JOINT PROGRAMMING INITIATIVE – A HEALTHY DIET FOR A HEALTHY LIFE EUROPEAN NUTRITION PHENOTYPE ASSESSMENT AND DATA SHARING INITIATIVE

PROJECT DELIVERABLE REPORT

DELIVERABLE NUMBER AND TITLE	D4.2
TITLE	REPORT DESCRIBING THE CONTRIBUTION TO THE UPDATE METADATA INTO MAIN DATABASES BY MEANS OF ONTOLOGIES
AUTHOR(S)	DUCCIO CAVALIERI, FRANCESCO VITALI, DAMARIZ RIVERO, ROSARIO LOMBARDO, FULVIO MATTIVI, PIETRO FRANCESCHI, FRANCESCO CAPOZZI, ALESSANDRA BORDONI, ALESSIA TRIMIGNO, GIOVANNI FELICI, FRANCESCO TAGLINO, CARL LACHAT, CHEN YANG, NATHALIE DE COCK, LARS OVE DRAGSTED, JILDAU BOUWMAN
WORK PACKAGE	WP 4
TASK	TASK 4.2
WP LEADER	D. CAVALIERI
BENEFICIARIES CONTRIBUTING TO THE DELIVERABLE	CNR-IBIMET, UNIFI, COSBI, FONDAZIONE EDMUND MACH, UNIBO, CNR-IASI, UGENT, UNIVERSITY OF COPENHAGEN, TNO
STATUS – VERSION	FINAL - VERSION 2.0
DELIVERY DATE (MONTH)	M18
SUBMISSION DATE	M20
DISSEMINATION LEVEL – SECURITY*	PU

* Security: PU – *Public*; PP – *Restricted to other programme participants (including JPI Services)*: RE – *Restricted to a group specified by the consortium (including JPI Services)*; CO – *Confidential, only for members of the consortium (including JPI Services)*

** Type: R – *Report*; P – *Prototype*; D – *Demonstrator*; - O - *Other*

DELIVERABLE TYPE**	R

* Security: PU – *Public*; PP – *Restricted to other programme participants (including JPI Services)*: RE – *Restricted to a group specified by the consortium (including JPI Services)*; CO – *Confidential, only for members of the consortium (including JPI Services)*

** Type: R – *Report*; P – *Prototype*; D – *Demonstrator*; - O - *Other*

CONTENTS

INTRODUCTION

With the remarkable progresses in analytical procedures of the last decades, researchers in the biomedical domain are able to produce great amounts of data, usually in many different formats and stored in heterogeneous databases. Efficiently structuring such great amount of information, and clearly state naming and mapping convention, is paramount to exploit such great amount of data, and ultimately produce new knowledge thanks to automated data mining procedures. The necessity to structure and formalize this knowledge adequately led to the development of ontologies.

Often researchers in the biomedical domain refer to an ontology using the terminologies more correctly pertaining to “controlled vocabularies”, “thesauri” (i.e. a list, often organized in a hierarchy or taxonomy, of concepts and their textual descriptions), or “taxonomies” (i.e. a hierarchy consisting of terms denoting classes linked by sub- super-classes relations). A proper ontology, on the other hand, can be defined as a formal representation of knowledge in a certain domain, with definitions of concepts, their attributes and relations between them expressed in terms of axioms in some well-defined logic (Arp, Smith and Spear, 2015, Rubin et al., 2008). Medical ontologies, in information science, are consensus-based controlled vocabularies, with logically formulated annotations of both concepts and their relations to promote automated reasoning (Cimino and Zhu, 2006). An ontology represents a certain domain of knowledge that different people, and computers, can understand it in a similar way (Arp, Smith and Spear, 2015). They consist of defined concepts, typically structured within trees or networks, where a concept is represented by nodes interconnected by semantic relationships (i.e. “is-a”, “part-of”) (Groß et al 2016).

