

Technology RoadMap: Shared Platform/ DIRISA

Introduction and Vision

Wim Hugo - January 2013

During the past few years, SAEON, as an implementing agent, has consolidated a number of funded initiatives into a platform for the provision of community specific web portals, mainly focused on the discovery, application, and visualisation of Earth and Environmental observation data. The data in question, although not exclusively so, was mostly geospatial data in some or other format.

SAEON and CHPC started collaboration in late 2011 on the establishment of DIRISA¹, which has as a technical aim the following:

1. Make use of DST-funded infrastructure such as HPC, SANReN, ad VLDB to create a platform for a collection of 'Science Gateways', serving specific communities, and capable of integration with a wider virtual research environment;
2. Extend the standards-based approach used by SAEON to date to include other domains, with specific focus on four pilot projects to establish Science Gateways:
 - a. Earth and Environmental Observation;
 - b. Human Geography, Social and Economic Sciences;
 - c. Astronomy and Cosmology,
 - d. Health and BioInformatics.

In addition, SAEON/NRF is involved in international initiatives to build global science data infrastructure, of which the ICSU World Data System, CoDATA, and the UNEP-Live/ Eye on Earth initiatives are examples. These initiatives are driving new thinking about the role, nature, and source of both data and meta-data.

The technology roadmap presented here is a synthesis of requirements and needs emanating from all three of these interlinked programmes:

1. Needs and requirements expressed on an ongoing basis by users of the SAEON Data Portal, SAEOS, the Risk and Vulnerability Atlas, and BioEnergy Atlas projects²;
2. The technology and usability vision (developed in part with inputs from the community) to underpin DIRISA;
3. New demands and considerations emanating from the global scientific data initiatives.

We can derive a high-level technology vision from these considerations. In essence, we believe that possible, desirable, and in the public interest to:

¹ Data-Intensive Research Infrastructure for South Africa

² As summarized in URS-IV (4th review of the platform URS).

