DATA MANAGEMENT PLAN (DMP)

PANINI Shared Dataset

What these data are:
Partners in PANINI will agree a minimum set of **common summary data** which all projects (including those using existing datasets and those planning to collect new data) will contribute to.

These data will be **numerical and tabular**, and include the results of biological assays and tests, total scores on self-report questionnaires/surveys, and function tests. These data will be collected from a range of older adult populations depending on the individual project which provides data; these include independent community dwelling older adults including those from minority ethnic groups, those in assisted living accommodation, those in care homes, outpatients from a mobility clinic, and short-term elective surgery inpatients. This will yield several data categories:

1. **Psychological** - wellbeing, depression, anxiety, stress
2. **Physical function/activity** - walking speed, balance, activities of daily living, accelerometry, self-reported activity, muscle function
3. **Cognitive function** (executive function, memory, reaction time),
4. **Nutrition information** (self-reported dietary intake, nutritional assessment, protein-energy balance assessment),
5. **Biological** (methylation of candidate genes, genomic variants, immunoassays),
6. **Anthropometrics & body composition** (BMI, DEXA data),
7. **Socio-demographics**. Data will be collected from a range of older adult populations.
8. **Social networks -** typology, size, composition and density of social networks influence eating and physical activity behaviours
9. **Qualitative –** data from individual interviews exploring eating behaviours, social capital, and identification of key social dynamics that promote or inhibit the ability of ethnically diverse older adults to change eating and physical activity behaviours.

Individual Beneficiaries within PANINI will hold and store the original source data for at least 10 years as per the Data Protection Act (questionnaires, surveys, tabular data of each response in questionnaires, laboratory books assays, interview transcripts, etc.) for this shared dataset as well as additional data collected or analysed as part of individual ESR research within the PANINI project, which are unique to each Beneficiary and may not contribute to the shared dataset.

Link with publications:
These data will underpin the scientific publications arising from PANINI and such publications will refer to the data repository in order to comply with transparency and research integrity. The PANINI website will indicate PANINI publications from the shared dataset and link to the dataset in the data repository (Research Data Archive (RDA)) which is fully searchable by external users. Links from the RDA will also be made through using research publication repositories e.g. PURE https://pure.bham.ac.uk/. The RDA and PURE adhere to the FAIR principle. The FAIR principle means that data must be findable, accessible, interoperable and reusable.

Data derived by Beneficiaries as part of PANINI but not contributing to the shared dataset will also be stored and linked to PANINI publications as above. This will be the responsibility of the PI at each individual Beneficiary site.

DATA MANAGEMENT PLAN (DMP)

Who is it useful for?

This will be useful to the beneficiaries of PANINI, and ultimately to anyone interested in healthy ageing research from a multi-disciplinary perspective, particularly in terms of the impact of physical activity and nutritional interventions in older adults. In addition, the individual projects within PANINI will have the option to further contribute unique data to the ultimate shared dataset on completion of the project, from their own specific projects which may have unique measurements not practiced across the network.

Overlap with other datasets:

PANINI will have some overlap with existing datasets e.g., NUAGE, MYOAGE, as certain of the projects within PANINI will be utilising or building upon these existing projects. This gives the potential for integration and the addressing of new research questions. Data from existing datasets will only be integrated into the PANINI shared dataset following agreement by these project Principle Investigators, and only if in ethical accordance with the original consent from participants/subjects or if additional consent has been agreed to cover this. The PANINI Beneficiary contributing these data will be responsible for obtaining these agreements and proof of consent.

The PANINI shared dataset will be created by ESR1 on PANINI under the supervision of Maier and Phillips (PI).

We will comply with EU fair principles and ethical compliance, for example the Charter of Fundamental Rights of the European Union, the European Convention on Human Rights, and its EU Data Protection Reform in January 2012.

We will evaluate the DMP on an annual basis.

The database will be generated by ESR1 in **IBM SPSS** for ease of day to day statistical analysis by the PANINI Early Stage Researcher initially working on these data. However, on making the data publically available it will be exported/transferred to long term storage as a csv file, which is suitable for long-term access and data sharing and can be opened in many different data and statistical packages including free open technologies.

Personal data generated during individual PANINI projects at each Beneficiary site, e.g., names, addresses, D.O.B., will be removed before entry into the shared dataset. Only individual PIs within PANINI will retain personal data, and only in instances where tracking of participants is essential to the project, e.g., there are several rounds of data collection, or data are to be linked to medical records information. Personal data will be stored separately from research data as per standard ethical guidelines. In line with EU ethical Regulations all data to be included and used will be anonymized.

Versions will be labelled with version number and date in the title. Rigorous checking for accuracy will be applied before the most up to date version is electronically archived for open access.

