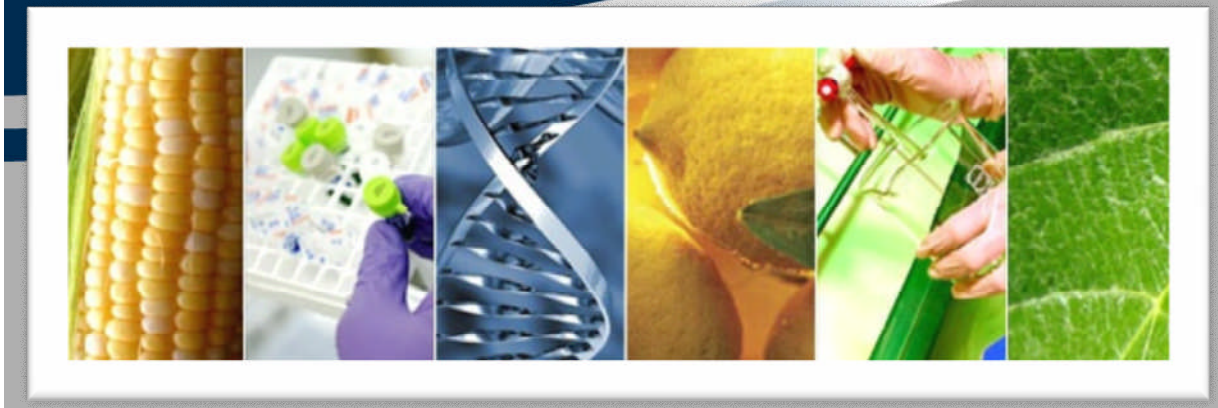
From theory to practice
Short phrase, long process!

Martie van Deventer

27 October 2010

22nd International CODATA Conference
Spier, Stellenbosch



our future through science

Purpose

- Personal experience in trying to establish a data management activity/ function within the CSIR information services.
- Perhaps assist those at the start of their institutional data management initiative.
- Unfortunately there is not a data management plug where all the necessary knowledge, skills and infrastructure could be downloaded – all in one go!!
- My paper does not imply that the CSIR does not manage data, nor does it mean that we do not contribute to national and international initiatives!

Roadmap

- The CSIR
- Background – theory of data management
- Rationale for records management approach
- 3 small but independent studies
  - Overview of our data management activity
  - Biosciences: challenges – as summarised by researchers themselves
  - Geospatial workflows
- Going forward

CSIR: A synopsis

People and demographics

- 2354 members of staff
- 1551 in SET base
- 457 with Masters'
- 286 with PhDs
- 53% of SET base black
- 33% of SET base female

