MULTIPLY (H2020-687320)

DATA MANAGEMENT PLAN

DOCUMENT REF:	WP1_D1.2_D2
DELIVERABLE REF:	D1.2
VERSION:	V2
CREATION DATE:	2016-03-22
LAST MODIFIED:	2019-12-28

MULTIPLY Data Management Plan

Name and contact details	Prof. P.M. van Bodegom Institute of Environmental Science, Faculty of Science, Leiden University Einsteinweg 2, 2333 CC Leiden, the Netherlands Tel.: +31 – 71 – 527 7486 Mail: p.m.van.bodegom@cml.leidenuniv.nl
Name of project	MULTIscale SENTINEL land surface information retrieval Platform (MULTIPLY)
Description of your research	The project developed an efficient, fully generic and **fully traceable** platform that uses state-of-the-art **physical radiative transfer models**, within advanced **data assimilation (DA)** concepts, to consistently **acquire**, **interpret** and **produce** a continuous stream of high spatial and temporal resolution estimates of land surface parameters, fully characterized. These inferences on the state of the land surface are the result from the coherent joint interpretation of the observations from the different Sentinels, as well as other 3rd party missions (e.g. ProbaV, Landsat, MODIS).
Project duration	Start: 01-01-2016 End: 31-12-2019
Names of people and their responsibilities for data management	Peter van Bodegom (UL): overall coordination of data management Jose Gomez Dans: documenting sources of data on irradiance in the Photosynthetic Active Radiation (PAR) regime Alex Loew: documenting sources of irradiance data in the passive microwave regime Phil Lewis: documenting, sharing and archiving derived data products on the vegetation and soil state. Jose Gomez Dans: documenting, sharing and archiving of data on disturbance indicators Leon Hauser: documenting, sharing and archiving of data on biodiversity indicators Daniel Kindred: documenting, sharing and archiving of data generated in relation to agricultural case studies within the programme other than derived data products on the vegetation and soil state. Philippe Peylin: documenting, sharing and archiving of data generated in relation to vegetation modelling case studies within the programme other than derived data products on the vegetation and soil state.
Funding body(ies)	European Union, Horizon 2020 research and innovation programme
Grant number	687320
Partner organisations	Ludwig Maximilians Universität München (Germany) University College London (United Kingdom) Brockmann Consult GmbH (Germany) Tartu Observatory (Estonia) Universidad de Alcala (Spain) Assimila Ltd. (United Kingdom) ADAS UK Ltd (United Kingdom) LSCE, Université de Versailles Saint_Quentin-en-Yvelines (France)

About this Data Management Plan

Date written	*18 June 2016*
Date last update	*8 March 2020*
Version	*Version 2.0*

Changes in this version of the Data Management Plan

Component	**Progress / Execution** *Please describe shortly what progress you have made, any questions or issues you have encountered and want to discuss, etc.*
1. Data collection	We added a protocol on collecting field campaign data
2. Data storage and back-up	We updated the data storage protocol (i.e. we only store data that will likely have a wider use)
3. Data documentation	We improved the coupling between the data documentation and the versioning system of our software
4. Data access, sharing and reuse	We added an additional audience for data re-use
5. Data preservation and archiving	We revised the data archiving section to describe how we arranged this through DIAS.

