approaches.

Prior to discussing such issues, we pause for a generic classification of the various approaches to integration.

1.2 Structural data integration: a taxonomy

Most integration attempts which have been applied in practice so far belong to the so-called **structural**, or **syntactic-level** integration. This assumes relatively well-structured data, which allows rather tightly-coupled solutions leading to a single global schema.

It is useful at this stage to have a taxonomy of structural integration regarding the architectural approach used. For this purpose, in the present section, we present the classification scheme

of Ziegler and Dittrich[1], which is based on the layered architecture for information systems given in the figure below.

Referring to the figure, an information system is composed of five layers. The top layer (layer 1) provides users with access to data and services through a variety interfaces that run on top of different applications (level 2). Applications may use middleware (layer 3), such as transaction processing (TP) monitors, message-oriented middleware (MOM) and SQL-middleware to access data (itself managed by a data storage system) via a data access layer (part of layer 4). Usually, database management systems (DBMS) are used to combine the data access and storage layer (part of layer 4). All those layers act on the data itself (layer 5).

Ziegler and Dittrich view the integration problem as being addressed at each of those system layers, which also define the various approaches to integration, shown in the figure as links between the layers of two different (heterogeneous) systems.

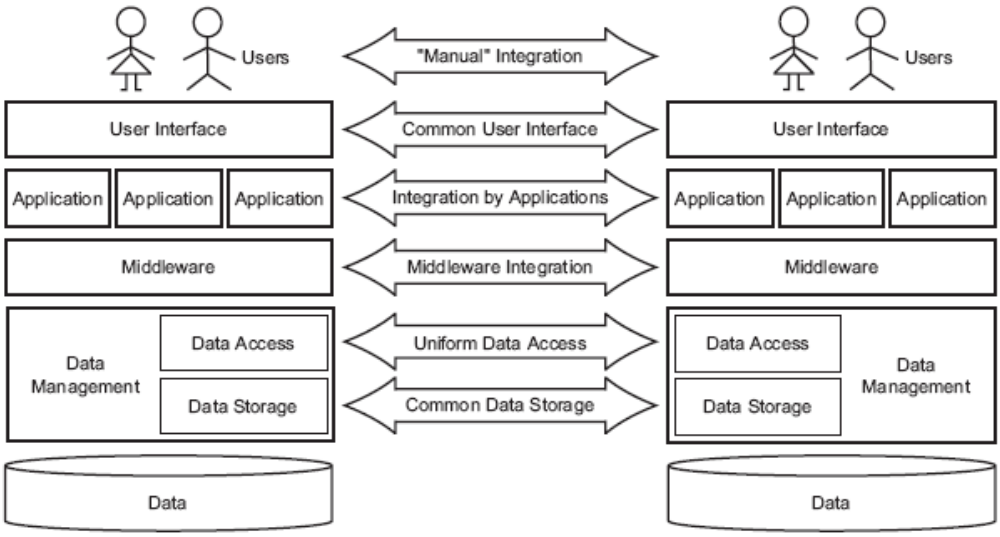

Figure 1. General integration approaches at different architectural levels (Ziegler and Dittrich)

The classification of Ziegler and Dittrich proposes the following types of integration:

1. **Manual Integration.** Here, users directly interact with all relevant information systems and manually integrate selected data, dealing with different user interfaces and query languages. As expected, detailed knowledge on location, logical data representation, and data semantics on behalf of the user is necessary.
2. **Common User Interface.** The next step into more automated integration is a common user interface such as a web browser. Although this provides a uniform look and feel, data from relevant information systems is still separately presented. Homogenisation and

[1] Patrick Ziegler and Klaus R. Dittrich, "THREE DECADES OF DATA INTEGRATION - ALL PROBLEMS SOLVED?", 18th IFIP World Computer Congress (WCC 2004), Volume 12, Building the Information Society.

integration of data must still be done by the user. A typical example of this approach is search engines.

3. **Integration by Applications.** This is the next level into automated integration. Integration applications access various data sources and return integrated results to the user. Although this is a practical solution for a small number of component systems, the integration applications used become increasingly large once the number of system interfaces and data formats to homogenise and integrate start to grow.
4. **Integration by Middleware.** Middleware provides reusable functionality that can generally be used to solve dedicated aspects of the integration problem (SQL-middleware for example). The advantage is that applications are relieved from implementing common integration functionality, although some integration effort is still needed in applications. For instance, SQL-middleware provides a single access point to send SQL queries to all connected component systems. However, query results are not integrated into one single, homogeneous result set. Another problem is that usually more than one middleware tools have to be combined to build integrated systems.
5. **Uniform Data Access.** In this case, logical integration of data is accomplished at the data access level. Global applications are provided with a unified global view of physically distributed data, though only virtual data is available at this level. However, global provision of physically integrated data can be time-consuming since data access, homogenisation, and integration have to be done at runtime.
6. **Common Data Storage.** Here, physical data integration is performed by transferring data to a new data storage; local sources can either be retired or remain operational. In general, physical data integration provides fast data access. If local data sources are retired, applications that access them have to be migrated to the new data storage as well. If local data sources remain operational, regular updates of the common data storage are needed.

Ziegler and Dittrich continue with characteristic examples which show that, in practice, most concrete integration solutions are based on a combination of one or more of those six general classes of integration approaches. We quote some of those examples:

- ***Mediated query systems*** represent a uniform data access solution by providing a single point for read-only querying access to various data sources. A mediator that contains a global query processor is employed to send sub-queries to local data sources; returned local query results are then combined.
- ***Portals*** are another form of uniform data access. They act as personalised doorways to the internet or intranet, where each user is provided with information tailored to his information needs. Usually, web mining is applied to determine user-profiles by click-stream analysis so that information the user might be interested in can be retrieved and presented.
- ***Data warehouses*** realise a common data storage approach to integration. Data from several operational sources (such as on-line transaction processing systems) are extracted, transformed, and loaded into a data warehouse. Then, analysis, such as online analytical processing can be performed on cubes of integrated and aggregated data.
- ***Operational data stores*** are a second example of common data storage. Here, a "warehouse with fresh data" is built by immediately propagating updates in local data sources to the data store. Thus, up-to-date integrated data is available for decision support. Unlike data warehouses, data is neither cleansed nor aggregated nor are data histories supported.
- ***Federated database systems (FDBMS)*** achieve a uniform data access solution by logically integrating data from underlying local DBMS. Federated database systems are

fully-fledged DBMS; that is, they implement their own data model and support global queries, global transactions, and global access control. Usually, the five-level reference architecture of Sheth and Larson (1990) is employed for building FDBMS.

- ***Workflow management systems (WFMS)*** allow implementation of business processes, where each single step is executed by a different application or user. Generally, WFMS support modelling, execution, and maintenance of processes which comprise interactions between applications and human users. WFMS represent an integration-by-application approach.
- ***Integration by web services*** performs integration via software components (web services) that support machine-to-machine interaction over a network using XML-based messages conveyed by internet protocols. Depending on their offered integration functionality, web services either represent a uniform data access approach or a common data access interface followed later by manual or application-based integration.
- ***Peer-to-peer (P2P) integration*** is a decentralised approach to integration between distributed, autonomous peers where data can be mutually shared and integrated. depending on the provided integration functionality. P2P integration constitutes either a uniform data access approach or a common user interface for subsequent manual or application-based integration.

Finally, for reasons of completion, we mention that the two most known database integration scenarios, namely Local-as-View (LAV) and Global-as-View (GAV) are examples of Uniform Data Access type of integration. In LAV the content of each data source is characterised in terms of a view over the schema of a virtual global database (global schema), while in GAV the content of each element of the global schema is characterised in terms of a view over the data sources.

1.3 Semantic integration

The tightly-coupled solutions which dominated integration approaches on data (be it relational or object-oriented) were proven inadequate for the semi or even fully unstructured nature of the data brought by the Internet. Associated web technologies have to cope with loosely-coupled mediator or agent systems handling heterogeneous sources and semantics. As none of the approaches mentioned so far (including XML) can supply a solution for integration, it is necessary to consider semantics and not syntax alone.

In the words of Ouksel and Sheth[2] semantics in this sense involves “mapping of objects in the model or computational world onto the real world” by taking into account of “the issues that involve human interpretation, or meaning and use of data and information”. Loosely speaking, the corresponding notion of semantic integration will have to involve merging of data which has the same meaning, i.e. which corresponds to the same (or for our purposes adequately similar) real-world entity or concept. This, in turn poses issues of resolving semantic ambiguities (an issue treatable by offering metadata which serves to resolve these ambiguities in a definite and explicit way) and also implicit context assumptions.

To address such issues and tackle heterogeneity, one would initially propose that an exhaustive specification is drawn capturing the intended real world meaning and associations of all data and schema elements in a database. This, unfortunately, is not possible for a data

[2] Ouksel, Aris M. and Sheth, Amit P., Semantic Interoperability in Global Information Systems: A Brief Introduction to the Research Area and the Special Section. SIGMOD Record, 28(1):5–12., 1999

base[3]. As the late Joseph Goguen explained[4]: "despite some optimistic projections to the contrary, the representation of *meaning*, in anything like the sense that humans use that term, is far beyond current information technology. As explored in detail in fields such as Computer Supported Cooperative Work (CSCW), understanding the meaning of a document often requires a deep understanding of its social context, including how it was produced, how it is used, its role in organisational politics, its relation to other documents, its relation to other organisations, and much more, depending on the particular situation. Moreover, all these contexts may be changing at a rapid rate, as may the documents themselves, and the context of the data is also often both indeterminate and evolving. Another complication is that the same document may be used in multiple ways, some of which can be very different from others. These complexities mean that it is unrealistic to expect any single semantics to adequately reflect the meaning of the documents of some class for every purpose."

It is evident from the above that the issue of semantics and semantic integration or even semantic interoperability are still widely open and actively pursued research areas. We shall refer to the directions and broad methods of such research in a subsequent section.

1.3.1 Ontologies in data integration

The use of the concept of ontology began in the early 90s, primarily as a tool of the knowledge-based methodologies, brought-in from the Artificial Intelligence community. The concept has subsequently gained popularity with the semantic web community and the relevant technologies. The idea behind this popularity is rather straightforward: we create a formal system of terms and subsequently attach elements of this system to those of machine-readable documents. In this way, loosely speaking, each class or domain of documents acquires another layer (semantic?) of logical relations which remain valid for all its instances.

Currently, it dominates all research in integration and interoperability around the world and certainly around Europe and FP6-FP7 research. We will concentrate on the latter in the next chapter.

The original idea of ontology, of course, belongs to philosophy (for example M. Heidegger[5]) and, the word itself to Greek, literally meaning "about the being" and implying a theory of existence. Although the interrelation between the two concepts i.e. philosophical and computer-scientific one is undoubtedly strong, the precise relationship between them is not straightforward[6]. In what follows we shall restrict ourselves to computer science ontologies.

It is of interest to note that even a computer science ontology does not have a generally accepted definition. In fact, the ontology community itself has agreed to differ in this respect!

[3] Sheth, Amit P., Gala, Sunit K., Navathe, Shamkant B, On Automatic Reasoning for Schema Integration. International Journal of Intelligent and Cooperative Information Systems, 2(1):23–50, 1993

[4] Goguen, J.A., "Information Integration, Databases and Ontologies", 2006, http://www.cs.ucsd.edu/~goguen/projs/data.html

[5] Heidegger, M. "Identity and Difference", Chicago, 1969. Translated and introduced by Joan Stambaugh

[6] Goguen, J.A., "Ontology, society and ontotheology", *International Conference on Formal Ontology in Information Systems (FOIS 2004)*, http://charlotte.ucsd.edu/users/goguen/pps/fois04.pdf, 2004.

An instructive description was first given by Gruber in 1992. We quote here a recent update of this definition and some commentary from the author[7]:

"In the context of computer and information sciences an ontology defines a set of representational primitives with which to model a domain of knowledge or discourse. The representational primitives are typically classes (or sets), attributes (or properties), and relationships (or relations among class members). The definitions of the representational primitives include information about their meaning and constraints on their logically consistent application. In the context of database systems, ontology can be viewed as a level of abstraction of data models, analogous to hierarchical and relational models, but intended for modelling knowledge about individuals, their attributes, and their relationships to other individuals. Ontologies are typically specified in languages that allow abstraction away from data structures and implementation strategies; in practice, the languages of ontologies are closer in expressive power to first-order logic than languages used to model databases. For this reason, ontologies are said to be at the "semantic" level, whereas database schema are models of data at the "logical" or "physical" level. Due to their independence from lower level data models, ontologies are used for integrating heterogeneous databases, enabling interoperability among disparate systems, and specifying interfaces to independent, knowledge-based services. In the technology stack of the Semantic Web standards, ontologies are called out as an explicit layer. There are now standard languages and a variety of commercial and open source tools for creating and working with ontologies."

Gruber continues to stress that an ontology when viewed as a tool and product of engineering is defined by its use: ontologies can provide the representational machinery with which to instantiate domain models in knowledge bases, make queries to knowledge-based services, and represent the results of calling such services. Furthermore, he points out that ontologies are part of the W3C standards stack for the Semantic Web (and the OWL formalism used in several variants of differing expressive power). In this context, they are used to specify standard conceptual vocabularies for data exchange among systems, query answering services, publication of reusable knowledge bases, and interoperability facilities across multiple, heterogeneous systems and databases.

Here we should make the distinction between uses of ontologies for document metadata (i.e. machine-understandable vocabularies) and those for knowledge bases. As Riichiro Mizoguchi explains[8], for metadata cases, instances are usually small, independent pieces of a web document, such as Mr. xxx:author, ICT:topic, 15/04/03:date, etc. This is far less stringent compared to models of real world objects, which require strict consistency for knowledge-base use. There, ontologies are specifications of a conceptualisation of the target world which require complex instance models. These must be built so as to guarantee consistency and fidelity with respect to the target world. Corresponding ontologies belong to the class of **heavy-weight ontologies** (as opposed to light weight ones).

For the purposes of the semantic web, a light-weight ontology (see same reference as before) will suffice as it only has to generalise the keywords in the query using an *is-a* hierarchy of concepts. In this case instance models mainly need to provide identification of the correspondence between concepts included in the two different sets of metadata under consideration.

[7] Gruber, T "Ontology", to appear in the *Encyclopedia of Database Systems*, Ling Liu and M. Tamer Özsu (Eds.), Springer-Verlag, 2008., preview in http://tomgruber.org/writing/ontology-definition-2007.htm

[8] Mizoguchi R., "Tutorial on ontological engineering"< Parts 1, 2, 3, New Generation Computing, 22, 2004.

