Institutionalising a GSDI at the CSIR

by Heidi van Deventer, CSIR

Institutionalising a Geospatial Data Infrastructure (GSDI) was a major watershed for research data management at the Council for Scientific and Industrial Research (CSIR). For the first time in the CSIR's history, the Information and Communications Technology (ICT) department took over the management and maintenance of research-related software and data. For the first time, geospatial data was related to publication outputs. Heidi van Deventer shares the experiences, lessons learnt and the remaining challenges.

Many predecessors at the CSIR have attempted awareness raising of geospatial data infrastructures (GSDIs) and the importance thereof for researchers in the CSIR. The GIS unit in Environmentek, who employed most of the champions of former GSDIs in the CSIR, took on the responsibility of managing servers, back-ups, archives, geoportals, metadata catalogues and software licences shared between operating units. These researchers also provided support functions in GIS and cartographic services to other units in the CSIR.

The Beyond 60 process however, saw a complete change in the strategy and structure of the GIS unit. Many of the staff have left the unit, either to other units in the CSIR or to external institutions. Owing to staff turnover and a strategy change to remote sensing (RS), this necessitated the move to a new era in the GSDI. The new Earth Observation (EO) Group in the Natural Resources and the Environment (NRE) operating unit, had a strategic focus on research only. There was also a general tendency for researchers to increase their skills in using geographical information systems (GIS), due to increased training courses available in the software in South Africa. It was estimated that the number of users at the CSIR increased in the early 2000 from about 20 to 80 people by 2009.

In 2006, the servers that used to store GSDI data were full and new disks could not be added due to the outdated technology. Management of the infrastructure related to the GSDI was no longer in the interest of the researchers, and the need for information and communications technology (ICT) to manage these was escalated to operational and executive management. Duplication of data increased as new researchers entered the CSIR and did not know about the existing server. The archives' CDs were becoming obsolete, and needed migration to managed infrastructure to retain geospatial data. The time for institutionalising the GSDI components at organisation level was ripe.

Establishing a link between support services and researchers through the GSDI

Two support sections in the CSIR are relevant to the GSDI, namely CSIR Information Services (CSIRIS) and Information and Communications Technology (ICT) (Fig. 1). CSIRIS is responsible for the curation and preservation of records published in the CSIR, and ICT's mandate and responsibility is to provide ICT-related services to the CSIR. Operating Units using and benefiting from a GSDI includes Built Environment (BE); Defence, Peace, Safety and Security (DPSS); Natural Resources & the Environment (NRE); and the Meraka Institute (MI).

CSIRIS have been managing the indexing of records in the CSIR. Reports, publication and documentation generated in the CSIR are indexed in a Technical Outputs Database (ToDB). Up to 2010, no records reflected a link to associated geospatial data and/or research artifacts. Large data sets were also not efficient to curate within the Document Management System (DMS) used by the CSIR. It would have meant that folders containing large geospatial data sets and projects, had to be zipped for back-up and archiving. In addition, ICT support focused primarily on

