Data Archiving and Networked Services

DANS

A Conceptual Approach to Data Stewardship and Software Sustainability

Scientists in charge, with a little help from their friends

Patrick Aerts (NLeSC)
Peter Doorn (DANS)

>>>

1 Contents

2 Abstract

Around the turn of the Millennium the term *Big Data* got introduced[1] in the science domain while independently also the construction of the Large Hydron Collider was initiated by CERN. From that moment on *data* started to become a significant element in the equations of science, impacting the research processes and questions, as well as the ICT-industry and all other business onwards.

Software has been around for 65 years now, but it is not ready for retirement. Influenced by the data stewardship discussions these days also attention is getting raised for software maintenance policies[2]. But data stewardship is also yet in its infancy. Available are many tools, secure storage, ideas and much drive to do something about data stewardship and software sustainability. What is needed now is improved coherence. So we present a *conceptual approach to the issues at stake*, from which not only the responsibilities and acting parties can be deducted, but also a practical implementation of most aspects of data management and software sustainability can be derived. The goal is *user involvement* and awareness in order to accelerate science by making it more transparent ("open").

The scope of this document is threefold:

- Stress the intrinsic coherence of software and data and thus of software sustainability and data stewardship;
- Design a macro framework to address the responsibilities regarding software sustainability and data stewardship for different stakeholders;
- Explicitly denote Software and Data sets as VALUE OBJECTS, allowing their positioning in the economy.

An important goal of the approach is to make scholars and scientists aware and involve them in the issue. The awareness can be gained by addressing the scientific communities (disciplines) and to ask them to write their own scenarios and protocols for data management and software sustainability through official publications, for instance in scholarly journals for later reference. Once well-established, these procedures will hardly put any extra burden on the science process workflow, but need to be spelled, documented and shared at least once.

The "Open" and "FAIR" movements are more concerned with data than with software. The approach proposed here involves making scientists aware that their data and software are *value objects* that deserve a certain minimum level of care.

[1] Bob Bishop, at the time CEO and Chairman of SGI, used the term from 1999 onwards, to refer to new science directions opening up by enabling Big Data handling, mainly by providing big memory systems and fast IO capabilities.

[2] In the recent "Guidelines on Open Access to Scientific Publications and Research Data in Horizon 2020 " dated 30/10/15 the Directorate-General for Research & Innovation writes: "At the same time, projects should provide information via the chosen repository about tools and instruments at the disposal of the beneficiaries and necessary for validating the results, for instance specialized software or software code, algorithms, analysis protocols, etc. Where possible, they should provide the tools and instruments themselves."

3 Software and Data: coherence and difference

3.1 Coherence and differences

Data and software are intrinsically connected. In fact data are completely inaccessible without software, unless the data takes the form of printed matter. This fact, however, is quite underappreciated. In addition both software and data undergo a lifecycle of collecting/creation via extensions/updates and upgrades to usage/processing. Both data and software are by nature *value objects*. And finally software code is ultimately also a kind of dataset – be it of executable nature.

The term value object is explicitly introduced to get across that data sets and software represent a certain value and should be treated with this representation. The value itself may a priori yet be undefined.

Data stewardship and software sustainability are distinguished mostly by the notion that data need to be kept *as is* while software needs to be maintained in order to remain useful for future purposes. There are exceptions to both statements: there are also volatile and dynamic data (websites that are generated on the fly for example) and one may also want to keep software *as is*, for instance because it is considered to be an element of cultural heritage.

The conclusion is that it is essential for the future use and re-use of data to process and manage data and software on equal footing, policy-wise and practically, up to the level where data and software need to be distinguished because of their differences.

3.2 Software appreciation and awareness

The underappreciation of software as an indispensable means to access and use data is attributable to the implicit nature of software. Just like laptops, PCs and other personal devices such as iPads are bought with most of the basic software (such as operating system and office-suites) on board, it is assumed that most software needed to read, handle and otherwise use data is on board devices.

In practice this may have been the case when conservation of data mostly concerned data in the form of *documents*, articles and archives serving those type of objects, but it can no longer be the case when addressing *data or digital objects of general nature* (for example resulting from measurements, simulations, streaming data from sensors, etc.). It is urgently necessary to activate the awareness that this is in fact the case, in particular if these data are to be provided with publications. These would be inaccessible without the proper software.

