## Working Package

| WP no | WP name | Lead partner | Start month | End month | Deliverable no |
|---|---|---|---|---|---|
| 2 | Preparing joint data analysis | **Lars Ove Dragsted** *ldra@nexs.ku.dk* | 1 | 24 | List of datasets available for integration,Minimal requirements for data entry,Description of best terminology for metabolomics data,Study validation criteria,Case studies development,Guidelines for nutritional data sharing,Training material on minimal requirements delivered to WP6 ,Studies relevant to case studies uploaded,Case studies analysis |

### Detailed information on work packages

**WP Leader:**

| | |
|---|---|
| Legal name of organisation: | University of Copenhagen - Dept. Nutrition, Exercise and Sports |
| Country: | DK |
| ZIP code: | DK-2200 |
| Town: | Copenhagen |
| Street name, number: | Norre Allé, 51 |
| Additional (e.g., department, building...): | |
| Website: | |
| Mrs/Mr: | Lars Ove Dragsted |
| Title: | PROF |
| First name: | |
| Last name: | |
| Function: | |
| Phone: | 24666694 |
| Fax: | |
| E-mail: | PROF |

**Additional information on person**

(max. 1500 characters) concerning personal background and explain responsibilities and tasks:

Lars O. Dragsted MSc PhD, is currently a professor in Biomedicine and Nutrigenomics at the Dept. Nutrition, Exercise and Sports (NEXS), University of Copenhagen (UCOP). He has currently a group of 15-20 scientists and other staff working with human dietary studies, metabolomics, biomarker development, and bio-banking. He has authored or co-authored 200+ scientific publications, reports and book chapters on subjects related to lifestyle, foods and health. He and collegues at NEXS conduct dietary intervention studies. Some of these studies are multi-center studies with data-sharing and NEXS has a data manager involved in setting up and controlling common databases similar to the Nutritional Phenotype Database system (Ommen et al., 2010). UCOP has a common sample repository (www.cube.ku.dk) and a metabolomics facility with an associated in-house standards database. In WP2, UCOP and other partners interested in preparing for joint analyses of existing data will form a framework for data sharing focusing on data quality, guidelines for data sharing and show-cases explaining the potentials of fused data. Lars O. Dragsted as a WP leader will have special responsibility a. for assuring that the work is coordinated across study types (observational and experimental), b. that there is a coherent workflow along the tasks, and c. that in the end many datasets are available for joint analysis to make data sharing and joint data analyses attractive for European researchers.

## Partners Involved

| **Legal name of organisation** | **Knowledge Hub member (main contact person/ project leader within the organisation)** | Person months # | Start month | End month |
|---|---|---|---|---|
| Ghent University - Faculty of Medicine - Department of Public Health | De Henauw Stefaan | 30.6 | 1 | 24 |
| CRNH Rhône-Alpes - Bt 1A | Laville Martine | 1.5 | 12 | 24 |
| CRA-NUT | Giuditta Perozzi | 42 | 1 | 24 |
| Alma Mater Studiorum - Università di Bologna - Department of Pharmacy and Biotechnology | Patrizia Brigidi | 17 | 1 | 24 |
| Istituto Superiore di Sanità - Department of Veterinary Public Health and Food Safety | Massimo D'Archivio | 32 | 1 | 24 |
| University of Rome "La Sapienza" - Department of Chemistry | Federico Marini | 19.5 | 1 | 24 |
| University of Bari Aldo Moro, Italy - Department of Soil, Plant and Food Sciences | Gobbetti | 25.6 | 1 | 24 |
| Institute of clinical physiology (IFC), UOS Lecce - Campus Ecotekne | Massaro Marika | 4.5 | 1 | 24 |
| National Research Council - Institute of Food Science (ISA) | Rosalba Giacco | 20 | 1 | 24 |
| IRCCS Burlo Garofolo - Medical Genetics | Paolo Gasparini | 24 | 1 | 24 |
| University of Naples "Federico II" - Department of Clinical Medicine and Surgery, Building 1 | Angela Albarosa Rivellese | 17 | 1 | 24 |
| Fondazione Edmund Mach - Food Quality and Nutrition Department | Fulvio Mattivi | 12.75 | 1 | 24 |
| Alma Mater Studiorum - Università di Bologna - Dept. Scienze e Tecnologie Agroalimentari | Francesco Capozzi | 2.4 | 12 | 24 |
| National Research Council - IBIMET | Duccio Cavalieri | 1.3 | 12 | 24 |
| Netherlands Organisation for Applied Scientific Research (TNO) - Department of Microbiology and Systems Biology | Jildau Bouwman | 32.3 | 1 | 24 |
| Max Delbrück Center for Molecular Medicine (MDC) - Molecular Epidemiology Research Group | Tobias Pischon | 20 | 1 | 24 |
| German Institute of Human Nutrition (DIFE) - Department of Epidemiology | Heiner Boeing | | | |
| Department of Nutrition and Food Sciences, University of Bonn | Ute Nöthlings | | | |
| Helmholtz Zentrum München (HMGU) - Research Group National Cohort | Jakob Linseisen | | | |
| Max Rubner-Institute (MRI), Federal Research Institute of Nutrition and Food - Department of Nutritional Behaviour | Ingrid Hoffmann | | | |
| University Heart Center, University Hospital Hamburg-Eppendorf - Preventive Medicine | Prof. Eberhard Windler, M.D. | | | |
| Bio-Competence Centre of Healthy Dairy Products (BioCC) | Andre Veskioja | | | |
| National Institute for Health Development (NIHD) - Department of Surveillance and Evaluation | Eha Nurk | 25.8 | 1 | 24 |