33% income – public funding

our future through science

The CSIR mandate

directed multidisciplinary research

technological innovation

industrial and scientific development

quality of life

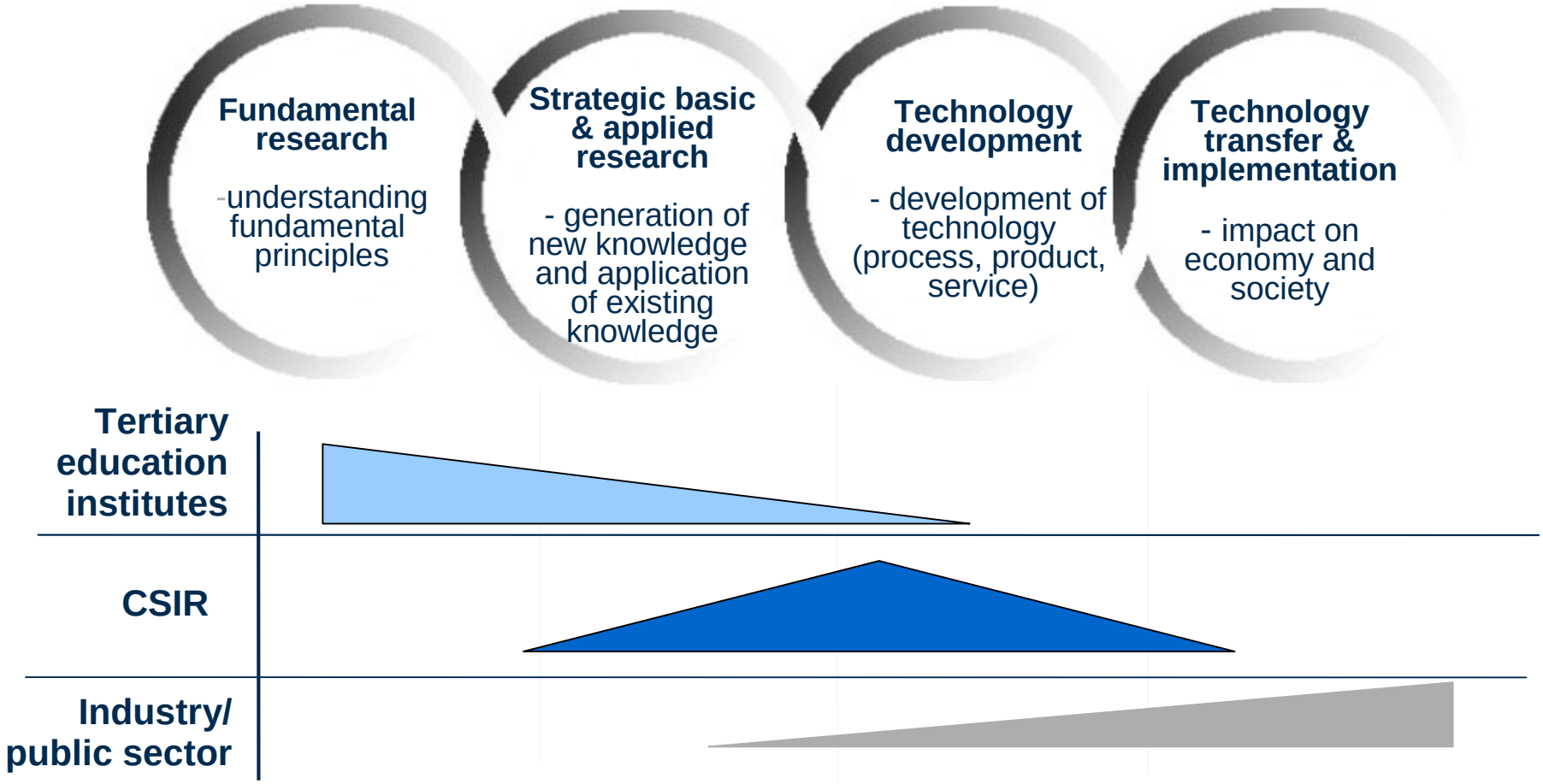
*'The objects of the CSIR are, through **directed** and particularly **multidisciplinary research** and **technological innovation**, to foster, in the national interest and in the fields which in its opinion should receive preference, **industrial and scientific development,** either **by itself or in co-operation with principals from private or public sectors**, and thereby to contribute to the **improvement of the quality of life** of the people of the Republic...'*

(Scientific Research Council Act 46 of 1988, amended by Act 71 of 1990)

Our line department is the Department of Science and Technology (DST)

our future through science

Strategic position in the National System of Innovation



The CSIR spans the research and innovation value chain but its role is differentiated from TEIs and industry/public sector

our future through science

CSIR areas of impact

Health
Nutrition & affordable novel treatments

Natural environment
Wise use & safe future

Energy
Alternative & renewable

Built environment
Transport & human settlements

Defence & security

Industry
Advanced manufacturing & mining

Applications

ICT

Materials

Research infrastructure

Technologies

CSIR
our future through science

After accepting the challenge of managing data at organisational level … the primary question is … where does one start?

Theory … the knowledge available

- Body of knowledge about data curation, data preservation and/or data management is comprehensive … but also daunting.
- The answer to every question you may have (especially at the start of your process) is freely available online.
- Considerable number of models to follow/ learn from.
- Each discipline has its own practices … so there is even more knowledge to tap into.

Why manage our research data?

- Data management is an essential component of the responsible conduct of research.
- This implies that, before starting a new research project, the researchers and or the research teams should address issues related to data management.
- By managing our data we will:
  - Increase our research efficiency.
  - Save time and resources in the long run.
  - Enhance data security and minimise the risk of data loss.
  - Prevent duplication of effort by enabling others to use our data.
  - Comply with practices conducted in industry and commerce.
  - Meet funding body grant requirements.
  - Enhance research integrity and replication.
  - Ensure that research data records are accurate, complete, authentic and reliable.
- Hence: increase our own credibility as reputable research organisation based at the tip of Africa.

based upon http://www.ed.ac.uk/is/data-management

Research data curation

We understand that there is more to this than

- Data archiving
- Data backups

Curation is about:

Selection … not everything;
Collection … some structure;
Preservation … formats;
Maintenance … active;
Archiving … safe;
Access … discovery;
Promotion … use;
… and back-ups.

