

Designing an Evolving Federated Geographic Data Repository

Dr. Philip Sargent

Visiting Scientist

European Commission Joint Research Centre
Ispra, Italy

Philip.Sargent@computer.org

<http://www.sargents.demon.co.uk/Philip/>

Abstract. With the development of less-proprietary and more interoperable geographic information systems (GIS), a common problem that faces organisations is how to mobilise their historical geographic data archives to address current demands and new requirements. This paper reviews the issues, technical and managerial, that must be faced when attempting to assemble sets of separately-organised and independently-collected geospatial data. The issues will be seen to be almost entirely independent of raster/object distinctions, and to depend heavily on meta-data maintenance and update administration.

1 Introduction

The ARIS (formerly AIS Unit) of the Space Applications Institute¹ produces a regular bulletin on agro-meteorological status of Europe and North Africa [1] and has accumulated significant archives of satellite image and ground measurement data. These data have been gathered using rectanguloid area samples² distributed evenly across Western Europe [2].

New responsibilities require a different system across a greater geographic range and for a wider variety of applications than crop-monitoring. The problem is to design a long-term information system and archive for the existing area sample site data, the new catchment-oriented datasets and all ancillary information that can be (or could be) georeferenced.

This paper does not address the desirability (or otherwise) of moving from an area sample to a catchment sample methodology.

The 5th Framework Programme (paragraph III.4c) explicitly says:

“A multidisciplinary database of harmonised and coherent geographical data for an enlarged European Union. JRC will provide the necessary technical co-ordination and will foster the development of standards for software and methods aiming at the complete interoperability of geographic information systems.”

<http://www.jrc.org/jrc/fp5.htm>

¹ SAI - at the European Commission Joint Research Centre at Ispra, Italy.

² These were aligned with SPOT image boundaries for convenience.

Agenda 2000 says: "...and for extending the Union's borders through enlargement as far eastwards as the Ukraine, Belarus and Moldova."

<http://europa.eu.int/comm/agenda2000/overview/en/agenda.htm>

So that's a *lot* of geographic data covering more than 30 countries with vastly different statistical and cartographic traditions - not counting any of the "Mediterranean Partnership countries" of North Africa.

2 New Requirements

There are a number of new requirements:

- Multiple data ownership and multiple data update authorities are expected to be important in the future.
- The number of applications which use this data is open-ended but will include topography, meteorology, soil, land-cover and land-use. The update frequency for the classes of data is not yet set.
- The constraints are that:
 1. Existing data be left in its original format and storage media unless migration costs can be justified and funded. This includes the satellite image library, aerial photo image library, soil database, meteorological station data, CGMS³ data and CORINE [3] land cover data.
 2. The catchments are hierarchical in three levels which requires consistency maintenance for some types of data, e.g. water flow volumes.

It is therefore clear that the problem is not yet well-enough defined to be able to specify any type of architecture, data-dictionary or schema, but nevertheless there is a great deal known about this general type of problem and about this stage of the requirements definition process.

A brief review of multi-database and federated database architectures is given in the next section.

Given the requirements and constraints, it is clear that we are in the "Preliminary Planning" phase of the specification process [4], and that we will be shortly moving on to the "Requirements Development" phase. At this time (preliminary planning phase) we therefore have the following immediate actions that must become true if the project is to succeed (from Steve McConnell [4]):

- The project has a clear vision.
- The project team has identified an executive sponsor with final authority over projectwide decisions.
- The project plans and progress compared to the plans are readily available to all project team members and upper management.
- The project has a risk officer.
- The project has a Top 10 Risks list.

³ Crop-Growth Monitoring System

- The project team develops risk-management plans for each risk on the Top 10 Risks list.
- The project leaders acquire well-qualified people (waiting for well-qualified people, if necessary), rather than just using whoever is available first.
- Time-accounting is begun before requirements development begins.
- All of the above considerations are formalized in a Software Development Plan.

An example of such a plan and illustration of how it can be easily published and maintained is available within the ARIS Unit at this internal website: <http://unit.ais.sai.jrc.it/people/philip/local/moolu/>.

A similar checklist is appropriate for the next phase, “Requirements Development”.

3 Multiple Database Architectures

The technical issues of connecting and maintaining multiple databases with separately evolving schemas have been well studied for many years [5–13].