| | | | | |
|---|---|---|---|---|
| University of Liège - Unit of Prof. Michèle Guillaume "Nutrition, Environment and Health" - Department of Public Health | Anne-Françoise Donneau | 16 | 1 | 24 |
| Scientific Institute of Public Health (known as "WIV-ISP") - Surveys, Life Styles and Chronic Conditions Service, Nutrition and Health Unit | Jean Tafforeau | | | |
| University of Copenhagen - Dept. Nutrition, Exercise and Sports | Lars Ove Dragsted | 14 | 1 | 24 |
| CIBER OBN - Instituto de Salud Carlos III | Dolores Corella | 1.9 | 1 | 24 |
| Health Research Institute Hospital La Fe - Endocrinology, Nutrition and Dietetic Unit | Jose M. Soriano | 1 | 1 | 24 |
| UCD - Institute of Food and Health - University College Dublin | Lorraine Brennan | 10 | 1 | 24 |
| Gent University | Carl Lachat | 12.7 | 1 | 24 |
| Hasselt University - Faculty of medicine and life sciences | Wim Pinxten | | | |
| KU Leuven - Clinical and experimental endocrinology | Christophe Matthys | | | |
| Technische Universität München - Molecular Nutrition Unit | Kurt Gedrich | 17 | 1 | 24 |
| Centro de Investigación Biomédica en Red - CIBERDEM - Pabellón 11 | Luis Castaño | 2 | 1 | 24 |
| Total | | 402.85 | | |

**Description of work package:**

General Description description-10-545c8ec2d42be.docx Show Uploaded Description

**Work package 2**

WP name: Preparing joint data analysis and sharing existing data

WP leader: Prof. Dr. Lars Ove Dragsted (Denmark)

**Description of work package:**

1. Scope of work package (including tasks, deliverables, risks) and interrelations with other work packages

This WP is concerned with data collection, setting standards for data quality and creation of guidelines and examples of use. WP2 is closely aligned with all other work packages. Expertise from both observational and experimental research is integrated in most tasks by leaders and co-leaders for each study type. The work package is divided into 5 tasks:

2.1 Collection of data sets for integration, subdivided into observational datasets (lead: EoI51) and experimental datasets (lead: EoI41) months 1-12. In this task we collate and describe potentially available datasets with a multitude of phenotypic outcomes coming from partners or already shared publicly. A preliminary list of partner projects, including relevant details of the datasets available for sharing through the ENPADASI KH by M12 of the project has already been compiled. No risks are therefore anticipated for this work. The output deliverable D2.1.1, *List of datasets available for integration* will be available on the internet after 12 months.

