Example answers to the 3 areas covered by the DMP exercises.
NB These are not THE answers but suggestions as a "Train the Trainer" exercise. Please feel free to disagree and/or update

File management **Data Backup** **Data sharing**

Project Name Drosophila Genetics - BBSRC Example

Description This project will investigate the role of Polo kinase in metaphase to anaphase transition in Drosophila melanogaster.

Funder Biotechnology and Biological Sciences Research Council **Institution** University of Glasgow

Data areas and data types
Outline the volume, type and content of data that will be generated e.g. experimental measurements, models, records and images

This project will generate three main types of raw data.

1. Images from transmitted-light microscopy of giemsa-stained squashed larval brains.

2. Images from confocal microscopy of immunostained whole-mounted larval brains.

3. Western blot data.

Measurements and quantification of the images will then be recorded in spreadsheets.

Micrograph data is expected to total between 100GB and 1TB over the course of the project.

Scanned images of western blots are expected to total around 1GB over the course of the project.

Other derived data (measurements and quantifications) are not expected to exceed 10MB.

Standards and metadata
Outline the standards and methodologies that will be adopted for data collection and management, and why these have been selected

All samples on which data are collected will be prepared according to published standard protocols in the field. All microscopes used for sample examination are serviced and recalibrated regularly. All Drosophila lines used in experiments are checked periodically for phenotypic markers. Drosophila are maintained in live culture according to standard methods in the field.

Files will be named according to a pre-agreed convention. The dataset will be accompanied by a README file which will describe the directory hierarchy and filenaming convention.

Each directory will contain an INFO.txt file describing the experimental protocol used in that experiment. It will also record any deviations from the protocol and other useful contextual information.

Microscope images capture and store a range of metadata (field size, magnification, lens phase, zoom, gain, pinhole diameter etc) with each image.

This should allow the data to be understood by other members of our research group and add contextual value to the dataset should it be reused in the future.

Relationship to other data

State the relationship to other data available in public repositories

This dataset will provide a novel characterisation of Drosophila Polo kinase mutants documented in the Flybase database. To the best of my knowledge, no other study has perturbed the metaphase to anaphase transition in these mutants, then examined the

phenotypes seen in mitosis.

Secondary Use
Outline the further intended and/or foreseeable research uses for the completed dataset(s)

The confocal and transmitted light images generated in this work may well be of use in the future. It is entirely possible that another study would want to measure a different aspect of mitosis in Drosophila (both the wild-type controls and the mutants) treated as per the protocols in this study.

I cannot see the western blot data being of future use.

Methods for data sharing
Outline the planned mechanisms for making these data available, e.g. through deposition in existing public databases or on request, including access mechanisms where appropriate

Datasets from this work which underpin a publication will be deposited in Enlighten: Research Data, the University of Glasgow's institutional data repository, and made public at the time of publication. Data in the repository will be stored in accordance with funder and University data policies. Files deposited in Enlighten: Research Data will be given a Digital Object Identifier (DOI) and the associated metadata will be listed in the University of Glasgow Research Data Registry and the DataCite metadata store. The retention schedule for data in Enlighten: Research Data will be 10 years from date of deposition in the first instance, with extensions applied to datasets which are

subsequently accessed. This complies with both University of Glasgow guidance and funder policies.

Enlighten: Research Data is backed by commercial digital storage with is audited on a twice- yearly basis for compliance with the ISO27001 Information Security Management standard.

The DOI issued to datasets in the repository can be included as part of a data citation in publications, allowing the datasets underpinning a publication to be identified and accessed. DOIs will also be linked with appropriate records in Enlighten: Publications, the University's publication repository, to enhance visibility of datasets.

Metadata about datasets held in the University Registry will be publicly searchable and discoverable and will indicate how and on what terms the dataset can be accessed.

Information about datasets from the Registry will be displayed on researcher profile pages on the University of Glasgow webpages which will also increase the visibility of the datasets.

Proprietary data
Outline any restrictions on data sharing due to the need to protect proprietary or patentable data

It is not anticipated that this study will generate any patentable data or proprietary data which would have to be protected.