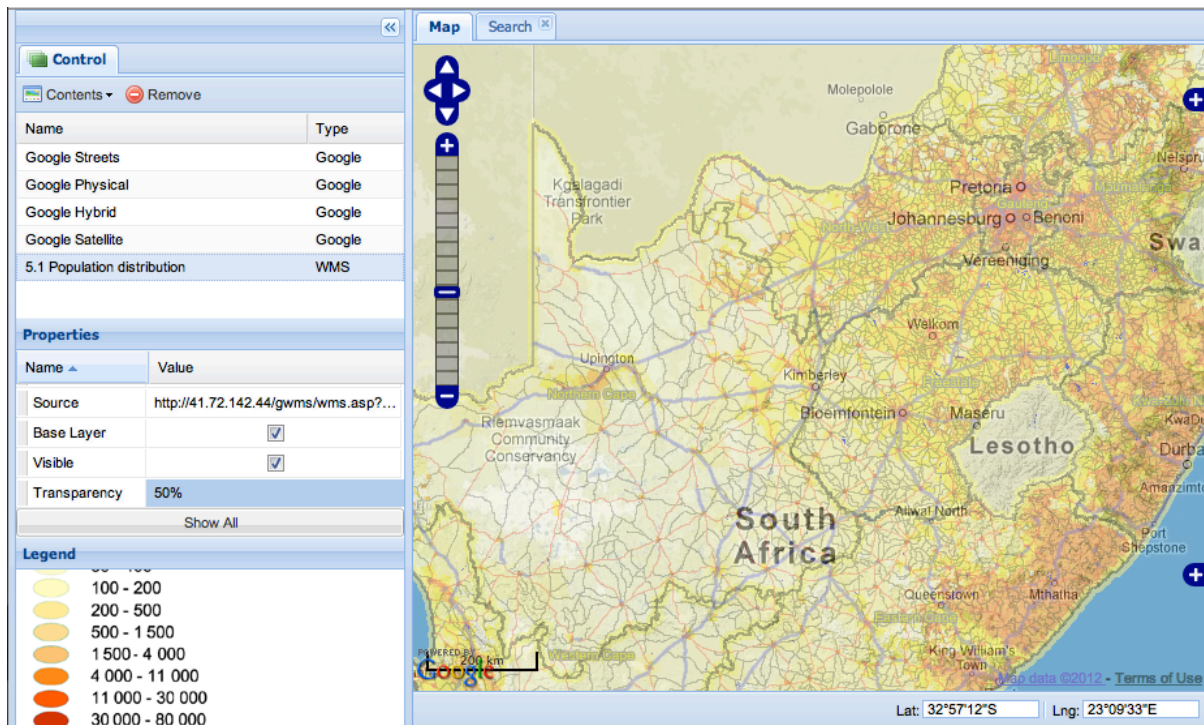
- **Ensure that scientific data is described properly, preserved properly, and discoverable:**
 - This implies continued technology focus and development on tools for standards-based meta-data management, harvesting, and improved search capabilities, including extension of the platform to accommodate new meta-data standards that are used by sizable communities.
 - In addition, we need to focus on the longer-term management of preservation, and as such, we need to develop or purchase licensing for preservation management software.
 - There is a need to issue and maintain persistent identifiers for data and meta-data objects.
 - There is an enormous growth potential in crowdsourced and voluntary contributed information, as well as in the mining of social media data – but this emerging field is poorly understood and will require some specific development to be useful to science;
- **Once discovered, its utility, quality, and scope can be understood, even if the data sets are huge:**
 - If data sets are standardised, and in use by a large community, part of our endeavour must be to assist with the exploratory visualisation and analysis of the data.
 - This, in turn, implies a consistent interface for the handling of a growing array of community-driven data formats, whether these formats are spatial or not, and to assist with its collation.
 - The collation requirement implies work in the emerging field of mediation and semantic interoperability.
 - The emerging standards for quality assessment, both in respect of automated (syntactic and schematic), and usability (semantic) metrics need to be accommodated.
 - Large data sets bring special challenges in respect of visualisation and assessment ;
- **Once understood; it can be accessed freely and openly:** In addition to the support for downloads and ordering of data sets, additional technology implications are
 - Subsetting and standards-driven query of large data sets with a view to partial downloads.
 - Cross-querying of data sets (for example in Sensor Webs or NetCDF repositories) to download collated data from multiple repositories.
 - Development of licensing and policy regimes that promote the least encumbrance in respect of access to data, and continued promotion of open access locally and internationally.
 - Direct support for data publication, citation, and linking with the online journal industry;
- **Once accessed, it can be included into distributed processes, preferably automatically, and on large scales:** The development of online, distributed processing, linked to discovery and visualisation components, is an emerging possibility.
 - Standards exist for web processing in the Earth and Environmental Observation science community, and it will present a significant challenge to develop this element.
 - In addition, the distributed processing of cross-domain data sources will require significant investment in semantic interoperability tools and development of ontologies.
 - For this reason tools for the integration of controlled vocabularies, thesauri, and ontologies into search and collation components need to be developed and supported ;
- **Once processed, the knowledge gathered can be re-used:** The re-use of knowledge is one of the main challenges on a meta-level of information management, because of the relatively poor

performance of scientific data systems in this regard. We have developed a knowledge definition schema (see figure 1), and based on this, will systematically start implementing knowledge preservation and extension technologies. These fall into the following categories:

- Semantic knowledge: users perform work in respect of manual linking and mediation of disparate dimensions or measures in data sources, but the knowledge gained is often not re-usable or persisted. By standardising the ***persistence of mediation***, this deficiency can be mitigated.
- Extension of formal meta-data and surfacing of implicit relationships: formal meta-data has traditionally ignored the rich relationships that exist within it implicitly (for example, co-authorship of an item implies a relationship that meta-data currently does not exploit), and, in addition, formal meta-data is not in a position to be extended with additional information contributed from the user community, social media, or websites created by institutions, initiatives, and projects.
- Documentation of scientific workflows in a manner that is standardised, extensible, and aligned with mainstream (commercial) distributed processing technology, to maximise its potential for re-use in an interoperable way.

