netherlands eScience center

www.esciencecenter.nl

Data-Stewardship in the Big Data Era:
Taking Care of Data

January 2013

On behalf of:

Netherlands eScience Center Management Team

Netherlands eScience Center Integrators

Contact Person: Scott Lusher (scott.lusher@esciencecenter.nl)

Data-Stewardship in the Big Data Era
Taking Care of Data

Executive Summary:

The purpose of this document is to propose a series of actions to NWO as it formulates its policies on data-stewardship. The contents of this document are deliberately provocative and formulated to encourage discussion and the eventual development of an agreed data-stewardship strategy. Above all, we hope it contributes to a climate in which researchers, funders and administrators recognise that "taking care of data" is a task for all.

A full list of recommendations with explanations is included in the final section of this document and summarised below:

A **NWO grants should include funds ring-fenced to support data-stewardship**

B **NWO grant proposals should require a named scientist or organisation with responsibility for data-stewardship throughout the project whose suitability will contribute to the evaluation of the proposal**

C **Generate guidelines outlining NWO's requirements for what data should be archived**

D **Encourage communities to establish partnerships to develop data repositories within framework of NWO guidelines for data-stewardship**

E **Develop/utilise persistent identifiers to enable tracking of data to allow future citation and to track provenance**

F **Don't let ethical/legal constraints in one field prevent action across others**

G **NWO should require all funded projects to cite datasets software and data curators in publications**

H **Develop a peer-administered online resource for sharing scientific software or datasets without writing full journal publications**

Defining Data-Stewardship:

Data-stewardship refers to the practice of storing data with provision for the required software and services to ensure its future accessibility, searching and analysis. Good data-stewardship governance is also concerned with ensuring the quality of stored data and the associated tools. It is therefore critical to ensure reproducibility of published claims, to develop larger populations of similar data to improve statistical analysis and for enabling data-driven research in the future.

Introduction:

As scientific research begins to adjust to the challenges of undertaking discovery in the era of "big-data" we are constantly warned to prepare for the oncoming "data-deluge" or "data-tsunami" that will flood all our lives and leave us drowning or clinging to life-rafts. This apocalyptic vision for the future of science ignores all the incredible opportunities that "big-data" and the continued development of data-driven research has to offer us in terms of discovery and optimisation of our current practices. More importantly, eScience (*enhanced* Science) approaches will allow us to undertake research projects never before possible. The excitement we should have as scientists from the promise of data-mining and data-driven

simulations should far outweigh our data-engulfing fears for the future. The trail-blazers and early adopters, both in academia and industry, are already benefitting from the wealth of data now at their finger tips, so why the general fear and apprehension?

The answer is that without strong data-stewardship practices ensuring access and management of the data that is generated from remote sensors, computer simulations and automated data-generation approaches, even the most IT savvy scientists will be unable to realise the potential of data-intensive research. Good data-stewardship, with its potential to enable transformative scientific IT approaches and stimulate data-driven research is the universal challenge of modern science.

Netherlands eScience Center:

The Netherlands eScience Center (NLeSC) supports and reinforces multidisciplinary and data-intensive research through creative and innovative use of ICT in all its manifestations. To stimulate this enhanced Science (eScience) NLeSC works as a network organization focused on collaboration, with the aim to change scientific practice by making large-scale data analysis possible across multiple disciplines. eScience is typified by experiments to identify subtle patterns in huge (often heterogenous) datasets or the undertaking of data-driven simulations with systematically varied parameters and inputs.

NLeSC stimulates creative data-driven research across all scientific disciplines by developing and applying enabling eScience tools and promoting knowledge-based collaboration between cross-disciplinary researchers.

The establishment of the National eScience Centre is an important step towards the goal of the Dutch government to coordinate data-intensive research in the Netherlands. SURF, the Dutch higher education and research partnership for ICT- and NWO – the country's principal science funding body – have combined their expertise to realize this goal by founding NLeSC and therefore creating an effective bridge between science and ICT.

Netherlands eScience Center Integrators

The work of the eScience Center is made possible by top researchers (mostly professors from around the Netherlands) from a variety of disciplines that have wide experience and understanding of the possibilities of eScience in their domains. These "eScience integrators", representing a defined eScience priority area, play an essential role in constructing an underlying coherent eScience architecture and strategy.