Timeframes
State the timescales for public release of data
Data will be made available at the point of publication of the associated paper or publication.

Formats

This document was generated by DMPonline (http://dmponline.dcc.ac.uk) 2 of 3

State the format of the final dataset

Images will be stored as .tif
Data in spreadsheets will be stored as .csv
Data in freetext documents will be stored as .txt.
These formats are platform agnostic and should support future access and reuse.

Any data which has to be stored in a proprietary format will have the necessary software (including version number) noted in the associated INFO.txt file.

File management **Data Backup** **Data sharing**

Making the final cut: dissecting cytokinesis-signalling pathways in bloodstream form Trypanosoma brucei

DATA MANAGEMENT PLAN

0. Proposal name

Making the final cut: dissecting cytokinesis-signalling pathways in Trypanosoma brucei

1. Description of the data

1.1 Type of study

Molecular, biochemical and imaging techniques will be used to determine the protein kinase pathways that regulate cytokinesis in T. brucei. Kinase assays will be developed, and kinase inhibitors (from the GSK HAT Box) sought, for future exploitation in drug discovery. 1.2 Types of data

Qualitative (e.g. phenotyping data for cell lines), quantitative (e.g. cell counts), mass spectrometry and image data will be generated. Raw data will be analysed and expressed as graphs, tables and annotated images, some of which, it is expected, will be published. 1.3 Format and scale of the data
Data generated will be in various formats and sizes of datasets, all of which will be accessible using common software allowing easy access and long term validity during and after the project, thus facilitating data sharing. The format/types of data include:

i) Cell images e.g. phase and fluorescence, and electron micrographs (~5,000 images over project). Software used includes OpenLab, Softworx and IN Cell Investigator, with data saved as software-specific files e.g liff and lg3 files, as well as generic formats such as jpeg, tiff etc.

ii) mass spectrometry spectra (from <50 samples). MS data will be analysed using Bruker Data Analysis or Thermo Excalibur software (generating xml and raw files) and proteins will be matched to the T. brucei genome dataset using the Matrix Science Mascot search engine. Each LC-MS data file is between 1-2GB.

iii) Cell line phenotyping data including growth curves and DAPI counts (Excel and GraphPadPrism files) and flow cytometry data (FlowJo and jpeg/tiff files) (~200 data sets).

iv)HAT Box screening data and EC50 curves which will be saved as Excel spreadsheets (~20 data sets)

v) enzyme assay data which will be recorded as Excel spreadsheets or jpeg/tiff scans of gels etc (<50 data sets)

vi)DNA and protein sequencing data, which will be saved as abi or MASCOT files (~200 data sets). Plasmid maps will be generated and oligonucleotide sequences stored using software such as Vector NTi or CLC.

2. Data collection / generation

2.1 Methodologies for data collection / generation

Data will be generated by the experiments described in the Case for Support. Some data will be collected manually by project staff e.g. DAPI counts while others will be collected automatically by the equipment used e.g flow cytometry or mass spectrometry profiles. Some data collected automatically e.g. In Cell High Content images will then be analysed manually e.g. to analyse the stage of cytokinesis in 2N2K cells. Data files generated will be labelled appropriately and placed in suitably labelled/organised folders and sub-folders, so that it is obvious which cell line, sample, experiment etc the data originated from, and to allow the project team to easily find files. Different versions of a data subset will be distinguished via a subscript (v1, v2, final etc) attached to the file name.

2.2 Data quality and standards

Standard protocols will be optimised and used to collect data to ensure they are reliable and consistent. All experiments will incorporate appropriate positive and negative controls to ensure validity; biological and technical replicates will be used to assess consistency of data. Project staff will be adequately trained in the techniques they use to ensure they generate high quality data. Data generated and the methods used will be scrutinised in weekly lab meetings to ensure procedures have been carried out correctly, that appropriate controls have been used, that all information is suitably recorded and that therefore there can be a high level of confidence in the data generated. These are routine procedures for conducting high quality research, which allow it to be judged and published in fully peer- reviewed journals, as well as discussed at relevant international conferences.

