Beyond spatial data infrastructure: knowledge and process extensions

by Wim Hugo, MBV Equsys

Starting in 2005, the CSIR, SAEON, and a number of other stakeholders have collaborated [1] to conceptualise, define, and create systems that aim to extend the conventional, but rather narrow definition of spatial data infrastructure (SDI). This view typically defines a spatial data infrastructure as a framework of spatial data/ themes, its metadata, user communities, standards, policies and supporting technologies [2].

The collaboration found its main impetus in a multi-stakeholder project, the "Collaborative Spatial Analysis and Modeling Platform (CoSAMP)", from which much of this article is sourced. The main focus of CoSAMP was to look beyond the normal "static" or "profiling" bias of spatial information use in the contributing organisations towards a more knowledge-driven, dynamic, and collaborative paradigm. Naude' and McFerren [3] provide a comprehensive overview of CoSAMP.

The concepts remain applicable and valid as a broad-based, abstract specification. This article provides an overview of the business requirements for such "extended" SDIs, conceptual or abstract solutions, and a summary of the scope of functionality that should be provided by ESDI (Earth Science Data Interface) implementations.

A number of initiatives have contributed to and continue to draw on the work initiated in CoSAMP: These include CoGIS [4] and the CSIR's ongoing work in the Sensor Web environment [5].

The business requirement

The CoSAMP project [1] evaluated the need for SDI-like resources from the perspective of research and development as a business: what were the goals of the organisation (such as CSIR, or SAEON, for example), and how did spatial data infrastructure fit into these goals? And, more critically, what were the most important problems that manifested in the use of spatial data in a research and development organisation? These are briefly summarised here [3]:

- There was, and still is, a widespread concern in the research community at large and in the sponsoring organisations specifically that despite a deep capability, a comprehensive data collection, and considerable exposure to geospatial projects, efficiency of use and re-use of these resources were inadequate or under-utilised.
- In addressing the inefficiency indentified above, there were serious problems of "collaboration divides and barriers" – ranging from bandwidth connectivity issues through data availability to knowledge and experience-related deficiencies.

Because of this perspective, the extended SDI needed to address not only spatial data management and its discovery, classification, and use, but also needed to include the perspective of the knowledge associated with processes and the data on which it operates. In addition, collaboration and knowledge transfer were identified as a crucial enhancement required of typical SDI implementations.

Setting the scene: conceptual overview

The outcomes envisaged by CoSAMP, and described in detail in the Business Requirement Specification [6], made a number of conceptual statements that were translated into a set of user requirements and specifications in April 2006 [7]. This has now been reviewed in the light of extended (and sometimes amended) requirements for CoGIS [8]. The outcomes are partly derived from seminal work done in this regard by Liping Di [9], and are heavily influenced by standards-driven interoperability ideals.

These outcomes, taken together, imply the following:

- Worldwide standardisation, mostly through OGC [10], has led to the ability to source (discover) spatial data from any number of compliant services. This is generally referred to as discovery or geoquery.
- Furthermore, that standardisation of interfaces has led to an environment where spatial data can be obtained from a collection of physically and logically disparate sources and combined for further application (geoassembly or aggregation), either: as a map that can be viewed and manipulated in a number of ways by end-users; or as a service that provides vector and attribute data to be consumed by downstream processes.
- That it is possible, with varying degrees of automation, to standardise, structure, and encode the way in which we process spatial data (geoprocessing). The things we do with the data range from:
    - The very simple (for instance depositing, discovering, aggregating, and displaying spatial data).
    - To the very complex (for example chaining a collection of spatial analysis processes together, either manually or automatically, to calculate difficult spatial result sets such as spatio-temporal optimisations).
- That we can improve the chaining and construction of complex processes by insisting on signatures, whether these are constructed of weak types, strong types, or a mixture of the two.
    - That future extensions of knowledge management will also be improved by the signature requirement – by allowing automated chaining, and
    - That at a time when automation becomes a reality, we require additional metrics on process capabilities and resource consumption to allow improved or automated selection decisions.
- That it will become possible to make these activities both:
    - Distributed, in the sense of obtaining either data, processes, or both, from a collection of remote sources
    - And collaborative, in the sense that different users and use communities may be able to construct knowledge and content.

- That the collaborative nature of data use can lead to a partial automation of peer review, which provides the mechanism for content to be formally published.