Annexure A: URS-IV-1: Refinements to Map Interface

The Mapping Component (Interface) is a standards-based viewer for a variety of standardised data feeds, and can persist composite maps as OGC Web Map Context Documents. It is based on OpenLayers and Geo-EXT. It is wholly JavaScript based and can be embedded into any website.



At present, a map interface exists that is functional, but substantial improvements can be made to layout and presentation. These improvements are required in respect of

1. The way in which legends are presented;
2. The link to meta-data repositories and detailed information about constituent layers in the map, including the ability to download the layer(s) being viewed;
3. Information about licensing and conditions of use, as well as citations;
4. Assistance with data downloads and a requirement to implement provider-dependent terms and conditions and registration prior to download.
5. Improved integration between mapping and portal functions
6. Automated Atlases: Once users have created collections of composite maps in the portal environment, they may want to present these as a single 'Atlas' entry.

These improvements are currently funded and development is under way (Target – Oct 2012).

Additional refinements, not currently funded, include:

1. Users are currently assisted with creation of composite maps by linking to search facilities. The search facilities currently find any resources meeting user criteria, but it will be useful to
 - a. Prioritise results in respect of type of resource (i.e. promote OGC, KML, NetCDF, SOS, etc. to a higher ranking.)

- b. Allow ranking and quality metrics to be used in prioritisation or filters.
- 2. At present, Google Map layers are used as default backdrops, but this is not always applicable or useful. The component allows for other backdrops, but we lack implementation experience or opportunity for other candidates such as OpenStreetMap or Bing.
- 3. The mapping component needs to be documented and published, and registered in the GEOSS repository.
- 4. One of the major unfunded extensions involves the presentation of 3-dimensional, spatially referenced data. This extension includes capabilities to view data in planes and slices, in partially transparent 3D views, and using z-dimension extrusion to represent an additional measure (e.g. DTM/ DEM – style maps but using another variable as a substitute for elevation). Bear in mind that this task has links to other technology threads:
 - a. In DIRISA, some domains have a similar requirement, albeit for non-spatial coordinate systems (for example brain imaging).
 - b. Video technology and 3D multidimensional spatial data is converging – with video essentially being a plane or volume for which very many time series instances exist.

Benefits		Positioning	
SAEON	Yes	Standards-Compliant	Yes
Risk Atlas	Yes	Open Source	Yes
BioEnergy	Yes	Local Collaboration	Maybe
SAEOS	Yes	International Partners	No
DIRISA	Yes	GEOSS Component	Maybe
NSID	Yes	Maturity	
ICSU WDS	Yes	URS	Yes
Other	Yes	SRS	Yes
Funding		UAT/ Tests/ Feedback	Partial
2012/13	Yes	User Manuals/ API	Partial
2013/14	No	Publicly Available	No

Annexure B: URS-IV-2: Improved Templates and Tools for Content Generation

The screenshot displays the SA Risk and Vulnerability Atlas website. The header includes navigation links: Spatial Search, Site Map, Log In, Register, and a search bar. The main title is "SA RISK AND VULNERABILITY ATLAS Electronic Spatial Database". Below the title, there are tabs for Home, Search, Case Studies, Data Providers, and Themes. The current page is the "Socio-economic Theme Page". The page content includes a sidebar with "Maps", "Documents", "Links", and "Value Added Products". The main content area has sections for "Introduction - purpose of theme page" and "Socio-economic dynamics, vulnerability and adaptive capacity". A map of South Africa is shown on the right side of the page. The footer contains the URL: www.csir.co.za/Built_environment/Planning_support_systems/index.html.