Despite progress at various levels witnessed in nutritional science, a proper ontology is still missing. From literature search and public ontological repositories queries (OBO foundry searched using ONTOBEE, and Bioportal, no proper ontology tailored on the need of nutrition scientist community were found. We were able to find a sole example of nutritional ontology on Bioportal The Bionutrition Ontology (BNO, http://purl.bioontology.org/ontology/BNO) which, however, mostly represents a controlled vocabulary of nutritional terms, without a proper annotation of terms or definition of properties, and lacking in orthogonality (i.e. no terms are imported or referred to external ontologies).

Main goal of our working package (WP4) was the construction of an ontology integrating the practical needs of a coherent description, annotation, and storage of nutritional experiments, with the practical needs of the specific bioinformatics infrastructure of ENPADASI; the DASH-IN. The DASH-IN is a distributed infrastructure able to manage raw data and metadata of nutritional studies, making them available for integrated analysis. To prevent access and retrieval of raw data, for the protection of data ownership and subject privacy, the system is configured so that the data is accessible only to those who have the correct permissions for the required access level. Moreover, non-anonymous data are never disclosed in the form of raw data, but only in the form of data aggregation for integrated analysis. The infrastructure leverages on two different communicating databases systems: a collection of Phenotype Database instances, mainly for the storage of interventional studies, and network of OPAL servers, mainly for the storage of observational studies, that have been seamlessly integrated together in the DASH-IN infrastructure. In the beginning an ontology was needed to harmonize biochemical, genetic, clinical and nutritional concepts typically found in intervention and observational studies in order to provide a coherent mean of data annotation and data query over the distributed infrastructure. Later on, the need for a proper ontology to be used by a broader nutrition community became apparent, so we committed to direct our efforts on the development of a general ontology for annotating nutritional studies.

This working group (WP4) inside the ENPADASI project, already produced a first document (D4.1) presenting the construction of a common vocabulary for the Ontology for Nutritional Studies (ONS). In this document (D4.2), the other aspect of ONS will be illustrated, presenting the concepts followed to structure the set of relations between the terms.

METHODS

ONTOLOGY STRUCTURE

The actions taken for ontology development can be basically divided in two type: i) collection of terms, development of a common vocabulary, and initial structuring of those terms in a class-subclass hierarchy (which was presented in deliverable D4.1 of our working package), ii) structuring of the common vocabulary in a proper ontology, by including more complex relations between terms and by framing it in general high-level terms pertaining to "reality", herein presented.

Figure 1 depict the high – mid levels of the ONS ontology. Each box represents a class and each arrow represents the relation of that class with others. Basic *is a* relation is indicated by dashed arrows, while solid arrows indicate other type of relations, in those cases a text annotation illustrates how the relation should be read. The green colored boxes highlight mid-levels of ONS ontology that where newly defined.

RESULTS AND DISCUSSION

HIGHER-LEVEL ORGANIZATION

The Ontology for nutritional studies (ONS) builds on a subset of the Ontology for Biomedical Investigation (OBI). OBI is an all-purpose ontology for the description of virtually any study in the biomedical domain; ONS directly fetch from OBI the higher level structure. As already illustrated in D4.1, this subset was constructed using ONTODOG tool.

Each ontology directly deals with a certain domain of discourse (in our case nutrition), hence for its nature can deal with very specific concept. Nevertheless, each domain of discourse is framed in the same reality. For this reason the topmost levels of a biomedical ontology should be the same across all ontologies, both for a conceptual reason (i.e. the reality is the same for all) and for practical reason of integration (orthogonality) between ontologies. For this reason, more and more ontologies, OBI included, are using (i.e. importing) the same ontology for their topmost classes: the Basic Formal Ontology (BFO) (Arp, Smith, and Spear, 2015). BFO is a genuine upper-level ontology, designed to be small and to support the creation of ontologies in virtually any domain of science, and to facilitate their integration. We will not go into the details of the BFO ontology, referring the reader to the above quote, but we will provide details on the topmost classes, of this "topmost" ontology. BFO refers to what is real, to everything, with the class "entity", a class that BFO developers considered appropriate both from the philosopher and the scientist point of view. BFO then divide what is real in "Continuants" and "Occurrent". The first class includes terms that continue or persist through time (for example material objects, but also immaterial one, such as a quality), while the second includes terms that occur or happen (for example events or processes)

Classes derived from BFO and imported in ONS are colored in orange in figure 1.

