midas

Meaningful Integration of Data, Analytics and Services

Grant Agreement No. 727721

Contract Duration: 40 months (1st November 2016 – 29th February 2020)

This project is funded by The European Union

H2020-SC1-2016-CNECT
SC1-PM-18-2016 - Big Data Supporting Public Health Policies

Deliverable 2.3

MIDAS Framework User Guide

Circulation:	Public
Nature:	Report
Version #:	1.0
Issue Date:	25/07/2019
Responsible Partner(s):	SET
Author(s):	Paul Carlin
Status:	Draft (Living Document)
Reviewed on:	31/07/2019
Reviewed by:	MIDAS Executive Board
Contractual Date of Delivery:	31/10/2018 (M24)

Grant Agreement No: 727721

Executive Board Document Sign Off

Role	Partner	Signature	Date
WP1 Lead	Ulster	Michaela Black	29/07/2019
WP2 Lead	SET	Paul Carlin	29/07/2019
WP3 Lead	VICOM	Gorka Epelde	30/07/2019
WP4 Lead	KU Leuven	Gorana Nikolic	30/07/2019
WP5 Lead	VTT	Juha Pajula	26/07/2019
WP6 Lead	DCU	Regina Connolly	26/07/2019
WP7 Lead	Ulster	Jonathan Wallace	30/07/2019
WP8 Lead	Ulster	Michaela Black	29/07/2019
Scientific-Technical Manager	Analytics Eng	Scott Fischaber	29/07/2019

Grant Agreement No: 727721

Abstract

This deliverable along with deliverable 2.2 is now overdue, the reason for this is that the original timelines agreed mitigated against a final guide that reflected in final MIDAS platform as the user guide would have been completed well before the final iteration of the product. This “Living Document”, reflects a baseline of current form, but will be added to and amended to meet the needs of the final MIDAS platform.

Grant Agreement No: 727721

Copyright

The MIDAS project is funded under the EC Horizon 2020 SC1- PMF-18 Big Data Supporting Public Health Policies

Grant Agreement No: 727721

Document History

Version	Issue Date	Stage	Content and Changes
1.0	25/07/2019	Draft - Living Document	First draft of living document

Statement of Originality:

This deliverable contains original unpublished work except where clearly indicated otherwise. Acknowledgement of previously published material and of the work of others has been made through appropriate citation, quotation or both.

Grant Agreement No: 727721

Executive Summary

Work Package:	WP 2
Work Package leader:	SET
Task:	T2.2 Create New Model
Task leader:	SET

This deliverable describes the process of developing a living document that reflects the current state and functionality of platform and process. This will be updated on a regular basis as change happens within the project to deliver a relevant and accurate user guide for project completion.

Grant Agreement No: 727721

Table of Contents

Grant Agreement No: 727721

Grant Agreement No: 727721

1 Introduction

1.1 Scope and Purpose

This document will act as your guide when learning how to explore and use the MIDAS system. MIDAS is a platform that allows users to access a variety of datasets, from different sources and bring them together to examine potential relationships, dependencies or causal associations that may impact health. Several tools exist in the platform, and you will be shown how this work both independently and within the system. The aim is to allow a variety of users such as politicians, analysts, policy aides and civil servants to have access to the system that delivers information to inform, deliver and evaluate policy.

The user guide will therefore be structured to provide information in a manner relevant to, and accessible by, a variety of users using an assortment of methods:

- Super-Text – This will provide a high-level overview of the functionality of the individual components and the system.
- Diagrams – These will provide an accessible visual representation of the functionality of the individual components and the system.
- Embedded Video/ Pictures – These will allow more detailed explanations and representations of actual use of the individual components and the system.
- Links to articles – Embedded links to articles of interest that support the MIDAS approach, throughout the document.

2 System Overview

The MIDAS platform brings together the following technologies and systems:

- Governance, Ethics and Quality Assurance
- Data Harmonization
- System Interoperability
- Data Analytics
- Privacy Preservation
- Application Programming Interface (API)
- Reporting

These will allow the various users to select the functions, tools and reports that meet their individual needs.

Grant Agreement No: 727721

2.1 Governance, Ethics and Quality Assurance

The MIDAS platform delivers a system of review and access for users of data that meets the requirements of individual member states legislative controls, within the overall context of the General Data Protection Regulation (GDPR)[1].

Each user can be assured that processes exist that allow data to be presented by the system while both: a) assuring the anonymity of any individual, and; b) allowing the system to gain enough insight to allow meaningful analysis and thus generation of knowledge for policy. The core principle is that each dataset that is accessed by the system is de-identified and cannot be re-identified using novel technologies and approaches. The data provider assures that the data that they release or allow the system to access contains no Personal Identifiable Data (PID) and is thus, under certain circumstances, exempt from the GDPR[2].

The notional use of MyData (D1), a component being examined within the context of MIDAS is an approach that could drive access to PID through a model of consent leveraged by technology; however, this is only mentioned as a potential model of wider access to data and system of control (figure 1).

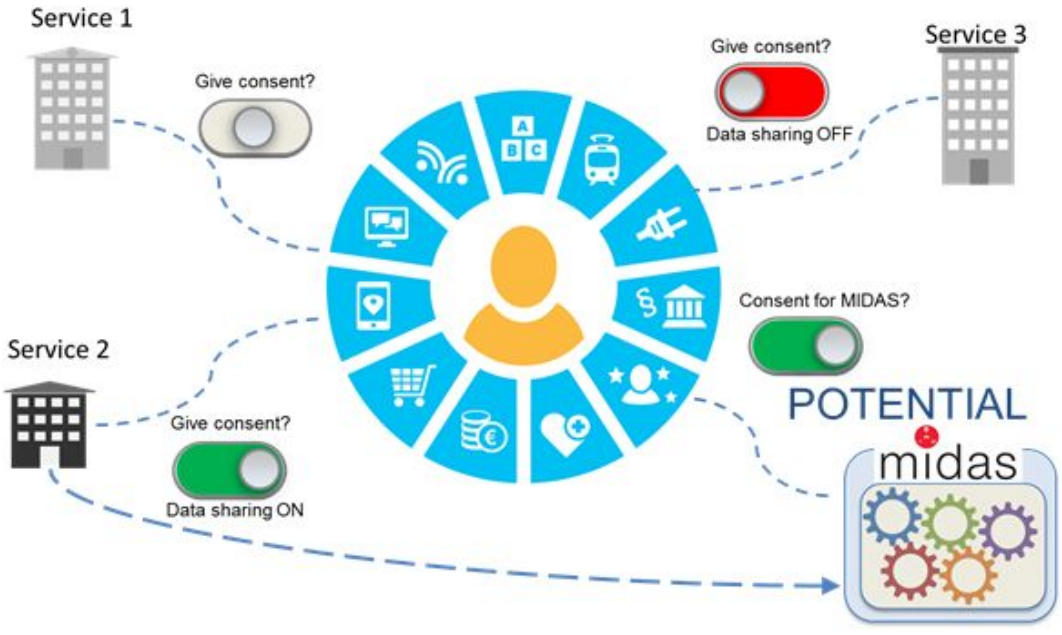


Figure 1: MyData Model

[1] https://eugdpr.org/the-regulation/

[2] https://ico.org.uk/media/for-organisations/guide-to-the-general-data-protection-regulation-gdpr-1-0.pdf

Grant Agreement No: 727721

2.2 Privacy Preservation

MIDAS uses Federated learning[3] (figure 2) and differential privacy[4] to drive privacy preservation. Rather than exchange of information to a centralised resource, learned parameters are shared between parties, updated and remodelled to refine analysis.

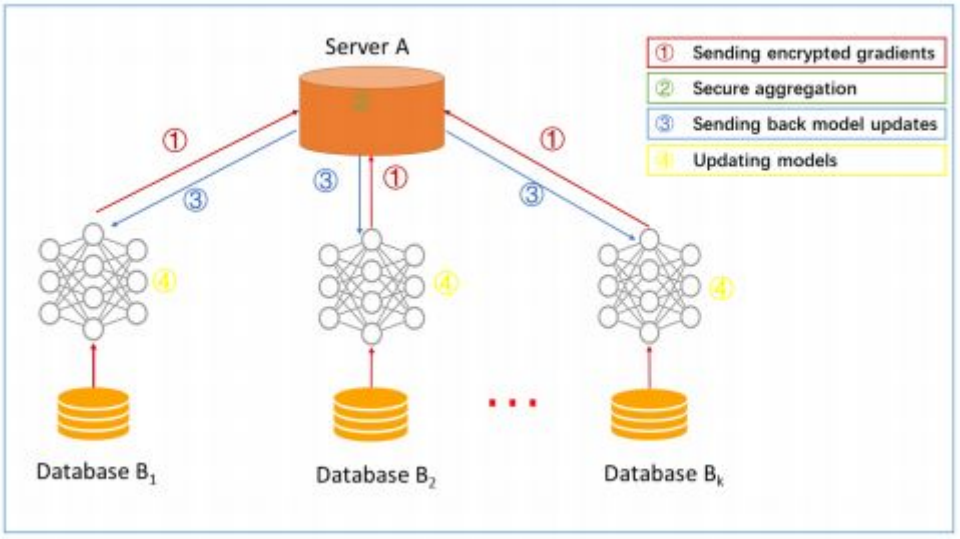


Figure 2: Horizontal Federated Learning System (reproduced from Yang et al (2019)

In data handling it is essential that privacy aspects are considered. The model supports describing data of various identity levels including, and especially, pseudonymised and anonymized data. This common model is also capable of handling person identifiable data (PID) although such data is not processed in the framework of the MIDAS project. For example, in a post-project scenario person identifiable data may be needed in order to combine person-level data retrieved from different sources. This is further enhanced using a brokerage model proposed by Apple that assures privacy, with a broker acting for all clients to refine the model, gain insight into the data and drive a linear regression modelling analysis for decision making.

2.3 Data Harmonization

Data harmonization refers to the different data preparation tasks to combine data from different sources (with different types, levels and sources) and provide users

[3] https://ai.googleblog.com/2017/04/federated-learning-collaborative.html
[4] https://ai.google/research/pubs/pub45428

Grant Agreement No: 727721

with a comparable view of data from different studies. The requirements for data harmonization, can be described as:

1. data cleaning,
2. data normalization,
3. data transformation,
4. missing values imputation as well as noise identification

There is a requirement on the data supplier to describe the characteristics within the proffered dataset and how each organization contributing to the MIDAS model manages this will remain that organizations responsibility. These characteristics must be communicated to the MIDAS platform owner following the established data description methodology and procedure. Data owners can then prepare their datasets through the project developed GYDRA (Appendix 1) tool for internal decision-making analytics or for cross-site analytics (having first agreed the target harmonised data structure with corresponding site partners).

A variety of data sources can be utilized within the MIDAS platform, and as such data privacy preservation should be a core focus for data providers, as well as for the platform owners and architects. Therefore, policy leads and services who will utilize the MIDAS platform should have the appropriate systems of governance, audit and control in place for appropriate sharing.

The main analytics platform can ingest tabular data from a variety of sources:

- City / Government generated datasets – Clinical data sets, social care data, economics data etc.
- 3rd Party generated data.
- Government open data.

Additionally, the MIDAS platform tool allows to monitor a variety of data sources with global coverage:

- Social media data – Twitter
- Media sources data
- Scientific publications – Medline/Pub-med
- Worldwide news monitoring in 60+ languages
- Crowdsourced data

The policy question under consideration will influence how data is ingested, managed and actioned within the MIDAS system and interface. Therefore, there is a requirement for the users to help identify the policy under review and invest in

Grant Agreement No: 727721

identifying a usable model for analysis. This is elaborated upon further when discussing the ISAACUS model.

2.4 System Interoperability

2.4.1 Coding Systems

Health Level 7 Clinical Document Architecture (HL7 CDA) forms the spine for clinical systems across the current consortia partners, however, a variety of health coding systems exist within multiple domains across the partner regions

Domain	Basque country	Finland	Northern Ireland	Republic of Ireland
Diagnosis[5]	ICD–10, NANDA-I	ICD-10, ICPC-2	Read v2, ICD-10, ICD-9, CTV3, SNOMED-CT v3.5, UDDA	ICD-10-AM/ACHI/ACS 8th Edition, ICPC-2
Procedures	Prescription: NIC-NANDA, Local code Execution: ICD – 10, NIC-NANDA	THL Procedure classification (national), ICPC-2	OPCS 4.8, NICIP (imaging)	HIPE, ICPC-2, ICD-10-AM/ACHI/ACS 8th Edition (Grouper for DRGs)
Measurements and observations	LOINC	Nomenclature of Laboratory Investigations (national), LOINC (FinLOINC)	Read v2	LOINC, Moving to Snomed-CT
Medication	DOE, ATC	ATC, VNR	dm+d	Data model for an electronic medicinal product reference Catalogue - a National Standard[6], Snomed-CT
Others	Pathology: SNOMED-CT Individual clinical variables: RIC (Local coding)			

[5]Symptoms are considered as part of diagnoses and are coded using R codes from ICD-10

[6] https://www.hiqa.ie/reports-and-publications/health-information/data-model-electronic-medicinal-product-reference

Grant Agreement No: 727721

	Image (X rays): Local coding Nursing outcomes: NOC-NANDA			

Table 1: Coding systems for Health

Some of the MIDAS data sets describe similar target groups. Despite this, most of the data is specific to one data set only, and it is not feasible to pursue a direct combination of the data sets from different partners. Therefore, a common data model extending to the lowest data item level is not the best approach for the project, rather, it is essential that the contents of each dataset is machine-readable allowing it to be automatically processed and visualized by the analytics platform. In order to facilitate this, a common metadata model is needed.