The question that naturally arises is whether and how ontologies can help with information integration. We have already asserted that syntax level efforts are not enough; what is needed is semantic integration.

We first examine the semantics of an ontology. Mizoguchi (ref. on page 11) remarks that although an ontology sometimes looks just like a set of labels, it has deeper computational semantics. He proposes three such levels of increasing sophistication (weight) in an ontology:

1. A structured collection of terms. The most fundamental task in ontology development is articulation of the world of interest, that is, elicitation of concepts, then, organising them in a hierarchy. These are indispensable for something to be an ontology. Typical examples of ontologies at this level include topic hierarchies found in internet search engines and tags used for metadata description. Only few, shallow definitions of the concepts are made (light-weight ontology)
2. In addition to the features of level 1, we can add formal definitions to prevent unexpected interpretation of the concepts and necessary relations and constraints also formally defined as a set of axioms. Relations are much richer than those at level 1, definitions are declarative and formal so as to enable computers to interpret them. The interpretability of an ontology at this level enables computers to answer questions about the models which have been built based on the ontology. This is the level most ontology engineering effort aims for.
3. The ontology at this level is executable in the sense that models built based on the ontology run using modules provided by some of the abstract codes associated with concepts in the ontology. Thus, it can answer questions about runtime performance of the models. Typical examples of this type are found in the task ontologies of Mizoguchi and most ontologies built in FP6 research projects for eGovernment.

Using the guidelines above, ontologies are employed in modern database systems as a higher level of data modelling. Gruber (ref. page 11) notes that this new level of abstraction is above specific database designs (logical or physical), so that data can be exported, translated, queried, and unified across independently developed systems and services. This offers obvious benefits to database interoperability, cross-database search and integration of web services.

One word of caution here however: there has been a built-up of disproportionate expectation surrounding ontologies and their “ability” to solve most of the information integration problems. In the words of Joseph Goguen (ref. [6] page 10) this is “partly due to the technology being new and relatively untried, partly due to the hyperbole that so frequently accompanies new technology, especially when it has a comforting reductionist flavour, and partly due to insincere marketing by some researchers and organisations.” He concludes that correct application and use of such technology necessitates understanding of its limitations. He continues with the observation that computer-processable ontologies, which consist of logical axioms that relate terms of interest, have genuine promise when restricted to appropriate, well-understood domains. An example of this can be transactions in the banking sector or a specific type of supply chain. What we should note here is that this situation is far from true for many of the domains of eGovernment. There are too many ambiguities, context-related and locale-bound issues which dominate such transactions.

Following the observations of Goguen on the limitations of ontology engineering, we remark that, essentially, ontologies cannot capture real world semantics but only logical relations between terms. Examples can be “all humans are mammals”, “each child has two parents”,

and others. What is important is that the actual meaning of "human", "child" etc. remains unformalised. This is also true for exceptions to logical relations (an example is hybrids such as humanoids). Moreover, a given domain may have several ontologies, each in some ways incomplete and/or ambiguous, and possibly written in different ontology languages, which in turn may be based on different logical systems. OWL and RDF are currently most prominent, but others include Ontologic, KIF, KL-ONE, XSB, Flora, OIL. Such languages are either syntactic in nature (OWL) or highly expressive but computationally intractable[9] (KIF). In addition, there are specialised ontology languages which tend not to have a formal semantics.

Goguen concludes that "**the ontology approach to data integration may require not just schema and ontology integration, but also ontology language integration, and even ontology logic integration**, such that semantics is respected throughout the entire "integration chain," from actual datasets or "documents," through schemas and ontologies, up to ontology logics."

This is certainly not what we need for data of low quality, high heterogeneity and arising out of heterogeneous context. The last example is characteristic of many eGovernment situations, where documents contain terminology and metadata which makes sense only within their own context, such as that of a particular regional or national administration.

The proposed solutions and problems mentioned above are indicative of the difficulty of the information integration problem in the web era. That said, semantic integration is a necessity and ontology engineering is perhaps the most promising approach currently available. Semantic integration methods have also penetrated the IT industry and, of course have been the result and the subject of extensive research. The next section presents the industry view, while the one following briefly surveys some dominant research trends in information integration.

1.4 Semantic integration - the industry view

Stephen Lahanas, a principal consultant and co-founder of Semantech Inc., offered this view[10] of semantic integration: it "represents a specialized field of practice dedicated to using Semantic Design Principles, Methodologies and technology as a facilitating mechanism (often alongside SOA) to help solve enterprise-level problems for IT."

Although there is widespread awareness of the need for semantic integration of information throughout industry, there is also the belief that such technology is still immature. As Gartner analysts wrote in a report, "semantic technologies are early in their maturity and market adoption" and "many organisations will struggle to understand semantic approaches and view such technology as 'bleeding edge,' avoiding it because they are risk-averse." That said however, Gartner themselves believe that semantic technologies will emerge as one of the 10 most disruptive technologies in the next four years.

[9] Goguen, J.A., "Information Integration, Databases and Ontologies", 2006, http://www.cs.ucsd.edu/~goguen/projs/data.html

[10] http://www.semanticreport.com/index.php?option=com_content&task=view&id=125&Itemid=1&ed=7

A recent article by Baseline magazine[11] points out that the inevitable fact of growing information sharing needs along with the massive size of human knowledge have established the belief in organisations that new methods are needed. There is reportedly wide acceptance of ontologies as a leading technology, but there is also awareness of the problems of the current semantic ecosystem and its language structure, which is in line with the difficulties we mentioned in the previous section. There are also complaints about the unnecessary complexity of current ontology languages.

Along similar lines, Loraine Lawson from IT Business Edge[12] reports that the main feature of semantics is ontology; wider applicability by industry, however, necessitates a simplification stage the advent of which may also render relational database systems obsolete.

Having shown the awareness exiting in industry, what is the prevailing method for achieving information integration today? A practical guide[13] is given by John Taylor, a prominent member of the Integration Consortium:

"We believe that it starts with a service- oriented architecture (SOA). This provides a universal access mechanism to all systems via Web services and a universal data representation via XML. Also, this allows access to data not conveniently located in a database - commercial packages, custom applications, Web content, documents, images, feeds, etc. Having an SOA as a foundation supports the integration and development of information from structured, transactional systems as well as unstructured, content-based systems. Beginning with the meta data describing access and representation, we can build an information model of an organisation's information set containing relationships and rules that represents the semantics of the data and it's interaction with other data and processes. This model is best represented with the Web Ontology Language (OWL) from the Semantic Web group of the W3C. While the Semantic Web may be years away from attainment (if ever), a model-driven enterprise is achievable today. By creating an active model of data entities and mapping those entities to their respective sources exposed as Web services, true enterprise information integration is finally realised."

The four key requirements for implementing information integration are according to Taylor:

- Service-Oriented Architecture
- Meta-Data Management
- Semantic Information Model
- Dynamic Aggregation

John Taylor remarks that the practical difference OWL brings is that it can capture business logic (previously programmed in each individual system but not available to other systems) directly in an information model read by an OWL inference engine. In this way, data instances that populate the model are read in, and a unified view of information is provided. The engine performs all of the "cleansing" and "correlation" previously done by hand (or not at all). Moreover, duplicates do not have to be removed, values do not have to be adjusted, and

[11] Baseline magazine, August 2008, http://www.baselinemag.com/c/a/Search/Understanding-Semantic-Web-Technologies/

[12] http://www.itbusinessedge.com/blogs/mia/?p=440&nr=EEB, August 11, 2008.

[13] http://www.dmreview.com/news/1009669-1.html

everything is left in its originating system. What essentially happens is that a semantic model of information is created and mapped to the underlying infrastructure via a service-oriented architecture. As Taylor observes, the fundamental difference is that **this loosely coupled system is built by knowledge workers instead of programmers**.

We draw the attention of the reader here that although these methodical steps for implementation of information integration represent current industrial practice, they cannot resolve the problems heterogeneous and/or highly unstructured data present. These belong to the domain of active research, as shown below.

1.5 Information integration: current research trends

In this section we will look at the general state of international research in information integration and will report on major issues, methods and trends.

To address the problems heterogeneous and unstructured data bring, research has to ensure that problems are posed in a general, logical and consistent manner so that no confusion and ambiguity arise. The way to achieve this is to first express the problem in a formal way and subsequently try to reach conclusions within the formalism used. We start by formalising the data integration problem using a well-known and quoted approach[14].

A data integration system is a triple (G,S,M), where *G* is a global schema in a language L_G over an alphabet A_G, *S* is a source schema expressed in a language L_S and an alphabet A_S and *M* is a mapping between *G* and *S* which consists of assertions between queries q_G over *G* and queries over q_S *S*. When users pose queries over the data integration system, they pose queries over *G* and the mapping then asserts connections between the elements in the global schema and the source schemas. We therefore define a data integration system I over (G,S,M) which answers a query q_G posed over the global schema *G*, via *M* and the source schemas.

An application of the formalism can be seen in the expression of the two most known architectures for the mapping *M*, namely Local-as-View (LAV) and Global-as-View (GAV), first mentioned in section 1.2.

A LAV approach is a set of assertions corresponding to each element *s* of *S* of the form:

$$M: q_S \rightarrow q_G \text{, such that } s \rightarrow q_G, \forall s \in S$$

while a GAV approach is a set of assertions corresponding to each element *g* of *G* of the form:

$$M: q_G \rightarrow q_S \text{, such that } g \rightarrow q_S, \forall g \in G$$

[14] Lenzerini M., "Data integration: a theoretical perspective", in PODS, 233-246, 2002.

One can now see that the LAV approach characterises each element *s* of the source schema in terms of a view (query) q_G over the global schema. The GAV approach characterises each element *g* of the global schema in terms of a view q_S (query) over the source schema.

A Global-Local-as-View (GLAV) approach has also been defined in the literature, where *M*, i.e. the mapping relating the source and the global schemas, is established by using GAV as well as LAV assertions.

It is evident from the above that the problem facing data integration is the specification of the schema mapping *M*. For source schemas belonging to heterogeneous sources, the specification of this mapping involves exploration of the semantics of data belonging to the source as well as the global schemas. To tackle semantic heterogeneity among schemas techniques such as schema integration, merging, mapping and others have been developed. This is where ontologies step in as explained in section 1.3.1, bringing, however, a new set of problems to be tackled. These are related to the fact that "ontologies unfortunately are proliferating almost as quickly as the datasets that they are meant to describe. Therefore, integrating datasets the semantics of which are given by different ontologies will require that their ontologies be integrated first."[15] To this end, current research has concentrated on the problems of multiple ontologies which need to be accessed by several applications, especially over a distributed environment such as the Semantic Web. Among them, ontology mapping (definition and survey of approaches are given by Kalfoglou and Schorlemmer[16]) and/or ontology merging are used as techniques of ontology integration so that semantically consistent information can be preserved and exchanged. Hybrid approaches, which involve use of multiple ontologies that subscribe to a common, top-level vocabulary defining the basic terms of a domain are also common.

Despite such advances, one quickly realises that the problem is even deeper. As mentioned in section 1.3.1 ontologies are expressed in certain logics. Following Goguen, to "integrate ontologies, it may be necessary first to integrate the logics in which they are expressed." Moreover, it is to be expected that schemas describing structure will be expressed in different languages and different underlying data models, such as relational, object oriented, etc.

Contemporary research tries to resolve these questions, in the aim to provide a solid and consistent theoretical foundation for semantic integration (and information integration by enlarge). This is a necessary step if the discipline is to eventually reach a level of applicability which is comparable to that of a mainstream engineering discipline. The main instrument in such an effort is abstract algebraic frameworks, such as those presented by Schorlemmer and Kalfoglou[17], Kent[18], Menzel[19] and, of course, Joseph Goguen[20].

[15] Goguen, J.A., "Information Integration, Databases and Ontologies", 2006, http://www.cs.ucsd.edu/~goguen/projs/data.html

[16] Kalfoglou Y., Schorlemmer M., "Ontology mapping: the state of the art", Knowledge Engineering Review, 18(1): 1-31, 2003

[17] Schorlemmer, M. and Kalfoglou, Y., "On Semantic Interoperability and the Flow of Information," in *ISWC '03 Semantic IntegrationWorkshop*, Sanibel Island, Florida, USA, October 2003, http://citeseer.ist.psu.edu/schorlemmer03semantic.html

[18] Kent R., "The IFF Foundation for Ontological Knowledge Organization," in Knowledge Organization and Classification in International Information Retrieval, Haworth, 2003.

What is common in these approaches is the high level of abstraction employed, which frequently goes beyond first order logic and model theory. This is indeed justified by the nature of the problems involved: information integration implies semantic integration, which, in turn, usually involves ontology integration. Goguen argues that “algebra can help master this integration chain." Ontology, formally, is a theory over a logic, i.e. logical axioms that relate predicates of interest (see footnote [15]). To integrate, one needs a **rigorous foundation not tied to any specific representational or logical formalism.** This led Goguen and other researchers to use the powerful tools of category theory, information flow and channel theory and the theory of institutions. Category theory[21] allows independence from any particular choice of representation, while the theory of institutions[22] generalises the information flow and channel theories of Barwise and Seligman[23] to any logic to provide an axiomatisation of the notion of a logical system.

What has been achieved so far, is a high degree of conceptual unification, which can lead to the creation of semi-automatic integration tools such as the interactive SCIA tool for XML Schema and Document Type Definition (DTD) matching, developed at the Meaning and Computation Laboratory of the University of California at San Diego. The tool was created by Joseph Goguen’s team based on their theory of abstract schemas and abstract schema morphisms, which provides a semantics for n-to-m matches with semantic functions and/or conditions over diverse data models. Research to extend the tool to handle ontology integration as well is currently ongoing.

Finally, we give some examples of the conceptual unification which can be achieved, as described by Goguen. First, with virtually every logical system having a corresponding institution, theories (i.e. sets of sentences over a common signature) over an institution can be defined. This, in turn, means that ontologies can be defined and theory morphisms can be used for translating ontologies over a given logic. For translating between different logical systems, institution morphisms[24] can be defined as well as morphisms of theories over different institutions[25,26], representing the most general form of ontology integration.

In a similar fashion, the LAV approach for database integration mentioned in previous sections corresponds to co-relations and co-cones in a category of schemas and views, while the GAV approach corresponds to their dual, namely relations and cones.