2.2 Minimal requirements for study data, subdivided into observational datasets (lead: EoI51) and experimental datasets (lead: EoI71), months 3-18. The major risk anticipated for this work is lack of efficient interaction with other players in the field. We will assure personal contacts within key collaborating initiatives, such as the JPI-HDHL BioNH and other relevant EU projects originating nutritional datasets or infrastructural design in the nutrition & health field (eg EuroDISH). Deliverable 2.2.1, *Minimal requirements for data entry*, and D.2.2.2, *Description of best terminology for metabolomics data* will either be posted as a report on the internet or published in an open access journal by the end of the task.

2.3 Validation of study quality, divided into observational studies (lead: EoI38) and experimental studies (lead: EoI71), months 1-12. No specific risks are anticipated for this work. Deliverable 2.3.1, *Study validation criteria*, will be a draft scientific paper; the criteria *per se* are posted as a report on the internet by the end of the task.

2.4 Case studies for existing data (lead: EoI41), months 12-24. The risk is that few of the datasets available overlap with respect to relevant endpoints making good examples of data fusion less persuasive. Deliverable 2.4.1, *Case studies development*, will be published as a report on the internet. Deliverable 2.4.2, *Case studies analysis*, will be published as a report on the internet and/or as a paper in an open access journal.

2.5 Guidelines for data sharing (lead: EoI71) months 16-24. No specific risks are anticipated for this work. Deliverable 2.5.1, *Guidelines for nutritional data sharing*, will be available on the ENPADASI internet site.

**Deliverables**

**D2.1.1** Description of datasets available for integration, availability due after 12 months.

**D2.2.1** Minimal requirements for data entry

**D2.2.2** Description of optimal terminology for metabolomics data will be posted as reports on the internet by the end of the task

**D2.3.1** Study validation criteria, will be a draft scientific paper for an open access scientific publication containing also information from task 2.1 and 2.2; the criteria developed in task 2.3 per se are posted as a report on the internet by the end of the task. The draft paper will contain a table of validated studies uploaded into the database system.

**D2.4.1** Case studies development. The description will be posted on the ENPADASI homepage (month 4) to increase the awareness of the possibilities in nutritional data sharing.

**D2.4.2** Case studies analysis. The output will be drafted as an open access journal article and posted on the ENPADASI homepage (month 24) to increase the awareness of the possibilities in nutritional data sharing and to answer the question that may be posed by many nutritional researchers, " what's in it for me?".

**D2.5.1** Guidelines for nutritional data sharing, will be posted on the ENPADASI homepage (month 24) to increase the awareness of the possibilities in nutritional data sharing.

**Milestones**

**MS2.1** Training material on minimal requirements delivered to WP6 (Month 18)

**MS2.2** Studies relevant to case studies uploaded (Month 18)

## 2. Concept and objectives

a. Objectives, vision including scientific/ technological challenges:

It is the objective of WP2 to facilitate data sharing in nutritional research across Europe and to guide researchers on the practical steps in preparing data for entry into a format that allows sharing through ENPADASI and with other relevant partners (e.g. JPI DEDIPAC and EuroDISH). The vision is that any researcher can bring their own data to sufficient quality for sharing and that researchers can clearly see the improved scientific possibilities and their own benefit from sharing data with others through ENPADASI. This vision includes experimental (mechanistic and intervention) as well as observational studies (epidemiological). The challenges are 1) to describe the prerequisites to which a potentially shared study should adhere for all major study types, 2) to show that the system gives new scientific possibilities for those who share, and 3) to provide facile tools and descriptions on how to share your data. The technical, legal, ethical, and governance requirements for data entry will be the main objectives of WP3, WP5 and WP1 (task2) while the integration of this work will be orchestrated within WP4 and training in WP6.

