



PROJECT DELIVERABLE REPORT

DELIVERABLE NUMBER AND TITLE	D3.2.1
TITLE	SPECIFICATION OF THE INFRASTRUCTURE
AUTHOR(S)	R. LOMBARDO, F. JAGERS, N. GILLAIN, H. BOEING, DR. U. HARTTIG, N. HULSTAERT, E. HUSSON, F. SANDØ, B. BALECH, M. SANTAMARIA
WORK PACKAGE	WP 3
TASK	TASK 3.2
WP LEADER	R. LOMBARDO
BENEFICIARIES CONTRIBUTING TO THE DELIVERABLE	COSBI, TNO, ULG, DIFE, UGENT, UCPH, CNR-IBBE, UNIBA
STATUS – VERSION	FINAL - VERSION 1.0
DELIVERY DATE (MONTH)	M18
SUBMISSION DATE	MAY 2017
DISSEMINATION LEVEL – SECURITY*	CO
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*



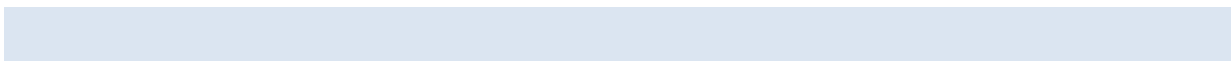
CONTENTS

Introduction	5
Scope of the work	5
Concept and objectives.....	5
Objectives, vision including scientific/ technological challenges	5
State of the art	6
Scientific/ technological concept.....	6
Potential impact.....	6
Overall strategy of the work plan	6
High level architectural overview	8
Introduction	8
Data Safe Havens	8
ENPADASI infrastructure.....	8
Dash-In infrastructure.....	10
Systems used.....	10
OPAL.....	10
Datashield	15
Phenotype database	18
Interactions between the systems.....	18
Datashield RServer.....	19
dbNP connector	20
The DASH-IN Interactive Federated Analysis system.....	21
Searching studies relevant for a research question.....	23
Setting up data sources.....	26
Federated explorative plots.....	28



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

Federated data analysis	30
Technical Specifications	33
OPAL.....	33
Installation prerequisites	33
Opal schematic overview	33
Opal database installation	33
DataSHIELD	35
DataSHIELD installation and usage	35
MICA.....	35
Mica schematic overview.....	35
Mica modules.....	36
Some Mica-server installation remarks	36
Mica demo's.....	36
Mica-python-client.....	37
Harmonization	37
Harmonization across Opal instances.....	37
Harmonized DataSchema.....	38
Harmonized DataSchema variable	38
More complex algorithm example.....	39
Harmonization dataset example.....	40
Phenotype database	40
Shinyapps	40
Installing the required software components	40
Linking into the Dash-In infrastructure.....	41
Server configuration and deployment of a multi-application server	43

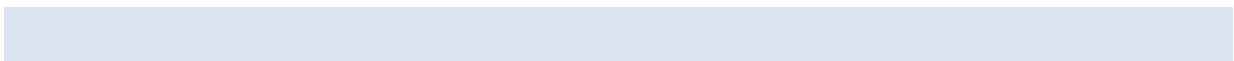




JOINT PROGRAMMING INITIATIVE – A HEALTHY DIET FOR A HEALTHY LIFE EUROPEAN NUTRITION PHENOTYPE ASSESSMENT AND DATA SHARING INITIATIVE

MAC..... 45

Availability of the systems 47





INTRODUCTION

SCOPE OF THE WORK

The purpose of this work is to connect and further develop data infrastructure for enhancing data sharing and exchange for systems nutrition research in order to facilitate research and reuse of data. This infrastructure should be scalable and sustainable over time. The work of EuroDISH, where the needs for research infrastructures are mapped, and the inventory of WP2 have been used as a basis for the design and the focus lied on the development of data solutions of nutrition specific data problems. The data solution has to facilitate both data storage and analysis. The initial work was to develop a functional and technical design, which will be based on the work of WP2, WP3 and WP4. This functional and technical design document forms the basis for the implementation of a first version of the data infrastructure that will be available to the systems nutrition community, which have to be tested by all partners of WP2. The usability of the user interface will be analyzed by a user survey, which will be filled out by all partners of WP2. The results will be shared with WP4 for further development of the analysis tools and bugs and feature requests from these tests will be used for further development of the infrastructure.

CONCEPT AND OBJECTIVES

OBJECTIVES, VISION INCLUDING SCIENTIFIC/ TECHNOLOGICAL CHALLENGES

The main objective of this work was to develop an infrastructure that allows the exchange of data in the field of systems nutrition at European level. A design phase in such a development is a fundamental step to move to the implementation of the infrastructure in such a way that different teams can work simultaneously (using github as sharing and versioning systems for communication and to prevent legacy problems). The main challenge was to identify the requirements of different groups (bioinformaticians, nutritionists, clinicians) in order to capture all the relevant information needed to capture experimental and observational studies in nutrition. From a technological perspective, we have to integrate several existing databases that store data in different format and that are geographically located in different countries. The challenges we faced here were the typical ones for the development of distributed system/ federated database, which require communication between the systems using web services (such as the Phenotype database). An additional difficulty is that the users of the infrastructures may have a limited computing literacy so that the user interface is the bottleneck for the adoption of the interface.



STATE OF THE ART

Biological data is heterogeneous in terms of format, protocols of transmission and interpretation. Providing an integrated framework to deal with the complexity of multi-level analyses is a strategic issue from both the computational and the algorithmic design point of view.

The development and the implementation of user friendly solutions is a necessary step towards the efficient and widespread use of these instruments. The ENPADASI services will aim to overcome existing limitations due to hardware resources and algorithmic limitations and will make them suitable to run in a shared environment such as the ENPADASI federation of institutions and research centres.

SCIENTIFIC/ TECHNOLOGICAL CONCEPT

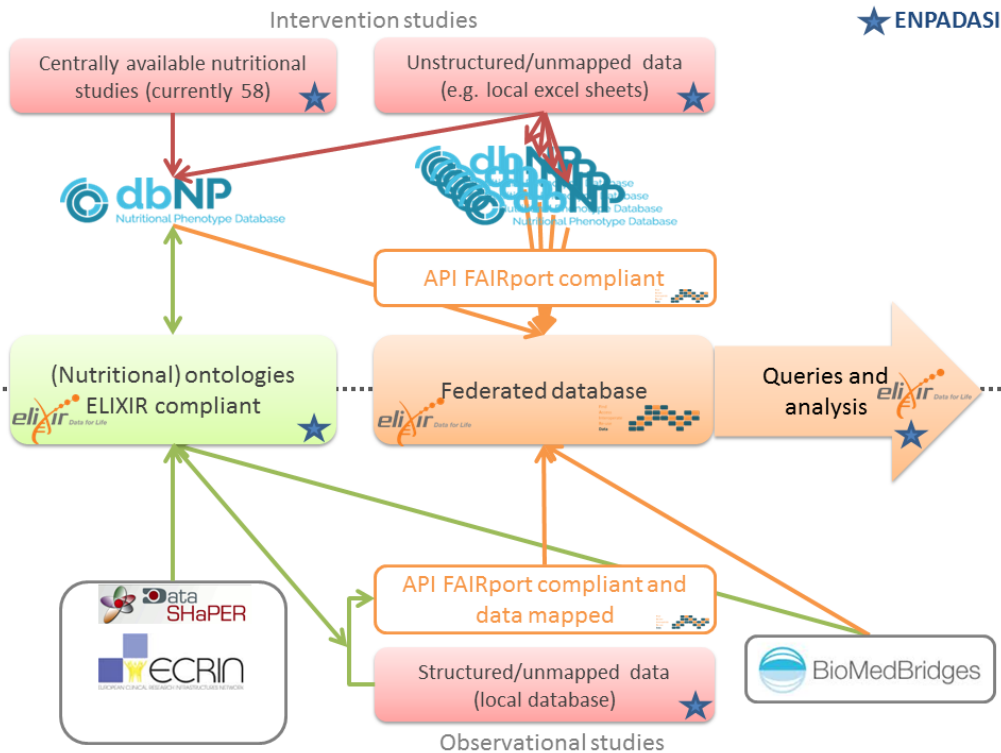
The use of distributed infrastructures, such as local computing clusters or the geographically distributed Computing Grid facilities, has been the first answer to the needs of a continuously growing community of biological data analysts, with interesting recent developments. The next step which naturally arises is the involvement of Cloud technologies. In fact, combination of workflow approach and cloud computing paradigms could facilitate big data access and processing, addressing high computation and complex storage issues coming from actual research. It is also necessary to provide technical capabilities to federate genetic and clinical data currently stored in hospitals, research archives and databases while guaranteeing strong protection for sensitive information. The implementation of such ideas within a shared platform will allow the ENPADASI Project to exploit large-scale initiatives and create data acquisition standards that could later be used by institutions and research centres to provide a common “language” and a fruitful interaction.

POTENTIAL IMPACT

The availability of a European infrastructure for nutrition research will enhance the capacity of small labs or purely biological labs to carry out high-impact research. It will constitute an atlas of findings in the systems nutrition area that will be useful to improve the health system. This possibility is even increased by the presence in the consortium of clinicians that will also help structuring the infrastructure in such a way that it is suitable to support translational medicine. The Phenotype database, developed by the nutritional community, is implemented in the Dutch initiative CTMM/TraIT), which focuses on medical translational research.

OVERALL STRATEGY OF THE WORK PLAN

Our first task was to consolidate the mapping work of EuroDISH and investigate the available data resources available within and relevant to this consortium (e.g. Phenotype database, www.metabolomexchange.org) with input from WP2. These data resources have to be integrated in the one infrastructure of the project which is based on the FAIRport concept (see figure below) and connected to relevant pipelines.



Then we produced a specification of the data infrastructure based on work done earlier in the field (e.g. Phenotype database). This has to be done in strict collaboration with the biologists (end-users) of the consortium and comparing our development with existing infrastructure for systems biology (e.g. SBML, OpenBEL, Combine community, etc.). The infrastructure has to connect to the ongoing work in complementary efforts (JPI DEDIPAC, ELIXIR, etc.) and will integrate the relevant structured data in the consortium. Moreover, we evaluated the cross-integrations with the EBI Metagenomic portal and the activities to be carried out in the EXCELERATE project within the ELIXIR RI framework. For intervention studies the infrastructure will 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) and measurement data (e.g. clinical chemistry, anthropometry, etc). This database makes use of templates and therefore will require limited adjustments depending on the selected study cases. Observational studies, for ethical reasons (see WP5), are generally stored in local databases; these will be made available by inclusion of an API.

Based on the specification, the interface (with sketches) of the common access point to the infrastructure has to be designed and a survey for usability and easiness of use with end-user has to be executed. The final goal of this task is to specify the interface that will be then implemented and tested.

Thanks to the testing phase recommendations, the first release of the infrastructure will be updated.



HIGH LEVEL ARCHITECTURAL OVERVIEW

INTRODUCTION

DATA SAFE HAVENS

Table 1. Proposed criteria for a Data Safe Haven

Data maintenance and release should be socially acceptable and appropriate

Criterion 1	Consistent with formal ethical and legal requirements
Criterion 2	Responsive to emerging issues
Criterion 3	Discoverable and accessible
Criterion 4	Transparent and accountable

Data should be veritable

Criterion 5	Data and metadata fidelity
Criterion 6	Quality assurance and control
Criterion 7	Curation and archiving
Criterion 8	Reliable availability including backup
Criterion 9	Effective audit

Data should be safe and secure

Criterion 10	Preserve confidentiality, integrity and availability of the repository
Criterion 11	Appropriate secure access to individually identifying data
Criterion 12	Appropriate protection of individually identifying data

Bioinformatics, 31(20), 2015, 3241–3248
 Data Safe Havens in health research and healthcare
 Paul R. Burton, Madeleine J. Murtagh, Andy Boyd, James B. Williams, Edward S. Dove, Susan E. Wallace, Anne-Marie Tasse, Julian Little, Rex L. Chisholm, Amadou Gaye, Kristian Hvee8, Anthony J. Brooke, Pat Goodwin, Jon Fistein, Martin Bobrow and Bartha M. Knoppers

Data Safe Havens

ENPADASI INFRASTRUCTURE

There are 2 kinds of studies that have to be covered by the ENPADASI infrastructure:



Observational studies

Large studies in terms of numbers
Often linked data (pseudo-anonymous)
Own steering bodies with many formal rules regarding collaboration and data sharing
BIOShare-project
Increasing data collection over time and sequential use of stored biomaterials

Intervention studies

Small studies in terms of numbers
Often not linked data (anonymous)
PI-specific rules regarding collaboration and data sharing
NUGO-collaboration
Detailed phenotypic data per subject

Intervention studies are covered by the Phenotype database.