In December 1997, Amit Sheth, Director of the Large Scale Distributed Information Systems Lab. at the University of Georgia (USA), gave the keynote speech [13] at “Interop’97” where he forcefully reminded the attendees that the computer science community had been working with federated database schemas and semantic problems for more than a decade, with little success, and that the GIS community should not expect to be able succeed in a difficult area in which they have little expertise. In fact, the strictly spatial aspects of database integration seem to be relatively well-understood compared with the semantic issues, but similarly complex, manual and intractable [14].

It is widely known that multiple database projects require continuous maintenance because the shared schema is extremely fragile [15]. To compensate for the fragility requires active and intensive management of schema evolution [16–18]. Necessary techniques for effective schema evolution include explicit semantic representations [19–21] and metadata and ontology management [22–24]. Even strictly limited schematic and syntactic heterogeneity between component databases is beyond the current state of the art, possibly because their resolution is either intensely manual or because automated tools at these levels would themselves require a solution to the semantic problem [19].

“...the basic assumption underlying any kind of automation, i.e. having a schema with expressive names and a low degree of inter-relationships between the entities, *is wishful thinking* in most practical environments.” [10] (my emphasis). It is tacitly assumed that the schema work in any project of this type would be entirely object-oriented [25], even if the underlying databases were strictly relational.

The problem is that even if a combined schema can be assembled at one time, as the system is used it acquires new needs which have to be reflected in either new attributes (new data) or new interpretations of existing data. New attributes

mean that a database schema has to be updated, and all the other software and other schemas in other databases need to be updated to reflect this. Also, before the update, all other schemas must be checked to see whether there is semantic overlap between the new proposed architecture and any data already existing in another system. New interpretations of old data have the same fundamental problem but are even more difficult to represent within schema definition tools.

Even when two schemas are developed using identical schema-design languages, on a common data model, it is necessary to use *schema mapping languages* to represent and compare the two schemas. These languages, e.g. Express-X, BRIITY [10], are current research topics, not proven tools. Transformation of *instance data* between schemas can be done if designed and implemented purely within a theme (domain), e.g. the pedotransfer rules system developed for soils [26].

Even if the semantics of the schemas can be captured and maintained, they still need to be distributed and interpreted by the other software which forms the federated database. Current metadata systems are not up to this task [27, 28].

The problems of imprecise and intrinsically indeterminate data [29] and spatio-temporal data [30], which should be areas under intensive research, become essentially impossible in an operational multiple database scenario with currently available technology.

3.1 Multibases and Federations

Historically, there have been two approaches to the multiple database problem where we want users to be able to find, access, read and perhaps update a set of different databases but where they have the illusion that they are dealing with one database:

- the “multibase” approach,
also called “Multibase Management Systems (MBSs)”,
- the “federated database” approach,

“Multidatabase” is the generic term used to describe the situation where independently created and administered databases are to be used together. These differ physically and logically, have different access procedures, protocols and different concurrency (update) controls. Because they are independently developed, they may have different data manipulation languages and (sometimes) different data models (e.g. object and relational) as well as different schemas. The biggest problem, however, is *semantic* conflicts [7, 13].

“Federated databases” are a subset of multidatabases where there are more controls on the constituent databases. A single data model is usually possible and a single organisation can be responsible for the integration mechanisms. Nevertheless, the individual databases will be to some extent autonomous and will have a large degree of administrative independence – including update responsibilities and concurrency and access control [9–11].

The federated approach has had very limited success in individual companies in the automotive, oil and aerospace industries.

The multibase approach has had very shallow and partial success using least common denominators only. Internet search engines operating over the free text of web pages do not find all the data that is there, and they are limited to a single unstructured datatype: text.

3.2 Global Schemas

The obvious way of attempting to present a single view of multiple databases is to construct a *global schema*. This is a single schema that includes all the schemas of the constituent databases. Users manipulate data, via some middleware interface, as if they were using a single database which had that schema. If all the databases have been constructed using the same *conceptual schema*, then the global schema is the conceptual schema.

Unfortunately, extensive practical experience has shown that creating a global schema is exceptionally difficult, even when it is constructed using a small number of databases which use the same data model and were developed for almost identical purposes for the same user community.