2.4.2 The Common metaData Model

The MIDAS Common metaData Model (MCDM) has adopted the ISAACUS model created by THL (as a result of study of different models for research and statistics data management) which is itself based on the Generic Statistical Information Model (GSIM)[7] Framework and largely exploits the (Data Document Initiative) DDI 3.2 concepts (DDI concepts are used for specifying the metadata model elements). It provides an architecture for data interoperability, assuring classification on levels of confidentiality and allowing access to systems for analysis and visualization.

The requirements the MCDM must meet is given in the table below:

Requirement	Description
Support for both micro and macro data	The source data sets include data at individual person level as well as data aggregated in various dimensions (e.g. time, region, age group).
Support for data from all relevant domains	The common model shall cover a wide spectrum of data related to health and wellness. Relevant data are e.g. clinical healthcare data, register data, research cohorts, biobank data, environmental data, data gathered by the individual (MyData).
Support for describing data confidentiality and access condition	Concerning, both automatic and manual use of the data, it is important that, along with the data, the confidentiality level of the data and as conditions for data access are defined.

[7] https://unstats.un.org/unsd/classifications/expertgroup/egm2015/ac289-22.PDF

Grant Agreement No: 727721

Scalability and sustainability	A common data set model is needed in order to achieve the direct objectives of the MIDAS project. However, it is highly important that the MIDAS architecture can be exploited after the project among the project consortium and beyond. Therefore, it is important that the common model is flexible enough to be adaptable to the existing infrastructure in different countries.
Exploitation of existing models and standards	As revealed in Section (2) a large effort has already been invested in the development of data set models and related standards. It is advisable to exploit the existing models and standards instead of developing a completely new MIDAS model.
Optional support for alternative coding systems	As revealed in Section (2.4.1) the coding systems are variable between countries. The common model seeks to adapt all data to unified coding systems. However, it is desirable that the common model enables alternative coding systems to be bound if such a need arises.

Table 2: Common data representation model requirements.

Using a common metadata model based around ISAACUS, data is described and catalogued, which takes place on an accessible server with two interfaces: an editor and a catalogue (for large dataset descriptions metadata import in CSV format is suggested).

Starting from the metadata description, the following data harmonisation approach has been defined:

- variable cataloguing and annotation,
- harmonizable variable identification
- data transformation for dataset harmonisation.

2.4.2.1 Datasets cataloguing approach following ISAACUS metadata model (variable cataloguing and annotation)

The following gives a brief overview when cataloguing in the ISAACUS model:

- MIDAS data sets should be described in a standard format, this will ensure homogenisation for the data providers thus allowing understanding and minimization of effort when ingesting. A general description of the dataset should be provided, as well as its context and structure, including any generally observed issues, especially regarding the quality of data. Then, each file which comprises the dataset should be described, including the way

Grant Agreement No: 727721

in which it is related to the rest (e.g. which are the common identifier variable(s)/field(s) used to link data among files).
- For each file, to understand the levels which may exist in the data, for example in spreadsheets, a description of the lower level structure and the sheets in the spreadsheet should be provided, with a clear process articulated for any linkage given.
- If the quality of data differs for a level (i.e. file, sheet), it should be described and include a description of the specific issues.
- Finally, for each variable, it should be described with an explanation of the content, i.e. the type of the data (e.g. integer, Boolean, date). When appropriate, additional information about the variable should be provided (e.g. format, coding standard, pre-processing).

There might be a great variability in the complexity of the data and structure which limits the applicability of this process. It gives a clear structure for tabulated data and should be used as a reference for other types.

The ISAACUS metadata model for insertion into the ISAACUS server allows for three methods:
- Assigning the variable (Instance Variable in the ISAACUS metadata model) to at least a concept variable or variable classification index (Variable in the ISAACUS metadata model). Variables can correspond to one or more classification index terms.
- Assigning topics (Adding Concept Scheme element to Concept in the ISAACUS metadata model), from platform loaded concept schemes (i.e. YSO, MeSH and TERO ontologies). This is reflected as keywords in the ISAACUS server tool.
- Adding free topics (Instance Variable's free Concepts field in the ISAACUS metadata model) to the variables, open terms are not restricted to the loaded vocabularies.

The variable annotation will allow a classification index that will guide the identification of harmonizable variables, restricting the initial search space. Based on the state-of-the-art analysis, the classification index selected for variable annotation is The Maelstrom Classification will be used as the classification index for variable annotation (Appendix 2).

As a complementary step, each variable will be annotated with at least a concept from the loaded concept schemes.

Grant Agreement No: 727721

2.4.2.2 Setting up the ISAACUS server

A dedicated metadata model server has been set up using an open source implementation of the ISAACUS model. The ISAACUS server GUI has been localised to English. The ISAACUS server application has two interfaces;

- an editor where the description of the datasets is carried out and
- a catalogue where, once the dataset is described, it can be published.

In addition to allowing the description of datasets, the ISAACUS editor also allows to import / export the description of the variables of a table in a csv file. Each of the fields of the variable description is represented by a column and each of the rows of the csv represents each of the variables.

The illustrations below show the dataset editor in an exemplar dataset, DIGS (Diabetes Insulin Guidance System):

Grant Agreement No: 727721

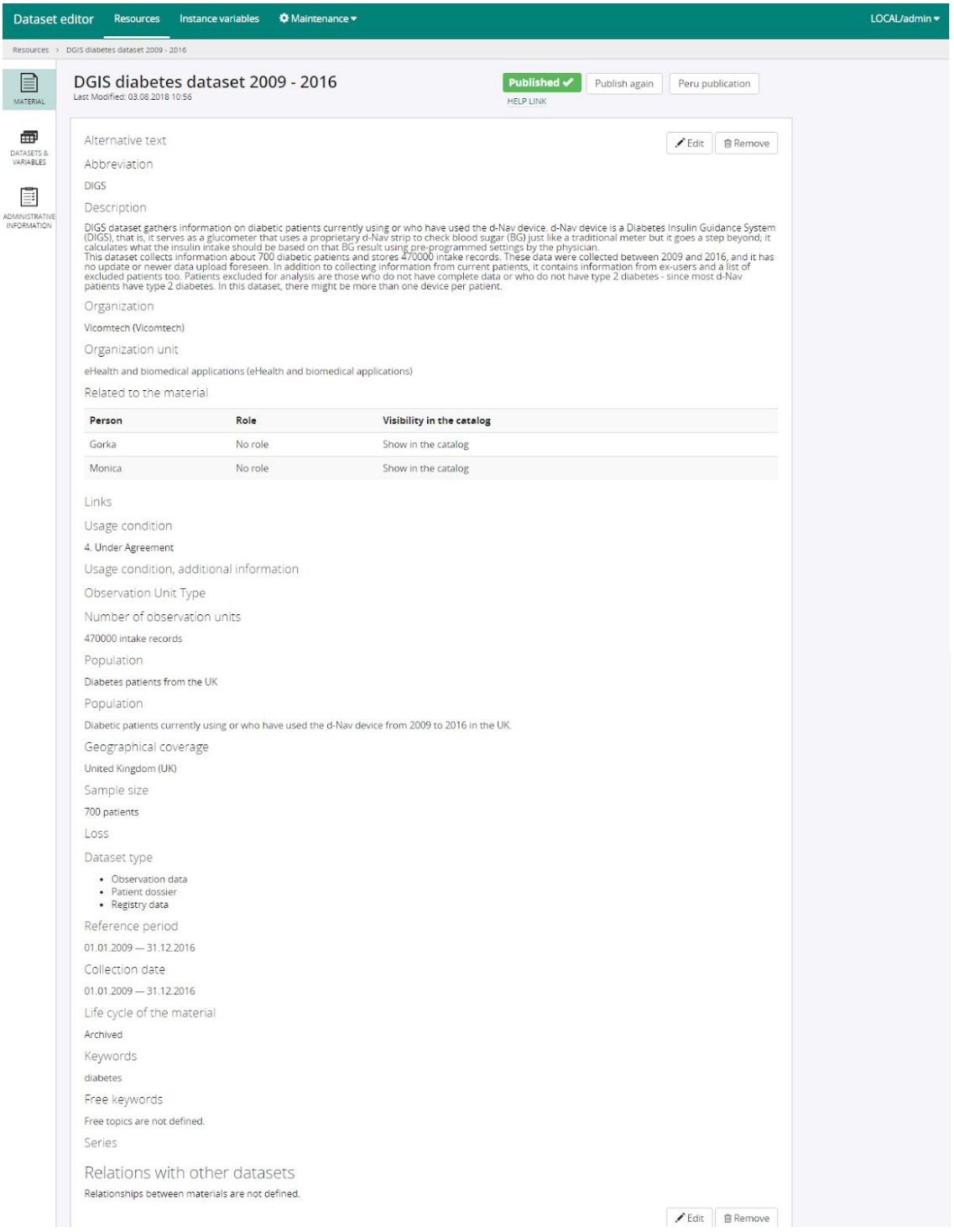


Figure 3: Isaacus data set editor - DIGS dataset description.

Grant Agreement No: 727721

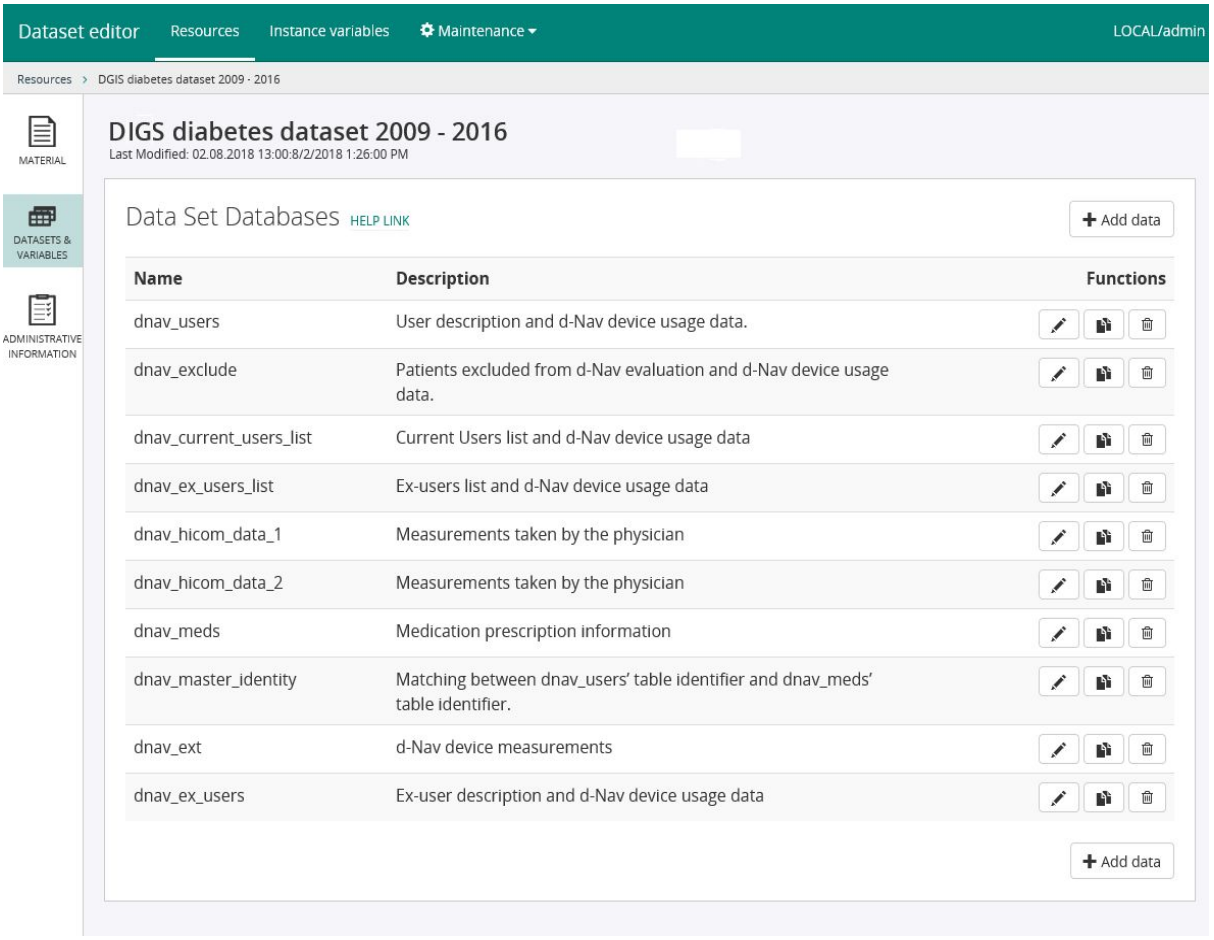


Figure 4: Isaacus data set editor - DIGS dataset tables.