Content providers, such as the Risk Atlas and BioEnergy Atlas collaborators, routinely create and publish 'Theme' pages – consisting of a number of structured, predictable sections (such a links to literature or data, maps, and atlases), and a number of free-form, unstructured sections (such as essays, web pages, blogs, and discussion forums). These can all be created in the portal, but users need some training and assistance to accomplish this, and layout results are not always consistent.

1. Define a series of standard content pages to support the creation of a theme or content collection.
2. Define a set of styles per 'portal' for these.
3. Create guideline content for these to allow potential contributors to follow by example.
4. Develop a style guideline in respect of content, layout, citations, and grammar elements.
5. Style guidelines for map creation.
6. Style sheet and template: Make provision for Figure, Chart and Map captions.
7. Assistance with citing third-party content.
8. Assistance with embedding of maps and atlases created elsewhere in the portal environment.
9. Assistance with the embedding of literature, link, and resource links in a consistent way.

These tasks are funded, and are currently under way for delivery in October 2012.

Enhancements that are not funded:

1. Extend the Plone Editor (TinyMCE) to allow seamless inclusion of maps, lists, and atlases into web pages. This involves extension (Python programming) of the TinyMCE product.
2. Analysis of the knowledge embedded into the theme, and collation with other themes in the wider set of portals to update an RDF(a)-based knowledge base³.
3. Presentation of RDF(a) knowledge in useful, graphical formats (for example network diagrams of links between researchers).⁴

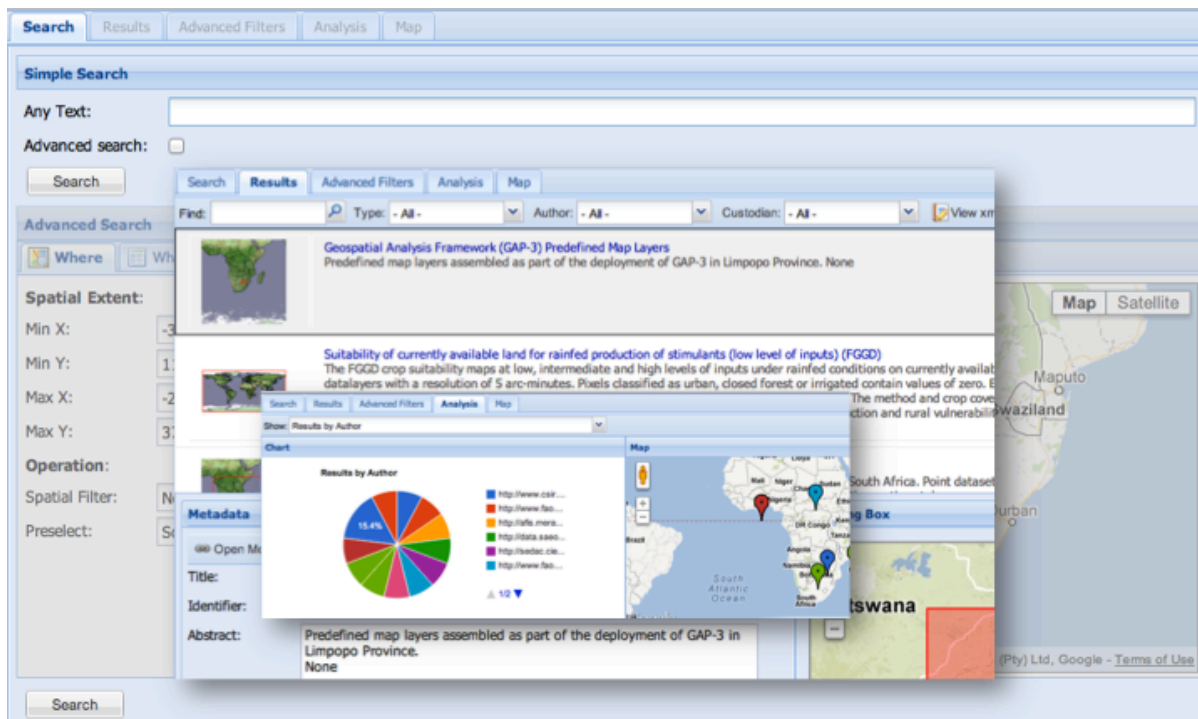
Benefits		Positioning	
SAEON	Maybe	Standards-Compliant	Partial
Risk Atlas	Yes	Open Source	Yes
BioEnergy	Yes	Local Collaboration	No
SAEOS	Maybe	International Partners	No
DIRISA	Maybe	GEOSS Component	No
NSIF	Maybe	Maturity	
ICSU WDS	Yes	URS	Yes
Other	Maybe	SRS	No
Funding		UAT/ Tests/ Feedback	Partial
2012/13	Yes	User Manuals/ API	Partial
2013/14	No	Publicly Available	No

