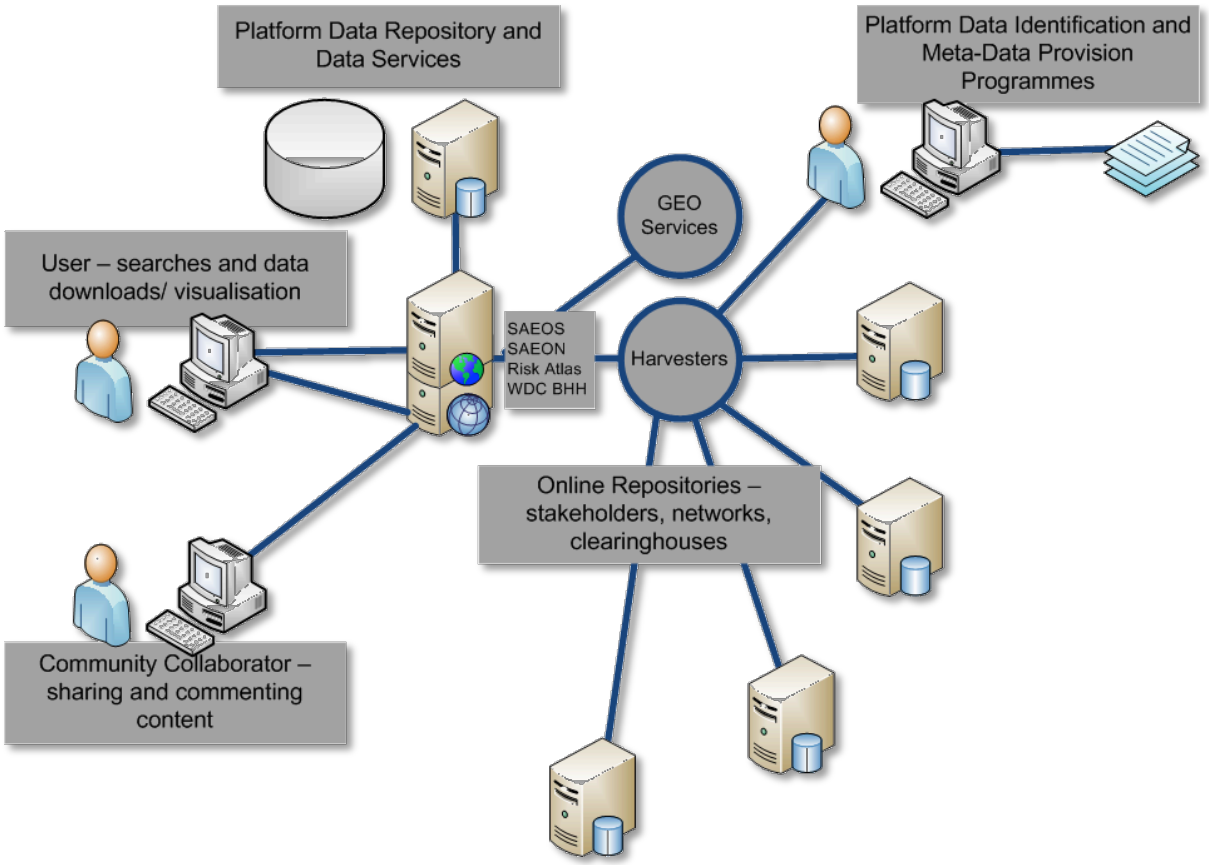
Options for Meta-Data Exchange

Wim Hugo, SAEON

2 August 2012, Updated 25 February 2013

SAEON manages and operates a shared platform on behalf of a number of stakeholders, including its own scientists and collaborators. The shared platform currently hosts the SAEON[1] Data Portal, The South African Risk and Vulnerability Atlas[2], the South African Earth Observation System[3] (SAEOS), the BioEnergy Atlas, the CSIR GSDI Geoportal[4], and a prototype World Data Centre for Biodiversity and Human Health in Africa (WDCBHH)[5]. It is also being extended to host DIRISA (Data-Intensive Research Infrastructure for South Africa).



[1] South African Environmental Observation Network, mandated to monitor the environment/ maintain observation data for the long term.
[2] Funded by the South African Department of Science and Technology as part of their Global Change Grand Challenge Research Plan.
[3] SAEOS is funded by DST as part of their contribution to GEOSS
[4] Previously known as 'CoGIS', funded jointly by CSIR, SAEON, and other stakeholders.
[5] Funded by ICSU, NRF/ SAEON and the US Geological Survey/ NBII.

