eResearch Africa

23-27 November 2014

Meeting the needs of the 21st century researcher

www.eresearch.ac.za | conference@eresearch.ac.za

Data Intensive Research Initiative for South Africa (DIRISA)

A Reinterpreted Vision

A. Vahed

25 November 2014

science
& technology
Department:
Science and Technology
REPUBLIC OF SOUTH AFRICA

CSIR
our future through science

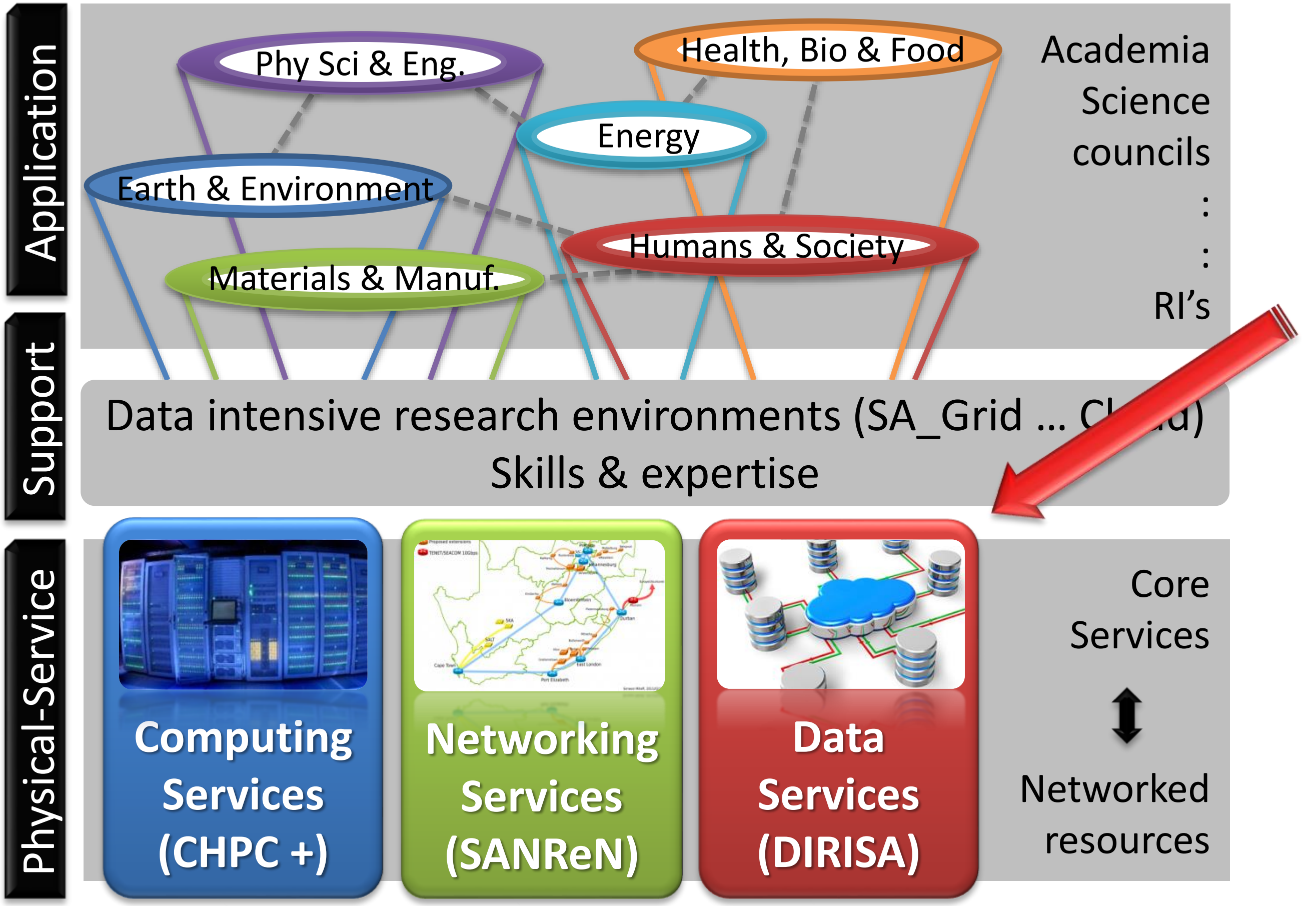
Outline

- Background
- Data Landscape
- Strategy & Objectives
- Activities & Outputs
- Organisational Structure & Implementation

NICIS

NICIS

- National data integrative enabler supporting
  - MTSF
  - RDP
  - SARIR,…
- Overarching coordination & national strategy
  - National (Tier1)
  - Institutional (Tier2)
- Amalgamated, physically distributed cyber platform for data intensive research
  - Data
  - Networking
  - Computing
  - Crosscut
  - S&T



Application

Support

Physical-Service

Phy Sci & Eng.

Health, Bio & Food

Energy

Earth & Environment

Humans & Society

Materials & Manuf.

