PROJECT DELIVERABLE REPORT

DELIVERABLE NUMBER AND TITLE	D4.4
TITLE	A REPORT ON PROTOCOLS FOR QUERYING DATA, DATA INTEGRATION AND USAGE OF PATHWAY TOOLS
AUTHOR(S)	DUCCIO CAVALIERI, FRANCESCO VITALI, DAMARIZ RIVERO, ROSARIO LOMBARDO, JILDAU BOUWMAN, M. MONICA SANTAMARIA, GRAZIANO PESOLE BACHIR BALECH, GIORGIO MAGGI
WORK PACKAGE	WP 4
TASK	TASK 4.4 (T4.4.1 AND T4.4.2)
WP LEADER	D. CAVALIERI
BENEFICIARIES CONTRIBUTING TO THE DELIVERABLE	CNR-IBIMET, COSBI, TNO, IBBA, IBBE, POLIBA, UNIFI
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

SUB-TASK 4.4.1

SCOPE

Development of procedures for semantic-based query of metadata; including a dedicated web interface and a dedicated web service.

CONNECTION WITH OTHER WP

Activities carried out for the tasks presented in this deliverable document (Task 4.4.1 and Task 4.4.2), were carried out in strict collaboration with WP3, as many action were overlapping and concomitant. Relevant information can also be found in deliverable document of WP3, namely D3.2.1 and D3.3.1.

METADATA TERMS

As already indicated in D4.1, the set of terms for metadata query was collected exploiting the "Metadata Terms collection" quest started by Rosario Lombardo (WP3) in April 2017. The "quest" was rather straightforward, asking to all the ENPADASI consortium members, which terms were considered of major importance to conduct a query over the distributed infrastructure DASH-IN. In other words, this question has been asked "While querying the DASH-IN for studies relevant to your research, for example study similar to a one you uploaded to the system and on which you want to conduct a federated analysis, which terms would you search for?".

Collected terms, if not already present, were inserted in the growing ONS ontology, following the same approach detailed in D4.1. Here we report the figure summarizing the approach, a modification of figure 3 from D4.1.

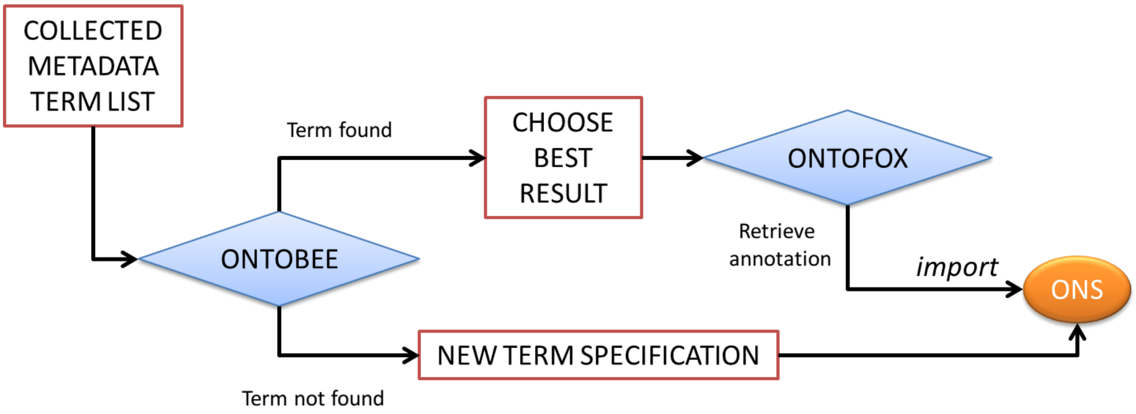


Figure 1: Schematic representation of the methodology used for the insertion of metadata search terms into ONS

Briefly, for each term we firstly checked its presence in other ontologies using ONTOBEE project web service (http://www.ontobee.org/), and defined it as new if no suitable matches were found. ONTOFOX web service (http://ontofox.hegroup.org/) with *includeAllAnnotations* option, was then used to fetch terms annotation and finally import terms in ONS.

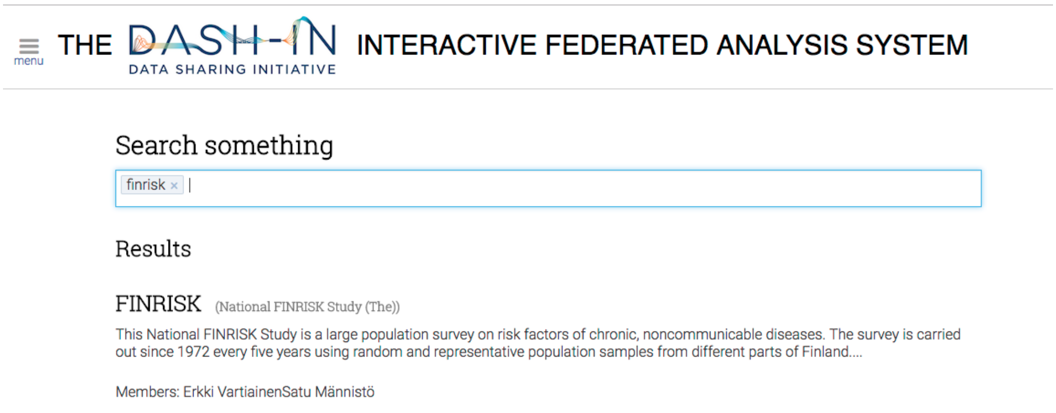
METADATA SEARCH TOOL (MUTUTED FROM D3.2.1)