---

[19] Menzel, C., “Basic Semantic Integration”, Dagstuhl Seminar Proceedings 04391 on Semantic Interoperability and Integration, 2005, http://drops.dagstuhl.de/opus/volltexte/2005/42/

[20] Goguen J., “Three perspectives on Information Integration”, Dagstuhl Seminar Proceedings 04391 on Semantic Interoperability and Integration, 2005, http://drops.dagstuhl.de/opus/volltexte/2005/38/

[21] Goguen J., “A Categorical Manifesto”, Mathematical Structures in Computer Science, 1 (1991) 49-67

[22] Goguen J. and Burstall R., “Institutions: Abstract Model Theory for Specification and Programming”, Journal of the Association for Computing Machinery, **39**(1) (1992) 95-146

[23] Barwise J, Seligman J., “Information Flow: Logic of Distributed Systems”. Cambridge 1997, Tracts in Theoretical Computer Science, 44

[24] Goguen J., Rosu G., “Institution Morphisms”, Formal Aspects of Computing, 13:274-387, 2002

[25] Diaconescu R., “Grothendieck institutions”, Applied Categorical Structures, 10:383–402, 2002.

[26] Goguen J., “Data, schema and ontology integration”, in Proceedings: Workshop on Combination of Logics, pp 21-31, Center for Logic and Computation, Instituto Superior Tecnico, Lisbon, Portugal 2004

2 Information integration in eGovernment

The public sector is characterised by a multitude of items which are amenable to different meanings, in domains such as laws and regulations, citizen services, administrative processes, various documentation and best practice examples. This is accentuated by the vast amounts of unstructured legacy information accumulated through the centuries. To this, one should add the multiplicity of languages spanning regions, nations and even whole continents. Technically, these differences belong to the sphere of semantics. As pointed out in the precious chapter, there is no currently available automated way to reconcile fully differences of this kind, primarily due to the dominant role which the associated context plays in the formation of the actual meaning. That said, it has long been recognised by all practitioners in the field of eGovernment that the most important problem in offering effective, context-aware eGovernment services to citizens is semantic information integration. The problem is even more dominant when such services transcend borders of culturally different regions or states.

The advent of the web enabled some progress towards solving the semantic integration issue: web services provided a hardware-software and location independent solution and concepts such as those of the semantic web gave a theoretical perspective of what is ultimately possible.

Indeed, when functional requirements of information integration are considered, eGovernment poses a set of demanding but exciting issues which surpass those of traditional web applications such as eBusiness or search engines. To mention some of them[27]:

- the need for inter-portal search, such as searching for additional resources on other portals to reply to a primary user or agent request
- a high degree of formality of key areas such as law
- extreme requirements to come to same decisions in similar situations
- high demands on security, privacy and trust
- occasionally extremely long-running process instances, as for example in urban and regional planning
- occasionally extreme informational imbalances between stakeholders, as well as many different stakeholders in the same process, for example citizen vs. city council, county council, federal government and others.

Although such requirements are far from trivial and in need of considerable research before solutions become available, the eGovernment research domain has to face the consequences of its unique relation to politics and policy making. A side-effect of this association is a sometimes short-termist attitude by funding bodies or decision makers which "channels" eGovernment researchers towards producing tangible results in given time frames. As a consequence, on a European scale, eGovernment research appears underfunded and without having realised its full potential. In the words of the eGovRTD2020 consortium[28]:

[27] Abecker A., Mentzas G., Stojanovic L. (eds), "Proceedings of the Workshop on Semantic Web for eGovernment", 3rd European Semantic Web Conference, Budva, Serbia & Montenegro, 2006

[28] "Roadmapping eGovernment Research Visions and Measures towards Innovative Governments in 2020", edited by Cristiano Codagnone and Maria A. Wimmer, eGovRTD2020 Project Consortium, 2007.

"*Certainly FP5 and FP6 have produced appreciable research, but our findings show that the current development of eGovernment has not reached outstanding results and that many challenges are still to be solved with the help of fundamental research. …"*

Information integration in European eGovernment research projects is no exception to this trend. The fundamental problems have been addressed, interesting and substantial results have been produced, but there is certainly room for more to be explored along the general research trends presented in section 1.5 of the previous chapter. In fact, as the eGovRTD2020 consortium mention: *"…there exist some key challenges which can only be overcome via basic fundamental research …"*

In line with these trends, information integration in eGovernment research has almost exclusively concentrated on ontologies and ontology integration in web services environments. This is justified by the fact that concepts the public sector uses are fundamentally similar in meaning between administrations and so are the relations between them. For example, a birth certificate usually mentions parents, time and place of birth, etc. and corresponds to each citizen. A formal terminological system (ontology) can therefore be created, and, at a high enough level can reach a certain level of stability and invariance to be used as a common domain ontology. The practical problem to be faced is agreement at both policy-making level and at administrative and technical level, given the vast size of the terminological system and the entities (communal, national, regional) involved.

One should also note that due to the nature of research projects to seek for what is new, there has been not much attention in FP6 in reconciling existing unstructured data associated with the public sector.

In what follows, we shall refer to the main features of FP6 research in ontology integration. As summarised by the Knowledge Web Network of Excellence researchers[29], heterogeneity problems in the semantic environments employed by FP6 research projects have to be solved at the ontology level, because data is described by ontologies. The general alignment-mapping-merging mechanisms between ontologies are used. These are usually created during design time to show how instances (data) from one ontology can be expressed in terms of another ontology.

At the web services level, data heterogeneity is tackled via data mediation, currently inherently supported only by the Web Services Modelling Ontology (WSMO) but not by the more common OWL-S.

2.1 Ontology development

Being a terminological system, an ontology is not only an important information integration enabler but also a knowledge base. This makes them particularly relevant to eGovernment as they allow flexible adaptation to changing and diverse environments and needs. Public administrations need such agility: they frequently face complex situations the analysis and response to which must comply with changing legal frameworks. This puts specific requirements to the development of suitable ontologies.

[29] "Data mediation in semantic web services", Deliverable D2.4.12, Knowledge Web NoE, 2007

The SAKE project[30] which develops a holistic framework and supporting tools for an agile knowledge-based **eGovernment** has supplied a guide for public sector ontology development. The ontologies built use "concepts" as their basic elements. For example, a concept of Public Administrator represents all people who are employed by the public administration and work at its premises. Specific public administrators are instances of this concept. The inclusion hierarchy continues via sub-concepts and super-concepts in the reverse direction. Using these, the generic concise guide developed includes the following main steps:

- **Step I.** Determine the domain and scope of the ontology: identify what the ontology will cover, and what the ontology is going to be used for.
- **Step II**. Consider reusing existing ontologies: consult existing documentation within the public administration, such as organisational charts, regulations, taxonomies or data models used by legacy systems. Available eGovernment ontologies and metadata standardisation efforts are also another option.
- **Step III**. Enumerate important terms in the ontology. Initially, it is important to get a comprehensive list of terms without worrying about overlap between the concepts they represent, relations among them, properties they may have or modelling of those properties.
- **Step IV.** Define the concepts and the concept hierarchy. Three possible approaches may be followed, namely, top-down, bottom-up or a combination of those. For top-down, start with the definition of the most general concepts in the domain and their subsequent specialisation. For bottom-up, start with the leaves of the hierarchy and gradually group these concepts into more general concepts. To apply the third approach, define the more salient concepts first and then generalise and specialise appropriately.
- **Step V.** Define properties of concepts. Once some of the concepts have been defined, it is necessary to describe their internal structure. Some of the terms identified in Step III that did not make it to the concept hierarchy should most probably become properties. For instance the concept "Address" can be modelled with the property "has Address" of the concept "Contact Info". Properties may be inverse; for example "Actor is Working on a Case" and "Case is being processed by the Actor": "is working on" and "is being processed by" are inverse properties. Another feature that a property may have is cardinality. This defines how many values a property can have. One may distinguish between single and multiple cardinality (one may have only one name, but several email addresses).
- **Step VI.** Create instances. This is the last step and concerns creating individual instances of concepts in the hierarchy.

These guidelines are essential not only for design nut also for maintenance of ontologies. The complexity of eGovernment concepts and processes has led to the use of multiple ontologies as shown by nearly all the FP6 projects. We give some characteristic examples in what follows.

2.1.1 Ontology maintenance and design examples

[30] SAKE: D9 Semantic Modelling of Knowledge Resources, 2007

The **ONTOGOV** project (Ontology-enabled e-Government Service Configuration) constructs services ontologies via a domain expert via a tool called **Service Modeller.** The domain expert also adds more semantics by creating instances of the following ontologies used throughout the project and represented in the next figure.

- **Domain ontology**, comprising concepts like data (e.g. name, first_name, municipality_from, municipality_to) and documents (e.g. application form, administration leaflet etc.).
- **Legal ontology**, comprising instances of process relevant law or regulations, e.g. basis of the new process is a regulation about settlement. Then several instances will be initiated in the legal ontology indicating the related law, paragraph and article.
- **Organisational ontology**, comprising instances of process relevant to organisational units, e.g. involved in the new service are the organisational units 'Registration Office' and 'Administration Office' with its roles and personnel.
- **Lifecycle ontology**, comprising instances of all (design) decisions relevant for the new service (e.g. technical or process immanent reasons), including instances of the legal and organisational ontologies.

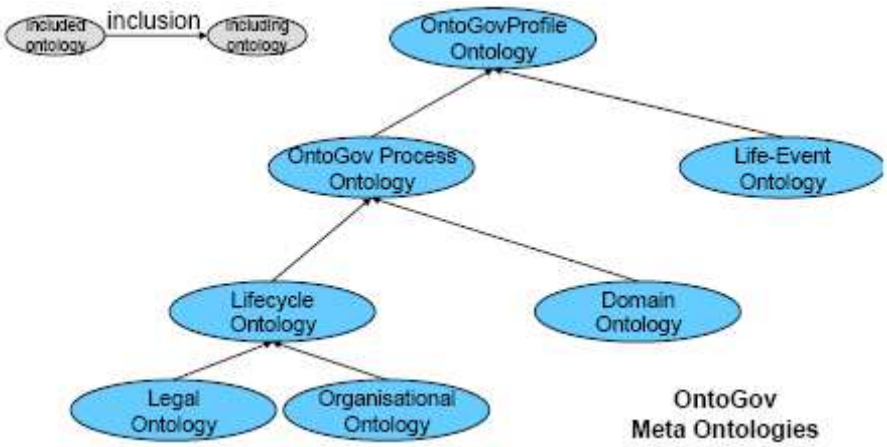

Figure 2. ONTOGOV ontologies used for modelling eGovernment services

We see here that this ontology structure goes beyond that of merely representing data at a higher level. The ontology itself is executable, belonging to the highest level 3 of ontology sophistication as described in section 1.3.1. As explained therein, the notion "executable" means that models built based on the ontology, such as the process ontology of the previous figure are ran using modules provided by some of the abstract codes associated with concepts in the ontology.

In a similar fashion, the **TERREGOV** (Impact of eGovernment on Territorial Government Services) ontology is also a multipurpose ontology: it models the domain of the public administration activities; it is used for semantic document retrieval, semantic discovery and

orchestration of web services, etc. In addition, the same ontology is handled by multiple actors responsible for providing the technologies supporting these various functionalities. The project's own developed Local Government Ontology enables local administrations to deal with information as a strategic resource. The ontology itself is reusable (a first example of data/knowledge reusability): it has been employed as the first step of the QUALEG project starting ontology, which was later expanded to fit the specific project needs.

One has to remember that a single ontology relating to public administration must provide multiple functionalities, and allow its maintenance and editing by multiple users, including the domain experts with no particular expertise in ontology engineering. This makes its handling and maintenance extremely complex.

Moving to other examples of ontologies in FP6 self-adaptable eGovernment frameworks demand specific ontology designs. The **FIT** project develops, tests and validates a self-adaptive eGovernment framework based on semantic technologies. To achieve this adaptation ability[31] a set of five distinct ontologies are used, namely Rules, Quality, Enterprise, Domain and Information ontologies. The ontologies developed are compliant with the generic development guide given earlier in this section and use "concepts" and their inclusion subdivisions as basic building blocks.

As FIT researchers explain, in order to be adaptive, the process models, which are part of the enterprise ontology, must be enhanced with adaptive information. Since the adaptation to the specific user and context is made at runtime, the criteria for adaptation have to be modelled explicitly. The information ontology must support adaptation of user interaction and information presentation on the public administration's web portal. Information presentation affects layout, content and linking. Therefore, the information ontology must describe the different kinds of information sources with their respective structure, access, and format. In addition there must be links to the content (which refers to the domain ontology) and the context in which the information source is used (which refers to the process model). The adaptation requirement demands that these relations are not fixed. Depending on a particular user in a specific process context the relevant content of the information sources is selected and the layout (e.g. font size) and additional links (e.g. to a specific help page) are determined. The following descriptions refer to the diagramme given in the figure below.

The **FITEnterprise** ontology represents different views on the enterprise, in particular the process models and the process participants which are described by their roles and — for internal participants — their places in the organisational structure

The **FITProcess ontology** is an extension of the OWL-S[32] process ontology. It extends the concept AtomicProcess with two data type properties:

- costs, which presents the costs of the process
- duration, which the duration of a process stores.

The **FITOrganisation** ontology represents the process participants which are described by their roles and—for internal participants—their places in the organisational structure.

[31] FIT: "D3: Requirements specification and process modeling formalism", July 2006

[32] http://www.daml.org/services/owl-s/1.2/Process.owl#

The **FITDomain ontology** represents domain-specific information which heavily depends on the chosen process. In the process of building permission the domain ontology contains concepts like building types (with subconcepts like one-family house, shopping mall, garage, outbuilding, and terrace), building projects (new building, rebuilding, renovation etc.) or geographical aspects. For registration and deregistration, concepts like city, states, foreigner etc. have to be modelled. The property presents of the concept 'Service' links to the 'ServiceProfile' of the OWL-S ontology to express the details of the service profile.

