Working Package

WP no	WP name	Lead partner	Start month	End month	Deliverable no
4	Integration	**Duccio Cavalieri** *duccio.cavalieri@fmach.it*	1	24	A report presenting a draft of the common language and ontologies for nutritional studies for describing the new terms and the modifications to the existing terms,Report describing the contribution to the update metadata into main databases by means of ontologies,Report describing the update of Nutrition Pathways Data Model,A report on protocols for querying data, data integration and usage of pathway tools,Guidelines for study reporting and study design,Organization of an Ontology meeting with the community of domain experts and ontology experts,One day workshop on metadata integration,Data model definition,Setup of dedicated website for web-based resource for data querying,Integration of guidelines for study reporting and study design into the ENPADASI website

Detailed information on work packages

WP Leader:

Legal name of organisation:	National Research Council - IBIMET
Country:	IT
ZIP code:	50145
Town:	Firenze
Street name, number:	Via Giovanni Caproni, 8
Additional (e.g., department, building...):	
Website:	
Mrs/Mr:	Duccio Cavalieri
Title:	PHD
First name:	
Last name:	
Function:	
Phone:	0039-0461-615153
Fax:	0039-0461-650218
E-mail:	PHD

Additional information on person

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

Duccio Cavalieri is associate researcher at IBIMET-CNR (Florence), Coordinator of the Computational Biology Department at Fondazione E. Mach (San Miichele all'Adige) and Assistant Professor at University of Florence. DC's studies deal with the interaction between genome, metagenome, nutrients and diet. The main activity is the development and application of bioinformatic tools, based on pathway and network analysis, for integrative analysis of omics data of observational nutritional studies. DC has been member of the FP7 NoE NUGO, and active in several EU funded initiatives in nutrigenomics, metagenomics, immunogenomics such as the NoE DC-Thera and the FP7 IP Sybaris. Within ENPADASI, consortium specific concerted action aims at the development of an integrated network of

computational infrastructures for nutrigenomics research. The NutriExp Directory will be an information system designed to support knowledge management and integrative data analysis for this research community and beyond, developing computational pipelines for data integration, building and sharing of methodologies and data (WP2, 3 and 4). Given the presence of internationally renowned experts in the fields of bioinformatics, nutrition, data modeling and ontology development within our Institutes, the joint institutes could play an important role to streamline bioinformatics tools enabling the development of state-of-the-art research, education and training through the network (WP4,6).

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	2	1	18
CRA-NUT	Giuditta Perozzi	2	1	12
IRCCS Burlo Garofolo - Medical Genetics	Paolo Gasparini	19	1	12
Institute of Biomembranes and Bioenergetics - c/o Dipartimento di Bioscienze, Biotecnologie e Biofarmaceutica	Graziano Pesole	2	1	24
Istituto di Analisi dei Sistemi ed Informatica Consiglio Nazionale delle Ricerche	giovanni felici	3	1	24
Fondazione Bruno Kessler - Center for Inf. Technologies / MPBA	Cesare Furlanello	7	1	24
The Microsoft Research - University of Trento Centre for Computational and Systems Biology	Corrado Priami	1.8	1	24
Institute of Food Science, National Research Council	Angelo Facchiano	23	1	24
Alma Mater Studiorum - Università di Bologna - Dept. Scienze e Tecnologie Agroalimentari	Francesco Capozzi	2.5	1	24
National Research Council - IBIMET	Duccio Cavalieri	19	1	24
Politecnico di Bari – POLIBA - The Interuniversity Physics Department	Giorgio Pietro Maggi	2	1	24
Netherlands Organisation for Applied Scientific Research (TNO) - Department of Microbiology and Systems Biology	Jildau Bouwman	12.2	1	24
Bio-Competence Centre of Healthy Dairy Products (BioCC)	Andre Veskioja			
National Institute for Health Development (NIHD) - Department of Surveillance and Evaluation	Eha Nurk	8.8	1	24
University of Copenhagen - Dept. Nutrition, Exercise and Sports	Lars Ove Dragsted	1.3	1	24
CIBER OBN - Instituto de Salud Carlos III	Dolores Corella	0.5	1	24
Gent University	Carl Lachat	15	1	24
Instituut voor Landbouw- en Visserijonderzoek - ILVO Technology and Food Science Unit	Lieve Herman			
KU Leuven - Clinical and experimental endocrinology	Christophe Matthys			

Centro de Investigación Biomédica en Red - CIBERDEM - Pabellón 11	Luis Castaño	1	1	18
Total		122.1		

Description of work package:

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

Work package 4

WP name: Integration

WP leader: Dr. Duccio Cavalieri (Italy)

Description of work package:

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

In particular WP4 will focus on definition of ontologies and common languages (Task 4.1) based on the needs defined by WP2. These ontologies will be integrated into bioinformatic applications for data and pathway analysis (Task 4.3) and intelligent interrogation of nutritional databases (Task 4.4) and the ENPADASI infrastructure by WP3. Only nutrition specific aspects will be covered by this WP, technical aspects and non-nutritional developments will be resolved in close collaboration with ELIXIR and the data FAIRport initiative.