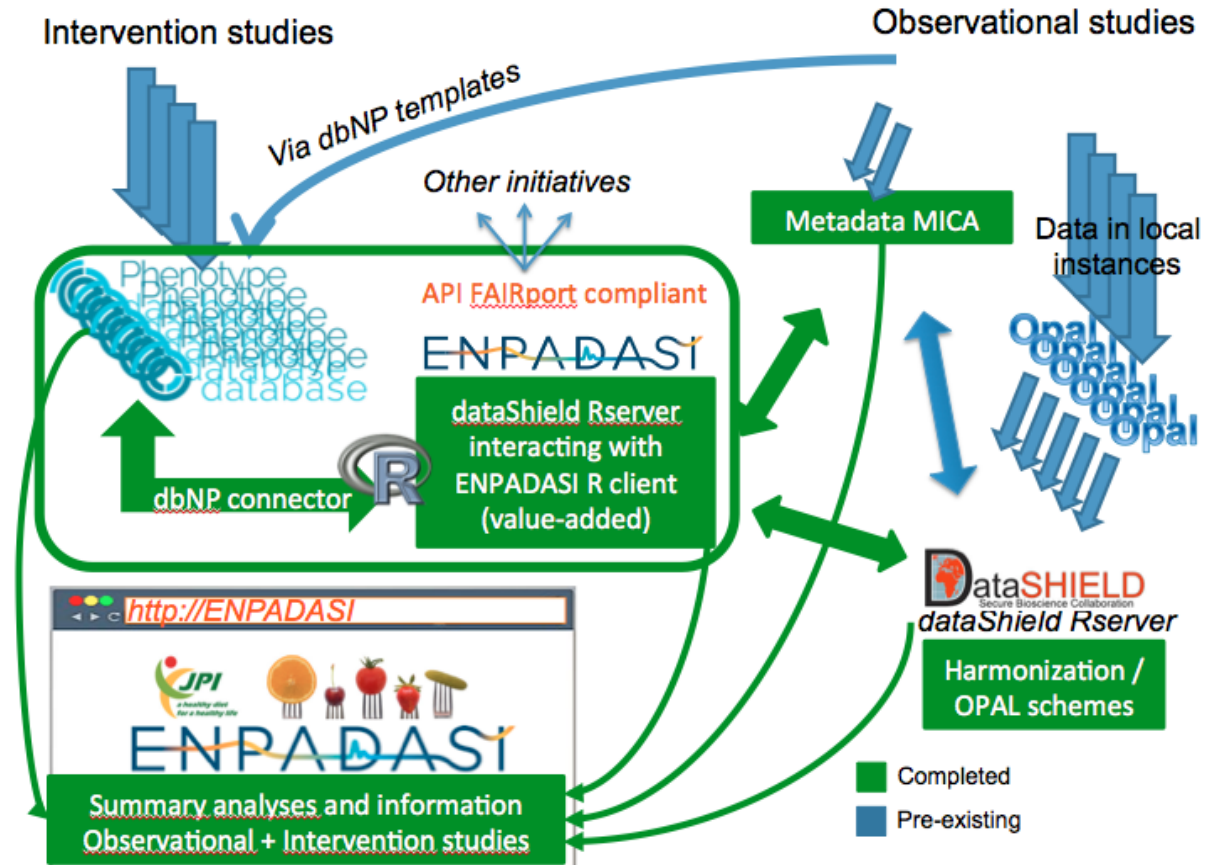
For observational studies, among a set of existing solutions, BioSHaRE was the infrastructure of choice.

Biobank Standardization and Harmonization for Research Excellence in the European Union (BioSHaRE-EU) is a European FP7 project funded from 2010 to 2015. BioSHaRE produced tools and methods for data harmonization and standardization, data sharing and analysis across multiple biobanks and databases.

BioSHaRE generated three types of foreground outcome: **tools**, **data** and **knowledge**. Software **tools** are developed for cohort studies and related databases to manage, present, catalogue, secure and share their data, facilitating a wider use and re-use of their data. The same tools are useful for the scientific community to explore and use the available data in an efficient and secure manner. In addition, broadly applicable guidance and recommendations have been made for sample handling and Ethical, legal and social implications (ELSI) of biobank research and sharing of data. A comprehensive overview of the tools and methods is available in the BioSHaRE Catalogue of tools and methods for data sharing. ENPADASI will use the OPAL/MICA combination from this catalogue for the storage of observational studies, data will be stored on opal servers. MICA will be used to share metadata. DataSHIELD will be used to analyze the data through R, as it facilitates the federated analysis. Pseudo-anonymized data cannot be disclosed in some of the ENPADASI countries and therefore requires a system that can work in a federated way.

The Phenotype database has been chosen for intervention studies (and/or anonymized observational studies) as it was explicitly developed for this purpose.

DASH-IN INFRASTRUCTURE



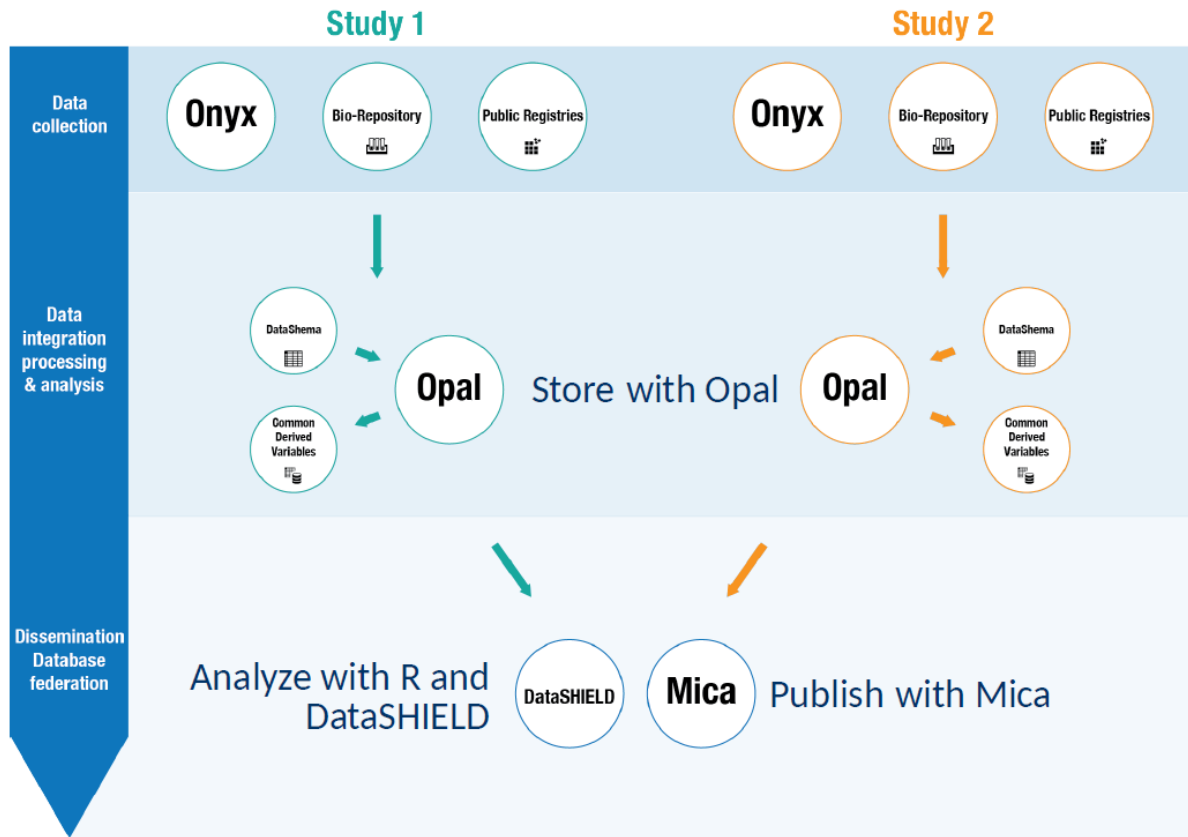
Overview of the Dash-In infrastructure

As it has been said before, 2 different systems will be used for intervention and observation studies. In order to be able to combine data, a dbNP connector has been developed to allow the Rserver to analyze both Phenotype databases and OPAL databases. Analyses are done on the local instances so that raw data are never shared therefore enabling a complete federate data analysis network. Finally, Shinyapp will be used to give users an easy, interactive online access to DataSHIELD-based analyses that include, among others, summarizes, plots and regression-based analyses.

SYSTEMS USED

OPAL

Opal is OBiBa's core data warehouse. This application provides all the necessary tools to import, transform and describe data. Subject's identifiers can also be managed at data import and export time.



<http://www.obiba.org/>

ObiBa applications suite

ANALYSIS WITH OPAL

Thanks to its integration with R, complex statistical analysis and reports can be performed. The implementation of the DataSHIELD process allows advanced statistical data analysis across multiple studies *without sharing and disclosing any individual-level data*.

INTEGRATION OF DATA IN OPAL

Being integrated with Mica, studies using Opal can seamlessly and securely import data into web data portals created with Mica that query Opal databases to obtain real-time aggregated reports on subject's data. Secured REST web services are also available allowing to automate server management (Python command line tools) or to access to data (from R or any tools that are web-capable).

OPAL'S FEATURES DATA WAREHOUSE



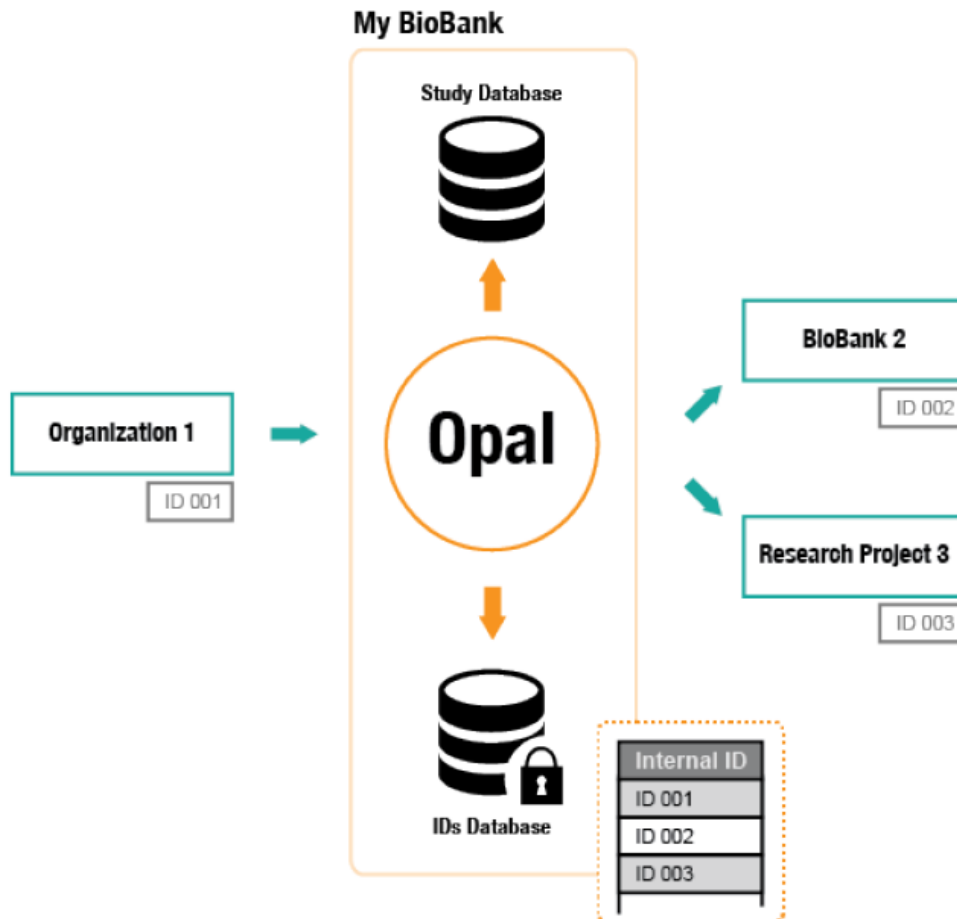
Here are some of the main features of the Opal's data warehouse technologies: Store data on an unlimited number of variables, Support MongoDB (recommended) and Mysql as database software backend, Customized variable dictionaries. Import data from CSV, SPSS and Opal XML file formats, Incremental data importation, connect directly to multiple data source software such as SQL databases and LimeSurvey, Store data about any type of "entity", such as subject, sample, geographic area, etc., Store data of any type (e.g., texts, numbers, geo-localisation, images, videos, etc.), Advanced indexing functionality using Elasticsearch.

OPAL'S VIEWS AND DERIVED VARIABLES

Opal provides the software infrastructure to create virtual tables called "views" of derived variables that can be persisted on disk or exported into files. Main features are: Comprehensive JavaScript library of util functions commonly used to derive new variables (e.g. unit conversion) See Magma Javascript API. User-friendly interfaces to recode variables without programming, Instant summary statistics computation of the new derived variables.

PRIVACY, CONFIDENTIALITY AND SECURITY

Opal provides a state-of-the-art software infrastructure for data encryption, participant identifiers management and user authentication/authorization. Main features are: Public Key Infrastructure (PKI) allowing Opal to manage public-private key pairs for encrypting and decrypting data, Authentication using either certificates or username/password mechanisms, Advanced participant identifiers manager enabling multiple identifiers per participant, Distinct and highly secure database for storing participant identifiers, Granular permission management down to the variable level, REST web services using HTTPS protocol.



OPAL TECHNICAL NOTES

Operations of studies involve file management and exchanges. Opal comes with its own file system to facilitate these processes. Main features are a centralized and file management and SFTP access.

A statistical analysis module using R is available and organized in a R server monitoring from Opal, Secured data access from R, Opal R package and DataSHIELD R packages. Imported data is indexed with the ElasticSearch search engine allowing fast retrieval and complex querying of the data. All data is REST-enabled allowing to access every data through an URL. Any client that can make an HTTPs request can be a client to an Opal server. Therefore, the resources can be obtained in JSON or binary form (Protobuf), client authentication can be done by providing a username and password or by establishing a Two-way SSL authentication. Clients are already available in Javascript, R, Python and Php.



Data Harmonisation Across Databases

Opal

Opal is a software application to manage study data, and includes a feature enabling data harmonisation and data integration across studies. As such, Opal supports the development and implementation of processing algorithms required to transform study-specific data into a common harmonised format. Moreover, when connected to a Mica web interface, Opal allows users to seamlessly and securely search distributed datasets across several Opal instances.

DESIGNED FOR

Database owners - Individual research studies/ biobanks and research study consortia: to manage and present data, to harmonise data, to give access to data in a federated database setting.

DEVELOPED BY

Opal development was initiated by OICR and is part of the Maelstrom Research suite of tools. Opal development is supported by BioSHaRE, Québec's Ministère de l'Économie, Innovation et Exportation, the Canadian Partnership Against Cancer, and the National Institutes of Health funded Integrative Analysis of Longitudinal Studies of Aging (IALSA) project.

APPLICATIONS

Opal software has been used in the BioSHaRE Healthy Obese and Environmental determinants of health projects to store the data used for combined analyses, develop and implement processing algorithms transforming study data into format, and create a federated infrastructure that allows researchers to jointly analyse harmonised data.