Academia Science councils : : RI’s

Data intensive research environments (SA_Grid … Cloud)
Skills & expertise

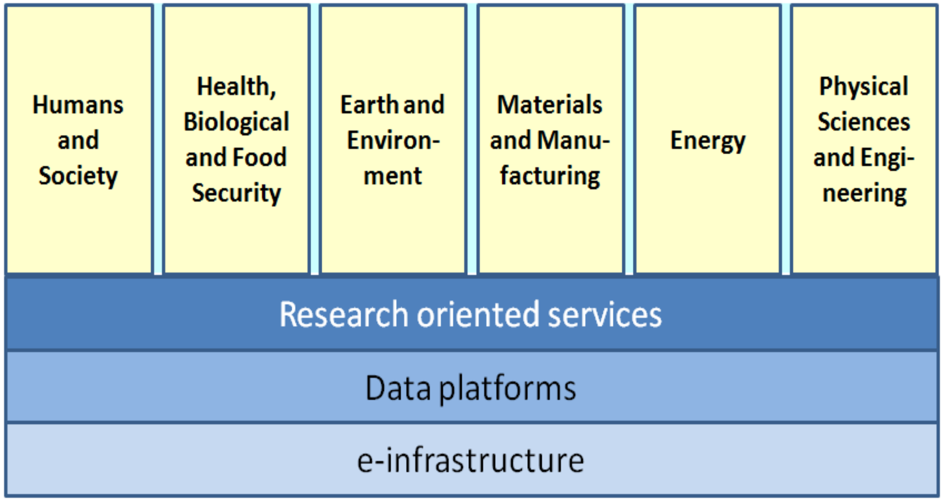
Computing Services (CHPC +)

Networking Services (SANReN)

Data Services (DIRISA)