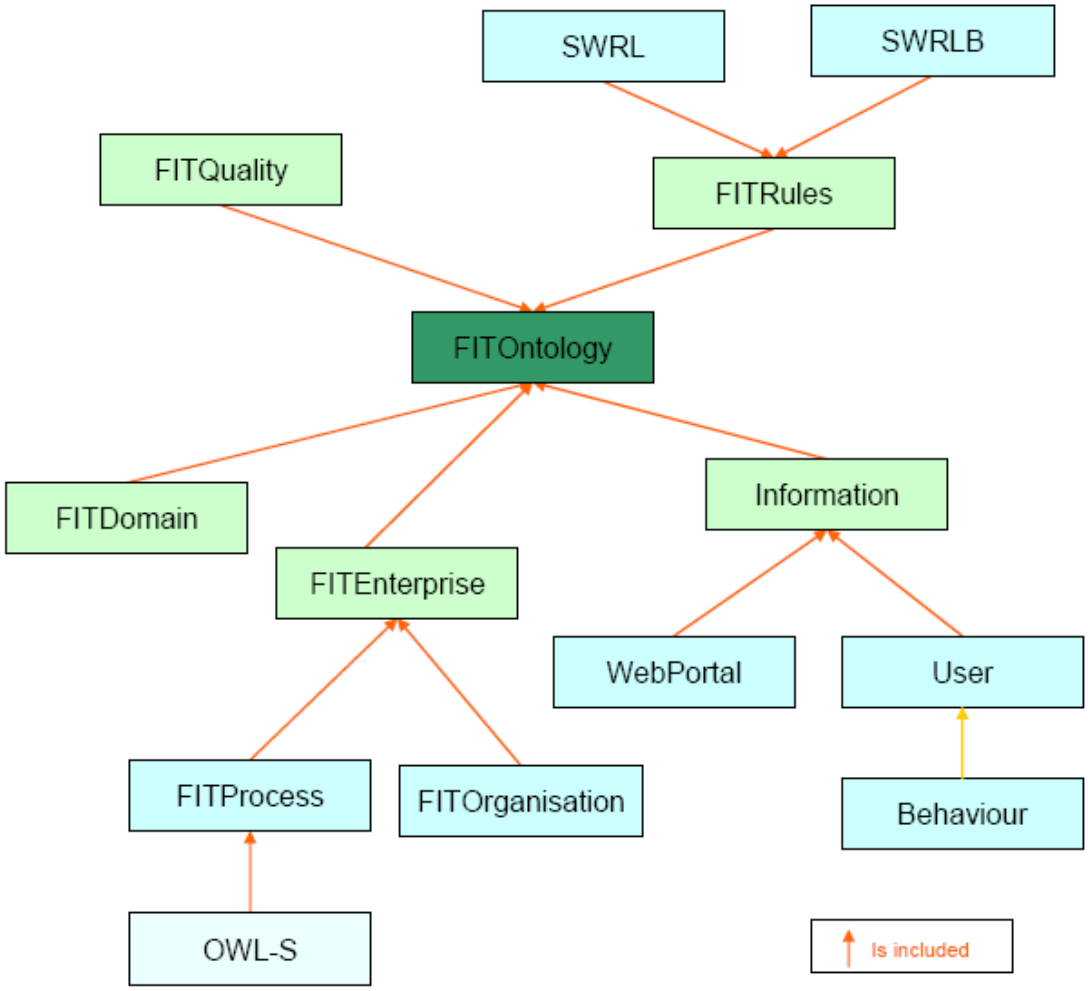

Figure 3. Overview of FIT ontologies[33]

[33] FIT:"D8: Semantic modeling of adaptable processes", January 2007

The **FITInformation ontology** describes various kinds of information with their structure and format. For the FIT project, information on the public administrations' web sites is of the utmost importance. The information of the user of the Webportal and his behaviour is modelled in the User ontology. The WebPortal ontology contains all elements, which a webpage can provide. Users interact with forms through named *controls*. All controls are stored as Elementsubconcepts. The filled in data are stored as values in the corresponding individual. The concept Document represents several documents which are available. The subconcept Form contains all available application forms. The concept CorrectForm contains as individuals all correct forms with their mandatory fields. ApplicationForm can be used to store all data of an application form temporary.

The **FITRule ontology** represents the rules and rulesets. It includes the SWRL and SWRLB ontology (see also below).

The **FITQuality ontology** contains three parts of the Quality of eGovernment services (QeGS) ontology:

- The aim of the top layer ontology is to define a minimal set of high level concepts and relations between them that are needed to describe the notion of quality of service. This layer concerns quality of service in general and models the theoretical foundations that the project has used for the development of the QeGS model.
- The middle layer ontology concerns quality of eGovernment services and models quality aspects related to e-government services.
- The third (lower) layer ontology is domain-specific. The aim of this layer is to support the different configurations of the three pilot users' systems.

Finally, to express SWRL rules in FIT two ontologies provided by Protégé-OWL and WRLTab are used, namely the **SWRL and SWRLB ontologies**.

2.1.2 Implementation of ontologies I

There are two parameters which mainly affect the extent to which ontologies such as those presented in the previous section are useful in practice.

- detection and orchestration of eServices
- transparency to public servants using the system, so that maintenance and update cab be done by non specialists as well.

Regarding the first issue, ONTOGOV services are heavily tied to QWL-S, which is not amenable to semantic data mediation and consequently to semantic integration. The same is true for FIT. That said, both projects have provided elaborate ontologies which incorporate many essential features of eGovernment processes and information. The manifestation of this is that the ONTOGOV ontology has been used by other FP6 projects such as QUALEG and SAKE. Now regarding the second issue, eGovernment ontologies are indeed complex and tedious to design due to the multitude of terms and relations among them.

ONTOGOV argues that business process flow specifications should be defined at abstract task levels, leaving open the details of specific service bindings and execution flows. This abstract level enables the definition of domain-specific constraints that have to be taken into account during the (re)configuration of a process flow. In order to model this abstract

representation of web services, the OWL-S and WSMO ontologies were extended so that they became able to better support process and lifecycle modelling.

The project illustrates their technical choices by application scenarios and shows the advantages of their chosen methods which utilise the principle of working only with **instances of meta-ontologies**[34]. This allows for strong governance of the modelling as a whole, with inherent semantic checks, which frameworks such as BPEL cannot provide. For example, adding the same organisational unit to two atomic services in a sequence will evoke a warning (as usually the activities will be performed as one) even though the process flow per se is correct.

As a verification of the genericity of the ontologies development process, we refer to the "Business Process Semantic Analysis and Modelling" task of the **SAKE** project (same reference as at the beginning of this section), which was implemented by using the ONTOGOV process ontology and their associated Service Modeller tool. The project researchers extended the standard set of metadata of the ONTOGOV process ontology with the legal, organisational and lifecycle aspects defined in the respective legal, organisational and lifecycle ontologies. All these ontologies were used for the annotation of eGovernment services so as to enable better and easier management of them. Additional modifications were carried out to the Service Modeller tool's source code so that it could import and use the SAKE ontologies. The result is a process ontology for each pilot, which gives a detailed description of the modelled process flow. It describes how the process works, which includes the specification of activities in the process, dependencies between them, information resources that are used as inputs/outputs of activities, the Public Administrators responsible for performing activities, the potential required communication interactions in activities between the responsible person for carrying out the activity and a third party, etc.

Technically, the ONTOGOV Service Modeller tool stores the process model directly in an XML file. This file constitutes the Process Ontology in OWL-XML presentation syntax, something which was the main reason for selecting the specific tool, besides its intuitive interface and the Open Source feature of the software. The OWL-XML presentation syntax differs from the standard OWL-RDF syntax, but can easily be manipulated, for instance by the KAON2 reasoning engine. This allows the thus developed process ontology to be fully compliant with the rest of the SAKE ontologies which are developed in Protégé.

In a similar fashion TERREGOV uses OWL-S for interoperability purposes within the project. As this cannot provide all the functionalities required for the semantic web services discovery, choreography and orchestration, researchers embedded proprietary instructions within OWL, enabled through the application of a **Simple Ontology Language (SOL)** with supporting such extended functionalities. This however, complicated maintenance for the TERREGOV ontology.

The solution chosen by the project was to establish a **central body** responsible for the maintenance of the TERREGOV ontology consisting of ontology engineers. This central body receives the so-called Requests For Changes (RFC) created by the domain experts, as well as technologists developing specific system components required to provide various functionalities of the system. These requests are then reformulated in SOL, and included into

[34] D. Apostollou *et al*, "Towards a Semantically-Driven Software Engineering Environment for eGovernment", in "E-Government: Towards electronic Democracy", M. Bohlen *et al* (Eds), Springer, Berlin 2005

the ontology through its compilation into the OWL format used by the system. Official ontology versions are released periodically.

The **Simple Ontology Editor (SOE)**, however, provides a user-friendly interface to end users, and in particular domain experts. It allows them to relatively effortlessly create, modify and edit the ontology. Such modifications are then exported as RFC and subsequently provided to the ontology engineers for further integration. The ontology maintenance utilities are provided in TERREGOV as a toolbox containing components supporting several functionalities required by the TERREGOV methodology for maintenance:

- **SOL2HTML** – a compiler producing HTML pages displaying the TERREGOV Ontology, and its various elements.
- **Simple Ontology Editor** (SOE) – a graphic user interface utility allowing display and editing of OWL-formatted ontology.
- **SOL2OWL** – a compiler transforming Simple Ontology Language (SOL) into OWL.
- **Search AND Retrieval Application (SANDRA)** – a natural language processing engine enabling automatic extraction of vocabulary used in ontology from free-text documentation.
- **SOL2RULES** – a compiler producing rules from SOL rules, or virtual objects.

2.1.3 Implementation of ontologies II

The group of FP6 projects chosen in the previous section is indicative of non-fully WSMO based implementations, in contrast with the newer group represented by Access-eGov and SemanticGov which exclusively relies on this newer and more flexible semantic technology which, due to its inherent data mediation features, can tackle issues of data heterogeneity.

Both projects adopt a service-oriented approach (in the sense of the table above) to implement semantic web services, a trend which is almost uniform for nearly all 4th IST call FP6 projects in eGovernment. The underlying concept and technology revolves around the definition of a semantic mark-up for web services so as to provide higher expressivity then traditional XML-based descriptions. The projects make use of the **Web Service Modelling Ontology (WSMO)**, which provides a conceptual model describing all relevant aspects of general services accessible through a web service interface. At the same time, it adheres to the principles of loose coupling of services and strong mediation among them. WSMO defines an underlying model for WSMX (see previous description on technology trends), a semantic web services execution environment as well as WSML an ontology language used for the formal description of WSMO elements. Thus, WSMO, WSML and WSMX form a complete framework followed by both projects facilitating all relevant aspects of the semantic web services.

Both projects address the WSMO top-level conceptual model which consists of ontologies, web services, goals and mediators. In this way, interoperability among some components (goals, web services) cab be achieved through a common (domain) ontology.

2.1.4 More ontology examples

The relationship between citizen's goals and services as well as the problem of matching citizen's goals to services is a known problem, which has been investigated in the literature.

Using a formal representation of the citizen's goals and the service's outputs a solution has been suggested[35,36] under the terms **Government Enterprise Architecture (GEA).** Access-eGov and SemanticGov have both considered using the architecture. Semantic-Gov exploits the abstract GEA-model and defines a mapping to WSMO constructs to make effective use of WSMO's semantic matching properties. This will also be included in the development of the Access-eGov platform[37]. (Cf. Wang et al 2007).

Semantic-Gov notes that according to the GEA architecture, a service model for public administration services, termed the PA service model has been defined. The PA service model however does not follow any formal service semantics thus the goal of introducing semantics for eGovernment architectures lies in a combination of the formal WSMO service model with the PA domain specific model from GEA. Based on these considerations the project has defined the WSMO-PA ontology. A map between GEA and WSMO-PA ontology concepts is given in the above table.

GEA-compatible concept	WSMO-PA concept
Service identifier	WSMO Web service Non functional Property dc: identifier
Public Service	WSMO Web Service
Service Description	WSMO Web service Non functional Property dc: description
Service Provider	Instances of PA Domain Ontology concept GEA_PA_Entity
Service Interface Location	WSMO Web service Non functional Property
Law	Instances of PA Domain Ontology concept GEA_Law
Output	Instance of PA Domain Ontology GEA_EvidencePlaceholder concept
Effects	WSMO Web Service/Goal Effects
Consequences	Web Service/Goal Orchestration
Evidence placeholders	Instances of PA Domain Ontology concept GEA_EvidencePlaceholder
Actors	Instances of

[35] Peristeras V., Tarabanis K., "Reengineering Public Administration through Semantic Technologies and a Reference Domain Ontology", in: Proceedings *AAAI Spring Symposium "Semantic Web Meets eGovernment"*, Stanford University, March 2006, Technical Report SS-06-06, AAAI Press, Menlo Park, CA, 2006, pp. 56-63.

[36] Goudos, S. K., Peristeras, V., Tarabanis, K., "Mapping Citizen Profiles to Public Administration Services Using Ontology Implementations of the Governance Enterprise Architecture (GEA) model", in: Abecker, A., Mentzas, G. and Stojanovic, L. (eds.), *Semantic Web for eGovernment,* Proceedings of Workshop at the 3rd European Semantic Web Conference (June 12, 2006, Budva, Serbia & Montenegro). http://www.imu.iccs.gr/semgov/index_files/Proceedings.html

[37] Wang X., Vitvar T., Peristeras V,, Mocan A., Goudos S., Tarabanis K., "WSMO-PA: Formal Specification of Public Administration Service Model on Semantic Web Service Ontology", in: Hawaii International Conference on System Sciences (HICSS), Jan. 2007, Hawaii. http://www.semantic-gov.org/index.php?name=Web_Links&req=visit&lid=63

GEA-compatible concept	WSMO-PA concept
	PA Domain Ontology concept GEA_Societal_Entity
Preconditions	WSMO Web Service/Goal Preconditions/assumptions
Evidence Provider	Instances of PA Domain Ontology concept GEA_PA_Entity / Instances of PA Domain Ontology concept GEA_Societal_Entity
Evidence Producer	Instances of PA Domain Ontology concept GEA_PA_Entity
Evidences	Instances of PA Domain Ontology concept GEA_Piece_of_Evidence or Attributes of PA Domain Ontology concept GEA_EvidencePlaceholder

Table 1. SemanticGov: GEA to WSMO concept mapping

Domain ontologies in Access-eGov are used to represent functional and non-functional properties of a particular service. The following domain ontologies are used to describe the concepts for non-functional properties of services for AeG:

- Fees: Describes fee that citizen has to pay in order to use the service.
- Forms: Services may require information and/or they might provide information in the form of documents or forms. The Forms ontology will be used to describe these kinds of (both mandatory and optional) input and output of service.
- Input and output artefacts: For inputs and outputs that cannot be described with the Forms ontology (for example, an artefact like a license plate), AeG will provide special ontologies that can be used to describe them See below for specifics.
- Administration: Every service is provided by one or more administrations. At least the following information related to service provision of an administration must be described:
    - Responsibility
    - Office hours/availability
    - Address and contact information
    - Physical accessibility constraints

Each domain ontology may have as many mappings as the number of ontologies in the Ontology repository. This way, an ontology can have m:n-relations to other entries in Ontology Repositories. All ontologies that the Access-eGov platform has to be familiar with are stored in persistent repositories that are accessible to all peer-instances. All ontologies consist of a core set of public service concepts to sufficiently describe services. Actual ontologies that are used for annotation processes and for lookup during automated service retrieval will have to be provided by the respective public service provider.