As the DASH-IN infrastructure developed by WP3 is growing in number of studies and storage sites across Europe, the need for a specific metadata search tool has arisen to allow finding studies whose characteristics best match the user's research questions. The metadata search tool is presented to the user as a simple search-box available at a central DASH-IN portal. The central portal interacts with 2 families of servers spread across Europe:

a) a number of independent Phenotype Databases (DBNP, for intervention and observational studies). Studies from different instances of DBNP can be found by a "search" access credential setup for each DBNP server. Metadata for public studies are accessible even without credentials. Appropriate users can search account-protected studies. Studies data cannot be accessed by the search tool itself.

b) a central Mica server which contain metadata information referring to studies stored on multiple Opal servers (for observational studies). Studies from the central MICA are found by providing a MICA credential. Only studies-related metadata are stored on MICA server and not the data itself. Therefore, studies cannot be accessed from the search tool.

The central search box (Figure) will allow to search user-provided query terms contained both in the Mica server (acting a central metadata point for OPAL servers) and in each independent DBNP, which have no centralized concept. Data-uploaders/ data-curators can annotate their studies with the metadata terms for each of the studies in the respective DB system (either DBNP or MICA). This action makes the corresponding studies visible from the metadata search-box at the central Dash-In server. The search results consist of a list of study entries each containing only metadata information on the studies, and particularly the owner/server allowing to get in contact and obtain credentials to access the data (if not already available to the user) .



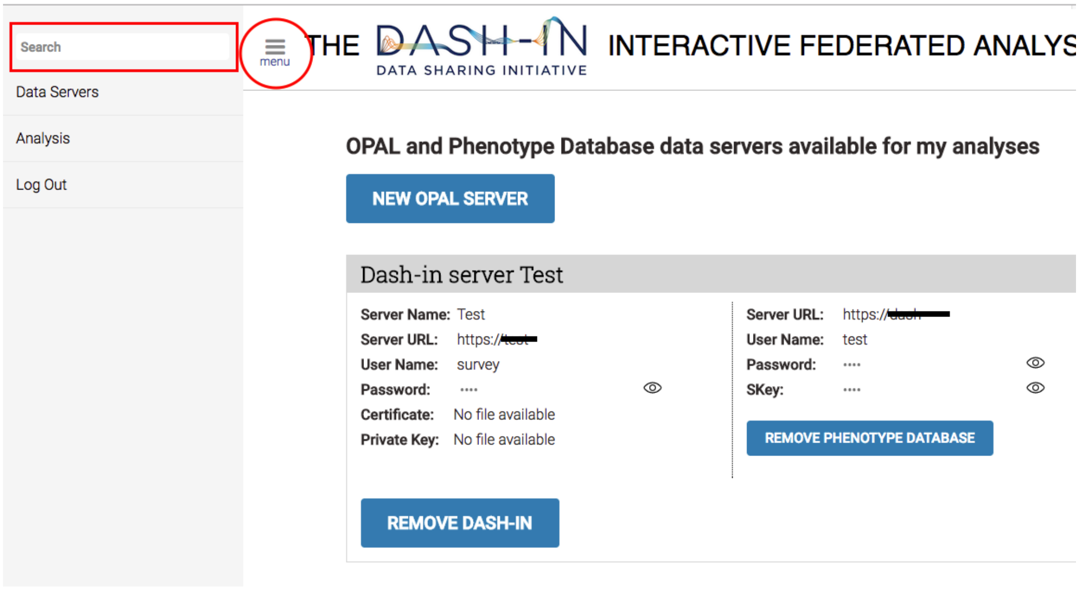
FINRISK (National FINRISK Study (The))

This National FINRISK Study is a large population survey on risk factors of chronic, noncommunicable diseases. The survey is carried out since 1972 every five years using random and representative population samples from different parts of Finland....

Members: Erkki VartiainenSatu Männistö

Figure 2: Study metadata search performed using the DASH-IN infrastructure looking for a sample data set loaded on the MICA central metadata server.

The task force identified a set of study search terms to be used for both study annotation and study search which are based on WP4 ontologies but might be extended with terms derived from: minimal study requirements (WP2), study quality appraisal tool (WP2) and potential biomarker ontologies from shared partnerships with the JPI HDHL FoodBALL project.