We stress the important aspect of strong typing[1] and signatures for a reason: while most researchers agree on the importance of ontology and interoperability specifications to enable the conceptual specification as outlined above, it is clear that automated processing and chaining of such processes cannot be achieved without some form of validation of inputs.

A model for building and describing knowledge

As part of the systems engineering work undertaken for CoSAMP, effort was expended on determining how knowledge is obtained and how it should be associated with typical SDI and geoportal implementations.

Firstly, the model makes a distinction between research-driven work, and delivery-driven work. We need to distinguish these because:

- Research is driven by perceived needs and gaps: gaps in understanding, gaps in knowledge, decision support needs, and so on. The measure whereby the success of research is evaluated is often based on peer review: norms such as relative novelty and innovation/ creativity are applied to test the ideas developed in the research and development effort. It is not without standards or checks and balances: research findings are tested for the defensibility of the research methods, repeatability of findings, and alignment with known facts. Clients who fund research are often prepared to pay for novelty and innovation.
- New ideas do not keep office hours: often the final, ground-breaking insight arrives only on completion of a research project or task.
- Delivery, on the other hand, is driven by documented requirements and specifications. The client, who evaluates output on the basis of fulfilment of the requirements and specifications, is in the boundary case only prepared to pay for stated requirements, and nothing more. The client is not interested in novelty or innovation. New ideas not included in the specifications are out of bounds.

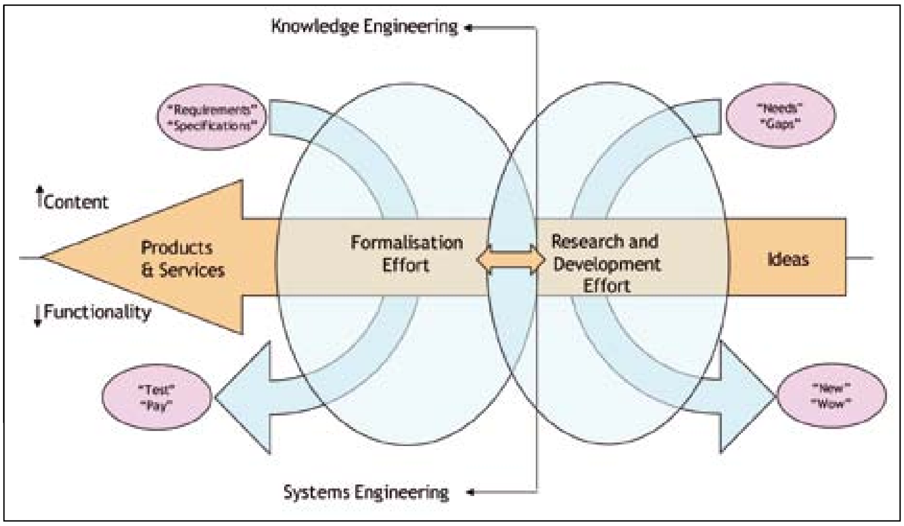

Fig. 1: From research and development to packaged knowledge products.

Because of this duality, one of the major challenges in the systems engineering task is to translate a wide scope of business requirements into a set of user requirements and systems delivery specifications that are matched with realistic release schedules.

Secondly, all efforts to provide systems are a mixture, to varying degrees, of providing functionality and content. Generally speaking, operational systems focus more on functionality with relatively simple content, while decision support systems often combine complex content and functionality. Knowledge management systems appear to combine relatively simple functionality (content-management type systems are not functionally complex) with very complex content.

Thirdly, it is likely that knowledge to be managed in such an ESDI environment will be in several or many states of "formalisation", and a distinction will be needed to inform potential users of the state of maturity of the knowledge being presented to them.

Knowledge management

Having established that the geoportal(s) envisaged by COSAMP primarily focuses on knowledge management with spatial references or linkages, we spend some