READ MORE

- Key publication: Doiron et al. 2013
- Demo website: <http://www.obiba.org/pages/products/opal/>.
- **BioSHaRE deliverable 2.2 at www.bioshare.eu**

USE

- Data storage and management
- Data harmonisation and curation through data processing algorithms
- Data search and query in study data and data dictionaries
- Data analysis: generate descriptive statistics and produce reports

STATUS AND ACCESS

Opal is freely available for download at www.obiba.org and is provided under the GPL3 open source licence. All studies or networks of studies using the Opal software for data storage, data management or data harmonisation must mention Opal in manuscripts, presentations, or other works made public and include a web link to the Maelstrom Research website (www.maelstrom-research.org).

When using Opal to implement data processing algorithms to harmonise or clean data, basic knowledge of the JavaScript programming language is required.

Opal is a Java-based application, so it should run on any platform for which a Java Virtual Machine is provided. Detailed installation and configuration instructions are available at www.obiba.org.

CONTACT

Dr. Vincent Ferretti
Ontario Institute for Cancer Research, Canada
vincent.ferretti@oicr.on.ca



DATASHIELD

Taking the analysis to the data, not the data to the analysis.

HOW DOES DATASHIELD WORK?

Analysis requests are sent from a central analysis machine to several data-holding machines storing the harmonized data to be co-analyzed. The data sets are analyzed simultaneously but in parallel, linked by non-disclosive summary statistics. Analysis is taken to the data – not the data to the analysis.

DATASHIELD INFRASTRUCTURE

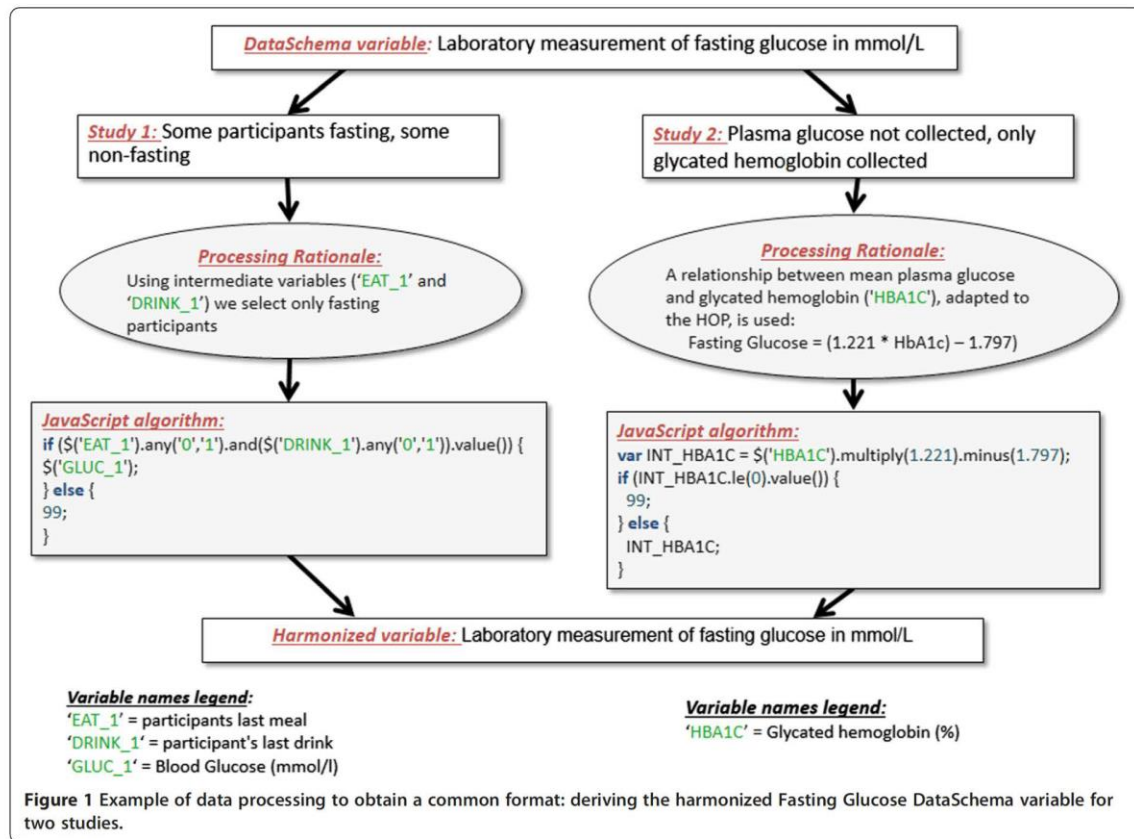
DataSHIELD is implemented entirely via free, open source software: at heart, a modified R statistical environment linked to an Opal database deployed behind the firewall at each data-holding organization. Analysis is initiated in a standard R environment at the analysis machine, with communication between the analysis and data-holding machines controlled via secure web services. The same infrastructure and approach may also be used with just one data source – this is then referred to as “single site DataSHIELD” providing a freeware-based approach to creating a secure data enclave.

Table 1 The Healthy Obese Project data harmonization and database federation step-by-step process

Step	Description
Study recruitment and documentation	Studies are recruited to participate in the HOP and their key characteristics (e.g. design, sampling frame) are catalogued on the BioSHaRE website (www.bioshare.eu).
Harmonized variable selection and definition	A set of “target” variables required to answer obesity-related research questions is identified at workshops bringing together BioSHaRE investigators.
Study variable identification and harmonization potential assessment	By analysing participating studies’ questionnaires, standard operating procedures, and data dictionaries, the potential for each study to generate this set of target variables is determined. Study-specific variables required to generate target variables are identified.
Data processing	Secure servers are set-up in each study’s host institution and the subsets of data required to generate target variables are loaded onto each of these servers. Processing algorithms transforming study data into the target (i.e. harmonized) format are developed and implemented for each study whenever harmonization is deemed possible.
Harmonized data federation, dissemination and analysis	A password protected web portal federates the servers found in the different study host institutions across Europe and allows remote retrieval of data summaries, descriptive statistics (frequencies, min, max, mean, standard deviation), and contingency tables. For more complex federated data analyses (e.g. linear regressions), the DataSHIELD method [28] is employed in the R software environment [36].

Emerging Themes in Epidemiology 2013, 10:12: Data harmonization and federated analysis of population-based studies: the BioSHaRE project

Dany Doiron^{1,2*}, Paul Burton⁴, Yannick Marcon¹, Amadou Gaye⁴, Bruce H R Wolffenbuttel⁵, Markus Perola^{6,7}, Ronald P Stolk⁸, Luisa Foco⁹, Cosetta Minelli¹³, Melanie Waldenberger¹⁰, Rolf Holle¹⁰, Kirsti Kvaløy¹¹, Hans L Hillege¹², Anne-Marie Tassé², Vincent Ferretti^{3†} and Isabel Fortier^{1†}



Emerging Themes in Epidemiology 2013, 10:12: Data harmonization and federated analysis of population-based studies: the BioSHaRE project

Dany Doiron^{1,2*}, Paul Burton⁴, Yannick Marcon¹, Amadou Gaye⁴, Bruce H R Wolffenbuttel⁵, Markus Perola^{6,7}, Ronald P Stolk⁸, Luisa Foco⁹, Cosetta Minelli¹³, Melanie Waldenberger¹⁰, Rolf Holle¹⁰, Kirsti Kvaløy¹¹, Hans L Hillege¹², Anne-Marie Tassé², Vincent Ferretti^{3†} and Isabel Fortier^{1†}



Data Analysis Across Databases

DataSHIELD

Data Aggregation Through Anonymous Summary-statistics from Harmonised Individual-level Databases

DataSHIELD was born of the requirement in the biomedical and social sciences to co-analyse individual patient data (micro data) from different sources, without disclosing identity or sensitive information. Under DataSHIELD, raw data never leave the data provider and no micro data or disclosive information can be seen by the researcher. The analysis is taken to the data – not the data to the analysis. It provides a flexible, modular, open-source solution ideally placed to serve a broad user and development community and to circumvent barriers related to ethical-legal restrictions, intellectual property and physical size of the data as a limiting factor.

DESIGNED FOR

- Database owners - biobanks, other studies: to allow analyses of individual level data while respecting ethical, legal and IP issues,
- Researchers - consortia: to share and analyse data in a consortium or between multiple studies without actual data pooling.

DEVELOPED BY

The following partners are involved in the ongoing development of DataSHIELD:
BioSHaRE partners UB (Data to Knowledge Research Group), OICR (including Obiba), McGill (including Maelstrom Research), NIPH, UMCG, ULEIC and external partner Eindhoven University of Technology, Netherlands.

APPLICATIONS

DataSHIELD is used for secure data analyses in BioSHaRE within the Healthy Obese Project and Environmental determinants of health projects. DataSHIELD will be used in InterConnect and other recently initiated projects.

READ MORE

- Publications and information at <http://www.datashield.ac.uk>
- Key publications: Gaye et al, 2014; Jones et al, 2012; Jones et al, 2013
- BioSHaRE Deliverable 3.2 at www.bioshare.eu

USE

Applied to a single site

- Create a “secure data enclave” in which data can be analysed but not seen, to collaborate in consortium-based analyses without revealing source data.
- Provide a “secure data enclave” to hold potentially sensitive data, created using record linkage, thereby making them accessible for secondary analysis.
- Provide a post-publication platform that enables the data underpinning all of the analyses in a paper to be made publicly available for extended analysis (including confirmation) without data being released into the public domain.
- Provide a publicly accessible web-portal that enables researchers to undertake simple preliminary univariate and bivariate analysis of data before application for full access to those data.

Applied to multiple sites

- Co-analysis of individual-level data or study level meta-analysis from multiple studies

STATUS AND ACCESS

All DataSHIELD packages are open source and in beta-testing. New packages, methodology and functions are also under development and will be tested and released into packages in due course.

Full information and access to DataSHIELD is available at <http://www.datashield.ac.uk> including access to the DataSHIELD wiki (<http://www.datashield.ac.uk/wiki>) that contains all technical documentation and tutorials to install and use DataSHIELD.

DataSHIELD Client Software:

- Runs in linux, Mac and Windows
- Requires R and/or R Studio and the DataSHIELD client packages
- Requires basic knowledge of epidemiological analyses / medical statistics methodology
- Requires experience analysing data in R

CONTACT

Professor Paul Burton
University of Bristol, UK
Paul.Burton@bristol.ac.uk



PHENOTYPE DATABASE

The (Nutritional) Phenotype database (www.dbnp.org) was an initiative of NuGO (NutriGenomics Organization) and NMC (Netherlands Metabolomics Centre), and was launched in 2007 and is a fully open source system under the apache license.

The 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 and is built in the framework Grails and data are stored in a PostgreSQL database.

Data can be uploaded via a wizard in a web browser or can be uploaded from txt or excel files. 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, via dropdowns or as ontology term from Bioportal.

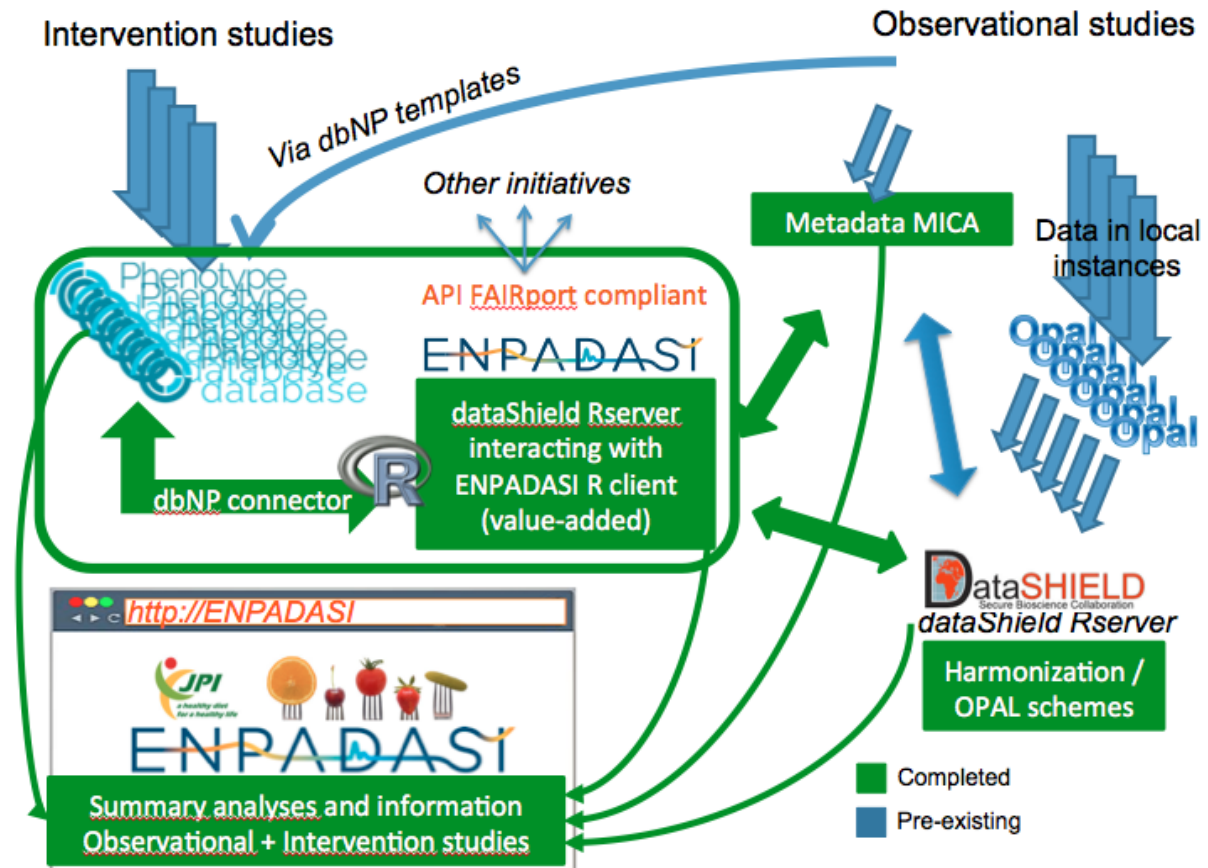
The system is secured with an authentication/authorization system. Only if you have a user account, you can include data. The person including data in the system (data owner) can give access to the data to others and can open the data to the world (e.g. upon publication). The Phenotype database is built in a modular way (e.g. a study design, metabolomics module; all modules are connected via REST APIs), making it possible to link the data and design in the system to data in other databases (e.g. ArrayExpress or other instances of the Phenotype database) and has been tested to work on a cloud solution.