	1. Data collection Describing the data you will be creating/collecting
1.1	**Will the project use existing or third party data?** No ☑ Own / group previous research Academic collaborators Commercial collaborators ☑ Publicly available database / archive Specialist commercial data provider Other (please specify) *Describe shortly provenance, type and format of this data. Are there any restrictions or requirements for use of third party data such as licensing conditions?*
	The project will deal with four types of data: 1. "raw" Earth Observation data e.g. back-scatter reflectance as derived from e.g. Copernicus services and NASA. These data are open-access (or otherwise made accessible to the Multiply project), but owned and archived by the data owners. These use of data from these sources will be documented as a meta-data catalogue within the project, but otherwise this Data Management Plan will not deal with these data. 2. "derived" Earth Observation data, describing i) the vegetation and soil state, ii) posterior estimates on disturbance and biodiversity, and potentially including iii) additional data products as derived within the case studies. 3. Validation data for vegetation and soil variables in as far as derived from field campaigns executed by participants of Multiply, within the context of Multiply. 4. Field data of third parties may be used within the Multiply project and its use is documented as a meta-data catalogue whenever applicable. Data storage and ownership will at all times remain the responsibility of the third parties and hence will not be part of this DMP.
1.2	**How will you collect and/or create your data?** *Please describe shortly. Name any relevant protocols and/or standard in your area of expertise.*
	Ad 2. Data will be derived by using the data retrieval and assimilation platform of Multiply. This platform is a central deliverable of the Multiply project. The protocols for deriving these data will be described in reports documenting this and connected deliverables. The codes will be made open access at Github. Ad 3. Field campaign data will be collected using protocols that are standard within this field. Data collection has taken place based on the international guidelines for sampling plant traits (Cornelissen et al. 2003 AustJBot) and fully complied to these protocols.
1.3	**What tools, instruments, equipment, hardware or software will you use to capture, produce, collect or create the data?** *Please give the names of the tools and state if they are already available. If not, state how you intend to acquire them. If applicable, describe whether you use a paper or electronic lab journal.*
	Ad 2. The data will be created using the data retrieval and assimilation platform of Multiply. The platform will be developed by Multiply and is available as open-access code at Github as part of the Multiply project. Ad 3. Data will be recorded in electronic lab journals before archiving them. Archiving is done as text files for further checks on data quality and data analysis. Upon publication of the validation results, the data will be made available through open-access repositories with a DOI, such as Dryad.
1.4	**Type of collected and/or created data per task** *Note that not all formats are long-lived. For sustainable access you best use the formats*

recommended by data archives, see for examples: *http://datacentrum.3tu.nl/fileadmin/editor_upload/File_formats/Preferred_formats.pdf* *or* *http://www.dans.knaw.nl/en/deposit/information-about-depositing-data/DANSpreferredformatsUK.pdf*

Data stage	Specification of type of research data	Software choice and file format	Data size now	Data size when project is finished
Vegetation state	*Description of the vegetation state, potentially including faPAR, spectral albedo, LAI, equivalent leaf water thickness, leaf chlorophyll and carotenoid concentration, leaf mass per area, leaf optical thickness*	*netCDF, ascii*	*2 GB*	*2 GB*
Soil state	*vegetation gap structure, soil surface vertical roughness, soil moisture*	*netCDF, ascii*	*2 GB*	*2 GB*
Atmospheric state	*Atmospheric correction features*	*netCDF, ascii*	*20 GB*	*20 GB*
Disturbance indicators	*Date of disturbance, type of disturbance (e.g. fire, logging…), change in vegetation parameters before and after the disturbance, long term disturbance tracking.*	*netCDF, ascii*	*1 GB*	*1 GB*
Biodiversity indicators	*Biodiversity indicators*	*netCDF, ascii*	*5 GB*	*5 GB*

2. Data storage and security

Ensuring that all research data are stored securely and backed up or copied regularly during your research

2.1 **Where will you store your data?**
Please describe how safe storage is guaranteed. Specify your method if your data is collected and / or transported in different locations / countries.

☐On university departmental network storage (J:)

X On university personal network storage (P:)

☐In a Virtual Research Environment (Sharepoint)

☐Physical storage (e.g. USB, external hard drive)

X Cloud service (e.g. SURFdrive)

☐Other, namely: …

We identify two stages:
Stage 1: local data storage on the university network storage of the individual partners. This local storage is used for “derived data” that are related to intermediate products and hence are not of direct use to others (see Table 4.2), for ‘’raw’’ data of field campaigns until data have been error-checked and cleaned and for processed data from field campaigns, that will not be archived. All other data are stored in the data archive (stage 2)
Stage 2: data archiving (further elaborated upon in section 5). Within this stage, we aim for a fully transparent and full open-access utility with a DOI to enhance finding and searching the