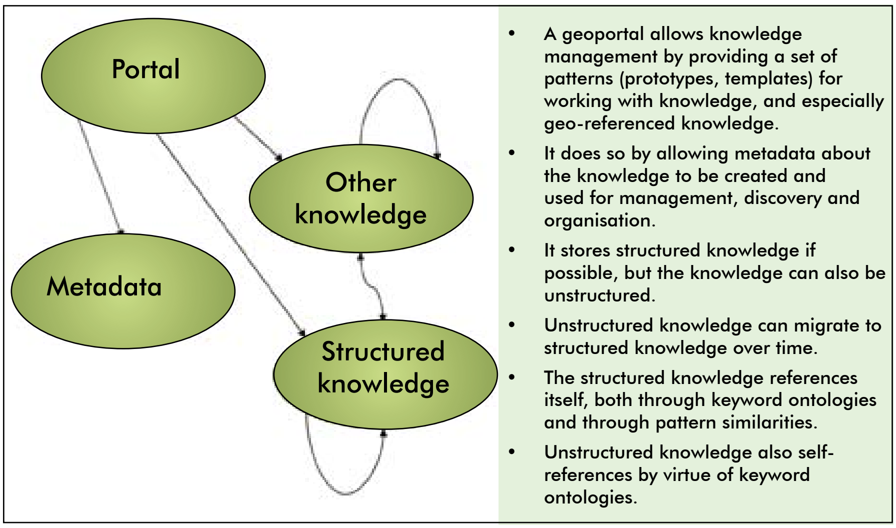

Fig. 2: High-level conceptual model for knowledge structures.

[1] *Analogous to the mandatory requirement, by a language definition, of compile-time checks for type constraint violations. That is, the compiler ensures that operations only occur on operand types that are valid for the operation.*

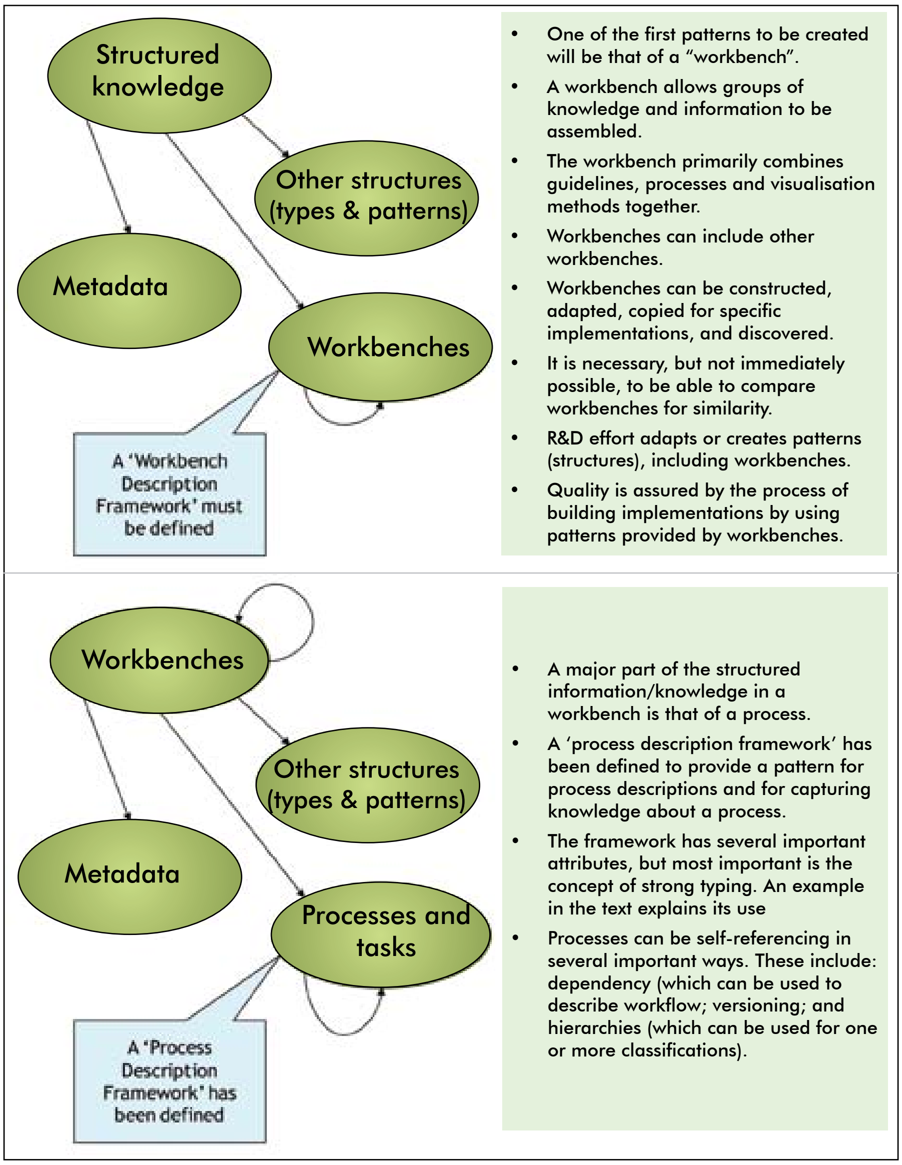

Fig. 3: High-level, conceptual model for workbenches and processes.

time on evaluating the way in which we can describe knowledge and its elements. Our ideas are crystallised in a set of diagrams, reproduced in Fig. 2.