The second focus of WP4 is the development of a common set of procedures and algorithms (Task 4.4) to facilitate data integration of nutritional studies (including nutrigenomics data: Task 4.2). Most effort will be on data and case studies collected and defined within WP2 focusing on the development of an integrated network of intercommunicating computational infrastructures (such as the NutriExp Directory). Part of the querying tools developed by WP4 will be automated and integrated in the ENPADASI infrastructure by WP3. The work package will also integrate the guidelines developed by WP5 and in this way improve the quality of data that are shared (Task 4.5).

Deliverables:

D4.1 A report presenting a draft of the *common language* and *ontologies for nutritional studies* for describing the new terms and the modifications to the existing terms (EoI74, EoI75, month 12)

D4.2 Report describing the contribution to the update metadata into main databases by means of ontologies (EoI74, all partners, Month 18, dependency on D4.1)

D4.3 Report describing the update of Nutrition Pathways Data Model (EoI45, month 24)

D4.4 A report on protocols for querying data, data integration and usage of pathway tools (EoI45, EoI38, month 24)

D4.5 Guidelines for study reporting and study design (month 18)

Milestones:

MS4.1 Organization of an Ontology meeting with the community of domain experts and ontology experts (EoI74, EoI45, month 6)

MS4.2 One day workshop on metadata integration (EoI45-**:** month 12).

MS4.3 Data model definition (EoI45, month 24)

MS4.4 Setup of dedicated website for web-based resource for data querying (EoI45, EoI38, month 24)

MS4.5 Integration of guidelines for study reporting and study design into the ENPADASI website (month 18)

Risks

Task 4.1. A risk concerning the building of the ontologies and common languages is represented by the normally scarce involvement of domain experts, who are the actual owners of the knowledge. Indeed, very often, the building of ontologies is considered as a technical task. On the contrary, organizational aspects, sharing of knowledge resources, and level of agreement among domain experts are crucial for defining a common knowledge reference. Considering the high quality of expertise present in the different consortia, the risk is very low.

Task 4.2. Metadata annotation of nutritional studies could be affected by the unavailability of correctly annotated nutritional studies. Referring to all the available databases (58 studies are already partly annotated in the Phenotype database) and on in house developed annotation to metadata will lower the risk of failure.

Task 4.3 The risk associated to this task is very low thanks to the availability of well-known and validated data models for pathway annotation (e.g. BridgeDB).

Task 4.4. The consortia involved have great experience on co-reference detection and data merging, definition of semantic rules, and data quality analysis (e.g. imperfect data, data redundancy, co-reference, data merging) as well as expertise on performing meta-analysis. The risk associated to querying data is indeed only associated to nutritional studies collection, thus very low.

Task 4.5. The collaboration with other EU funded initiative keeps the risk very low.

2. Concept and objectives

a. Objectives, vision including scientific/ technological challenges

The main task of this WP will be to harmonize the analysis and querying tools to the benefit of the nutrition community.

Annotating metadata by means of Semantic Web technologies will effectively further help to organize the knowledge generated by modern collaborative research. This part should be developed in close collaboration with ELIXIR where interoperability is a large task. Only the nutrition specific parts should be developed within this project. We will help redirect the users to “best practice” tools enabling effective data analysis and data management solutions during and beyond the project lifecycle.

Standardization is essential for relations with other databases and analysis. This WP will deal mainly with analysis of combined studies, to focus the development of future research on the most relevant biological questions (WP2).