Netherlands eScience Center Scientific Advisory Committee

The eScience Center has an independent scientific advisory committe (eSAC) whose composition, tasks, and responsibilities include encouraging a new and open form of scientific endeavor and building bridges between NLeSC and its applications in business and industry.

Netherlands eScience Center Management Team

The eScience center's management team is responsible for its operation, scientific direction and delivery as well as its technology portfolio, business development, communication and community building.

Background to this document:

In November 2011, NWO chairman Jos Engelen and SURF chairman Amandus Lundqvist gave a joint interview for ScienceGuide.nl entitled "Time is Ripe for eScience" in which they outlined their joint vision for Dutch eScience and data-driven research.

As a response to this interview, the Netherlands eScience Center Integrators, Scientific Advisory Committee and Management Team prepared a joint response which was later discussed directly with Jos Engelen and Amandus Lundqvist. Key areas covered in this discussion included:

1- Provision for data-handling in all funded projects

Data-intensive science is not to be considered the preserve of certain projects or disciplines, but should become a fundamental requirement within research & scholarly pursuits.

2- eScience education

All scientists delivered by higher education institutes should be knowledge workers with a minimum level of computer literacy and able to appreciate the potential of advanced computational methods to impact their research. eScience training should therefore be embedded in all scientific education curricula in the same way as mathematics.

3- Recognising importance of HBO in developing eScience practitioners

To further broaden the Dutch eScience community, Dutch Universities of Applied Sciences (HBO) must continue to be recognised as a key partner. Alumni of curricula such as bioinformatics and (applied) informatics from Universities of Applied Sciences are obvious candidates to become eScience engineers in the application- and solution-oriented approaches of eScience.

4- Career tracks for eScience practitioners and metrics for scientific impact

For some eScience practitioners the traditional peer review measures of scientific output and achievement will remain relevant. However, for others, their accomplishments will only be apparent if we also consider their scientific output in terms of their impact on data stewardship and application development. Unfortunately these measures alone will not support a scientific career at the present time, and we therefore need to find novel ways to ensure eScientists are able to sustain long term professions.

5- Maintenance of infrastructures, open source and open access

Open access to key generic technologies and data repositories for all Dutch researchers from industry and academia will further facilitate the eScience revolution and help achieve the data-intensive research paradigm that we are striving for. Coordinated access, open source and training will extend the toolboxes and repositories available for Dutch scientists in industry and academia, encourage collaboration and prevent the needless development of new applications that provide limited advantages compared to existing tools (reinventing the wheel).

6- Public availability to publicly funded data

Publicly funded data generators should be compelled, within security constraints, to share their data within the Dutch academic community.

7- Cooperation of project calls

Coupling future eScience calls directly with calls from existing scientific and scholarly disciplines will provide an opportunity to reach scientists currently unfamiliar with eScience and send a clear message that all scientific discovery in the future will include an eScience component.

The key outcome of this round-table discussion was recognition from all parties that the introduction of suitable data-stewardship requirements for NWO funded projects has the potential to directly impact on each of these seven key issues. The rationale for this conclusion is that requiring all projects to put in place provision for data-stewardship will ensure the advancement of necessary infrastructures for data management and sharing as well as require the prerequisite training, education, recognition and career development strategies for future data-stewards (often coming from HBO institutes). Future access to public data will be ensured and open source/open access approaches supported as a key component of any sustainable data-stewardship strategy. In order to comply with data-stewardship requirements it is envisaged that collaborations will develop to share the costs and responsibility of long-term data-management, perhaps across institutions or disciplines, depending on preference.

Most importantly, the path towards further embracing data-driven approaches to Dutch research will be transformed as a new culture of data-access is able to mature in combination with a growth in the availability of scientists able to explore and exploit data and access to data-driven research tools.

Developing a cross-stakeholder framework for data-stewardship:

The principles, challenges and recommendations set out in this document have been developed based on internal discussions, the recent joint NLeSC-NWO-SURF meeting and from studying the data-stewardship recommendations of other nations and funding bodies world-wide (in particular from the National Science Foundation of the USA). For any series of data-stewardship recommendations to be successfully implemented and adopted they must have significant cross-disciplinary and cross-institute support. We therefore propose a scheme of presentations, discussions and panels to develop and refine a series of principles, challenges and recommendations we feel underpin the concept of modern data-stewardship. We hope this document can help structure these discussions.