Core Services ⬍ Networked resources

Other views


Humans and Society

Health, Biological and Food Security

Earth and Environment

Materials and Manufacturing

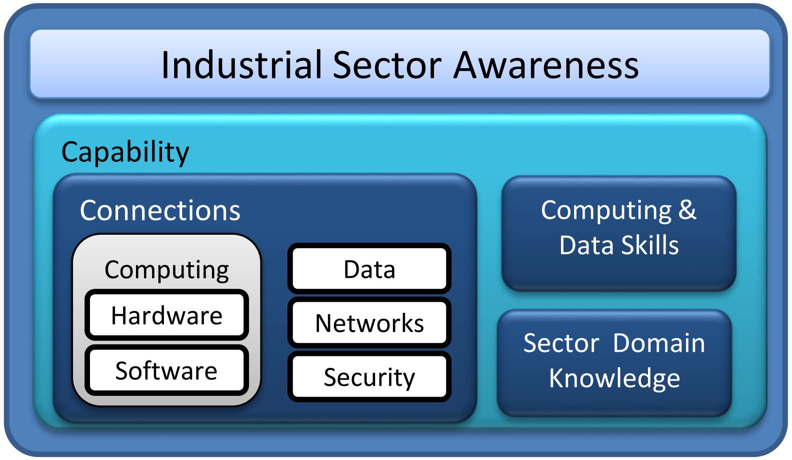
Energy

Physical Sciences and Engineering

Research oriented services

Data platforms

e-infrastructure


Industrial Sector Awareness

Capability

Connections

Computing

Hardware

Software

Data

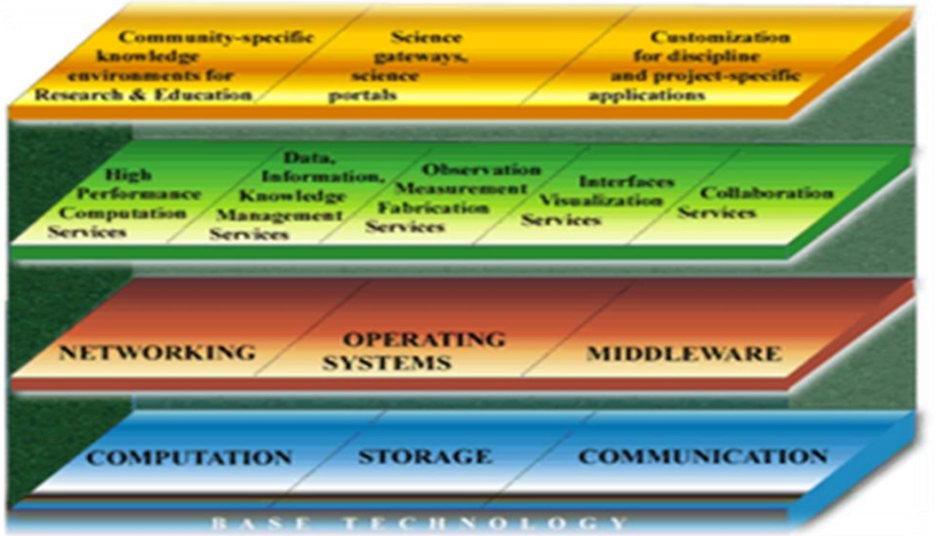
Networks

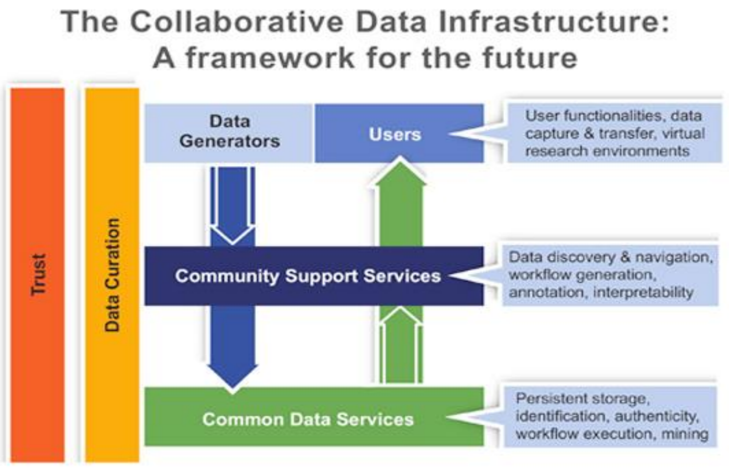
Security

Computing & Data Skills

Sector Domain Knowledge

D. Tildesley: Vision of integrated e-infrastructure ecosystem

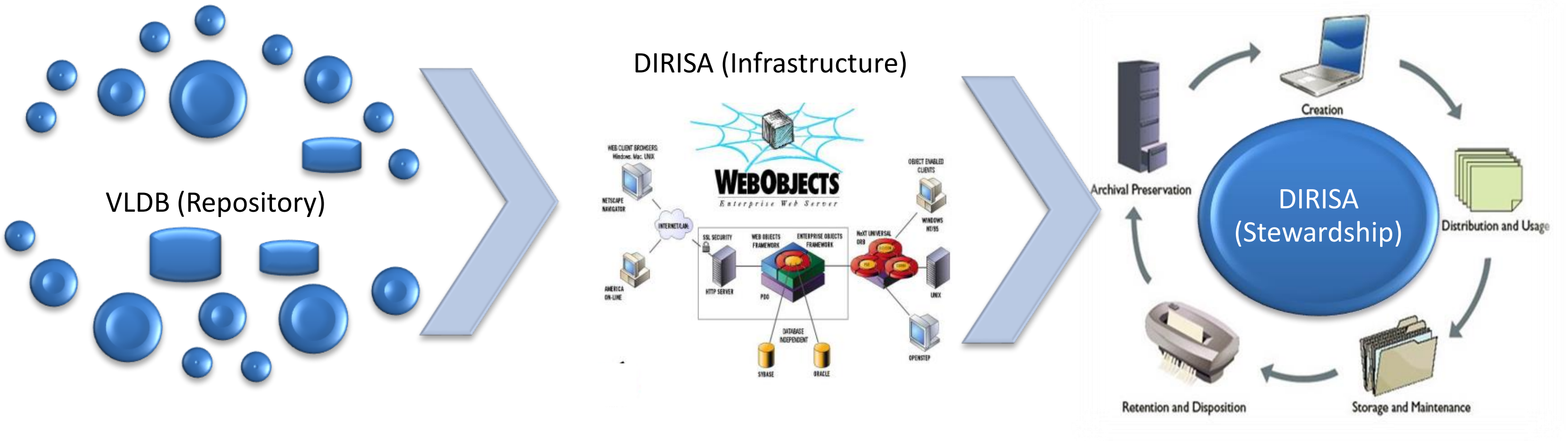

Community-specific knowledge environments for Research & Education
Science gateways, science portals
Customization for discipline and project-specific applications
High Performance Computation Services
Data, Information, Knowledge Management Services
Observation Measurement Fabrication Services
Interfaces Visualization Services
Collaboration Services
NETWORKING
OPERATING SYSTEMS
MIDDLEWARE
COMPUTATION
STORAGE
COMMUNICATION
BASE TECHNOLOGY


The Collaborative Data Infrastructure: A framework for the future

Trust
Data Curation
Data Generators
Users
User functionalities, data capture & transfer, virtual research environments
Community Support Services
Data discovery & navigation, workflow generation, annotation, interpretability
Common Data Services
Persistent storage, identification, authenticity, workflow execution, mining

Source: High Level Expert Group on Scientific Data, Riding the wave, 2010.

NICIS: DIRISA evolution

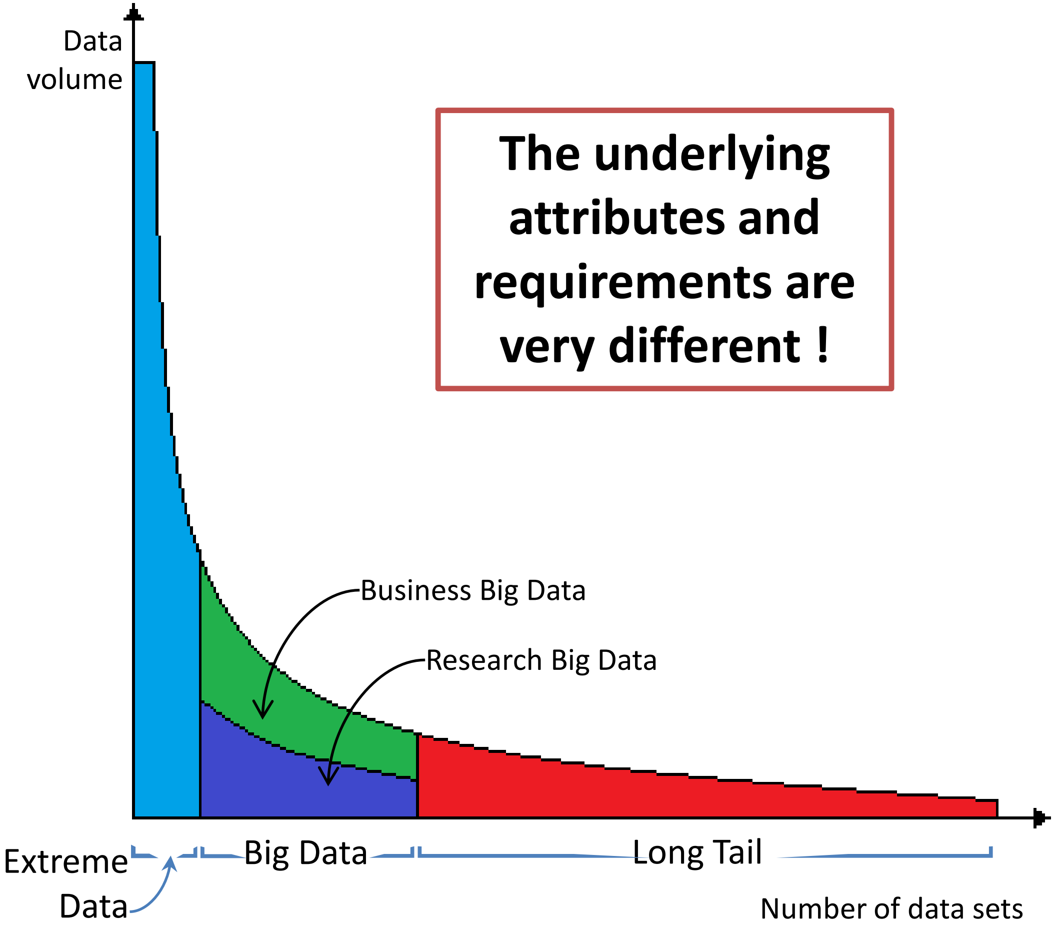


NICIS

- Recommendations
  - *"… expanded Data Services (DIRISA)…"*
  - *"… ambitious proposal on data services […] predicated on economic competitiveness, human resource development and industrial benefit"*
- Innovation for socio-economic development & knowledge economy
  - Step change: NDP, MTSF, NSI,…