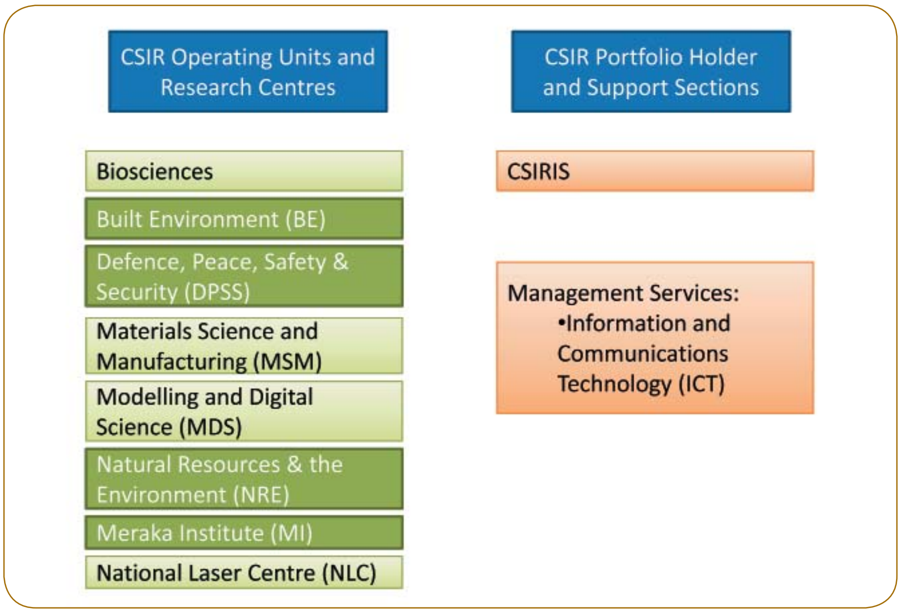
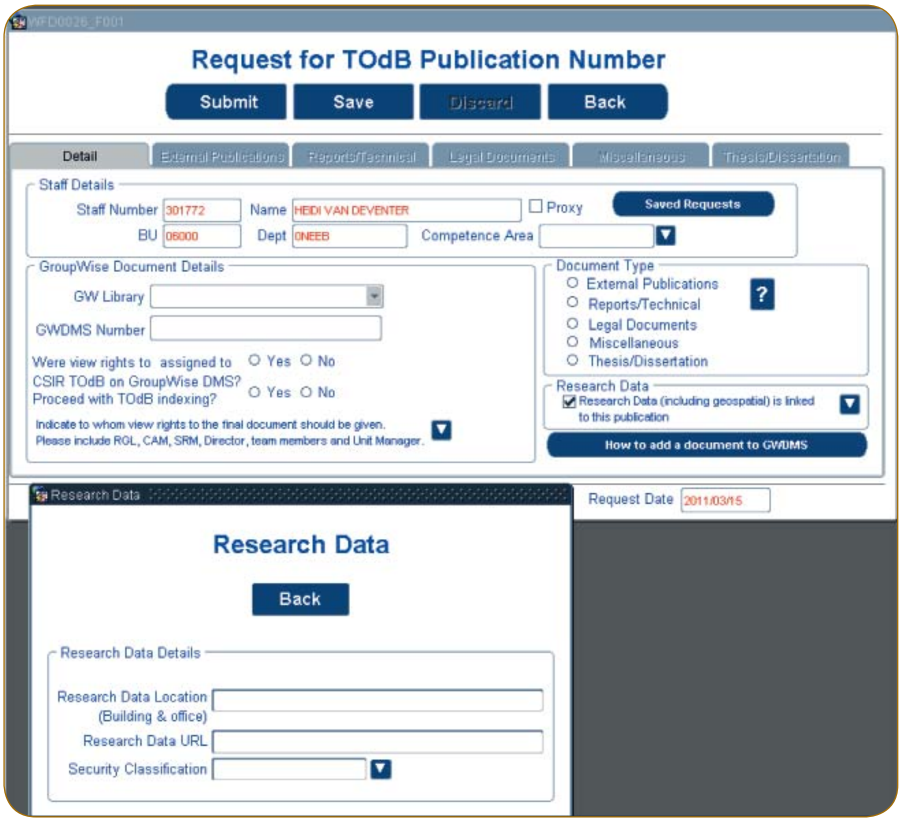

CSIR Operating Units and Research Centres
CSIR Portfolio Holder and Support Sections
Biosciences
Built Environment (BE)
Defence, Peace, Safety & Security (DPSS)
Materials Science and Manufacturing (MSM)
Modelling and Digital Science (MDS)
Natural Resources & the Environment (NRE)
Meraka Institute (MI)
National Laser Centre (NLC)
CSIRIS
Management Services:
•Information and Communications Technology (ICT)

Fig. 1: Simplified view of the operating units, CSIRIS and ICT.



Request for TOdB Publication Number

Staff Details

Staff Number: 301772 Name: HEDI VAN DEVENTER □ Proxy

BU: 06000 Dept: 0NEEB Competence Area:

GroupWise Document Details

GW Library:

GWDMS Number:

Were view rights to assigned to CSIR TOdB on GroupWise DMS? ○ Yes ○ No

Proceed with TOdB indexing? ○ Yes ○ No

Indicate to whom view rights to the final document should be given. Please include RGL, CAM, SRM, Director, team members and Unit Manager.

