EXAMPLE OF A DATA MANAGEMENT POLICY

Copying a data management plan from another project is impossible, but a good *procedure* to make a data management plan may be copied. Such a procedure, made by Rob van Nieuwpoort of the Netherlands eScience Center, is presented below:

<ProjectNameHere> data management planning

The <ProjectNameHere> data management plan is a living document by design. In the first year of the project, we will write a full data management plan. This will subsequently be evaluated and revised yearly. The <ProjectNameHere> Steering Committee will be end-responsible for the data management plan.

<ProjectNameHere> will implement the FAIR guiding principles for scientific data management and stewardship [1], where FAIR stands for four foundational principles: Findability, Accessibility, Interoperability, and Reusability. The FAIR principles enhance the ability of machines to automatically find and use the data, in addition to supporting its reuse by individuals. This for instance means that we will annotate the data with high-quality keywords and metadata (using the Dublin Core Schema, http://dublincore.org/), as well as generate DOIs (Digital Object Identifiers) for data sets. We will make the data available through the well-defined API and web-based user interface of the <SubprojectNameHere>. Finally, the software used to generate and process the data will be made available under the open source Apache 2.0 license whenever possible, if the software is developed in <ProjectNameHere>, thus enabling reproducibility.

By default, we will make all data publicly and openly available. For financial market data, or for social media data, for example, we may not have the right to make the data available to third parties. The use of large-scale real-world data in the social domain comes with natural concerns over privacy, trust, and responsible use. We will therefore strictly comply with the policies of the data sources from which data was collected. When an open data policy is not applicable, we will publish derived, aggregated and/or anonymized data whenever possible. Moreover, if we cannot publish the raw data, we will in all cases publish the queries that generated the data. Together with the full provenance and data versioning supported by <ProjectNameHere> , this guarantees reproducibility of the scientific results. In addition, we will make the analysis that the system performs publicly available, providing added value for policy makers, and stimulating other parties to reuse the data in sensible and responsible ways.

Data will be made available at the latest one year after producing or gathering it, even if scientific publications related to the data are still in progress or pending. The implementation in <SubprojectNameHere> will guarantee that the data are accurate, complete, authentic and reliable at all times. The modular design of the system will enable a concerted backup policy as well, even though the data itself may be distributed. Data will be kept for at least 10 years after capture, if allowed by the policies of the source. One possible partner for data archiving and storage is the <ArchiveNameHere>. However, at this moment, it is unclear if they can cope with the volumes produced by <ProjectNameHere>. Other parties may be <Archive2NameHere> or commercial providers. In the first year of the project, a choice for one or more providers will be made. Privacy, security and ethical aspects, as well as the ability to handle large real-time volume data will play key roles here.

[1] The FAIR Guiding Principles for scientific data management and stewardship. Mark D. Wilkinson *et al.*, Nature publishing group, Scientific Data 3, No 160018 doi:10.1038/sdata.2016.18, March 2016