Data landscape

- **Extreme Data**
  - Global, massive, well-typed, homogeneous volumes
  - LHC & SKA
- **Research Big Data**
  - Large, mixed-typed volumes
  - Imagery, text, audio, etc
- **Business Big Data**
  - Lots of (closed) transactional, serialised data
  - Sentiment data (Facebook, Twitter, etc)
- **Long Tail Data**
  - Lots of (poorly managed) relatively small data sets



The underlying attributes and requirements are very different !

Data class characteristics

Class	Ownership	Big Data Vs	Technology	Skills	Research Env
Extreme	International	Vol, Vel, Open	Exascale	Comp Maths / Stats / Astro, Visual	Distributed teams
Big Data – Business	Businesses	Vol, Vel, Var, Closed	Clusters, SAS, Cloud, Hadoop	Data Engineers	Team
Big Data – Research	National, Institutional	Vol, Vel, Var, Ver, “Open” access	HPC, Clusters, Grid, Cloud, data transfer	Data Scientists, Domain Researchers, Comp Scientists, Maths, Model	VRE, multi-disc, RIs
Long Tail	Department, Individual	Var, Ver	Grid, cloud	Stats, Comp Science	Individuals, PhD, PD, Ris

Vision & Strategy

Vision: Vibrant communities of research and industry

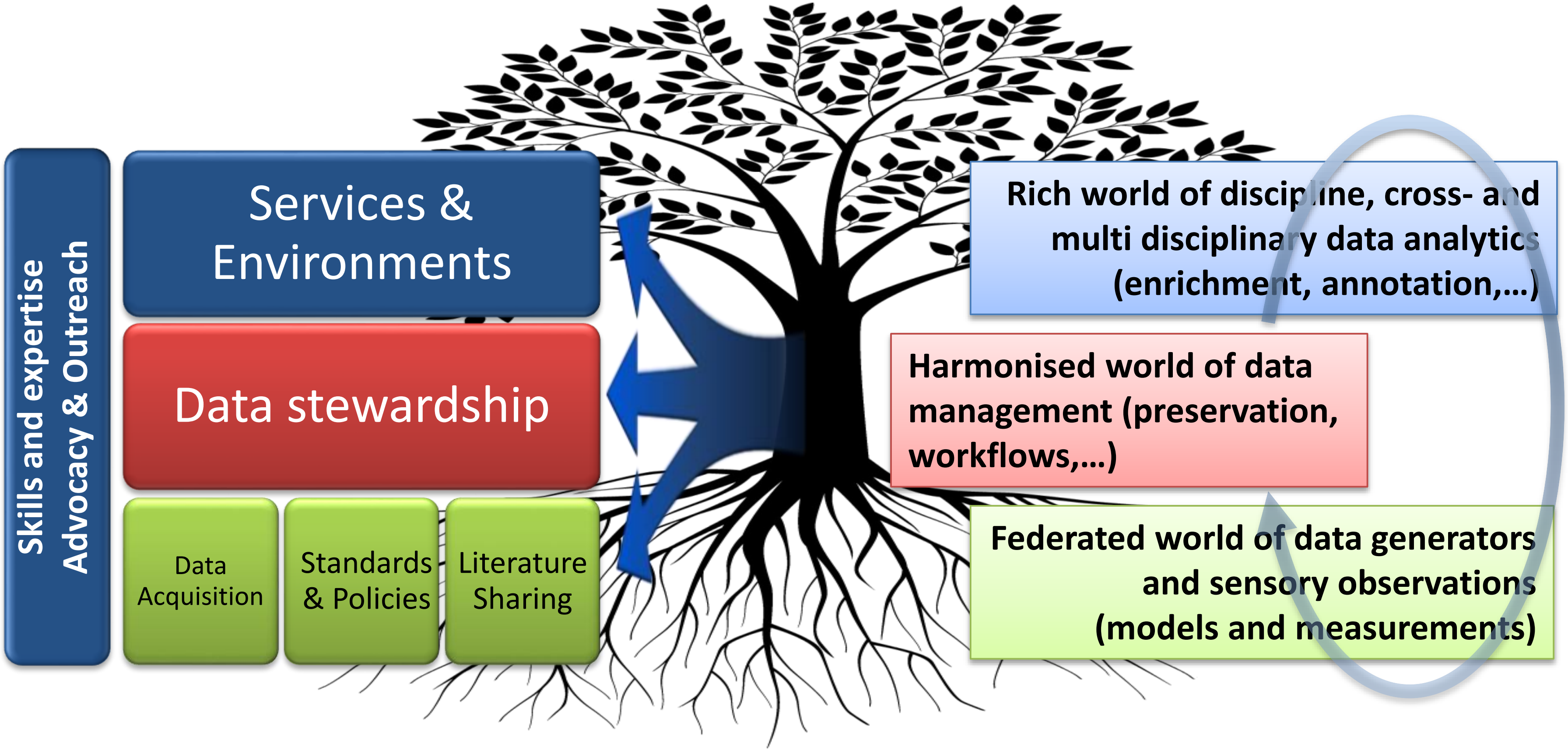
- access, share, reuse, combine data in a cohesive network of data repositories,
  - governed by sound data stewardship policies and principles,
  - supported by robust services and environments,
  - managed by expert and skilled people, and
- produce data intensive research output that support innovation for socio-economic growth and improved service delivery