The final aim of a large part of ENPADASI is to dynamically maintain analysis pipelines directly querying the public and custom created databases (WP2, 3), with the aim to correlate data including clinical, ‘omics’, food intake and endpoint data. By identifying, learning from and contributing to best practices from across Europe, predictive models could be developed by integrating functional genomics, metagenomics, nutrigenomics, dietary information and clinical information. This will facilitate nutrition research and thereby increase the knowledge and understanding how food and nutrition can improve human health.

b. State of the art

New ‘omics’ technologies offer a different type of data to nutrition research and stimulates collaboration between research groups. An important challenge for successful collaboration is the management and structured exchange of information that accompanies data-intense technologies. Major efforts for standardization have been achieved in NuGO, the FP6 Network of Excellence “NUtriGenomics Organization”, the major collaborating network in molecular nutritional sciences. NuGO developed and implemented the NuGO Information Network, a distributed system for data exchange based on standard web technology, as a tool for data management and computing infrastructure that supports collaboration between nutrigenomics researchers. This resource does benefit from several efforts that standardize ontologies in the biomedical field (e.g. OBI (ontology for biomedical investigation) and ISA-TAB standard) for the annotation of multi-omics information and their relation with metadata. Proper ontologies should be developed, preferably with partners who are experts in this subject within projects such as ELIXIR, FAIRport and OpenPhacts. NuGO (NutriGenomics Organization) also initiated the (Nutritional) Phenotype database (www.dbnp.org). This application is developed to store nutritional intervention studies with complex design (including cross-over) and is meant to facilitate standardized data output and study comparisons. The system is accessible via a web interface. Flexibility of the system is guaranteed in the system by templates. The templates make it possible to adjust the information that can be stored by adding additional fields to the database via the user interface (if the user is template administrator). The type of the field can be defined by the template administrator, making it possible to store information in text format, dates or via dropdowns. Ontologies are included to facilitate standardization, linking to the ontologies available in Bioportal (https://bioportal.bioontology.org/). For several types of data (mainly related to study design) no standards are yet available. Using the dropdown menus in the system several vocabularies specific for nutritional intervention studies have been developed, these can be used for ontology development in this WP.

Thus, it is crucial to further develop a common ontology for dataset structuring. Centralized approaches for ontology development include work performed in Methontology (Fernández-López M et al., 1997) and On-To-Knowledge (Sure Y et al., 2003). Neither approach envisages the involvement of a community in the process. A second category of methodologies is based on techniques that focus on collaboration and consensus building (e.g. Holsapple et al.; Addison-Wesley, 2002; Karapiperis et al, 2006; Delbecq et al,

1986; Noy et al., 2006). These approaches are mainly oriented towards the social revision and maintenance of existing ontologies, and, provide a limited support for the creation of a new ontology from scratch, as well as its substantial enrichment and extension. The Ontolingua Server (Farquhar et al, 1997) and WebODE (Vega et al., 2001) represents seminal works on distributed Ontology Engineering platforms. The Web version of Protégé (Diaz et al., 2006) is the most widely used ontology editor, which provides an open collaborative platform for ontology modeling and knowledge acquisition. However the collaboration facilities are limited to the concurrent editing of OWL ontologies, leaving out features such as voting and discussions. In the ENPADASI project, several institutes has a wide expertise in data and knowledge representation and management. characterized by a collaborative and social participation approach, for building reference Ontologies: the SemSim semantic similarity search method for the retrieval of semantically annotated digital resources has been herein developed (http://bioconductor.wustl.edu/bioc/html/SemSim.html).

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.

Several partners of ENPADASI already developed algorithms for the transformation of raw *omics* data (e.g. transcriptomics, metagenomics, metabolomics) in parameters suitable for annotation and meta-analysis with other *omics* datasets. Examples are the *foodomics approach* (Bordoni et al., 2013; Capozzi and Bordoni, 2013) and the multiscale approach to investigate type-2 diabetes (Castellani et al., 2013). All types of *omics* datasets require standardized processing and storage to make cross-study comparison possible. Optimal procedures in defining predictive biomarkers from microarray (MAQC-II) and next generation sequencing NGS data (SEQC) have been designed and fully implemented in the EU FP7 HyperDART project (SEQC/MAQC-III Consortium. Nat Biotechnol. 2014)

Analytical pipeline and new analytical methods are necessary to analyse multi-source, multi-level data sets in the fields of systems nutrition. The facilities required in ENPADASI include high-performance clusters to run analysis and many proprietary software tools to help researchers extract knowledge from big data (link to ELIXIR).

The analytical goal is to define multi-omic systems that integrate data for multiple studies to link diet to health. This implies that it possible to use the data from well-structured studies to answer new biological questions based on old data.

In NuGO, a pathway analysis tool EuGene (Cavalieri et al., 2007) has been developed and then improved as Pathway Processor 2.0 (Beltrame et al., 2013), a pathway analysis web application designed to analyze high-throughput datasets, including but not limited to microarray and next-generation sequencing (RNA-seq). The tool can perform two different types of analysis: the first covers the traditional Fisher's Test used by Pathway Processor and topology-aware analyses, which take into account the propagation of changes over the whole structure of a pathway, and the second is a new pathway-based method to investigate differences between phenotypes of interest, like dietary regime, metabolic syndrome as so on.