The Phenotype Database facilitates sharing of data within a research group or consortium, as the study owner can decide who can view or access the data. In addition, the Phenotype Database can stimulate collaborations by making study information and data publicly visible. New studies can be based on study data within the database, as standardized storage is stimulated by the system. Upon publication of the data, studies can be made publicly accessible (under these license terms).

Analysis can be performed via a API link to R, but also clients for PHP, python and Grails are available.

TNO was in the lead for the functional requirements and responsible for steering the development work. The system has now been fully operational for 2 years (at four different locations). One of the instances is publicly available (studies.dbnp.org), on which daily and weekly back-ups are made. The database currently includes 99 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).

INTERACTIONS BETWEEN THE SYSTEMS



DATASHIELD RSERVER

Used for the interaction between an R statistical environment and Opal, the R module allows pushing data from Opal into an R environment and back. It can also execute arbitrary R code within these environments.

Opal interacts with an R server through Rserve's protocol. This allows the R Server to be on a different machine than the Opal server. It also allows maintaining R separately from Opal.

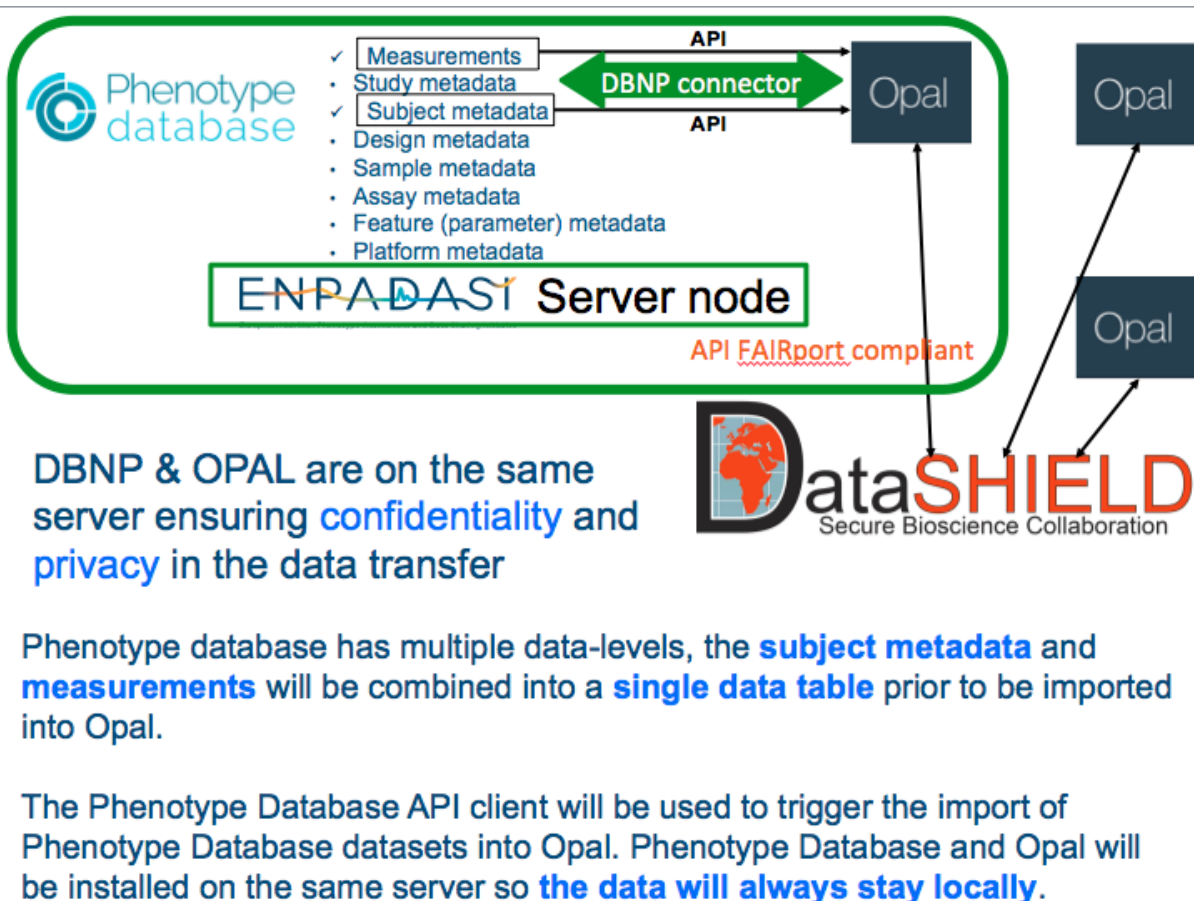
Built "on top" of the R module, the DataSHIELD module provides a constrained and customizable access to the R environment. Specifically, this module allows pushing data from Opal into R, but does not allow reading this data unless it has first been "aggregated".

The term "aggregated" here means that the data in R must go through a method that will summarize individual-level data into another form that removes the original individual-level data. For example, obtaining the length of a vector, obtaining the summary statistics of a vector (min, max, mean, etc.)

It is these methods that are customizable. That is, administrators of the Opal server can add, remove, modify and create completely custom "aggregating" methods that are provided to DataSHIELD clients.

DBNP CONNECTOR

The value-added leveraged by the WP3-designed ENPADASI Research Infrastructure comes from the federated integration of the datasets stored either on OPAL or dbNP. A dbNP connector, running solely on the same machine that hosts both data servers, securely enable dbNP-stored data to enter a DATAShield federated analysis. Selected data fields from dbNP can become part of a wider geographical analysis comprising data safely stored on several local OPAL and dbNP servers. Intrinsic security of dbNP and OPAL datasets integration is achieved by constraining such intercourse only between databases that are running on the same machine therefore ensuring that no single bit ever leaves the same physical hardware.



```
> ## client-side installation
install.packages('opaladmin', repos='http://cran.obiba.org', dependencies=TRUE)
install.packages('dsBaseClient', repos=c(getOption('repos'), 'http://cran.obiba.org'), dependencies=TRUE)
install.packages('dsModellingClient', repos=c(getOption('repos'), 'http://cran.obiba.org'), dependencies=TRUE)
install.packages('dsStatsClient', repos=c(getOption('repos'), 'http://cran.obiba.org'), dependencies=TRUE)
install.packages('dsGraphicsClient', repos=c(getOption('repos'), 'http://cran.obiba.org'), dependencies=TRUE) |
```

In case a running R environment is already installed and running, within an active R session the installation of the DataShield client (the piece of software that enables to access the data and run the analysis) is performed in a console environment issuing the commands listed above.

```
> # load libraries
library(opal)
library(dsBaseClient)
library(dsStatsClient)
library(dsGraphicsClient)
library(dsModellingClient)

# load the login file
my_login<-read.table('../logins.txt', sep=" ", header=TRUE)

# log in to the remote servers
opals <- datashield.login(logins=my_login, assign=TRUE, symbol = 'D')

# analyse the outcome variable DIS_DIAB (diabetes status)
# and the covariates PM_BMI_CONTINUOUS (continuous BMI), LAB_HDL (HDL cholesterol)
# and GENDER (gender), with an interaction between the latter two.
ds.glm(formula=D$DIS_DIAB ~ D$PM_BMI_CONTINUOUS + D$LAB_HDL * D$GENDER, family='binomial') |
```

To run a federated analysis it is required to enter a set of commands. The bare minimum set of commands are presented in the first screen while the second screen shows the results. The type of analysis involves a Generalized Linear Model regression using the binomial link function.

Using a standard R installation it is thus possible to install and use appropriate R packages to run textual-based analyses as in the following figure. This analysis requires technical and analytical expertise, which unfortunately it might not be the most accessible way to enable a wide audience of researchers to explore research questions in nutrition.

THE DASH-IN INTERACTIVE FEDERATED ANALYSIS SYSTEM

The Dash-in Web based federated analyses is an innovative research tool allowing to design, execute and assess analytical results from federated analyses of datasets stored on OPAL servers or both OPAL and Phenotype Databases. The easy to use DASH-IN interactive analytical system can be accessed from <https://dashin.cosbi.eu> providing an authentication credential (Figure 1) that can be requested simply sending an email to lombardo@cosbi.eu .

Project-wide shared credentials – already used for the survey – can be found at <https://dashin.cosbi.eu/data/survey/supportinginformation.html>.

Figure 1: Secured login access to the interactive federated analysis system.

After logon, the list of own data servers is presented. Every user has her/his own list of data servers and credentials which are encrypted and accessible only by the logged in user. The list of servers include all OPAL servers for which the user has her/his own credential but also for those DASH-IN-enabled sites where the ENPADASI-connector has been installed and therefore is linking data from an associated Phenotype Database (Figure 2). The user credentials for those systems can be safely stored in the profile enabling to run federated analyses of any data accessible with those credentials.

OPAL and Phenotype Database data servers available for my analyses

NEW OPAL SERVER

Dash-in server Test

<p>Server Name: Test</p> <p>Server URL: https://test</p> <p>User Name: survey</p> <p>Password: 👁</p> <p>Certificate: No file available</p> <p>Private Key: No file available</p>	<p>Server URL: https://dash</p> <p>User Name: test</p> <p>Password: 👁</p> <p>SKey: 👁</p> <p style="text-align: center; background-color: #0070c0; color: white; padding: 5px; margin-top: 10px;">REMOVE PHENOTYPE DATABASE</p>
--	---

REMOVE DASH-IN

Figure 2: The data server page lists all servers for which the user has an access credential. Therefore, multiple OPAL credentials can be safely stored along with associated Phenotype Database ones.

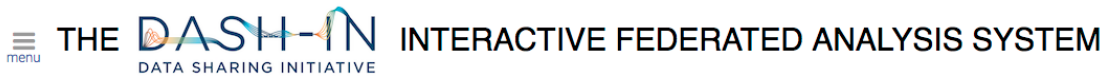
SEARCHING STUDIES RELEVANT FOR A RESEARCH QUESTION

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 4) 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) and perform analytical tasks such as those described above. Additional user-friendly statistical analyses such as regressions, generalized linear models and exploratory data-undisclosing plots can be setup and applied from the web-based Dash-In portal (see more details below).



Search something

finrisk x |

Results

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 Vartiainen Satu Männistö

Figure 3: 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 FoodBALL project.

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.