Data Management includes …

- **Data Ownership** This pertains to who has the legal rights to the data and who retains the data after the project is completed.
- **Data Collection** This pertains to collecting project data in a consistent, systematic manner (i.e., reliability) and establishing an ongoing system for evaluating and recording changes to the project protocol (i.e., validity).
- **Data Retention** This refers to the length of time one needs to keep the project data according to the sponsor's or funder's guidelines. It also includes secure destruction of data.
- **Data Storage** This concerns the amount of data that should be stored -- enough so that project results can be reconstructed.
- **Data Protection** This relates to protecting written and electronic data from physical damage and protecting data integrity, including damage from tampering or theft.
- **Data Analysis** This pertains to how raw data are chosen, evaluated, and interpreted into meaningful and significant conclusions that other researchers and the public can understand and use.
- **Data Sharing** This concerns how project data and research results are disseminated to other researchers and the general public, and when data should not be shared.
- **Data Reporting** This pertains to the publication of conclusive findings, both positive and negative, after the project is completed.

(Steneck, 2004) http://ori.dhhs.gov/education/products/clinicaltools/data.pdf

Moving from theory to actual practice … getting past inertia … is complex for newcomers.

Records management provided us with a place to start!

Rationale for using records management to manage research data

- CSIR receives public funding and therefore the National Archives and Records Management Act is applicable.
- Research is our core function – our main body of records.
- Both the researcher need for managing the research record more efficiently and the national requirement are addressed simultaneously.
- Records management required that we understood where we were:
  - Gained overview information regarding the use of research data
  - Conducted data management practices research in Biosciences as the pilot group.
  - Investigated the workflows used by researchers responsible for generating geospatial data

Overview regarding data use and practice

- Short survey done in collaboration with the CHPC.
- Data collected at research unit level.
- Intent: High level first investigation into our data management activities.
- Questions related to:
  - Current data related challenges experienced
  - Current data storage size
  - Perception - how long data should be preserved/ kept?
  - Future plans to address data management needs?
  - Future data storage needs
  - What other forms of data support was required

Findings

- Much concern was expressed regarding
  - The obsolescence of media (tapes, stiffies, CDs, flash drives)
  - Synchronisation of datasets
  - Duplication of datasets
  - Financial burden associated with curation of data
- Data needs to be stored between 5 years and ‘for ever’.
- Quantitative data is often supplemented by video, photographic and documented information. Data cannot be curated without the ‘context giving’ supplementary material.