Strategic principles

- Provide national capstone coordination
  - Data intensive research initiatives; Stakeholder engagement
- Promote & support data intensive research
  - Higher education; PPPs
- Data stewardship (more than DAAS)
  - Robust infrastructure & enabling environments; E2E research data lifecycle
- Strategy & leadership
  - Priority domains; Cross-cutting (inter- and multi-disciplinary)

our future through science

Value proposition



Skills and expertise

Advocacy & Outreach

Services & Environments

Data stewardship

Data Acquisition

Standards & Policies

Literature Sharing

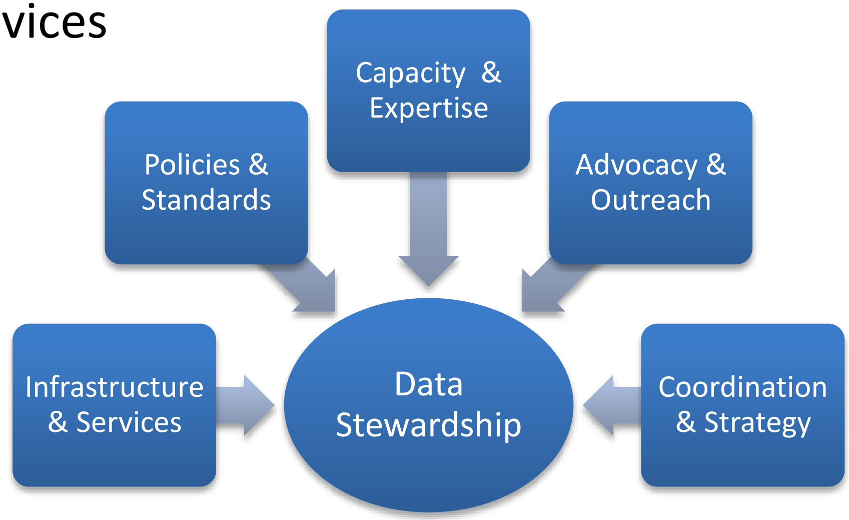
Rich world of discipline, cross- and multi disciplinary data analytics (enrichment, annotation,…)

Harmonised world of data management (preservation, workflows,…)

Federated world of data generators and sensory observations (models and measurements)

Key Objectives

1. Provide robust infrastructure and services
   - Federate Tier 1 & Tier 2 repositories
   - Enabling environments
   - Journal licencing
2. Ensure good data stewardship
   - Policies, protocols & standards
   - Internationally benchmarked
3. Develop capacity & expertise
   - Data intensive research and
   - Data science programmes with HEIs & private sector
4. Advocacy & outreach
   - Data stewardship and data sharing
   - Stakeholder engagement – establish and leverage existing forums
5. Coordination & strategy
   - National data intensive research activities
   - Inform on and guide aligned & consolidated strategic agenda



Scope

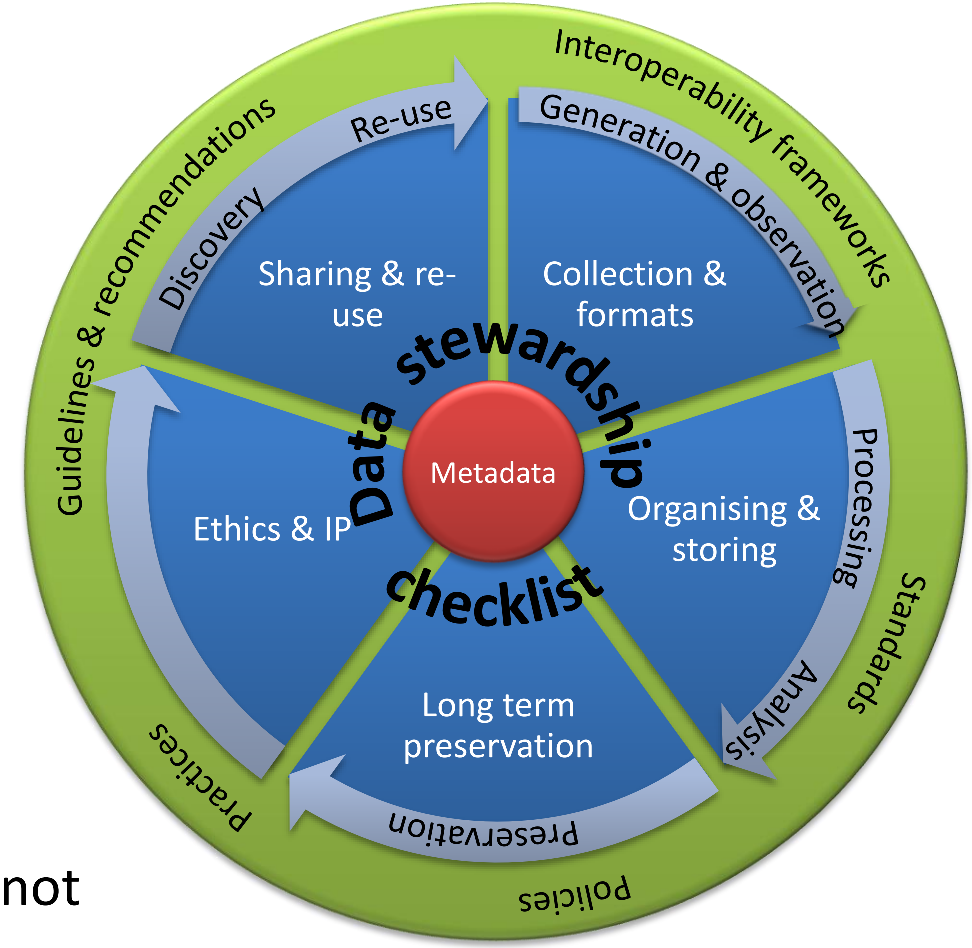
Primarily
a national capstone orchestration, enabling,
supporting and facilitation role