	archive. The default format for derived data will be netCDF and the default format for field campaign data will be ascii, unless the field campaign deal with spatially explicit information in which a netCDF format will be used. The data archive is directly coupled to the Multiply platform. However, we chose only to archive retrievals that we considered to be of wide use for other users, such as the atmospheric correction fields and retrievals associated to validation sites and retrievals used in the e-learning platform. No other data were stored and instead the software versions and satellite data versions used to derive those data sources were documented as this turned out to be cheaper and faster.
2.2	**Will your data be backed up?** *Please specify shortly for each storage device frequency, location of backups and who is responsible.* *Describe how you can restore your data in the event of data loss and who is responsible.*
	All stage 1 data are backed up within the university network on a daily basis. Stage 2, data within the archives are also backed up by the archive, such as by DIAS
2.3	**Are there any commercialisation, ethical or confidentiality restrictions about handling your data?** *Please specify shortly.*
	☐Contractual obligations ☐Requirements by law : protection of personal data (e.g. privacy law) : specify in 4.1 ☐Requirements by law : copyright, intellectual property : specify in 4.1 ☐Ethical restrictions (e.g. ethical review) : specify in 4.1 ☐Commercial considerations (e.g. patentability) ☐Formal security standards X No requirements ☐Other, namely: ………
	We will make the entire archive open access as a cloud service connected to our platform. Stage 1 data are not publicly available.
2.4	**How will access to the data be managed during the project?** *Please specify for each storage device, from different locations / countries.*
	Each participant within the project is responsible for delivering his/her own data from the university data storage to the centralised repository in due time (see conditions in section 4). Likewise, each participant is responsible for checking that the uploaded data is complete. Also “derived” data produced by predecessors of the Multiply will be made available within this centralised repository. The project coordinator will regularly check the archive for completeness.
2.5	**What are the main risks to data security?** *Please list risks, e.g. accidental deletion, falling into the wrong hands.* *Please describe what would happen if the data get lost or become unusable.*
	Given that all data in each stage are being backed up, we consider the risks minimal.
2.6	**What measures do you take to comply with the security requirements and to mitigate the risks?** *Describe how you can restore your data in the event of data loss and who is responsible.* *If applicable, please describe procedures to ensure personal data are handled confidentially and who is responsible.*

	☐Access restrictions ☐Encryptions ☐Data processing ☐De-identification / Anonymisation X Regular back-ups ☐Master copy stored on university network storage ☐Master copy stored elsewhere ☐Other, namely: …
	During stage 1, this is the responsibility of each individual partner.
2.7	**How do you differentiate between raw and processed data?** *Please explain shortly why you (do not) differentiate.*
	☐I will not differentiate ☐I will create a new file for processed data X I will create a new file for processed data and I will lock raw data ☐Other, namely: …
	Ad 2. The platform will create processed data only (by definition). Data assimilations for the same “raw” dataset run with different settings (different priors or models) will be given different file names. This will be made clear in the meta-data (see section 3). Ad 3. Field campaign data will first be error checked and cleaned, outliers -where possible- re-analysed. After this stage, the raw data will be locked and the data will be released.
2.8	**Is there any non-digital data or outputs that the project will generate? Where will these outputs be stored?** *Please specify shortly* ***and describe who is responsible for storage of these outputs***.
	No, there will be no non-digital data.
2.9	***Do you expect to have any supplementary costs for storage not covered by the project budget?*** *Please specify*
	No supplementary costs are currently envisioned.

	3. Data documentation Documenting your data to help future users to understand and reuse it
3.1	**How will files be named? Include the dataset identifier** *Please describe shortly.*
	For “derived” data, each file name is composed of: - Region of Interest (time and space) identifier - version of the platform identifier This avoids the need of adding the date of processing within the file name. For field campaign data, each file name is composed of: - Region of Interest (time and space) identifier - variable(s) - date of measurement This file structure is compulsory within the data archive and will be heavily promoted for the