³ SAEON will endeavor to develop this as part of the DIRISA implementation in 2013/14.

⁴ A prototype of this will be developed for the ICSU WDS using NRF funded time. In that context, it will be used to maintain and visualize resources in respect of WDS membership.

Annexure C: URS-IV 3: Refinements to the Search Component

The shared platform uses a search interface that is based on Open Source software (Geo-Ext), developed in JavaScript, and can serve as a modular component in any web-based system. It serves as a client for OGC CS/W search endpoints (harvestable services), and can be pointed to any standard CS/W service.



The search interface not only returns records, but also allows filtering and analysis of the current search result. Furthermore, it allows visualisation of the linked online resources described in the meta-data, provided the services are standardised. It currently supports a set of sources as described in the table.

Service	Description	Standards Organisation	Support
WMS	Web Map Service	OGC	Download/ Map
WFS	Web Feature Service	OGC	Download/ Map/ Chart
WCS	Web Coverage Service	OGC	Visualise
KML	'Keyhole Markup Language'	Google	Map/ Download/ Link to Resource
GeoRSS	Geo-aware Syndication, Geotagged images, ...	Open	Map/ Link to Resource
SOS	Sensor Observation Services	OGC	Map/ Download/ Chart

NetCDF	Large Multidimensional Data Sets	UniData/ UCAR	Map/ Chart
--------	----------------------------------	---------------	------------

Not all of these implementations are equally mature – for example the WCS, SOS, and NetCDF implementations can be described as prototypes of experimental implementations.

Additional non-spatial extensions to the visualisation and download capabilities are described in Annexures D and F.

During the current round of development, these aspects can be addressed in part:

1. Develop more extensive implementations for WCS, NetCDF, and SOS visualisations and downloads.
2. Predefine a number of standardised services to simplify searches and filtering.
3. Multi-tier (2-tier) searches should be available to allow for very large meta-data collections and to manage granularity of search results. This is a necessity given the nature of some data sets (SANSA, WAMIS, ADU, SADCO, and others).
4. Keeping track of search request history for subsequent analysis.
5. Allowing additional meta-data sources (not aonly CS/ W) to serve as a basis for the search – specifically OAI-PMG (Dublin Core), and GeoRSS.