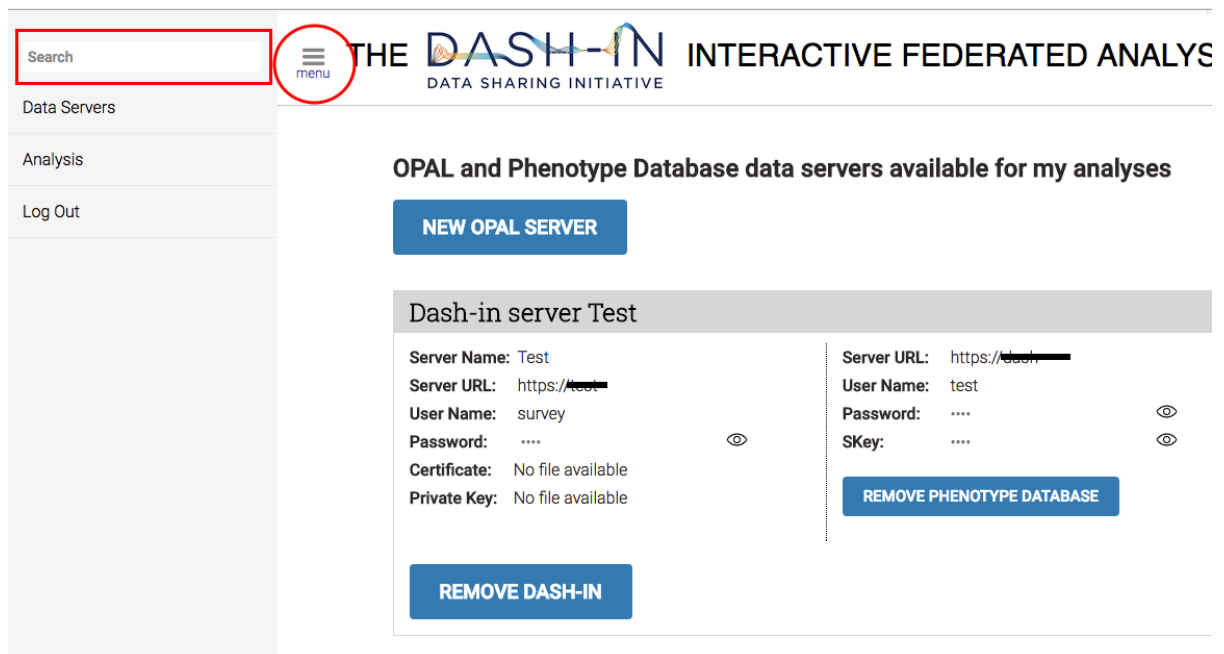
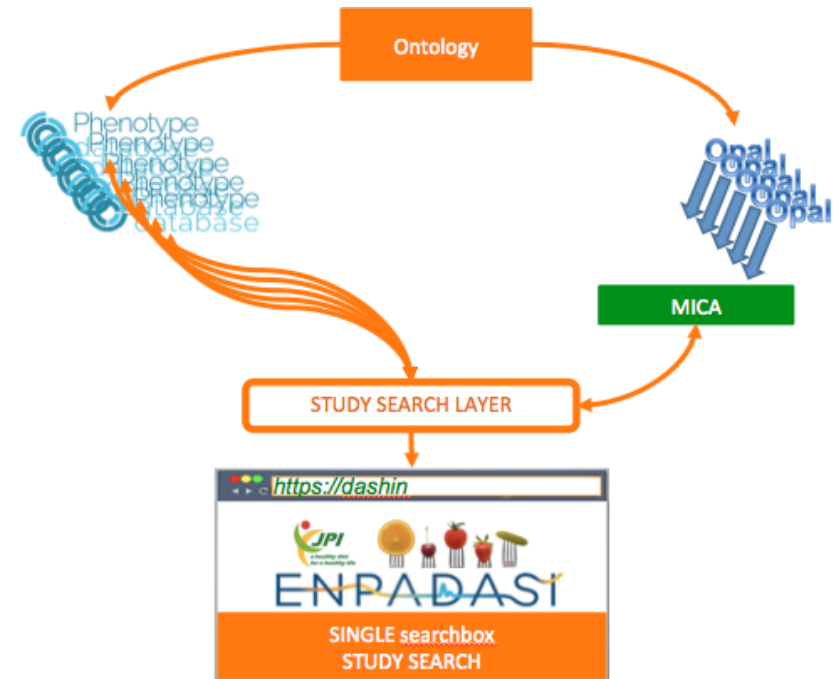
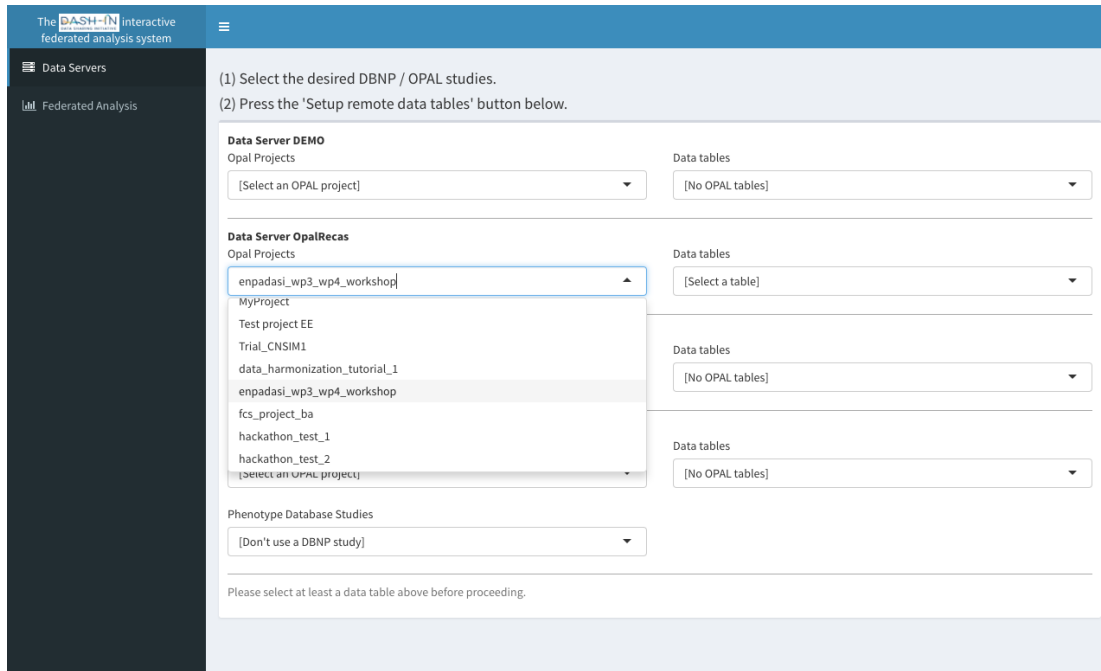


Figure 4: Clicking on the top-left menu icon a sidebar menu slides in. The first search box gives access to the metadata search feature that allows to enter predefined metadata search terms to finely identify available studies annotated with the same search terms.

SETTING UP DATA SOURCES

Once appropriate studies have been identified, possibly after having obtained appropriate credentials from the study/dataserver owner, the researcher can move on setting up the actual datasets needed to be included in the federated analysis. For each of the configured data server the user can easily see listed in dropdowns all studies (s)he has access to (Figure 5, Figure 6, Figure 7, Figure 8).



The screenshot shows the 'Data Servers' configuration interface. At the top, there are instructions: (1) Select the desired DBNP / OPAL studies. (2) Press the 'Setup remote data tables' button below.

The interface is divided into three main sections:

- Data Server DEMO:** Includes a dropdown for 'Opal Projects' (currently showing '[Select an OPAL project]') and a dropdown for 'Data tables' (currently showing '[No OPAL tables]').
- Data Server OpalRecas:** Includes a dropdown for 'Opal Projects' with a list of projects: 'enpadasi_wp3_wp4_workshop', 'MyProject', 'Test project EE', 'Trial_CNSIM1', 'data_harmonization_tutorial_1', 'enpadasi_wp3_wp4_workshop', 'fcs_project_ba', 'hackathon_test_1', 'hackathon_test_2', and '[Select an OPAL project]'. To the right, there are two 'Data tables' dropdowns, one showing '[Select a table]' and the other showing '[No OPAL tables]'. The 'enpadasi_wp3_wp4_workshop' project is highlighted in the dropdown menu.
- Phenotype Database Studies:** Includes a dropdown for 'Phenotype Database Studies' (currently showing '[Don't use a DBNP study]').

At the bottom, a note states: 'Please select at least a data table above before proceeding.'

Figure 5: Selection of the Projects into OPAL servers.

Figure 6: Selection of tables from OPAL projects.

Figure 7: Selection of Phenotype database data hosted within DASH-IN-enabled Phenotype Database computers.

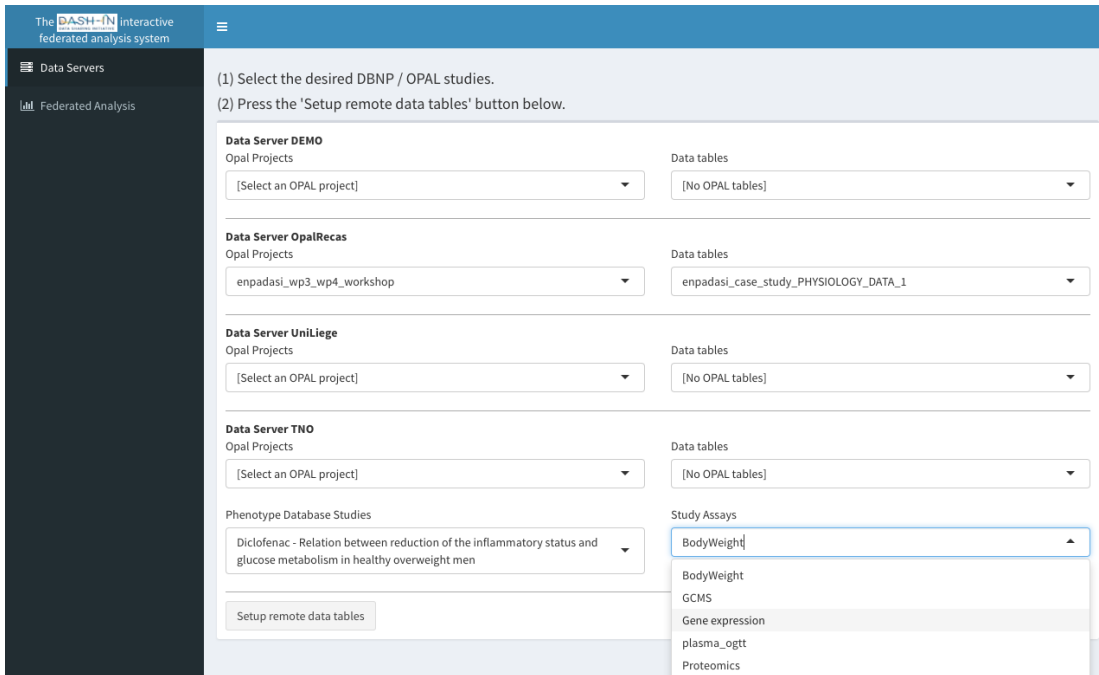


Figure 8: Selection of the study Assays associated to Phenotype Database DASH-IN enabled machines.

After having selected all OPAL and Phenotype Database pressing the button “Setup remote data tables” will prepare the DASH-IN technical infrastructure to make federated analysis on different tables hosted on different server but without moving or disclosing any single information out of each hosting server. When the setup process is ready a green message will inform the user (Figure 9).

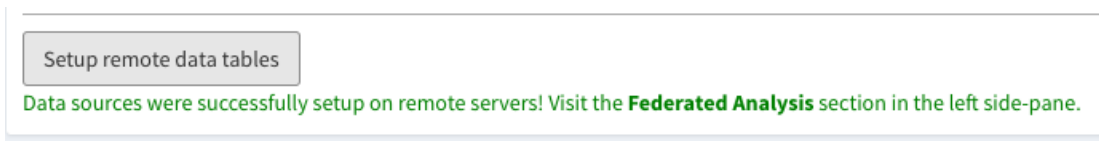


Figure 9: Federate data setup is needed to retrieve the variable names essential to make sense of the analysis the user wants to perform.

FEDERATED EXPLORATIVE PLOTS

Prior to diving into data analysis, it recognized to be a good practice making some exploratory data plots to figure some overall properties of the data under study. To this aim a number of explorative analyses from federated data are easily available to the final researcher including **Histograms** (Figure 10), **Contour Plots** (Figure 11) and **Heatmaps** (Figure 12). All data analysis are non-disclosive 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/data points.

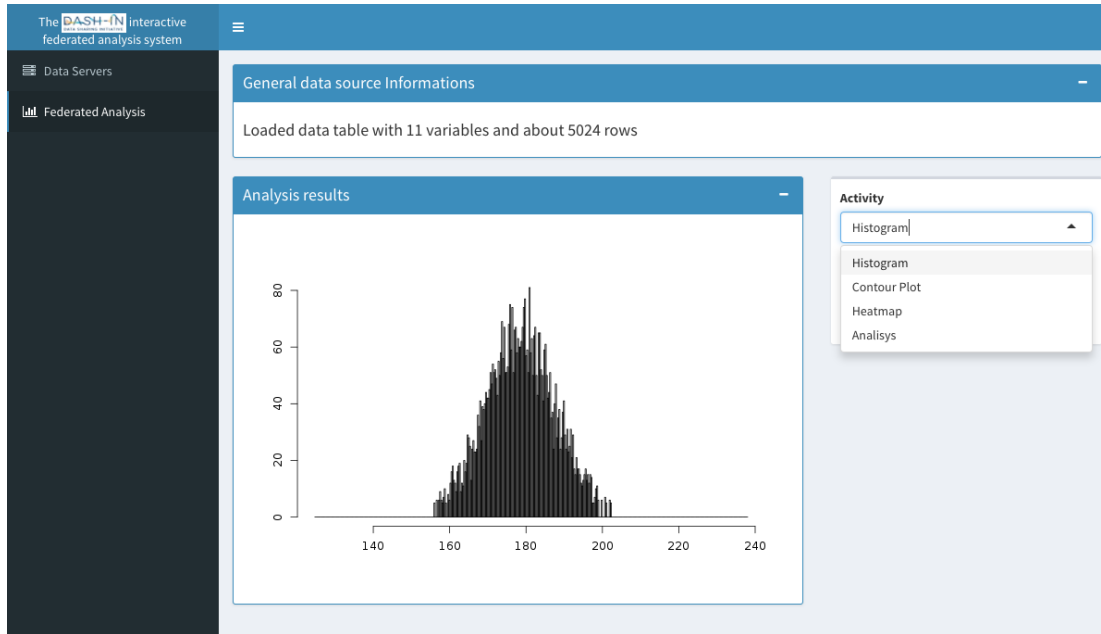


Figure 10: Federated histograms.



≡ The **DASH-IN** interactive federated analysis system
DATA SHARING INITIATIVE

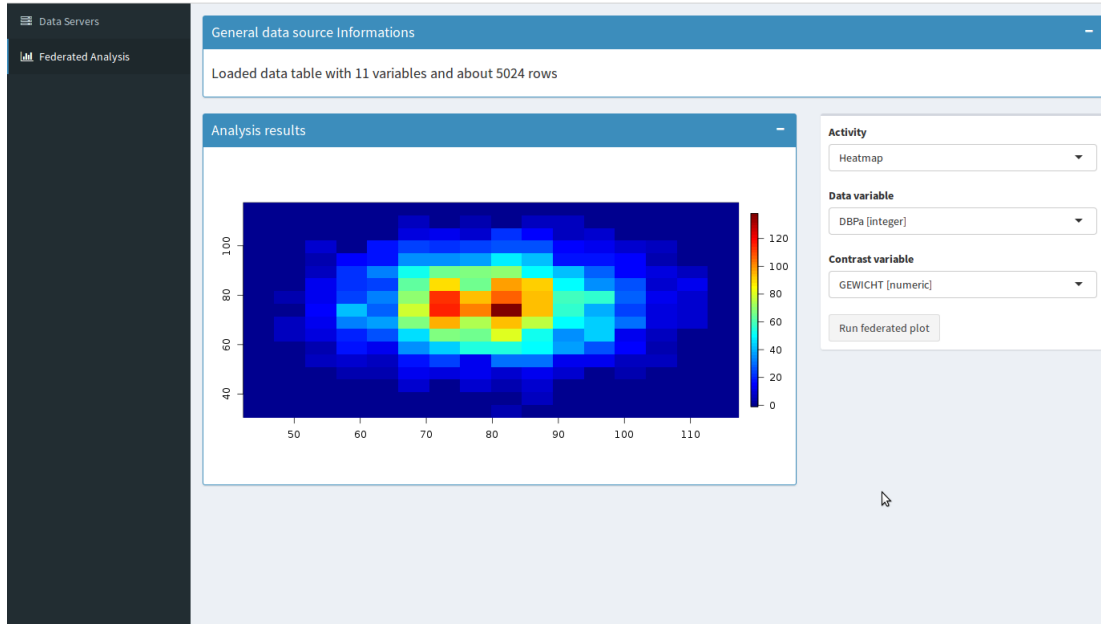


Figure 11: Federated heatmaps.