Grant Agreement No: 727721

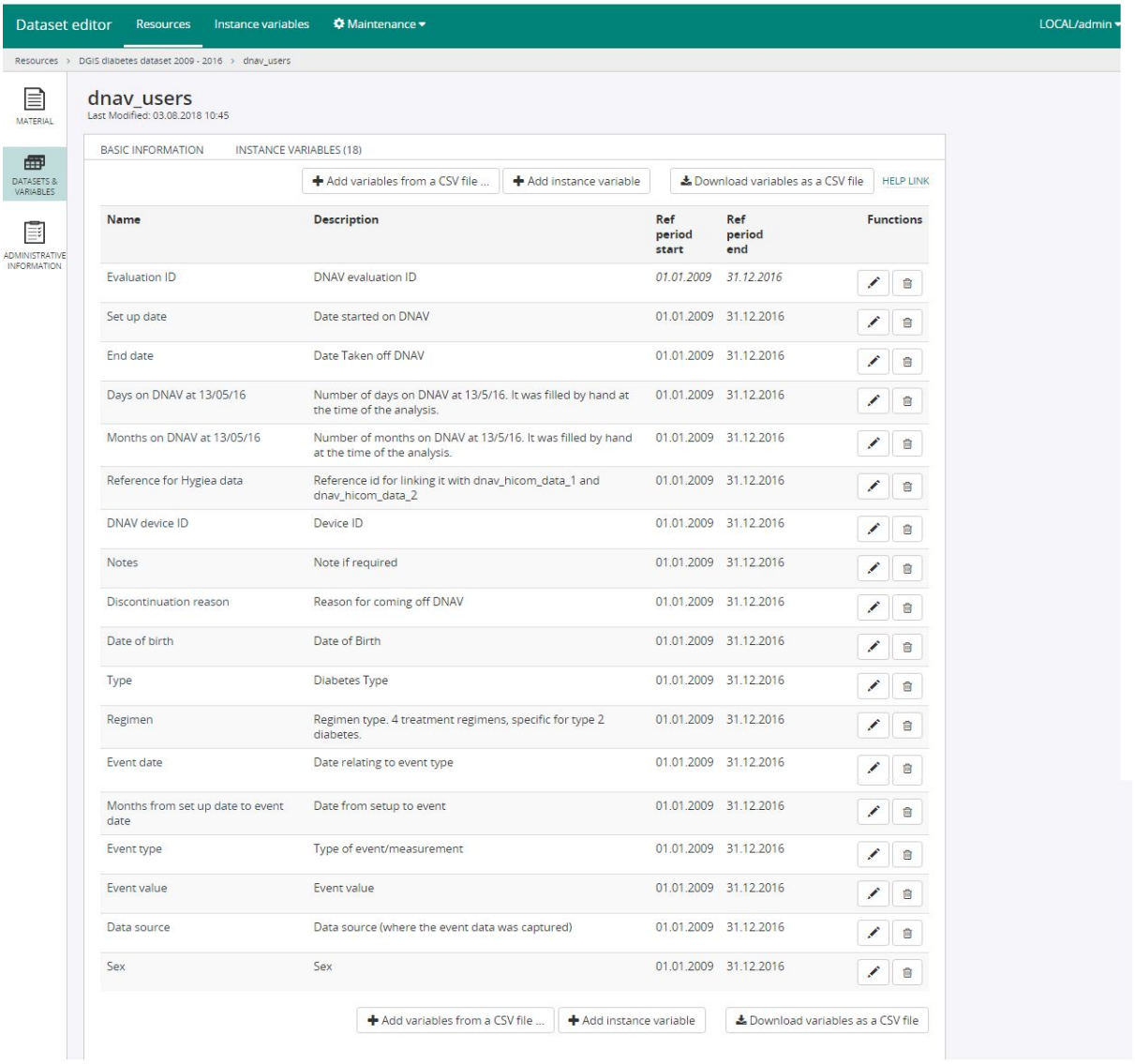


Figure 5: Isaacus data set editor - DIGS dataset "dnav_users" table variables.

Grant Agreement No: 727721

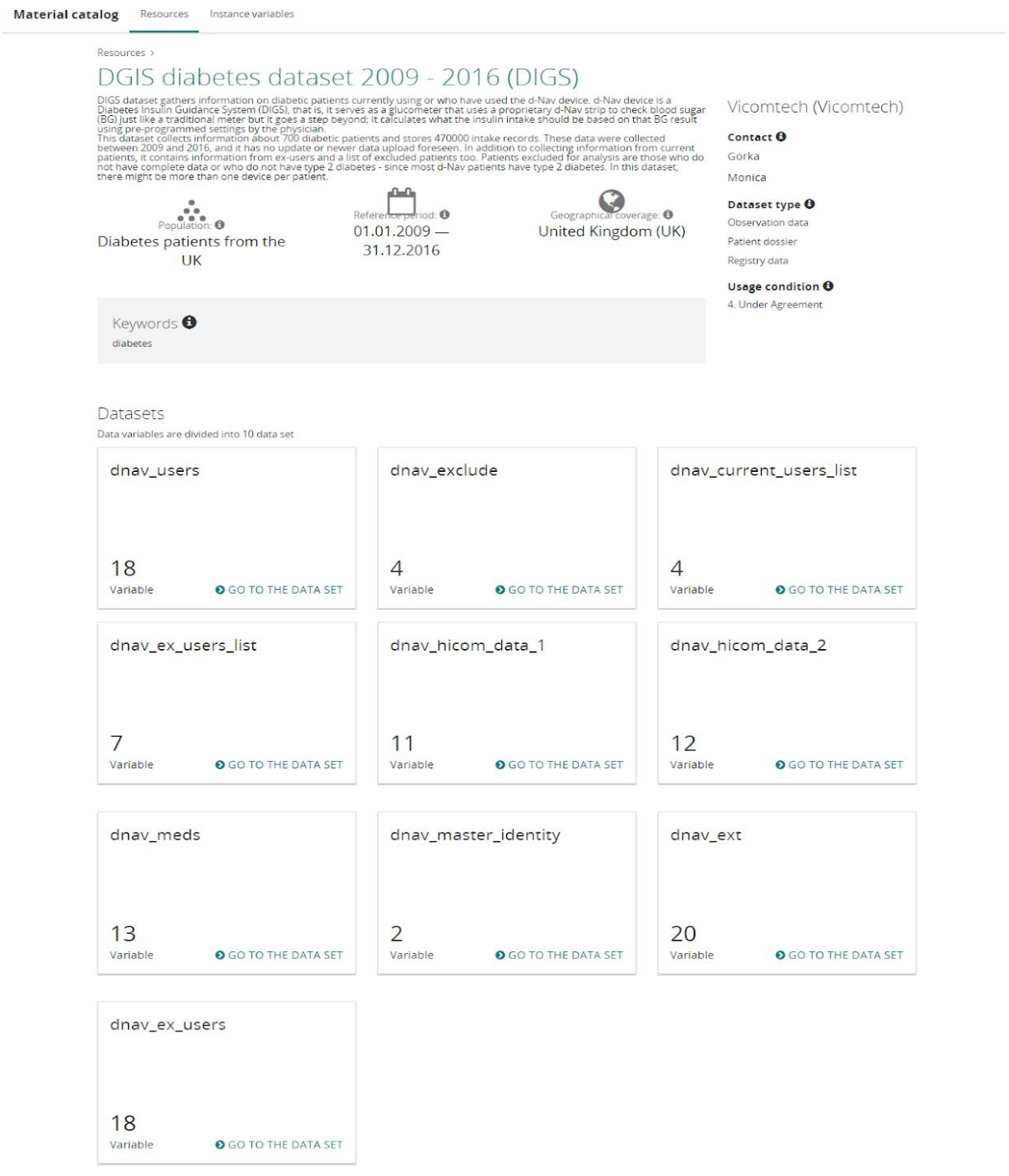


Figure 6: Isaacus data set catalogue - DIGS dataset description.

Grant Agreement No: 727721

2.4.2.3 Harmonizable variable identification strategy (harmonizable variable identification)

The harmonizable variable identification strategy begins with the annotated classification index, which examines domains and subdomains using Maelstrom Classification, and identifying where different data sources share elements.

Based on this summary and per variable classification index information, the ISAACUS server catalogue obtains further information on the variables. Additionally, the ISAACUS server's search functionality allows users to search among variables by name, description and keywords.

The following figure shows the results of the ISAACUS server's search functionality:

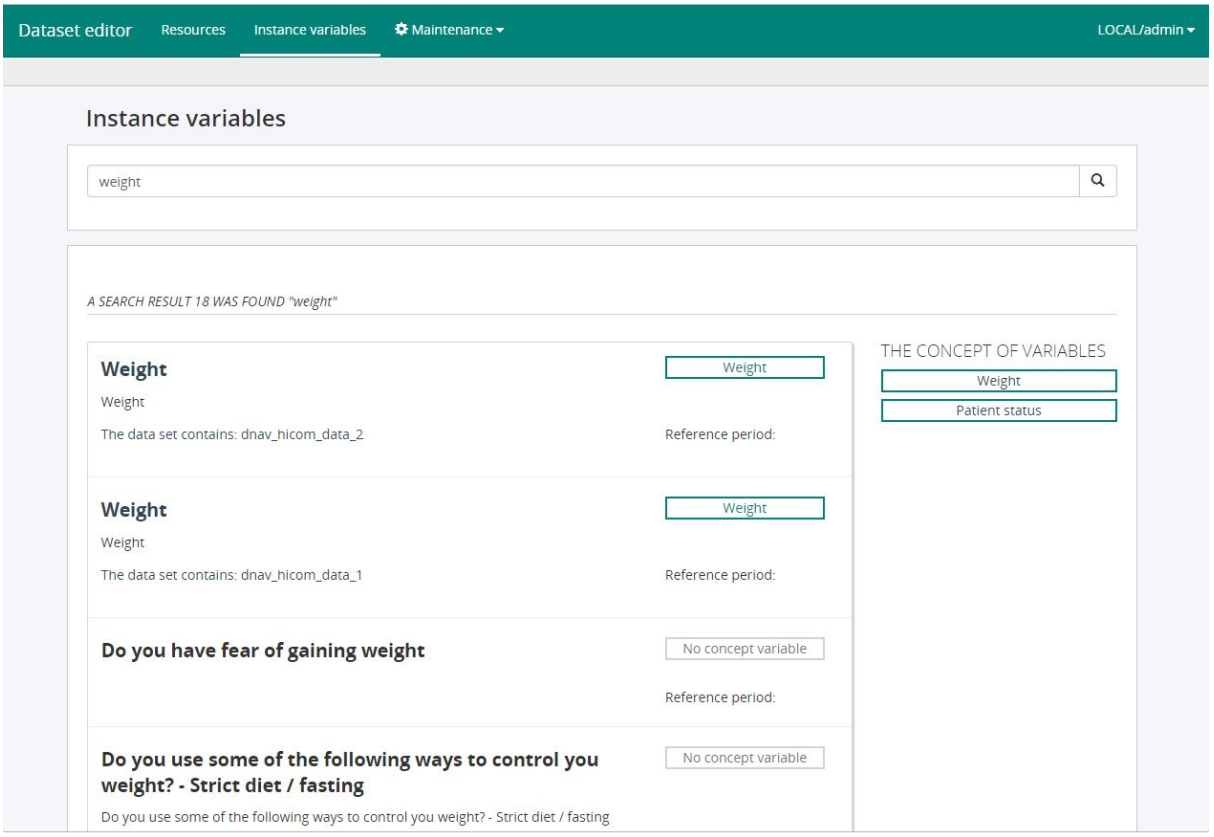


Figure 7: Results of the ISAACUS server's search functionality.

From the search results of the ISAACUS server's search functionality, we can also identify concept variables related to the search and identify misclassifications or navigate to variables under certain classification index by clicking on the concept variable item. The following figure shows the resulting variables for a classification index:

Grant Agreement No: 727721

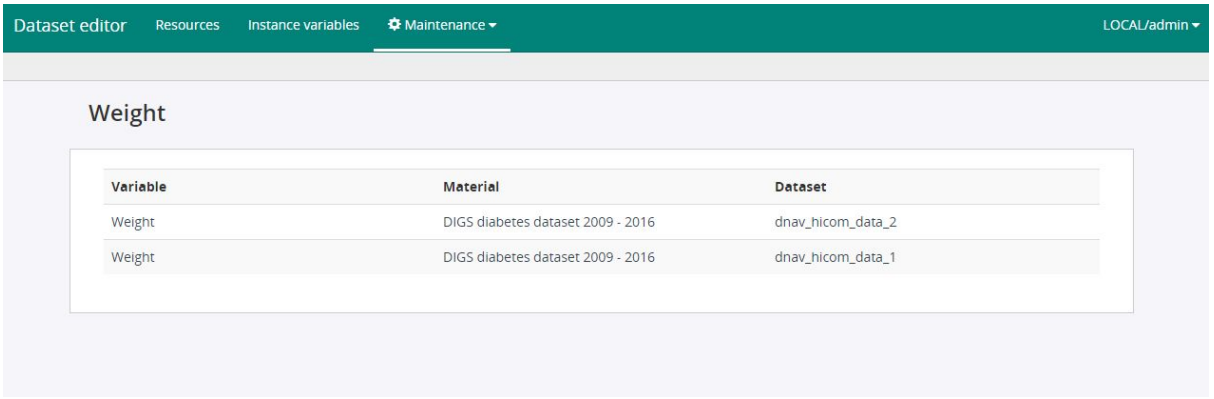


Figure 8: Resulting variables for a classification index.

