The CoGIS Portal

Background and History

The CoGIS[1] portal has developed from initial project work undertaken by the CSIR and SAEON, together with other stakeholders, in 2004/5. This initiative was called CoSAMP[2], and provided the backbone for a scaled-down, more operational platform called CoGIS, which was developed between 2006 and early 2008. To some extent, CoGIS can be viewed as a component of the larger CoSAMP environment.

While CoSAMP was conceptually geared towards management of a wide variety of content items, and its associated meta-data (ranging from geospatial data sets through images, documents, and databases to knowledge repositories and process descriptions), CoGIS focuses more directly on the provision of data (mostly spatial data, but not limited to it), and its associated meta-data in a collaborative environment. Specifically, the more ambitious goal of knowledge and process management was excluded from CoGIS.

CoSAMP and CoGIS were formally audited on two occasions[3], and the results from these audits were used to determine gaps between user community expectations, formalized in a series of URS documents, and the CoGIS implementation at that time. The gaps, as identified, formed the basis of a development and extension program that was embarked on by CSIR and SAEON in 2008/9, and is currently under way. This program aims to achieve the following:

1. Meta-Data (Completion: March 2009): Extending CoGIS to work with **multiple meta-data standards**, based on the premise that not all data sets and content items will be adequately described by one standard only. Specifically, earth and environmental observation data that are spatially referenced in South Africa will typically require at least two standards:
    a. SANS 1878 to describe its geospatial meta-data,
    b. and an overlapping, domain-specific EML record to describe it to environmental scientists.
    c. This extension is now available and operational.
2. Spatial/ Image Extensions (Completion: November 2009): CoGIS requires extension to improve its capabilities in respect of **client-side mapping**, and the manner in which map composition is handled, supporting spatial data sets that are not only available locally, but in standards-compliant services world-wide.

[1] Collaborative Geographic Information System
[2] Collaborative Spatial Analysis and Modelling Platform
[3] G294.2.1.3 CoGIS Audit Report.doc, G294.2.1.2 URS Portal Integration.doc, G294.2.1.3 URS Mapping Clients.doc, 2 October 2007

3. Structured Data (Estimated: November 2010): This work package has not been scoped, but is required to assist with the standards-based storage, querying, extraction, and collation of large observational data sets (mostly time-series data). It is aligned to the growing interest in Virtual Research Environments (VRE's), and the original goals of the CoSAMP project.

4. Citations, Meta-Data Sources, Copyright (Estimated: March 2010): This requirement stems from the need to be able to, *inter alia*,

    a. cite data sets and documentation obtained from the portal correctly and automatically;

    b. preserve the chain of meta-data provision in cases where the meta-data was harvested from other sources;

    c. properly acknowledge copyrights, usage restrictions, or both, by way of example acknowledgements and license inserts, as applicable.

5. Meta-Data Extensions: Phase II (Estimated: November 2009): Additional meta-data standards need to be addressed, and integration established between overlapping standards to minimize duplication and contradiction.

Other, smaller extension projects are envisaged, but are less pertinent to the current discussion.

It must also be noted that the majority of the governance issues raised in the audit report (systems engineering approach, management of issue resolution, licensing and ownership, alignment of stakeholders) have since been addressed.

Meta-Data Extensions

CoGIS is ***ready for operationalization as a meta-data clearinghouse***, supporting several meta-data standards. In the main, this extension offers the following functions to the general user community:

(1) The ability to 'harvest' and aggregate meta-data records from all stakeholders and partner organizations in a ***multi-standard meta-data repository***[4], with automated processes to maintain synchronization with sources.

(2) The ability to ***search for meta-data*** records in a number of ways, ranging from very general phrase-based searches to tailor-made, very specific searches, and to package the search definition for future use or embedment in other sites and applications. This allows targeted searches, for example, to be directed to the CoGIS portal from corporate, node, or project websites.

(3) The ability to ***navigate to or download data*** referenced by the meta-data records.

In addition, system and community administrators can configure search facilities, controlled vocabularies, and content for a community to use.

[4] CoGIS currently supports ISO 19115, ISO 19115 p2, ISO 19139, EML, Dublin Core, FGDC, and SANS 1878.

Some progress has been made in respect of making CoGIS more service-oriented, allowing other systems to interact with it in an automated fashion. Specifically, two possibilities are of interest:

(1) The ability to embed ***the results of saved searches into collaborating systems*** as RSS feeds or GEO-RSS Feeds, allowing seamless integration of CoGIS search results into other environments;

(2) The ability to ***expose subsets of CoGIS meta-data to other harvesters, specifically the NSIF***, so that the ***legal obligation to lodge SANS 1878 meta-data with them can be satisfied via the CoGIS Portal***. This is of particular use to CSIR spatial data publishers.