We will use a standard file naming convention:

FILENAME-YYMMDD-CATEGORY-VnMn-STATUS.EXT

FILENAME   the name or title, preferably without embedded spaces or special characters except dash '-' or underline '_'

YYMMDD   Publication or version date

CATEGORY   the information category, one of 'CONFIDENTIAL', 'RESTRICTED' or 'OPEN', uppercase

DATA MANAGEMENT PLAN (DMP)

VnMm   'version n modification m' where n and m can be any meaningful numbers
STATUS   'Draft', 'Published' etc.
.EXT   File extension Operating system dependent

- Each participant included in the dataset will be assigned a unique ID code which will be the first variable. Their ID from each individual project will be stored and matched to the unique PANINI ID in a separate document in order to ensure participant matching when entering follow-up data.
- Variable 2 will indicate which beneficiary in PANINI has provided the data (labelled from 1-8).
- Variable 3 will indicate which ESR project provided the data, where beneficiaries host more than one ESR (labelled from 1-11) or which existing dataset provided the data.
- This will enable analyses within specific projects as well as across the dataset, taking into account the potential influence of different countries and investigators in the data collection.
- Variables preceding each set of data will include: date, time of day, whether this is baseline or follow-up – the best way of coding this will be decided within Work Package 1: Standardisation by ESR1.

Personal data will be stored separately from any research data collected in any of the PANINI projects, and only accessible by members of the Research Team at each beneficiary's site in order to link participant contact details with their research data during data collection for purposes of follow-up or data linkage in accordance with local and national Ethical Guidelines and permissions. Files linking personal data with ID numbers will be stored electronically separately from PANINI dataset and individual beneficiaries' own local datasets as encrypted files on password protected systems, and only accessible by members of the Beneficiary's Team at the site in accordance with the Data Protection Act. Thus, there will be no personal data stored in the shared PANINI dataset. Hard copy data and stored biological samples e.g., questionnaires, blood samples, serum etc. will be identifiable only by participant ID numbers and stored securely at the site of each beneficiary according to local, national and European regulations including the Human Tissue Act, Data Protection Act, and Ethical Guidelines.

Metadata:

Given that the shared dataset in PANINI will be multi-disciplinary, it is difficult to choose and apply one specific metadata standard for overall dataset. Consequently, to aid data discovery we will implement the metadata profile CERIF (Common European Research Information Format) is the standard that the EU recommends to its member states for recording information about research activity – http://www.ukoln.ac.uk/rim/documents/Introduction_to_CERIF_1.0.pdf. This will make PANINI searchable within the University of Birmingham Research Data Archive software in which it will be deposited and PURE, the current research information system developed by Elsevier that implements the CERIF standard and is utilised by the Coordinating institution (University of Birmingham) and other systems using this standard. This will include information such as the title, dates/years covered, geographical coverage, authors, manager, publisher, access options, contact person, projects linked to the dataset, units, and links to publications from these data and any other documents necessary to describe how the data were derived and essential for interoperability and data comparison e.g., protocol for the intervention, SOPs. These will be standardised using ontologies. Where terms that are essential for interoperability of the PANINI project are not yet captured in ontology, standards will be defined in the project. Metadata will be held in searchable repositories such as PURE, and also within Cadmus. The PI will ensure metadata is created by recording the research data within PURE. We will also investigate the viability of sharing the data more widely through registration with other database systems, e.g., DASH-IN, a nutritional database.

DATA MANAGEMENT PLAN (DMP)

Interoperable: the dataset initially produced in SPSS during the course of the PANINI project will be made interoperable following the completion of the PANINI grant through easy transformation to and storage as a csv file. This will allow for data exchange between researchers institutions, organizations, countries, etc. through adhering to standards for data annotation, data exchange, compliant with available software applications, and allowing re-combinations with different datasets of different origins

For initial use within PANINI, documentation about specific variables will be recorded very simply in separate manual parallel to each of the dataset files. Each variable will be listed in the same order as each file within the dataset, and labelled first by its name in the dataset, followed by the beneficiary the data were derived from, followed by the ESR number the data were derived from, followed by the year the data were derived, followed by a free text short description of the measure. This documentation will also be available alongside metadata to support users. Documentation will include:

- o **Who** created it,
- o **When**;
- o **Why;**
- o **Description** of the item;
- o **Methodology** and **methods;**
- o **Units** of measurement;
- o **Definitions** of jargon, acronyms and code;
- o **References** to related data.

Data sharing:

Data will be shared between PANINI beneficiaries during the course of the project via email of password protected encrypted zipped datafiles in IBM SPSS format as described above. This format will be used during the course of the project, but for long term storage and open access, the data will be converted to open csv format. The working dataset will be stored at the PI's institution on password protected computers which automatically save files to the university server rather than on each machine's hard drive. It will also be backed up weekly on external hard drives by ESR1 or the PI, locked away separately from the main computer containing the data.

It is difficult to know at this stage how large the shared dataset will be, but we estimate around 1GB. We will update the DMP annually with a better estimate as this becomes apparent. The accompanying files e.g. protocols, etc. will be several MB maximum.

The full PANINI shared dataset will become Open Access within one year of completion of the project grant, embargoed only until the final PANINI publications, submitted during the course of the grant, have been accepted and are available online via journal websites, institutional repositories and Cadmus. The database will be designed and held in Amsterdam for the initial half of the PANINI project then transferred to Birmingham for long term storage.