There are several aspects to the difficulty:

- when constructing the global schema, there is no general solution to the *semantic conflicts* that have to be solved manually, and
- in operation, data conflicts arise if the independence of the databases is to be preserved: if two databases represent the same value for the same object, which is correct ?
- in operation, the middleware technology is not capable of bridging the different update and synchronisation semantics required by either the different databases, or the data itself,
- in operation, the global schema, constructed with such difficulty, is discovered to be extremely fragile as new interpretations are applied to the data in the individual schemas, and especially as additions to the schema become necessary as the system evolves.

The last point concerning the discovery of new relationships to be represented becomes clearer by considering the example of “dependencies”:

Dependencies: a relational database is constructed by analysing the data dependencies and creating a table schema that respects these, e.g. in Fifth Normal Form, etc. However, some multi-attribute dependencies require complex *manual* update procedures, essentially arbitrary computations, which are outside the data model, e.g. topological constraints on land parcel survey data inside a database which has no spatial predicates in its update language, or the dependency of the strength of Nylon components on the level of humidity when it is a matter of research and experiment to discover if such a dependency exists [31, 32].

Therefore, except for systems which are designed from scratch to be essentially a single system, the general approach is to assume that the multiple databases have no global schema and that users (which may be software agents) must be aware that they are dealing with multiple databases. The multibase therefore needs to provide functions for manipulating data that are in visibly distinct schemas and may be mutually non-integrated.

Several research groups are attempting to construct “mediators”, middleware to mediate between multiple databases [9, 19, 24], but the effort and expertise involved is vastly more than that available in the ARIS Unit.

3.3 Feature Identity Issues

The Global Schema federated database approach, even if it could be effectively implemented, does not provide some capabilities commonly required by multiple geospatial data repositories.

The federated database architecture does not address issues of duplication and matching across the component databases - the approach tacitly assumes that foreign keys exist that can unambiguously identify features and ensure integrity. With geospatial data no such assurance can be given and the Open GIS Consortium [33] currently has as its highest priority the production of standards for “Feature Identity and Relationships” [34].

The problem of unique identifiers for objects within multiple databases and within the entire Internet is an extremely active area of standards development [35–37].

4 Management

Managing a federated database system requires an *operational* management style rather than a *research* style. The Joint Research Centre at Ispra may have had such management expertise in the past when it was running nuclear reactors. The attention to repetitive detail, strict regulations and work that is carefully checked by several different people are characteristics that are required by nuclear safety management. However, nuclear operations standards have increased in strictness and cost over the decades and their increasing cost, and difficulty with this style of management in the JRC, could have been a factor in closing down the reactors.

Two examples illustrate the current situation in the ARIS Unit.

1. The existing static archive of image data is not managed: it is simply stored. The tapes are not in order, not indexed and not re-wound every three years as is necessary to prevent “print through”.
2. The extremely simple management task of monitoring and enforcing trivial controls on the ARIS missions budget is observably beyond the management capabilities of the Unit.

An organisation cannot attempt a mission which is in flat contradiction to its organisational culture – whether that culture is officially recognized or not [38].

5 What Can Be Done ?

If neither the technical nor the managerial tools exist to run the catchments data and information archive as a coherent entity, the only practical solution is to run the information archive as separate thematic projects. Integration between pairs of themes is not then impossible, it just requires to be properly planned and managed (and costed) as a project in its own right; a project with defined targets and an end- date.

The separate themes should be run with as common a data model as can be afforded, in design and on-going maintenance time, but it should be recognised from the start that coping with the multiple viewpoints inherent in the multiple data sources is *essentially a manual task*.

6 Conclusions

The deep technical problems have no automated technical solution; but there are manual management solutions. However these management solutions are out of reach of the ARIS Unit.

The update management, metadata management and on-going ontology maintenance tasks require, in the context of the European Commission Joint Research Centre, a dedicated Unit with its own independent budget. This dedicated group requires a specific mission to obtain and manage the data, with specific service level agreements (SLAs) to the other research-oriented Units within the JRC who will develop applications which use the data.

It is extremely unrealistic [38] to expect that the goal of providing a reliable managed data service can be combined with the other goals of the ARIS Unit.

Author's Qualifications

The author is well qualified to assess this problem in both its technical and managerial aspects. Prior to working in geographic information, he was head of software applications development and deployment at a company installing Product Data Management software [39] in large engineering companies in the UK.

