DIRISA: Data-Intensive Research Infrastructure for South Africa.

# Workshop Recommendations

DIRISA: Data-Intensive Research Infrastructure for South Africa.

1. **Technical and Infrastructure Requirements**: A number of requirements were identified in respect of improving usage and utility of existing infrastructure. See Annexure A for details.
2. **Governance Recommendations**:
   1. **Clarification of Funding Models**: DST is providing seed funding and a number of products and services can be offered using this. The nature of these products and services, and the way in which they will be supported free of charge need to be communicated. Future sustainability needs to be clarified.
   2. **Clarification of Mandates**: The mandates and strategic positioning of initiatives such as DIRISA, CHPC, SANREN need to be clearly stated. Future funding horizons need to be clarified. One also has to define how national, local, and global initiatives integrate and complement one another.
   3. **Alignment**: At this point, the alignment between DST funded projects in support of the Global Change Grand Challenges is perceived to be organic and largely left to the individual projects: better alignment can possibly be achieved in respect of this. Alignment with and of other Grand Challenges in a similar way is necessary. Publication of a Technical Roadmap for the initiatives listed in Annexure A might also assist with this. It is assumed that DST will endorse such a Roadmap.
   4. **Advisory Committee**: This may assist with a number of issues and requirements. The Committee will provide monitoring and evaluation. A formal channel of communication between DST and this committee needs to be established. The Advisory Committee needs to be guided by a Terms of Reference. Its relationship to the DST Sector Reference Working Groups must be clarified.
   5. **Global Initiatives:** It is recognised that well-managed international initiatives such as GEO and GEOSS have made a significant contribution to the ease of infrastructure implementation in the Earth and Environmental Sciences domain. This leads to two recommendations:
      1. To investigate the existence of similarly successful initiatives in other domains, and align with these should they exist;
      2. To develop similar (albeit on a much lesser scale) standards architecture and interoperability specifications in cases where such resources are lacking;
3. **Capacity Building and Training Recommendations**: The workshop identified lack of knowledge about digital data management and preservation as a major obstacle to implementation. For this reason, it recommends that all of the following be considered at National level:
   1. **Including data and meta**-data management concepts as a prerequisite for study at Honours level at universities, focusing on both primary and secondary use of data;
   2. **Promotion of degree study** in informatics and library science, and development of a Degree course in Digital Research Data Management; Data visualisation, Data modelling and Data ‘mashing’.
   3. **Alignment of policies** in in respect of research output management and preservation between DHET, NRF (including RIMS), other government departments, and DST;
   4. **Development of technical human resources** to assist with the management, support, and improvement of infrastructure**.**
   5. **Raising awareness** within the **research ethics area** to consider data related issues as part of the ethics approval protocol.
   6. **Support for data creation and use:** Measures to enable data producers to manage data according to best practise and incentives to share and use secondary data are necessary.
   7. In all of these, proper consideration should be given to **curation and preservation management** in addition to cataloguing and dissemination concepts.
4. **Collaboration and Extension into Africa**: This is supported in principle by CHPC and the workshop endorsed this. Frameworks for institutional collaboration will be required to support such initiatives. This effort should be based on the results of the workshop hosted by CHPC on 9 December 2011 (“African Partnership on HPC and Data Research”).

# DIRISA Work Plan

DIRISA will formalise a work plan in January 2012, and in broad terms will aim to achieve the following:

