PROJECT DELIVERABLE REPORT

DELIVERABLE NUMBER AND TITLE	D4.3
TITLE	REPORT DESCRIBING THE UPDATE OF NUTRITION PATHWAYS DATA MODEL
AUTHOR(S)	D. CAVALIERI, D. RIVERO, F. VITALI, CHRIS EVELO, MONICA SANTAMARIA GRAZIANO PESOLE, GIORGIO MAGGI
WORK PACKAGE	WP 4
TASK	TASK 4.3
WP LEADER	D. CAVALIERI
BENEFICIARIES CONTRIBUTING TO THE DELIVERABLE	CNR-IBIMET, UNIMAAS, IBBE-CNR, IBBE-CNR, POLIBA
STATUS – VERSION	FINAL - VERSION 1.0
DELIVERY DATE (MONTH)	M24
SUBMISSION DATE	M30
DISSEMINATION LEVEL – SECURITY*	PU
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

The progress of research in nutrition has generated a large amount of datasets and information, described as metadata, both for observational and intervention trials. The promise for the integration of these datasets relies on Nutritional System Biology approaches. Central to the success of such big data integration are efforts to define common languages for sharing information in multidisciplinary areas In order to map data from different studies on their biological relevance. In WP4 pathways currently annotated in Pathway databases were made compliant with the ontologies developed in ONS. To this aim nutritional related terms and biomarkers in ONS were mapped on the datamodels from the major pathway databases and pathway annotation standards. Existing biological networks can be classified into four categories, depending on the nature of their nodes and their interactions: metabolic pathways, molecular interactions, gene regulatory networks and signaling pathways. The major nutrition related pathway repositories considered in this study (among the large number of existing ones) were Wikipathways (Kelder et al., NAR, 2012), KEGG (Kanehisa and Goto, 2000), Reactome (Vastrik et al., 2007), Biocarta (www.biocarta.org), Pathway Commons (www.pathwaycommons.org). The major problem we encountered is the lack of a golden standard on how biological pathways should be represented. A consensus representation of a pathway is important to enable efficient knowledge management and integration of data coming from multiple sources in the nutrition field, in particular integrating the names of the genes and proteins with the names of the reactions in different species. Recent efforts moved from simple graphical representation to machine-readable formats from which a graphic representation could be then generated. On the basis of the existence and the use of graphical and machine- readable formats, pathway representations can be classified into: static-non-modifiable; semi-dynamic, representing information not only as a graphical map, but also using a corresponding machine-readable format, which is not, however, strongly interconnected with the graph; dynamic, where the graphical representation format depends directly on the underlying data model, and thus any modification in the latter can be immediately translated to the former. Our assessment regarding the nutrition related pathways is that those currently stored in public databases are either static or semi-dynamic.

The working group (WP4) inside the ENPADASI project, assessed the existing standards for pathway annotation selecting the most suited for the integration and dynamic relation of network information with the common vocabulary developed into the Ontology for Nutritional Studies (ONS). In this document we will illustrate the consideration behind the choice of the model for integration and the first results obtained.

METHODS

INTEGRATION FO THE PATHWAY DATA MODELS IN THE ONS STRUCTURE

We reviewed existing literature regarding existing data models and assessed the most suitable writing a report justifying the solution proposed.