The author is experienced with specifying, designing and implementing software database systems to hold nuclear engineering data and North Sea oil platform design data [40]. Both industries require assured access, resilience and flexibility of their information systems for a minimum of thirty years. However, both industries are prepared to pay the costs and prepared to set up specific management systems to ensure that the information stays available and current.

Acknowledgements

The paper is a personal view written while the author was a Visiting Scientist at the Space Applications Institute (SAI) of the European Commission Joint

Research Centre at Ispra, Italy. At the time of writing, the author had worked nearly 10 months within the in the Agricultural and Regional Information Systems (ARIS) Unit of SAI.

References

1. MARS - Monitoring Agriculture with Remote Sensing - Bulletin. European Commission publication, ARIS Unit, SAI, JRC Ispra, Italy. <http://www.ais.sai.jrc.it/marsstat/bulletin>
2. Taylor, C., Sannier, C., Delincé, J., Gallego, F.J.: Regional Crop Inventories in Europe Assisted by Remote Sensing. Report EUR 17319 EN, European Commission Brussels 1997. <http://www.ais.sai.jrc.it/coordination/ar96/2-2-96.html>
<http://www.ais.sai.jrc.it/coordination/ar96/figs/fig3-1.gif>
3. European Commission Joint Research Centre and European Environment Agency: Technical and methodological guide for updating CORINE land cover data base. EUR 17288 (1997),
4. McConnell, S.: Software Project Survival Guide. Microsoft Press 1998. ISBN 1-57231-621-7. <http://www.construx.com/stevemcc/sgbrief.htm>
5. Gligor, V.D., Luckenbaugh, G.L.: Interconnecting Heterogeneous Database management Systems. IEEE Computer, January 1984, 33-
6. Cardelli, L., Wegner, P.: On Understanding Types, Data Abstraction, and Polymorphism. ACM Computing Surveys **17** (4) December 1985.
7. Litwin, W., Abdellatif, A.: Multidatabase Interoperability. IEEE Computer, December 1986, 10-18.
8. Mark, L., Roussopoulos, N.: Metadata Management. IEEE Computer **19** (12) 26-36 (1986)
9. Loeser, H. Härder, T. dLIMIT - A Middleware Framework for Loosely-Coupled Database Federations. Proc. 2nd Int. Conf. on Worldwide Computing and Its Applications, (WWCA'98), LNCS 1368, Springer, March 1998, pp. 412-427. <http://www-agdvs.informatik.uni-kl.de:18070/publications/LH98.WWCA.html>
10. Härder, T., Sauter, G., Thomas J.: Design and Architecture of the FDDBS Prototype INFINITY. Proc. Int. CAiSE'97 Workshop "Engineering Federated Database Systems (EFDBS'97)", Barcelona, pp.57-68.
11. Sauter, G. Käfer, W.: EXPRESS as the Common Data Model in Federated Database Systems. Proc. of the 5th Int. Conf. of the EXPRESS User Group, (EUG'95, Grenoble, France, October) 1995.
12. Morgenstern, M.: Metadata for Heterogeneous Databases. in: The Second IEEE Metadata Conference, Silver Springs MD. (USA), September 1997. http://www.llnl.gov/liv_comp/metadata/md97.html
13. Sheth, A.: Semantic Interoperability in Infocosm: Moving beyond infrastructural and data interoperability in federated information systems. Interop'97: Intl. Conf. on Interoperating Geographic Information Systems, Santa Barbara, California, USA, Dec. 3-4, 1997. <http://lstdis.cs.uga.edu/~amit/>
<http://www.ncgia.ucsb.edu/conf/interop97/program/papers/sheth.html>
14. Laurini, R.: Spatial multi-database topological continuity and indexing: a step towards seamless GIS data interoperability. Int.J. Geographical Information Science **12** (4) 3733-402.
15. Frost, R.A., Whittaker, S.: A Step Towards the Automatic Maintenance of the Semantic Integrity of Databases. The Computer Journal **26** (2) 124- (1983)