MID-LEVEL ORGANIZATION

As already stated, ONS build upon a subset of OBI ontology; and most of the mid-levels terms are imported as part of this subset. Such an approach has several advantages. First of all, OBI ontology is a well-established and widely used ontology, hence re-using those terms and their relations have the added value of orthogonality of ONS with other ontology. In practical terms, this means that virtually any ontology that built upon OBI and BFO is interoperable with ONS, and could directly imports sections of ONS with minor adjustments. In parallel it also means that ONS can easily imports sections of other ontologies, and easily harmonize them with the rest of the ontology exploiting the presence of the same super-classes from OBI. In addition, ONS includes various mid-terms encompassing different topics (i.e. study design, how data were retrieved, experimental procedures, etc.) that were connected with the nutritional field and were deemed necessary for organizing lower levels terms part of the common vocabulary (D4.1).

The Phenotype database, one of the two databases part of the informatics infrastructure of ENPADASI, was developed as an extension of the, widely accepted, ISA-TAB standard (Sansone et al. 2012). Another advantage of the described approach was that some terms in the OBI ontology are annotated with property "ISA alternative term" which serves as a mapping between OBI ontology terms and ISA-TAB standard terminology. As a consequence, also ONS has a partial mapping with ISA-TAB terminology. Despite that, it is planned as future maintenance of the ontology to extend this mapping to the highest possible number of terms.

Diet, nutrient, foods are the central concepts for an ontology aimed at effectively assist researchers in the standardized description of the nutritional study they are conducting. To a certain extent, they are the central terms in ONS as their definition and relations were not present in other resources. We could say that, if it is true that ONS is very similar to OBI, being a subset of it, those terms are the most nutritional-related, and represents one of the main contributions of ONS respect to other ontologies. Figure 1 illustrates how those concepts were included and connected in ONS. Following, we will describe those terms and their connections.

DIET: definition was borrowed from the Mesh term thesaurus, and was defined as "the regular course of eating and drinking adopted by a person or animal". This concept was further detailed, and three sub-classes were created.

- Usual diet: the regular course of eating and drinking adopted by a population in a certain geographical area, or in a certain cultural setting, or following certain common eating behavior. It is also intended as the diet a person would follow without further prescription or indications, i.e. vegetarian diet
- Prescribed diet: a diet prescribed by a physician/nutritionist to meet specific nutritional needs of a person
- Intervention diet: the diet administered during an intervention study. It usually comprises the adoption of a certain nutritional intervention, intended as the prescription of consuming or not consuming certain food, and follows a precise study design. Intervention studies usually compare at

least two subgroups of a population, one control group receiving a null nutritional intervention, and one or more test groups receiving the intervention

FOOD COMPONENT: definition was borrowed from the CHEBI ontology, and was defined as "Any substance that is distributed in foodstuffs. It includes materials derived from plants or animals, such as vitamins or minerals, as well as environmental contaminants." This concept was further detailed, and three sub-classes were created.

- Nutrient: A nutrient is a food component used by the body for normal physiologic functions that guarantee survival and growth. It must be supplied in adequate and defined amounts from foods consumed within a diet. Malnutrition occurs when the right amount of nutrient is not provided
- Food bioactive: A food bioactive is a food component other than those needed to meet basic human nutritional needs (nutrients). Food bioactives modulates one or more metabolic processes, possibly resulting in the promotion of better health. The daily required intake for food bioactives is not established yet, and there is no demonstration that malnutrition occurs when the right amount is not provided
- Contaminant: Unwanted food component that makes the food no longer suitable for use
- Additive: component added to food to improve or preserve it.

