Advances in Metainformation Definition and Management of Information and Services in Geospatial Data Infrastructures¹
(TIN2007-65341)

Pedro R. Muro Medrano²
Depto. de Informática e Ing. de Sistemas
Universidad de Zaragoza

Abstract

INSPIRE is an initiative for a European directive to create a Spatial Data Infrastructure (SDI) at the European level, its objective is to facilitate the access to spatial data, information which is recognized of vital importance mainly in the environmental sector. The development of such SDI involves many technological problems mainly related with interoperable geoprocessing and related areas. This project, stated in dec-2007, is the continuation of a previous project and it focuses on important SDI aspects related with metadata, catalogue and other geoprocessing services and conceptual and architectural aspects of SDIs. In addition to the scientific results the indicators show an outstanding effort in knowledge and technology transfer.

Keywords: OGC, Catalog Services, Metadata, Thesaurus, Geographic Information, GIS, Web Services, Geoprocessing Services, Web Map Servers, Interoperability, Software Architectures, Spatial Data Infrastructures, Open Standards

1. Project Objectives

1.1. Context and coarse objectives of the project

In recent years, a significant and growing proliferation of geo-spatial information or information susceptible to be georeferenced (GPS and navigation devices, availability of more complete and more precise cartography, increase of satellites for Earth observation, interest of administrations and companies in the creation of thematic geo-spatial information...) is arising. Moreover, it is growing the gateways to this information through Internet, even through mobile devices, and the demand of applications and geo-processing services that can take advantage of this information. To deal with these new possibilities and derived technological challenges the concept of Spatial Data Infrastructures (SDI) has emerged. A commonly cited definition of Spatial Data Infrastructure (SDI) is provided by the SDI cookbook version 2.0: ‘The term “Spatial Data Infrastructure” (SDI) is

¹ Project reference TIN2007-65341
² E-mail: prmuro@unizar.es
often used to denote the relevant base collection of technologies, policies and institutional arrangements that facilitate the availability of and access to spatial data’ (GSDI Tech. Working Group and contributors 2004 p. 8).

A key issue now underscores the need for advancement in the R+D of SDI technical aspects: the 5th of May of 2007 the European Commission launched INSPIRE (INfrastructure for SPatial InfoRmation in Europe), an European Directive to create European legislation to promote national and regional SDI development within Europe. This initiative, sponsored at the highest levels within the EC, (see Memorandum of Understanding for the cooperation between Commission services: Directorate General of Environment, EUROSTAT and Directorate General of the Join Research Center), will mandate how and when each member state should create its national SDI. We must also add at this time, the convergence between different standardization organizations in the field of geographic information (OGC, ISO, CEN, FGDC, …), institutions (GSDI, EUROGI, AGILE, JRC, …) and European initiatives (GETIS, GINIE, INSPIRE, …).

Inside the Spanish framework, AENOR is pushing for ISO geographic information standards, and the Consejo Superior Geográfico (High Geographic Commission-Commission of Geomatics) is in charge of the development of the National Spatial Data Infrastructure in Spain and the Spanish contribution to the European Spatial Data Infrastructure.

The project has been established in the context of continuing the technological advancement toward the construction of a Spatial Data Infrastructure (SDI) making possible the development of national technology in this field. From the technological point of view, a SDI consists of reference geographic data, metadata, and diverse web services (standard and interoperable), especially access via catalogs, that allow that users (scientists, public sector technicians, general public) are able to locate and access these data and their value-added products. The core of a SDI are the use of international standards, distributed open systems, criteria of interoperability and an architecture of services centered in the use of metainformation as key tool for the knowledge of data, services and their interoperability. Many challenging technological problems arise to deal with SDI in areas such as distributed geoprocessing, technological and semantic interoperability, open geoprocessing services and standards, web services and architectures, information recovery, metadata, geodata models, geodata harmonization, …