3. Data management, documentation and curation

3.1 Managing, storing and curating data.

All laboratory experimental details and data will be recorded daily and dated in dedicated laboratory notebooks and scrutinised at weekly lab meetings. Electronic records of the recorded data will be generated (e.g. by scanning or photographing gels/blots, by plotting cell counts and DAPI data in Excel spreadsheets etc) and saved to the University server (which is automatically backed up daily) and also, where appropriate, on the [PI's] lab website. File names/locations will be recorded in lab notebooks to allow electronic records to be linked to the raw data; likewise, file names will have an appropriately descriptive title, including the date the data were generated to allow the corresponding raw data records to be easily found. Details of all oligonucleotides, plasmid constructs and parasite cell lines etc used/generated during this project will be recorded in existing electronic searchable databases located on the University server. Additionally, regular electronic progress reports summarising key data will be compiled by project staff.

3.2 Metadata standards and data documentation

Standard operating procedures will be saved on the University server. Details of SOPs used and any deviations from them will be recorded in laboratory notebooks alongside records of the data generated. Project data will be published, with suitable annotation

and experimental details, in open access journals and/or on the [PI's] lab website to allow it to be easily accessed and used by researchers in the field. Phenotypic information obtained for any protein studied will be released to the field by open access publication and/or by submitting appropriately annotated data to TriTrypDB (www.tritrypdb.org) where they can be easily accessed by the research community. Large research data sets e.g. proteomic data and High Content images will be deposited with suitable annotation in the University of Glasgow Data Registry (Enlighten – Research Data; DataCite standard).

3.3 Data preservation strategy and standards

Electronic data generated and written records will be retained for at least 10 years after the project ends. Open access publication of data and deposition in the Enlighten repository will ensure longevity of the data in the long-term.

4. Data security and confidentiality of potentially disclosive information

4.1 Formal information/data security standards

No human participants will be used in this research. The University of Glasgow is ISO compliant (certification no. ISO27001).

4.2 Main risks to data security

The main risks to data security are loss or damage to laboratory notebooks and loss or corruption of electronic data. Data will be safeguarded by the following measures:

1. a) Data in lab notebooks will, as described above, also be recorded in electronic form that is backed up daily to secure against loss or damage of the notebook.
2. b) Access to electronic data (prior to publication as described above) will be limited to the members of the research group and relevant collaborators via limiting access to shared drives on the University server.
3. c) Access to laboratories and offices are controlled by card access to reduce the likelihood of malicious loss/damage; all computers used in this project will run Standard Staff Desktop, whereby firewalls and antivirus software are automatically upgraded and secure remote access to data is enabled; staff will lock their workstation whenever they are away from it.

5. Data sharing and access

5.1 Suitability for sharing

Yes – data generated from screening HAT Box compounds for effects on cytokinesis, and proteomic data sets are likely to be of interest to the trypanosome community.

5.2 Discovery by potential users of the research data

Project data (with accompanying metadata) will be made available to all interested researchers via open access publication, the Enlighten: Research Data Repository or via the [PI's] lab website. Data in Enlighten: Research Data will be issued with a Digital Object Identifier (DOI). This can be included as part of a data citation in publications, allowing the datasets underpinning a publication to be identified and accessed. DOIs will also be linked with appropriate records in Enlighten: Publications, the University's publication repository, to enhance visibility of datasets.

Metadata about datasets held in the University Registry will be publically searchable and discoverable and will indicate how and on what terms the dataset can be accessed. Information about datasets from the Registry will be displayed on the PI's University of

Glasgow webpage which will also increase the visibility of datasets. The team's approach to data sharing will be outlined in publications and on the lab website.

5.3 Governance of access

The PI will make the decision on when to submit data to Enlighten: Research Data (publically available; see above) and whether to supply research data to a new user.

5.4 The study team's exclusive use of the data

Data will be made available at the time of publication, at the latest. Depending on the nature of the data itself, data may be made available earlier, either on an individual basis to interested researchers and/or potential new collaborators, or (for e.g. negative data) publically via deposition in e.g. Enlighten: Research Data.