FOOD: food term was actually present in other ontologies, however we did not found a term with a suitable definition. For this reason we defined the food as "The food is a complex matrix that is consumed by a person through the process of eating or drinking. Food are bearer of the nutrients, bio actives and other food components. Food consumption, through the meal consumption, follows a certain dietary pattern, which define the diet. Nutrients and bio actives contained in food can be exploited by the human organism thanks to the process of digestion, absorption, metabolization, or through the intervention of the gut micro flora"

Over this initial scheme, including only *'is_a'* relations, additional complexity has been added with other type of relations and other classes. Following an example trying to put all the concept together:

- Food, which is a processed material entity, is *bearer of* food components. As a material entity it *participates in* the processes of eating and drinking, which *realize* the realizable entity of meal. The meals, their composition and order, *define* the dietary pattern that an individual follows, which is ultimately *related to* its diet. On the other end, the food is the *specified output* of the process of food preparation. This planned process, *realize some* and *concretize some* recipe providing instruction for food preparation, usually *realized by* a person *having role* of cook

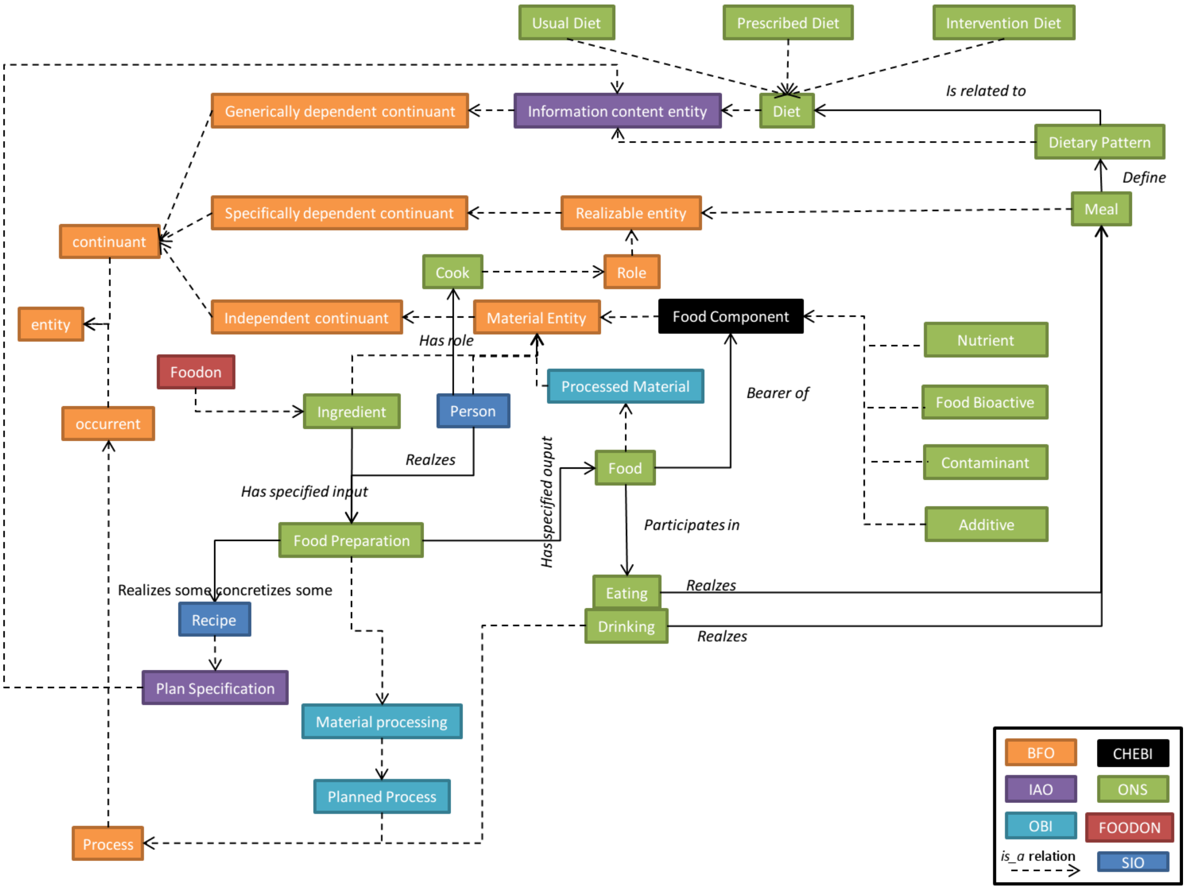

Figure 1: High-Mid Level View of ONS. The presented figure does not include all mid classes of ONS, mainly derived from OBI. Here we only include the higher-level of ONS (in orange), some newly defined mid-level classes meaningful for the nutritional domain (in green), and some other "bridging" mid-level classes form other ontologies (in purple, light blue, black, red, and blue)

Ontology metrics:

Metrics	
Axiom	42981
Logical axiom count	4203
Declaration axioms count	3727
Class count	3450
Object property count	62
Data property count	9
Individual count	104
Annotation Property count	104
DL expressivity	SROIQ(D)
Class axioms	
SubClassOf	3836
EquivalentClasses	81
DisjointClasses	0
GCI count	0
Hidden GCI Count	82