The important feature in the diagram (see Fig. 2) is the concept of keyword ontologies that link otherwise unrelated islands of knowledge. Keyword ontologies and semantic relationships between them provide a relatively under-utilised, but powerful extension to basic search engine functionality.

Note that it is foreseen that the major contribution from research and development into this environment will be to adapt existing or create new workbenches, (see Fig. 3).

Applying knowledge

The following section is based largely on inputs provided by Naudé [3]. The diagram in Fig. 4 provides a perspective on how knowledge is applied in a spatial problem context, and how the concept of workbenches relates to the process of knowledge creation and application. The diagram is useful on more than one level, because it defines the way in which a specific user community expects to interact with the systems provided for knowledge management (of which a geoportal providing access to workbenches is an example), and it defines a starting point for generalising concepts such as processes, guidelines, workbenches, and the like.

Individual aspects of the process diagram are discussed below. In a certain sense, this represents the highest level process that the knowledge management portal should know about: all other processes derive from it by virtue of being included or referenced in various locations in this parent process.

Problem structuring

The first step in the process involves problem structuring. Problem structuring is domain-dependent, involves knowledge and experience, and is difficult to automate. In our view, one of the major research challenges lies here: expressing problems in such a way that it can be matched with known solution methods. For this, one needs a problem and solution typology that extends across knowledge domains and boundaries.

The work done to date on a functional typology structures problems into four broad categories, all of which need to be addressed in one way or the other by establishment of the geoportal(s) envisaged by CoSAMP:

- Basic good practice problems
- Data availability, compatibility and inference problems
- Predictive problems
- Application problems

Clarification of models, ontologies, perspectives

This involves the ability to search for, and identify, ontologies, models (theories, algorithms), and real-world representations (data, variables, processes) that have to be modelled or understood as a preamble to solving the problem. This is an important step if the problem cannot be well structured or is partly unstructured, less so for structured problems for which known solution methods are available and known.

The possibility also arises that the ontologies, models, data, or processes need to be extended or created from basic principles, in which case the knowledge base will have to grow.

Workbenches

The next step generally involves application of a 'workbench'. In the CoSAMP environment, a workbench is described as having the following components:

- One or more set of guidelines
- One or more set of nested processes
- Visualisation and reporting capabilities
- Other workbenches

If a suitable workbench is not available (for example because the problem has not been encountered before), it must be possible to derive new workbenches, either by modifying or extending existing ones, or by creating new ones. It is foreseen that workbenches, in their simple or null implementation, consist of a template or pattern that can be extended. It also makes sense that some guidelines (for example, dealing with quality control) may be a non-removable or mandatory property of all workbenches.

As shown in the Fig. 4, research and development often involves extending or creating these workbenches and the tools that are referenced in their processes.

Guidelines

Guidelines are not processes, but provide insights on how to approach the processes to be executed when solving a problem. The typical guideline is best described through one or two examples:

- A guideline will provide pointers on the selection of a boundary for a study area. Because of interaction and spatial linkages, the exact administrative or physical boundary suggested by the problem statement may not be appropriate. A guideline can be useful in providing best practice examples or case studies in support of a decision.
- A guideline can provide information on the application of statistical experiment design (deciding how large a sample is required for a desired confidence interval).

Processes and tools

- In its simplest form, a process is a serial description of steps to be taken in solving a problem, and the associated tools, if any, are abstract (i.e. not implemented or computable).
- In its most complex form, the process is a cyclical graph (an example is a Gantt Chart) with decision points and multiple threads of execution, and the entire procedure is automated over distributed services, many of whom are not under the ownership or control of the actor requesting the process to be executed.