2.4.2.4 Metadata usage for semi-automatic generation of analytics and visualisation required information files

The MIDAS Dashboard and the analytics widget wizard require the analytics platform to feed Open VA logic with metadata on the available data, analytics and visualisations.

The metadata of the datasets described on the ISAACUS server, is converted using generated scripts into the format required by the analytics platform to feed the MIDAS Dashboard with the required information.

The script is semi-automatic since the ISAACUS metadata model currently does not provide information on how datasets could be joined (e.g. possible primary / foreign keys or through which variables the datasets could be co-analysed), therefore this information is being extracted and integrated manually from dataset descriptions. Despite the fact that the ISAACUS metadata model has a DataSetRelation object (Figure 3.3a), in the current ISAACUS server implementation, only predecessor (i.e. previous study covering similar topic) type relation is implemented (which is generic level information and not providing exact information on how datasets could be joined).

Also, currently the ISAACUS CSV export does not provide descriptions of the tables, so this information is also being added manually from dataset descriptions. Additionally, some MIDAS analytics platform deployment and dataset setup variables (e.g. Apache Hive deployment configuration or Hive database under which datasets have been loaded) need to be set as script parameters for the correct creation of the information files for the analytics platform.

Grant Agreement No: 727721

Database table creation queries are also generated from the metadata using a script, helping in the task of reloading modified datasets.

2.4.2.5 Data transformation approach for dataset harmonization

Upon identifying harmonizable variables across datasets and agreeing on a final set of variables along with their units and coding, the corresponding metadata is inserted into the ISAACUS server allowing datasets to be transformed.

To allow this to happen MIDAS includes the Get Your Data Ready for Analysis (GYDRA) tool (Developed from TAQIH).

Currently the GYDRA tool allows the user to define and run a pipeline of dataset modification actions that target quality improvement, which supports data harmonization tasks and related data transformation, to obtain a target dataset on CSV starting from a source CSV.

Supported harmonization functionalities include:

- Deleting a variable (GYDRA DROP feature)

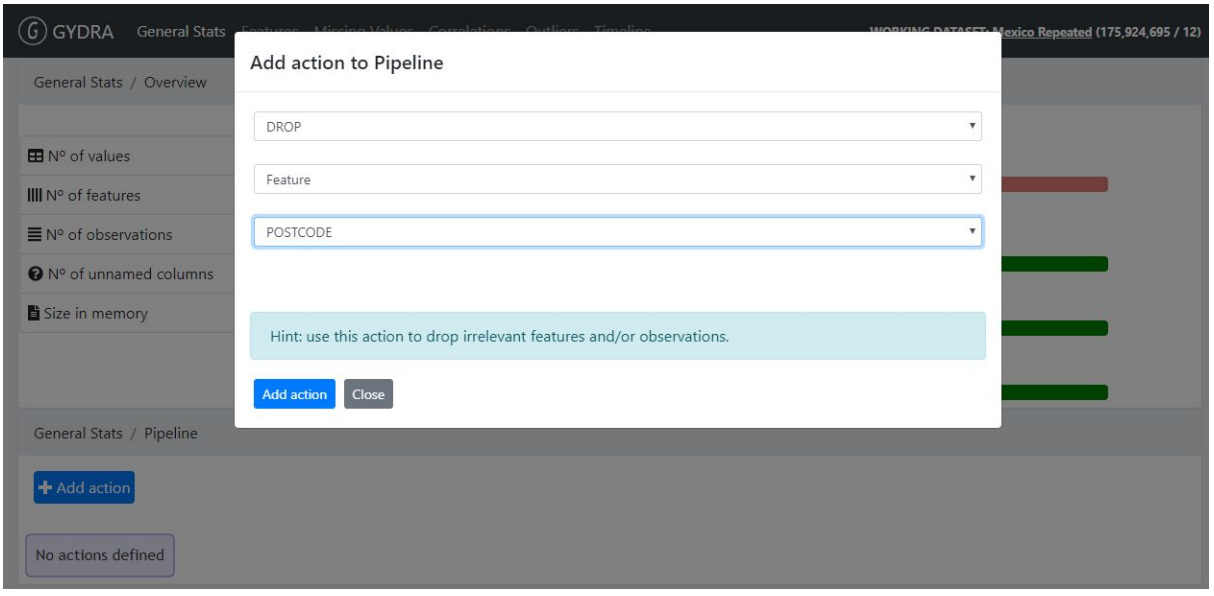


Figure 9: GYDRA DROP feature

Grant Agreement No: 727721

- Renaming a variable (GYDRA RENAME feature)

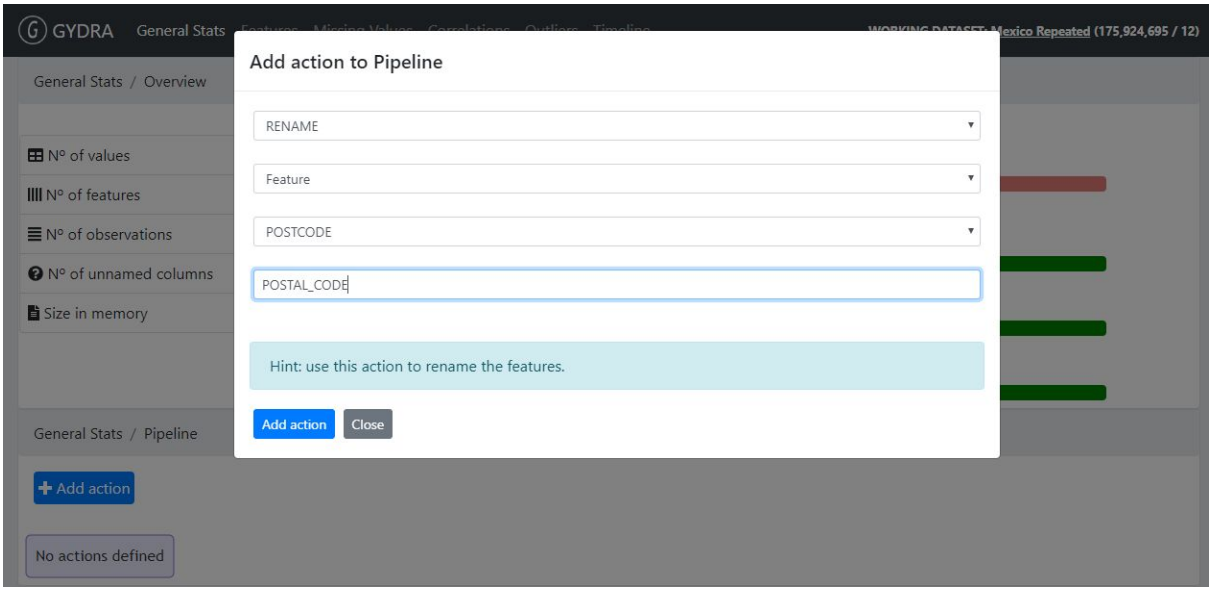


Figure 10: GYDRA RENAME feature

- Changing coding of categorical values (GYDRA CHANGE_VALUE)

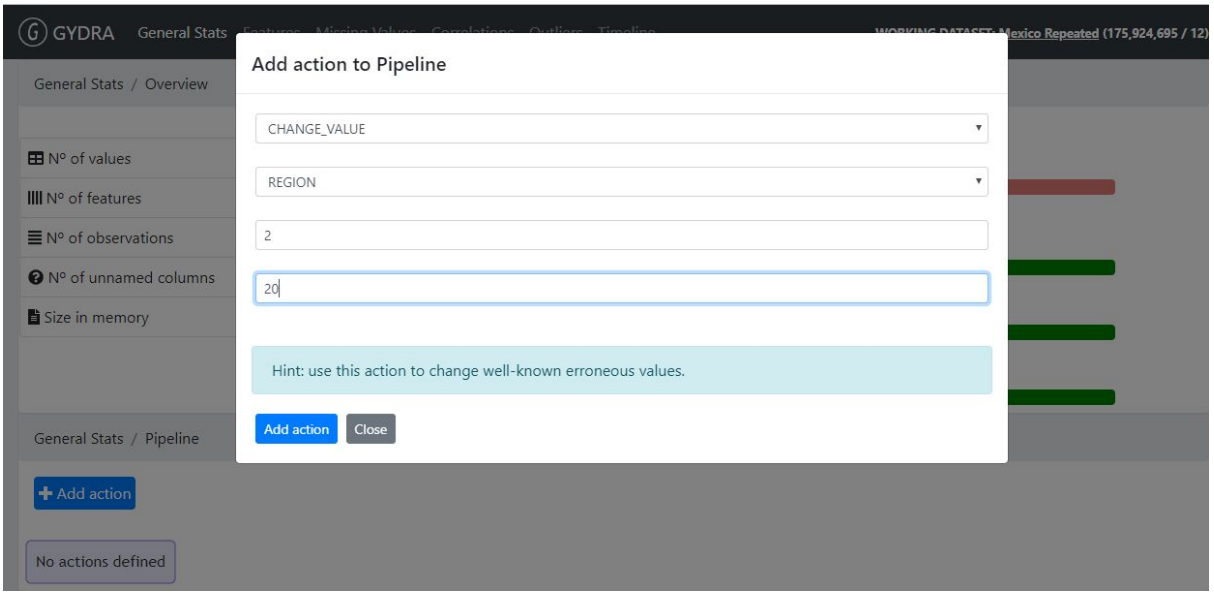


Figure 11: GYDRA CHANGE_VALUE feature

- Continuous unit changes (GYDRA OPERATE with add / subtract / divide / multiply operators)

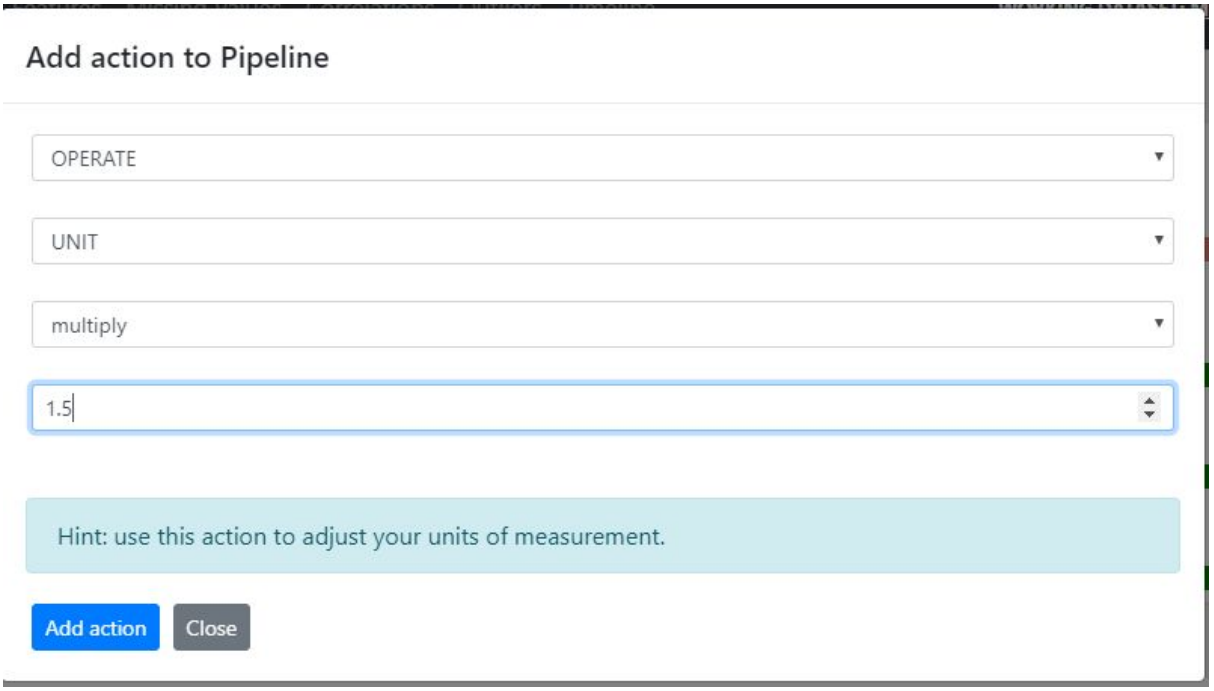


Figure 12: GYDRA OPERATE feature with add / subtract / divide / multiply operators

A number of data transformations can be achieved within existing GYDRA functionalities, although new functionalities will be developed such as operating over two variables to generate a new one (e.g. BMI value based on height and weight) and more complex transformations to target integrating external services (e.g. transformation of clinical codes).

Additionally, GYDRA allows the user to define and update dataset transformation pipelines visually and interactively.