In this document Software Sustainability and Data Stewardship are therefore handled on equal footing, except where the differences between the two require otherwise.

4 A conceptual model for Data Stewardship and Software Sustainability

4.1 Introduction

In this conceptual model for Data Stewardship and Software Sustainability, we put the scientists in charge, with a little help from their friends. The concept as such introduces a break down of the total challenge (to get feasible and well accepted procedures and processes in place) into components, coinciding with the identified interests and responsibilities of three distinguished stakeholder groups. The goal of this breakdown is to give handles to actors and actions and place them in a macro context to make their roles apparent. The basic idea and the intent of the approach are to actively involve the researchers in setting up a structure of well documented descriptions for behavior, conduct and processes that will lead to secure data and software for future use, re-use, heritage, inspection, openness and the like.

4.2 Researchers in charge

Data and software are assets developed by scientists in the context of their research activities. Scientists are the conceivers and creators of these value objects so they should be in charge of their "wellbeing". Yet some basic procedural behavior can add tremendous value to these assets and minimize cost for future maintenance. Once well-established, these procedures will hardly put any extra burden on the science process workflow, but need to be spelled, documented and shared at least once. For this the science disciplines need to resort to their peers in the field to compile these procedures and publish them as reference for the community. To this end they will be provided with templates, best practices, advice on standards and alternatives and where applicable with existing legislation and agreements. But before we dive into the details, we provide the bigger picture in which these activities have their natural role.

4.3 The conceptual approach

First we start with a graphical representation of the breakdown into stakeholder categories, their responsibilities and roles and the domain of their actions. The reason why this breakdown is supportive to the goal (namely user involvement) is that the stakeholders have been approaching the topics data stewardship and software sustainability for some time now, from *their* perspective without proper reference to the role of the scientists and other stakeholders. Yet the support of all stakeholders, scientists included, is indispensable to achieve the ends.

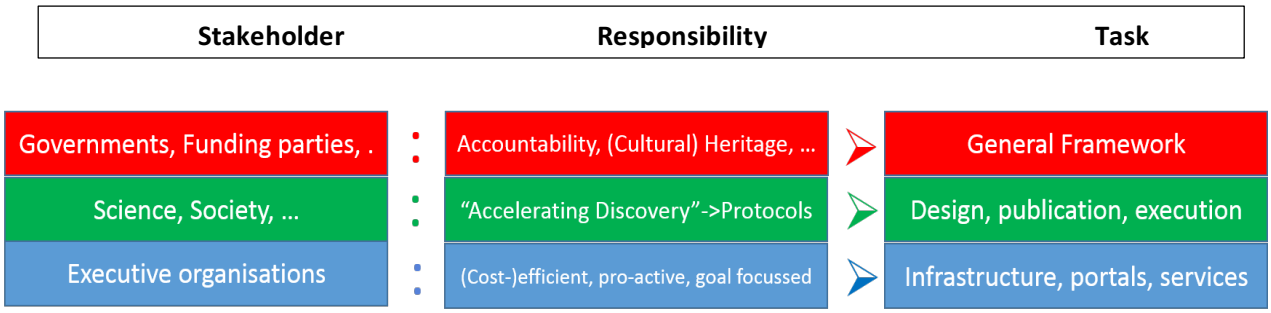


Figure 1 Overview categories of stakeholders and their roles

4.4 Stakeholders

For the sake of the concept, three top level groups of stakeholders are identified:

- Governments, research organisations and granting organisations;
- Science and society;
- Executive and other parties.

The roles and responsibilities of these stakeholders will next be elaborated.

4.4.1 Governments, research organisations and granting organisations

The interest of this category of parties concerns the following aspects:

- Being accountable for the spending of budgets, taking into account the recognition of the efforts by scientists regarding data and software;
- Being accountable for the way in which allocated funds are actually spend, including the verifiability of research results;
- Expediency of the allocated funds, including the integrity of the processes that are being applied and maintained;
- Act on behalf of society to serve the general interest, including economy and cultural heritage.