≡ The **DASH-IN** interactive federated analysis system
DATA SHARING INITIATIVE

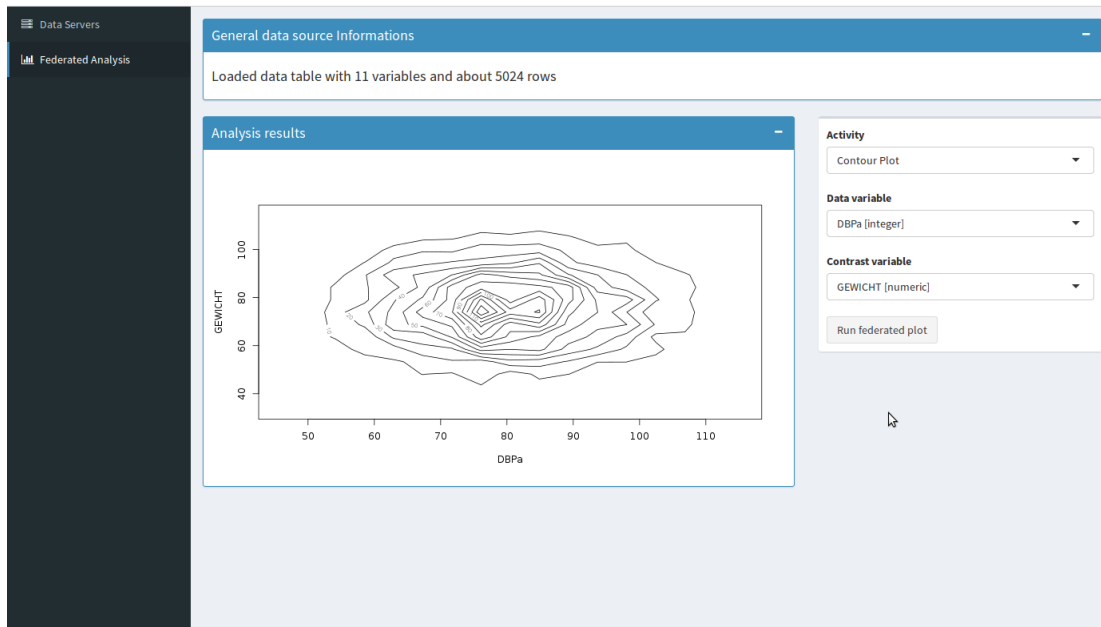


Figure 12: Federated contour plots.

Analysis for phenotype-trait associations is a fundamental cornerstone in research and the DASH-IN offers an easy way to access the great power of pooled-federated analyses through a simple web page. The Analysis toolbox offers **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.

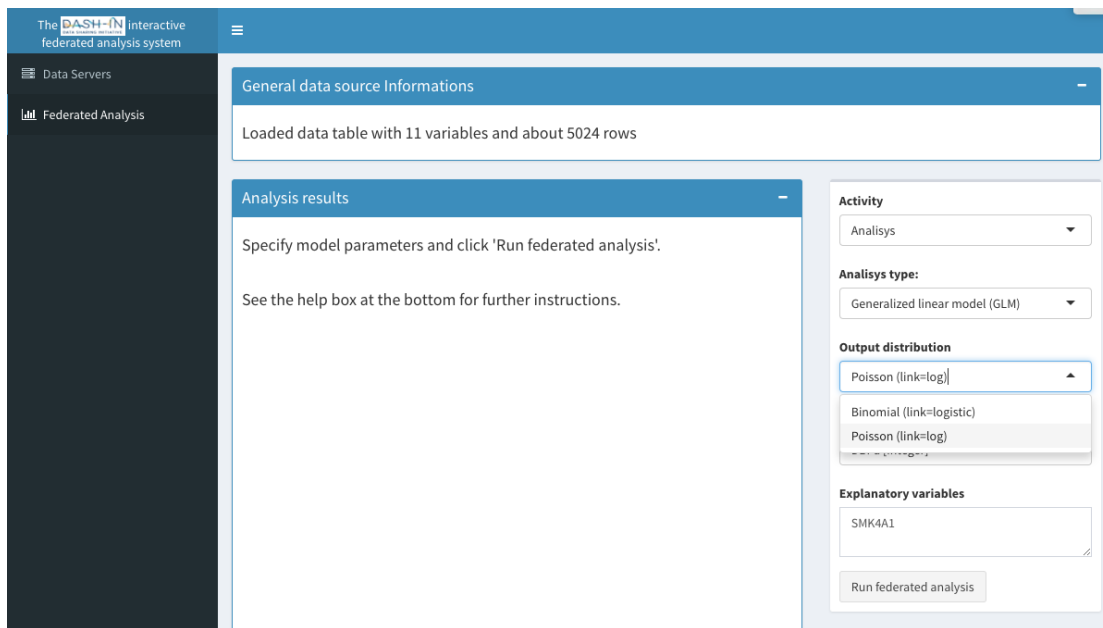


Figure 13: Setting up a GLM using the Poisson distribution it is a matter of few mouse clicks. Total control on the meaning of the performed analysis is left to the researcher who needs to know and understand what the computed association means. Specific type conversion (e.g. categorical to numerical) are allowed to offer full savvy control on the research needs. Full documentation is provided at the bottom of the screen (see for example [Figure 14](#) and [Figure 15](#)).



The **DASH-IN** DATA SHARING INITIATIVE interactive federated analysis system

General data source Informations
Loaded data table with 11 variables and about 5024 rows

Analysis results
Evaluated model 'GESLACHT ~ LENGTE + SMK11#num + DBPa * GEWICHT' family 'gaussian' link 'identity'

Variables	Estimate	Std. Error	z-value	p-value	low0.95CI	high0.95CI
(Intercept)	1.699	0.284	5.991	0	1.143	2.255
LENGTE	0	0.001	-0.1	0.92	-0.002	0.001
SMK11_conv	-0.04	0.018	-2.249	0.025	-0.074	-0.005
DBPa	-0.002	0.003	-0.687	0.492	-0.008	0.004
GEWICHT	-0.002	0.003	-0.664	0.507	-0.009	0.004
DBPa*GEWICHT	0	0	0.874	0.382	0	0

Null deviance: 1255.7 on 5033 degrees of freedom
Residual deviance: 1253.56 on 5038 degrees of freedom
Multiple R squared: 0.0017 Adjusted R squared: 5e-04
F-statistic: 1.7126 on 5 and 5018 DF; p-value: 0.128

Activity
Analysis

Analysis type:
Linear Regression

Dependent variable
GESLACHT [factor]

Explanatory variables
LENGTE + SMK11#num + DBPa * GEWICHT

Null explanatory variables
1

Null model used to compute R² and F-test
Run federated analysis

Help
Dependent variable:
The variable we would like to predict using the explanatory variables. If the selected dependent variable is categorical and the output distribution is continuous (as in linear regression or GLM+Poisson case) the variable will be automatically converted to numeric.
Explanatory variables:
These are variable names we want to test for correlation with the dependent one. More than one explanatory variable can be generated by the "+" operator. The interaction between two variables can be tested as well generating

Figure 14: Federated linear regression results and full documentation on how to setup and interpret analytical results.

The **DASH-IN** DATA SHARING INITIATIVE interactive federated analysis system

General data source Informations
Loaded data table with 11 variables and about 5024 rows

Analysis results
Evaluated model 'GESLACHT ~ LENGTE + SMK11#num + DBPa * GEWICHT' family 'binomial' link 'logistic'

Variables	Estimate	Std. Error	z-value	p-value	low0.95CLP	high0.95CLP	P_OR	low0.95CLP_OR	high0.95CLP_OR
(Intercept)	0.799	1.137	0.703	0.482	-1.429	3.028	0.69	0.133	0.954
LENGTE	0	0.003	-0.1	0.92	-0.006	0.006	1	0.994	1.006
SMK11_conv	-0.159	0.071	-2.248	0.025	-0.298	-0.02	0.853	0.742	0.98
DBPa	-0.009	0.013	-0.687	0.492	-0.033	0.016	0.991	0.967	1.016
GEWICHT	-0.009	0.013	-0.664	0.507	-0.034	0.017	0.991	0.966	1.017
DBPa*GEWICHT	0	0	0.874	0.382	0	0	1	1	1

Residual deviance: 6954.97 on 5018 degrees of freedom

Activity
Analysis

Analysis type:
Generalized linear model (GLM)

Output distribution
Binomial (link=logistic)

Dependent variable
GESLACHT [factor]

Explanatory variables
LENGTE + SMK11#num + DBPa * GEWICHT

Run federated analysis

Help
Dependent variable:
The variable we would like to predict using the explanatory variables. If the selected dependent variable is categorical and the output distribution is continuous (as in linear regression or GLM+Poisson case) the variable will be automatically converted to numeric.
Explanatory variables:
These are variable names we want to test for correlation with the dependent one. More than one explanatory variable can be generated by the "+" operator. The interaction between two variables can be tested as well generating

Figure 15: Federated GLM modeling the output as a binomial distribution (logit link function).



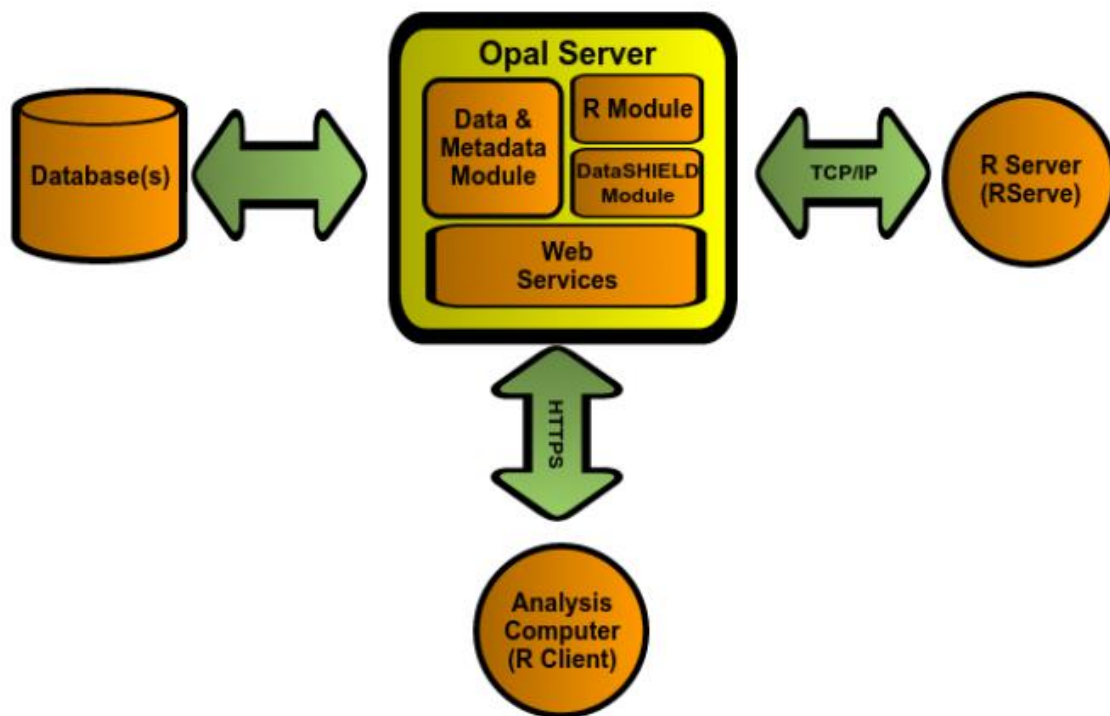
TECHNICAL SPECIFICATIONS

OPAL

INSTALLATION PREREQUISITES

- Ubuntu 14.04
- 3 Virtual machines (XenServer) in private network
 - 2 opal instances
 - 1 mica instance
- HTTP only, no HTTPS

OPAL SCHEMATIC OVERVIEW



<http://wiki.obiba.org/display/OPALDOC/Opal+R+and+DataSHIELD+User+Guide>

OPAL DATABASE INSTALLATION

- <http://wiki.obiba.org/display/OPALDOC/Databases+Administration>
- MongoDB or MySQL/MariaDB, PostgreSQL



Database Engine	Data Schema	Storage	Import	Export
MySQL	Opal SQL	✓	✗	✗
MySQL, PostgreSQL	Tabular SQL	✓	✓	✓
MySQL, PostgreSQL	Limesurvey	✗	✓	✗
MongoDB	Opal Documents	✓	✗	✗

SOME MONGODB INSTALLATION REMARKS

- MongoDB on Ubuntu:
<https://docs.mongodb.org/master/tutorial/installmongodb-on-ubuntu/>
- Create a MongoDB “root” user according to
<http://stackoverflow.com/questions/20117104/mongodb-root-user>
- Follow the Opal wiki documentation

SECURITY AUTHORIZATION IN THE /ETC/MONGOD.CONF FILE

Some MySQL installation remarks

- Database creation
 - CREATE DATABASE opal_ids CHARACTER SET utf8 COLLATE utf8_bin;
 - CREATE DATABASE opal_data CHARACTER SET utf8 COLLATE utf8_bin;
 - CREATE USER 'opal' IDENTIFIED BY '<opal-user-password>';
 - GRANT ALL ON opal_ids.* TO 'opal'@'localhost' IDENTIFIED BY '<opal-user-password>';
 - GRANT ALL ON opal_data.* TO 'opal'@'localhost' IDENTIFIED BY '<opal-user-password>';
 - FLUSH PRIVILEGES;
- Data schemas:
 - Opal SQL: flexibility over performance
 - Tabular SQL: SQL limitations, import/export

OPAL WEB APPLICATION DEMO

- <http://opal-demo.obiba.org/ui/index.html#!dashboard>



OPAL R SERVER SIDE NOTE

Can be accessed without dataSHIELD!

```
# Load Opal R library
require('opal')

# Then, create an opal object with the login information
# Change the login credentials and url with the appropriate values!
o <- opal.login(username='administrator', password='password', url='https://localhost:8443')

# To verify if the connexion with Opal works,
# get the list of all projects
opal.datasources(o)

# Assign the content of the LifeLines table (from the test project)
# into a data frame in a R session of the R server
opal.assign(o,'D','test.LifeLines', missings = TRUE)

# Get the summary of this data frame from the remote R session
opal.execute(o,'summary(D)')

# Terminate the remote R session
opal.logout(o)
```

DATASHIELD

DATASHIELD INSTALLATION AND USAGE

- <http://wiki.obiba.org/display/OPALDOC/How+to+install+and+use+Opal+and+DataSHIELD+for+Data+Harmonization+and+Federated+Analysis>
- On Opal server (web application):
 - install dataSHIELD packages in admin section
 - create a user and set the correct permissions on both dataSHIELD and project/table level
- On client side:
 - install necessary R packages
 - connect for example with Rstudio (demo?)