b. State of the art: Almost all larger nutrition research institutes worldwide have established their own data management systems, or systems are developed *ad hoc* for each new project. While public data storage after study finalization is often a prerequisite for funding and for publication the (usually national) systems are in many cases not well suited for receiving the data structures produced by systems biological-nutrition studies. Moreover, the access to stored data is constrained and may include ill-defined application procedures. Shaping a common system for sharing in real time as envisioned by the ENPADASI initiative must build on well-defined descriptions of the minimal requirements to which data should comply (including the necessary information on

study design, endpoints, other measurements, data ownership, data availability and ethical limitations). Such requirements have not yet been defined and developing these criteria must therefore be one of the foundations for building ENPADASI. There are only few examples of scientific achievements obtained by nutritional data sharing. An example is the (Nutritional) Phenotype database (www.dbnp.org), which was an initiative of NuGO (NutriGenomics Organization) and NMC (Netherlands Metabolomics Centre), which was launched in 2007. This data infrastructure was developed to store data from nutritional intervention studies with complex design (including cross-over) and is meant to facilitate standardized data output and study comparisons. One of the instances is publically available (studies.dbnp.org), on which daily and weekly back-ups are made. The database currently includes 58 biological studies, mostly nutritional intervention studies. The system is connected to other databases containing metabolomics data on www.metabolomexchange.org and is also used in the biomedical research area (CTMM/TraIT). Within the EU project EuroDISH data analysis of combinations of studies (from the Phenotype database) with multiple measurements and interpretation of this analysis was performed. It was shown that a virtual cohort can be developed based on data of independent studies, making use of specific statistical tools. This indicates that it is possible to use the data from well-structured studies to answer new biological questions based on old data. However still most nutritional data sets are not shared. This is a direct consequence of the lack of appropriate infrastructures but may also rely on reluctance on behalf of the involved researchers to share their hard-earned work and lack of a scientific culture for data sharing in nutrition research (Tenopir et al., 2011). In order to overcome this challenge there is a need for more good examples of sharing with important outcomes and instructions for use that are easy to grasp and include the workflow for most researchers in the field. Public guidelines is one of the important tools that need to be in place to meet this challenge but other tools to solve political, legal and ethical issues are equally important and need to be solved in concert (WP5).

c. Scientific/ technological concept:

WP2 shapes the scientific/biological part of the foundation for data sharing. Many data sets from nutritional studies exist at the ENPADASI partner institutes and other places within Europe and beyond. More and more journals require deposition of data as a prerequisite for publication but the few and scattered systems for sharing data are not adequately designed for open querying. Partners and other researches who wish to share their data therefore need a system to define and unify the data formats. In order to create a system for querying data from many different studies of different design, duration, etc., there is a need for classifying study types and defining the variables that will characterise each unique study. The central scientific and technical concept of the current WP is to outline these classifications and definitions:

i. Creating a global system to classify all study types in nutrition
ii. For each study type to create the vocabulary (in collaboration with WP4 and 5) and metadata variables (e.g. time line information, participant characteristics, interventions (if any), etc.) necessary to minimally define the study setup.
iii. To define the modalities of use for the data in each study, including ethical limitations, description of data ownership and rules for data use (typically defined by the project consortium), etc.
iv. Defining variable names for nutritional, biological and physiological endpoints that include the sample type, time point and endpoint.
v. Defining matrix structures for ‘omics data and preferred variable names within matrices, e.g. point to well mapped ontologies for genes, proteins and metabolites to facilitate interoperability.

ENPADASI will link to interoperability consortia such as FAIRport for defining sample types, global variable names, ‘omics’ databases of genes, proteins and metabolites, etc. will be established and inform such consortia of input for nutritional specific requirements. This will assure that vocabularies and ontologies are in line with work in those consortia and that no double work is performed

An additional concept is to create a number of case studies that highlights the potentials of using a shared querying system for nutritional datasets. These should include very simple cases as well as more complex ones. This should facilitate the interest in data sharing since it would highlight that sharing normally leads to more scientific output (knowledge, publications, visibility) for those who share. Case studies can also help identifying solutions for proper attribution of credits when merging shared data deposited in a database (links with WP5). This aspect is of special relevance when omics datasets are used, most of which are largely unexploited by authors in their published studies. This issue can be confronted in connection with Editors/EB members of relevant journals in the health/nutrition field, as Journal and database policies need to be aligned to drive data sharing and maximize its advantages.