The principles, challenges and recommendations listed below should provide the framework for future discussion on data-stewardship with the goal to develop actionable targets to be recommended to NWO.

Principles of Data-Stewardship:

Underpinning the data-stewardship recommendations should be agreement from all stake-holders that the following eleven principles are at the heart of any data-driven research strategy. Translating these principles into achievable targets, including the challenges and recommendations below, will require continued discussion and revision, but these principles should be widely agreed.

1- ***The sharing of data is at the heart of all scholarly scientific research***

The requirement for sharing data is not new to science, and has always underpinned the scientific process. Data-sharing is required to ensure reproducibility of results and for enabling cross-site and cross-disciplinary research. The emergence of the “Big-Data” challenge has however ensured that challenges in this respect, such as the legal and ethical problems involved in utilising citizen-sensitive data, become more acute. Data-sharing refers not only to data, but also to the technologies and materials needed to verify, replicate and interpret it.

2- ***Data-driven research approaches are fundamental to all modern scientific disciplines and is a positive driver for change and discovery***

We are often warned that the scale of information generation is now so great that science has to adapt or drown in a data-deluge. Whilst true, this analogy fails to appreciate the huge opportunity that “Big-Data” can provide. It will however require better and more wide-spread application of digital technologies if we are to harness the potential benefits to experimental

design, data-analysis and communication that it makes possible. Whilst the precise mechanisms for this change are open for discussion, the belief that data-driven research will provide new opportunities for scientific discovery is beyond question.

3- Data-storage, preservation and curation are fundamental to the scientific process

Turning the challenge of dealing with the so-called data-deluge into a series of opportunities for more rapid scientific discovery will require high-standards of data-management throughout research. This requires high compliance rates and suitable curation to ensure that deposited data retains value.

4- Data-release improves quality

Requiring researchers to share their data, especially when including unique persistent data identifiers, will result in more careful experimentation and analysis by primary researchers.

5- eInfrastructure (including hardware, software and personnel) is a crucial component of any capital investment in science

Funding new scientific endeavour, without sufficient focus on the information-technology systems needed to underpin these investments, results in technology islands and should be considered malpractice.

6- Publicly funded data-generation is a public good and should be publicly accessible

Scientific breakthroughs are already made at the interface of disciplines. Ensuring researchers have access to data from disparate research groups will further stimulate such discoveries.

7- Interoperability of data formats enables data-driven research

Storing heterogeneous data, in specialist or inaccessible formats, with insufficient metadata and in autonomous databases is in contradiction to good data-stewardship practices. Data and its location must be readily identifiable, searchable and accessible. However it is recognised that existing and future data-formats and standards (where they exist) vary across disciplines and universal solutions are unlikely to be developed.

8- Data-stewardship is a multi-stakeholder challenge

Stakeholders should include, but perhaps not be limited to:

- Active researchers currently engaged in data-driven research i.e *eScience Integrators*.
- Active researchers from across a wide selection of disciplines
- Universities and NWO institutes
- NWO and other funding organisations
- Publishers
- Data-driven research specialist companies
- Commercial data-management enterprises

9- Capacity to undertake data-driven research should increase throughout the scientific community

Data-driven research processes, such as cross-type data integration, data-mining, visualisation and analytics must become fundamental to all scientific disciplines and to all researchers in the

way that mathematics and statistics currently is. Whilst not every scientist is a mathematician, it is necessary for all scientists to have some mathematical capability.

10- Expert level data-handlers will still be required to undertake the most complex data-driven research

Despite mathematics being a prerequisite for all scientists, there remains a need for domain-experts able to drive forward advances in this field in the same way as eScience experts will be required to continue driving forward this scientific discipline.

11- A "one-size-fits-all" approach to data-stewardship will prove too inflexible

We have to recognise that whilst shared rules for data-stewardship should be defined, the approaches and methods to satisfy these rules are likely to differ across disciplines and domains.

12- Storing all data may not be necessary if easier to re-measure

In some disciplines, storing data will prove more difficult and less efficient than its re-measurement. This approach still requires provision for storing the detailed parameters under which the measurements are traceable and reproducible.