MICA

MICA SCHEMATIC OVERVIEW



Servers



Clients



Python



Drupal



Browser

<http://wiki.obiba.org/display/MICADOC/Overview>

MICA MODULES

- Mica server: RESTful Java server for managing, storing and searching Mica content
- Mica Web Application: user interface for administration and configuration of a Mica server
- Mica Drupal Client: Mica CMS providing tools to create custom web portals
- Mica Python Client: command-line interface for administration and automation of a Mica server

SOME MICA-SERVER INSTALLATION REMARKS

- Install MongoDB first!
- <http://wiki.obiba.org/display/MICADOC/Mica+Server+Installation+Guide>
- Connection with Opal:
 - On Opal server: create Mica user and set the necessary project/table permissions
 - On Mica server: add (multiple) Opal credentials

MICA DEMO'S

- Mica Web Application demo
 - <http://mica-demo.obiba.org/>
- Mica Drupal Client demo
 - <http://demo.obiba.org/mica2/>



- Mica Drupal Client example
 - <https://www.bioshare.eu/>

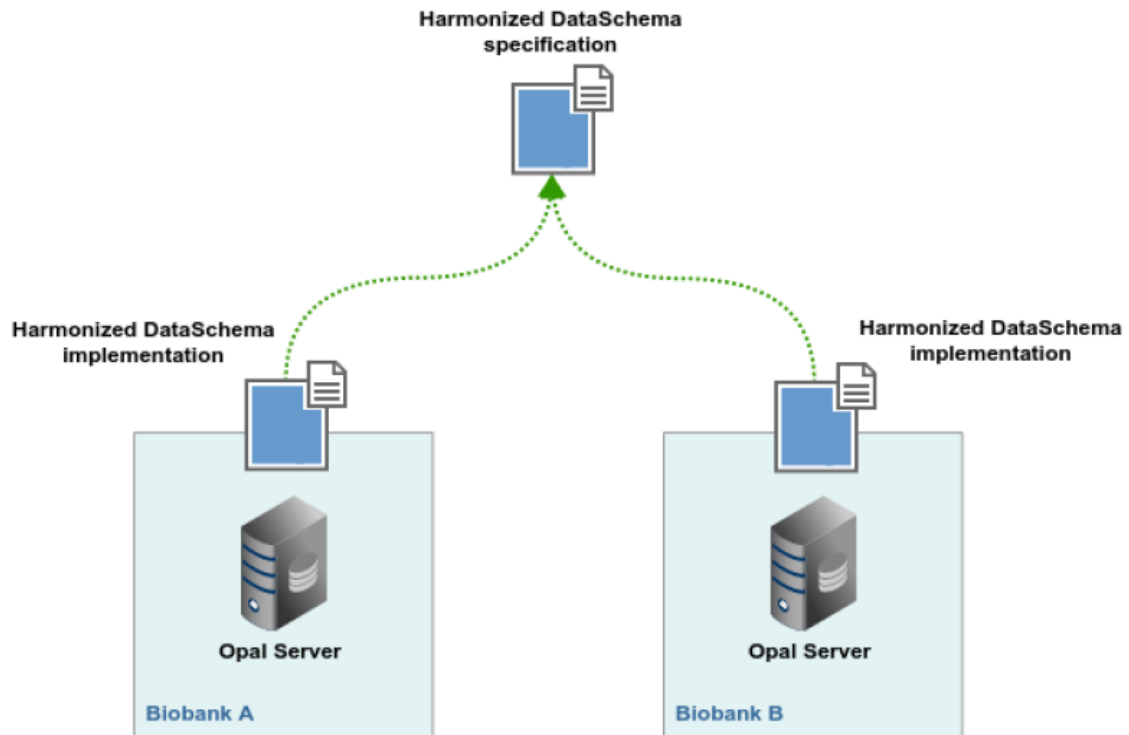
MICA-PYTHON-CLIENT

- Command line scripting tool written in Python, enables automation of tasks in a Mica server
- Provides access to the Mica REST-API

```
compomics@compomics-xubuntu:~$ mica rest -mk http://mica-demo.obiba.org -u administrator -p password -j /study/finrisk-2007
{
  "acronym": [
    {
      "lang": "en",
      "value": "FINRISK 2007"
    }
  ],
  "contacts": [
    {
      "email": "helena.kaariainen@ktl.fi",
      "firstName": "Helena",
      "institution": {},
      "lastName": "Kaariainen",
      "phone": "+358 (0) 20 610 6000"
    },
    {
      "email": "markus.perola@ktl.fi",
      "firstName": "Markus",
      "institution": {
        "address": {
          "city": [
            {
              "lang": "en",
              "value": "Helsinki"
            }
          ],
          "country": {
            "iso": "FI",
            "name": [
              {
                "lang": "en",
                "value": ""
              }
            ]
          },
          "zip": "251"
        }
      }
    }
  ]
}
```

HARMONIZATION

HARMONIZATION ACROSS OPAL INSTANCES



<http://wiki.obiba.org/display/OPALDOC/Data+Harmonization+with+Opal#DataHarmonizationwithOpal-DataHarmonizationacrossBiobanks>

HARMONIZED DATASHEMA

- Create on Opal (or Mica Drupal Client) web application
- Export/Import as XML to other Opal instances
- Extensive tutorial:

<http://wiki.obiba.org/display/MLSTRM/Data+Harmonization+and+Database+Federation+Tutorial>

<input type="checkbox"/> Name	Label	Value Type	Categories
<input type="checkbox"/> Gender		integer	0, 1, 9
<input type="checkbox"/> PM_BMI_CAT		integer	1, 2, 3, 9
<input type="checkbox"/> DIS_DIAB		integer	0, 1, 9
<input type="checkbox"/> PM_DIASTOLIC		integer	999
<input type="checkbox"/> SMK_CIG_CURRENT		integer	0, 1, 9

HARMONIZED DATASHEMA VARIABLE

- Has categories (from taxonomy if available)
- Project/table data processing algorithms



- Javascript
- Wizard for simple transformations

Categories

[Edit Categories](#) Total 3

Name	Label	Missing
0	Male <input type="checkbox"/> Male	
1	Female <input type="checkbox"/> Female	
9	Missing	✓

Attributes

[+ Add Attribute](#)

Name	Value	Actions
<input type="checkbox"/> script	<pre> \$('SUKUP').map ({ '1': '0', '2': '1' }, null, '9'); </pre>	Edit Remove

MORE COMPLEX ALGORITHM EXAMPLE

Categories

[Edit Categories](#) Total 4

Name	Label	Missing
1	Less than 25 kg/m2 <input type="checkbox"/> Less than 25 kg/m2	
2	25 to 30kg/m2 <input type="checkbox"/> 25 to 30kg/m2	
3	Over 30 kg/m2 <input type="checkbox"/> Over 30 kg/m2	
9	Missing	✓

```

/*
If Q69 = 1 or 2 and TUP11 is greater than 0,
then SMK_CIG_CURRENT = 1*/

function cig_smoker() {
  || var cig_day = $( 'TUP11' );
  || if ( cig_day.gt(0).value() ) return 1;
  || else if ( cig_day.eq(0).value() ) return 0;
  || else return 9;
}

$( 'Q69' ).map ( {
  || '1': cig_smoker(),
  || '2': cig_smoker(),
  || '3': 0
}, 9, 9 );

```



HARMONIZATION DATASET EXAMPLE

- <http://demo.obiba.org/mica2/mica/datasets/harmonization-datasets>

Statistics

Crosstabs

Value	Atlantic PATH 1	Atlantic PATH 2	BCGP 1	BCGP 2	BCGP 3	CaG	OHS 1	OHS 2	TTP 1	TTP 2	All
Valid Values											
0 None	17 12.90% (12.90%)	15 12.50% (12.50%)	13 12.30% (12.30%)	14 12.70% (12.70%)	13 12.50% (12.50%)	15 11.90% (11.90%)	23 12.40% (12.40%)	14 12.20% (12.20%)	20 12.70% (12.70%)	24 12.70% (12.70%)	168 12.50% (12.50%)
1 Elementary school	16 12.10% (12.10%)	15 12.50% (12.50%)	13 12.30% (12.30%)	14 12.70% (12.70%)	13 12.50% (12.50%)	16 12.70% (12.70%)	23 12.40% (12.40%)	14 12.20% (12.20%)	20 12.70% (12.70%)	23 12.20% (12.20%)	167 12.40% (12.40%)
2 High school	17 12.90% (12.90%)	15 12.50% (12.50%)	13 12.30% (12.30%)	13 11.80% (11.80%)	13 12.50% (12.50%)	16 12.70% (12.70%)	24 12.90% (12.90%)	15 13.00% (13.00%)	19 12.00% (12.00%)	23 12.20% (12.20%)	168 12.50% (12.50%)
3 Trade, technical or vocational school, apprenticeship training or technical CEGEP	16 12.10% (12.10%)	15 12.50% (12.50%)	13 12.30% (12.30%)	14 12.70% (12.70%)	13 12.50% (12.50%)	15 11.90% (11.90%)	23 12.40% (12.40%)	15 13.00% (13.00%)	19 12.00% (12.00%)	23 12.20% (12.20%)	166 12.30% (12.30%)
4 Diploma from a community college, pre-university CEGEP or non-university certificate	16 12.10% (12.10%)	15 12.50% (12.50%)	14 13.20% (13.20%)	14 12.70% (12.70%)	13 12.50% (12.50%)	16 12.70% (12.70%)	24 12.90% (12.90%)	14 12.20% (12.20%)	20 12.70% (12.70%)	24 12.70% (12.70%)	170 12.60% (12.60%)
5 University certificate below bachelor's level	17 12.90% (12.90%)	15 12.50% (12.50%)	13 12.30% (12.30%)	14 12.70% (12.70%)	13 12.50% (12.50%)	16 12.70% (12.70%)	23 12.40% (12.40%)	15 13.00% (13.00%)	20 12.70% (12.70%)	24 12.70% (12.70%)	170 12.60% (12.60%)

PHENOTYPE DATABASE

See for all information the Phenotype database website (and github): www.dbnp.org

SHINYAPPS

INSTALLING THE REQUIRED SOFTWARE COMPONENTS

DEBIAN

RSTUDIO DESKTOP

Is the software that helps us working with R – hence with shiny as well.

Install gdebi (used to install RStudio server and the shiny server)

```
$ sudo apt-get install gdebi-core
```

Download and install RStudio Desktop:

```
$ wget https://download1.rstudio.org/rstudio-0.99.902-amd64.deb
```

```
$ sudo gdebi rstudio-0.99.902-amd64.deb
```




If Debian is not your OS, then you can the other binaries or the sources here:

<https://www.rstudio.com/products/rstudio/download/>

INSTALL R AND SHINY

Add the CRAN repository to get the latest version of R. In this tutorial we use the GARR repository, but you should choose the one that best fits you: <https://cran.rproject.org/mirrors.html>

Add the following statement in the file `/etc/apt/sources.list.d/cran.list`

```
deb http://cran.mirror.garr.it/mirrors/CRAN/bin/linux/debian jessie-cran3/
```

Then add the key for this Debian archive:

```
$ sudo apt-key adv --keyserver keys.gnupg.net --recv-key 381BA480
```

And update the packages list:

```
$ sudo apt-get update
```

Install R from the command line:

```
$ sudo apt-get install r-base
```

Then install the shiny package from either the command line:

```
$ sudo R -e "install.packages('shiny', repos='https://cran.rstudio.com/')
```

Or, from the R prompt:

```
> install.packages('shiny', repos='https://cran.rstudio.com/')
```

MAC

On Macs it is enough to download the latest R version from one of the mirrors at <https://cran.r-project.org/mirrors.html> and the latest RStudio version from <https://www.rstudio.com/products/rstudio/download/>.

LINKING INTO THE DASH-IN INFRASTRUCTURE

To use external data in a shiny application, you must use the `global.R` file, whatever is declared in this file, it is parsr first of any other Shiny file and it also accessible from both `ui.R` and `server.R` files – so let's put the definition of our functions there.



Below we see the complete global.R, ready to interact with the Dash-In infrastructure and namely the DataShield system. Most of these commands have been covered in previous DataShield tutorials of the Hackaton so let's briefly say that the first commands perform a distributed login across all the sites from which we want to fetch data.

As prerequisites the following DataShield R packages should be installed system-wide (Debian):

```
sudo apt-get install r-cran-rjson  
sudo apt-get install libcurl4-gnutls-dev libcurl4-openssl-dev
```

in R console:

```
install.packages('RCurl', repos='http://cloud.r-project.org',  
dependencies=TRUE)
```

Additionally the following packages need to be installed on any OS:

```
install.packages('opaladmin', repos='http://cran.obiba.org',  
dependencies=TRUE)  
install.packages('dsBaseClient', repos=c(getOption('repos'),  
'http://cran.obiba.org'), dependencies=TRUE)  
install.packages('dsModellingClient', repos=c(getOption('repos'),  
'http://cran.obiba.org'), dependencies=TRUE)  
install.packages('dsStatsClient', repos=c(getOption('repos'),  
'http://cran.obiba.org'), dependencies=TRUE)  
install.packages('dsGraphicsClient', repos=c(getOption('repos'),  
'http://cran.obiba.org'), dependencies=TRUE)
```

NOTE: soon also an ENPADASI R package will be needed to fully connect the Dash-In infrastructure.