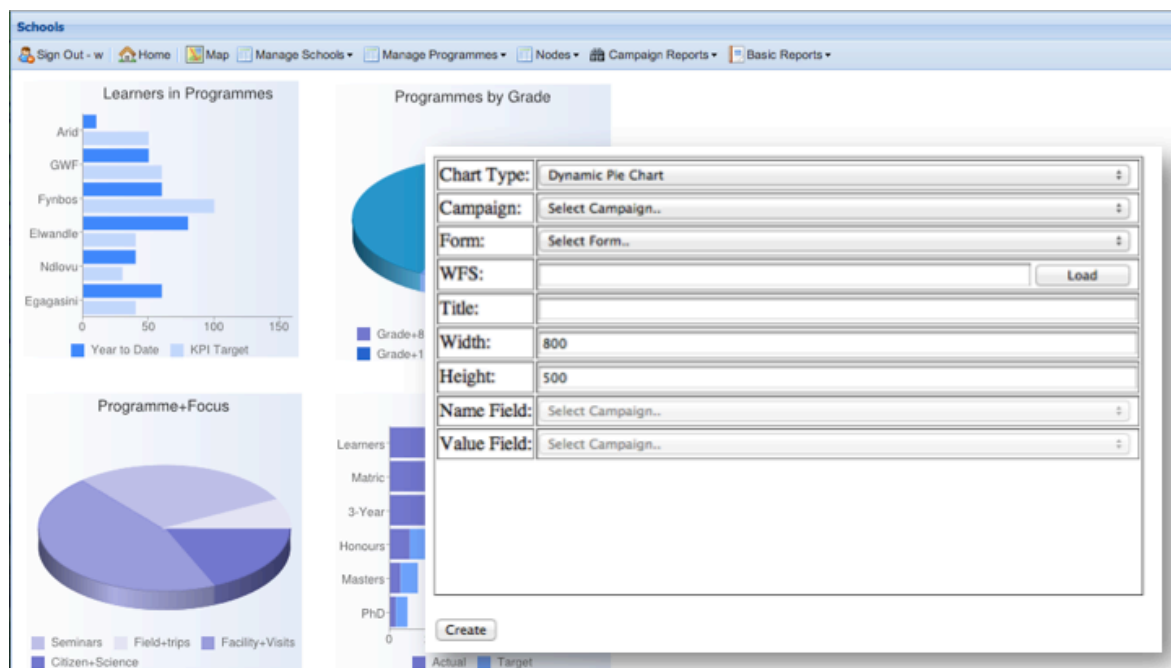
Should additional funding be available, the following extensions can be considered:

1. The initial implementations of NetCDF and SOS are likely to be functional and operational, but extensions and refinements will be possible. Specific aspects include:
 - a. Completely automated, generic requests to OpenDAP and NetCDF repositories, without an intermediary translation service.
 - b. Querying multiple Sensor Observation Services based on the same criteria.
 - c. More complete filtering and querying capabilities for SOS and NetCDF sub-setting prior to visualisation.
2. Migration of the current object-relational based meta-database to a RDBMS/ pointer-based implementation, with two goals in mind:
 - a. Improved indexing and search speeds, and
 - b. More open schema to accommodate less formal meta-data.

Annexure D: URS-IV-4: Improved Support for Analysis, Charts, and Tables

The shared platform currently supports simple graphical and chart-based representations, using Google Charts, and based on WFS feeds. These feeds are mediated and persisted as HTTP REST calls, limiting the size of the data that can practically be represented in this way. A toolkit has been developed to assist with mediation, but users require substantial background information to be able to create chart and graphics views. To date, the development has been largely funded by SAEON, aimed at development of visualisation for generic corporate management systems and for configurable citizen science campaigns.

In addition, the author has been involved in assisting CSIR with development of desktop-based software that offers considerable data visualisation capabilities (GAP, GeoSpatial Analysis Platform) that extends the paradigm established by GapMinder. These approaches are scheduled for inclusion into a web-based component that will allow the shared platform to produce similar representations on request.



Many standardised data formats do not imply or are not limited to two-dimensional spatial views of the data. This obviously becomes more pertinent outside the domain of Earth and Environmental observation. Examples include:

1. Traditional Earth and Environmental Observation Data:
 - a. SensorWeb Data: essentially time series data presented as comma-delimited or GML-style data.
 - b. NetCDF, GRIB, and HDF-5 data: three-dimensional, time-based data with multiple attributes.