Taking the scientific-technological developments in the area of SDI performed by our research team as starting point, the project centers on some specific problematic and new ideas originated in previous projects that are of great interest to us and focuses on architectural issues and the use of metainformation for geoinformation management and geoprocessing services. In addition to maintain our implication in commissions of standardisation institutions related with the project topics (OGC, ISO, DC, CEN, AENOR, CSG-CG), and within a strategy of metadata intensive systems, the progress will be made in the characterization of geospatial information with metadata representation models. These models facilitate interoperability and robustness, and face the problematical issue of the metadata specification of geoprocessing services. Among others, the project will deal with issues such as the definition of metadata as hierarchies and their support, the transformation between standards, profiles, relationships, aggregations, derivations, multilingual problems, spatial representation schemas based on geographical identifiers, XML and RDF representations, improvements in thesauri representations, etc.. The project will progress in the generalization of the characterization of heterogeneous georeferenced information on the base of the metadata standard Dublin Core. Indexing and ranking models for metadata searching services will be adapted and improved, aimed to very large storage systems. Referring to services, the project will propose a methodology and design guidelines for a metainformation centred access strategy, and additionally analyze aspects such as multilingualism, harmonization, distribution, …
patterns for integration of services and applications and the election of adequate linking technologies. As a paradigmatic example, the project will study the functional aspects of a remote sensing data sets catalogue application realized on top of the services provided by an SDI, application that is of special interest for the entities related to the project. The project will also rely on the development of a pilot system that will provide feedback of scientific-technological aspects, as demonstration of technical adequacy, feasibility of possible results and to set the bases for the demonstration of potential knowledge and technology transference.

1.2. Detailed objectives of the project

Many expectations exist on the use of metainformation but it is necessary to advance in the characterization, management and added value of this metainformation; it is needed to improve the management of aspect as geospatial, thematic, relations, aggregations, multilingualism, harmonization, indexing problems, ranking, etc.

Advances in the definition, comprehension, use and interoperability of geospatial services and the use of metadata as the central core of information infrastructure (in discovery, use and to facilitate interoperability and linking) are also needed. In addition, some conceptual development and a higher knowledge, experience and innovative ideas to use geospatial e-services in applications is also needed. On the other hand, progress on the conceptual view of the architecture of spatial data infrastructures is complementary to advance in the understanding of these systems of systems. An application of remote sensing dataset catalog is very interesting because of the entities with interest in the results that collaborates in the project that offers an excellent example of use of metainformation and geospatial services.

The work of investigation of this project wants to advance in the knowledge of information management and geo-spatial services, on a strategy based on the use of metadata in an infrastructure distributed and open based on interoperable components that, in the area of geospatial information, has been denominated spatial data infrastructures. This knowledge has to be complement with proposals to fit the software architecture of these systems, and in applications that use the geo-spatial data and services. More specifically this basic objective is divided in the following issues:

1. To improve the characterization of geo-spatial information with metadata representation models which facilitate the interoperability and robust respect to metadata in different languages and management of problems of misspelling. In addition, the project will develop improvements for the specification and robustness of fields with the intensive use of thesaurus. The project will undertake the specification of geospatial metainformation, with special attention to the support of schemes of spatial reference based on geographical identifiers, semantic disambiguation between terms and geographic concepts, vague spatial aspects, etc. There will be advances in the generalization of the characterization of e-services of geoprocessing and georeferenced heterogeneous information on the base of the Dublin Core standard (standard ISO 15938), studying the way to make them compatible with other ISO proposals, with the objective to facilitate the interoperability based on metadata.

2. To advance in aspects related to management and structuring of knowledge in SDIs working on the base of the use and management of ontologies, aspects of multilingualism of the ontologies and by extension of an SDI, and topics maps as tool to structure knowledge.
3. To adapt, to improve and to develop models of indexing and ranking for the creation of metadata search services that can take more advantage of the expressive potential of the metadata, specially the geospatial aspects. Additionally, it is needed a viable implementation and put into operation of algorithms related to models for storage systems with huge volumes of metadata so that they improve the efficiency of present approaches.