Grant Agreement No: 727721

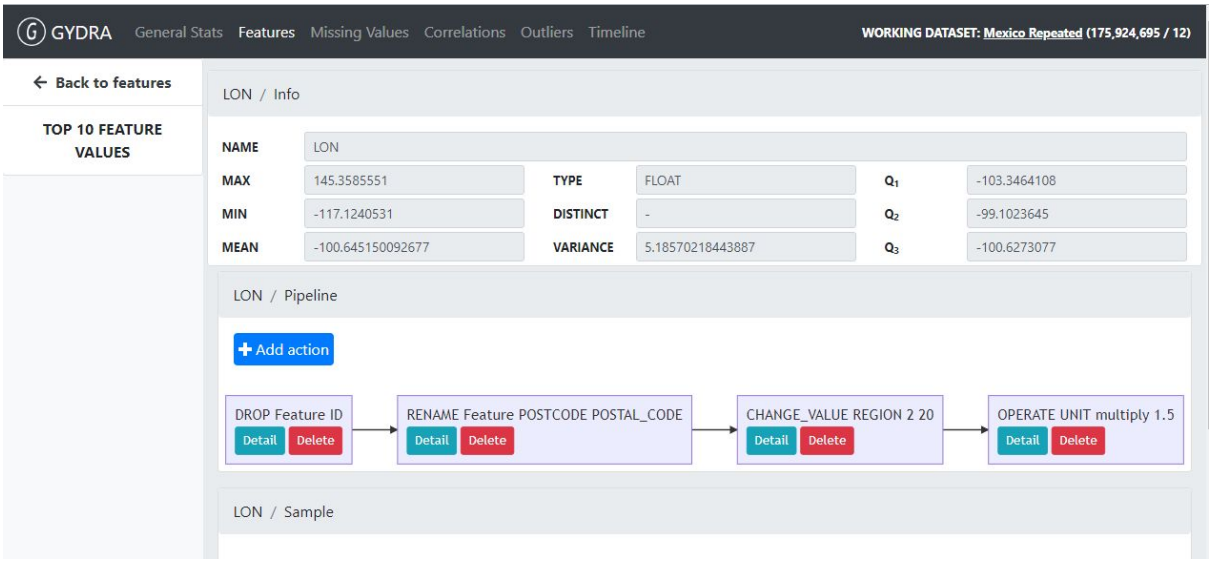


Figure 13: GYDRA visual and interactive dataset transformation pipeline definition

The resulting transformed dataset in CSV format, will be placed in Hadoop Distributed File System (HDFS) and loaded into Apache Hive.

The video available here demonstrates overall functionality.

2.5 Data Analytics

The Data analytics of the MIDAS platform is built upon a common data analysis model, which cascades through a defined sequence to final visualization within the platform:

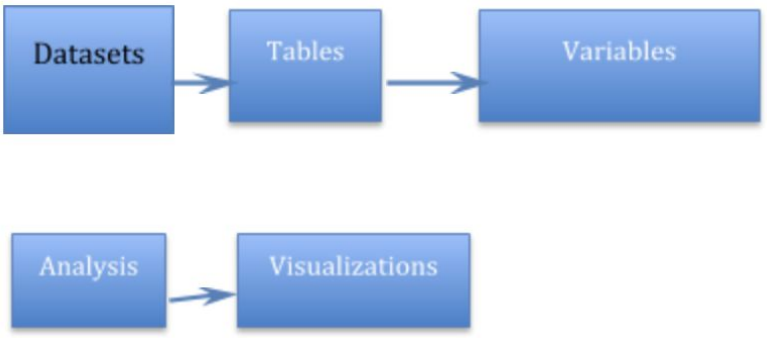


Figure 14: Common Data Analysis Model

Using several sophisticated technologies, a variety of users can interact with the system to analyse and present results for review and decision making.

Grant Agreement No: 727721

These are driven using Application Programming Interfaces (APIs) to allow functionality for:

- Datasets
- Tables
- Variables
- Analytics
- Visualization
- Render visualization

Visualizations can include:

- Histograms,
- Scatter plot
- Time series
- Correlation matrix
- Bar Chart
- Pie Chart
- Bubble plot
- Lexis Rate analysis

This creates functionality as follows:

Analytics	***Variable Types***	***Visualization***
Scatterplot	Numerical	Line graph
Correlation matrix	Numerical	Heatmap matrix
Histogram	Numerical	Histogram
Time Series	Numerical, Datetime	Line graph
Bar Chart	Numerical, Categorical	Bar graph
Pie Chart	Numerical, Categorical	Pie graph
Bubble Plot	Numerical, Categorical	Bubble plot
Lexis rate analysis	Categorical	Line graph, Bar graph, Chronopleth map, Heatmap matrix

Table 3: Analytics/ Variable/ Visualisation Matrix

Grant Agreement No: 727721

3 User Interfaces

3.1 The MIDAS Dashboard

The system is designed for several users, with functionality tailored to role and need. For example, users could include policy makers, data scientists or civil servants to create insight into policy. Wedded to this is the ability to create analytics and reports that can be shared with key decision makers as needed.

The system has an admin function that allows control and permits a user to be registered. When an account is created it must be by someone with Admin rights, which will allow the generation of an individual user account on the platform.

3.1.1 Account Generation

The user must have a username and password which is created in the system as shown in figure 14 below.

3.1.2 Login

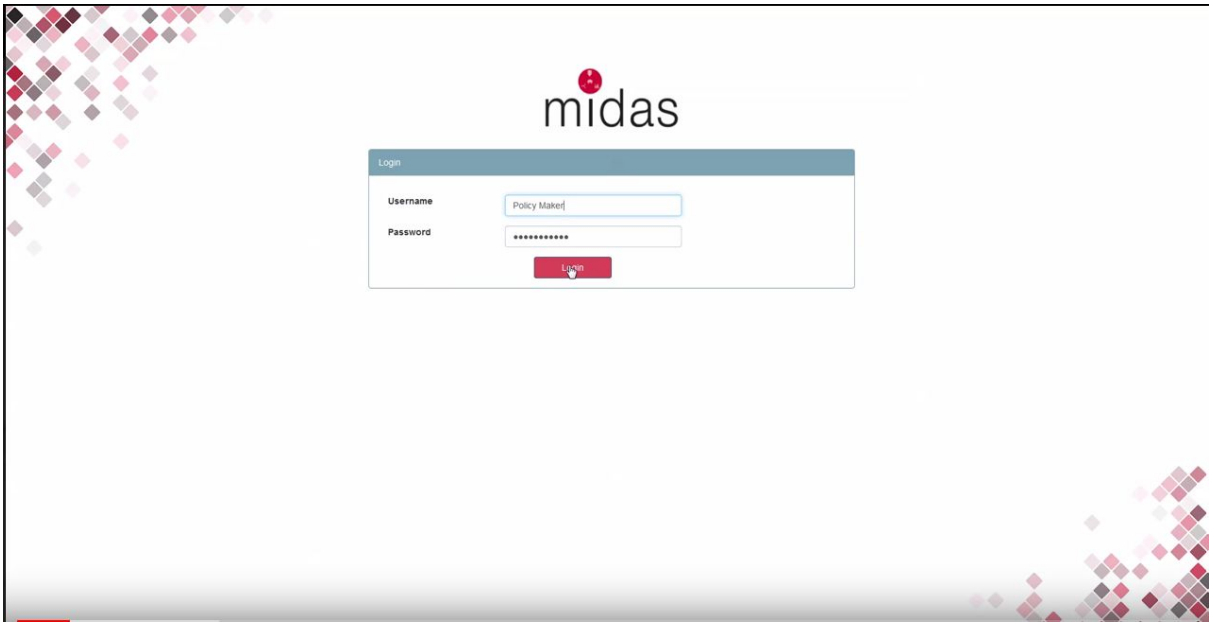


Figure 15: Login Screen

Once Logged in the user is brought to the MIDAS dashboard

Grant Agreement No: 727721

3.1.3 The Dashboard screens

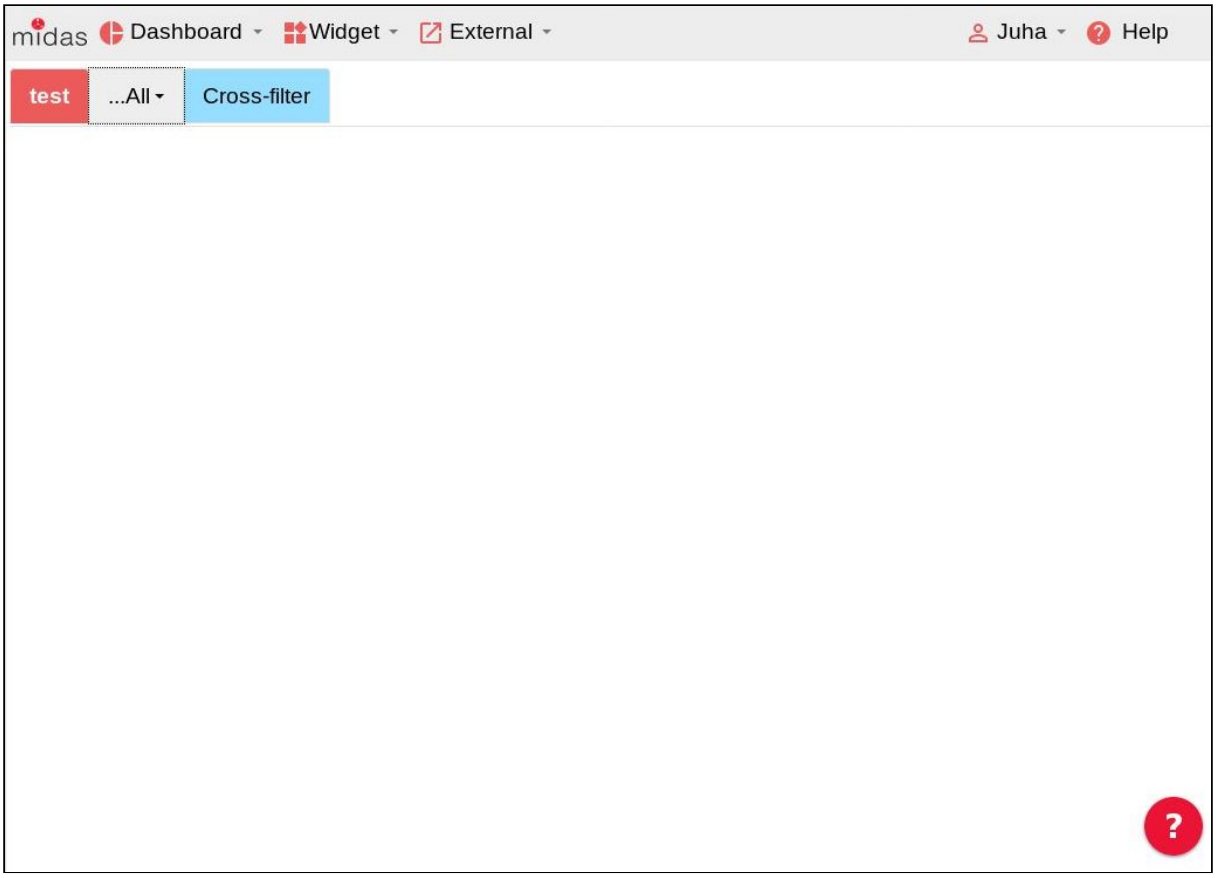


Figure 16: Dashboard Screen

On the Dashboard screen there are 4 Tabs across the top, these Tabs are:

- Dashboard
- Widget
- External links
- Help

Grant Agreement No: 727721

3.1.4 Dashboard Tab

This tab allows the user to select a defined dashboard from previously created work

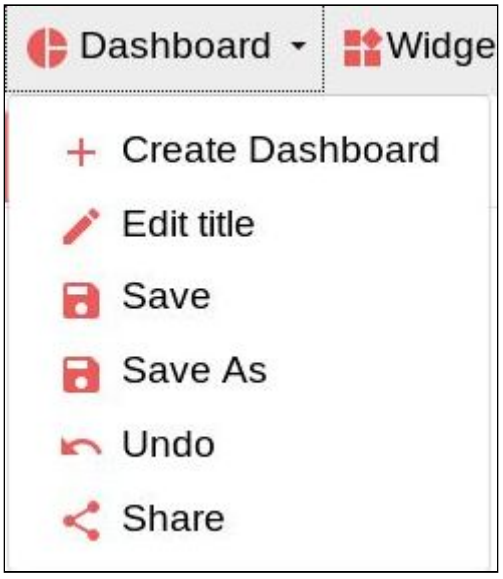


Figure 17: Dashboard Tab

3.1.5 Add Widget Tab

There are four widgets that drop down in this Tab. There are as follows: Analytics, MEDLINE search, News and Social media widget:

Figure 18: Add Widget

The functionality associated with these tabs will be explored later.