Input and outputs are defined in special ontologies. These ontologies are not only used to define the requirements for Goals but are also used to define the service profile of a service. By using the description of inputs and outputs in both the definition of Goals and of service

profiles, Goals and services can be automatically matched by the Access-eGov Matching Component when a user searches for a specific Goal.

Besides the Goal and input/output-ontologies, Access-eGov will provide ontologies that define concepts, which are related to other properties of services. For example, a user may want to locate services, which he can pay for by a credit card or which provide a certain level of data-encryption. These properties are called *non-functional* properties of services.

The last example of an ontology comes from the **BRITE**[38] project, which deals with business registers in Europe as its operational domain. These registers can supply critical information for many inter-business processes, for example the majority of eProcurement processes. Information resources, typically managed by these registers, must bear clear evidence of authenticity and accuracy as part of their core mission and operating business entities should be able to rely on this commitment.

The BRITE approach uses domain and process ontologies for interoperability and process-oriented information support. The BRITE Domain Ontology (BDO) aims at facilitating information integration, e.g. for Business Register (BR) spanning queries or formal policy rule checking, enabling communication between multiple business registers, and easing the implementation of BR, e.g., in new EU countries. The main focus of the BDO development is the proper balance between local perspectives taken by the various National Business Registers and the more global perspective required to enact the cross-border directives imposed by the EC. Another objective of the BDO is to provide a reference framework for assessment of local registry information structure.

To establish the BDO, the domain expertise of National Business Registers and the EBR together with the modelling work already performed by these institutions is taken as seed for a minimum agreement on relevant concepts and their semantics. These concepts are then used to build the **formal BRITE Domain Ontology**. Mapping rules are defined that map the BRITE Domain Ontology onto the local scenario.

BRITE faces the multi-domain context by analysing the existing differences from an ontological perspective. Ontologies allow mediation techniques on a semantic level by providing semantic descriptions of resources to enable resolution of mismatches at data-level. The mediation strategy[39] considers three options:

- **Ontological mappings**, used in the same sense as SemanticGov and Access-eGov to map definitions while none of the ontologies involved are changed
- **Ontology alignment** used to find classes of data that are semantically equivalent, but indeed not necessarily logically identical (e.g. full name and first name). Ontological alignment requires the necessity to alter at least one of the concepts involved to guarantee that the overlapping parts of the ontology are being aligned.
- **Ontological merging** which describes the creation of a new ontology used in place of the old ontologies, via operations such as unification or intersection.

[38] Milani P., Mondorf A., Process Ontology for a collaboration framework among Business Registers in Europe (BRITE project), eGovINTEROP'07 Conference, Paris, October 2007

[39] Milani P., Mondorf A., Process Ontology for a collaboration framework among Business Registers in Europe (BRITE project), eGovINTEROP'07 Conference, Paris, October 2007

In addition to the usual data and process level mediation BRITE provides another tool for information integration: **service-level mediation**. This also uses ontologies to enable resolution of mismatches at service-level, by addressing issues such as:

- **Differing functionalities**
  - Functionalities of a provider and a requester often do not match exactly.
  - Matching of given requests; e.g. a requester wants to travel from Cologne to Vienna without specifying the means of transportation (e.g. bus, train, plane) and a provider only offers tickets for the train. Thus, the ticket of the provider should be only available under the condition that the requested ticket is a train ticket.

- **Differing processes**
  - Address conflicts that occur due to varying processes which should cooperate automatically.
  - Mismatches often occur as a result of different interaction behaviours intended by the business process and the client. A successful interaction for the purpose of consumption or interaction then fails on a behavioural level, e.g. deadlock situation: at some point a client may expect an acknowledgment while the process waits for input.

2.2 Ontologies for knowledge management

The definitive knowledge management project in FP6 is **QUALEG** (Quality of Service and Legitimacy in eGovernment) which proposes a knowledge management model for the support of multilingual applications in the field of eGovernment. The model is based on a global ontology, manually designed for a specific domain, and local contexts, associated with ontology concepts. The combination of ontologies and contexts lends itself well to multilingual applications in which a single ontology fails to capture all nuances that stem from language and cultural differences. The single ontology system proposed, with associated concepts in multiple languages, provides a framework that is both versatile and flexible. The system, functions simultaneously in multiple languages, is low-maintenance, and is easily extended in and adapted to different languages. The model captures cultural as well as lingual differences using contexts, thus allowing easy customisation across cultures and languages. It is a prime example of a context-aware, meaning-capturing system, which can extract information over unstructured data.

The QUALEG system is modular and uses web services and BPEL coupled with workflow models. The main components are as follows:

- **Workflow Management System (WMS)**, which integrates web services and workflows. WMS consists of three specific modules (the composer, execution, and repository) that aim at creating and maintaining workflows. The outcome of this component is a set of published Web Services.

- **Datamart,** a component which stores indicators that relate both to performance of government services, and satisfaction of citizens collected through questionnaires. This component comprises of services and interfaces with other components.

- **Ontology Management System,** used for building and maintaining ontologies suitable to the QUALEG scope. The existence of a domain ontology built by knowledge experts is essential since a lot of components use it in order to classify existing knowledge. The modules of the ontology management system can be called from the external modules of QUALEG system but they have also the capacity to call other components. The ontology management system depends directly on the Datamart and Agora system; the Datamart is the knowledge repository from which the system can retrieve the information, while Agora is the search field of the ontology words.

- **Questionnaires Composer**, which provides questionnaires constructed in a semi-automated way according to the user that accesses the system. A user profiling mechanism is also a sub system of this component.

- **AGORA Service**, which offers a knowledge space for debates between citizens, politicians, parties and civil society. It provides access to policy evaluation indicators and a mechanism for semantic document search. Agora also includes the published questionnaires area and the common topics area such as information and news. This component is based on open source software (portal/content management systems) as far as the front-end is concerned.

- **Knowledge Extraction,** a component which extracts semantic descriptors from debates, mails, docs, using existing free text based algorithms. It uses the QUALEG ontology for parsing electronic data. Furthermore, the knowledge extraction component will implement methods in association to ontology representation. The Ontology Management System format has a direct dependency on the format of the information that the Knowledge Extraction can handle. The Ontology Management System also includes an interface that will allow the Knowledge Extraction module to access the ontology and the information it contains. The Datamart dependency includes the predefinition of the types of possible queries in order to verify that all the information can be accessed via the Agora Service.

- **Intelligent Agents**, to provide synchronisation among several components running asynchronously.

For the deployment in QUALEG the first step included starting with an existing ontology and expanding it according to project needs. The ontology used was the local government ontology developed for TERREGOV.

2.2.1 Ontologies for legal knowledge

Knowledge interchange formats are ontologies of concepts which include knowledge bases, specific terminologies, rules and normative statements. Existing ontologies are used in social sciences and the law[40] such as DOLCE (Descriptive Ontology for Linguistic and Cognitive Engineering), SUMO (Standard Upper Merged Ontology), Stamper's Norma Formalism, LLD (Language for Legal Discourse Ontology), FBO (Frame-based ontology of law), FOlaw (Functional Ontology for Law) Valente's Functional Ontology of Law, LRICore, CLO, JurWordNet.

[40] ESTRELLA project: D5.2 Survey of existing document databases and legal ontologies, 2006

FOLaw is a 'functional' core ontology for law, supporting the building of legal knowledge systems, due to the fact that it reflects an understanding of types of knowledge and dependencies existing in legal reasoning. FOLaw presents two distinctive sets of concepts that are typical for law: normative terms (and their definitions and axioms), and responsibility terms (liability, guilt, causation, etc) found in legal theory.

The main function of a legal system is to regulate social behaviour (law as a social control system). By using this functional view of law, categories of legal knowledge are distinguished which are represented in the ontology. One may distinguish six categories of legal knowledge:

- Normative knowledge is knowledge that defines a standard of social behaviour. It thereby sets down behaviour of the people in the society. The standard is defined by issuing individual norms, expressing what ought to be the (compliant) case.

- Responsibility knowledge is legal knowledge that either extends or curtails the responsibility of an agent for its behaviour. Its function is to provide the legal means to reject the common idea that someone is only responsible for what one causes. Prime examples of this 'fault-causation-responsibility' problem are to find in tort law and fraud cases.

- World knowledge is legal knowledge that describes the world that is being regulated. It delineates the possible behaviour of persons and institutions in a society. Thereby it provides a framework to define what behaviour ought (and ought not) to be performed.

- Reactive knowledge is legal knowledge that specifies which reaction should be taken (and how) if an agent violates a primary norm. Usually, this reaction is a penalty (fine and/or imprisonment, etc).

- Meta-legal knowledge is legal knowledge about legal knowledge, or legal knowledge that refers to other legal knowledge. It deals with legal principles in case of conflict of norms.

- Creative knowledge is legal knowledge that allows the creation of previously nonexistent legal agents, bodies and entities (e.g. law enforcement agency, a contractual agreement).

The **ESTRELLA** project is to create a tool that will facilitate interchange between major proprietary formats already existing in the market, without becoming dependent on them. This is done by developing and validating an open, standards-based platform, allowing public administrations to develop and deploy comprehensive legal knowledge management solutions. Features of the platform include:

- Semantic Web compatibility, which means that LKIF should be compatible with emerging XML-based standards of the Semantic Web serving reusability and interchangeability of information, modularity and layering.

- Legal-oriented expressiveness, as LKIF is intended for users in the broadly understood market of legal knowledge systems.

- Domain specific format (law), which is however task-independent so as to provide universal support for tasks without dependence on the choice of a particular legal domain.

Using these prerequisites as a starting point, the EXTRELLA researchers established the characteristics and specifications[41] which language for LKIF should have. Unambiguous representation necessitates a language with constructs of precisely defined meanings. This will allow syntax and semantics to determine the way conclusions are derived: via a calculus associated with the chosen language. This calculus will also be able to specify whether a set of expressions is sufficient to obtain the given information. A challenging question that emerges is to optimally choose the minimum necessary set of expressions needed to derive all the remaining implicit knowledge.

The dominant trade-off is the conflict between increased expressiveness and computational tractability and compatibility with established standards. Computational tractability is the boundary where expressiveness and complexity should be limited to. Unless the language is simple (reduced expressiveness and complexity), automated reasoning may become an intractable problem. The least that should be done is to guarantee that some fragments of LKIF maintain tractability or find a variant of representation preserving these fragments.

The architecture of LKIF is a resulting compromise of requirements, sometimes mutually exclusive. Terminological knowledge and rules should be distinctly layered. The modularised approach to representation eases interchangeability and reusability and is in line with the Semantic Web's philosophy. Within the legal context, modularity implies a clear distinction between ontologies, which represent common knowledge about applicability of law, and normative knowledge, which is expressed in rules. The former may be exchanged between different jurisdictions; yet the latter is less reusable, as rules may vary between legislative acts and systems.

Translations between LKIF and other proprietary formats should be done with minimal loss of information, which means that the language should be sufficiently generic to embrace expressiveness of other formats. Nevertheless, there may be formats where a common denominator for a correct translation may not be established[42]. The reader is reminded here of the remarks made in the previous chapter (sections 1.3.1 and 1.5) regarding the computational intractability of highly expressive ontology languages and the need to integrate ontology languages and even logics for true information integration.

Moreover, the language should handle exceptions to norms, which is a common case in the legal domain. Normative rules should not only appear as formulae constructed of terms, but also as objects with properties, embracing information such as date of enactment, period of validity and date of repeal etc.

Mechanical jurisprudence occurs when legal reasoning is naively seen as a representation of norms, in terms of logical formulae and application of deductive rules of inference. LKIF should provide a representation that goes far beyond than that, namely, to provide sufficient expressiveness in order for external engines to be able to cover deductive as well as inductive forms of arguments, by generalising the concept of an inference rule.

Implementation issues

[41] ESTRELLA: D1.1 Specification of the Legal Knowledge Interchange Format, 2007

[42] ESTRELLA: D1.2 Formal specifications of the knowledge formats of the participating lkbs vendors, 2006

The Legal Knowledge Interchange Format (LKIF) of the ESTRELLA project is an OWL ontology of legal concepts allowing legal knowledge bases, encompassing specific terminologies, LKIF rules, and normative statements to be represented in OWL and stored as OWL files.

Within the LKIF format one may distinguish the terminological part (the ontology) and the language. The language part has been defined as a layered language, whose building blocks are: OWL-DL, DLP, DL-safe subset of SWRL, but also proper LKIF rules. Users can select the combination they prefer, in view of desired computational features or expressive power.

OWL DL is one of the Semantic Web languages recommended for ontological modelling. Among the features that make it attractive for LKIF are: firm logical underpinning, decidability, compatibility with XML-based standards, good support by existing tools and growing popularity in web applications.

The rule layer of LKIF is supported by LKIF rules, a partly novel formalism developed under the ESTRELLA project. LKIF rules are not fully compatible with W3C standards; however none of the recommended rule languages, predominantly SWRL, seemed to satisfy the high expressiveness requirements derived from the goals of the project. However, a new W3C proposal has been recently put forward.

Normative statements are handled at the level of terminological knowledge by a set of predefined deontic terms concentrated around the concept of *subjunctive betterness*. Deontic reasoning is inherent to the domain of law. LKIF supports the representation of deontic aspects of legal knowledge by providing the necessary terminology equipped with the minimal semantics required for modelling normative statements. All elements of the deontic vocabulary, including the operators, are represented as OWL properties and classes in the module norm of the LKIF-Core ontology. Central to deontic reasoning is the notion of *deontic choice* which states that if "*it ought to be **a** given **b**"*, then the agent should prefer any choice compliant with ***a***. Choices that meet this criterion are deemed *better* than others, introducing the most fundamental axiological intuition, namely that of the *normative preference* or *subjunctive betterness.*

As elaborated in the previous chapter LKIF is rather too expressive for computational tractability, therefore interoperability and integration suffer. The ESTRELLA researchers are aware of this, but mention instead other vendors may be attracted into migrating to LKIF[43], as LKIF should enable interchange between proprietary formats already in use by vendors in the legal markets and therefore be sufficiently generic to embrace expressivity of other formats.

2.3 Data mining and ontologies for fraud detection

Data mining is used in the **iWebCare**[44] project to automatically extract structures from data and generate predictions in order to assist fraud inspectors in identifying novel cases of fraud. The method helps them concentrate their search on the most suspicious cases in large databases of possible fraud cases. Data mining techniques can extract automatically structures from data and generate predictions on new fraud instances in order to assist fraud

[43] ESTRELLA: D1.2 Formal specifications of the knowledge formats of the participating lkbs vendors, 2006

[44]Dimakopoulos *et al*, "iWebCare: an Integrated Web Services Platform for the Facilitation of Fraud Detection in Health Care e-Government Services", iWebCare project http://iwebcare.iisa-innov.com

inspectors in identifying novel cases of fraud. This helps them concentrate their search on the most suspicious cases in large databases of possible fraud cases. Data mining techniques and approaches that are used in the iWebCare project are based on the CRISP process model[45] and available open-source environments like R[46], Yale[47] and Weka[48].