4. To analyze the possibilities of use of standard metadata to facilitate the construction of interoperable systems on the base of discovery of data and services using metadata catalogues and to propose patterns for its integration in applications, in an SDI strategy intensive in metainformation. Implications in services of multilingual aspects, harmonization of models, strategies for management of distribution (including harvesting), strategies for linking and its specification, architectural patterns of analysis and design for the integration of services in applications are aspects in which a conceptual and technological advance also is considered.

5. To deepen in the functional problematic of a remote sensing datasets catalogue application based on services provided by an SDI.

6. To make a pilot system that will provide feedback of scientific aspects, to demonstrate the fitness and technical viability of the possible results and to lay the foundations to show the potential transference of knowledge and technology.

The work to achieve the previously mentioned objectives is structured into seven activities according to the following timetable:

<table>
<thead>
<tr>
<th>Activities/Tasks</th>
<th>First year (*)</th>
<th>Second year (*)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Coordination and management of the project</td>
<td>X X X X X X X X X X</td>
<td>X X X X X X X X X X</td>
</tr>
<tr>
<td>Study, follow-up and participation in the development of standards and technologies that support them</td>
<td>X X X</td>
<td></td>
</tr>
<tr>
<td>Advances in the definition and support of metadata</td>
<td>X X X X X X X X X X X X</td>
<td>X X X X X X X X X X X X</td>
</tr>
<tr>
<td>Advances in the structuring and management of knowledge in an SDI</td>
<td>X X X X X X X X X X X X</td>
<td>X X X X X X X X X X X X</td>
</tr>
<tr>
<td>Advances in search and recovery of metainformation and contents of geospatial data</td>
<td>X X X X X X X X X X X X</td>
<td>X X X X X X X X X X X X</td>
</tr>
<tr>
<td>Network of geoprocessing services</td>
<td>X X X X X X X X X X X X</td>
<td>X X X X X X X X X X X X</td>
</tr>
<tr>
<td>Prototype of remote sensing dataset catalogue</td>
<td>X X X X X X X X X X X X</td>
<td>X X X X X X X X X X X X</td>
</tr>
</tbody>
</table>
2. Success level reached in the project

The following paragraphs illustrate some relevant results of the project.

- Characterization of geo-spatial information and services. Important advances in the automatic characterization of geographic services has been made. A Dublin Core profile for this kind of services has been specified and it is in use in the Spanish National Spatial Data Infrastructure. In addition, the reference tool for metadata creation has been improved with several new features that are going to be provided as open source. Maybe, one of the most relevant aspects has been the work done within the CatMDEdit tool towards the synergy with GIS tools. Up to now most metadata have been created with the main purpose of uploading the contents of catalogs accessible through an SDI initiative. Obviously the visibility of data holdings by means of metadata is very important. But there is no sense in producing metadata if these metadata are not going to facilitate the processing of data in a more effective way. Furthermore, if data processing is combined with metadata management, the quality of metadata will increase and it will be easier to maintain metadata catalogs up-to-date. Based on the metadata describing a resource, this work has described the mechanism to derive dynamically the information required to create a gvSIG project with the adequate viewing parameters. The connection mechanism between CatMDEdit and gvSIG will appear in the new version of CatMDEdit to be released in 2009.

Additionally, other tools to facilitate the creation of metadata content have been studied and used. It is remarkable the use of WOS (see next section) since it provides access to terminological ontologies (concepts, properties, definitions and relations between concepts and other ontologies) recommended by metadata standards.

Finally, some techniques to automatically generate a classification of a collection of metadata records using the elements of their keywords section have been specified. These techniques are based on hard (K-means) and fuzzy (C-means) clustering and have into account the hierarchical structure of the concepts contained in the Knowledge Organization Systems (KOS) used to select these keywords. Two different approaches have been analyzed. In the first approach, the keywords in the metadata records picked up from a single thesaurus have been encoded as a numerical value that maintains the inner structure of the thesaurus. Then, this encoding has been used as the property for the definition of the thematic clusters. In the second approach, all the keywords in the metadata collection have been considered as free text. Additionally, in order to improve the results, when it has been recognized that a term belongs to a KOS, the terms in the hierarchy of ancestors have been added to the set of keywords. Then, these sets of expanded keywords have been clustered using hard and fuzzy clustering techniques. This last approach aimed at avoiding the deficiencies of the first approach that assumed the use of a single selected vocabulary.