Document Type

○ External Publications
○ Reports/Technical
○ Legal Documents
○ Miscellaneous
○ Thesis/Dissertation

Research Data

☑ Research Data (including geospatial) is linked to this publication

Request Date: 2011/03/15

Research Data

Research Data Details

Research Data Location (Building & office):

Research Data URL:

Security Classification:

Fig. 2: Example of linking research data to a report when indexing records.

procurement, human resources and other financial software systems. Up to 2008 no ICT support was provided on applications used for research purposes or on data sets produced as a result of research outputs.

The GSDI was the first project of its kind at the CSIR, which necessitated a work relationship between these two support sections and researchers in the operating units. Researchers required the curation of research artifacts associated with publications, and organisational ICT support for the management of hardware and software used for research outputs. A three-year project was embarked on to institutionalise a GSDI and become fully operational by 1 April 2011. ICT was tasked with the management and support of infrastructure to meet the needs of the researchers, while CSIRIS had to assist with records management needs. Meeting these requirements in the GSDI project as a start, has set the scene for wider implementation of records management and ICT support of research data and software within the CSIR.

ICT – implementing GSDI as an operational function in the CSIR

An 11 TB server for geospatial data was procured and installed by ICT for the GSDI following rigorous test specifications. All hardware related costs and maintenance will be managed by ICT in future, similar to the other infrastructure that already supports corporate systems for human resources, procurement and finances. Software licences are managed by system administrators and where issues occur, the server administrator obtains support directly from relevant vendors.

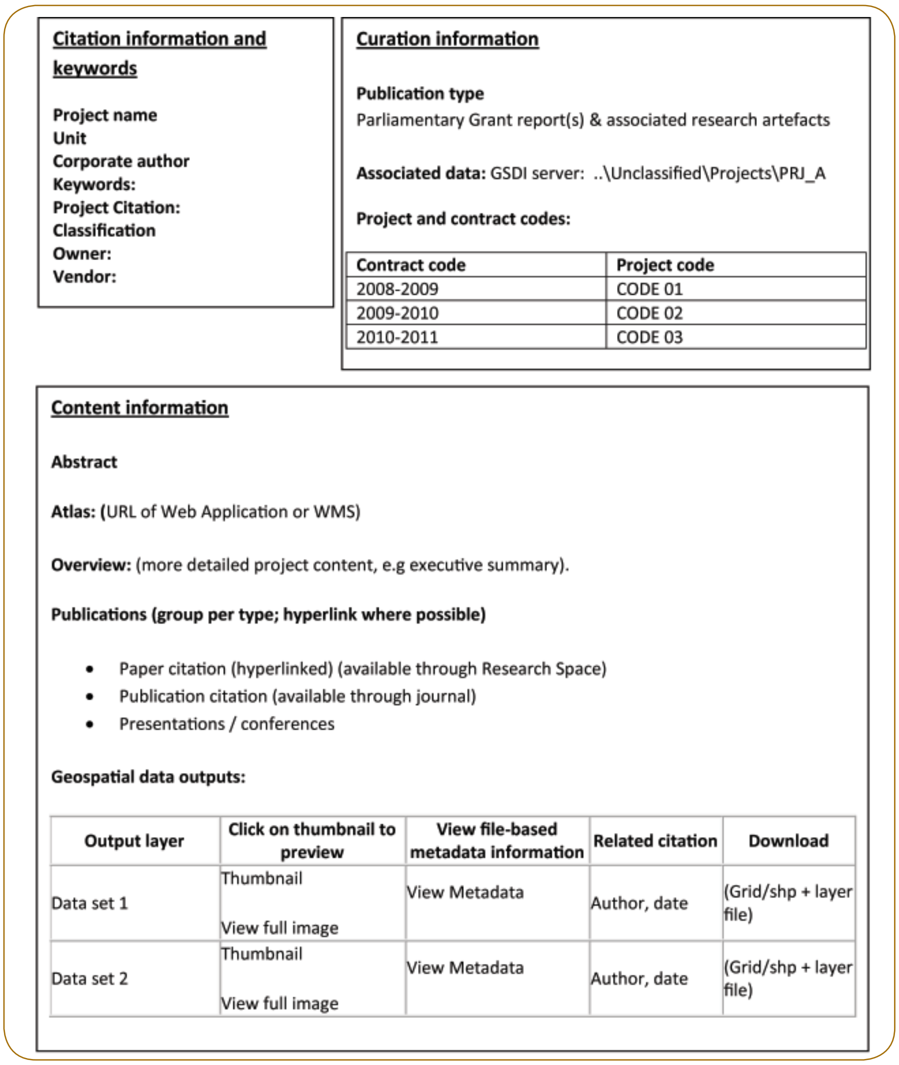
The server structure allows for a restricted section where folder permissions are assigned by the systems administrator as per user request for read, write and delete access of project participants (not accessible to all). In the unrestricted section, read access is allowed for all users, with less than eight people having loader access, to minimise uncontrolled upload of data and duplication. The "loaders" follow a structuring schema and file naming convention to upload data to the server. Folders in both sections are sorted per operating unit acronym and subfolder project acronyms. External data is sorted per acronym of external data custodian or owner, or in the case of satellite imagery, per sensor type, e.g. SPOT, MODIS, Landsat et cetera.