- Coordinate, not prescribe, data science capacity development
- Funded capacity development limited to DIRISA's remit
- Promote & support priority research
  BUT with caveats of data stewardship plan and capacity building
- Guide research strategy and funding (Big Data, NRF,...)
- Provide services and research environments
  BUT not a domain research funder
- Promote , not enforce, data contribution and adoption of Open standards & Open data where feasible
- Support data stewardship in federated context

CSIR
our future through science

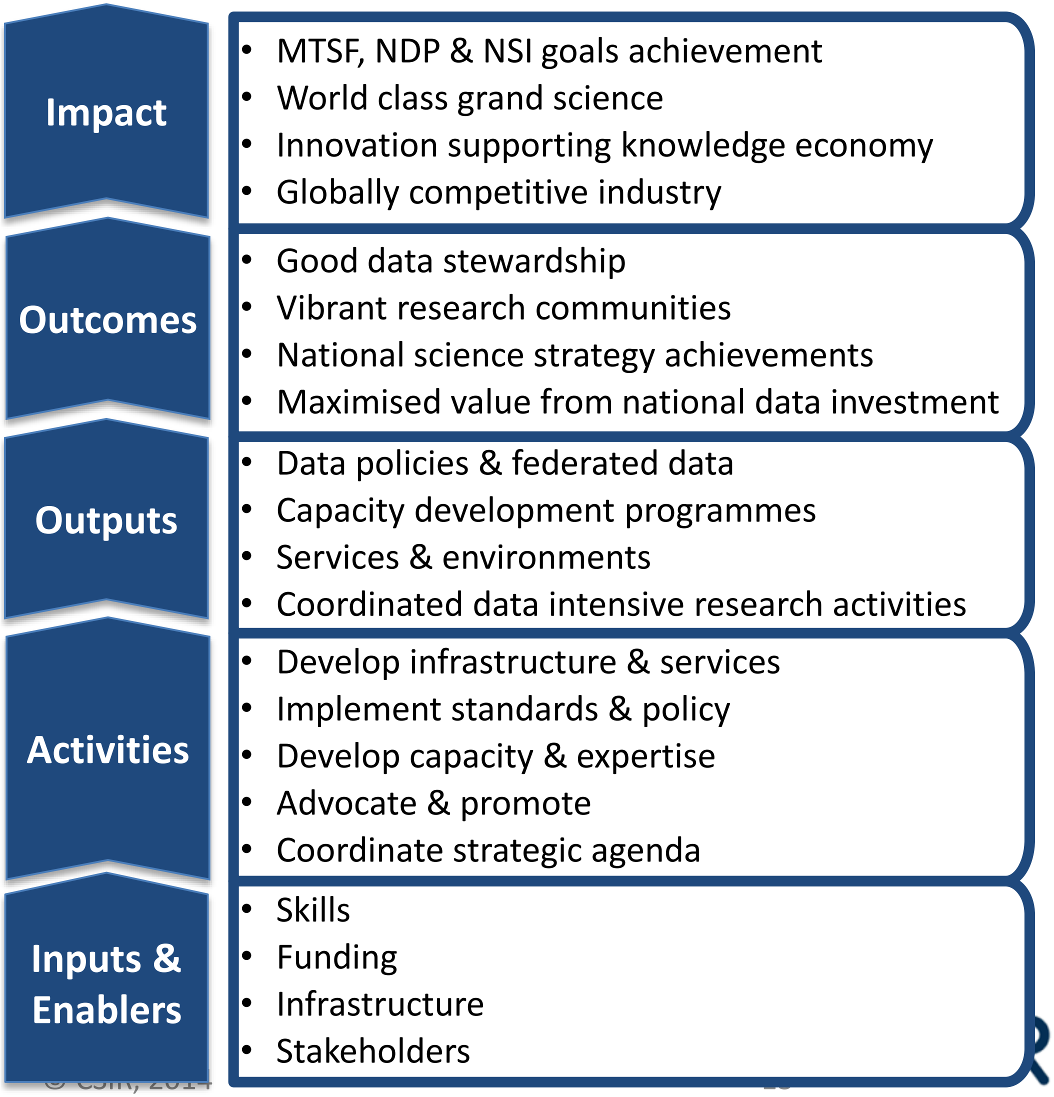
Issues

- Research data lifecycle
  - Observation / Generation
  - :
  - Preservation/ Expunction
- Ethics & privacy
  - Re-identification
  - Discriminative profiling
  - Who watches the watchers?
- Access spectrum
  - Trust & security
  - IP & Copyright
- Data sharing mind-set (What's in it for me?)
- Laws have borders; data does not



Stakeholder Engagement

- Stakeholders
  - RIs & Champions
  - Academia & research councils
  - Funders
  - Industry
  - International forums
  - …
- Engagement is critical
  - Strategic research agenda
  - Data stewardship policy & frameworks
  - Coordination of initiatives
  - Contribution & participation



Impact

- MTSF, NDP & NSI goals achievement
- World class grand science
- Innovation supporting knowledge economy
- Globally competitive industry

Outcomes

- Good data stewardship
- Vibrant research communities
- National science strategy achievements
- Maximised value from national data investment

Outputs

- Data policies & federated data
- Capacity development programmes
- Services & environments
- Coordinated data intensive research activities