- c. OGC-WFS and WCS: potentially many attributes associated with the same feature record.
 - d. For all modes of presentation, if the data is scientific data with some error estimate, more and more attention is being given to the simultaneous presentation of error or quality. This is commonplace for chart- and table-based representation, but not for spatial representations.
2. Health and Bio-Informatics-Related Data:
 - a. Three-dimensional representations of the body, including ability to rotate images and represent plane sections,
 - b. Genome sequencing visualisations,
 - c. Molecular graphics,
 - d. MRI scans,
 - e. Phylogeny,
 - f. Audio and Video files obtained as a result of a number of scanning procedures.
 3. Astronomy and Cosmology: Not investigated yet.
 4. Human and Economic Sciences:
 - a. Substantial focus on statistical analysis results (correlation, distributions, etc.).
 - b. Use of cartograms (adjustment of areal representation of the earth to reflect a non-areal measure, such as GDP per capita).
 5. General purpose representations:
 - a. 'Knowledge representations', in the form of word clouds, trees and directed graphs, dendograms and clustering, heat maps, and the increasing use of 'mind maps', concept maps, and other relational representations.
 - b. Increased use of 'Infograms', representing several allied concepts using a mash-up or composite view that relies heavily on graphics,
 - c. The special case of 'indicator representation', in which aspects of the representation have very specific roles, such as 'actual', 'planned', 'forecast', 'context', 'events', and so on.
 6. Large Data Sets: We are aware of three specific cases of large data sets that require some form of meso-level visualisation so that the user can understand the underlying data and its quality.
 - a. Search results on meta-data, which, increasingly, can run into millions of hits;
 - b. Large audio and video files, as well as frequently sampled remotely sensed data;
 - c. Meta-data and data mining operations.

There is a modest programme scheduled to end 2012/13 financial year, in support of anticipated use within the Risk and Vulnerability Atlas, and in the BioEnergy Atlas:

1. Operationalisation of support for GAP-like data analysis and charting, using WFS, WCS, NetCDF, and SOS as input services.
2. Development of 'source-agnostic' collation and analysis services for Web Coverage Service inputs – see below.
3. Development of knowledge visualisation components for the ICSU WDS and related projects, which will also be used in DIRISA, Risk Atlas, and other portals.

4. Development of indicator exchange protocols and visualisation techniques. This will form part of the development of a Business Intelligence system for NRF, u tis usable beyomd the NRF.

Additional funding will be required to develop the visualisation techniques required for the non-E&EO domains:

1. Visualisation of non-spatial 3-dimensional data, such as body scans and model results;
2. Increased support for and integration with mainstream statistical analysis software such as SAS, R, and CASS.
3. Improved visualisation of large data sets, specifically NetCDF, HDF-4, and GRIB data sets. This will benefit from international collaboration.
4. Visualisation of genome sequences.

Benefits		Positioning	
SAEON	Yes	Standards-Compliant	Yes
Risk Atlas	Yes	Open Source	Yes
BioEnergy	Yes	Local Collaboration	Maybe
SAEOS	Yes	International Partners	Maybe
DIRISA	Yes	GEOSS Component	Maybe
NSIF	Yes	Maturity	
ICSU WDS	Yes	URS	Yes
Other	Yes	SRS	Yes
Funding		UAT/ Tests/ Feedback	Partial
2012/13	Yes	User Manuals/ API	Partial
2013/14	No	Publicly Available	No

Annexure E: URS-IV-5: Additional Static Page Content and General Portal Functionality

The shared platform offers standard content management functionality, which includes the ability to create, upload, and optionally publish content – either generally or to specific individuals and groups. In addition, it offers considerable opportunity for community collaboration – through blogs, commenting of content, and discussion forums.