This means that this category of stakeholders sets frameworks, creates boundary conditions, considers costs and benefits and acts to achieve societal goals –in this case regarding data stewardship and software sustainability. To this end data and software are to be considered value-objects, which helps to account for their well keeping, sustainability and valorization. While setting these frameworks also attention is to be paid to the recognition of the efforts involved in generating and/or collecting data and the creation of unique software.

4.4.2 Science and Society

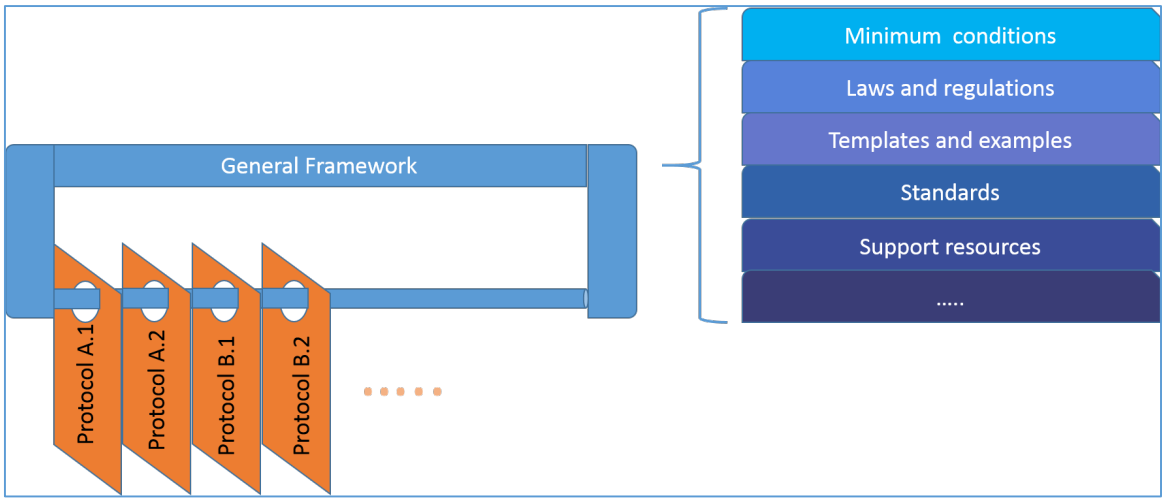
Society, business included, and science have an interest in accelerating innovations that may come from the science domain, by broadening, deepening or speeding up the pace of research. This interest may be of economical nature but also concern the public physical and mental wellbeing. These accelerations can to a high degree be supported by the increased availability of research results, the open access to these results, the communication about this, the re-usability of the results and the underlying materials, including data and software. Within the frameworks set out by governments, research organizations and funders, science itself is responsible for shaping the data and software landscape, in order to maximize re-use and achieving the desired acceleration and deepening of the research.

4.4.3 Executive and other parties

These are -very important- parties involved in and providing the infrastructure for research data management and software maintenance, storage and provisioning: service providers, intermediaries, professional software developers, media and more. Their role is to a large extend derived from the requirements set forth by funders, governments, universities and the scientific community or the community at large, to enable what is required by these parties.

These parties need to be involved in the discussions on the framework and the protocols where these touch their direct interests and where their knowledge and experiences are required with respect to feasibility, cost levels and other important details. They can also be part of the checks and balances, to see if agreements are lived up to, check if what's supposed to be open is really open, etc. The knowledge on how to do things in practice resides at these parties.

4.5 The concept of a framework with protocols



Graphically depicted, the proposed concept is the following.

Figure 2 Concept of general Framework with Protocols

While the category 1 stakeholders ("governments, funding agencies..") - provided with knowledge and information by the other categories stakeholders, agrees on a basic framework, the category 2 stakeholders ("scientists, …"), the scientific disciplines (or sub-disciplines if so required) will start a process of defining so called *protocols* that describe the way data and software will be treated during and after project end and *publish those protocols* as scientific papers for future reference and re-use. The process is to be managed by the category 3 stakeholders (the executive organizations) may go like described below.

It is important to note that in this process the researchers themselves as a collective (their discipline) set their own rules with no more overhead then minimally required to achieve the goal.