Ontologies can play a vital role in both rule-based and data mining fraud detection approaches. A knowledge base may use an ontology to specify its structure (entity types and relationships) and its classification scheme. In such a case, the ontology, together with a set of instances of its classes constitutes the knowledge base. Ontologies can capture and represent domain knowledge due to their expressive power. Ontologies can be used in the data mining area as they can select the best data mining method for a new data set: when new data is described in terms of the ontology, one can look for a data set which is the most similar to the new one and for which the best data mining method is known. This method is then applied to the new data set; this means that there is no need for trying out every known method on the new data set; instead, the ones which appear to be the most promising can be directly selected.

The iWebCare methodology defines a process for identifying, measuring and treating fraud in the context of eGovernment services. This process comprises three steps:

- establishment of the fraud context
- identification of fraud within this context
- transformation of this information into an ontological model.

Establishment of the fraud context within an organisation is done through a business process modelling procedure that records fraud susceptible business processes of the organisation and their context. Fraud identification involves description of potential fraud cases and corresponding detection methods, accomplished via intra-organisational knowledge and/or via data mining methods in order to extract unknown fraud patterns.

[45] http://www.crisp-dm.org

[46] http://www.R-project.org

[47] http://www.sf.net/yale

[48] http://www.cs.waikato.ac.nz/~ml/weka

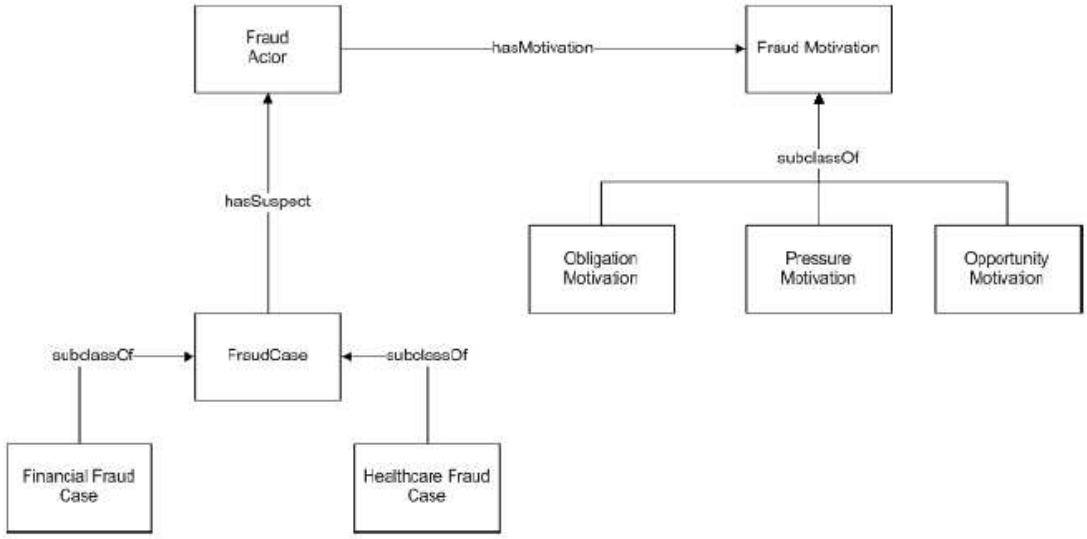

Figure 4. iWebCare: Generic fraud ontology

The methodology is an iterative procedure. In order to minimise the effort required by each iteration, the project created a **generic fraud ontology** that acts as the basis for building domain specific fraud ontologies.

Ontologies are built in a layered architecture in grades of genericity in order to maximise modularity, reusability and extensibility. In this way, the highest layer, namely the Generic Upper Ontology, captures generic and domain-independent knowledge that helps minimise redundancy and duplication of knowledge within the overall ontology. The next layer, namely the generic fraud ontology contains concepts representing fraud actors, fraud cases, etc. and relations linking actors with motivations and cases with actors. Part of the the generic fraud ontology is shown in the figure above.

iWebCare runs pilots in social security organisations, namely NHS, UK and TSAY, Greece.

3 International Dimension

3.1 The global perspective

3.1.1 Introduction

Information integration plays an important role in every attempt to implement eGovernment globally. Understandably, the degree of progress varies significantly from country to country as a result of many parameters ranging from the most obvious such as technological level and economical and political situation to the least obvious such as population density and geographic characteristics. The next paragraphs present current eGovernment-related research initiatives mainly focusing on the international dimension of information integration. A coarse classification of these initiatives distinguishes the following categories:

- International cooperation, further split in:
  - Cooperation under the auspices of international organisations;
  - Cooperation stemming from joint initiatives;
- Non-EU national initiatives.

This classification groups together initiatives with similar rationale:

- International cooperation sponsored initiatives mainly focus on cross-border eGovernment issues, exchange of practical information (lessons learnt) and further perspectives.
- National initiatives usually form part of a wider medium to long term eGovernment policy or strategy and focus on the specific hurdles existing at a specific country level.

Both classes of projects are important when trying to address the fundamental problem of information integration. Examination of the results achieved will help eGovernment practitioners benefit from international research and practice and gain valuable experience.

3.1.2 Scope and direction

“Information integration” is a rather loose term since:

- The term “Integration” may be defined in a variety of ways, mainly with respect to the different angles information is observed from: technical (directly related to data), semantic (related to meaning), or organisational (context within which information exists);
- It may refer to integration at different levels: from user’s interface to information storage;
- The definition of “Information” may be very wide or very narrow, e.g. with respect to cultural, lingual, social aspects of the relevant time and space.

There are two main directions in which information integration is addressed in various current research activities:

- Policy level, examining **cross-border** legislative, social or cultural differences, trying to establish specifications, regulations or policy documents.

- Actual use cases, where the main challenge is to form a multidisciplinary team of experts and provide all the enabling science and technology to address the **real life cases** up to the actual service deployment.

EU and Australian research usually falls into the first category; USA and Asian projects mainly follow the second approach.

3.2 International research topics in information integration

A serious persistent problem in eGovernment is how to integrate each government agency's resources to form cross-agency services for citizens. The next paragraphs present the different dimensions in which this integration can be accomplished and the research directions currently followed internationally.

3.2.1.1 Technical aspects

Integration is the single most important issue that needs to be solved when trying to access information residing in different systems that are independently administered and have independently evolved. "Information islands" is the term commonly used and eGovernment aims at bridging the space between them. Luminita Hurbean[49] promotes the use of ERP technology in eGovernment; however its success is directly linked to correct business strategies and processes and not to technological solutions.

It seems that there is an increasing interest in the integration of specialised types of information, and more specifically geospatial information[50]. Geospatial information has been an important type of information, which until recently faced significant problems due to the unavailability of large amounts of memory and storage at affordable prices. The continuous move towards smaller, faster and cheaper storage solutions has resulted in more attractive geographical information systems. In addition, the recent evolution of low-cost high-speed connectivity has identified a much broader audience for systems integrating information from many different locations, within the borders of a country or internationally.

An interesting observation is that eGovernment research generally does not focus on Information Quality (IQ), although this is central to government agencies' willingness to share or to use shared information. In a study by Ralf Klischewski and Hans Jochen Scholl[51] it is

[49] Hurbean Luminita, "Information Integration - An Essential Pillar in e-Government Development", West University of Timisoara, May 30, 2008, http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1282028

[50] Jun-san Zhao, Xue Li, Yaolong Zhao, Tao Xu, Xiaodong Fu, "Methods And Implementation Of The Geospatial Databases Integration And Update Towards E-Government", http://www.commission4.isprs.org/workshop_hangzhou/papers/203-208%20Junsan%20Zhao-A049.pdf

[51] Ralf Klischewski, Hans Jochen Scholl, "Information quality as capstone in negotiating e-government integration, interoperation and information sharing", Electronic Government, an International Journal 2008 - Vol. 5, No.2 pp. 203 – 225, http://www.inderscience.com/search/index.php?action=record&rec_id=16647&prevQuery=&ps=10&m=or

demonstrated how IQ serves as an indispensable capstone and common ground in cross-agency information-sharing and interoperation projects. In particular, there is a need for distinction between desired, negotiated and emergent IQ and how these are linked to the choice of organisational arrangements and utilised standards. The parallel follow up of these three types of IQ provides significant information, which can be used in order to draft more efficient information integration policies and procedures.

The "International Workshop on Intelligent E-government"[52] was held as part of the 2008 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology[53] includes "Information Integration" in its research topics. Its proceedings have not yet been published.

3.2.1.2 Semantic aspects

Semantic aspects of information integration at the level of eGovernment are an active research subject in the international scientific community. Several research papers address this issue, and the following list is indicative of the most relevant recent announcements:

- Graciela Brusa[54] discusses how information integration must be performed through the use of eGovernment-specific ontologies, which show significant benefits in comparison to standard ontologies addressing a wider scope.

- Wenyu Zhang[55] proposes the use of grid technology for the storage and retrieval of diverse information in the context of eGovernment, as this technology supports better utilisation of hardware and network resources without the need for significant central investment. In fact, grid technology seems to be ideal for the eGovernment area as it concurs with the general decentralisation trend of most public administrations worldwide. What remains to be tackled is the sensitive issue of proven adherence to the necessary security policies that form an integral part of eGovernment.

- Jingtao Zhou et al propose the SGII[56] semantic grid-based information integration. In this work the information integration fundamental problem is addressed using the decentralisation properties of grid technology which allows better control to the integration process at semantic level.

[52] http://management.dlut.edu.cn/WI-IAT%2708_IEG/index.htm

[53] http://maebashi-it.org/wi-iat08/wi08/index.html

[54] Graciela Brusa, María Laura Caliusco, Omar Chiotti, "Enabling knowledge sharing within e-Government back-office through ontological engineering", Journal of Theoretical and Applied Electronic Commerce Research, Volume 2, Number 1, p.33-48, April 2007.

[55] Wenyu Zhang, Yan Wang, "Towards building a semantic grid for E-government applications", WSEAS Transactions on Computer Research, Volume 3, Issue 4, pp 273-282, April 2008.

[56] Jingtao Zhou, Shusheng Zhang, Han Zhao and Mingwei Wang, "SGII: Towards Semantic Grid-Based Enterprise Information Integration", Grid and Cooperative Computing - GCC 2005, Volume 3795/2005.

- Grid technology is also discussed by Xiufen Fu et al.[57] and Ying Li et al.[58]

- Niels Barnickel et al[59] propose the use of cross-ontology semantic web service composition to achieve interoperability in eGovernment.

The integration of geospatial information is a typical example of a problem efficiently addressed by semantic approaches: users in general have to cope with distributed heterogeneous data sources in their quest for appropriate resources suited to particular situations[60].

The use of metadata in eGovernment specifics is addressed in detail in a research project funded by the Australian Research Council:[61] Descriptive metadata, i.e. structured context-rich information about business processes, agents, and information resources, is a vital tool in managing business transactions and related information objects in complex intranet/internet environments to support eBusiness and eGovernment. Implementation of record keeping metadata standards is problematic as metadata generation and deployment are resource-intensive and application-specific. This project develops a proof of concept prototype to demonstrate how standards-compliant metadata can be captured ONCE in particular application environments and then reused MANY times across business applications and in different environments. Implementation of the prototype in a test-bed site provides a model for best practice.

3.3 Activities

3.3.1 International Cooperation

3.3.1.1 UN

The United Nations (UN) does not directly fund research in eGovernment. However, eGovernment is considered as being very important in a country's development process and, as such, the UN closely monitor the advancement of eGovernment among their constituent nations.

[57] Xiufen Fu, Ding Peng, Haishui Xu, Yansheng Lu and Yinwei Zhan, "Research and Implementation of E-Government Information Portal Based on Grid Technology", Computer Supported Cooperative Work in Design II, Volume 3865/2006.

[58] Ying Li, Minglu Li and Yue Chen, "Towards Building E-Government on the Grid", E-Government: Towards Electronic Democracy, Volume 3416/2005

[59] Nils Barnickel, Matthias Fluegge, Kay-Uwe Schmidt, "Interoperability in eGovernment through Cross-Ontology Semantic Web Service Composition", Proceeding of the Workshop on Semantic Web for eGovernment 2006 Workshop at the 3rd European Semantic Web Conference 12 June 2006, Budva, Serbia & Montenegro.

[60] Vlad Tanasescu, Alessio Gugliotta, John Domingue, Leticia, Gutierrez Villarias, Rob Davies, Mary Rowlatt, Marc Richardson, "A Semantic Web GIS based Emergency Management System", Proceeding of the Workshop on Semantic Web for eGovernment 2006 Workshop at the 3rd European Semantic Web Conference 12 June 2006, Budva, Serbia & Montenegro.

[61] Create Once, Use Many Times - The Clever Use of Metadata in eGovernment and eBusiness Recordkeeping Processes in Networked Environments, Australian Research Council (ARC) Grant LP0347844, 2003-2005

The UN report[62] large differences between the five regions in terms of eGovernment readiness, with Europe (0.6490) having a clear advantage over the other regions, followed by the Americas (0.4936), Asia (0.4470), Oceania (0.4338) and Africa (0.2739). Asia and Oceania were slightly below the world average (0.4514), while Africa lagged far behind.

The UN monitor eGovernment in several different aspects including the presence of administrative information in the World Wide Web and the availability of electronic services to citizens, among others. They also monitor the effects of eGovernment in society. In addition, the United Nations General Assembly (UNGA) promoted Resolution 58/199 that encourages national and international research and development initiatives to create a general ICT security culture and protect critical ICT infrastructures.[63]

As commented by researchers of the eGovRTD2020 Specific Support Action[64], the UN have indeed shown strong support for the use of ICT in government in the Asia Pacific region. This is manifested by the establishment of the Asia Pacific Development Information Programme, which also includes a portal to other Asia Pacific eGovernment websites[65]. The programme is not specifically oriented towards eGovernment research, but its early conclusions indicate that by 2020 some countries in this region could potentially participate in an eGovernment and eGovernance network.

3.3.1.2 OECD

A research focus of the OECD, relevant to eGovernment, is reducing the administrative burden, with particular emphasis on how to achieve administrative simplification and how to measure the progress in achieving it[66]. The study revealed that that although in the past administrative simplification was often carried out on an ad-hoc or sectoral basis, most OECD countries today follow a holistic government approach. Yet, governments still place more emphasis on reviewing existing regulations than on reforming them. The OECD came to the conclusion that basic approaches to administrative simplification are single access points for public eServices and business process re-engineering. For the latter, a reference layer of data integration is necessary in order to render eGovernment services independent from the underlying information sources.