SAEON will commence with implementation of the COGIS portal in May 2009, and has embarked on a training program for its node data managers to introduce them to the functions, and how to best make use of these functions.

CSIR will commence implementation in August 2009, and develop, resource, and execute operationalization plans to support such implementation. The approach followed by SAEON and its attendant support materials/ documentation is available as a starting point.

Future Work

In addition to these operationalization tasks for the meta-data management functions, CoGIS will need investment in three streams of work in the near future (End March 2010):

(1) A ***program for functional extension and improvement*** has been compiled, and blends the shortcomings identified in the Audit Report with the requirement for wider use of the portal by SAEOS, The Risk and Vulnerability Atlas, the World Data Centre, and similar initiatives.

(2) A program to enable SAEON and the CSIR to assist stakeholders and contributors with the ***provision of meta-data***, and, ultimately, ***data sets for curation and long-term preservation***. This effort, from a CSIR perspective, can be focused on extending internal facilities for storage and backup of large data sets.

(3) A program to review ***and validate the capabilities of the hardware infrastructure*** (connection speeds, feasibility of transfer of large data sets, ability to manage large meta-data collections, …), and recommendations in respect of future hosting arrangements and infrastructure.

In the longer term (2010/11, 2011/12), the investment in CoGIS should focus less on functionality extension and more on extension of the data and meta-data holdings, value addition to the available data, and ongoing refinement/ alignment of technology.

Alignment with other Initiatives

CoGIS has the potential to be extended and utilized for three other initiatives:

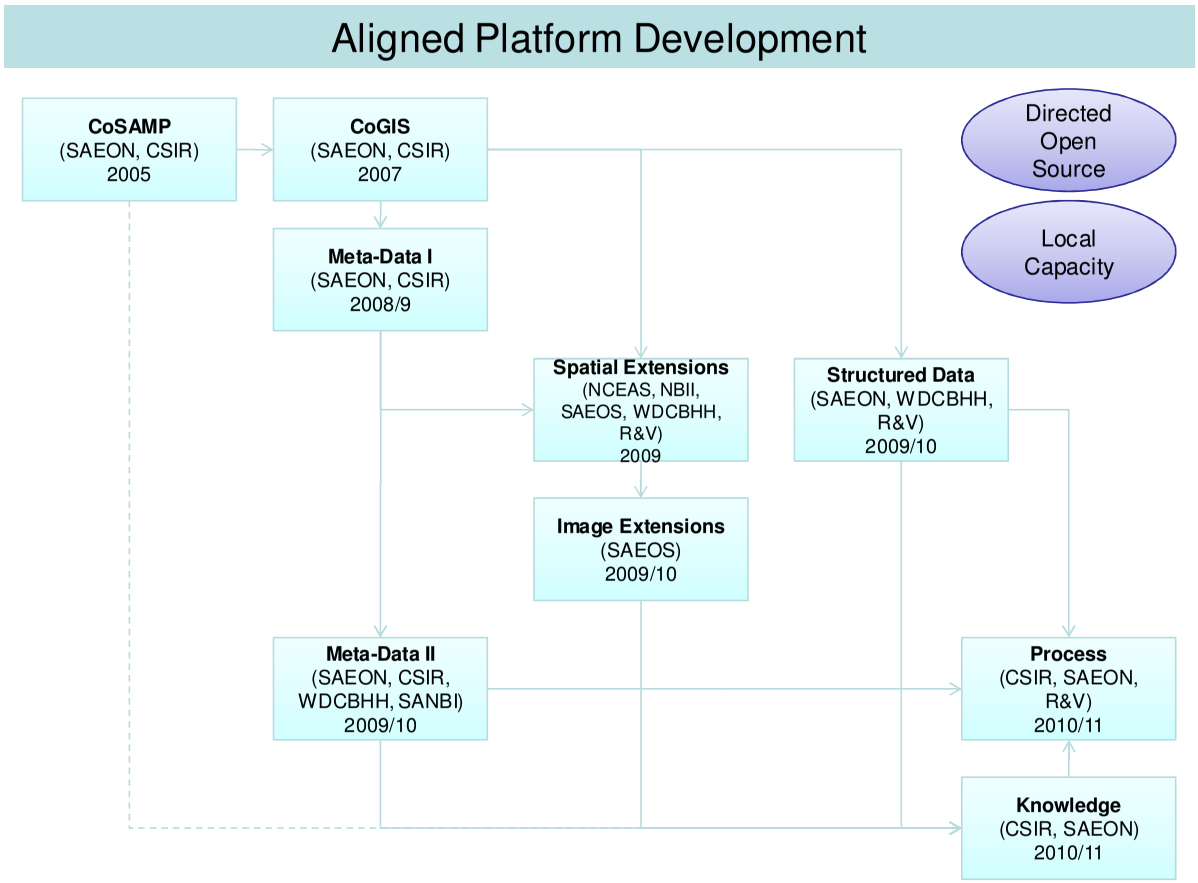
1. To act as the portal environment for SAEOS (a proposal in this regard has been requested by DST and provided through SAEON).