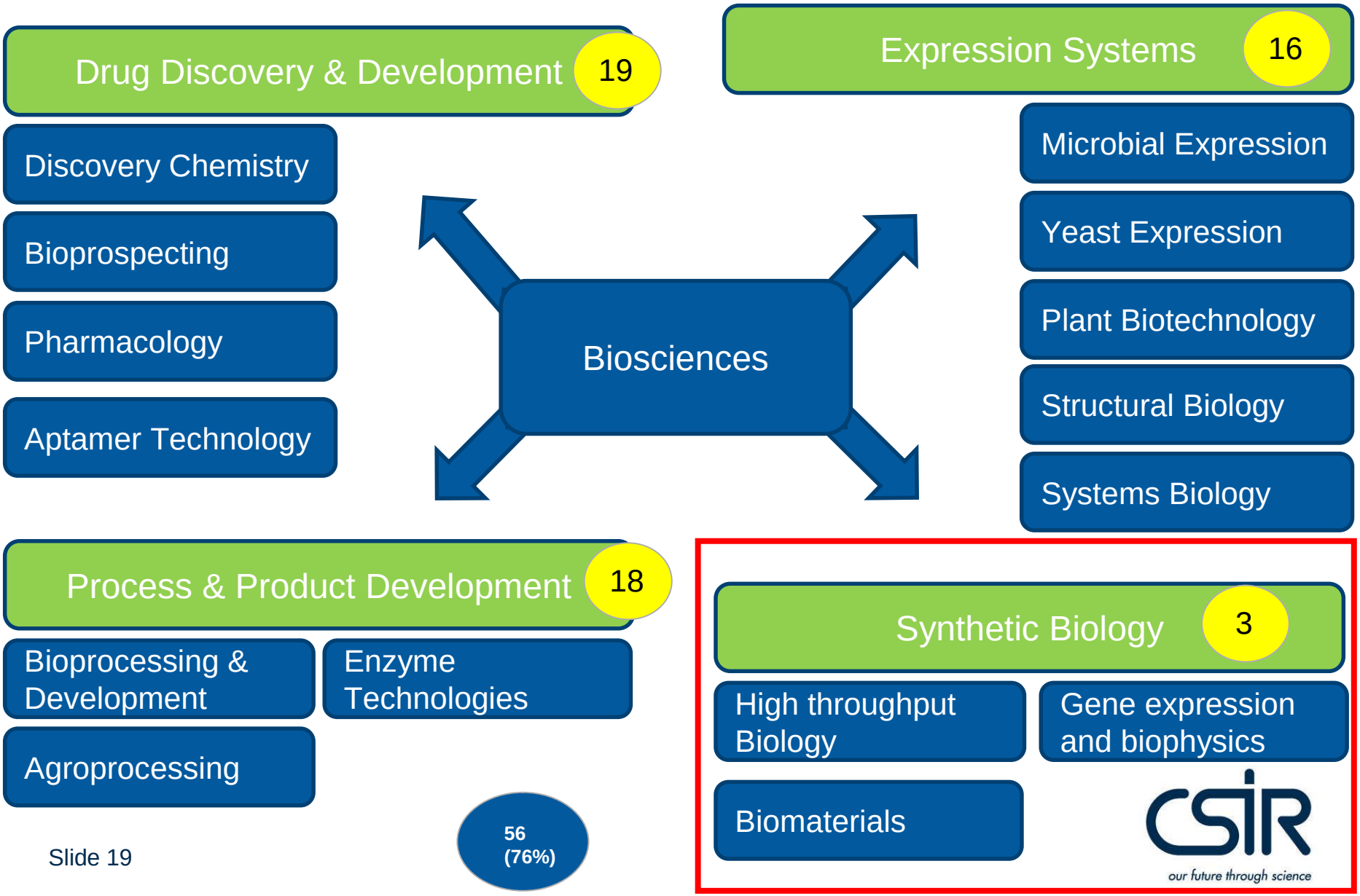
Findings 2

- Managed storage of data is a huge need.
- Similarly data discovery was reported to need attention.
- Usual infrastructure requirements regarding sharing and transfer of data sets – especially then the transfer of large data sets.

Zooming in – investigating one Unit

- A structured interview was developed making use of existing studies at the universities of Oxford, Purdue and Pretoria.
- All relevant research staff were identified by Biosciences management.
- All staff were not available but fifty-six (76%) interviews were conducted over a period of eight days.
- Naïve interviews were conducted by 10 students in their final year of studies (after receiving appropriate training).
- It was understood that results have to be verified and additional information would need to be collected where necessary.