As reported by eGovRTD2020 researchers[67] the OECD's eGovernment research is driven by a strong economic focus and, therefore it is not surprising that eProcurement is the main topic of interest. Other benefits such as increased transparency, integrity and accountability in public spending also combat corruption. In order to achieve the stated goals, regulated

[62] UN e-Government survey 2008: From E-Government to Connected Governance, ST/ESA/PAD/SER.E/112, ISBN 978-92-1-123174-8, p xiii, http://unpan1.un.org/intradoc/groups/public/documents/un/unpan028607.pdf

[63] United Nations, "Creation of a global culture of cyber-security and the protection of critical information infrastructures", General Assembly, Resolution 58-199. http://www.apectel29.gov.hk/download/estg_13.pdf.

[64] http://www.egovrtd2020.org

[65] United Nations. "Asia Pacific Development Information Programme", portal to Asia Pacific eGovernment websites. http://egovaspac. apdip.net/references/online/.

[66] OECD, "Cutting the red tape: National strategies. Policy Brief", http://www.oecd.org/dataoecd/12/9/38016320.pdf

[67] eGovRTD2020, Deliverable 5.2.

access to information coming from different sources must be achieved so that properties such as integrity, controlled access and chain of custody are maintained.

In addition, one should note the OECD's promotion of the project WiMAX (Worldwide Interoperability for Microwave Access) to support long-distance wireless connectivity for broadband access and interoperability. WiMAX has the potential to overcome the digital divide by reducing costs and therewith prices.

3.3.1.3 EICTA

The European Information & Communications Technology Industry Association (EICTA) and the ICT industry at large support eGovernment efforts in Europe. EICTA underlines the need for effective identity management systems to enable real-time eGovernment, to assist the launch of new eGovernment services and to free governments to innovate and develop new business models of delivering services.

The "eGovernment issue group" of EICTA states[68] the importance of new approaches in data integration and consistency, along with the need for research on semantic pan-European eGovernment activities.

3.3.1.4 Joint initiatives

3.3.1.4.1 InterPARES

Information integration has been addressed from a different viewpoint with respect to eGovernment. The perspective of long term preservation of digital records poses specific requirements for information being handled by administrations. This information comes from different sources, on different media and generally requires a reliable storage and access method that can last for many years. The long term preservation of digital records has been addressed by several projects including the international consortium InterPARES[69].

The main focus of the **Inte**rnational **R**esearch on **P**ermanent **A**uthentic **R**ecords in **E**lectronic **S**ystems (**InterPARES**) projects is the long term preservation of digital records. It started in 1999 and its current phase is scheduled to end in 2012. Its international scope and direct relation to public administration requirements in the preservation and processing of information can be a significant contribution in information integration research for eGovernment.

As presented in the project's home page, InterPARES-3 is an international collaborative endeavour composed of several regional, national and multinational teams with international funding: InterPARES-3 was initiated in 2007 and will continue through 2012. This third phase of the project builds upon the findings of InterPARES 1 and 2, as well as of other digital preservation projects worldwide. It will put theory into practice and work with small and

[68] EICTA, "EICTA Reflections/ Comments on eGovernment research in FP7", p 4, http://www.eicta.org/fileadmin/user_upload/document/Issues/DEPG/EICTA_reflection-comments_on_egovernment_research_in_FP7.pdf

[69] http://www.interpares.org

medium-sized archives and archival records units within organisations. Besides, it will develop teaching modules for in-house training programmes, continuing education and academic curricula.

The organisation of InterPARES-3[70] is built around local groups that use the name TEAM (derived from the specific title given to this third phase of the Project, **T**heoretical **E**laborations into **A**rchival **M**anagement). The InterPARES-3 Project International Alliance comprises TEAM organisations named: Africa, Brazil, Canada, Catalonia, China, Colombia, Italy, Korea, Malaysia, Mexico, Netherlands and Belgium, Norway, Singapore, Turkey, and United Kingdom and Ireland.

InterPARES addresses all aspects of information integration related to digital record preservation mainly focusing on technical and semantic level interoperability among storage and retrieval systems.

3.3.1.4.2 South-east Asia

The geographical characteristics of the region have offered a strong incentive for the international collaboration on eGovernment issues, as most of the states in the area face the same or at least similar challenges. There is strong influence from Australian similar activities; however they are not directly exploitable as every state has its own peculiarities mainly in the social and political dimensions.

An Economist article[71] on "Opportunities and obstacles for intergovernmental collaboration in South-east Asia" presents the intricacies of the region, where the main problem is identified to be the political instability and lack of trust towards the government. Another issue that has to be considered is that the evolution of eGovernment should not widen the social gap between those who are more wealthy and educated with access to modern IT infrastructure and those who do not.

A factor resisting the evolution of eGovernment is the inherent conservatism of public administrations. Many bureaucrats in the Philippines reportedly "fear a loss of power if government functions are decentralised and streamlined through technology". There is no denial however in these reports that eGovernment promotes transparency, trims bloated structures and consequently reduces the scope for corruption.

3.3.2 Non-EU national initiatives

Although the main focus of this report is to present international cooperation activities related to information integration, certain research activities occurring at national level have been included below. The inclusion was based on the following criteria:

- Scale: The activity must have a significant geographical scope.
- Relevance: The activity must handle information integration issues and propose solutions

[70] http://www.interpares.org/ip3/ip3_index.cfm

[71] http://osrin.net/docs/seamless_administration.pdf

- Importance: The activity must contribute in the global know-how on information integration

3.3.2.1 USA

3.3.2.1.1 Research activities

The majority of research in USA is funded by NSF[72]. An NSF funded project addressing the information integration issue is "Modelling the Social and Technical Processes of Inter-organisational Information Integration". According to eGovRTD2020 "This project develops and tests dynamic models of information integration in multi-organisational government settings in law enforcement and public health, combining organisational behaviour, computer and information science, and political science perspectives; it uses both system dynamics and social process modelling."

US Department of Justice (DOJ) sponsors research projects that address improvements in law enforcement and criminal justice. Interagency and intergovernmental information sharing is a strong theme in much of this work. Projects are carried out by university-based investigators as well as by professional law enforcement and information technology management associations. The National Institute of Justice (NIJ), the research arm of US DOJ, sponsors technology research in several areas pertaining to law enforcement including crime mapping and communications technologies.

In addition, the National Institutes of Health (NIH) are part of the US Department of Health and Human Services (US DHHS) which is responsible for many programmes that address public health, social services, and related areas.

The National Academies (NA) are chartered by Congress to serve as an independent advisor on scientific topics of importance to the nation. Study panels made up of leading scientists assess various topics and issue reports, usually at the request of a federal government agency. The Computer Science and Telecommunications Board (CSTB) of the NA has issued reports on information technology research for crisis management, federal statistics, and innovation and eGovernment. Also, the National Historical Publications and Records Commission (NHPRC), the research arm of the National Archives and Records Administration, supports a research programme aimed at improving archival and records management theory and practice.

Examples of research projects in government modernisation funded by NSF include:

- "Connecting to Congress", concerning the adoption and use of web technologies among congressional offices, and performing research on how members of the congress use, or should use, the Internet to provide information to and interact with their constituents.

- "Modelling the Social and Technical Processes of Inter-organisational Information Integration": This project develops and tests dynamic models of information integration in multi-organisational government settings in law enforcement and public health,

[72] eGovRTD2020 Deliverable D5.2 – Final book: Roadmapping eGovernment Research, http://www.egovrtd2020.org/EGOVRTD2020/navigation/results/book.

combining organisational behaviour, computer and information science, and political science perspectives; it uses both system dynamics and social process modelling.

- "COPLINK Center: Information and Knowledge Management for Law Enforcement" develops knowledge management technologies and methods for capturing, analysing, visualising and sharing law enforcement information and studies the organisational, social, cultural and methodological impacts and changes needed to maximise and leverage in information and knowledge management investments.

- "Knowledge Management Over Time-Varying Geospatial Datasets" focuses on integration of spatial data collected by many government agencies in various formats and for various uses, thus providing for new uses; includes development of a knowledge management framework to provide syntax, context, and semantics, and explores the introduction of time-varying data.

3.3.2.1.2 Vertical activities

In the US, there are organisations that focus primarily on learning how information technology has affected and will affect society and culture. The Pew Charitable Trusts is a non-profit foundation that sponsors the "Internet and American Life Program" which explores the impact of the Internet on Americans and disseminates research-based information on the Internet's growth and societal impact. Recent work has addressed broadband adoption, on-line activities, social networks, and the demographics of Internet use. Also, the Markle Foundation is a non-profit organisation that focuses on the impacts and potential of information and communication technologies to change people's lives. The Foundation conducts research and social change projects in partnership with selected collaborators from the public, private, and civic sectors. Its current priorities are health care and national security.

Also, examples of NSF funded projects that have more of a civic and societal perspective include:

- "Policy Made Public: Technologies of Deliberation and Representation in Rebuilding Lower Manhattan". This project examines how old and new advocacy groups are adapting to new deliberative technologies that may challenge traditional mechanisms of citizen participation in public policy decisions.

- "Digital Government: Harvesting Information to Sustain Our Forests". An initiative to design and produce a prototype of an "Adaptive Management Portal" to make information available in an open, natural and useful way to all parties interested in forest lands.

In addition to purely research topics, there are several publicly funded projects addressing information integration subjects from research to actual implementation. The following paragraphs present such initiatives.

3.3.2.1.2.1 Consolidated Health Informatics (CHI) Initiative

The Consolidated Health Informatics (CHI) initiative adopts a portfolio of existing health information interoperability standards (health vocabulary and messaging) enabling all agencies in the federal health enterprise to "speak the same language" based on common enterprise-wide business and information technology architectures.

Through the CHI governance process, all federal agencies will eventually incorporate the adopted standards into their individual agency health data enterprise architecture used to build all new systems or modify existing ones.

Progress reported up to 2006 includes:

- Government-wide health IT governance council established;
- Portfolio of 24 target domains for data and messaging standards identified;
- Partnered with 23 federal agencies/departments who use health data for agreements to build adopted standards into their health IT architecture;
- Messaging and terminology standards adopted for 30 domains, yielding 14 sets of standards to be used in federal IT architectures;
- Domains that did not have standards ready or mature enough to adopt produced follow-up recommendations;
- Regular meetings with industry to prevent major incompatibilities in partnership with the National Committee on Vital and Health Statistics;
- Defined change management role for the initiative's merger into Federal Health Architecture (FHA);
- CHI goals incorporated into the FHA and activities coordinated through the Office of the National Coordinator for Health Information Technology (ONC) http://www.hhs.gov/healthit;

The Office of the National Coordinator for Health Information Technology (ONC) is in charge of the continuation of the activity.

3.3.2.1.2.2 Stewardship

The NSF Strategic Plan for FY 2006-2011 defines the Stewardship[73] strategic goal as supporting excellence in science and engineering research and education through a capable and responsible organisation. Excellence in NSF's stewardship is essential to achieving the Foundation's mission and accomplishing its goals.

[73] http://www.nsf.gov/about/budget/fy2008/pdf/31_fy2008.pdf

E-Government Initiatives

The National Science Foundation (NSF) has provided about $4.000.000 for eGovernment projects in 2007 and 2008. About one quarter of this goes to the **Grants.gov** project that benefits NSF and its grant programs by providing a single location to publish grant (funding) opportunities and application packages, and by providing a single site for the grants community to apply for grants using common forms, processes and systems. Currently, beginning in 2007, NSF posts all its discretionary grants programs in "Grants.gov Find" and all of its funding opportunities in "Grants.gov Apply".

Another NSF-funded initiative that relates to information integration in eGovernment is the **Geospatial Line of Business** activity that ensures the effective and efficient provision of geospatial data to the research community. NSF is able to realise cost savings by not having to process individual requests for data in an ad-hoc fashion. The public frequently requests maps and other geospatial data from NSF, particularly during emergency response situations. The Geospatial portal provides an integrated environment to coordinate these requests, making the agency's response more efficient. It has the potential to reduce the cost of supporting such data requests.

NSF has had significant impact on the nation's research in the area of Geographic Information Systems. The "National Center for Geographic Information and Analysis" (NCGIA) which resides at the University of California at Santa Barbara, the State University of New York at Buffalo, and the University of Maine-Orono has developed and demonstrated powerful practical applications of geospatial data and technology. The NSF "Geographic and Regional Science Program" sponsors research on the geographic distributions and interactions of human, physical, and biotic systems on the surface of the Earth surface utilising GIS at state, county and city level.

Information integration topics are discussed in several other initiatives such as:

- the **Enhanced Human Resource Integration (EHRI)** initiative, which develops policies and tools to streamline and automate the electronic exchange of standardised human resource data (such as the electronic office personnel file) needed for creation of an official employee record;
- the **Integrated Acquisition Environment**, which provides tools and services allowing the NSF to improve its ability to make informed and efficient purchasing decisions and allows it to replace manual processes. If the NSF were not allowed to use these systems, they would need to build and maintain separate systems to record vendor and contract information, and to post procurement opportunities. Agency purchasing officials would be unable to have access to databases of other agencies on vendor performance and would not be in a position to use systems to replace paper-based and labour-intensive effort.

- The "**Budget Formulation and Execution Line of Business**" initiative enhances the NSF budgeting capabilities by establishing:

    - a community of practice,
    - a clearinghouse for sharing best practices,
    - tools for government-wide budget exercises and collaboration,
    - standards for data and data exchange,
    - modularity to facilitate flexible solutions, sharing, and re-usability.

3.3.2.2 Australia

The eGovRTD2020[74] Specific Support Action addresses the Australian eGovernment status from several points of view. As far as information integration is concerned, much work has been done, especially in the area of unified access by the citizens to information residing in different repositories. The Australian Research Council (ARC) is the main organisation funding eGovernment research.

As noted in the project's final book, "an approach by several federal government agencies towards their implementation of eGovernment applications is to consider the citizen as a customer for their services and to have a business approach to the implementation of these services. This is understandable in many cases since the services include the payments of benefits and pensions."

ARC has been funding a project named "Create Once, Use Many Times - The Clever Use of Metadata in eGovernment and eBusiness Recordkeeping Processes in Networked Environments"[75] where the concept of standards-compliant metadata is used to capture semantic information once in particular application environments then reused many times across business applications and in different environments.