5.5 Restrictions or delays to sharing, with planned actions to limit such restrictions

Restrictions to data sharing (e.g. with competitors) will be in place where necessary to ensure novelty for publication or where appropriate, to protect IP but otherwise data will be shared as widely as possible.

5.6 Regulation of responsibilities of users

Where data or resources are provided to an external user ahead of their publication, it will be stipulated that the external user, prior to publishing any work using the data/resources, must consult the PI to determine whether it would be justified for the PI and project team members to be included as authors on that publication.

6. Responsibilities

Along with the PI, members of the project team will have responsibility for study-wide data management, metadata creation, data security and quality assurance of data. Research colleagues within the department will assist with quality assurance by criticising data presented at joint (confidential) lab meetings. The University Research Data Management team will be able to advise on best practice in data management and security.

7. Relevant institutional, departmental or study policies on data sharing and data security

Data Security Policy	IT Security Policies http://www.gla.ac.uk/services/it/informationsecurity/policies Confidential Data Policy http://www.gla.ac.uk/services/it/informationsecurity/confidentialdata Data security issues are also covered to some extent by the UoG's Data Protection policy http://www.gla.ac.uk/services/dpfoioffice/policiesandprocedures/dpa policy Policy on Confidential Data in the UoG http://www.gla.ac.uk/media/media_180727_en.pdf
Data Sharing Policy	Data sharing is covered in the UoG's Code of Good Practice in Research http://www.gla.ac.uk/media/media_227599_en.pdf Data sharing is covered in a number of Ethics related policies. The policies are accessible via the Ethics Committee homepage http://www.gla.ac.uk/research/aims/ourpolicies/ethics

Institutional Information Policy

Other:

The UoG lists all policies relevant to research conduct and interaction with the wider community http://www.gla.ac.uk/research/aims/ourpolicies
Details on relevant RCUK policies and draft UoG EPSRC roadmap available at http://www.gla.ac.uk/services/datamanagement/rdm- atgu

File management **Data Backup** **Data sharing**

Research Ideas and Outcomes 3: e11624 doi: 10.3897/rio.3.e11624

Data Management Plan

Data Management Plan for a Biotechnology and Biological Sciences Research Council (BBSRC) Tools and Resources Development Fund (TRDF) Grant

Laurent Gatto ‡

‡ University of Cambridge, Cambridge, United Kingdom

Corresponding author: Laurent Gatto (lg390@cam.ac.uk) Reviewable v1 Received: 23 Dec 2016 | Published: 05 Jan 2017

Citation: Gatto L (2017) Data Management Plan for a Biotechnology and Biological Sciences Research Council (BBSRC) Tools and Resources Development Fund (TRDF) Grant. Research Ideas and Outcomes 3: e11624. https://doi.org/10.3897/rio.3.e11624

Abstract

Background

This Data Management Plan (DMP) was created for Laurent Gatto's BBSRC Tools and Resources Development Fund award (BB/N023129/1).

New information

The DMP describes the management and sharing of all data and code associated with the grant, including software dissemination and release schedule, source code development and open source licensing, software documentation, reproducible framework and data annotation and dissemination.

Keywords

Spatial proteomics, Bioconductor, machine learning, mass spectrometry, proteomics, software

Products of research

The participants have a long history of successful collaboration and open source development and are fully committed to abiding by the BBSRC's policy on data management. Specific outputs of this project and how they will be made available to the community are listed below.

Software

All software infrastructure and statistical routines developed in this project will be submitted to the Bioconductor project (Huber et al. 2015). We will continue to follow the well established standards for packaging, versioning, documentation, updating and installation. Bioconductor will be the official distribution channel for the software, thus benefiting from existing dissemination infrastructure, support channels, user base and the developer community. In addition to biannual official releases in March and October, tested and documented development versions of the software will also be available through the dedicated Bioconductor development branch.