Grant Agreement No: 727721

3.1.6 External Tab

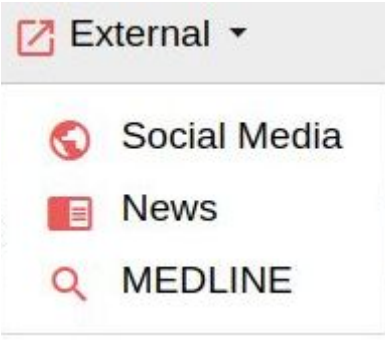


Figure 19: External Tab

Three options are available in the Drop down:

- Social media
- News
- Medline

All three options offer the user the ability to explore datasets described in the next section within external tools, some of which can then be included in the dashboard. Overall, the interface is mouse driven and is therefore point and click. Resizing, close, opening etc are modelled on standard Windows functionality.

3.2 Open and Social data

3.2.1 Social Media Campaigns

MIDAS provides a credible system for capturing the “voice of the public”, through an interface that manages a variety of social media API’s through a common model and presenting these on the MIDAS dashboard. This tool uses IBM’s Watson, for natural language and cognitive processing, as well as providing security through OAuth 2.0[8]. This functionality is mediated through a Chat bot interface.

The system requires the policy team to drive the policy question cycle and prepare the Chatbot for interaction with the user group.

EVP – Creating a Campaign video

[8] https://oauth.net/2/

Grant Agreement No: 727721

3.2.1.1 Stage 1

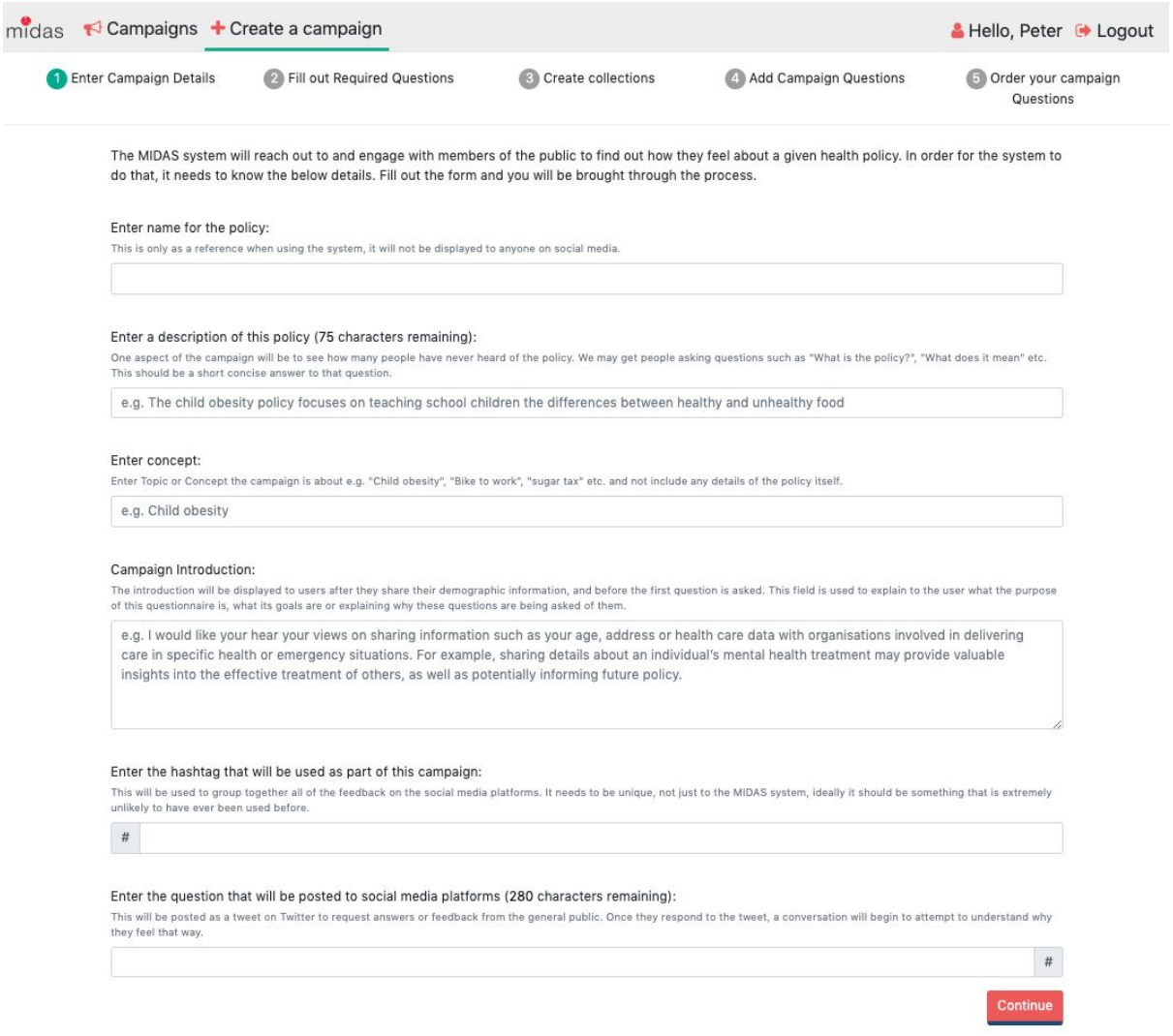


Figure 20: Campaign Creation

The user names and describes the campaign that needs to be set up, this forms the information about the campaign. that forms the basis of the work.

Grant Agreement No: 727721

3.2.1.2 Stage 2

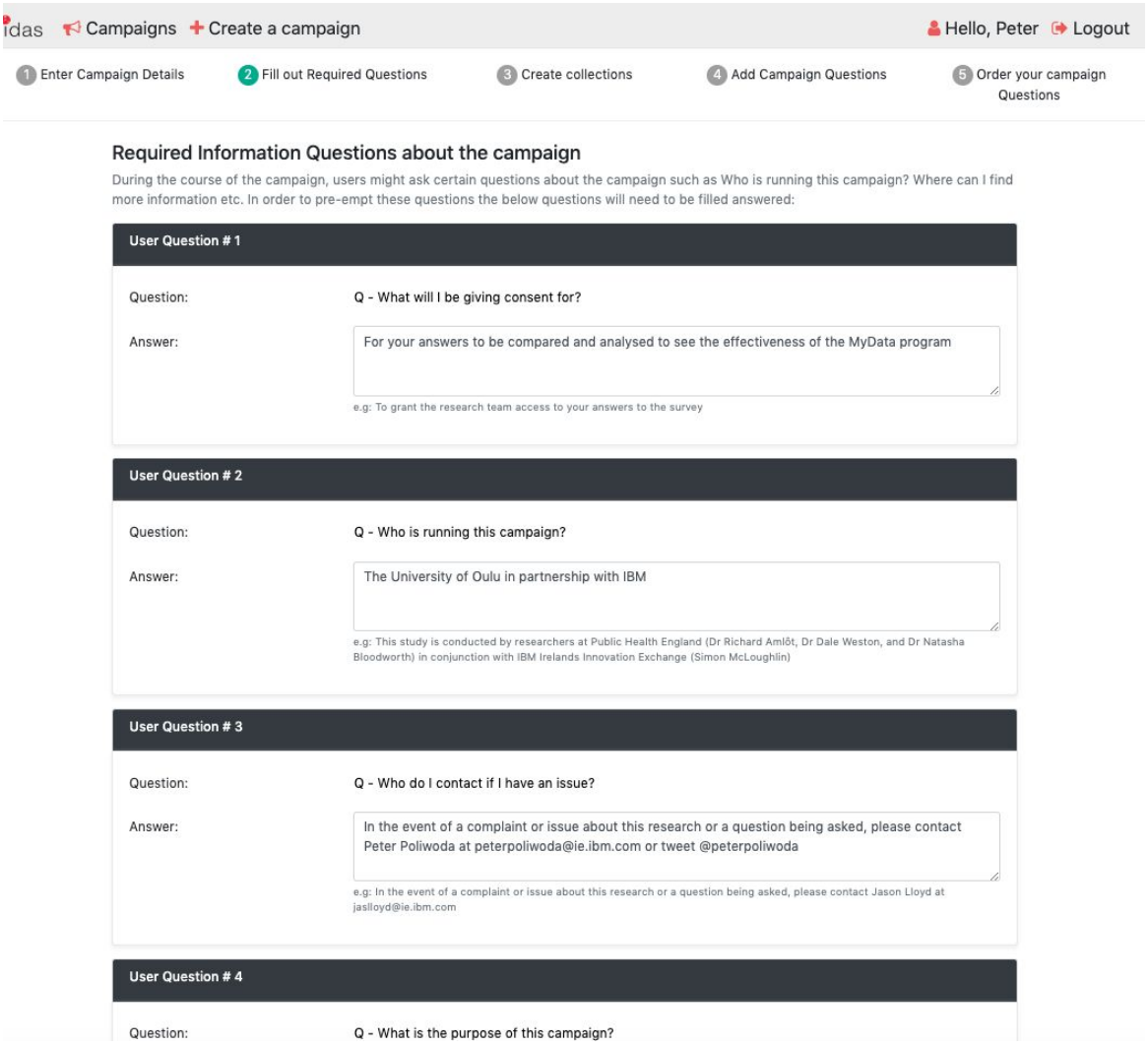


Figure 21: Social dashboard campaign creation step 2 - Providing answers to default user questions around consent

The user creating the campaign defines what the potential participant is consenting for within the context of the campaign.

Grant Agreement No: 727721

3.2.1.3 Stage 3

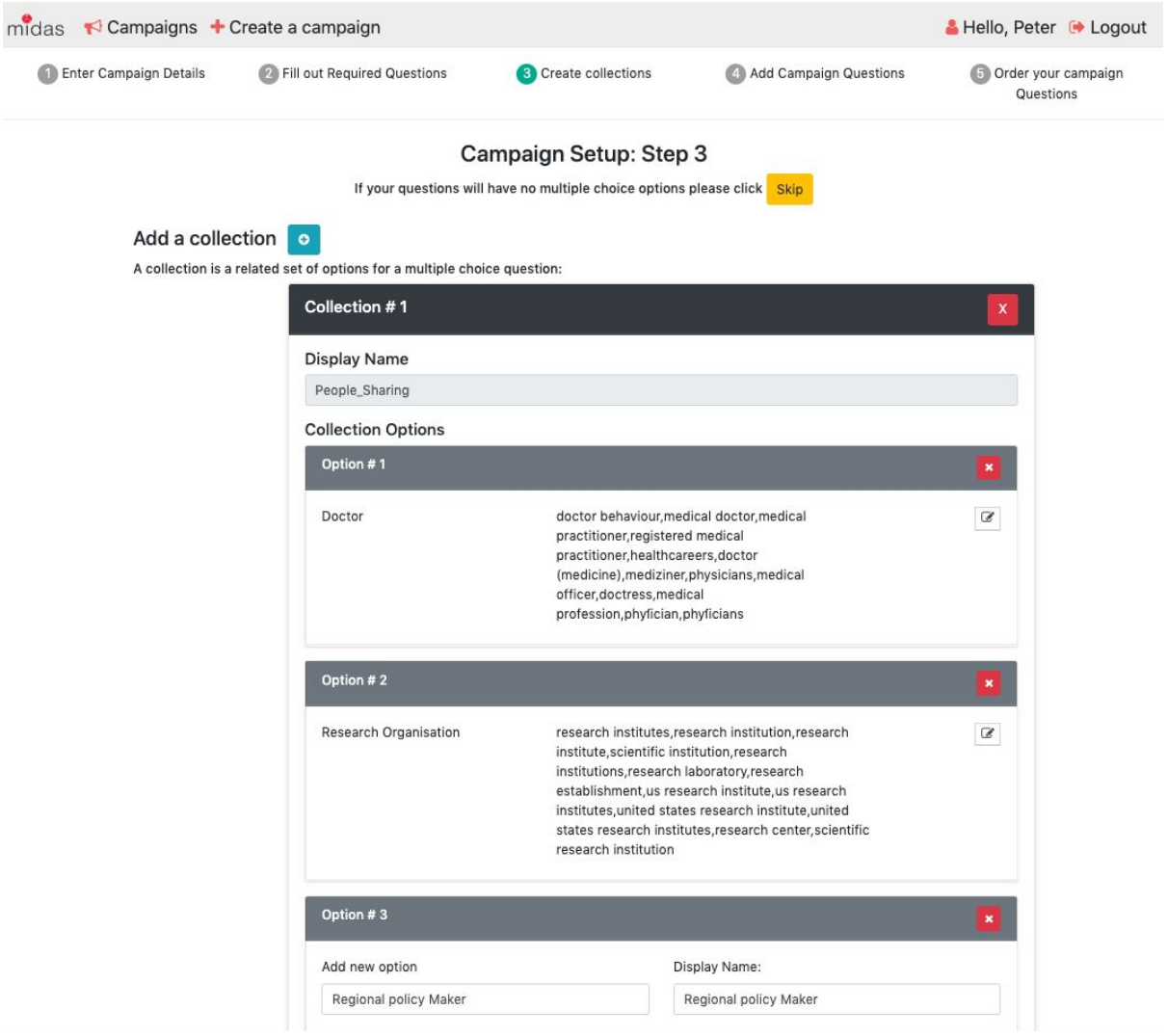


Figure 22: Social media dashboard campaign creation step 3 - Creating a multiple-choice list

The user creates a multiple choice list to define how potential answers can be used to further drive the narrative and response with the participant.