The platform is based on a shared and aggregated meta-data repository, and the meta-data repository is capable of accepting and working with a range of well-established meta-data standards. These include Dublin Core, SANS 1878, the ISO 19115 family, EML, and FGDC. The list is likely to be extended from time to time to accommodate other standards in widespread use by a user community or new provider.

The platform and its hosted portals are designed to serve a stakeholder community as a resource for the ***referencing, discovery, management, and optional archiving of relevant data sets*** and information objects. It also allows the composite visualization of distributed data sets, provided that access to these sets is automated and standardized.

The platform also provides ***collaboration, sharing, and content composition facilities*** for the distributed creation and management of value-added themes, discussions, blogs, community pages, and more.

SAEON is currently the implementation agent for the platform and its hosted portals, and can assist with

- guidelines in respect of choice of meta-data standard and supporting open source software,
- creation of meta-data,
- guidelines in respect of applicable data standards and services,
- guidelines in respect of software that can be implemented in support of the GEO architecture,
- deployment of components to assist with meta-data exchange and synchronisation,
- deployment of platform components for search and data visualisation within external sites,
- and hosting or archiving of data in specific circumstances.

In the recent past, the development programme has extended the platform to accommodate a ***range of meta-data standards*** (ISO19115, SANS 1878, EML, Dublin Core, and FGDC), and to allow automated harvesting of such meta-data from repositories offered by data providers and stakeholders.

The remainder of the document deals primarily with meta-data and data exchange.

Data Custodians and Providers

Access to data sets and the mechanisms whereby it is implemented are determined by the preferences of the provider community, and can range from the ideal situation of open, automated standards-compliant services (WFS/WMS, NetCDF, etc.) through download links or redirection to provider-determined websites.

Under ideal circumstances, ***data providers curate their own data holdings*** on a platform that is acceptable to them, and the links to data sets are included into meta-data records. ***The platform can also host data on behalf of a provider should the requirement exist.***

Options for providing access to data include the following:

1. Automated access to data, in any of the following standardized formats:
    a. A Web Map Service (WMS) or Web Feature Service (WFS), commonly used to serve spatial data over the web.
    b. KML files located anywhere at a valid web address (URL);
    c. GeoRSS services located anywhere at a valid URL;
    d. Any link to a valid web-based document (such as a PDF, ZIP, or data file).
    e. NetCDF and HDF formats (in the pipeline).
    f. Any text or comma-delimited data file.
    g. SensorWeb-enabled services.
2. Any valid link to a web page maintained by the provider, where a prospective user
    a. Can be asked to register or log in if already registered;
    b. Fills in a request for data to be processed off-line;
    c. Constructs a query to subset data for download or visualization (See next section);
    d. Is provided with contact details so that a data set can be obtained manually.

SAEON has developed a comprehensive data policy that regulates meta-data and data access, and can be used as a serves as a basis for formalized agreement between a data provider and the platform.

SAEON has defined the user requirements for ***access frequency reporting*** to providers and have implemented services to support this towards the latter half of 2012. This will allow data providers to automate the periodic update of usage statistics from the platform to their own systems.

Query Parameters

In some cases, the search parameters that the platform has obtained in one of the portals needs to be passed to a web page or service under control of the provider. At present, this can be achieved by constructing an appropriate link in the meta-data record. There is a standard interface for this exchange (based on CS/W), but providers are free to include any link and query string that suits their own systems. See Annexure A.

A meta-data record can point to multiple resources on any number of sites: See Annexure B.

Options for Meta-Data Exchange

The platform supports a number of automated meta-data exchange options and methods, and manual interactions can also be supported. The table below provides a summary of these:

Mode	Method	Description/ Notes	Example/ Specification
Manual	Upload an XML Meta-Data File	User needs to be registered as a contributor, and a repository must exist for the data provider. XML can be uploaded and validated (single record at a time)	
	Create a Meta-Data Record from Template	User needs to be registered as a contributor, and a repository must exist for the data provider. XML can be created within a repository of choice,	