The development of software tools and databases for the taxonomic and functional analysis of microbial populations living in different environments (i.e. gastro-intestinal tract as well as food samples) and its applications for studying the impact of specific diets, physical exercise as well as its implication in different physiological or pathological conditions is becoming urgent. Pipelines for microbiome analysis have been developed at EoI45-NutriExp hub.

Modeling and predictive machine learning are the ultimate goal of integrative analysis. MLPY Python package for machine learning one of the most downloaded open source systems from the European MLOSS platform for predictive machine learning and most recently Nettools - published as a R package-, and the web interface ReNette have been developed for complex network analysis (https://renette.fbk.eu/, doi: http://dx.doi.org/10.1101/008433).

C. Scientific/ technological concept

WP4 aims to create a common language and integrate the annotations and analysis tools within the ENPADASI activity, focusing on building a Knowledge Hub for joint trans- and multi-disciplinary activities. The system we propose will help the development of a de-centralized databases (WP3), providing a common annotation of experiments and metadata and their integration with flexible tools to share data, also during experiments, while preserving ownership. Standardization and integration will be accompanied in the DB by development of metadata tracking tools searching nutritional datasets generated employing one or a combination of technologies (link with WP3).

The standardization and integration efforts will target both the project structure, the study seen as a unit of research and the assay, the analytical measurement, with the different technologies (i.e. sample characteristics, technology and measurement types, sample-to-data relationships) so that the resulting data and discoveries are reproducible and reusable.

The flexibility of the infrastructure will allow a wide variety of services for data processing and integration by combining several web services, including public services. Sharing resources will boost research standards and will permit full exploitation of the informative potential of these datasets.

Sharing bioinformatics resources and keeping them up to date with the current developments in the field and acquisition of multi-dimensional data will profit all consumers, as they may lead to novel healthier foods, programs for personalized nutrition and self-sustainability of health.

3. Management

The WP4 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 the project much of the management will be performed using teleconferences with involved partners. Annual meetings will be conducted as part of the annual project 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 into public health practice or policy (in 2 years, with a perspective on a longer term).

Building a common language for federating the nutritional databases and parallel interrogation of all the publically funded resources present within the EU will permit a change of paradigm in the way scientific datasets are used by the lawmakers, medical doctors, funding institutions. Our project will grant intelligent access and categorization of big data, including negative results and unpublished information, often unstructured and poorly accessible. The potential outcome will be easier access to medically and nutritionally relevant datasets and exploitation of previously unreleased datasets as well as published ones to the benefit of the wider EU based scientific community.

5. Overall strategy of the work plan

WP4 is organized in 5 tasks, each of which has a task leader responsible for coordinating the activities of the task. The WP leader will oversee the task leaders and keep strict connection with the other WP leaders to ensure the needed level of information flow and collaboration to deliver data integration useful for the community. The survey on the specification of data integration and tools and the testing phase with the end-users are all preventive actions to limit the risk of delivering algorithms and ontologies that is not used. WP4 activity will be performed as collaborative projects with “wet” biology units, to merge the computational and bioinformatics expertise with experimental data.

Task 4.1: Define a common language and building ontologies for nutritional studies - EoI74 and EoI75 [lead], EoI41, 45, 50, 56+66, 71, 73, 75 – Month 1-12

In this task, we will (1) analyse the existing and available resources (e.g. dropdowns of the Phenotype database, documents, datasets) to be used as input for the definition of the common language and ontologies and (2) apply the Ontology Engineering methodology for defining the common language and building the corresponding ontologies on the nutritional field. The construction will be based on: (1) the identification of, and social/collaborative participation of a community of domain experts contributing with their biological backgrounds and resources in the specific domains of their expertise as input towards definition of common annotations; (2) a step-wise approach, that goes through the production of incremental outcomes, such as, for instance, a lexicon, a glossary, a taxonomy, and a semantic network. Defining the common language and building ontologies will see a strong engagement of domain experts, which will be co-ordinated by ontology engineers to clearly identify the purposes of such artefacts, the level of formality detail, as related to the scope. Selected studies in WP2 will show the needs for development of languages and ontologies and the studies uploaded in WP2 should adhere to the developed languages and ontologies (task 2.1 and 2.2). For this reason, WP4 will integrate the languages and ontologies in the infrastructure. We will connect to other initiatives such as ELIXIR to prevent double work and to adhere to common standards.