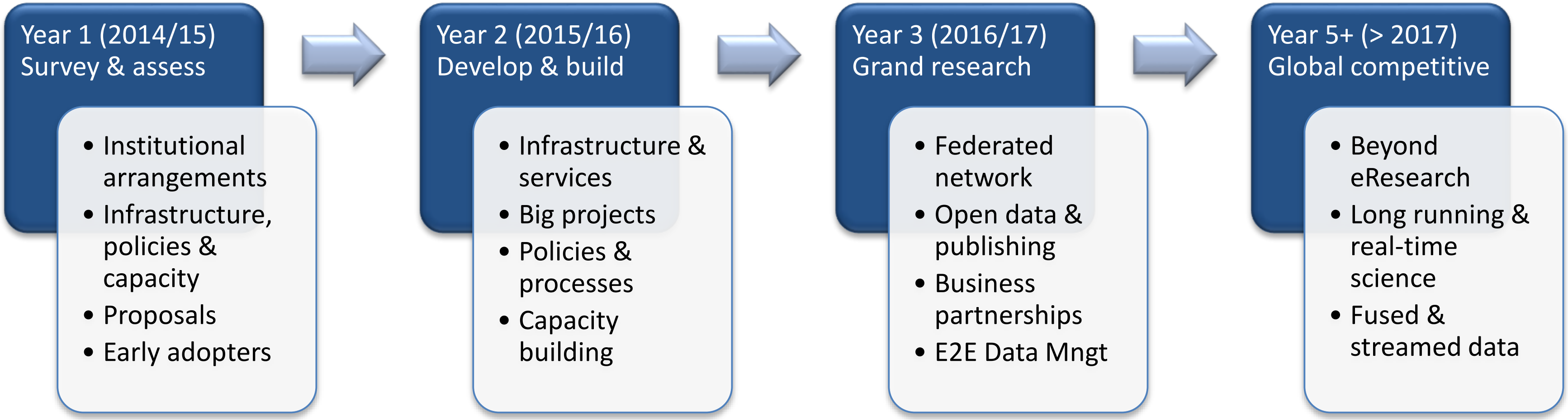
Activities

- Develop infrastructure & services
- Implement standards & policy
- Develop capacity & expertise
- Advocate & promote
- Coordinate strategic agenda

Inputs & Enablers

- Skills
- Funding
- Infrastructure
- Stakeholders

DIRISA Roadmap



Year 1 (2014/15)
Survey & assess

- Institutional arrangements
- Infrastructure, policies & capacity
- Proposals
- Early adopters

Year 2 (2015/16)
Develop & build

- Infrastructure & services
- Big projects
- Policies & processes
- Capacity building

Year 3 (2016/17)
Grand research

- Federated network
- Open data & publishing
- Business partnerships
- E2E Data Mngt

Year 5+ (> 2017)
Global competitive

- Beyond eResearch
- Long running & real-time science
- Fused & streamed data

Year 1	
Action/Task	**Outputs**
- Institutional arrangements - Set up forums & events - Engage & consult - Survey, assess "As-Is" situation - Prioritise areas & needs - Coordinate new & ongoing projects	- Tier 1 & core services - Data stewardship policies & framework (RDA, etc) - University data science programmes - Solicited proposals in data stewardship - Data intensive research strategy coordinated with funders, strategies and key initiatives

Conclusion

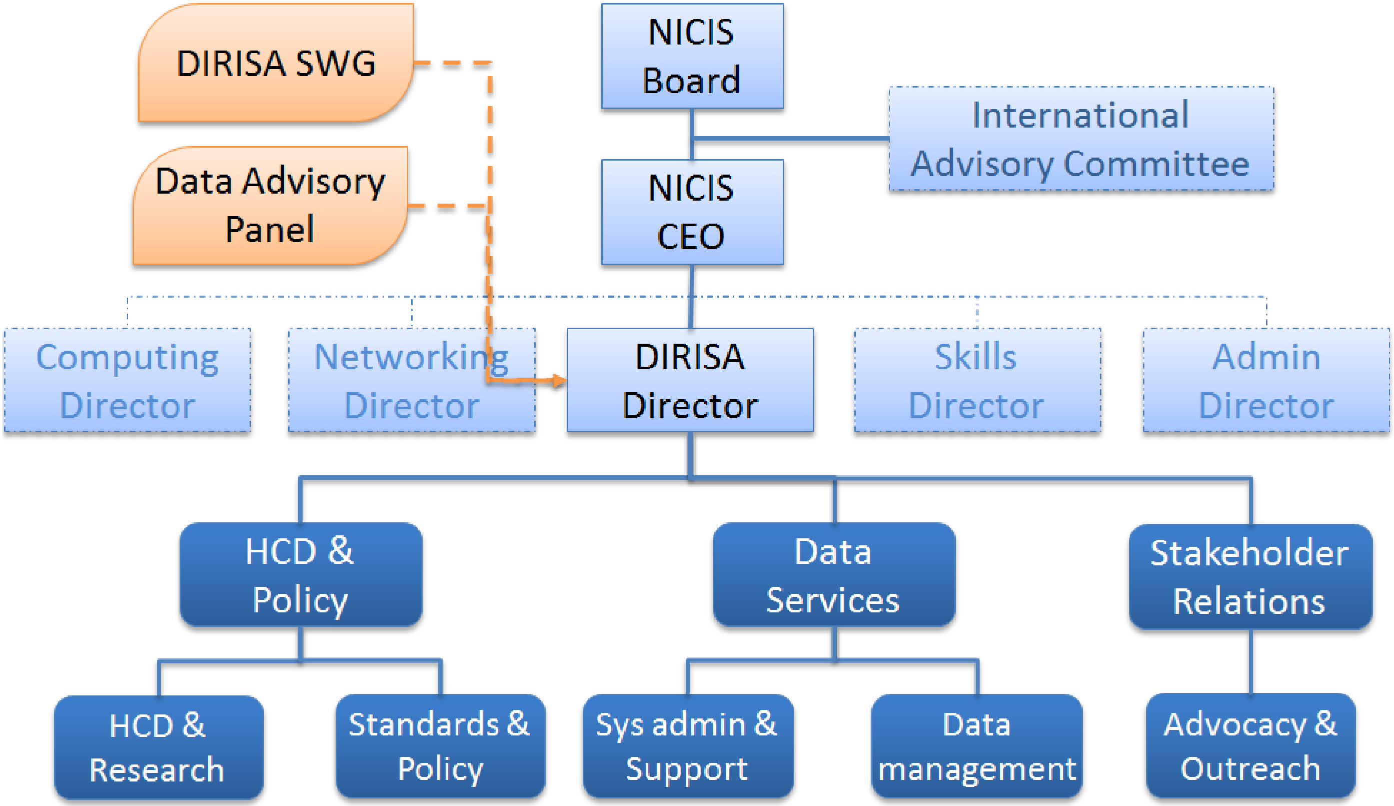
- Business plan further provides
  - Detailed activities, outputs detailed over 3-year timeframe
  - Governance and managerial structure; institutional arrangements and organizational structure
  - Major premises, risks and contingencies
- DIRISA's new remit being formalised
- Implementation plan with stakeholders