Knowledge workbenches

One of the fundamental problems addressed by the CoSAMP initiative is the conceptual framework and mechanisms whereby knowledge is to be structured for re-use by others. The conceptual framework states a number of 'desirable outcomes' for the framework:

- The structuring process is aided by the concept of workbenches: essentially a handy name for domain-specific templates that allow us to discover previously successful approaches, apply them, and adjust them to reflect new capabilities, use cases, and enhancements.
  - The workbenches will reference the processes (both structured and unstructured) as described above.
  - That workbenches, their properties, and their links to processes form a special type of structured knowledge, but it is envisaged that it will be necessary to store and reference unstructured or poorly structured knowledge as well.

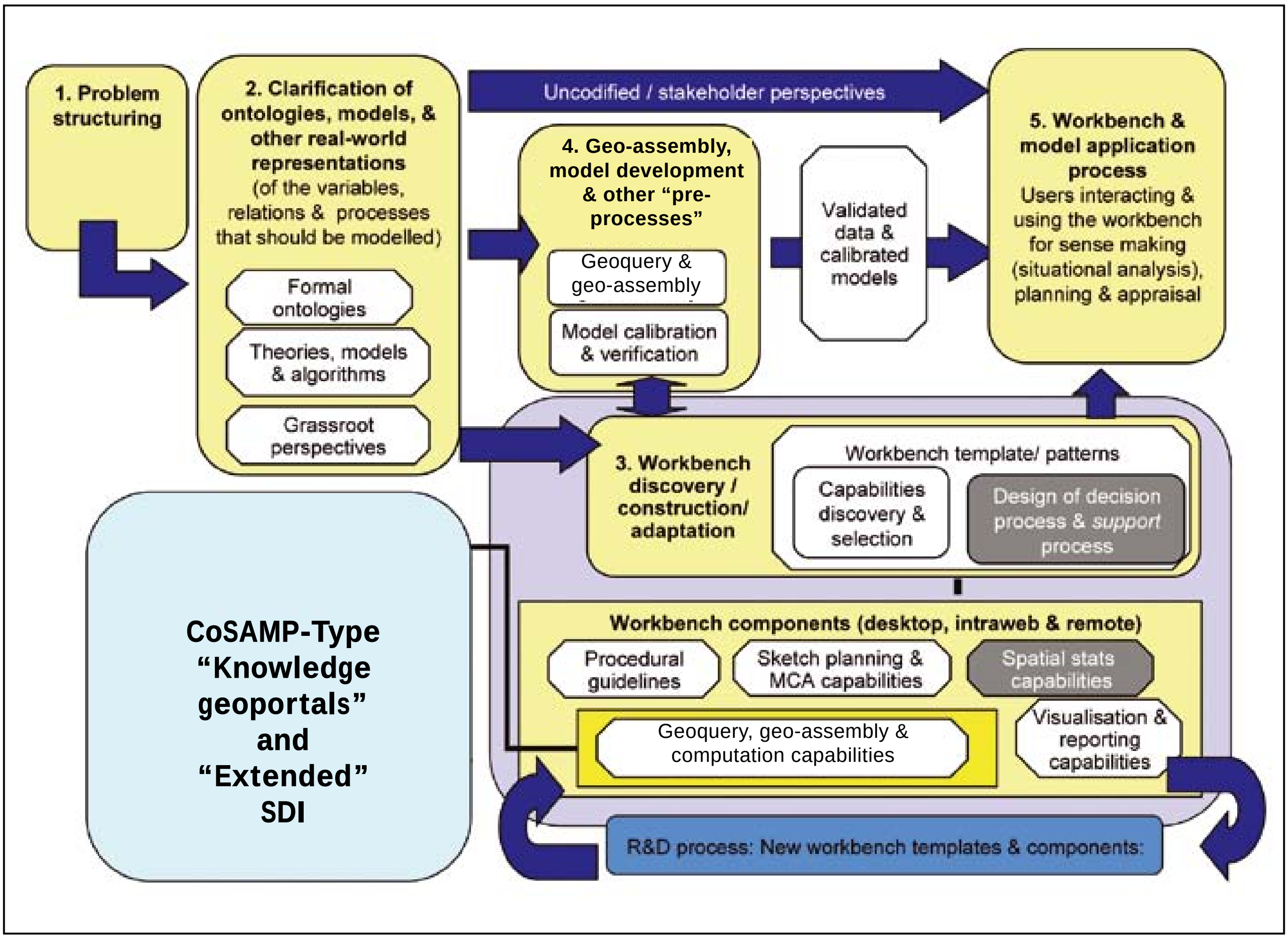

Fig. 4: Moving from problem definition, through knowledge workbenches, to geoportal applications [3].