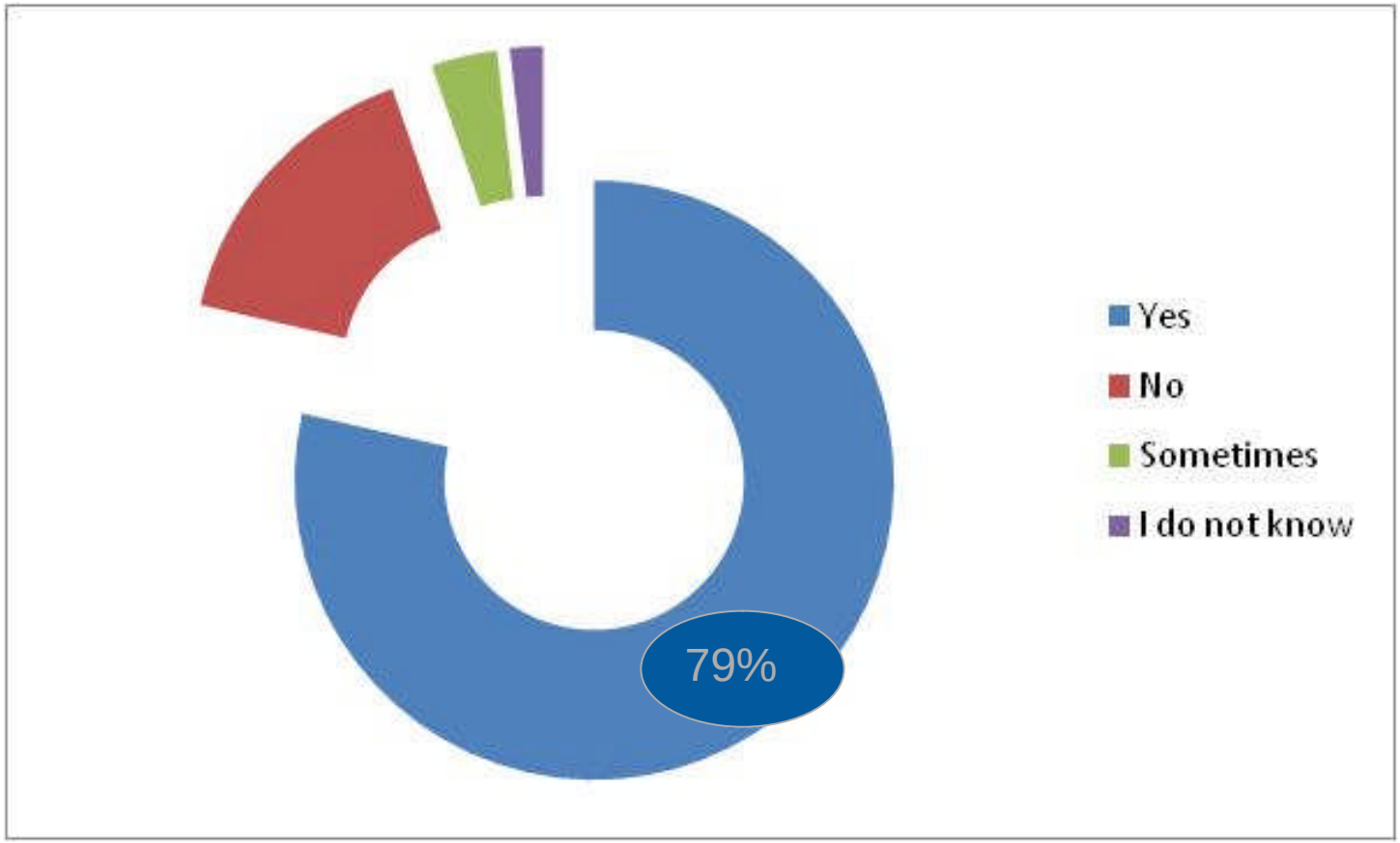
Biosciences: areas investigated



Data management planning

- 68% did no formal data management planning.
- Almost 20% acknowledged that research funders are requiring formal data management plans.
- Approximately 9% felt they knew what they were doing – they do not need to submit formal management plans.

Is the research data sensitive?



Sensitivity: is it because of IP? … matter needs further investigation.

Challenges for researchers (1)

- Data management
  - Duplication of data sets
  - Transferring large data sets
  - Data volumes & formats
  - Remote monitoring of lab
- Access
  - When staff leave or new staff join
  - Paper is inaccessible
  - Distributed DMS libraries
- Retrieval
  - No system available to assist
- Storage
  - Insufficient
  - No guidelines
  - No centralised managed storage
- Security
  - Confidentiality
  - Data is not tamper proof
  - Manual back-ups are slow

Challenges for researchers (2)

- Archiving
  - Inefficient/ Is not done
  - Curation not done
- Skills
  - Lack knowledge to properly manage data
  - Skills to do statistical analysis of data
  - Dedicated resources to manage data
- Software
  - Too few licences
  - Simulation/ analysis tools lacking
- Hardware
  - Outdated/ slow
  - System crashes = loss of data
- Funding
  - Finding projects with commercial value that will allow for curation costs to be added
- Electronic lab book
  - Seen as route to data access
  - Efficiency improvement

Requests expressed by the research staff

- Standardise the process of controlling stored data.
- Connect all instruments to the internal infrastructure - to facilitate the easy storage of data.
- Do not add more administration to our workload!
- Make old data (paper, Afrikaans) accessible.
- Empower us - give us the necessary skills/ training.
- Although duplication is a serious problem ensured **ACCESS** to existing data is the real issue!
- “I manage my own data and would not let anyone else near it!”

Six researchers experience no challenges!