Source code

We understand the value of open source development practices within the scientific community. The source code of the software will be freely available in code repositories under permissive open source licenses and hosted on the Bioconductor subversion server. In addition, we will continue to use the GitHub social coding infrastructure to facilitate collaboration within the team and promote contributions from the community. The two repositories will be clearly documented (for example by using software versions) to avoid any confusion and kept in sync using dedicate tools such as git-svn. As well as being good practice, open source and collaborative development of our software will enhance the visibility and sustainability of what we produce.

Documentation

All software that will be released as part of this project will be thoroughly documented in multiple ways. Individual functions and data containers will be described in detail to allow users and developers to understand and use them in their own pipelines. In addition, we will produce vignettes, dynamically generated documents that offer a general overview of the functionality of the software and flexibility of the pipelines, advise on how to explore the data and understand the results, information on data preparation and import into the R environment and links to relevant resources. We will also produce educational material that will be broadly distributed independently of the software through workshops and courses to maximise visibility of the software and analysis methodologies and facilitate adoption by new users less familiar with the R/Bioconductor environment and community. In particular, the material for our second workshop dedicated to the analysis and interpretation of spatial proteomics will be made publicly available.

Reproducible framework

End users will gain access to accurate, biologically relevant results and experimental data through existing resources, dedicated data packages and wider databases, and experts interested in the analytical process will gain open access to relevant elements of a key proteomics methodology. The combined distribution of annotated data and well- documented software bundled in analysis scripts will offer users and developers a complete reproducible environment.

Data

While no new data will be generated specifically in the frame of this project, statistically sound (re-)analysis and reliable (re-)interpretation of published or private data will be produced. These data will be made available through multiple existing community resources using established standards and annotated withample meta data. They will be distributed as dedicated R object (in well-established data structures defined in MSnbase

Gatto and Lilley 2011), as used and manipulated through the pRoloc (Gatto et al. 2014) and pRolocGUI software and included in the open pRolocdata (Gatto et al. 2014) data package. All datasets will be thoroughly annotated with meta data to provide users with all necessary details on the origin or manipulation of the data in order to favour and facilitate re-use and reproducibility. Several exporters are already available, to offer these same data as spreadsheets or in the mzTab (Griss et al. 2014) format. When available, raw and identification data will be distributed using the mzML and mzIdentML Proteomics Standards Initiative (PSI) community formats and disseminated through the ProteomeXchange (PX) project (Vizcaíno et al. 2014) and the PRoteomics IDEntifications (PRIDE) resource. We will also distribute the data and results through the online resource SpatialMap.org that we are currently developing, which will enable users to interactively visualise, explore and search the data and annotated results stemming of our state-of-the-art statistical learning pipelines.

The refined and novel protein sub-cellular localisations will be communicated to the wider proteomics community via relevant protein databases and annotation providers like Swiss-Prot, the Gene Ontology Annotation database as well as more specialised resources. The improved localisation information will be distributed with all technical details regarding the analysis and interpretation/evidence, including algorithm specifications and parameters and assignment probabilities.

Data will be made available as soon as it has been quality controlled and converted into usable computational objects. Once validated on various datasets, the algorithms will be included and distributed through the relevant software packages. The multiple sources and formats will be cross-referenced to maximise utility and availability to the research community.

Acknowledgements

The author would like to thank Dr Marta Teperek and Dr Ross Mounce for their enrouragements to publish this DMP, as well as the Research Data Management team at the

University of Cambridge for their efforts in promoting open data and good data management practice.

Grant title

Understanding protein multi- and trans-localisation at the full proteome level

Hosting institution

University of Cambridge

Author contributions

Laurent Gatto wrote the Data Management Plan.