	data storage stage too.
3.2	**How will folders be named and structured?** *You are invited to draw a folder structure and describe it shortly.*
	Folders are named and structured following the Region of Interest (time and space) identifier
3.3	**How do you handle version control to maintain all changes that are made to the data?** *Please explain your choice shortly. Remember to also document any deletion of data, if applicable.*
	☐No version control (e.g. original files are overwritten) ☐Version control software, namely: … X Data/version number in filename/folder ☐'Track changes' feature in software ☐By saving the script with which I process my data ☐Other, namely: …
	We developed a comprehensive data version system (described in another deliverable). This data versioning system fulfils the state of the art in terms of software engineering and ensures an appropriate coupling between data and software. Different files will be distinguished based on the criteria outlined above. No separate version management of the data is therefore needed.
3.4	**What metadata standard will be used, if any?**[i] *Please explain why you use this standard (most used in my discipline, required by the data archive where I will deposit my data). Please outline how the metadata will be created (read me file, spreadsheet, in the data). If no standard exist, please specify which metadata is needed to understand the data.*
	X No metadata standard is used ☐Generic metadata standard (e.g. Dublin Core) ☐Standard automatic Windows metadata (e.g. from Word, Excel) X Specialised metadata standard, namely: … ☐Other metadata standard, namely: …
	For all derived data, the NetCDF+ CF-1.7 metadata standard will be implemented. Within this standard variables are defined within the code with their units. The big advantage of CF-1.7 is that it deals with projections in a consistent way which is understood by other software. The use of variables that might not have standard names (e.g. leaf equivalent water thickness or real part of the dielectric constant of the soil) is dealt with, provided that units of the variables are defined. For field campaign data, metadata will be generated based on the data management protocols available for each participant. For all data, each participant is responsible for the generation of any metadata.
3.5	**What supporting information / documentation will you create to enhance understanding of the data ?** *Please describe shortly how peers should be able to understand the data. Examples are a readme.txt, lab journals, a codebook, survey questions etc. Is there a standard for documentation in your field? Describe at what moment in your research process you will add the documentation necessary to make sure the data is understandable for peers.*
	The data archive is coupled to the Multiply platform within which the data retrieval can be executed. Coupled to this platform is an e-learning environment, explaining how to use the platform and how to interpret the derived data. All codes and code documentation are made

	available, and can be accessed through Github. In addition, users can run a virtual machine on DIAS to recreate the data.

	4. Data access, sharing and reuse Managing access and security, sharing your data
4.1	**Are there any restrictions placed on sharing / reuse of some / all of your data?** *Please account for not sharing your data. Reasons may be ethical, commercial, security-related, protection of personal data rules, intellectual property, copyright,*
	There will be no restrictions on sharing and accessing the data produced by the Multiply project. For "raw" data and field data owned by third parties (see section 1.1), the data sharing and access rules of the data owners apply.
4.2	**With whom will you share your data at which stage in your research? You can use the table below.** *Please state any sharing requirements, e.g. funder data sharing policy. Please describe shortly how you will share your data: on request, pro-actively, etc.. Please specify how your data can be accessed.*

	Would not share with anyone	Would share with my immediate collaborators	Would share with others in my research centre or at my institution	Would share with scientists in my field	Would share with scientists outside of my field	Would share with anyone
Immediately after the data has been generated		Vegetation state Soil State				
After the data has been normalized and/or corrected for errors				Vegetation state Soil State		
After the data has been processed for analysis		Disturbance indicators Biodiversity indicators Field campaign data				
After the data has been analysed			Disturbance indicators Biodiversity indicators Field campaign data			Vegetation state Soil State
Immediately before				Disturbance indicators		

publication				Biodiversity indicators Field campaign data		
Immediately after the findings derived from this data have been published					Field campaign data	Disturbance indicators Biodiversity indicators

Based on: Interview worksheet, Jake Carlson, Purdue University Libraries / Distributed Data Curation Center

4.3 **If intending to share any part of the data, do your participant consent forms include information about intentions for sharing, retention of data and steps taken to protect participants privacy and confidentiality?**

X Not applicable.

☐Yes. *Please specify the relevant formula in the consent form.*

…………

4.4 **Who has authority to grant (additional) access to your data?**
Please describe shortly.

X Only you

☐A colleague from the project, namely: …

☐Supervisor

☐Funder

☐Collaborator / research partner organisation

☐Other, namely: …

…………

4.5 **How will you manage copyright and Intellectual Property Rights issues?**
Who owns the data? How will the data be licensed for reuse? Please describe shortly your choices and their consequences.

Copy and Intellectual Property rights have been arranged in the Consortium Agreement.