RESULTS AND DISCUSSION

DATA MODEL ORGANIZATION

The most successful standard is currently the System Biology Graphical Notation (SBGN) (Le Novere et al., 2009). SBGN splits the representation of a biological network into three different levels (the process definition, the entity relationship and the activity flow language). The three representations are constructed in order to capture different aspects of the biological systems, defining a set of glyphs and constraints to reduce ambiguity and improve interpretation. The SBGN Process Diagram specification defines a comprehensive set of symbols with precise semantics, together with detailed syntactic rules defining their use. It also describes how such graphical information is to be interpreted. Process diagrams show how different entities in a system transition from one form to another, as a result of chemical reactions or interactions of biological entities. SBGN has two types of symbols, nodes and arcs that characterize the quantitative effect of a substance on a process or vice versa. The symbols that represent concepts in SBGN are uniquely identified by terms from the Systems Biology Ontology (SBO), thus to integrate ONS into SBGN we simply mapped all the ONS terms on SBO, defining which term is a node and which an arc. Biomarkers from metabolomics studies derived from WP2 and WP3 become Entity pool nodes (EPNs) that represent molecules, defined with distinct glyphs: unspecified entity, simple chemical, macromolecule, nucleic acid feature, perturbing agent (such as light, temperature, etc.) and source and sink. Semantics of EPNs can be modified by auxiliary units, which represent a particular state. Finally, tags can be used to identify an EPN used in two or more physically different maps, thereby allowing the modular decomposition of diagrams. Process nodes (PNs) describe the way in which molecules taking part in a reaction, including nutrients or metabolites present in food, blood, urine or faeces, are transformed. One of the tasks we undertook was to integrate ONS framework and ontological construction into BCML machine readable format (Beltrame et al 2011). In addition to a full implementation of the SBGN specification, BCML provides a series of optional features (defined as extensions of the main schema). First of all, BCML can include additional information on the entities that compose the network: each entity be described by a series of database identifiers, e.g. Entrez Gene or Uniprot accession numbers and each species can have its independent set of identifiers. Furthermore, condition-specific information, called ‘Findings’, can be associated to each entity or reaction. ‘Findings’ are collections of biological information that are relevant to that entity or reaction. The current specification includes support for organism, organism part (tissue), cell type, the specific biological environment in which the evidence was proven and the type of the experiment used to gather evidence. To reduce ambiguity and promote consistency among different ‘Findings’, the schema enforces a controlled vocabulary built from current medical ontologies. Nutritional pathways integrating ONS ontologies can define five PNs: process (used to represent most of the transformations between EPNs), omitted process, uncertain process, association, and dissociation The Nutritional Phenotype, is classified as a higher level PN, which can be modulated but does not consume or produce anything, since it has been annotated as a final stage, a manifestation of the metabolic activity in the underlying PNs, by connecting arcs link EPNs and PNs, indicating how entities influence processes. In addition to consumption and production arcs, which indicate the effect on the flux of matter through PNs, specific arcs can represent different possible modifications of a process, such as modulation, stimulation, catalysis, inhibition and trigger (or absolute activation). ONS driven logical operators provide the means of indicating boolean combinations of influences from EPNs onto PNs. The three possibilities are conjunction (and), disjunction (or), and negation (not). The BCML schema also provides support to split pathways into subpathways called ‘macro modules’, representing independent units of a signaling section. The structure and compartimentalization of the processes is a multilayer of submodules "folded" in the main module, represented by a symbol. The unfolded submodules can be retrieved, for instance, as files or web pages, or printed on paper.

ONS integrated BCML also contains support for a number of graphical hints, such as border, background and text colors of the elements (while the original SBGN specification is monochromatic). These hints are recognized and processed by the tools we developed that can read and parse BCML files. Since BCMl files are XML files they already contain all the needed information, which is then used by the software suite to produce a SBGN-compliant graph. The current implementation converts BCML files into GraphML, a widely used format for graph representation. BCML files converted to GraphML can be opened by programs such as the yEd graph

editor, or SPIA where they can be exported to vector graphs or bitmap images, or imported and visualized in PathVisio.

Lastly, the BCML format can incorporate any kind of experimental measurements that can be matched to the identifiers of an element. This permits, for example, to map high-throughput data coming from transcriptomics or proteomics experiments and to determine which elements of the pathway are affected in a given condition/tissue/organism. Measurements can be coupled with graphical hints so that when the pathway is converted to graphical representation, elements with experimental measurements will be colored accordingly. The presence of identifiers associated to the entities of a network described with BCML permits the transformation of the pathway in different data format suitable for data analysis. The tools provided with the suite permit the extraction of identifier (gene) lists from a BCML file, enabling their use with analysis methods such as Gene Set Enrichment Analysis (GSEA) and Fisher's exact test. Additionally, the format can be converted in a form amenable for impact analysis through the SPIA R package, Pathway Processor, SPIA or Pathway Inspector enabling a topology-aware analysis of the network.

REFERENCES

Kanehisa,M. and Goto,S. (2000) KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.*, **28**, 27–30.

Novere,N. *et al.* (2009) The Systems Biology Graphical Notation. *Nat. Biotechnol.*, **27**, 735–741. Li,E. and Davidson,E.H. (2009) Building developmental gene regulatory networks.

Beltrame L, Calura E, Popovici RR, Rizzetto L, Guedez DR, Donato M, Romualdi C, Draghici S, Cavalieri D. The Biological Connection Markup Language: a SBGN-compliant format for visualization, filtering and analysis of biological pathways. Bioinformatics. 2011 Aug 1;27(15):2127-33.

Pico,A.R. *et al.* (2008) WikiPathways: pathway editing for the people. *PLoS Biol.*, **6**, e184.

Sales G, Calura E, Cavalieri D, Romualdi C. graphite – a Bioconductor package to convert pathway topology to gene network. BMC Bioinformatics. 2012 Jan 31;13(1):20.

Vastrik,I. *et al.* (2007) Reactome: a knowledge base of biologic pathways and processes. *Genome Biol.*, **8**, R39.

Draghici,S. *et al.* (2007) A systems biology approach for pathway level analysis. *Genome Res.*, **17**, 1537–1545. Grosu,P. *et al.* (2002).

Hucka,M. *et al.* (2003) The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. *Bioinformatics*, **19**, 524–53.

Beltrame L, Bianco L, Fontana P, Cavalieri D. Pathway Processor 2.0: a web resource for pathway-based analysis of high-throughput data. Bioinformatics. 2013 Jun 8.

Bianco L, Riccadonna S, Lavezzo E, Falda M, Formentin E, Cavalieri D, Toppo S, Fontana P. Pathway Inspector: a pathway based web application for RNAseq analysis of model and non-model organisms. Bioinformatics. 2016 Oct 6.

.