16. Teorey, T.J., Wei, G., Bolton, D.L., Koenig, J.A.: ER Model Clustering as an Aid for User Communication and Documentation in Database Design. *Comm. ACM* **32** (8) August 1989, 975-.
17. Osborn, S.L.: The Role of Polymorphism in Schema Evolution in an Object-Oriented Database. *IEEE Trans.Knowledge and Data Engng.* **1** (3) September 1989, 310-.
18. Hurson, A.R., Pakzad, S.H., Cheng, J-b.: Object-Oriented Database Management Systems: Evolution and Performance Issues. *IEEE Computer* February 1993, 48-.
19. Bishr, Y.: Overcoming the semantic and other barriers to GIS interoperability. *Int.J. Geographical Information Science* **12** (4) 299-314.
20. Hull, R., King, R.: Semantic Database Modeling: Survey, Applications, and Research issues. *ACM Computing Surveys* **19**, (3) Sept.1987.
21. Abiteboul, S., Hull, R.: IFO: A Formal Semantic Database Model. *ACM Trans. Database Systems* **12** (4) December 1987, 525-565.
22. Mark, L., Roussopoulos, N.: Interformation Interchange between Self-Describing Databases *Information Systems* **15** (4) 393-400 (1990)
23. Ontology Markup Language,
Available at <http://wave.eecs.wsu.edu/CKRMI/OML.html>.
24. Wiederhold, G., Genesereth, M.: The Conceptual Basis for Mediation Services. *IEEE Expert*, Sept./Oct. 1997, 38-47.
25. Celko, J., Celko, J.: Debunking Object-Database Myths. *Byte*, October 1997, 101-105.
26. Daroussin, J., King, D.: A Pedotransfer rules database to interpret the soil geographical database of Europe for environmental purposes. Workshop proceedings, "The use of pedotransfer in soil hydrology research in Europe", Orléans, France, 10-12 Oct.1996.
27. ISOTC 211WG 3, Geospatial data administration, Part 15: Metadata, <http://www.statkart.no/isotc211/wg3/wg3welc.htm>.
28. Committee on Earth Observation Satellites: CIP - Catalogue Interoperability Protocol, <http://lcweb.loc.gov/z3950/agency/profiles/cip.html>.
29. Burrough, P.A., Frank, A.U., (eds.): *Geographic Objects with Indeterminate Boundaries. GISDATA2.* Taylor and Francis, London (1996) ISBN 0-7484-0386-8 (series editors Ian Masser and François Salgé)
30. Langran, G.: Issues of implementing a spatiotemporal system. *Int. J. Geographical Information Systems* **7** (4) (1993) 305-314
31. Sargent, P.M.: *Materials Information for CAD/CAM.* Butterworth-Heinemann Publ., August 1991, 172 pages, ISBN 0-7506-0277-5.
32. Sargent, P.M.: Engineering information handling technologies, Chapter 27, 634-663, in *Handbook of Engineering Management* (2nd. edition) D.Lock (ed.), Butterworths Heinemann Publ., October 1993, ISBN 0-7506-0786-6.
33. The Open GIS Consortium: The OpenGIS[®] Implementation Specification, OpenGIS[®] Simple Features Specifications for OLE/COM, CORBA and SQL, <http://www.opengis.org/techno/specs.htm>.
34. Sargent, P.M.: Feature Identities, Descriptors and Handles. Draft paper (submitted to Interop'99).
<http://www.sargents.demon.co.uk/Philip/feature-ids/base.html>
35. Berners-Lee, T, Fielding, R., Irvine, U.C., Masinter, L.: Uniform Resource Identifiers (URI): Generic Syntax, Internet Engineering Task Force (IETF) Request For Comment (RFC) 2396, August 1998, <http://info.internet.isi.edu:80/in-notes/rfc/files/rfc2396.txt>.

36. World Wide Web Consortium: Naming and Addressing: URIs, <http://www.w3.org/Addressing/Addressing.html>.
37. Sun, S.X.: Handle System: A Persistent Global Name Service — Overview and Syntax, Internet Engineering Task Force (IETF), Work in progress — Internet Draft July 16, 1998, Document `draft-sun-handle-system-01.txt` in <http://www.ietf.org/internet-drafts/>, see also <http://www.handle.net>.
38. Handy, C.: Understanding Organisations. Oxford University Press, 4th Edition (1993) ISBN 019-5087321
39. Hewlett-Packard (HP), Understanding Product Data Management. <http://www.pdmic.com/undrstnd.html>
40. Quillion Information Systems Ltd., Cambridge UK. http://www.quillion.com/q_web_3/stud02.htm