Object property axioms	
SubObjectPropertyOf	32
EquivalentObjectProperties	0
InverseObjectProperties	21
DisjointObjectProperties	0
FunctionalObjectProperty	1
InverseFunctionalObjectProperty	0
TransitiveObjectProperty	6
SymmetricObjectProperty	0
AsymmetricObjectProperty	0
ReflexiveObjectProperty	0
IrrefexiveObjectProperty	1
ObjectPropertyDomain	30
ObjectPropertyRange	31
SubPropertyChainOf	2
Data property axioms	
SubDataPropertyOf	0
EquivalentDataProperties	0
DisjointDataProperties	0
FunctionalDataProperty	1
DataPropertyDomain	8
DataPropertyRange	8
Individual axioms	
ClassAssertion	144
ObjectPropertyAssertion	0
DataPropertyAssertion	1
NegativeObjectPropertyAssertion	0
NegativeDataPropertyAssertion	0
SameIndividual	0
DifferentIndividuals	0
Annotation axioms	
AnnotationAssertion	34694
AnnotationPropertyDomain	0
AnnotationPropertyRangeOf	0

Figure 2: Summary metrics of ONS ontology from Protègè "Ontology metrics" tab.

REFERENCES

- Arp, Robert, Barry Smith, and Andrew D. Spear. Building ontologies with basic formal ontology. Mit Press, 2015.
- Cimino, J. J., and X. Zhu. "The practical impact of ontologies on biomedical informatics." Yearb Med Inform 2006 (2006): 124-35.
- Groß, Anika, Cédric Pruski, and Erhard Rahm. "Evolution of biomedical ontologies and mappings: Overview of recent approaches." Computational and Structural Biotechnology Journal (2016).
- Hoehndorf, Robert, Paul N. Schofield, and Georgios V. Gkoutos. "The role of ontologies in biological and biomedical research: a functional perspective." Briefings in bioinformatics 16.6 (2015): 1069-1080.
- LAMY, Jean-Baptiste. "Ontology-Oriented Programming for Biomedical Informatics." Transforming Healthcare with the Internet of Things: Proceedings of the EFMI Special Topic Conference 2016. Vol. 221. IOS Press, 2016.
- Ong, Edison, and Yongqun He. "Community-based Ontology Development, Annotation and Discussion with MediaWiki extension Ontokiwi and Ontokiwi-based Ontobedia." AMIA Summits on Translational Science Proceedings 2016 (2016): 65.
- Sansone, S. A., Rocca-Serra, P., Field, D., Maguire, E., Taylor, C., Hofmann, O., ... & Begley, K. (2012). Toward interoperable bioscience data. Nature genetics, 44(2), 121-126.