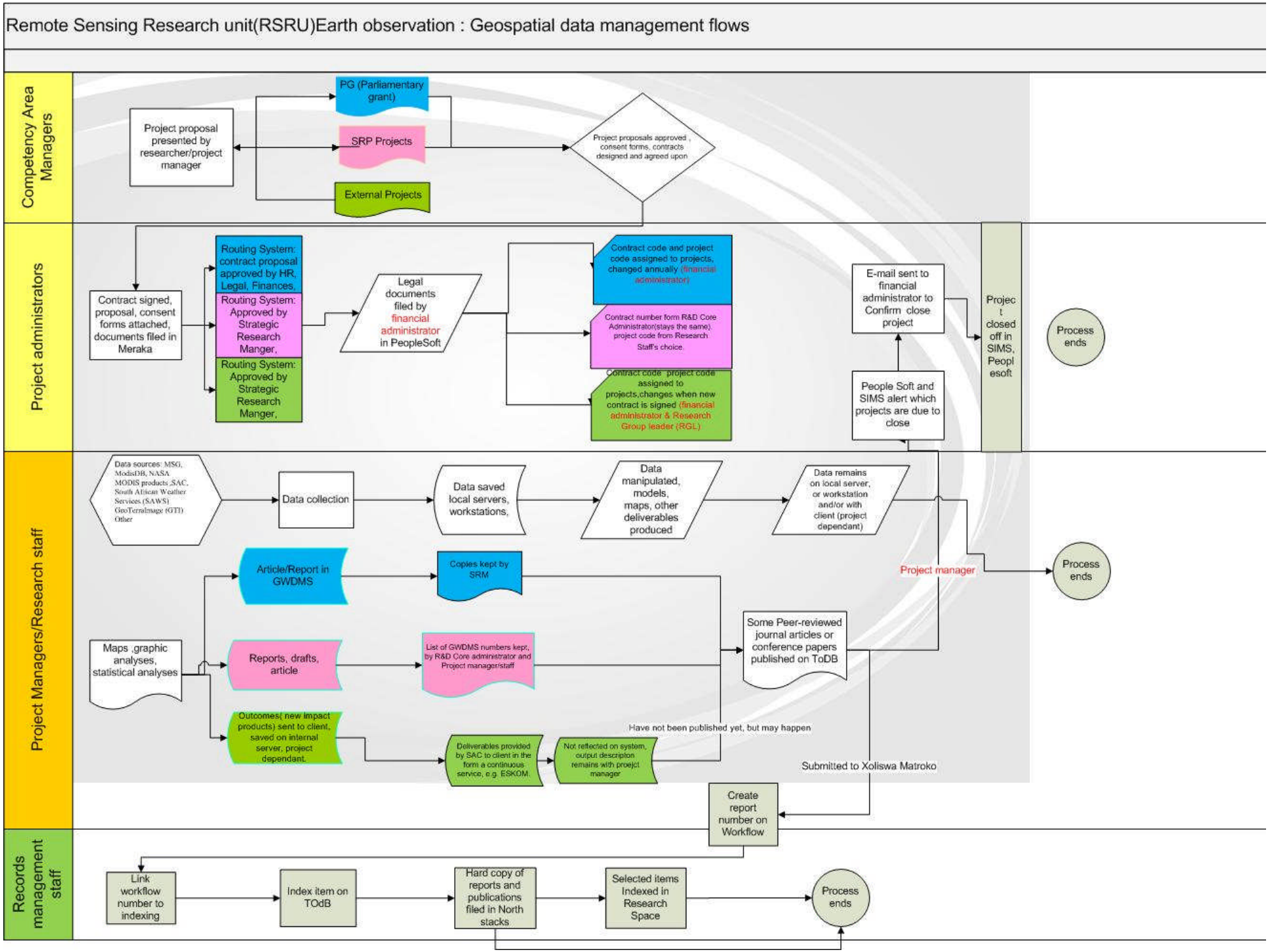
… which is a challenge in itself!

Deduction

- The research staff within the unit already know what the basic activities are that are necessary – makes the change management process easier. However … this does not mean that change management can be ignored!