The focus in this tutorial is the definition of the **get_study_variables()** function – as a demonstration of the web-based interactive analysis system offered within the Dash-In infrastructure for both intervention and observational studies.

GLOBAL.R

```
library(opal)  
library(dsBaseClient)  
library(dsStatsClient)  
library(dsGraphicsClient)  
library(dsModellingClient)
```



```
##  
## DATASHIELD commands  
# load the login file  
my_login<-read.table('../logins.txt', sep=";", header=TRUE)  
# log in to the remote servers  
# assign=TRUE will have the remote opal server instruct the remote R  
# instance to assign the dataframe into variable 'D'  
opals <- datashield.login(logins=my_login, assign=TRUE, symbol = 'D')  
# detect the list of variables in the study  
get_study_variables <- function(symbol="D") {  
  tryCatch({  
    ds.colnames(x=symbol)[[1]]  
  }, error = function(e) {  
    print(e)  
    return( list("No data was loaded! See error messages!") )  
  }  
)  
}
```

In the logins.txt files a list of different OPAL and DBNP DataShield-enabled servers can be entered.

The ***get_study_variables()*** function fetches the variables in the study. It also handles some error condition, for example no internet connection or remote servers not reachable.

At the same time we extend the application with all DataShield supported plots, i.e. **histogram**, **contourPlot** and **heatmap**.

SERVER CONFIGURATION AND DEPLOYMENT OF A MULTI-APPLICATION SERVER

DEBIAN

Install the shiny server

```
$ wget https://download3.rstudio.org/ubuntu-12.04/x86_64/shiny-server-  
1.4.2.786-amd64.deb  
$ sudo gdebi shiny-server-1.4.2.786-amd64.deb
```

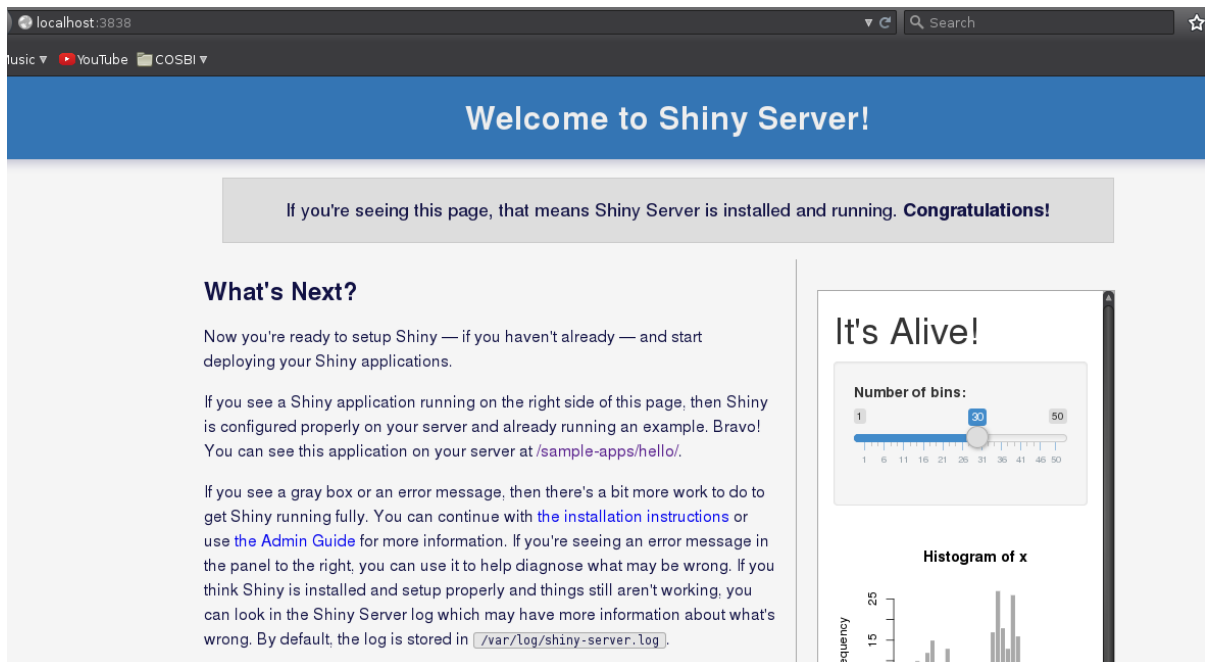
At this point the server should be automatically up running.



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

```
Applications Places System - The Comprehensive R ... how-to-shine.txt (~/Do...
Shiny Server
Shiny Server is a server program from RStudio, Inc. that makes Shiny applications available over the web. Shiny is a web application fra
putation language.
Do you want to install the software package? [y/N]:y
Selecting previously unselected package shiny-server.
(Reading database ... 161744 files and directories currently installed.)
Preparing to unpack shiny-server-1.4.2.786-amd64.deb ...
Unpacking shiny-server (1.4.2.786) ...
Setting up shiny-server (1.4.2.786) ...
Creating user shiny
Adding LANG to /etc/systemd/system/shiny-server.service, setting to en_US.UTF-8
Created symlink from /etc/systemd/system/multi-user.target.wants/shiny-server.service to /etc/systemd/system/shiny-server.service.
● shiny-server.service - ShinyServer
   Loaded: loaded (/etc/systemd/system/shiny-server.service; enabled)
   Active: active (running) since Wed 2016-06-01 14:20:57 CEST; 10ms ago
     Process: 22087 ExecStartPost=/bin/sleep 3 (code=exited, status=0/SUCCESS)
    Main PID: 22088 (shiny-server)
      CGroup: /system.slice/shiny-server.service
              └─22086 /bin/bash -c /opt/shiny-server/bin/shiny-server --pidfile=/var/run/shiny-server.pid >> /var/log/shiny-server.log 2>&1
                └─22088 /opt/shiny-server/ext/node/bin/shiny-server /opt/shiny-server/lib/main.js --pidfile=/var/run/shiny-server.pid
```

Test if it's running with the default configuration test: <http://localhost:3838>



localhost:3838

Music YouTube COSBI

Welcome to Shiny Server!

If you're seeing this page, that means Shiny Server is installed and running. **Congratulations!**

What's Next?

Now you're ready to setup Shiny — if you haven't already — and start deploying your Shiny applications.

If you see a Shiny application running on the right side of this page, then Shiny is configured properly on your server and already running an example. Bravo! You can see this application on your server at </sample-apps/hello/>.

If you see a gray box or an error message, then there's a bit more work to do to get Shiny running fully. You can continue with [the installation instructions](#) or use [the Admin Guide](#) for more information. If you're seeing an error message in the panel to the right, you can use it to help diagnose what may be wrong. If you think Shiny is installed and setup properly and things still aren't working, you can look in the Shiny Server log which may have more information about what's wrong. By default, the log is stored in `/var/log/shiny-server.log`.

It's Alive!

Number of bins:

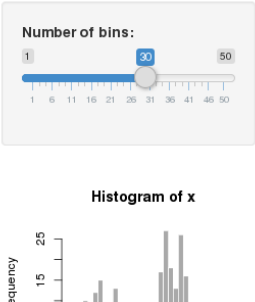
1 30 50

1 6 11 16 21 26 31 36 41 46 50

Histogram of x

frequency

15 25



The configuration file is located at `/etc/shiny-server/shiny-server.conf`

The file is well commented, so it will be easy to understand what to edit in order to get the desired configuration.

To change the port, search and edit the line:

```
listen 3838;
```

To change the address:

```
location /put/here/your/address { ...
```



To reload the server with the new configuration:

```
$ sudo service shiny-server stop  
$ sudo service shiny-server start
```

For the deployment of a multi-application server simply prepare different folders each containing its own ui.R, server.R (and optionally global.R) and the server will treat each such folder as a different application.

MAC

On Macs the shiny server needs to be compiled from source. It all passes through homebrew. Install homebrew with the following command:

```
$ /usr/bin/ruby -e "$(curl -fsSL  
https://raw.githubusercontent.com/Homebrew/install/master/install)"
```

Using homebrew install the following software:

- python 2.6 or 2.7 (Really. 3.x will not work)
- cmake (>= 2.8.10)
- gcc
- g++
- git

typing commands as the following:

```
$ brew install python
```

Install a development version of R available from ATT: <http://r.research.att.com/>

Then install the shiny package in the system-wide library:

```
$ install.packages("shiny", repo="http://cran.rstudio.org", type="source")
```

Now proceed with the first steps – **stopping before the CMAKE step** – under “Installation” on the official page at

<https://github.com/rstudio/shiny-server/wiki/Building-Shiny-Server-from-Source>



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

The current *launcher.cc* source file must be edited to use the `proc_pidpath()` function on OSX instead of Linux `proc` (see [this thread](#)). Use [this version](#) from Nathan Weeks instead.

After replacing the file, you can proceed with `cmake` and all subsequent installation steps.

See references:

<https://github.com/rstudio/shiny-server/wiki/Building-Shiny-Server-from-Source>

<http://www.ducheneaut.info/installing-shiny-server-on-mac-os-x/>

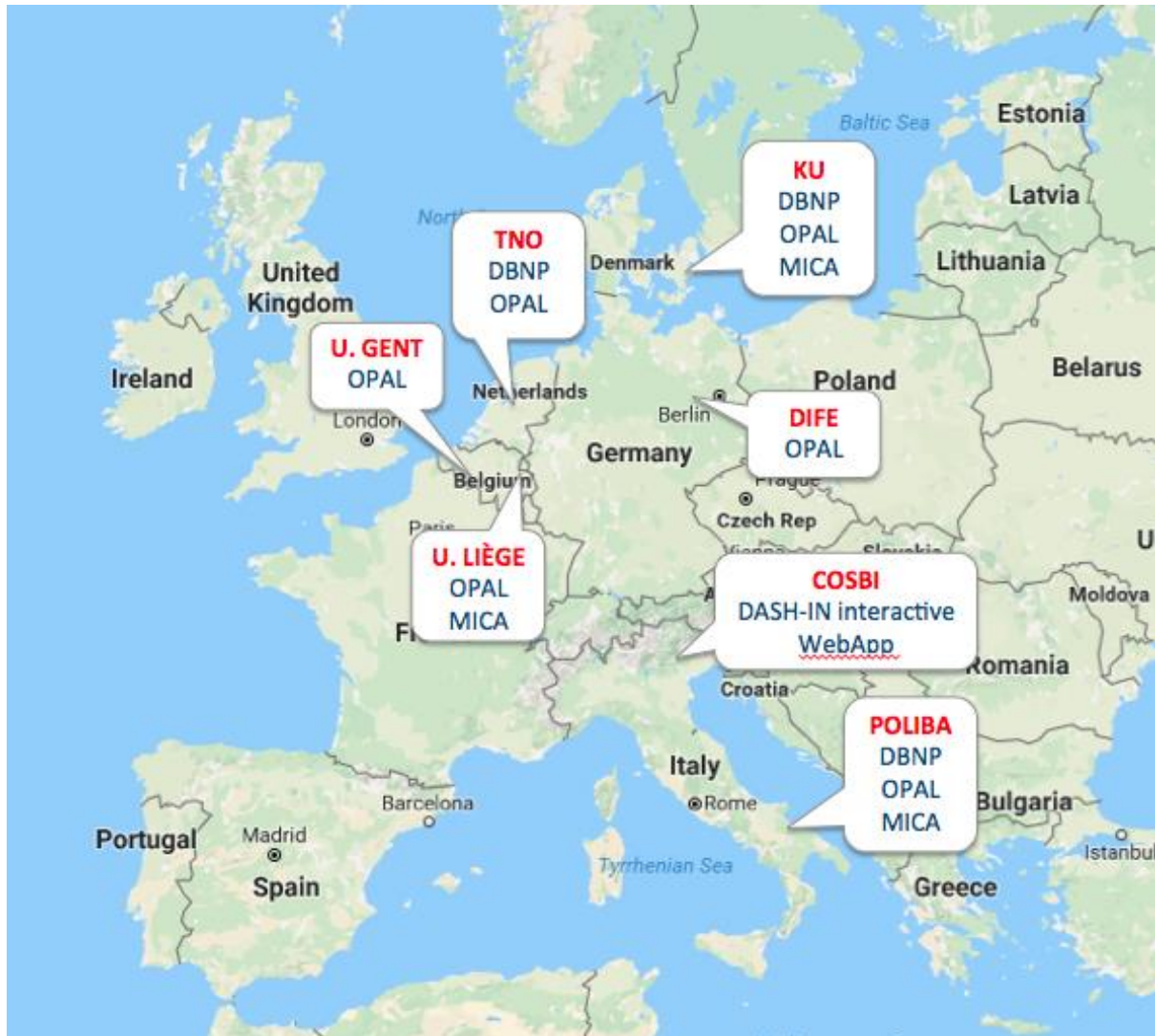
<https://groups.google.com/forum/#!topic/shiny-discuss/WTXFtrEnR-k>

<https://github.com/nathanweeks/shinyserver/>

[blob/d5240ef6d795dafc89c74a49d6f14d7fe0509541/src/launcher.cc](https://github.com/nathanweeks/shinyserver/blob/d5240ef6d795dafc89c74a49d6f14d7fe0509541/src/launcher.cc)



AVAILABILITY OF THE SYSTEMS



Within ENPADASI, the systems described above have been installed on different computational platforms. The instances are currently available to the entire ENPADASI community as follows:

- Phenotype Database: it is hosted by TNO [NL], ReCaS Data center [IT] and the University of Copenhagen [DK].
- Opal/DataShield instances are installed on ReCaS Data center [IT], TNO [NL], University of Liège [BE], Gent University [BE], the University of Copenhagen [DK] and DIFE [DE].
- A centralized instance of Mica server is accessible at ReCaS Data center.
- The Dash-In interactive web-based analytical platform is available at COSBI.