References

• Gatto L, Lilley KS (2011) MSnbase-an R/Bioconductor package for isobaric tagged mass spectrometry data visualization, processing and quantitation. Bioinformatics 28 (2): 288-289. https://doi.org/10.1093/bioinformatics/btr645

• Gatto L, Breckels LM, Wieczorek S, Burger T, Lilley KS (2014) Mass-spectrometry- based spatial proteomics data analysis using pRoloc and pRolocdata. Bioinformatics 30 (9): 1322-1324. https://doi.org/10.1093/bioinformatics/btu013

• Griss J, Jones AR, Sachsenberg T, Walzer M, Gatto L, Hartler J, Thallinger GG, Salek RM, Steinbeck C, Neuhauser N, Cox J, Neumann S, Fan J, Reisinger F, Xu Q-, Toro Nd, Perez-Riverol Y, Ghali F, Bandeira N, Xenarios I, Kohlbacher O, Vizcaino JA, Hermjakob H (2014) The mzTab Data Exchange Format: Communicating Mass- spectrometry-based Proteomics and Metabolomics Experimental Results to a Wider Audience. Molecular & Cellular Proteomics 13 (10): 2765-2775. https://doi.org/10.1074/ mcp.o113.036681

• Huber W, Carey VJ, Gentleman R, Anders S, Carlson M, Carvalho BS, Bravo HC, Davis S, Gatto L, Girke T, Gottardo R, Hahne F, Hansen KD, Irizarry RA, Lawrence M, Love MI, MacDonald J, Obenchain V, Oleś AK, Pagès H, Reyes A, Shannon P, Smyth GK, Tenenbaum D, Waldron L, Morgan M (2015) Orchestrating high-throughput genomic analysis with Bioconductor. Nature methods 12 (2): 115-21. https://doi.org/10.1038/ nmeth.3252

• Vizcaíno JA, Deutsch EW, Wang R, Csordas A, Reisinger F, Ríos D, Dianes JA, Sun Z, Farrah T, Bandeira N, Binz P, Xenarios I, Eisenacher M, Mayer G, Gatto L, Campos A, Chalkley RJ, Kraus H, Albar JP, Martinez-Bartolomé S, Apweiler R, Omenn GS, Martens L, Jones AR, Hermjakob H (2014) ProteomeXchange provides globally coordinated proteomics data submission and dissemination. Nature Biotechnology 32 (3): 223-226. https://doi.org/10.1038/nbt.2839

File management **Data Backup** **Data sharing**

EPSRC Violence study

Data management plan

Existing Data. The current project will extend the existing z-proso study (http://www.cru.ethz.ch/en/projects/z-proso.html) which, by the end of the current project, will include ten waves of survey data. It will also (depending on the success of other funding applications) potentially include brain imaging, gene expression and other biological data (such as substance use levels derived from hair samples) on a subset of 200 participants. Whilst z-proso has rich data over longer time lags, it does not contain any data at the 'day-to-day' level to inform on momentary processes. Combining the newly collected experience sampling data described in the current proposal with the ten waves of existing longitudinal data will be critical to answering key outstanding questions relating to how criminal and aggressive behaviour develops over momentary and developmental timescales. To the best of our knowledge, there are no existing datasets that combine longitudinal data with experience sampling data for the study of crime and aggression. The principal investigators of z-proso have agreed that for our research we can combine the 10 waves of z-proso data with our experience sampling data (see Data access, sharing and re-use agreement). Investigators on the current project will have priority on and responsibility for research projects using the experience sampling data. The newly collected data can be easily integrated into existing z-proso datasets and its collection, processing and storage implemented in accordance with existing z-proso protocols. These data management protocols have ensured the security and fidelity of the existing data across z-proso's lifetime. The use of this existing data source and its integration with the new data from the current project will present no additional difficulties.

New Data. The current project will generate ~3360 new variables yielding a dataset with ~1000 rows by 3361 columns (including participant unique identifiers). The majority of items will use a 5-point Likert scale response format, resulting in ordered-categorical data. The data will be in the form of a comma-delimited ('.csv') file. As the data dimensions are relatively low, no new specialist processing or storage will be required; the data can be accommodated within existing z-proso IT systems at the Swiss Federal Institute of Technology Zurich (ETH), Switzerland with the existing and other to-be-collected data. Data will be collected via participants' smartphones using an application provided by LifeData LLC. Data is then transmitted to and securely stored on a LifeData LLC server during data collection period. As detailed in their Privacy Policy: https://www.lifedatacorp.com/lifedata-privacy-policy/, LifeData LLC will treat this data as confidential and safeguard it with security protections and precautions.