Level	Example	Notes	Inputs and outputs
Informal process description	Textbook/ encyclopedia style process descriptions	The process description provides general background on the process, the typical problems it is designed to address, and the typical solutions (outputs) it produces. It is not possible to guarantee that a process will be found, and the naming on outputs, inputs, problems, and solutions is not standardised.	Description only
Formal process description	A methodology is a perfect example	A formal process description is not structured, but it conforms to an ontology. This means that the names of inputs, metrics, outputs, and other aspects of the process can be discovered and linked through keyword ontologies.	Weakly typed
Structured process description	A 'structured' or 'encoded' methodology, such as a process that has been stored	This process description level is structured. It contains generic parts (i.e. shared by all processes, irrespective of ontology), and application-specific parts. In the current context, this implies that the process is described in terms of the process description framework proposed for CoSAMP.	Weakly or strongly typed
Technique(s) (one or more)	Algorithm	An algorithm is a pseudo-code description of a process, its inputs, and outputs. The pseudo-code can be in a formal encoding such as UML or any other accessible standard, and must be complete.	Weakly or strongly typed
'Method'/ Function (one or more)	Implementation-ready Class(es)		Strongly typed

Table 1: Considerable effort was expended in the CoSAMP business and user requirements analysis phase to understand, typify, and stratify the processes that are implied by "geoprocessing".

- That in future, it may be possible to automate the methods used to match problems to pre-defined solutions to an increasing degree, and that the structured workbenches may play a key role in this.
- That some processes and workbench implementations will be aided by standardised data models and real-world implementations of these: specifically looking at predefined "geoframes" both in an abstract and implementation sense.
- That it will be possible to collect the various components of our knowledge (data sources, datasets, process definitions, workbenches, documentation, and so on) into portal environments that we refer to as a geoportal, that allows us to do the following:
  - Organise the knowledge in one or more logical structures;
  - Apply ontology to improve the discoverability of the knowledge;
  - Use portal capabilities to derive, create, and manage new instances of the knowledge components in a controlled and predictable way.
- That it will be possible to implement a collection of distributed, interlinked geoportal nodes, sharing some of their resources more directly with one another.
- That it will be possible to specify and manage the visibility, access, and conditions of use of all portal resources for a collection of user communities.

Processes ("workflows") for geoprocessing

Considerable effort was expended in the CoSAMP business and user requirements analysis phase to understand, typify, and stratify the processes that are implied by "geoprocessing" (see Table 1).

It is clear from Table 1 that any system that attempts to provide portal-like, knowledge-based access to spatial analysis processes will require a rigorous classification and definition capability – geoprocessing as described above will not work without it.

Description	Discussion
Transport "Connect"	Not directly applicable: achieved by the internet and its protocols.
Security "Secure"	Network security, standards-imposed security, and our own security measures.
Service description "Publish"	Standards and enabling technology for description of services. In the TCP/IP – HTTP environment, this is to a large extent dominated by XML Web Services descriptions – utilising WSDL.
Service discovery "Find"	Standards and enabling technology for locating services and content. As a rule, the generalised service is catalogued in UDDI-compliant resources, while in the geospatial domain this is accomplished by the OGC Catalogue Services specification.
Service binding "Bind"	Service bindings are provided in general by XML Web Services and SOAP, and specifically in the case of the geospatial domain by OGC services – WMS, WFS, WCS being the most important.
Modelling and workflow "Chain"	Again: provided in general by XML–based standards (such as BPEL), and specifically in the case of the geospatial domain by OGC process specifications – WPS (Web Processing Service).
Problem and solution "Match"	Modelling of knowledge: it is difficult to match solutions to problems without human intervention, and the encoding of the knowledge to be able to do so formed a part of the original CoSAMP research. The section in this document that deals with weak and strong types makes a substantial argument for some form of signature-based matching of problems to solutions. To our knowledge, no firm standardisation exists in this domain.
Knowledge application and extension "Apply"	This set of enabling standards and technology must ensure correct application and context. The emerging "semantic web' movement is an example of standardisation efforts to support this aim. The standards are aimed at 'shared meaning' in the interoperability space

Table 2: The definition of a continuum of standards and specifications by which such an extended SDI might be achieved.