Concurrent software licence dongles are also hosted on a windows machine to enable rapid transfer of the dongle, should the machine crash. Software licences for Linux machines use the machine number and are less transferable in a crisis. The geospatial data are hosted on a Windows server with a file-based structure. This enables easy access to data from all GIS and RS software and faster back-ups. Back-ups are done according to CSIR policy: in a fire protected area with daily differential back-ups, monthly full back-up and off-site storage. On sub-folder level static and dynamic data is distinguished with differential back-ups of dynamic sub-folders done daily and full back-ups of dynamic data every week. Static data is backed-up in full once a month, or more regularly as per user request. Initially an Oracle database on a Linux machine was used to facilitate the ArcGIS Server software, though after thorough testing and securing PostGreSQL support, the plan is to migrate fully to PostGreSQL from 1 April 2011.

Operating units contribute pro ratio to the full cost of a new (concurrent) software licence or annual maintenance costs, whereas ICT facilitate the payments to the vendors. Software licence monitoring is done through freeware PhpLicensewatcher v. 1.4.2 (http://freshmeat.net/projects/phplicensewatcher/) per software type. Annual usage tracking per operating unit is however not sufficient and cannot indicate the total time period usage per operating unit per licence. This will enable a more accurate model indicating the financial contribution of each operating unit to the concurrent software licences.

Support has been implemented at the CSIR Call Centre for GSDI calls. Users have to use the keyword "GSDI" in the subject heading, when requesting support on installation instructions, software support, creation of folders in the restricted section, and server-related support.

A new geoportal was established http://gsdi.geoportal.csir.co.za which enables newsletters to CSIR users; links to external data repositories, institutions, open source software sites and atlases; links to software installation files, notes and vendor support links; support documentation and to Web Applications and Web Map Services (WA & WMS) for



Citation information and keywords

Project name
Unit
Corporate author
Keywords:
Project Citation:
Classification
Owner:
Vendor:

Curation information

Publication type
Parliamentary Grant report(s) & associated research artefacts

Associated data: GSDI server: ..\Unclassified\Projects\PRJ_A

Project and contract codes:

Contract code	Project code
2008-2009	CODE 01
2009-2010	CODE 02
2010-2011	CODE 03

Content information

Abstract

Atlas: (URL of Web Application or WMS)

Overview: (more detailed project content, e.g executive summary).

Publications (group per type; hyperlink where possible)

- Paper citation (hyperlinked) (available through Research Space)
- Publication citation (available through journal)
- Presentations / conferences

Geospatial data outputs:

Output layer	Click on thumbnail to preview	View file-based metadata information	Related citation	Download
Data set 1	Thumbnail View full image	View Metadata	Author, date	(Grid/shp + layer file)
Data set 2	Thumbnail View full image	View Metadata	Author, date	(Grid/shp + layer file)

Fig. 3: A proposed project-level metadata file.

CSIR-produced atlases. Further developments to the geoportal capabilities are done in partnership with the South African Earth Observation Network (SAEON). A metadata catalogue is being populated to be hosted on the geoportal and harvested to the National Spatial Information Framework (NSIF) automatically.