We will download data from these servers in comma-delimited ('.csv') files. Question numbers will form the column headings and columns are populated with participant responses. At the point of download, personal data will be deleted in order to anonymise the dataset. Data will be downloaded at regular intervals during the data collection period to create back-up copies in the event of software failures. These back-ups will be given names such as 'z-proso_ESM_data_BU_DATE.csv' to differentiate them from the final datasets where 'DATE' indicates the date of data download.

The full experience sampling datasets for each burst will be downloaded at the end of the two data collection periods. In these, column headings will be re-labelled manually by the research assistant and checked by the PI. Column headings will distinguish the same variable measured in the first and second experience sampling bursts using the suffixes '_ES1' and '_ES2'. Responses will be stored in character form e.g. 'strongly agree' with numerical recoding performed at the point of data analysis. Once processed, the new data (the 'ESM data') will be stored as datafiles independent of the larger z- proso dataset containing the pre-existing data under filenames such as 'z- proso_ESM_data_master_VX.csv' where 'master' is used to distinguish from other versions of the datasets that will be kept for quality control and back-up purposes. The ESM data or specific variables from it will be merged with other variables from the main dataset via unique participant identifiers contained in all z-proso datasets at the point of data analysis and sharing.

Meta-data will be created and maintained in accordance with the UK Data Service guidance. Basic variable information will be stored in the ESM data dictionary in the form of a comma-delimited ('.csv') file that includes: variable labels (corresponding to the labels in the dataset); items as presented to participants (in German); English translations of items; coding information (e.g. '999'= missing); variable classes (e.g. 'nominal'); and brief explanations of variables. Explanations will include information such as the scale to which an item belongs (plus citation) or, in the case of summed scores or other derived variables, explanations of how they were derived. In addition, details on data collection, cleaning, coding, quality and version control procedures will be provided in a ESM data protocol document. This document will also record the version history of the ESM data. Finally, .doc and .pdf

copies of the questionnaires as they were administered to participants (ESM questionnaire ES1 and ESM questionnaire ES2) will be stored with the data.

Datasets will be named in accordance with the following convention, with filenames taking the form: z-proso_ESM_data_type_VX.csv'. 'Z-proso' identifies the project to which the data belong and 'ESM' identifies the sub-project. This is desirable because z-proso includes a number of sub-projects which are- due to the collective volume of data involved- stored separately from one another and linked only as required. The 'data' term identifies the type of file (to distinguish it from other types such as draft manuscripts, meta-data, policy documents etc.). The 'type' term will vary across datasets in order to distinguish temporary back-up copies created during data collection ('BU'); back-up copies stored long- term for quality control and back-up purposes ('QC'); data variants created for the purposes of sharing for specific projects ('SUB_XXX'), where 'XXX' identifies a short project code recorded in the ESM Data Protocol; and the master dataset ('master'). The 'VX' term identifies the file version. Major versions will be given names such as 'V1, V2, V3' etc.; minor versions names such as 'V1.1', 'V1.2' etc.

Quality Assurance. Pilot studies. Quality assurance will begin with pilot studies conducted before the start date of the proposed research. These will test the items, data collection method and the data management protocol, allowing us to optimise our methodology for the main study. Data will be collected on the acceptability, reliability and validity of items, response rates and times, and participant experiences. At time of writing, one small pilot study with n=20 participants has been completed and a larger n=200 study is underway.

These data will be treated as independent of the data created in the current project and managed according to a separate data management plan.

Manual data entry and coding. The only manual data entry will be re-labelling dataset columns to ensure that labels are intuitive, brief and include no characters that would create difficulties when the data is imported into data analysis programmes. Re-labelling will be completed by the research assistant and checked by the PI.

Data checking. Ten per cent of responses will be selected and manually checked against the data held on the application server by the research assistant to ensure no errors have been introduced in data conversion, download, and column re-labelling. Data will be screened for respondents with large numbers of missing responses, out of range values, and random responding or responding according to response sets. This will use simple functions written and implemented by the PI in R Statistical Software and supplemented by manual checking by the research assistant. a priori protocols for dealing with suspect values will be written into the ESM data protocol. As the newly produced data will be integrated with the existing z-proso data, the same conventions as have been used in the existing datasets will be used to code for missing data, anomalous responses etc. No missing data will be imputed; all missing data treatment will occur at the point of data analysis.