				Least Ontological Meaning								Most Ontological Meaning
	Abstract Layers			"Connect"		"Secure"	"Publish"	"Find"	"Bind"	"Chain"	"Match"	"Apply"
	OSI-RM			TCP/IP		Distributed processing, including distributed spatial analysis					Knowledge management	
	Stack-->			Transport		Security	Service description	Service discovery	Service binding	Modelling and workflow	Match problems to solutions	Apply/ extend knowledge
Most Physical	Layer 1	Physical layer	Devices, cables, interfaces, voltages, and similar	Wire Radio Fiber Optic	Link							
	Layer 2	Data link layer	Physical data exchange: switches, network cards and bridges	Ethernet WIFI ISDN								
	Layer 3	Network layer	Routers	IP	Inter-network							
	Layer 4	Transport layer	Transports transparently between end users	TCP	Data Transport							
	Layer 5	Session layer	Management of dialogue between end users	TCP								
	Layer 6	Presentation layer	Definition of presentation from communication streams, elimination of syntactic differences	SSL	Application	SSL						
	Layer 7	Application layer	Semantic conversion	HTTP FTP SMTP		HTTP	HTTP	HTTP	HTTP RPC	HTTP RPC		
Least Physical	Layer 8	Signature layer	Matching applications on the basis of schema, signatures or interfaces	W3C OGC OMG, ...	URI/ URL		XML WSDL	UDDI	XML SOAP	XML BPEL	Schema/ types/ signatures	XML semantic web

Process elements and standards identified by Liping Di (2005)

Additions derived from CoSAMP contributions

Layers specified by TCP/IP

Layers specified by OSI-RM

Fig. 5: Standards and specifications for ESDI and knowledge geoportals.

Enabling standards and architecture

The final major contribution from the CoSAMP project in respect of extended SDI involves the definition of a continuum of standards and specifications by which such an extended SDI might be achieved. In short, it provides an overview of the technology base on which such an ESDI can be built. The schema discussed in Table 2 extends the work first published by Liping Di [9].

The schema as described in Table 2 can be ordered and linked to specifications by way of Fig. 5.

Conclusion

The CoSAMP project provided insightful work on the way in which typical SDI can be extended to include knowledge-based objects in a workbench. The technology framework for implementation of ESDI exists or is being created by way of standardisation efforts driven by several international organisations (such as OGC and ISO).

The future achievement of the conceptual ideal described in this paper hinges, though, to a large extent on:

- The effectiveness of unifying ontologies to describe the processes and techniques implied by knowledge workbenches.
- The extent to which it will be possible to define inputs to processes in enough detail for automated processing and process chain ing.

References

[1] CoSAMP: Collaborative Spatial Analysis and Modelling Platform – refer to: www.csir.co.za/publications/pdfs/annualreport_2006/Built.pdf.

[2] As defined by the United States National Spatial Data Infrastructure – refer to: www.whitehouse.gov/omb/circulars/a016/print/a016_rev.html#background.

[3] Andries Naudé and Graeme McFerren: "Collaborative Spatial Analysis and Modelling in a Research Environment", Internal Meraka Institute Publication. Obtainable at http://ict4eo.meraka.csir.co.za:8081/ict4eo/publications-and-research-outcomes/documents/paper4CSIRseminar-final_draft.pdf/view.

[4] CoGIS – refer to: www.cogis.co.za.

[5] Sensor Web activities are undertaken by the Meraka Institute of the CSIR – refer to http://ict4eo.meraka.csir.co.za:8081/ict4eo.

[6] Wim Hugo, Andries Naudé, Graeme McFerren, Louis Waldeck: "BRS (Business Requirement Specification) for Interlinked GeoPortals" - December 2005 – Internal CoSAMP Project Documentation.

[7] Wim Hugo: "URS for Portal Integration" and "URS for Distributed Spatial Analysis" – February 2006 – Internal CoSAMP Project Documentation.

[8] Wim Hugo: "URS for Portal Integration (II)" and "URS for Distributed Spatial Analysis (II)" – November 2007 – Internal CoGIS Project Documentation.

[9] Liping Di: "Distributed Geospatial Information Services-Architectures, Standards, and Research Issues". 2004, The International Archives of Photogrammetry, Remote Sensing, and Spatial Information Sciences. (7 pages - Invited)

[10] Open GeoSpatial Consortium – refer to: www.ogc.org.