Data from existing datasets contributing to PANINI may be under different embargo rules or may be already made open access in a different format than the PANINI shared dataset. Beneficiaries contributing such data will be responsible for ensuring that embargo and/or sharing of the PANINI dataset will not be in contravention of existing agreements/rules, and where possible, will gain agreement from existing project Principal Investigators. If this is not possible, the shared PANINI dataset may need to have two sections, existing data and new PANINI-generated data with different embargo and sharing rules if agreement with existing projects cannot be matched to the PANINI embargo and sharing rules described above. Where it is

DATA MANAGEMENT PLAN (DMP)

not possible to share certain aspects of the data as Open Access (e.g. from existing datasets under different embargo rules), we will specify which data will not become Open Access in a future version of this Data Management Plan when this is known.

Data generated during PANINI specific to individual Beneficiaries' projects will be published where possible by storage in research data archives at each Beneficiary site and linked to publications. We will also seek to enhance data sharing of these individual smaller project datasets within PANINI by publishing in journals which support publishing datasets alongside articles where possible.

We will comply with updates on EU legislation on data sharing.

Findable: The data are findable given that PURE is metadata searchable and the data deposited in PURE will also be made searchable within the Research @ Birmingham page, FindtIt@Bham and Google. The RDA is currently being developed to be able to be searchable from outside UoB.

All publications from the PANINI shared dataset and also from specific Beneficiaries' projects within PANINI will include a short statement describing how and on what terms any supporting research data may be accessed.

Data access:
The PANINI dataset will be searchable through data repository searches based on the CERIF metadata profile e.g., PURE.

Accessible: The data are accessible, as the metadata will ensure that the project is assessable and intelligible to third parties. Where possible, the PANINI shared dataset will become Open Access (under the CC-BY license) as recommended by the EC. Initially the study will only be open to the owner of the study (and the administrator of the system). The original owner of the data has management rights and controls who else has access to the data during the course of the PANINI project. The study can be made open access upon publication of the results within one year of completion of the PANINI project grant.

Archiving:
As well as being contained in the Research Data Archive (searchable), attached to published papers from the PANINI shared dataset, the PANINI shared dataset will also be backed up in long term storage in the institutional archive: Research Data Archive at the University of Birmingham (the project coordinator) where 3 terabytes of data can be stored free of charge to university staff during the course of the project and is automatically backed up by the University of Birmingham.

In order to make the final PANINI shared data set more easily Findable and Accessible following completion of the PANINI grant, we will also seek to deposit the dataset and accompanying metadata and files in other appropriate open data repositories. We are currently researching which might be most appropriate and are considering dash-in.eu (part of ENPADASI project) a nutritional phenotype database, and other repositories via re3data.   Also, papers generated at University of Birmingham are deposited in our research e-papers repository which we are now requesting be included in re3data, making PANINI papers more accessible.

The data will be stored throughout the grant and for a further 10 years. If the data is requested during the 10 years, the 10 year clock will be reset and the data will then be stored for a further 10 years. It is not

DATA MANAGEMENT PLAN (DMP)

expected that the final database will exceed 3 terabytes; therefore no additional costs for archiving/storage are expected.

Usability: the PANINI dataset will be usable by third parties even a long time after data collection through storage in the certified repositories for long-term preservation described above. The PANINI dataset will be stored alongside the minimum metadata and documentation needed to make the dataset useful beyond specialists only.

Raw data will be stored separately with mirrors to prevent data loss, until finalization of processing. Once the data is processed and finalized, raw datasets will be stored as backup until the end of the project grant.

Standardized pipelines will be used for data processing, these pipelines will be carefully described to improve reproducibility and their actual implementation will be made public where possible. Data will be curated during data entry similar to a clinical study protocol and re-validated for analysis by ESR1.
The underlying non-digital data (questionnaires, surveys, lab books) not containing personal data will be stored at each Beneficiary site that generated the data. Where this is substantial, local records stores may be accessed and funded via PANINI before the grant terminates, e.g. Modern Records Store, University of Birmingham: £15 per box for up to 7 years storage.
https://intranet.birmingham.ac.uk/as/specialcollections/modernrecordsstore/index.aspx.

Data sustainability and Back-up plan:
Following completion of the dataset, the PANINI shared dataset will be backed up in long term storage in the institutional archive: Research Data Store at the University of Birmingham (the project coordinator). During the course of creating the database, in order to prevent data loss through both equipment failure and human error, data will be backed up overnight automatically by the host institution. Each file will be kept for 90 days and have a maximum of 3 versions.
https://intranet.birmingham.ac.uk/it/teams/infrastructure/research/bear/research-data-service/RDS/BackupRetentionPolicy.aspx

Dated versions of the dataset will be backed up on a weekly basis to the secure local server at the site of creation (initially Vrije University for 18 months then eventually the University of Birmingham). These copies will be separate from the working dataset which will be stored on local institutional computer hard-drives also automatically backed up to the local server as described above. At the end of each month, when the working database has been screened and checked for errors, the latest version will then replace those stored each week.

> **Note:** An annual review of the document will be carried out to adjust systems accordingly and to ensure that all Data Protection and FAIR principles are adhered to.