Grant Agreement No: 727721

3.2.1.4 Stage 4

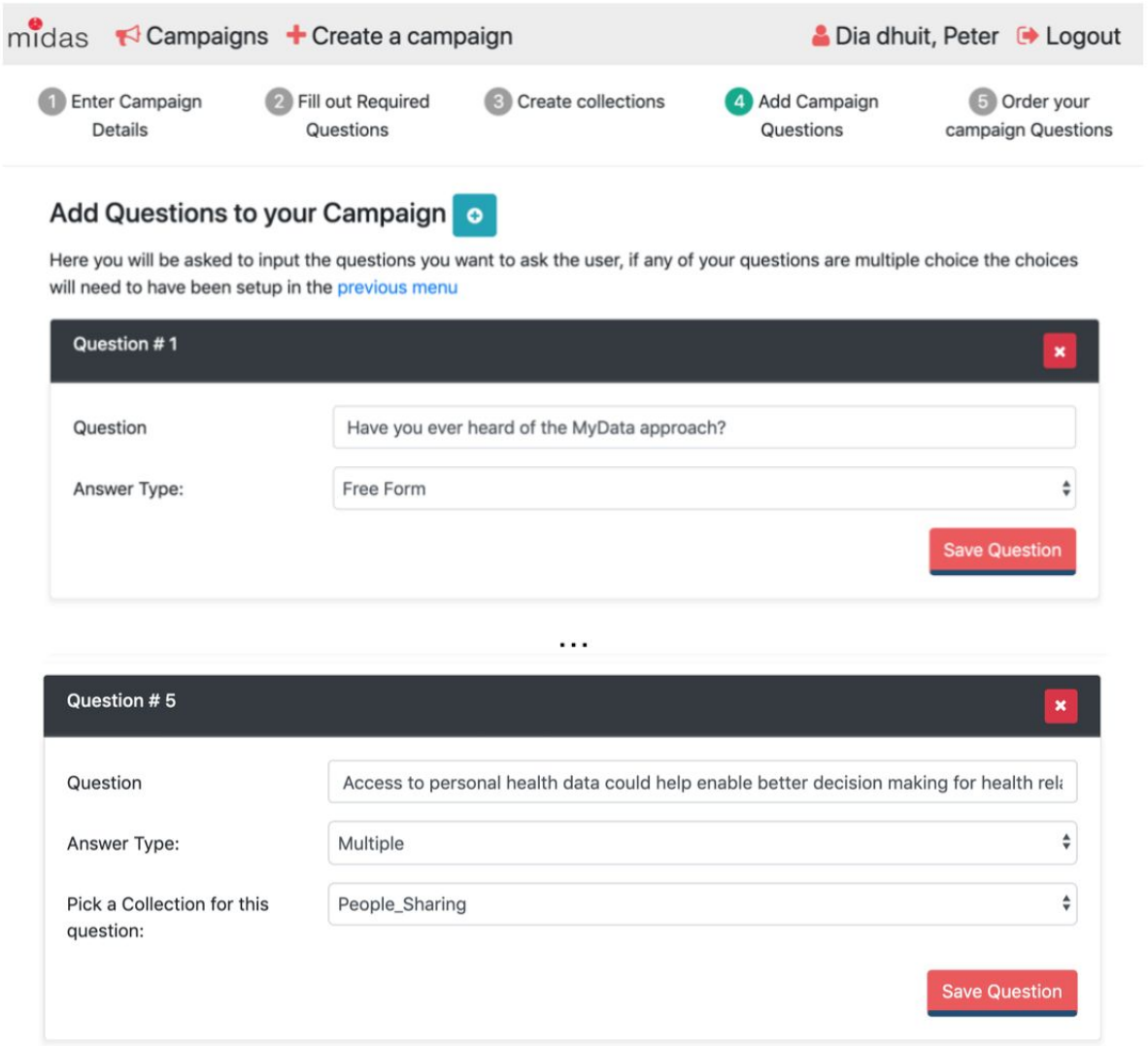


Figure 23: Adding questions

The questions are then entered into workspace. Once the questions are set the campaign is active, and interaction with users can be tracked, collated and analysed.

Grant Agreement No: 727721

3.2.1.5 Campaign Creation

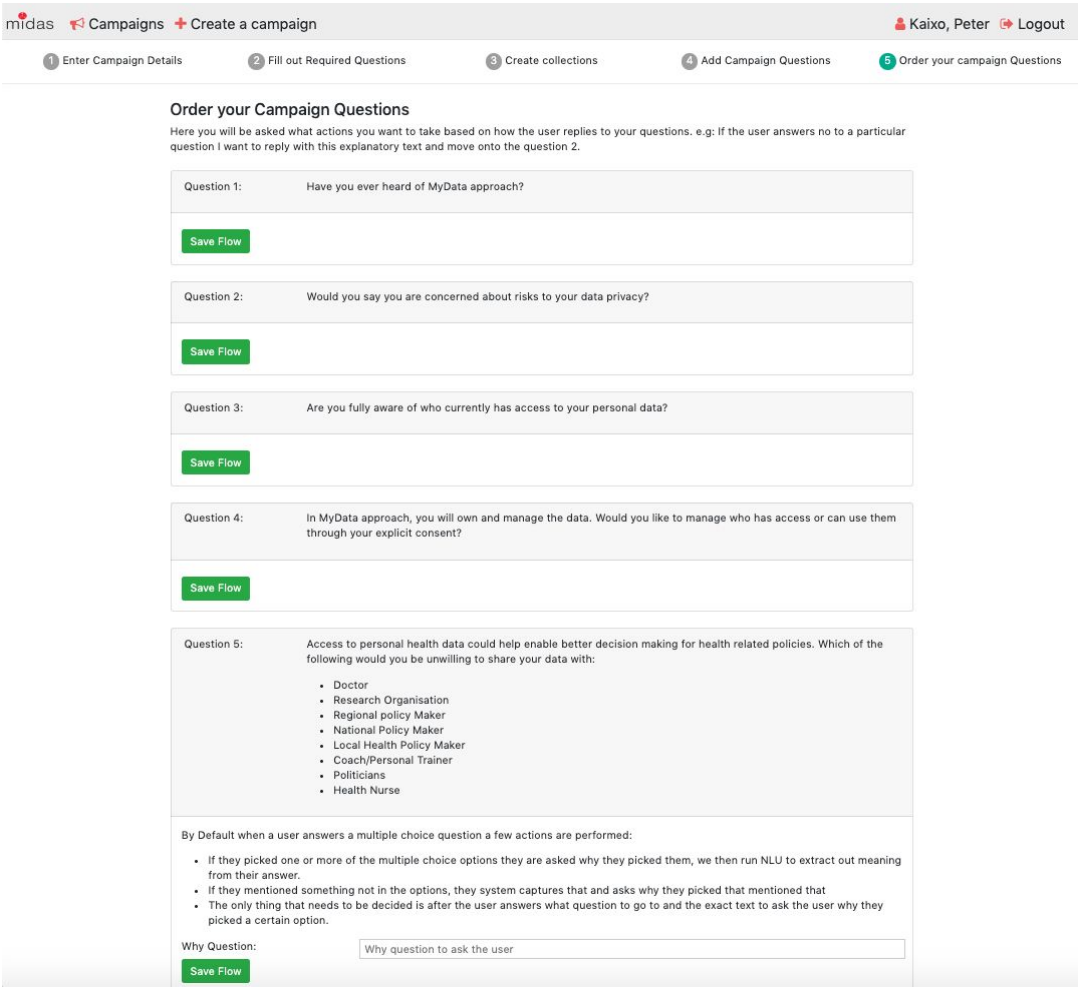


Figure 24: Campaign Creation

3.2.2 Complex visualisation of scientific knowledge

MIDAS has two methods that can be utilised by users for interaction:

- the custom widget
- the dedicated dashboard.

The user using the custom widget can construct a public health panel which will rest in the dashboard. When the user needs to explore further, the dedicated MEDLINE dashboard can be accessed and used in parallel with the well-established PubMed search engine.

Grant Agreement No: 727721

3.2.2.1 MEDLINE custom widget

In this widget the user can explore the MEDLINE database, using its own search keywords.

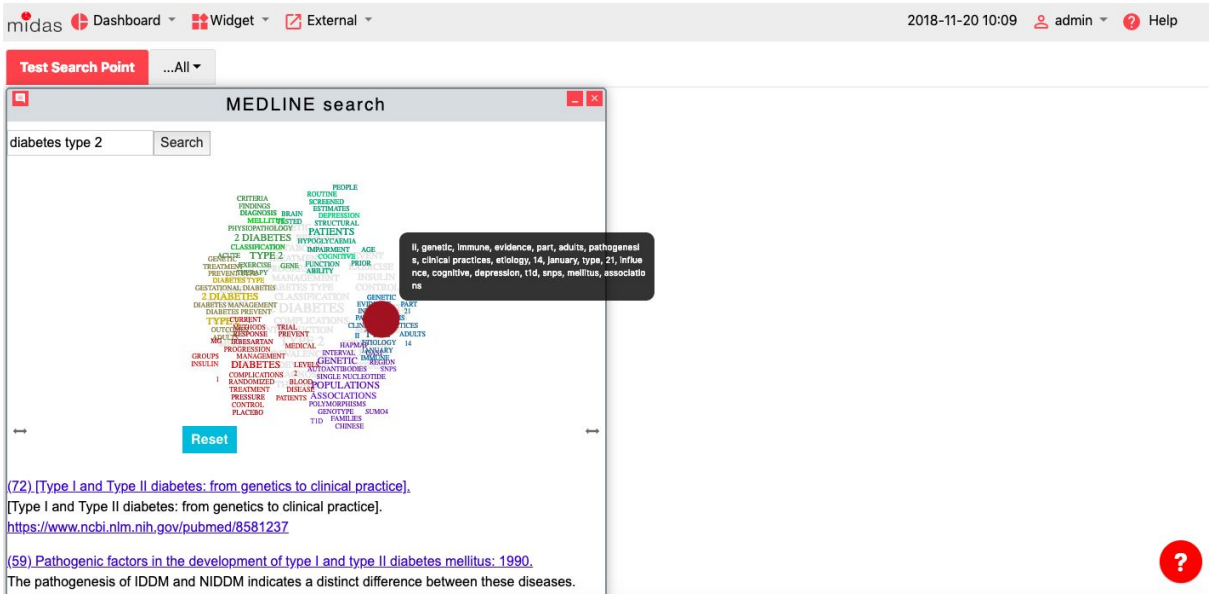


Figure 25: Cluster search model on the MIDAS dashboard

Functionality is achieved by:

1. Choosing the MEDLINE Widget from the Widget drop down menu
2. Typing the search keywords in the displayed search box
3. Moving the red pointer over the keywords in the word cloud that are most related to the particular search.
4. Clicking on the title of the articles that are of interest, which will then redirect the user to the appropriate article page in PubMed

3.2.2.2 MEDLINE exploratory dashboard (with public instance)

In the exploratory MEDLINE dedicated dashboard, the user can directly visualise key attributes of the MEDLINE data that reflects the user's interests. The user can create new visualisation modules that present the search outcomes, querying the data directly. The available dashboard feeds on the dataset through the elasticSearch index. It is composed of several interactive visualisation modules that utilise the mouse hover interaction and provides information through mouse-over messages on several key aspects of the data, based on particular queries of interest (e.g. a pie chart representing the "public health" citations that refer to "childhood obesity" during

Grant Agreement No: 727721

a selected period of time; or a bar chart showing different concepts included in the articles related to “mental health” in Finnish scientific journals).

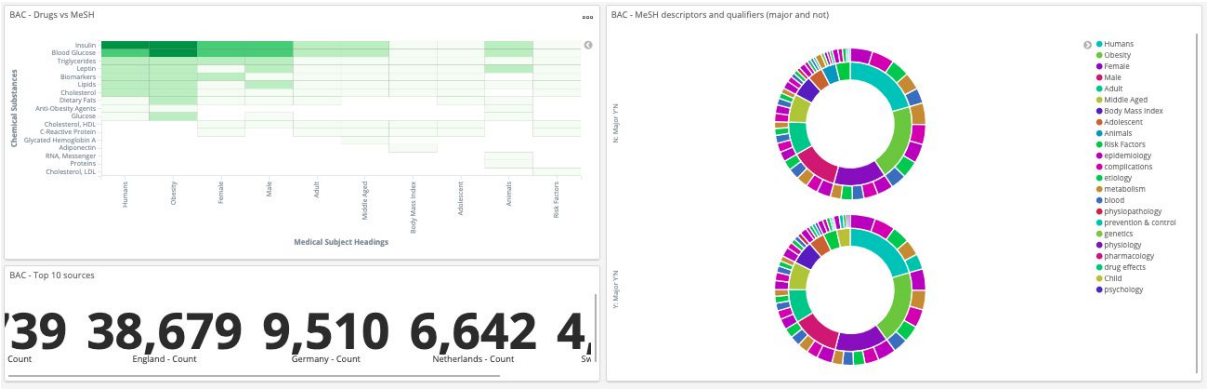


Figure 26: Dashboard of visualisation modules on articles in MEDLINE discussing childhood obesity (named “Paediatric Obesity”, term introduced only in 2014, with complementation on “Obesity” for earlier queries) to support the Basque use-case at MIDAS.