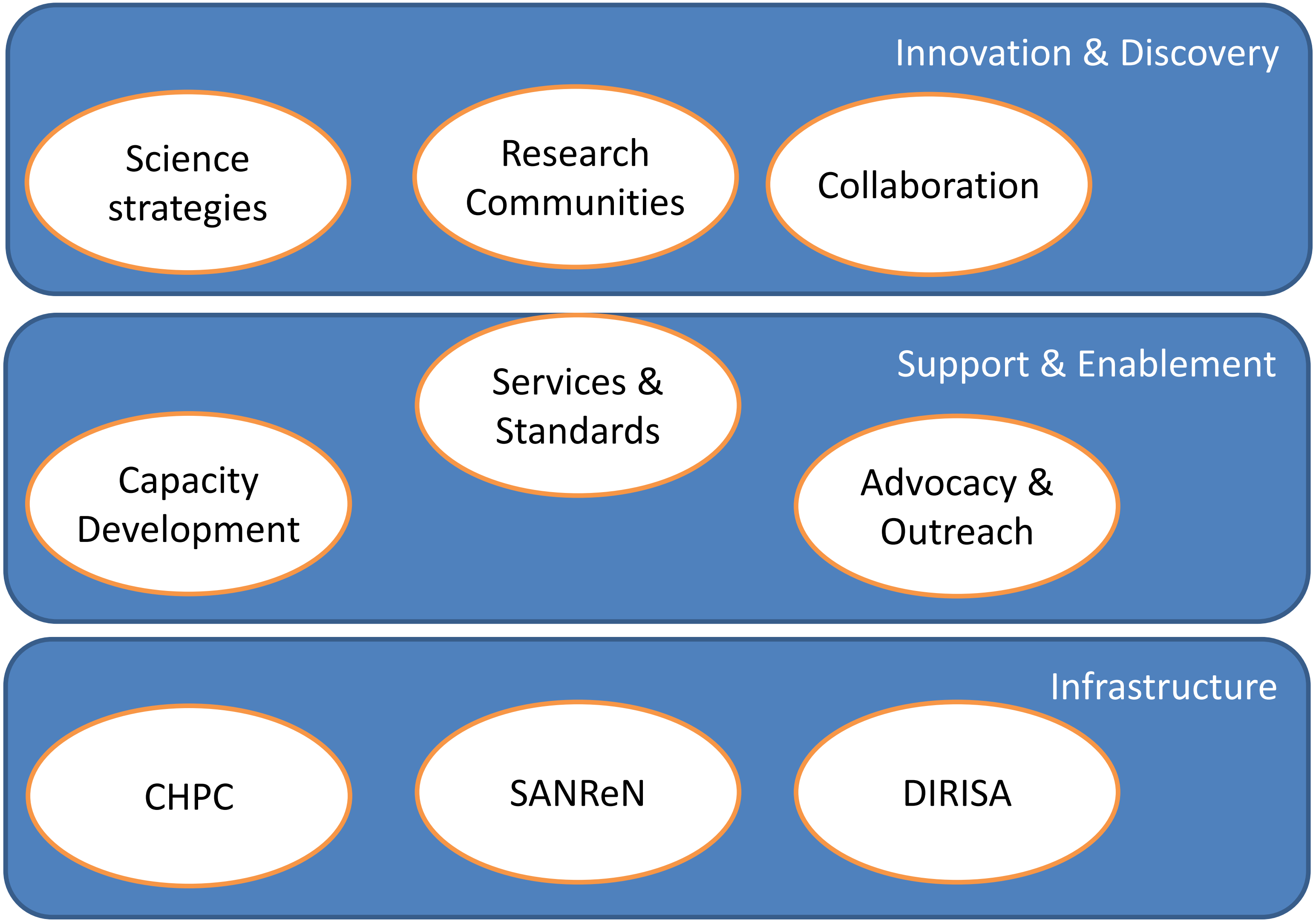
Thank you

"The good thing about data is that there's so much of it

The bad thing about data is that there's so much of it"

Organisational structure




Innovation & Discovery
Science strategies
Research Communities
Collaboration
Support & Enablement
Services & Standards
Capacity Development
Advocacy & Outreach
Infrastructure
CHPC
SANReN
DIRISA