- Management and structuring of knowledge in SDIs. In this research line, the work done can be separated in two main areas:

  - SDI conceptualization. It has been identified a pattern to design and document distributed geographic information systems following SDI design principles. The pattern has been presented as an architectural style, defined inside the component-and-connector
view type, extending two well-known styles in distributed information systems: the client-server and shared-data styles. The style has been created analyzing several important SDI architecture proposals, finding their common elements, and giving them a unique name and a definition.

- **Ontology Management.** Taking as initial work the ontology service developed in a previous project, it has been extended as a Web service compliant with the OGC Web Services Architecture specification and whose purpose is to facilitate the management and use of ontologies in an SDI. Designed as a centralized service, the architecture of this service aims at reducing the cost of creation of a new ontology, improving reusability and avoiding duplicities and inconsistencies. It is planned to submit the specification of this Web Ontology Service (WOS) as a new OGC Web Service specification that could be integrated in the future with the rest of Web Service specifications already issued by the Open Geospatial Consortium, to at least obtain the required feedback to improve, if necessary, the functionality offered by this service.

In addition, within this project, the ThManager Open Source project has been launched. ThManager is a tool for creating and visualizing SKOS RDF vocabularies, a W3C initiative for the representation of knowledge organization systems such as thesauri, classification schemes, subject heading lists, taxonomies, and other types of controlled vocabulary. ThManager facilitates the management of thesauri and other types of controlled vocabularies, such as taxonomies or classification schemes. SKOS has been selected, from the available interchange formats for thesauri, as the most promising format to become a standard for SKOS.

- **Models for indexing and ranking.** One aspect worked in this research line is the necessity of having reference context-based-knowledge models for indexing and ranking. In this way, two kind of works have been done: specification of process for building networks of concepts, and specific ontology models. Let’s see them in detail:

  - **Network of concepts in the urbanism area.** A process to generate a network of concepts has been developed, and it has been applied in the urbanism area. The objective has been to use this network as a basis for a first draft of an urban domain ontology. The process uses as input a set of multilingual thesauri from different knowledge areas (e.g., GEMET, AGROVOC, UNESCO and EUROVOC) and a thesaurus specialized in urbanism (URBISOC) to be able to select the urban terminology present in the other thesauri. The main steps of the generation process are the harmonization of the input formats, the mapping between the concepts to generate clusters of equivalent concepts using linguistic similarity measures, and the establishment of relations between the clusters on the basis of the original relations between the concepts contained in different clusters. Finally, in order to facilitate the visualization and reusability of the generated output, it is transformed into XTM and OWL formats.

  - **Ontology model for managing the administrative structure of a country.** An administrative unit ontology model for managing the administrative structure of a country has been built. This model has been used to create an ontology for Spain, France, United Kingdom and Portugal. The importance of this model, and the ontology created, could be remarked by the general problem found for developing spatial-disambiguation methods. Although techniques for disambiguation in the geographic area had been developed by
this project team in previous projects, we found that the main limitation was the necessity of having a reference ontology within the political organization of the States. Many information sources can be found with this information, but all of them have different data models. The ontology model proposed unifies the most popular models and provides semi-automatic mechanisms to generate data from different sources. It is important to stress that the high frequency of changes in the administrative organization (shape, structure, name or administrative capabilities) makes necessary to establish specialised policies and techniques for updating all the elements of the infrastructure that uses this model. Also the quality of the data sources is an issue that should be considered. If the official name, code or coordinates of a unit are not accurate, associated services in the SDI that use this information will obtain poor results.