The current work programme includes the following, but it is unlikely that all the requests will be accommodated with currently available funding:

1. Terms and Conditions that can be added per data provider, and is invoked as part of the data download or visualisation process.
2. Links to standard data policies and licenses – such as may exist, as well as contributions to national and international efforts to establish these standardised policies and licenses.
3. Privacy statements: the shared platform needs to establish a well-defined and internationally acceptable privacy policy in respect of user data.
4. Support for meta-data as an embeddable reference:
 - a. Citation Management;
 - b. QR-codes and shortened URLs;
 - c. Download citations as one of a number of reference exchange formats, and upload references as meta-data objects (probably Dublin Core).
 - d. Digital Object Identifiers.

5. Benefits of Registration: assist users with useful additional functionality:
 - a. Saved Maps and Charts, and building a personal space within the portal environment;
 - b. Saved Searches: configuring these as named URLs that can build a repository of tailored meta-data views;
 - c. Usage Statistics: in respect of own activity, responses to data sets contributed, etc.
 - d. “My Network”: inferred links to researchers that share the same interests (as derived from search keywords), possibly interesting data sets, and other useful knowledge-type derivations;
 - e. Allowing registered users to create and embed RSS feeds into other clients, such as mail clients and newsfeed clients.
6. Commenting on and rating content, including meta-data content. This is not a simple matter, since meta-data is harvested and replaced at regular intervals from feeder systems and data providers. This requirement implies that some form of synchronisation should be possible – since the shared platform will retain user statistics ‘locally’, while the corresponding meta-data record is maintained externally and periodically refreshed locally.

Benefits		Positioning	
SAEON	Yes	Standards-Compliant	Partial
Risk Atlas	Yes	Open Source	Yes
BioEnergy	Yes	Local Collaboration	Maybe
SAEOS	Yes	International Partners	No
DIRISA	Yes	GEOSS Component	Maybe
NSIF	Yes	Maturity	
ICSU WDS	Yes	URS	Yes
Other	Yes	SRS	Yes
Funding		UAT/ Tests/ Feedback	
2012/13	Yes	User Manuals/ API	Partial
2013/14	No	Publicly Available	No

Annexure F: Enhanced Support for Online Resources

D.1 Earth and Environmental Sciences

One would be considering the following generic data representations:

Base Data Schema	Source Service	Default View	Other Views
XY-A	WMS	2D Map: [XY][A]	
	WFS (GML)	2D Map: [XY][A]	
	WCS	2D Map: [XY][A]	
	KML	2D-Map: [XY][A]	3D-Map: [XYD][A] 2D-Map: [XY][M] (Z)
	GeoRSS	2D-Map: [XY][M]	
XYZ-A	WFS (GML)	3D-Map: [XYZ][A]	
	KML	3D-Map: [XYZ][A]	
XYZT-A	SOS	2D-Map: [XY][C]	Trend Chart: [AT] (XY) Correlation: [AB] (XYT) Bar Chart: [AB] (XYT) Pie Chart [A] (XYT)
XYZT-A	NetCDF, HDF4		

Guide:

1. X, Y, Z: Spatial Coordinates in three dimensions
2. T: Time dimension
3. M: A collection of attributes formatted as a marker data element.
4. A: an attribute associated with a record in the data (any cell value).
5. B: an attribute associated with a record in the data (any cell value).
6. D: An attribute A used as a depth/ height/ Z coordinate.

To be completed.

Annexure G: Specific Portal Enhancements and Additional Services

1. Version Request
2. RSS Request
3. User List Request
4. Statistics Request
5. Automated Meta-Data Generation
6. Content Registration Service
7. Improved News Item and Blog Support
8. Content Folder Request

To be completed.

Annexure H: RDF and RDFa Support

1. RSS to RDF and RDFa
2. RDF(a) to Tag Cloud
3. RDF(a) to Ordered List
4. RDF(a) to Network Diagram
5. RDF(a) to Chart Collection`

To be completed.

Annexure J: Adapters for Standardised Harvesting

To be completed.

Annexure K: SensorWeb Enablement Adapters