

Concept Note: Working for Data

Wim Hugo, SAEON

One of the major obstacles to greater availability of data, both from the state and academic or research environments is the lack of capacity to properly describe, standardise, and publish data. This situation, coupled to the general paucity of data publication infrastructure create significant obstacles to a constant stream of published data, preserved and curated in such a way that it can be found today and re-applied in future.

The process model in Figure 1 serves to discuss the problem further, and to design possible solutions.

Figure 1: Process Model for Research, Development, and Innovation Assets



We believe that while there is significant investment by DST in the creation of Research, Development, and Innovation outputs, the return on this investment (largely dependent on the rate at which the outputs can be re-used and serve as a stimulus for further innovation) is constrained by relatively poor performance[[1]](#footnote-1) in some of the other critical process steps.

* Outputs are ***rarely described properly*** using accepted meta-data standards, limiting its discoverability and usefulness, now and in future;
* Outputs are often ***not compliant with simple domain or community standards*** in respect of vocabularies, syntax, semantics, and schema, reducing its value to end users who have to invest time and money to correct these deficiencies;
* Should outputs be published, the ***services offering the outputs should be standardised*** so as to improve its usefulness, and critically in a data-intensive future RDI world, improve the extent to which it can be ***applied and processed automatically***.
* Provision for long-term preservation and curation of these outputs is virtually non-existent, except on an ad-hoc basis in more mature research or academic institutions.

The solution to these issues requires policy, capacity in the institutions, and incentives to individuals in the following ways:

* ***Knowledge of the applicable standards***: applicable to meta-data, community, and publication standards. These capacities are largely absent in the South African context, and longer-term solutions include making coursework on these topics compulsory for honours studies in the sciences and engineering.
* ***Incentives***: we believe that incentives to follow a more defensible approach will slowly improve as data publication and citation becomes more common in the research environment, but employees in the state sector, for example, are not measured in this way – compliance to a basic process will have to form part of key performance areas, and to enable this, policy and regulation will be required. Furthermore, policy compelling proper curation and publication of RDI outputs as a condition of funding will have a significant impact.
* ***Infrastructure***: publication and preservation is additionally constrained by lack of appropriate infrastructure, again this is commonplace outside of mature institutions. This problem is relatively simple to address, and initiatives such as SAEON, SAEOS, and DIRISA address this requirement in part.

These measures, especially in respect of capacity building by conventional means, will take a long time to bear fruit, and can be viewed only as the long-term solution. In the short and medium term, an interim solution is required

# ‘Working for Data’

A solution that follows the successful models of ‘Working for Water’ and ‘Working for Fire’ is envisaged. In this solution, use is made of suitably qualified ‘crowd-sourced’ resources to perform steps in the process. Small contracts (‘hits’) are published in a web environment, and participants follow on-line instructions and use on-line tools to shepherd RDI outputs through the process. Incentives can range from compensation to ‘street credibility[[2]](#footnote-2)’, or a combination of both. Effectiveness is improved by double-checking and quality assurance mechanisms - and participants are also matched to tasks based on prior performance – in other words, some form of ranking.

This approach has significant advantages:

* It requires little or no office and physical infrastructure;
* Developing a supporting software platform can be speeded up by re-factoring components already in use for the Risk Atlas, SAEOS, and SAEON.
* It makes copious use of recent graduates and current students, a sector of the workforce that is currently underemployed and may be well served by extra income;
* In doing so, it is simultaneously educating the participants in the standards, principles, and processes involved in proper management of RDI outputs;
* By careful opening of the underlying software to selected open source development, some participants will be able to make significant refinements and improvements to the platform;
* The programme will greatly assist in extending the capacity of non-research or academic institutions, such as state-owned enterprises and government departments, but also serves as an enormous improvement to the likely situation at provincial government departments, district, and local municipalities.
* The program is ambitious, but can be initiated on a small scale.

# Estimate of Likely Costs

SAEON currently manages a team to process data in this way for the SAEOS and Risk and Vulnerability Atlas projects. The experience with these data sets, mostly spatial data, suggests that:

1. Some data sets are obviously difficult to standardise, process, and describe, but the majority take between 2 and 4 hours to publish on the SAEOS infrastructure if some form of documentation exists.
2. Quality assurance may take another hour or two.
3. Based on this, one can assume that most data sets could be described, standardised, published, and committed to proper infrastructure within a working day – 8 hours. At the standard rate for graduate interns, this equates to approximately R 500 per data set – in our view, a small amount compared to the ensuing benefits, both for the data provider and the community at large.
4. The costs will be raised by a requirement to provide software, marketing and promotion, and support for the user community.

Implementation will require some capital investment on software development, our estimate being in the order of R 200,000, and some provision for support of users – estimated to be about R 30,000 per month. Table 1 provides a high-level estimate of the costs that will be associated with a programme to process 12,000 data sets over a period of three years.

Bottom line: Process 12,000 data sets in 3 years, at an average cost of R 800 per data set.

Table 1: Estimate of Costs based on SAEOS/ Risk Atlas experience.



1. The rate of publication is further constrained by IP and data publication issues, but this is a separate (though related) policy issue. See a companion document prepared by SAEON on behalf of DST. [↑](#footnote-ref-1)
2. One of the most successful recent examples involved ‘Old Weather’, where online resources were used to digitise and capture historical weather observations from Royal Navy logbooks stretching back 300 years. No payments were involved, and 18,000 logs were processed in 18 months – driven by a desire to gain recognition only. See <http://oldweather.org> [↑](#footnote-ref-2)