Data maintenance. One master copy of the ESM data will be kept on the z-proso servers at ETH. Editing rights will be restricted to the research assistant, the PI and the z-proso main study PIs with all changes to be authorised by the PI. Overall responsibility for this dataset will be with the PI. Changes will be documented in the ESM data protocol and corresponding updates made to the data dictionary. The consistency of these files will be checked at six-month intervals. Previous versions will be archived on the z-proso servers for reference and as a back-up in case errors are introduced into later datasets. Old versions of ESM data will be discarded in accordance with existing z-proso protocols and documented in the ESM data protocol.

Security and Backup. ESM data will be stored securely in accordance with existing z-proso protocols. Data will be stored on password protected computers located at the Swiss Federal Institute for Technology (ETH) along with the other z-proso scientific datasets and the database containing participant personal details. These PCs are in locked offices in secure University buildings at ETH. The data files will be password protected. The separate files containing participant personal information and the ESM data will be linkable to the participants' data via unique identifiers. The ESM data files will be anonymised, with only unique identifiers given. Only the PI, research assistant and z-proso PIs will have access to the personal information files. Access to the anonymised scientific data will follow the procedures already in place in z-proso; namely the completion of a confidentiality agreement and project proposal to be approved by a project PI. Data will be backed up by saving to an external drive (encrypted) systematically every 2 weeks, after any alterations are made to any files, and after any new data is downloaded. Data on the ETH servers are also backed-up automatically at regular intervals.

Data Sharing – Issues and Solutions. Prior to data deposit, data access and sharing will follow existing z-proso protocols. Files will be shared securely through direct download from the secure z-proso

server accessed via a username and password or shared as password protected files on a USB or by email. At the end of the grant period or on publication (whichever is earlier), the ESM Data and associated meta-data and documentation will be prepared for deposit in the UK Data Service data repository to provide long term open access to the data. This will be completed by the research assistant under the supervision of the PI. GPS data will be deposited only in summary form because individual GPS data could risk identifying participants.

Consent and Anonymity. Expressed written consent will be collected for all participants. The request for consent will make clear that their data may be used and shared with other researchers in anonymised form but that their individual data will not be identifiable in any outputs. All participants will have unique identifiers assigned to them. Only this identifier will appear in the datasets. A separate file with personal details will be kept secure. These files will not be shared with other researchers.

Copyright and Intellectual Property Ownership. Copyright and intellectual property will be held by the PI of the current project.

Responsibilities. The PI of the current project will have overall responsibility for ensuring the integrity and security of data. The research assistant will have responsibility for the day to day management and upkeep of the datasets, meta-data and supporting documentation production. This will be supervised by the PI under the guidance of the mentor and z-proso project PIs (also see Staff Duties).

Appendix: Data access, sharing, and reuse agreement

The project D2M will generate two bursts of experience sampling data ‘D2M data’ that can be combined with existing and to-be-collected z-proso data ‘main data’ to form ‘combined data’. In all cases, the data excludes any identifying participant information. The arrangements for data reuse, access and sharing of these three datasets will be as follows:

 The D2M and z-proso study principal investigators agree to combining the D2M and main data to form the combined data.

 The combined data can be used and accessed by the principal investigators of the D2M and z- proso main study without restriction.

 The D2M and combined data can be shared with other members of the z-proso team and external researchers in accordance with the existing z-proso protocols.

 The D2M data will automatically be subject to existing z-proso data management, access and sharing protocols unless otherwise stated in the D2M Data Management Plan

 The D2M data will be made publicly available on publication or on the end of the D2M grant funding period by deposit in the UK Data Service repository.

 An exception is the GPS data from D2M which can only be shared with external researchers in summary form due to risk of de-anonymization.

 The main data may become publicly available at a later date, in which case the full combined data will be publicly available.