		using a standard of choice.	
Automated	Upload from FTP	Collections of meta-data records, in a zipped file, can be harvested with a specified frequency from an FTP site. To create a dedicated harvester, the user needs to be registered as a contributor, and a repository must exist for the data provider.	
	Upload from Web Folder	Collections of meta-data records, in a zipped file, can be harvested with a specified frequency from a Web Folder (HTTP Address). To create a dedicated harvester, the user needs to be registered as a contributor, and a repository must exist for the data provider.	
	Harvest Using CS/W	Collections of meta-data records can be harvested with a specified frequency from a standardised CS/W end point. To create a dedicated harvester, the user needs to be registered as a contributor, and a repository must exist for the data provider.	GeoNetwork exposes a standardised CS/W harvesting end point. CS/W:
	Harvest Using OAI-PMH	Collections of meta-data records can be harvested with a specified frequency from a standardised OAI-PMH end point. To create a dedicated harvester, the user needs to be registered as a contributor, and a repository must exist for the data provider.	MetaCAT exposes a standardised OAI-PMH end point. OAI-PMH:
	Harvest Using SAEOS/ SAEON Adapter	The platform can provide an adapter that translates a database view or table into a harvestable CS/W endpoint. Take note that (1) Only Windows platforms supported at present (2) The adapter may require customisation for the provider's circumstances	Contact SAEON for details

The typical options that can be invoked in respect of different meta-data standards are as follows:

Standard	**Protocol**				
	FTP – ZIP	**HTTP – ZIP**	**CS/W**	**OAI-PMH**	**DB Adapter**
ISO 19115 & p2	Yes	Yes	Yes		Yes
FGDC	Yes	Yes	Yes		
SANS 1878	Yes	Yes	Yes		Yes
Dublin Core	Yes	Yes	Yes	Yes	Yes
EML	Yes	Yes		Yes	

Implementing Meta-Data Management in Your Organisation

Organisations vary widely in respect of maturity, availability of resources (human, infrastructure, funding), and management methods. In addition, there is an overlap between the typical content of meta-data standards and the information that organisations need to maintain to manage their data assets – and meta-data management systems often do not accommodate these institutional requirements.

The following options generally apply:

<table>
<tr><th colspan="3">Situation</th><th>Best Option</th><th>Second Option</th><th>Assistance from Platform</th></tr>
<tr><td rowspan="9">Own Infrastructure and Human Resources</td><td rowspan="2">In-House Meta-Database</td><td>Windows-based</td><td>Deploy Adapter</td><td>See 'Postgres' below</td><td>Limited assistance to customise adapter</td></tr>
<tr><td>Linux/ Ubuntu/ …</td><td colspan="3">See 'Postgres' below</td></tr>
<tr><td rowspan="4">Standardised Meta-Data System</td><td>GeoNetwork or GeoNode</td><td>Adequate: CS/W Harvester</td><td rowspan="3">Export, ZIP, and harvest (FTP or HTTP)</td><td rowspan="3">Assist to set up harvester and test</td></tr>
<tr><td>MetaCAT</td><td>Adequate: OAI-PMH Harvester</td></tr>
<tr><td>DSpace/ Fedora</td><td>Adequate: OAI-PMH Harvester</td></tr>
<tr><td>Other</td><td>Export, ZIP, and harvest (FTP or HTTP)</td><td>Evaluate on case-by-case basis</td><td>Assist to set up harvester and test</td></tr>
<tr><td rowspan="3">Spatial Database</td><td>Oracle, MySQL, SQL Server</td><td rowspan="3">Deploy GeoNode or GeoNetwork with some manual work to update. CS/W automated harvester.</td><td rowspan="2">Create Spreadsheet-based inventory</td><td rowspan="2">Assistance to map maximum data from spatial data repository to meta-data. Spreadsheet can be imported in semi-automated way</td></tr>
<tr><td>Other</td></tr>
<tr><td>PostGres</td></tr>
<tr><td></td><td>ArcGIS/ SDE</td><td>Export to FGDC or ISO 19115 for HTTP/ FTP Harvesting</td><td></td><td>Project being considered to create automatable adapter for this</td></tr>
<tr><td>File-Based Systems</td><td>No Current Implementation</td><td>Deploy GeoNode or GeoNetwork</td><td>Create Spreadsheet-based inventory</td><td>Spreadsheet can be imported in semi-automated way</td></tr>
</table>

Human Resources Only	Platform can Provide Data and Meta-Data Repository		SAEOS Data Store and GeoNetwork Node		Automatically harvested by platform
	File-Based Systems	No Current Implementation	Deploy GeoNode or GeoNetwork	Create Spreadsheet-based inventory	Spreadsheet can be imported in semi-automated way

Annexure A: Linking Back to Parameterised Services

Table A.1: Typical Search Parameters ('Common Queryable Elements')