Finally WP2 has a very close connection with WP3, which should provide the technical solutions to implement the systematics provided here, with WP4 to define the common language for data integration and to execute the case studies defined in this work package, with WP5 for the definitions of commonalities of data usage and WP2 will also feed into the training activities (WP6), especially with the definitions and case studies.

## 3. Management

The WP will be managed by the WP leader in close collaboration with leaders of the individual tasks. Since collaboration has to take place continuously over the time span of ENPADASI much of the management will be performed using teleconferences with involved partners and task leaders and co-leaders for the tasks will be involved in order to assure that the broadest possible expertise is included. Annual meetings will be conducted as part of the annual ENPADASI meetings. Each partner will be responsible for their own

reporting to their funders but the work input from each partner will be made visible as a reference for all reporting activities.

## 4. Potential impact on the advancement of the research area, capacity building, plan for *translation of research* (suitable for ENPADASI) into public health practice or policy (in 2 years, with a perspective on a longer term)

The impact of WP2 must be seen in connection with all of ENPADASI, however the work classifying nutritional study types and unifying the use of terms from vocabularies generated by others will have a large impact on future data sharing. There is also a significant impact by the contributions to visualize the potentials of data sharing in the form of case studies and training materials that will impact on a number of nutrition researchers who will be informed also of their own gains from data sharing (coming 2 years). The nutrition and health research area has developed heterogeneously throughout Europe, largely reflecting national funding priorities. We are facing therefore a correspondingly heterogeneous, country-specific infrastructural development, as well as differential awareness of the advantages of data sharing towards the advancement of the EU research area as a whole. Unifying the different methodological approaches and study designs in WP2 will therefore lead to an increase in the overall quality of European research and optimize the overall research impact and translational capacity.

There is a great need for a system to allow generalized data sharing and our impact will be the fulfilment of these needs and a much better use of nutritional study data for future analyses across data sets and study types. This should form better and more informed hypotheses in future studies and make it possible sometimes to actually test new hypotheses without having to conduct additional studies.

## 5. Overall strategy of the work plan

WP 2 is divided into five tasks, each with a task leader/co-task leader and a team of collaborating partners.

**Task 2.1:** Collection of data for sharing – EoI51 [lead: observational studies]: EoI41 [lead: intervention studies], EoI35, 42, 50, 56, 66, 71, 73, 74 - Months 1-18

This task will initially identify relevant studies (e.g. studies related to chronic diseases with different dietary intake) that are already posted publicly for sharing or that will be made available by the partners (this list will be used by WP3 task 3.1 to identify technical needs of the data infrastructure); this will also incorporate collaboration with studies identified in the JPI DEDIPAC and BioNH projects and the EuroDISH mapping activities (Fig. 2). The identified studies will be subdivided into a number of study classes, beginning with the major groups, observational and experimental, then subdividing into additional categories until all study types are clearly distinguished from others and with a number of class division characteristics that can be used to identify the class for any additional nutritional study entered. The available studies will then be classified into study types .

Related deliverable: D2.1.1

**Task 2.2:** Minimal requirements for study data – EoI51 [lead: observational studies], EoI38 [lead: intervention studies], EoI35, 41, 42, 50, 56, 64, 66, 71, 73, 74 - Months 3-24

This task will identify the minimal requirements for studies within each class (according to 2.1), including the necessary meta-data (i.e. study description including design and SOPs used), modalities of use for the study data, and systems for terming the endpoint variables. This will also form the basis for training in task 6.2. There will be a particular weight on metabolomics data since this field is the least developed among the ‘omics’ technologies, still lacking many standard terminologies while being very important for the nutritional field. This work on metabolomics will build on the work performed in the COSMOS and Metabolight projects and in Metabolomics Society task groups. For other ‘omics’ technologies this task will connect to other initiatives (e.g. ELIXIR). This task will lead the development of the data infrastructure (WP3 task 3.5) and will bridge forward to work in the BioNH (FoodBAll and Mirdiet) joint programming activity. The studies listed in task 2.1 will be analysed with respect to the minimal requirements set up in task 2.2.