2. To act as a portal environment for the World Data Centre for Biodiversity and Human Health (WDCBHH) in Africa. A Project planning exercise, in which CSIR and SAEON were active participants, has been conducted in Washington DC (Feb/ Mar 2009) and a project design has been published.

3. To act, with extensions, as the basis for the Risk and Vulnerability Atlas, a longer-term program managed by CSIR, but involving researchers and contributors in many other organizations.

The WDCBHH, Risk and Vulnerability Atlas, and SAEOS requirements align well with work that is programmed for traditional stakeholders in the extension and refinement of the current CoGIS portal. The alignment and collaboration foreseen, should SAEOS/ WDCBHH form part of the stakeholder group that builds the general capabilities of the platform, is summarized in the diagram.

In addition, the NSIF already shares a substantial part of the opens source infrastructure on which CoGIS has been built, and alignment of their future development with that of CoGIS is of critical importance.



The following additional notes:

1. SAEON and CSIR have co-funded the meta-data extensions currently under way (beta-testing in March/April 2009, operational in May 2009).

2. In terms of spatial capabilities, it should be possible to

    a. secure the collaborative involvement of USGS for strategic reasons[5] (this has been agreed to);

    b. attempt to re-use and extend components already available within the USGS/ GBIF/ GEOSS environment and adjust them for use with Plone (also agreed with USGS);

    c. fund the specific WDCBHH requirements in the prototype development, and use the basic functionality as an input for operationalisation for other stakeholders. The benefit to WDCBHH should be a more mature technology as and when operationalisation commences,

    d. Re-use and extend GAP-3, developed for the CSIR Built Environment, as a primary analysis and data visualization extension to the portal.

3. The next round of meta-data extensions, while not funded by SANBI, will involve them more directly, through inclusion of the Darwin Core standard so as to accommodate the SANBI species-related data sets. This is an imperative for the WDCBHH, and some of the funding available for WDCBHH should be applied to this end.

4. The World Data System (WDS) and its initiative in Africa, in which NRF and SAEON are set to play a supportive role, may benefit substantially from improvements in the handling of large structured data sets, but it is not a prerequisite for the prototype.

In broad terms, a shared infrastructure is in all stakeholders' interest, and each initiative is likely to extend and refine the common set of tools available to all initiatives.

Note that the intention in terms of CoGIS is to package it, over time, as a formal open-source project (or extensions to a current project), but that this process is dependent on availability of resources and funds. In the interim, the intention is to make the CoGIS portal as widely available as possible though a network of collaborators, and to ensure that new extensions are capable of being deployed as open-source extensions ('products') for Plone[6].

Vision

This short discussion Sketches a perspective on the possibility of a National Knowledge Management Infrastructure (NKMI), and specifically indicates how the proposed CoGIS extension projects support this ideal.

[5] As indicated in the SAEON IT strategy, SAEON and CoGIS should seek to assist major international organizations with extension and refinement of their platforms rather than developing their own, unique solutions.

[6] Plone (www.plone.org) is an open source content management system that forms the basis of CoGIS.

There is a strong trend[7] towards more and more portals aimed at the preservation, description, discovery, and dissemination of knowledge of all types. In support of this trend, there has been a rapid development of meta-data standards to aid with the description and discovery of the knowledge, diversified according to knowledge types and domains.

Three major problems are associated with this growth.

- There is fierce competition, sometimes wasteful, for the provision of technology infrastructure to support the large number of portals. A general counter-trend of standardization provides some direction.
- There has also been some divergence of meta-data standards within knowledge types and domains, although large organizations such as ISO (3) and OGC (4) attempt to mitigate this.
- There is some competition for domain-related mandates between different organizations.

The first two issues can be addressed in part by the establishment of a National Knowledge Management Infrastructure, while the NRF already plays a role in the management of the third problem in the local context.