4.6 **What is the audience for reuse?**
Please list possible audiences and purposes. Consider who might use it now and who might use it later.

We have identified four possible audiences for use and re-use:

1. Fellow scientists, institutions and companies active in the development of new satellites and new (radiative transfer) models to generate satellite products. With the platform and the data generated by the platform, the in-silico performance of satellites and retrieval models can be evaluated and improved.
2. Fellow scientists, institutions and companies applying (our improved) satellite products to inform agricultural practices, nature management and land use management.
3. Fellow scientists, institutions and companies that need internally consistent high quality satellite products to parameterise and/or validate their models in the fields of soil and vegetation modelling.

	4. Fellow scientists, institutions and companies who would like to use our data products, e.g. on atmospheric correction, or who are interested in the field campaign data for validation of remote sensing products.

	5. Data preservation and archiving Preserving your data
5.1	**Which criteria will you use to decide which data has to be archived?** *Please shortly describe your choices.*
	☐Type of data (raw, processed) and how easy it is to reproduce it X Relevance of content for others ☐Usability of format for others X Data underlying publications X Verification of research ☐Available time ☐Available money ☐Other, namely: …
	Through the platform, “raw” EO-data will be processed. We chose only to archive retrievals that we considered to be of wide use for other users, such as the atmospheric correction fields, prior databases, the emulators and retrievals associated to validation sites and retrievals used in the e-learning platform. No other data were stored and instead the software versions and satellite data versions used to derive those data sources were documented as this turned out to be cheaper and faster. We consider it essential to archive data underlying publications and publicly available reports to allow full transparency and reproducibility. The field campaign data are an important source of verification and have been preserved for this purpose. These data will be archived in specialised repositories (such as DRYAD) and directly linked to the publications and will have their own DOI.
5.2	**How long should your data be preserved? Are there any requirements regarding the disposal of data?** *State obligations you have by law, funder, university, etc. if any.* *Describe how you will dispose of the data, e.g. how you will get approval, what people and/or tools you need, etc.*
	As far as we are aware there are no legal obligations (yet) for the period over which the data will be stored. We aim for a 5-10 years’ time frame.
5.3	**Which data repository is appropriate for archiving your data?** *Please describe shortly. Does this archive have a ‘data seal of approval’ or another form of certification?*
	☐Discipline specific (international) repository, namely ... ☐SurfSara ☐DANS X Other (international) repository, namely : DIAS ☐Other, namely: …
	Ad 2. For processed data, we archived data that are considered of wide use directly in connection to our virtual machine on DIAS. This comprises results from SIAC as well as prior databases

	and emulators. Ad 3. For field campaign data, the cleaned and error-checked data will be archived in specialised repositories, with their own DOI and directly linked to the accompanying scientific publication. Our data will not have a certification.
5.4	**Does the archive have specific requirements concerning file formats, metadata etc.** *Provide relevant urls to the documentation on these requirements. Describe how you intend to meet those requirements, e.g. converting the file formats, providing supplementary documentation.* *Will there be extra costs to prepare your data for archiving? Please specify. See http://www.data-archive.ac.uk/media/247429/costingtool.pdf*
	Data are stored in netCDF and ascii formats. However, note that we developed routines in the MULTIPLY platform that allow reading other file formats and checking compliance to file formats. All outputs are netCDF and asci. There are costs involved in archiving the data. However, to promote the use of the MULTIPLY, we decided to bear those costs (paid by Leiden University) and other costs associated to hosting a VM on DIAS for the next 5 years.
5.5	**Rules for accessibility of the archive** *Provide agreed upon rules (including costs & payment)*
	Our archive will be open-access without additional costs: Users can download our VM including archived data free of charge. Use of the VM for retrieval and other calculations will be done through the account of the user (and thus paid by the user).
5.6	**Who is responsible for the data after the project ends?** *Please state a position and the current person in that position.*
	The data archive is available on DIAS. Leiden University (Peter van Bodegom, project coordinator) guarantees the availability of the VM with all associated data for the next five years.

[i] See http://www.dcc.ac.uk/resources/metadata-standards or http://en.wikipedia.org/wiki/Metadata_standards or the relevant repository.