OPAL and Phenotype Database data servers available for my analyses

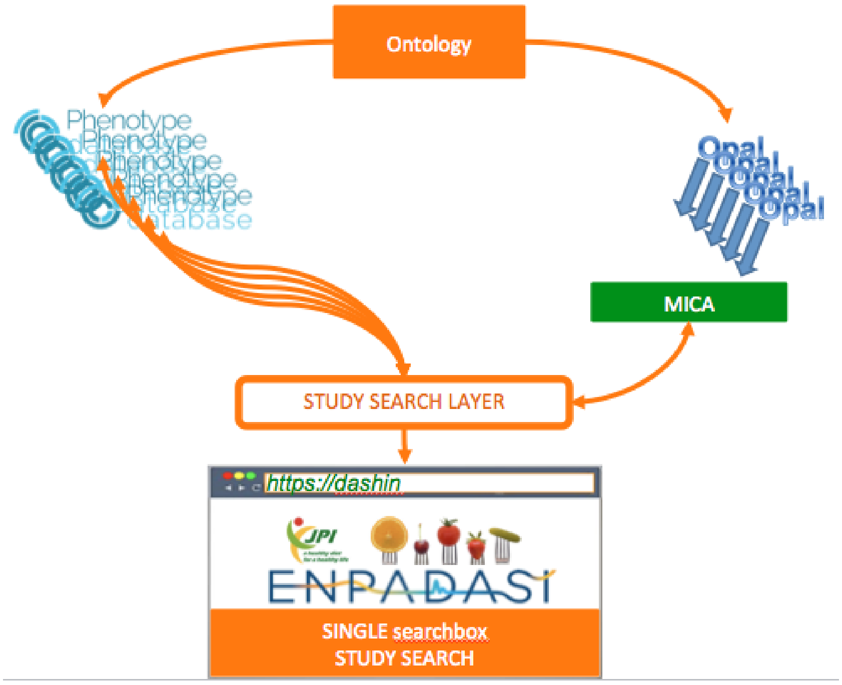
Dash-in server Test

Server Name: Test
Server URL: https://[redacted]
User Name: survey
Password: ····
Certificate: No file available
Private Key: No file available

Server URL: https://[redacted]
User Name: test
Password: ····
SKey: ····

Figure 3: Clicking on the top-left menu icon a sidebar menu slides in. The first search box gives access to the metadata search feature that allows to enter predefined metadata search terms to finely identiy available studies annotated with the same search terms.

Overview of the metadata search tool where the constituent components from WP3 infrastructure are interacting together to provide the best data for the user research needs.



The above schema also indicates the role of ONS ontology in integrating and harmonizing the two databases system of the distributed infrastructure.

SUB-TASK 4.4.2

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. 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.

SCOPE

Using developed query tools (D4.4.1) search among the available real case studies (D2.4.1) uploaded in DASH-IN, for adequate studies and to perform, leveraging on the developed infrastructure (Task 3.2 and D3.2.1), an integrated analysis to finally come to new biological insights.

IMPLEMENTATION AND DRAWBACKS

The infrastructure for federated data analysis is extensively presented in D3.2.1 from WP3.

Beside the informatics distributed infrastructure organization, which lies behind the surface, the final user can interact with the DASH-IN by means of a simple and user-friendly Shiny web app. The Shinyapp allows interactive online access to DataSHIELD -based analyses that include, among others, summarizes, plots and regression-based analyses. The Shinyapp actually allows explorative analyses from federated data in the form of Histograms, Contour Plots, and Heatmaps. All data are non-disclosed because no single pixel or number is ever returned on the web page if that is not the result of a summary data for at least 5 different individuals. Moreover, it allows more complex analysis of phenotype-trait associations with linear regressions and generalized linear models for the binomial and Poisson distributions. The analytical results are the aggregation of statistics computed at each hosting institution increasing the power while at the same time ensuring complete data privacy and confidentiality as the data never leaves the hosting data servers.

One of the main challenges encountered during the course of ENPADASI project, was the substantial funding delay. This problem has affected also other partners (WP3 and WP2) and has ultimately brought to a replanning of the initially proposed activities. Despite the renewal of the ENPADASI project deadline, likely due to perceived ethic and intellectual property issues, the upload of real studies to the DASH-IN system and the definition of a common case study was strongly delayed. For other reference see D2.4.1

This task (as well as other tasks of other WPs) was originally designed as an highly interconnected, integrating the actions of all the WPs. For this reason, the above mentioned delay deeply affected the actions to be taken for carrying out this subtask. Moreover, the absence of a real case study has limited our ability to formulate hypotheses for the integrated analysis, and thus has limited our capacity to integrate new algorithms that will facilitate in resolving chronic diseases with lifestyle related solutions.

OUTLOOK AND FUTURE DEVELOPMENT

As soon as the number of nutritional studies stored in the ENPADASI DASH-IN system will grow, more and more complex analysis will become possible. At this moment, for now we have just shown the possibilities of the system.