In addition, the use of the WOS service has been analyzed and “architected” in order to be integrated within an information retrieval system to facilitate the construction of user queries and improve the recall of such systems. The WOS service can be used to exploit the knowledge of lexical ontologies to expand the original concepts with translations, synonyms and related concepts in similar lexical ontologies.

- Network of geoprocessing services. It has been developed a method that allows adding the optional SOAP platform-specific interfaces to any Web Processing Service (WPS), without requiring access to the source code or the binaries of this service; it only needs network access to its mandatory interfaces. Using this method in order to provide optionals interfaces, it is possible to avoid errors that could be committed when programming different platform-specific interfaces to offer the same functionality. It also makes it easier to integrate WPS with other Web Services, facilitating thus interoperability. A tool has been designed to validate this method. In order to test it, a BPEL composition has been implemented within the context of the Spatial Data Infrastructure of Spain (IDEE). Due to the fact that WPS services in the IDEE have not got any SOAP/WSDL interface, the incorporation of this tool in the IDEE is currently being studied as a promising resource to facilitate interoperability.

- Prototype of remote sensing datasets catalogue. This is the less advanced working area. The difficulties found in other research lines and the delay on the beginning of the project has made that there is not special remarkable results. At this moment, we are working in the characterization of the remote sensing products, and analyzing the functionality and services that this kind of systems should provide. Our plan is to work in this research line even after the end of the project because we think that it could provide us with an important knowledge that could be transferred with success to the industrial environment.

- Standardization. We continue our participation on standardization processes. As a result of this experience, we have been working with the OGC (Open Geospatial Consortium) and we have been collaborating in the definition of the metadata profile which has been adopted for the Spanish Spatial Data Infrastructure as well as with the Spanish gazetteer model recently adopted.

- Technology transfer. This is a very important part of our research strategy: the demonstration of our results is both an objective and a feedback mechanism. As it can be seen in section 1.5, we are having clear success on that.
3. Result indicators

The research work undertaken in this project became reality in a variety of ways as techniques, processes, methodologies, models, architectures, concepts, software components, systems, papers, projects, courses, … We will focus below in the most relevant:

3.1. PhD Thesis

Some PhDs have been partially developed during the project.

1. Contributions to the modelling of Spatial Data Infrastructures and their portrayal services. PhD student: Rubén Béjar Hernández. PhD advisors: Pedro R. Muro Medrano y Javier Nogueras Iso. PhD program: Systems Engineering and Computer Science. Title: PhD in Computer Science, University: University of Zaragoza. This thesis has already been admitted for its defense. PhD defense planned by the 2nd of April.


Other 2 PhD thesis are being developed during the project.

3.2. Publications

The following list of papers present results directly related with this project (see bibliography for full references).

3.2.1. Journals and book chapters

Publicados: [1], [2], [3], [4], [5], [6]

Sometidos: [12], [13], [21], [22]

3.2.2. Conference papers

Publicados: [7], [8], [9], [10], [11]

Sometidos: [14], [15], [16], [17], [18], [19], [20]
3.3. Projects

All the research and development projects we are involved are related in some way with the topics in project TIN2007-65341, we list above live projects during the life time of project TIN2007-65341 so far.

3.3.1. European R&D Projects


3.3.2. National projects and R&D activities sponsored in public calls


3.3.3. Projects for management of R&D activities sponsored in public calls
1) Consolidated research group – Advanced Information Systems Lab. Sponsored by the Aragón Government, Department of Education and Science. Directorate General of High Education
(attendance to support Operative Units of Research of the Autonomous Community of Aragón ref T56) 2008.

3.3.4. Projects and R&D activities with companies and/or administrations


3.4. Diffusion

Presentations of project work were made at GSDI, JIDEE, GVSig, SeCoGIS, Townontology and Congreso Nacional de Riegos.