1. Q1 2012:
   1. Confirmation of requirements of different user communities and initiatives, expanding the detail in Annexure A, and selection of technical solutions;
   2. Finalisation of a Technical RoadMap;
   3. Engagement with DST to communicate the workshop recommendations as part of a strategic plan for future management and funding of DIRISA;
   4. Operationalisation of Earth and Environmental Observation Portal;
   5. Clarification of service and product offerings;
   6. Well-supported and resourced hosting environment;
   7. Initiate Technology Pilot Projects in respect of
      1. Preservation – in collaboration with HSRC;
      2. Content Management Systems – in collaboration with SAEON/ SAEOS;
      3. File-based distributed data: WOS, NetCDF, and HDF4 file formats, in collaboration with MRSU, Meraka, and CSIR.
   8. Establishment of a body of knowledge in a community-maintained wiki.
2. Q2 2012:
   1. Preparing promotional material and launching a national road show to promote the existence of the facility and to communicate its purpose and mandate.
   2. Launching an initiative to identify and prioritise training needs.
   3. Operationalisation of the Socio-Economic Data Portal;
   4. Initiation of the development of prioritised training courses;
   5. Implementation of Disaster Recovery procedures, policies, and implementation;
   6. Establishing a well-supported and resourced data archiving service – able to advise and guide the national initiative;
   7. Pilot implementation (requirements, solution selection) of a Digital Preservation Management platform;
   8. Establishment of distributed generic web processing services[[1]](#footnote-1);
   9. Establishment of an Advisory Committee for DIRISA, based on a Terms of Reference, and formally endorsed by DST. The Terms of Reference will be based on this document and the proposal to be submitted to DST;
   10. Possible implementation of the Network Data Centre for Socio-Economic Data in Africa.
3. Q3 2012:
   1. Operationalisation of a Portal for BioInformatics and Health;
   2. Investigation and potential optimisation of bandwidth requirements and SANREN usage;
   3. Development and publication of formal training materials;
   4. Investigation into 3-dimensional spatial data visualisation and visualisation of multi-dimensional data sets.
4. Q4 2012:
   1. Operationalisation of a Portal for Astronomy, Cosmology, and related Sciences;
   2. Implementation of on-line ratings, peer review, and linking to electronic journal/ publishing industry, using existing and emerging initiatives;
   3. Investigation of linkages to citation management software and platforms;

Each operationalization pilot project is driven in respect of user input by collaborating institutions, at present these organisations are expected to be:

* + - 1. Earth and Environmental Science: SAEON and DST (Risk Atlas, SAEOS);
      2. Socio-Economic Data Portal: HSRC;
      3. BioInformatics and Health: Department of Medicine, UCT;
      4. Astronomy and Cosmology: SAAO, SAEOS, and SANSA.

The work plan will be accompanied by a funding proposal to be presented to DST, and ***some of the objectives of the work plan will be dependent on availability of such funding***. The work plan, as it stands, has a reasonable overlap with the recommendations and requirements expressed by the Workshop.

# Annexure A: Requirements

The following set of requirements was gathered from user communities during the following events:

* CHPC Annual Meeting: WOS and iRODS workshops, 5 and 6 December, 2011.
* CHPC Annual Meeting: ‘Infrastructure for Scientific Data Dissemination’ workshop, 8 December 2011.
* Views expressed by CHPC user community, CHPC Strategic Planning Session, October 2011.

These requirements can be summarised as follows:

|  |  |  |  |
| --- | --- | --- | --- |
| **#** | **Requirement** | **Description** | **Note** |
| W.1 | User Support | CHPC provides a variety of services and resources, but the general feeling is that it is a substantial learning curve exists for potential users at institutions such as Universities and Research Councils to use the infrastructure. A number of things can be done to improve the situation, including:   1. Provision of sandbox-type facilities where users can experiment; 2. Clear definition of available services, products, and attendant support expectations/ offerings. This includes defining the scope of services available from DIRISA. 3. Training should be available to prospective users. 4. See also ‘Best Practice’ below. |  |
| W.2 | Participation | User communities need to more involved, and as a result an advisory committee for DIRISA users is proposed.  Communities to be made visible to participants – so that potential users could establish who else is working on a similar initiative. |  |
| W.3 | Integration | It is felt that technical integration between initiatives and projects can be improved. The advisory committee may be able to assist with this. |  |
| W.4 | Generic Approaches | A large number of initiatives overlap in respect of technical and infrastructure requirements – these can also be addressed by documenting, and publicising generic approaches and reference implementations. The Technical RoadMap will also assist with explaining the shared nature of the infrastructure. It must be complemented by a non-technical version for general distribution. |  |
| W.5 | Best Practice | A repository of best practice, case studies, and records of successes and failures is required – possibly made available as a Wiki, editable by the communities of practice. This would include recommended and/or available standards. CSIR has some specific contributions to make as a result of their GSDI implementation. This should be complemented by formal training documentation for DIRISA. |  |
| W.6 | ‘Collateral’ | The assets that are required to support the use of the infrastructure by potential clients:   1. Service Level Agreements; 2. Definition of Products and Services; 3. Data Policies; 4. Support Infrastructure and Help Desk/ Issue Resolution Mechanisms. 5. A defined plan to attract and develop essential skills to support DIRISA |  |
| W7 | Training | Regular opportunities to gain training  A booking system that allows for scheduling the training ahead of time  Making use of a variety of training locations (or appointing training agents at specific dispersed locations  The development of an online training course |  |
| W8 | Promotion | Regular road shows that communicate and promote new functionality  Discipline-linked conference presentations – to show application & share positive stories  Newsletter/ social media  Linking to similar initiatives (international as well as national) |  |

Additional mainstream requirements derive from the following, but are not discussed in detail in this document.