4.5.1 The Framework

The Framework consists of the following elements:

- Minimal conditions applicable to Protocols for Software Sustainability and Data Stewardship that are to be created by experts from the various scientific disciplines;
- Existing legal directives and regulations imposed by national law, European regulation and laws and other bounding rules applicable to software sustainability and data stewardship;
- Templates and models to guide and help the disciplines to set up their protocols;

- References to standards in Research Data Management or Software creation, maintenance and management (comply or explain);
- References to "OPEN" and "FAIR" (comply or explain);
- Other resources for support, including best practices and already published Protocols.

Although the Framework is to be set by the category 1 stakeholders (the governments and funding organizations), they will need input from the scientific community and support organizations to be to the point, factually correct and practical.

Next the Framework is made official by agreement among the category 1 stakeholders at large. Ideally this would involve stakeholders globally, but for Europe one could think of Science Europe, DG Research etc. This could be achieved by agreeing on a set of basic principles of which this framework is part, similar to the Concordat[3] as was recently undersigned in the UK. The final step in this process is to set up, together with the scientific communities, expert groups that take the task to create the required protocols.

The Concordat mentioned could for this very purpose look like the following:

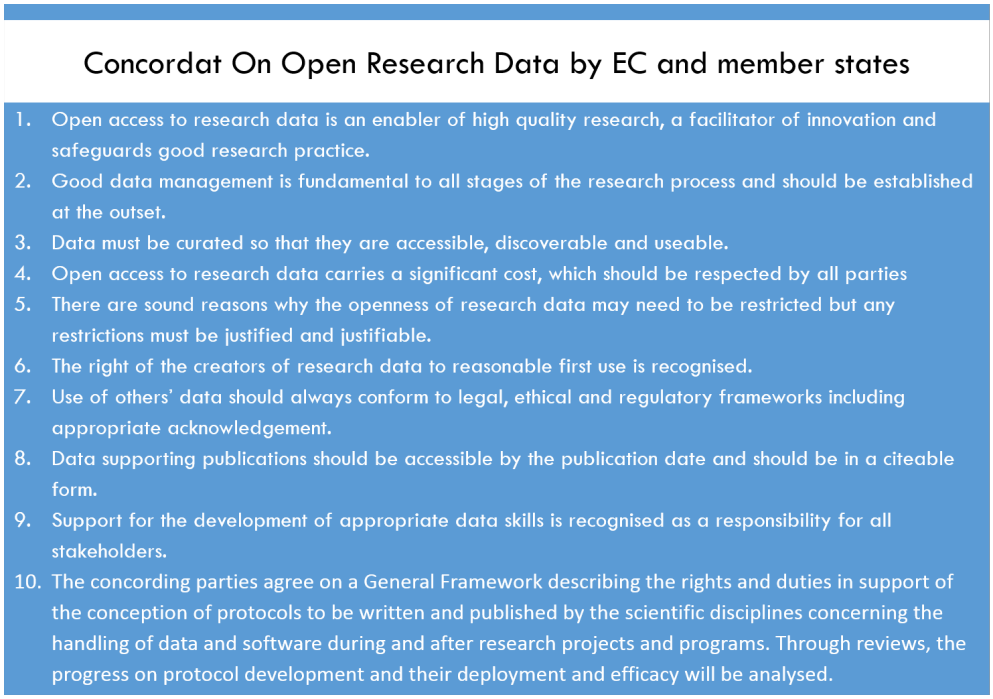


Figure 3 Concordat on Data Stewardship, Software Sustainability and Open Access, based on[2]

The Framework is not to be considered as a top-down directive to enforce actions by scientists. Rather the Framework itself is a minimum condition to get some coherence of the many actions that are taking off to manage research data and software. It makes clear which responsibilities best reside where and gives freedom and encouragement to scientists to design their own practices that suite

[3] Concordat by a multi-stakeholder group, draft version 10 can be found at http://www.rcuk.ac.uk/RCUKprod/assets/documents/documents/ConcordatOpenResearchData.pdf and was open for discussion until September 28 2015. The heading principles from this draft are taken as a lead for this proposal, to which some extensions are included for the sake of the proposed Framework Concept.

their own interests best as well as those of society, while avoiding re-inventing the wheel. Because much is already somewhere out there, but not yet as part of a bigger picture: universities are taking actions, often around acquired data management systems (physical environments with a particular software solution), RDA is doing much good work on standards and community forming, funding agencies are requiring data management plans in grant applications, European research infrastructures (ESFRI Projects and the like) have to conceive data management plans, etc.