The University of Technology, Sydney, Australia has published its research activities on semantic web for eGovernment service delivery integration. Farzad Sanati and Jie Lu[76] discuss the importance of a repeatable methodology for e-service composition projects.

3.3.2.3 Korea

Korea (6th in the global eGovernment readiness index) was second to the US in the eParticipation index of the UN eGovernment survey[77] and performed well in eConsultation assessments. It was also second only to Australia in the "e-Information" criterion, which assesses whether governments do provide the basic information that serves as the foundation for citizen participation.

[74] Deliverable D5.2 – Final book: Roadmapping eGovernment Research, http://www.egovrtd2020.org/EGOVRTD2020/navigation/results/book

[75] http://www.sims.monash.edu.au/research/rcrg/research/crm/index.html

[76] Farzad Sanati, Jie Lu, "Semantic Web for E-Government Service Delivery Integration," itng,pp.459-464, Fifth International Conference on Information Technology: New Generations (itng 2008), 2008, http://www2.computer.org/portal/web/csdl/doi/10.1109/ITNG.2008.120

[77] UN e-Government survey 2008: From E-Government to Connected Governance, ST/ESA/PAD/SER.E/112, ISBN 978-92-1-123174-8, p xiii.

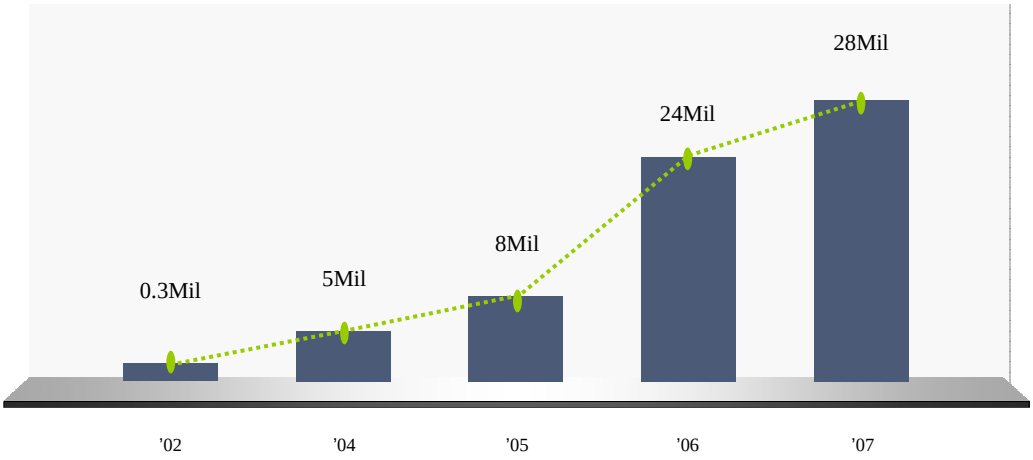

Figure 5. Adoption of Information Sharing

In Korea, information integration was channelled through the "National Basic Information System". This was separated into five areas and corresponding national databases were created[78]. Furthermore, an "eGovernment Special Committee" was formed and eleven key projects were initiated.

In the presentation[79] made at the Asia e-Government Forum 2008, information integration was identified as one of the main goals of the Korean future plans. The plans include extension of information available to citizens (see Figure 5) to include information provided by public institutions, in addition to currently provided administrative information. The number of documents available to the public was expected to increase by no less than 50% during 2008.

3.3.2.4 China

There is significant research work performed in China on eGovernment mainly focusing on GRID technology, as already noted in paragraph 3.2.1.2 above. Following the GRID paradigm requires that simultaneously emerging information (resulting from national or regional initiatives) is unified. China, however, has the advantage that its eGovernment initiatives can benefit from cases and practices already tested in the US and Europe[80].

3.3.2.5 South-east Asia

[78] Kijoo Lee, "The Strategy for Building Information Society in Korea", The Conference for Financing Information Society, Santiago, Chile, July23, 2003

[79]Kang-Tak Oh, "Government Information Sharing in Korea", Asia e-Government Forum 2008, http://www.korea.go.kr/eng/_eng_inter/pdf05_down.jsp

[80] Ding Feng, Wang Yanzhang, Ye Xin, "E-government for the People: Learn from North America and European Union", IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, 2008. WI-IAT '08. http://ieeexplore.ieee.org/xpl/freeabs_all.jsp?arnumber=4740763

The Economist Intelligence Unit has published a special research report[81] on the situation in eGovernment in South-East Asia. There is significant variance among the different states in the region, with Singapore leading (23rd place globally), while four countries lagging (below the 130th place globally). Regarding these laggard states, as mentioned in the report, a change in mindset must first be achieved before actually any significant progress in eGovernment occurs. This is a prerequisite for any plans for information integration. For example, the report states that in the Philippines, the lack of orientation towards customer service is one of the stumbling blocks to eGovernment integration. Furthermore, there is an important lack of strong leadership which is coupled with the fear of job losses. The latter is feared to be a consequence of agencies becoming more streamlined and more efficient thanks to eGovernment. All these factors contribute to the observed slow advancement.

3.3.2.6 Japan

In a paper by Björn Niehaves[82] it is stated that: "public sector reform in Japan, especially decentralisation, has significant impact on the governance structure. There is a growing trend towards strengthening local governance capabilities and shifting tasks, functions, authority, financial revenues, and responsibility from the central to the local government level. Such changes in the governance structure have a strong effect on eGovernment and NPM innovation and diffusion processes. First, decentralisation creates a greater motivation for local governments to innovate and adapt innovations (account-ability). Second, financial and managerial reforms support a decentralisation of innovativeness. Here, for instance, greater financial autonomy allows for a decentralised establishment of eGovernment and NPM knowledge in local entities, e. g. in terms of own research units. Furthermore, managerial and educational reforms build up a greater body of NPM and eGovernment knowledge among local government officials."

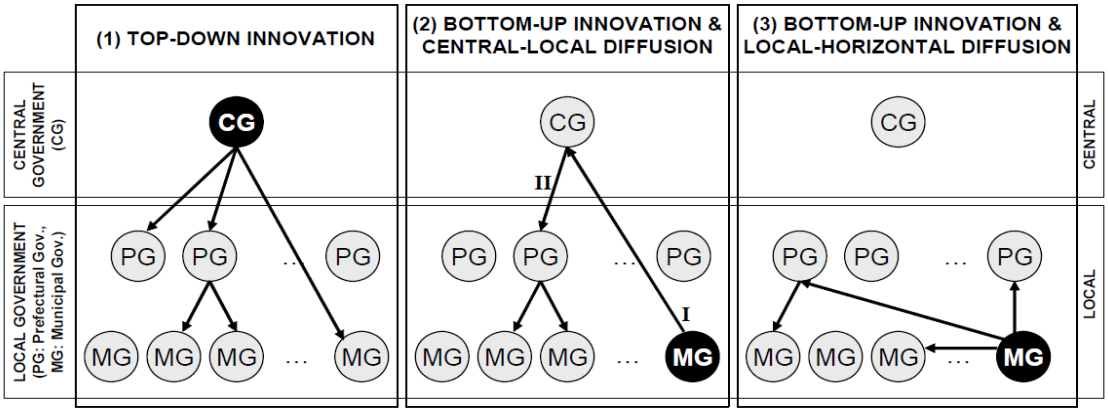

Figure 6: Types of innovation and diffusion in Japan

[81] http://osrin.net/docs/seamless_administration.pdf

[82] Niehaves Björn, "Institutional Change and eGovernment Innovation Processes", The Twelfth Annual Conference of the International Research Society for Public Management (IRSPM XII) at http://www.irspm2008.bus.qut.edu.au/papers/documents/pdf/Niehaves%20-%20Institutional%20Change%20and%20eGovernment%20Innovation%20-%20IRSPM%20-%202008.pdf

The public sector reform and its consequences are reflected in the information integration process as well. Different types of dissemination paths can be used, in both top-down and bottom-up approaches, as depicted in page 6 of the aforementioned paper and shown in Figure 6.

3.4 Other resources

The e-GISE network[83] initiative, funded by the **UK** government explored IST strategies and their embedded mechanisms for evaluation and integration within government and the service or public sector.

Estonia has designed and implemented a nation-wide eGovernment infrastructure[84] based on their X-Road communications network that separated the user interface from the underlying databases. Information integration was achieved through the use of a single authentication mechanism (digital certificate) embedded in the Estonian identity card, and the use of standardised messages between the applications providing services to the citizens and the applications maintaining the relevant information. This standardised methodology in information exchange has assisted Estonia to overcome information integration issues.

At present in **India**, personal information sharing among government departments and their agencies does not exist and there is an urgent need to not only accelerate information distribution, but also to broaden the scope of organisations so that they can share data[85]. This issue is not only related to information integration but also to trust; it is not a technical problem as the technological solutions exist at large. For example, citizens' demographic information can be collected by government departments via the various states under a series of security options or directly by the central government via a uniform standard, The latter could lead to a the adoption of a national ID.

There has been reported research work on the integration of GIS and eGovernment in **Kuwait**. This research revolves around the technical and informational integration between GIS and eGovernment. The importance of this research becomes apparent once the following points are considered[86]:

- The lack of information and spatial infrastructure in many Arab countries pertaining to different service patterns and its relation to other factors such as population, resources, housing and others.

- The lack of modern data communication mechanisms and the adherence to legacy systems that result in ineffective use of those GIS systems already successfully established in a number of Arab countries.

[83] Network for eGovernment Integration and Systems Evaluation (eGISE), http://www.egise.org/

[84] www.ria.ee/public/x_tee/xtkalja.ppt

[85] Velamala Ranga Rao, Rakhi Tripati, "Personal Information Integration in e-Government", http://egovonline.net/articles/article-details.asp?Title=Personal-Information-Integration-in-e-Government&ArticalID=2190&Type=FEATURES

[86] Mohamed Aziz, "Integration of GIS and e-Government in Kuwait", at http://www.mapmiddleeast.org/magazine/2006/mar-apr/16_1.htm

- The fact that in a number of non-Arab countries successful integration between both technologies has been achieved; in parallel, there is a significant increase in the awareness level for information technology in the general Near East region. Both facts indicate that considerable improvement may be expected in the following years based on the current trends.

Additionally, work by the IBM Software Group in **Singapore**[87] highlights the concept of enterprise vertical integration in eGovernment that addresses the information integration subject focusing on vertical interoperability at all layers (technical, semantic, organisational).

[87] Raymond Cheng-Yi Wu, IBM Software Group, Singapore, "Enterprise integration in e-government", at http://www.emeraldinsight.com/Insight/ViewContentServlet?Filename=Published/EmeraldFullTextArticle/Articles/3260010106.html

4 Conclusions

It has long been recognised by all practitioners in the field of eGovernment that the most important problem in offering effective, context-aware eGovernment services to citizens and businesses is semantics and semantic information integration. The latter has been extensively researched in a varied theoretical context spanning classical database theory, abstract algebraic constructions and computational ontologies.

Information integration in eGovernment research has almost exclusively concentrated on ontologies and ontology integration in web services environments. Being a terminological system, ontologies are not only important information integration enablers but also flexible knowledge bases. This makes them particularly relevant to eGovernment as they allow agile adaptation to changing and diverse environments and needs. Concepts and terms used by the public sector are fundamentally similar in meaning between administrations and so are the relations between them; the difficulty lies in the multiplicity of sources and the varied quality and heterogeneity of data available. A formal terminological system (ontology) created at a high enough level can reach a certain level of stability and invariance and serve as a common domain ontology. Heterogeneity can then be solved at ontology level via data mediation, such as that inherently supplied by the Web Services Modelling Ontology (WSMO) or by extensions of other ontology languages. Both techniques are used by FP6 projects.

Research projects employ ontologies as modelling tools at all levels of eGovernment, namely enterprise, domain, process and information (data). To describe complexity, multiple ontologies, managed by purposely-built editors and even centralised clearing houses for large scale applications, are frequently employed. Also, niche applications, such as extraction of meaning from multi-lingual content, or legal knowledge representation have benefited from combining ontologies with relevant models of context. Similarly, ontologies coupled with data mining techniques have been used to combat fraudulent use of public funds.

The information integration problem is addressed outside Europe as well. There is substantial diversity observed when studying the various approaches followed, funding schemes and results obtained in different countries under different circumstances. As expected, significant results have been published by the most advanced countries; however, there are specific examples of important research and application of information integration in several not-so-obvious cases, including Australia, south-east Asia, and the Far East.

When it comes to large-scale application the practical problem to be faced is agreement at both policy-making level and at administrative and technical level, given the vast size of the terminological system and the entities (communal, national, regional) involved.

ANNEX: eGovernment readiness index Top 35 (2008)

A significant reference when handling the international dimension of eGovernment progress throughout the world is the UN eGovernment survey published in 2008. As a general indication of the main countries where eGovernment has progressed more, the eGovernment readiness index top 35 countries are presented below. This index takes into account many aspects of information integration, including unified access to information coming from different administrations through a single web portal, the provision of services relying on information residing in different departments, the dispatching of information provided by the user to different departments etc.

Rank	Country	Index
1	Sweden	0.9157
2	Denmark	0.9134
3	Norway	0.8921
4	United States	0.8644
5	Netherlands	0.8631
6	Republic of Korea	0.8317
7	Canada	0.8172
8	Australia	0.8108
9	France	0.8038
10	United Kingdom	0.7872
11	Japan	0.7703
12	Switzerland	0.7626
13	Estonia	0.7600
14	Luxembourg	0.7512
15	Finland	0.7488
16	Austria	0.7428
17	Israel	0.7393
18	New Zealand	0.7392
19	Ireland	0.7296
20	Spain	0.7228
21	Iceland	0.7176
22	Germany	0.7136
23	Singapore	0.7009
24	Belgium	0.6779
25	Czech Republic	0.6696
26	Slovenia	0.6681
27	Italy	0.6680
28	Lithuania	0.6617
29	Malta	0.6582
30	Hungary	0.6485
31	Portugal	0.6479
32	United Arab Emirates	0.6301

Rank	Country	Index
33	Poland	0.6117
34	Malaysia	0.6063
35	Cyprus	0.6019

Table 2: EGovernment readiness index (UN 2008)

Prepared by:

Lead Contractor:
EUROPEAN DYNAMICS

http://www.eurodyn.com

Authors:

Dr. Ioannis KOTSIOPOULOS

Dr. Panagiotis RENTZEPOPOULOS

Contract No.:

Contract No. 30-CE-0043035/00-16

European Commission
Information Society and Media Directorate-General
ICT for Government and Public Services Unit

Tel (32-2) 299 02 45
Fax (32-2) 299 41 14

E-mail infso-egovernment@ec.europa.eu
Website http://ec.europa.eu/egovernment