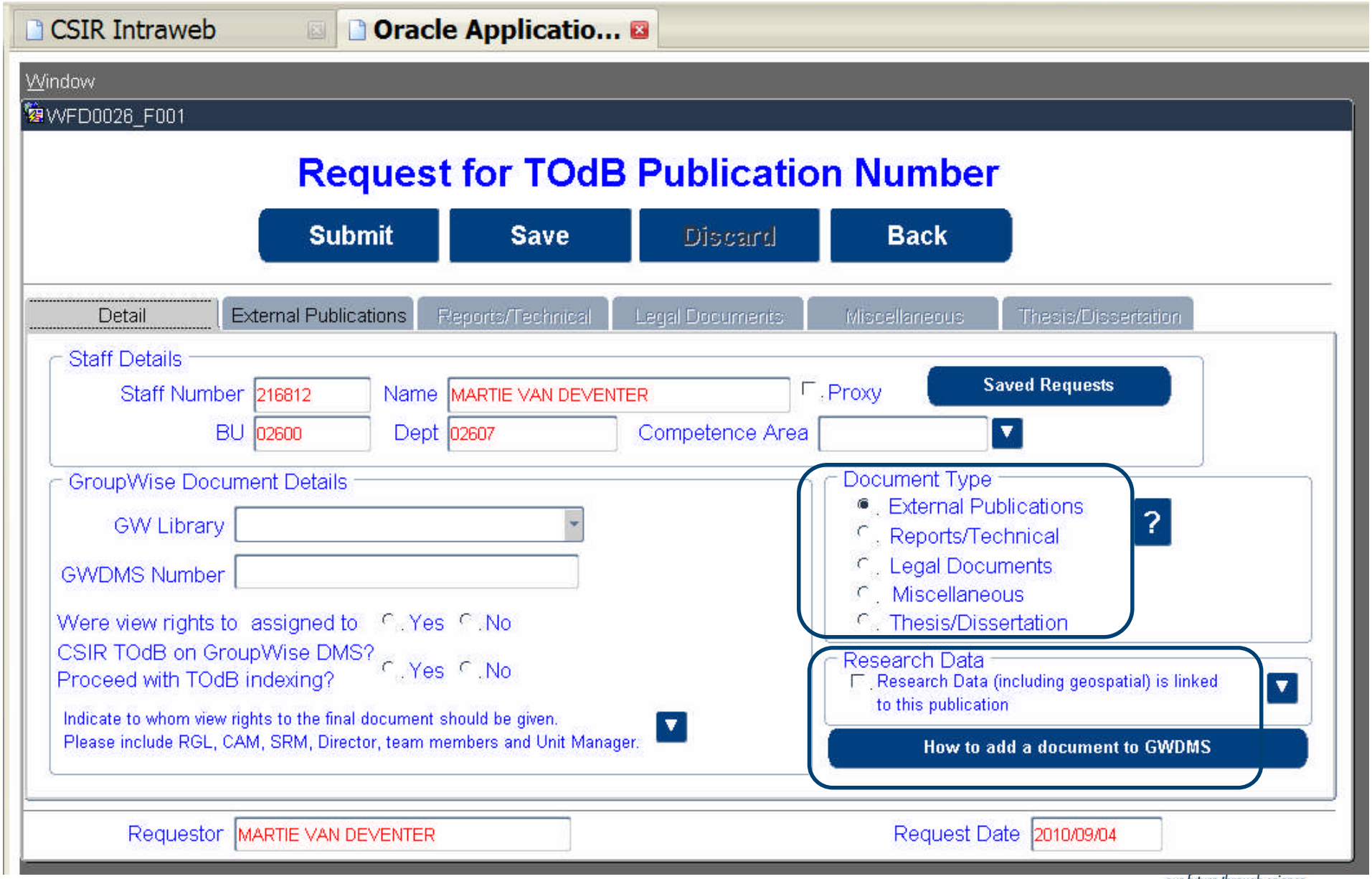
Geospatial data workflows

- Methodology: personal interviews with selected/ identified candidates across three research units.
- Followed by focus group discussion & e-Mail feedback from research staff unable to attend.
- Purpose: to establish if the discipline follows similar procedures (workflows).