Common Queryable Elements		
Name	**Definition**	**Data type**
Subject [a]	The topic of the content of the resource [b]	CharacterString
Title [a]	A name given to the resource	CharacterString
Abstract [a]	A summary of the content of the resource	CharacterString
AnyText	A target for full-text search of character data types in a catalogue	CharacterString
Format [a]	The physical or digital manifestation of the resource	CharacterString
Identifier [a]	An unique reference to the record within the catalogue	Identifier
Modified [c]	Date on which the record was created or updated within the catalogue	Date-8601
Type [a]	The nature or genre of the content of the resource. Type can include general categories, genres or aggregation levels of content.	CodeList f
BoundingBox [d]	A bounding box for identifying a geographic area of interest	BoundingBox, See Table 2
CRS [e]	Geographic Coordinate Reference System (Authority and ID) for the BoundingBox	Identifier
Association	Complete statement of a one-to-one relationship	Association, See Table 3

a Names, but not necessarily the identical definition, are derived from the Dublin Core Metadata Element Set, version 1.1:ISO Standard 15836-2003 (February 2003)

b Typically, a Subject will be expressed as keywords, key phrases or classification codes that describe a topic of the resource. Recommended best practice is to select a value from a controlled vocabulary or formal classification scheme.

c DCMI metadata term <http://dublincore.org/documents/dcmi-terms/>.

d Same semantics as EX_GeographicBoundingBoxclass in ISO 19115.

e If not supplied, the BoundingBox CRS is a Geographic CRS with the Greenwich prime meridian.

f A "CodeList" is a CharacterString taken from an authoritative list of CharacterStrings or Identifiers. The authority may optionally be identified in the value.

Table A.2 Typical Encoding as a Key Value pair (REST) Query String

Table A.2: Service Interface Implementation (Key Value Pairs): **Fields marked as *: Optional** **ReturnType: Applies only if the Service returns catalogue records.** **Taxonomy: Applies only to EML services** **In many cases, pointing to a Unique ID in the target catalogue or system will be adequate for data identification.**	
Description	**Interface Definition**
URL	http://<baseURL>/<ServiceName>?
Unique ID	&UID=<sometext>
ReturnType	&returntype=<1,2,3> 1. Unspecified Format (Default) 2. GeoRSS 3. Dublin Core
AnyText	&anytext=<some text>
Date Range* (Temporal Coverage)	&fromdate=yyyy-mm-dd
	&todate=yyyy-mm-dd (default is always today's date)
Modified – Publication Date	&modified= yyyy-mm-dd

Title	&title=<some text>
Abstract	&abstract=<some text>
Subject	&subject=<some text>
Extent (Spatial Bounds)	&extent=<MinX - West>%2C<MinY - South>%2C<MaxX - East>%2C<MaxY - North>
WestBoundLongitude	&WestBoundLongitude=LO
SouthBoundLatitude	&SouthBoundLatitude=LA
EastBoundLongitude	&EastBoundLongitude=LO
NorthBoundLatitude	&NorthBoundLatitude=LA
Spatial Search Type*	&spatialtype=<Contains, Equals, Intersects, Touches, Within, Outside> (by default, if absent, equates to "Contains".
Taxonomic Coverage*	&eml_TaxonomicCoverageRankName=<Kingdom, Phylum, Genus, Species>
	&eml_TaxonomicCoverageRankValue=<some text> a
a: Add to anytext	

Annexure B: Multiple Resource References in Meta-Data

In ISO 19115/ SANS 1878 Meta-Data, multiple references to online resources are allowed. These can be used to reference

- More than one data resource in the same meta-data record (for example a set of scenes or a data table with lookup tables);
- A natural collection of resources (data service, descriptive document, corporate or project website).

The platform currently attempts to match the collection of URLs in meta-data with a range of typical services. For example, if one or more of the URLs designated as online resources point to OGC web mapping services or sensor observation services, the platform will invoke a suitable (pre-) viewer.

We are constantly refining this function. Currently, we should correctly allow previews and download of:

1. OGC WxS and Sensorweb data
2. GeoRSS services
3. KML files (single layers)
4. Documents, PDFs, and ZIP files
5. EML Meta-data
6. EML/ MetaCAT CSV Data Sources
7. Pass-through to any website or web page.
8. WAMIS Time Series Data
9. SAEON Text File Data

An example can be found at http://urlmin.com/saeos_signature