13- Completing NWO project proposals should encourage researchers to ask themselves "What am I doing with my data?"

All researchers should consider themselves to be data-stewards of their projects even if they do not have this formal responsibility.

The challenges to achieving the ten principles:

Below is a list of challenges to the implementation data-driven research that still require addressing:

1- ***Should data be managed across discipline or across location?***
2- ***Who actually owns the data?***
3- ***Does all data have equal value, does all data need to be saved and should it be saved forever?***
4- ***Who is going to manage the data?***
5- ***When should data be released?***
6- ***How do we ethically and legally manage sensitive citizen data?***
7- ***How do we ensure the data-generator receives credit over time?***
8- ***Will the data-generators get credit from the work of the data-aggregators?***
9- ***Is there value for one country implementing data-stewardship rules unilaterally in a world dominated by international collaboration?***
10- ***How do we begin to standardize/harmonize data-formats, code or data-definitions?***
11- ***How do we ensure we educate sufficient data experts of the future?***
12- ***How do we provide data scientists with sustainable career paths?***
13- ***How do we ensure all scientists receive basic data-stewardship training?***
14- ***Will requirements for including data-analysis tools along with data prevent researchers using commercial solutions even when desirable?***
15- ***How do we objectively measure software quality and what quality is actually needed?***

Recommendations:

A **NWO grants should include funds ring-fenced to support data-stewardship**

The precise percentage of project funds require for data-stewardship requires further investigation, but by making it a prerequisite on all projects it will be possible for domains or more likely institutions to develop repositories at scale. It will also be possible for the private sector to develop business models to provide this service.

B **NWO grant proposals should require a named scientist or organisation with responsibility for data-stewardship throughout the project whose suitability will contribute to the evaluation of the proposal**

This will ensure accountability at the level of individual projects but also ensure the role of data-curator becomes acknowledged. Institutes will need to ensure proper training and education of data-stewards to be successful in future grant applications which will further support this career path.

C **Generate guidelines outlining NWO's requirements for what data should be archived**

Based on the experience of DANS, NWO should appoint a team to develop rules for data-storage, but then allow institutes and communities of practice the opportunity to develop the partnerships they need to comply with these requirements.

D **Encourage communities to establish partnerships to develop data repositories within framework of NWO guidelines for data-stewardship**

As above.

E **Develop/utilise persistent identifiers to enable tracking of data to allow future citation and to track provenance**

This technology, already being developed by DANS, will guarantee scientists remain associated with their data into perpetuity and ensure maximum care of data is taken before deposition and that future citations are correctly awarded

F **Don't let ethical/legal constraints in one field prevent action across others**

In a number of research areas, particularly in biomedicine and social sciences, issues related to the ethicacy and legality of data-sharing patient and citizen data are often correctly raised. It will be the task of researchers in these fields to eventually find the careful balance needed between open data access and protection of individual citizens. However, whilst these discussions continue, they should not be used as an excuse to delay data-stewardship activities in other areas.

G **NWO should require all funded projects to cite datasets software and data curators in publications**

In this new data-driven world it may be the case that the scientific contribution of software engineers, developing tools to enhance discovery, may be of equal or greater relevance than a traditional researcher conducting very narrow research with little interest outside his own sub-domain, but for whom publication is possible.

H **Develop a peer-administered online resource for sharing scientific software or datasets without writing full journal publications**

Such a resource would allow researchers to be cited (by providing a citation) but also record how often their software is downloaded etc. This will support the careers of data-scientists. Researchers utilising tools or datasets would be able to leave constructive feedback, initiate collaborations etc.

Authors:

Netherlands eScience Center Management Team:

- Jacob de Vlieg & RUN
- René van Schaik
- Patrick Aerts
- Scott Lusher (Contact) & RUN
- Frank Sienstra

Netherlands eScience Center Integrators:

- Barend Mons NBIC & LU
- Nic van de Giesen TUD
- Marco de Vos Astron
- Scott Lusher NLeSC & RUN
- Paul Tiesinga RUN
- Jeff Templon Nikhef
- Henk Dijkstra UU
- Willem Bouten UvA
- Henri Bal VU
- Antal van den Bosch RUN