Firstly, one requires competition to improve quality and choice, but undirected competition is often counterproductive, and choice brings with it selection and evaluation dilemmas, especially in the information technology field. Here, decisions need to be made with some degree of specific knowledge. Any solution to this problem needs to preserve choice and competition, take specific advantages and disadvantages of the open source ideal into account, and aim for optimality in terms of cost benefit. The simplest and least intrusive manner in which to do this is by way of ***shared implementation,*** a term used to describe a framework containing architecture guidelines, abstract specifications and user requirements, recommended interoperability standards, and reference implementations. The main aim is the creation of a ***directed open source community*** that contributes to, extends, and maintains the shared implementation. The impetus for baseline operation of such a community and its shared implementation is the NKMI, and special initiatives (WDCBHH, R&VA) or ongoing operations (SAEON, SAEOS) will make both continued contributions to the NKMI and draw from it.

Secondly, the potential divergence of meta-data standards and the knowledge it is based on requires an ongoing initiative to promote ***shared meaning.*** This finds its expression in meta-data standards selection, crosswalks, and in the controlled vocabularies, thesauri, and ontologies associated with these standards.

Thirdly, there is a requirement for ***shared processes*** – ranging from coordinated training and awareness creation on the importance of meta-data, to data sharing and meta-data provision policies to be implemented at research institutions and as a condition of funding, to coordination of and assistance with contributions to the infrastructure, to creation of forums for the validation of decisions and guidelines.

[7] It is not possible to review either recent trends in establishment of such systems, the WDCBHH initiative, or other context-providing material in such a short document, but some references and a short discussion on each of them are provided in an Annexure to the document.

Finally, there is the need for ***institutional arrangements*** and an explicit framework to guide the development of contributing portals, with the NRF funding all, some, or often none of such development.

Parallels are easy to find: The US National Biological Information Infrastructure [5] (NBII), DataNetOne [8], GEOSS [6], and GBIF [7] are all ready examples of how such a directed development focus can lead to a shared resource. The difference is that we recognize that ***the shared infrastructure is not domain-specific***.

It is matter of perspective rather than scope.

A. References and Further Reading

(1) "Project Charter: Portal Establishment for WDCBHH", SAEON and NRF, March 2009.

(2) ISO 19115 and related standards for spatial meta-data, the SANS 1878 family of standards for spatial meta-data as applicable in South Africa, EML (applicable to environmental data), the Dublin Core meta-data standard (applicable to all data), the Darwin Core standard (mainly applicable to taxonomic collections), and the FGDC with Biological Profile (applicable to resources obtained mainly from the US, and requiring translation to ISO or EML, and vice versa).

(3) ISO Technical Committee 211: Standardization in the field of digital geographic information. This work aims to establish a structured set of standards for information concerning objects or phenomena that are directly or indirectly associated with a location relative to the Earth. These standards may specify, for geographic information, methods, tools and services for data management (including definition and description), acquiring, processing, analyzing, accessing, presenting and transferring such data in digital/electronic form between different users, systems and locations. http://www.isotc211.org/

(4) Open Geospatial Consortium is a private international geospatial standardization organization. Consisting of over 300 member organizations from government, industry or education, the OGC develops interoperable consensus-based geospatial standards. The OGC provides standards up to ISO as well as developing its own standards. http://www.opengeospatial.org/

(5) NBII: National Biological Information Infrastructure - The National Biological Information Infrastructure (NBII) is a broad, collaborative program to provide increased access to data and information on the nation's biological resources. The NBII links diverse, high-quality biological databases, information products, and analytical tools maintained by NBII partners and other contributors in government agencies, academic institutions, non-government organizations, and private industry. www.nbii.gov

(6) GEOSS: GEO is a voluntary partnership of governments and international organizations. It provides a framework within which these partners can develop new projects and coordinate their strategies and investments. As of March 2009, GEO's Members include 77 Governments and the European Commission. In addition, 56 intergovernmental, international, and regional organizations with a mandate in Earth observation or related issues have been recognized as Participating Organizations. www.geoss.org

(7) GBIF: GBIF aims to be the preferred gateway, worldwide, to a comprehensive, distributed array of primary species-occurrence data. Much of GBIF's methods and approaches are applicable in the context of a national infrastructure. http://www2.gbif.org/visionen.pdf

(8) DataNetOne: Will consist of a large number of geographically distributed Member Nodes that house data archives and metadata describing those data. Member Nodes will be linked together by contributing metadata to a series of replicated Coordinating Nodes that provide valuable services to Member Nodes. The Coordinating Nodes provide a common infrastructure to handle, for example, distributed authentication, fault tolerance, geographic, taxonomic, and temporal search services, and data replication services. http://www.tdwg.org/proceedings/article/view/406