4.5.2 The protocols

Once – as a result of the actions leading to the Framework – expert groups are formed for and from disciplines (or sub-disciplines or even more refined communities) and supported by the executive parties, the expert groups start writing one or more protocols for coping with data stewardship and software sustainability, during and after the end of a project. The protocols describe matters such as version control, documentation and transparency and future publication when writing software, or matters such as source documentation, formatting data, meta data, future storage, acknowledgement of the awareness that the data need to be archived and otherwise managed somewhere by some party to be named, etc. during and after a project ends. These protocols, once checked against the (minimum) requirements set forth in the Framework are then to be published, preferably as scientific paper, for openness and future reference. Depending on the particular situation one protocol may suffice for a collection of data types and/or (sub-) disciplines or solutions per project or type of data or type of software may have to be formulated and published. Of course the fewer protocols the more uniformity and lesser work, but this will have to be discovered along the way.

Examples of this process can be found in archeology, where the process of searching, finding and storing objects of (pre-)historic and cultural relevance is already well regulated and described. Also in the medical domain, conducting experiments is highly regulated through scientifically published protocols.

5 Software Sustainability

5.1 Software Seal of approval

Software Sustainability requires not only criteria for service levels, but also a positive stimulus for good behavior. In addition future users of software want insight in the quality and trustworthiness of the software. An example of such a stimulus from the data archive world is the introduction of a Software Seal of Approval. An international committee already certifies data archives with a data seal of approval[4], based on measurable criteria. That model can be taken as an example. A first step is to formulate and agree on the very criteria to set for software sustainability. A Software Seal of Approval is to be based on the process that has been applied and implemented for the proper future maintainability of software rather than on the (scientific) quality or functionality, in order to avoid discussions not related to the sustainability of the software.

5.2 European Software Sustainability Organization

[4] See Datasealofapproval.org for details

The maintenance of software is no easy task. Maintaining all software resulting after project end is practically next to impossible. But taking care of future maintainability during software creation certainly helps, and involving the community from which the software originated is prerequisite for any level of success. So the focus policy-wise should point to self-support from the onset, with exceptions under circumstances.

The UK's Software Sustainability Institute (SSI) provides an excellent example of what can be achieved along these lines with limited resources. A physical institute as central part in a distributed consortium of participating institutes (universities), SSI provides of advice, courses, materials, guides, best practices. Internationally, the Netherlands and Germany are considering implementing similar concepts in association with the UK's SSI. This as yet embryonic European initiative deserves a serious European follow up through a more robust European Software Sustainability Organization, which consists of a network of national SSI-like organizations.

6 Proposed next steps for implementation of the Framework concept

The implementation of the concept involves two steps:

Getting the stakeholders of category 1 (governments, funding agencies, …) to agree on the minimum conditions and perhaps the basic model of templates. Which parties should be addressed will depend on country or level of scope (national, European, global). To get to a set of minimum conditions and the rest of the information, experts from the executive organizations and preferably also the scientific community are needed to get a start, if only because there should be a guaranteed link to the practical world. Things should be reasonably doable and yet be state-of-the-art. This kind of knowledge may be expected available in the executive organizations. As an estimate the agreement on the top-level framework's minimum conditions, etc. should be achievable in about eight months.

Next one could select one or two (sub-)disciplines to start with: select members of the community at hand with sufficient insight in the matter and accepted expertise in the field and form a small committee, completed with technical experts from the executive organizations and start from the very templates, flow charts, check lists available. This process may take a few months to compile a draft, a few more for external consultation and a few for acceptance for publication.

After the initial pioneering, gradually all disciplines at any granularity as required, can have its set of protocols finished. An *a priori* estimate for the whole science domain this process should take about three years to finish. Of course it should be kept in mind, that the process should not be static and procedures should be put in place for updating the protocols and possibly design new ones for better performance or as a response to new formats, standards, workflows, etc.