Related deliverable: D2.2.1, D2.2.2, MS2.1

**Task 2.3:** Validation of study quality – EoI41 [lead: observational studies], EoI71 [lead: intervention studies], EoI35, 38, 42, 50, 51, 56, 64, 65, 66, 73, 74, 75 - Months 1-12

Quality criteria for human studies will be defined in this task on the basis of guidelines issued by scientific societies. Most such guidelines are related to medical study designs and must be amended in order to fit nutritional studies. This will lead to a quality check system for nutritional studies. In this task the studies identified in task 2.1 are further sorted based on the minimal requirements set up in task 2.2 and scored by the new quality system. In addition, these studies are uploaded (if not yet entered in a machine readable format that adheres to the criteria) in the data infrastructure (training on data upload is taken care of by WP6 task 4). For intervention studies the infrastructure (see Fig. 3) will in most cases make use of the Phenotype database (www.dbnp.org), which can store metadata on the study design (e.g. all details on the intervention that would also be needed in a publication), subject details and measurement data (e.g.

clinical chemistry, anthropometry, etc.) including details on measurement techniques and data cleaning. This database makes use of templates and therefore will require limited adjustments depending on the selected study cases (task 2.4). Some larger intervention studies and most observational studies are generally stored in local databases; these will be made available by inclusion of an API (WP3), depending on the study cases. A list of potentially available studies in the project can be found in table2. Data can be contributed via several ways and in different levels of detail. Studies can be completely shared including measurements or only shared at the level of the study description (all meta-data). Integration of datasets derived from in vivo/in vitro models will also need to be integrated to fully exploit their outputs for human mechanistic nutrition research.

Data will be delivered by the partners and curated by a dedicated data manager (gatekeeper) of the project. The data manager will guard completeness of the captured data. Only the final version of the data (ready for publication) will be stored in the ENPADASI infrastructure. The templates of Phenotype database will be adjusted based on these quality criteria and will be shared via the ENPADASI website for broader usage.

Related deliverable: D2.3.1, MS2.2

**Task 2.4:** Case studies with existing data - EoI41 [lead], EoI35, 37, 42, 45, 50, 56, 64, 66, 71, 74 - Months 12-24

This task will define a number of case studies (for which shared data can be used scientifically and practically and later will select one or two of them as show cases where the cross-cutting analyses are actually performed using the new querying system (developed in WP4) on studies fulfilling the requirements set up in tasks 2.1-2.3. The case studies will be focused on resolving chronic diseases with life style related solutions. Close collaboration will be needed with WP3, 4, 5 and 6 since this task integrates most efforts in ENPADASI.

Related deliverable: D2.4.1, D2.4.2

**Task 2.5:** Guidelines for data sharing - EoI71 [lead], EoI35, 41, 42, 50, 56, 64, 65, 66, 73, 74, 75 - Months 16-24

This task will produce a step-by-step guideline for how to go forward in order to share study data from a nutritional study. Guidelines will include procedures for sharing of observational as well as experimental studies. This work will be done in close collaboration with WP1 (sustainability) WP4, WP5 and WP6. All partners generally support open access policies for non-commercial data and several have already shared data sets. Procedures to safeguard privacy protection of ENPADASI must be defined and procedures for this will be further developed in this WP and WP5. The database must be processed in line with the rights of consortia, which produced the data and the subjects who participated in the studies. Currently, many data cannot be used publicly due to privacy constraints and confidentiality rights. It is expected that the procedures developed in ENPADASI to address these issues will be applied and tested on this database, potentially leading to open access use of these resources in due course (WP5). The most common solution is to make sure subjects are anonymous by following procedures in the countries where the trials were conducted.

Related deliverable: D2.5.1

| **Budgetary table** | | | | |
|---|---|---|---|---|
| | **2014** | **2015** | **2016** | **Total Costs** |
| personnel **[k€]** | | | | 0 |
| travel **[k€]** | | | | 0 |
| consumables **[k€]** | | | | 0 |
| equipment **[k€]** | | | | 0 |
| dissemination **[k€]** | | | | 0 |
| others **[k€]** | | | | 0 |
| direct costs **[k€]** | | | | 0 |
| indirect costs **[k€]** | | | | 0 |
| requested funding **[k€]** | | | | 0 |
| Total **[k€]** | 0 | 0 | 0 | 0 |