Remote Sensing Research unit(RSRU)Earth observation : Geospatial data management flows
Competency Area Managers
Project administrators
Project Managers/Research staff
Records management staff
Project proposal presented by researcher/project manager
PG (Parliamentary grant)
SRP Projects
External Projects
Project proposals approved , consent forms, contracts designed and agreed upon
Contract signed, proposal, consent forms attached, documents filed in Meraka
Routing System: contract proposal approved by HR, Legal, Finances,
Routing System: Approved by Strategic Research Manger,
Routing System: Approved by Strategic Research Manger,
Legal documents filed by financial administrator in PeopleSoft
Contract code and project code assigned to projects, changed annually (financial administrator)
Contract number form R&D Core Administrator(stays the same) project code from Research Staff's choice.
Contract code project code assigned to projects,changes when new contract is signed (financial administrator & Research Group leader (RGL)
E-mail sent to financial administrator to Confirm close project
People Soft and SIMS alert which projects are due to close
Project closed off in SIMS, Peoplesoft
Process ends
Data sources: MSG, ModisDB, NASA MODIS products ,SAC, South African Weather Services (SAWS) GeoTerraImage (GTI) Other
Data collection
Data saved local servers, workstations,
Data manipulated, models, maps, other deliverables produced
Data remains on local server, or workstation and/or with client (project dependant)
Project manager
Process ends
Article/Report in GWDMS
Copies kept by SRM
Maps ,graphic analyses, statistical analyses
Reports, drafts, article
List of GWDMS numbers kept, by R&D Core administrator and Project manager/staff
Some Peer-reviewed journal articles or conference papers published on ToDB
Outcomes(new impact products) sent to client, saved on internal server, project dependant.
Deliverables provided by SAC to client in the form a continuous service, e.g. ESKOM.
Not reflected on system, output description remains with proejct manager
Have not been published yet, but may happen
Submitted to Xoliswa Matroko
Create report number on Workflow
Link workflow number to indexing
Index item on TOdB
Hard copy of reports and publications filed in North stacks
Selected items Indexed in Research Space
Process ends

Findings

- Confidentiality of data is not the biggest issue.
- Researchers are using the same infrastructure but processes differ.
- Input data varies considerably – questions regarding the curation of input data sets.
- The community is active in formulating its own data management procedures.
- We are actively contributing towards transferring data from old storage formats onto dedicated storage server.


CSIR Intraweb
Oracle Applicatio...
Window
WFD0026_F001
Request for TOdB Publication Number
Submit
Save
Discard
Back
Detail
External Publications
Reports/Technical
Legal Documents
Miscellaneous
Thesis/Dissertation
Staff Details
Staff Number 216812
Name MARTIE VAN DEVENTER
Proxy
Saved Requests
BU 02600
Dept 02607
Competence Area
GroupWise Document Details
GW Library
GWDMS Number
Were view rights to assigned to Yes No
CSIR TOdB on GroupWise DMS?
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
How to add a document to GWDMS
Requestor MARTIE VAN DEVENTER
Request Date 2010/09/04

What next … getting to practice?

- Collaboration between ourselves and our ICT to:
  - In terms of storage - distinguish between active and ‘record’ data
  - Make available a repository for recording the existence of ‘record’ data – with at least Dublin Core metadata
  - Standardise data storage for ‘record’ data
  - Follow open standard formats where possible
  - Limit the output formats for ‘record’ data as far as is possible
- Data management plans to be submitted as part of the project planning record - as of April 2011.
- Initiated a process to develop appropriate skills linked to discipline specific data management requirements.

Getting from theory to practice is a long haul but … it is the inertia (fear of the unknown) which is the difficult hurdle.

The practice is interesting and rewarding!

My advice: Go out and start! Give it a shot! … you'll discover it is not brain surgery and it is not rocket science either!

Acknowledgement

- Findings of the scoping study interviews and the research data management workshop. Scoping digital repository services for research data management. A Project of the Office of the Director of IT www.ict.ox.ac.uk/odit/projects/digitalrepository/" by Luis Martinez-Uribe (luis.martinez-uribe@oerc.ox.ac.uk), Digital Repositories Research Co-ordinator, University of Oxford, UK
- Pienaar, H. 2010. Survey of research data management practices at UP: October 2009 – March 2010 (unpublished)
- Steneck, 2004. http://ori.dhhs.gov/education/products/clinicaltools/data.pdf
- DCC Lifecycle model. Available: http://www.dcc.ac.uk/resources/curation-lifecycle-model
- Data Curation and Libraries: Short-Term Developments, Long-Term Prospects http://digitalcommons.calpoly.edu/cgi/viewcontent.cgi?article=1027&context=lib_dean
- A new role for academic librarians: data curation http://www.era.lib.ed.ac.uk/handle/1842/3207