7 Glossary and abbreviations

CERN	Centre Européenne pour la Recherche Nucléaire (home.cern)
DANS	Data Archiving and Networked Services, http://dans.knaw.nl
Data management	Concerns the development, execution and supervision of plans, policies and programs with the goal to keep, protect and enlarge the value of data and allow their delivery. (Free after Wikipedia)
Data Stewardship	For the purpose of this document data stewardship is the thoughtful and careful handling of data focused on sustainability, re-usability and exchange of data during and after (research)projects. This term to be changed to include the protocolled handling of data. But such protocols are not yet in place.
FAIR	Acronym denoting the principles for Findability, Accessibility, Interoperability and Reusability as established during the Lorentz Workshop in Leiden, The Netherlands, January 2014, named: Jointly designing a Data FAIRport. See: http://datafairport.org/
ICT	Information and Communication Technology
KNAW	Koninklijke Nederlandse Academie van Wetenschappen, Royal Netherlands Academy of Sciences www.knaw.nl
NLeSC	Netherlands eScience Center, founded by NWO and SURF, http://www.esciencecenter.nl
NWO	Netherlands Organisation for Scientific Research, http://www.nwo.nl
OPEN	reference to the collection of movements that use the term "open" as central to that activity (like Open Science)
Open Data	A subset of all data is called Open Data if those data are publicly available without further restrictions. Restrictions can encompass licenses, use policies in rules to protect privacy. Other, more extensive, definitions may be used elsewhere.
Open Science	Open Science is a not yet strongly defined overarching term referring to the various "open" developments, such as open data, open research, open access, open source. Open Science presently is a movement the scope of which can only be appreciated in due course. (Other definitions may be used elsewhere)
Protocol	A public and referenceable document describing the formal procedures.
RDA	Research Data Alliance (RDA) builds the social and technical bridges that enable open sharing of data. https://rd-alliance.org/
Software Sustainability	Software Sustainability is a policy concerning coding practices for re-usability, verifiability and maintainability of computer codes and tools (including "apps") and the systematics of access and up-to-date keeping of software for later (re-)use.
SSA (SSoA)	Software Seal of Approval, a predicate/certification to be given to software that satisfies certain minimum conditions.
SSI	Software Sustainability Institute (UK), http://www.software.ac.uk/

SSO	Software Sustainability Organization, a prospect European organization carried by national initiatives with goals as defined by the SSI.

Nowadays many tools, secure storage, ideas and much drive are available to do something about data stewardship and software sustainability. What is needed now is improved coherence. We present a conceptual approach to the issues at stake, from which not only the responsibilities and acting parties can be deducted, but also a practical implementation of most aspects of data management and software sustainability can be derived. The goal is user involvement and awareness in order to accelerate science by making it more transparent ("open"). Please contact us for further information.

The Netherlands eScience Center (NLeSC)

The Netherlands eScience Center (NLeSC) is the national hub for the development and application of domain overarching software and methods for the scientific community. NLeSC is a joint initiative of the Dutch national research council (NWO) and the Dutch organisation for ICT in higher education and research (SURF).

Data Archiving and Networked Services (DANS)

DANS promotes sustained access to digital research data. For this, DANS encourages scientific researchers to archive and reuse data in a sustained form, for instance via the online archiving system EASY (easy.dans.knaw.nl) and DataverseNL (dataverse.nl). With NARCIS (narcis.nl), DANS also provides access to thousands of scientific datasets, publications and other research information in the Netherlands. The institute furthermore provides training and consultancy and carries out research on sustained access to digital information. Driven by data, DANS ensures the further improvement of access to digital research data with its services and participation in (inter)national projects and networks. Please visit dans.knaw.nl/en for more information and contact details.

Data Archiving and Networked Services (DANS)
Postbus 93067 | 2509 AB Den Haag
Anna van Saksenlaan 51 | 2593 HW Den Haag
+31 70 349 44 50
info@dans.knaw.nl | dans.knaw.nl

DANS is an institute of KNAW and NWO

KNAW NWO

Driven by data