Implemented

- 11 TB GSDI server in Pretoria with back-ups complying to CSIR policy
- Software licence maintenance and support; monitoring software
- Call centre support for GSDI-related support requests
- A new geoportal with WA and WMS

Remaining challenges

- Software licence usage-tracking reporting per operating unit
- Metadata catalogue and harvesting to the NSIF
- Geoportal feature and capability expansion

CSIRIS – ensuring geospatial data curation

Researchers obtain credit in their annual performance and career ladder applications only for records reflected in ToDB. Up to 2010, these indexed records did not reflect any link to the geospatial data or any other research artifacts (e.g. cartographic or process diagrams, scripts, metadata or shapefiles) associated with the report or publication output. No link was retained between the research artifacts produced and records indexed in ToDB.

An information audit was done to assess the workflow between project registration and ToDB indexing to assess (i) whether a unique code was generated along the project work flow which can be used as a unique identifier for a project in its life span; and (ii) which artifacts are generated in a project and should be kept related to one another. A unique code could enable folder names and archiving, and potentially a document object identifier number (doi#) could retain location when moving folders between the active GSDI server and potential future archive server, as well as between contracts, reports and research artifacts. The study uncovered a huge amount of differences in project registration and work flow, not only between operating units, but also between external and internal projects. No unique code was assigned for the duration of a project: even project codes changed in some instances per financial year. Document object identifier numbers had to be generated at a cost outside the CSIR, and would not be an immediate solution to internal records management. It was therefore decided to add a rudimentary link when indexing publications and reports by allowing users to copy and paste a folder location related to the DMS publication (Fig. 2). Whether this will be sustainable for records management, remains to be seen.

Archiving of the former GIS unit's CD archive was done and procedures documented. Where possible an attempt was made to link the CD codes (per year) to potential ToDB reports or publications prior to 2004. These were sent out to some of the research staff who contributed to these projects. Additions or corrections to the list will be done as per requests. The CSIR's hardcopy map collection was also minimised. All hardcopy maps after the year 2000 and duplicates of those prior to 2000 were donated to Unisa for distribution. The use of remaining collection will be monitored and indexed as being used.

Archiving of remaining and newly generated geospatial data sets is still outstanding, and is an issue under discussion in the CSIR. Publicly published documents are sent to the Legal Depositories and the National Archives, though the associated research artifacts that are large in size, such as the geospatial data, cannot be managed by these institutions. The requirements for archiving are therefore under investigation: which data sets should be archived, at which point in time, where and which back-up procedures to follow.

Harvesting metadata from the GSDI server is under discussion. While the ArcGIS software enables file-based metadata through ArcCatalogue, the

Location of GSDI users	%
Pretoria	55,1
Stellenbosch	31,9
Johannesburg	2,9
Durban	5,8
Pietermaritzburg	2,9
Rosebank, Cape Town	1,4

Table 1: Estimated percentage of research personnel using geospatial data per location.

majority of the RS software does not enable metadata capturing within the software itself. Harvesting can be set up from the CSIR's Geoportal to collect metadata as a type of file (e.g., *.xml) or from a particular folder. The author has however, questioned the value of harvesting metadata at file-based level: many people would for example, clip orientation data layers such as roads, rivers and wetlands to smaller project study areas. Reflecting all of these in a metadata catalogue does not carry value for an external person, looking for national wetlands. One would rather refer requesters to the appropriate data custodian to obtain such a data set. It is therefore critical that only a key selection of metadata is harvested for the CSIR's Geoportal metadata catalogue, which will enable external requesters to view what the CSIR has produced.

The author suggested a project-level metadata template to be compiled and used (Fig. 3). At this level, all the related publications and geospatial data outputs (or research artifacts) are concatenated in one document, preferably through hyperlinks to relevant databases. This proposed document is being discussed at the CSIR to evaluate the value of this to all operating units, their respective research niches and how it can be potentially implemented.

It is envisaged that the project-level metadata will not only be useful for harvesting project metadata to the CSIR Geoportal, but will add value to records management of these projects internally to the CSIR, and perhaps other non-GSDI projects too.

Implemented

- A link between geospatial data and the documents indexed in ToDB

Remaining challenges

- Workflow procedures to improve records management and curation of research artifacts and particularly large data sets
- Archiving of large (geospatial) data sets
- Project-level metadata for harvesting

Branches – what to do?