Related deliverable and milestone: D4.1, MS4.1

Task 4.2: Mapping for terms: metadata integration - EoI74 [lead], EoI38, 50, 56+66, 71 and 75 – Month 1-24

WP2 will identify relevant nutritional studies, some of which will be uploaded already in a database system (e.g. ArrayExpress, Phenotype database, Metabolights or a study specific database). In order to be able to integrate the outcome of these studies, the metadata (i.e. study design description, how data are retrieved etc.) of the different systems will be mapped against the reference ontologies developed in Task 4.1. We will apply a semantics-based annotation/description/enrichment of datasets by using the developed ontology. Application of the Linked Data approach will guarantee a valuable solution, adherence to standard technologies, and a particular attention to issues concerning open access to the data. In collaboration with WP2 (task 2.1 and 2.2) this task will relate to the technical developments accomplished by FAIRport and ELIXIR (Fig 2).

The cross annotation of the experiments with a unique meta-annotation will be used by WP3 for the DB construction.

Related deliverable and milestone: D4.2, MS4.2

Task 4.3: Integration of ontologies and pathways - EoI 45 [lead], EoI50, 56+66, 74 – Month 1-24

In order to map data from different studies on their biological relevance, pathways should be made compliant with the ontologies

developed in task 4.1. Nutritional related pathways will be re-annotated using the ontology terms defined in task 4.1. Pathways from wikipathways (Kelder et al., NAR, 2012) will be represented and re-annotated according to Biolayout using a SBGN compliant architecture (Le Novere et al., 2010), paying attention on integration of the topological structure of the pathways.

Related deliverable and milestone: D4.3, MS4.3

Task 4.4: Querying data and Data Integration - EoI38 and EoI45 [lead], EoI50, 56+66, 74 – Month 1-24

Task 4.4.1: Together with WP3 (task 3.4-3.7) we will develop semantics-based services (specifications and implementation) for querying data by exploiting metadata (semantic annotation) and allowing to extract and integrate different types of information (e.g. retrieving information on dietary intake and phenotypic outcomes). Depending on the complexity level of the built ontology, querying services could include exact matching, similarity matching, and sophisticated reasoning mechanisms for data filtering and checking conditions. Data and case studies will be derived from WP2 (task 2.4). Implementation of methods will enable combined queries on metadata and extraction of relevant information. EoI38 will share their expertise in co-reference detection and data merging. Semantic rules (e.g. definition of words, units, conversion factors, word relations) and data quality analysis (imperfect data, data redundancy, co-reference, data merging) will be developed, taking into account existing standards in biosciences, e.g., ISA framework (Sansone et al. 2012). Specifically, based on a collection of validated algorithms and services, EoI 45 (FBK) will structure, engineer and test an in-cloud prototype version (link to ELIXIR for the hardware needed), including a dedicated web interface and a dedicated web service.

Task 4.4.2: We will develop adequate procedures to integrate data in ENPADASI focusing on algorithms that will facilitate in resolving chronic diseases with lifestyle related solutions. These algorithms will use the queried data of task 4.4.1. EoI50 will share the expertise on performing meta-analysis by using the Phenotype database (www.dbnp.org). The queried data from subtask 4.4.1 will be analysed to come to new biological insights (making use of Meta-analysis), these outcomes will be interpret by WP2 task 2.4. Tools to integrate data of different *-omics* platforms will be further developed. In addition, integration problems of non-absolute measurements (e.g. metabolomics data) will be considered (together with the COSMOS project). The consortium will leverage knowledge in machine learning and development of algorithms to extract information from complex and heterogeneous databases, like pathways analysis. Tools which are broadly applicable will be integrated in the ENPADASI infrastructure of WP3 (see Fig 3). Other tools will be made available via a link on the ENPADASI website. All developed code will also be shared via github (via the phenotype foundation: https://github.com/PhenotypeFoundation) under an open source license (e.g. Apache license). This will facilitate reuse of the code and will make local installation of the developed tools possible.

Related deliverable and milestone: D4.4, MS4.4

Task 4.5: Integration of guidelines - EoI71 [lead], EoI35, 45, 56+66, 75 – Month 1-18

By interacting with WP1 (task 1.2) and WP5 (task 5.4), we will collaborate with the European initiative ELIXIR, ECRIN and with the EU funded initiatives on reporting guidelines (e.g. STROBE, CONSORT) and involve them to improve quality of data reporting.

Related deliverable and milestone: D4.5, MS4.5

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