The research group has organized the 3rd Townontology Workshop of Cost Action C21, hold in Zaragoza in October 2008.
3.5. Participation in

3.5.1. INSPIRE experts
Another relevant indication of the expert knowledge got by the members of the project is that two of them have been selected as experts in the call for experts for the development of INSPIRE draft Implementing Rules (5 drafting teams of 6 members each have been selected from a total of 180 all along Europe -see http://inspire.jrc.it). An additional expert has also been selected in a support expert group (only 4 experts have been selected from Spain).

3.5.2. WG-IDEE
We are also participating as experts (the only University Members) in the working group for the Spatial Data Infrastructure of Spain (IDEE standing for Infraestructura de Datos Espaciales de España), a working group (of about 100 persons in representation of all institutions involved in the IDEE) established by the Specialized Commission on Spatial Data Infrastructures from the High Geographic Council (Spain), which is in charge of the development of the national SDI. Additionally we are part of the Specialized Commission on Spatial Data Infrastructures in representation of Universities.

4. Final considerations
We do not want to finish before making some considerations that very probably has positioned this project in inferior conditions with respect the other projects:

- This is a two year project instead of the three year term which are most of the projects and at this time the project is starting its second year.

- The project had an additional delay of two more months because only it could be started on the 1st of December of 2007 (instead most of the projects that could be started on the 1st of October), so we only can show the results from these 13 months, compared with 27 months for most of the projects. This delay was due to the delay to answer the allegation made to the resolution of the project (although it should have been done by jul-07 and it was not done until the end of nov-07).

- The project got a substantial support reduction with respect to the project proposal. The reduction covered all items from financial aid (about the 20% of the proposed budget) to technical personal (none) as well as research fellowships (none). Quite obvious errors were detected in the evaluation report; however the allegations were rejected with no reason provided.
5. References


Jornada de Seguimiento de Proyectos en Tecnologías Informáticas

Pedro Larrañaga, Sascha Ossowski y Víctor Robles (editores)

jspTIN2009
IDEATIO: Intelligent Devices and Tools for Health Care Quality Control in Integrated Inpatient/Outpatient Follow-up
TIN2006-15460-C04................................................................................................................615
Roque Marín, Alfonso Palma, Paulo Félix y Ramón Ruiz

Techniques to enhance the performance, reliability, and services delivered by high-performance servers based on a cluster architecture
TIN2006-15516-C04................................................................................................................633
José Francisco Duato Marín, José Manuel García Carrasco, Antonio Garrido del Solo y Juan Manuel Orduña Huertas

Management, Access and Exploration of Large and Heterogeneous Data in the Web and Relational Databases
TIN2006-15536-C02................................................................................................................661
Ricardo Baeza-Yates y Josep Lluis Larriba-Pey

WEB Services and Testing: foundations and applications (WEST)
TIN2006-15578-C02................................................................................................................673
Manuel Niñez y Valentín Valero

Desarrollo de Aplicaciones de la Computación Celular con Membranas y la Computación Biomolecular en la Biología
TIN2006-15595.......................................................................................................................685
Alfonso Rodríguez-Patón

AmbienNet Ambient Intelligence Supporting Navigation for People with Disabilities
TIN2006-15617-C03................................................................................................................699
Julio Abascal, José Luis Sevillano y José Ignacio Artigas

DESAFIOS: Desarrollo de software de alta calidad, fiable, distribuido y seguro
TIN2006-15660-C02................................................................................................................711
Narciso Martí Oliet y Juan José Moreno Navarro

Autonomic Electronic Institutions
TIN2006-15662-C02................................................................................................................729
Juan A. Rodríguez-Aguilar y Teresa Alsinet

iDoc: Interactive Analysis, Transcription and Translation of Old Text Documents
TIN2006-15694-C02................................................................................................................741
Alfons Juan y Josep Lladós

DEVLENA: Descripción de Escenas de Video en Lenguaje NAtrual
TIN2007-62568.......................................................................................................................753
Juan Moreno García

Advances in Metainformation Definition and Management of Information and Services in Geospatial Data Infrastructures
TIN2007-65341.......................................................................................................................761
Pedro R. Muro Medrano