A number of CSIR branches host users of geospatial data, including head office Pretoria, Stellenbosch, Johannesburg, Durban, Pietermaritzburg and Rosebank (Cape Town). During the planning phase of 2008-9 the percentage of users were recorded and a significant amount of users were present at both Pretoria and Stellenbosch (Table 1). At that stage a 14 Mbps connection between Pretoria and Stellenbosch would not have enabled direct linking to a single server, and no particular dates were set for the envisaged South African Research and Education Network (SANReN) upgrade between Pretoria and the branches.

Originally mirroring of data sets to the branches were considered, but this had huge cost implications. Not only would this have meant replication of a full 11 TB server at Stellenbosch and Durban, but software such as ArcGIS Server would require two additional licences to enable WA and WMS from these. Geoportal instances would have required replication at each site, and data management and back-ups would have required complex synchronisation and version management, which would have added to costs.

Fortunately during 2010 a 1 Gbps SANReN pipeline between Pretoria and Durban opened up, followed by a 100 Mbps SANReN pipeline, extended from Stellenbosch University to CSIR Stellenbosch, between Stellenbosch and Pretoria in November 2010. The GSDI immediately pursued tests to establish whether these links will enable transferring large data sets between the main GSDI server in Pretoria and the two branches. Tests proved that no additional implementation funding will be required for Stellenbosch, Durban or any other branch, and that all the pipelines are sufficient to carry the reading and upload of data from and to the main server.

The GSDI will also benefit from a planned network upgrade at all branches planned for the next year and a half, to be completed in 2012. The upgrade plans for at least 1 Gbps switches and CAT6 cables between all buildings. Tests were done at each building where GSDI users were located, to note the time it took to transfer < 1 GB, 1 GB, 2 GB and 5 GB data sets to and from the server. Switches and cables in the ICT building which hosts the servers, were already upgraded, and significant improvements in speed were noted when uploading large files (>= 5 GB) to the GSDI server. The test results will be considered to determine if 1 Gbps switches will suffice at the relevant buildings or whether better ones will be required.

Implemented

- Tests showed SANReN between Pretoria, Stellenbosch and Durban to be acceptable
- No need for additional servers/ software at branches

Remaining challenges

- Internal network upgrade requirements per building tested for all branches

Governing the GSDI and capacity building

The GSDI was implemented through an Implementation Plan and Operational Strategy (IPOS) and guided by a steering committee and a technical committee. Two strategic documents were created during the implementation phase of the GSDI: a Spatial Data Policy and Spatial Data Best Practices Procedures and Knowledge Management. Concepts from the Spatial Data Policy were fully incorporated as research data into the Records Management Policy, which was approved and implemented in the CSIR in 2010.

Responsibilities per contributing partner of the GSDI are listed in the Spatial Data Best Practices Procedures and Knowledge Management. The GSDI will be further governed by representatives from the researchers (Heidi van Deventer), ICT (Trudi van der Walt) and CSIRIS (Martie van Deventer). Relevant requirements from the IPOS will be translated to a service level agreement between all the relevant units.

A number of students benefited in capacity building of skills during the implementation phase of the project. These include:

- Ramapulana Nkoana (general)
- Dumisani Nkosi (hardcopy map indexing)
- Lindenhle Lushozi (CSIRIS-related work for the GSDI)
- Zoleka Milongo (ICT-related work for the GSDI)

Implemented

- GSDI project and governance structure; implementation plan and operational strategy
- Best practices procedures and knowledge management

Remaining challenges

- Service level agreement (SLA) between all parties

Remaining challenges within GSDI

Improvements and challenges are envisaged for the future of the GSDI, as listed per section above. An immediate benefit to other operating units, is the management of other concurrent software licences, used in research practice by more than one operating unit. The organisational wide interest in archiving and project-level metadata will be beneficial to both the GSDI and other units. Further development in software licence usage tracking and developments on the geoportal may benefit not only the CSIR, but other institutions in South Africa, who use geospatial data.

Acknowledgements

Acknowledgement is given to the CSIR personnel who contributed to the GSDI project Group and Operating Unit Executives, personnel from ICT and CSIRIS; and researchers from NRE, BE, MI and DPSS. We would also like to thank Wim Hugo from SAEON who contributed to our geoportal development.

Contact Heidi van Deventer, CSIR,
Tel 012 841-2507,
hvdeventer@csir.co.za