* Specific requirements derived from specific initiatives, stakeholders, and domain-based projects:
  + Earth and Environmental Sciences Portal: joint project between SAEON and CHPC;
  + Social and Economic Sciences Portal: joint project between CHPC, SAEON, and HSRC. Future partners, dependent on funding proposals being successful, may include NRF, DataFirst at UCT, and StatsSA. Should the funding be made available, it will result in the creation of a ‘Network’ Data Centre for SocioEconomic Data, aiming to become a member of the World Data System (WDS);
  + Astronomy and Cosmology Portal, a joint project between CHPC and SAEON;
  + BioInformatics and Health Portal, a joint project between CHPC and SAEON;
  + World Data Centre for Biodiversity and Human Health, a joint project between SAEON, NRF, and CHPC;
* Technical requirements of a number of linked initiatives, funded by DST, and which includes the Risk and Vulnerability Atlas and the BioEnergy Atlas.

The main requirements emanating from these initiatives in respect of infrastructure are the following:

|  |  |  |  |
| --- | --- | --- | --- |
| **#** | **Requirement** | **Description** | **Note** |
| P.1 | Preservation | The majority of initiatives not only require data services, but also preservation services in the long term. | WDS Requirement |
| P.2 | Disaster Recovery and Fail-Over | Adequate provision for disaster recovery and fail-over is required for almost all users and initiatives. | WDS Requirement |
| P.3 | Hosting | Support services to enable high quality hosting of portal environments, enabling high availability, bandwidth, and ease of initiation are required | Portal Requirement |
| P.4 | SensorWebs and Semantic Interoperability | There is significant potential to implement SensorWeb architecture in non-traditional domains, such as Social Sciences. This will lay the groundwork for cross-domain, queryable databases distributed in the web. Semantic interoperability can assist with cross-domain queries and data collation. | Portal Requirement |
| P.5 | Data storage | Providing a service to researchers not connected to large institutions with their own facilities or providing a home for orphaned data.  Providing a facility for the storage of open data – with very definite specifications regarding formats, storage periods, costs – if any, review of usability, etc. | Research community |
| P.6 | Harvesting of metadata | Establish a South African repository of harvested metadata for all data collections (distributed collections) accessible to fellow researchers – irrespective of discipline | Research community |
| P.7 | Service provision | As the benefit from infrastructure is directly linked to the use of related services, infrastructure provision must be coupled with appropriate resources to deliver these services. | Sustainability Requirement |
| P.8 | Certification requirement | There is a requirement for best practise which in the case of data curation implies certification. The requirement is for a certification model to be established, followed by a purposeful implementation. This extends to multiple certification environments, including ‘Trusted Digital Repositories’, OAI, and the WDS | Certification Requirement |

Comments and Notes

1. A key motivation for the funding of DIRISA is that the many of the data sets (though obviously not all) will have many and varied applications, so funding DIRISA will have a better return on investment (multiplicative effect).
2. Similarly, providing the data sets through one, properly managed, portal will facilitate integration of data sets and spawn synergistic and unexpected applications of the data.
3. For (2) and (3) to be possible, though, standards are essential, such as for data formats, quality assurance, metadata, reference models, terminology, maintenance (update cycles), etc. It might also be useful to have registers. Obviously, many data sets will have a spatial framework, but not all. Hence, determining the standards could be complex.
4. Similarly, for (2) and (3) to be possible, collecting and making available metadata is essential. The complication will be in presenting the diverse metadata in a manner that makes it easy to use! :-)
5. The work plan probably should include something on dealing with intellectual property rights and promoting open data access. Obviously, we would all hope that DIRISA will have open data, but this is becoming a big issue, what with SOPA/PIPA in the USA, and the Protection of State Information Bill and the South African Weather Service Amendment Bill down here. Also, my notes say that Heila feels that most university researchers are opposed to open access to data!
6. There should be something in the work plan about establishing the spatial framework for the data sets, to facilitate integration of data sets. I think that it would be useful to establish a relationship between DIRISA and the Committee for Spatial Information (CSI) in this regard.
7. DIRISA should recognise the growing importance of citizen science and user-generated content (which includes volunteered geographical information). For example, the Second South African Bird Atlas Project (SABAP2) currently has 3462366 records with 58.4% of the country covered, with the collection done by 1008 observers, most of whom are amateurs.

1. These will assist users to implement processes, using DIRISA infrastructure for computation, but which can be invoked and managed remotely through standardised interfaces. [↑](#footnote-ref-1)