This dedicated MEDLINE dashboard serves the less technical user to explore the available data (over a subset of the data generated by a topic of interest). Other options are available that permit more control of the data by the data scientists at a more detailed level.

These include:

I. the management dashboard, where the technical user can perform the appropriate subsampling based on the topics of interest as well as the optional advanced options over the available data features;
II. the visual modules creator, which permit the less technical user to easily create new interactive visualisation modules; and
III. the live dashboard, that can be set up through iframe as a live window in the decision-maker’s workflow, enabling the monitoring of the status of the KPIs represented at each visualisation module.

This dashboard comprises:

I. a series of dashboards that can be used by decision-makers as monitoring dashboards each of which representing one study/topic (e.g., childhood obesity)
II. a set of visualisation modules that can be used to construct the dashboards in (I), each of which represents a question being monitored, with access to the tool to create new modules in predefined formats

Grant Agreement No: 727721

III. a dashboard allowing visualisation of the main aspects of the raw data, and a query box that can be used to directly query over the dataset

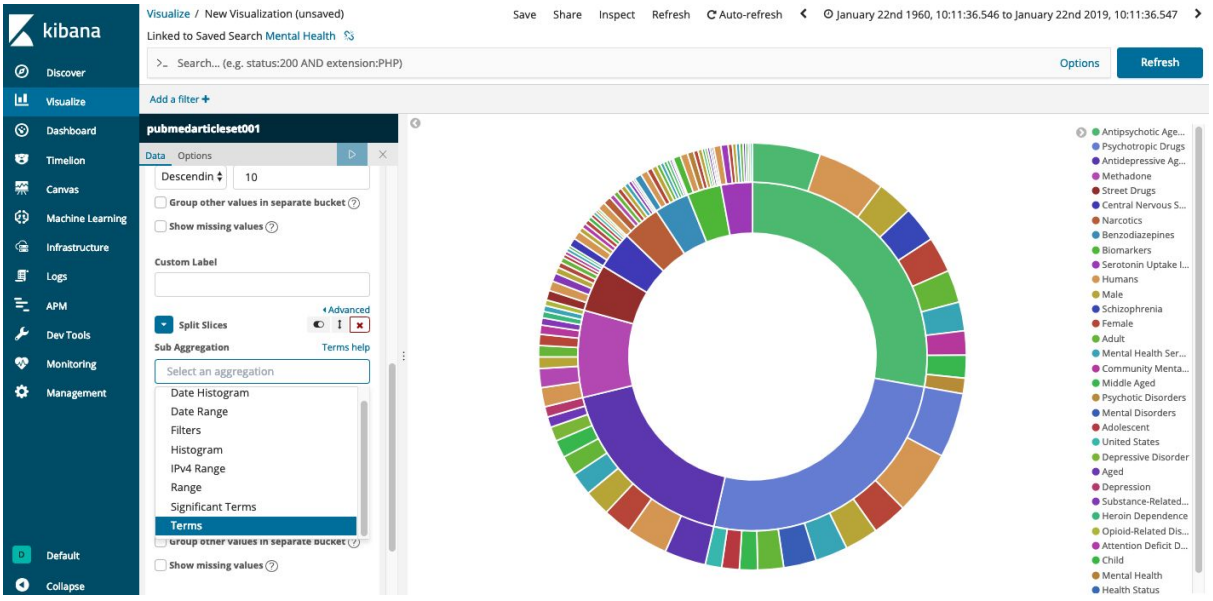


Figure 27: Construction of a visualization module based on a query related to mental health within the MEDLINE dataset

To use this dedicated dashboard, the user must:

1. Select MEDLINE from the External Tab in the drop-down menu
2. The user will then be presented with the monitoring dashboard with options on a side bar displayed on the left of the screen
3. The user explores the available visualisation with mouse over
4. The user can then edit the dashboard adding other available visualisations by choosing “edit” in the top bar
5. The user can also share the live dashboard over iframe (to integrate, e.g., a website or a monitoring tool) by choosing “share” in the top bar
6. To edit an existing visualisation module the user must click on Visualise on the left side bar and then the visualisation name
7. To build a new visualisation module the user must click on Visualise on the left sidebar and then the red plus button
8. To directly query the dataset, the user chooses Discover on the left side bar and then uses the search box (the accepted query language is Lucene)

Grant Agreement No: 727721

3.2.3 News media monitoring

The news custom widget takes place in the MIDAS platform side to side with other widgets like SearchPoint or the heatmap. This custom widget is available through the MIDAS platform to monitor topics of interest and in line with the public health study of the overall dashboard. In this widget the user can explore the worldwide news dataset according to its own search selected on the topic pages of the dedicated news dashboard.

3.2.3.1 News widget

In this widget the user can explore the worldwide news dataset related to the search selected on the topic pages of the dedicated news dashboard.

The widget comprises:

I. a word cloud that represents the main topics of the listed news
II. a list of news titles and first lines that are linked to the original news source
III. search choices based on the “Media Monitoring” option of the dedicated news dashboard

Grant Agreement No: 727721

Figure 28 The news widget of the MIDAS platform, showing the dropdown menu for the choice of the use-case pilot to be sourced from.

To use this widget the user must:

1. Chose the MEDLINE Widget from the dropdown menu
2. Click Select
3. Scroll over the news and click on the news that the user is interested in to be taken directly to the news source location
4. To tune the choices of the news to be displayed at the custom widget, the user must enter the dedicated news dashboard by choosing “News” in the dropdown “External” menu
5. Then, the user must choose “Media Monitoring” in the main menu of the dedicated news dashboard, and there (s)he must choose the use-case/topic of interest.
6. Once the topic is chosen, the user is presented with a menu of sliders, dropdown menus and search boxes to fine tune the choices filtering the news stream. The pink button below permits the user to see the results of the choices made.

7. Finally, the user can change the main slider to obtain the new source more/less close to the choices made, which in the case of small subsets (e.g. a rare disease in a small location) can be important because of the low frequency of news over that topics

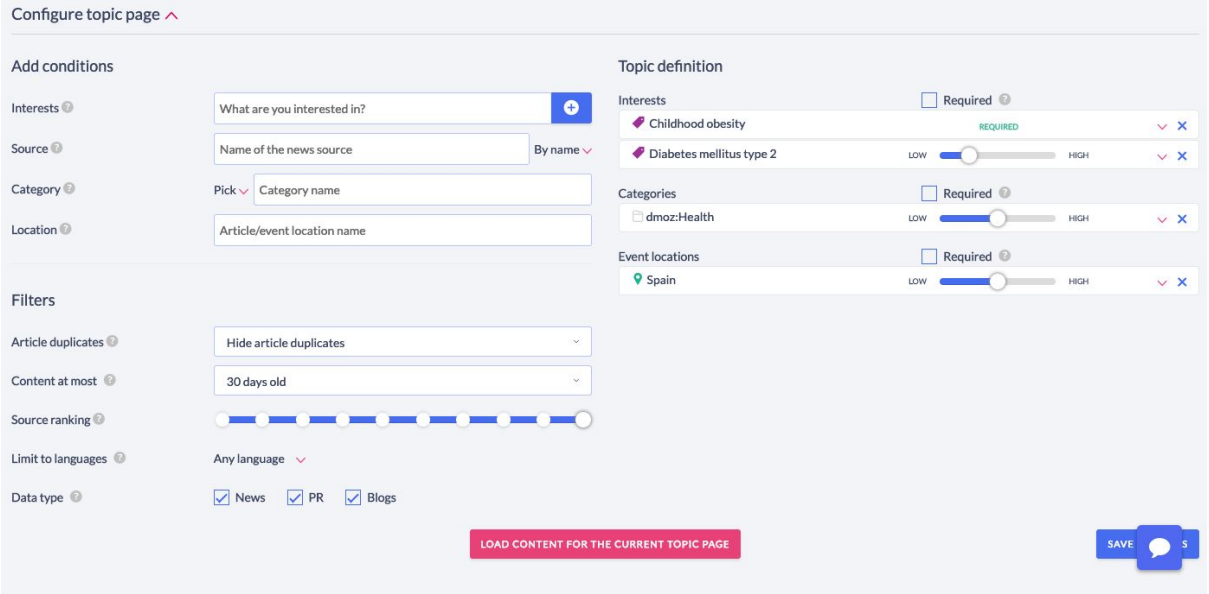


Figure 29: The advanced filters in the News Monitoring board for Event Registry, underpins the news engine, allowing the user to update the topics of interest in the news feed.

3.2.3.2 News exploratory dashboard

This dashboard allows the user to explore the worldwide news in 60+ languages using a variety of filters to target topics of interest. It provides the user with resources to better visualise and interact with search results, to enable deep exploration of news, and includes integration with the MIDAS MeSH classifier to refine the search of news through MeSH Heading classes, mirroring the techniques used in PubMed when searching scientific articles.

To use this dedicated dashboard the user must:

1. Chose the “News” option in the “External” dropdown menu
2. The user is then presented with a dashboard which includes a search box and several dropdown menus (e.g. locations, languages, categories, etc) to perform a search of news articles.
3. The user can choose from a variety of languages, although the search item used can be maintained. Search items can be terms under “” (that are

Grant Agreement No: 727721

matched exact strings), Wikipedia concepts (that are then multilingual by nature), or keywords, etc.

4. The user can also choose a window of time, to limit parameters for the displayed items
5. The location is either provided in the news article or, if there is no location mentioned, the system adopts the location of the news source
6. When choosing “Top Concepts” in the menu on the left of the screen, the themes associated with the articles are displayed.
7. The user can also choose the option “Tag Cloud” to display a word cloud that represents the main topics in the choice of articles (this concept is reproduced in the news widget)
8. The option “Timeline” allows the user to display the number of news articles on the topic of interest, showing the evolution of that topic in the media over time
9. The user can also explore several aspects related with the news search in the option “Categories” (similarly to that in the exploration over the MEDLINE widget)
10. The user can also explore sentiment reflected in the media within a news topic using “Sentiment” (functionality is limited in that the search should be simple in scale and not involve a significant number of filters)
11. To change from the global exploration view (i.e., “Media Intelligence”) to the topic pages view (i.e., “Media Monitoring”) the user must select the appropriate option in the top left displayed menu
12. After selecting the monitoring topic the user can use filters to refine the search and examine a real-time news stream
13. The user can also create a new monitoring topic, can label with a descriptor, change the icon, choose the type of privacy (public or private) and provide a general description.
14. To make it public the user must copy the below selected code and add it to a URL as follows: http://eventregistry.org/topic/<code> This will create a public page including the title and description of the topic, the stream of news, the word cloud and the most relevant entities

Grant Agreement No: 727721

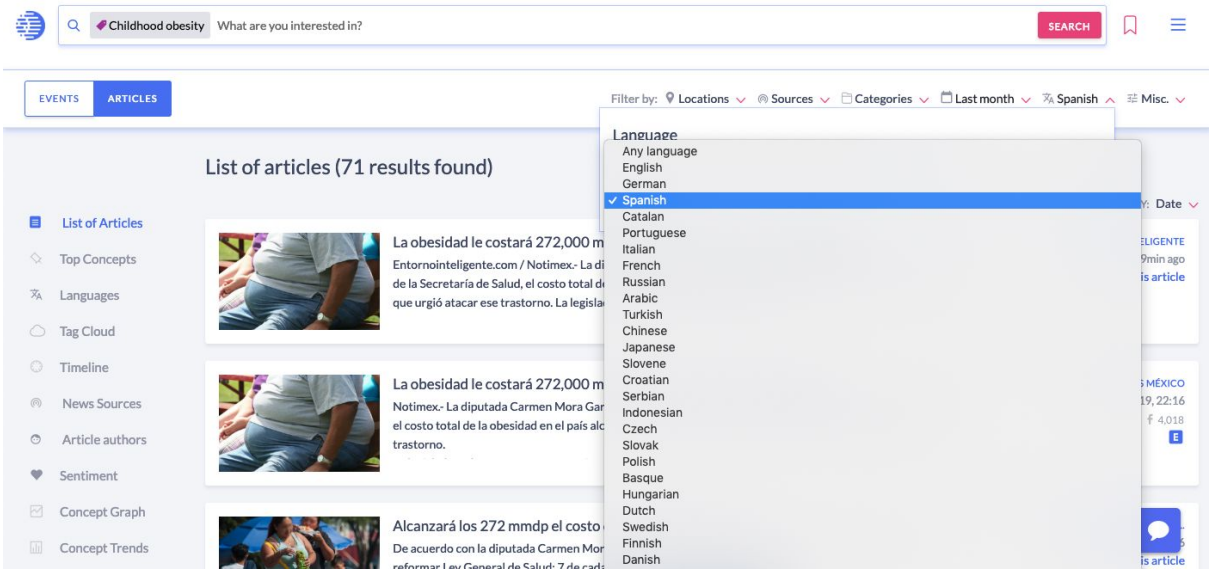


Figure 30: The search engine of the MIDAS news exploratory dashboard showing the list of possible languages after subjects have been selected.

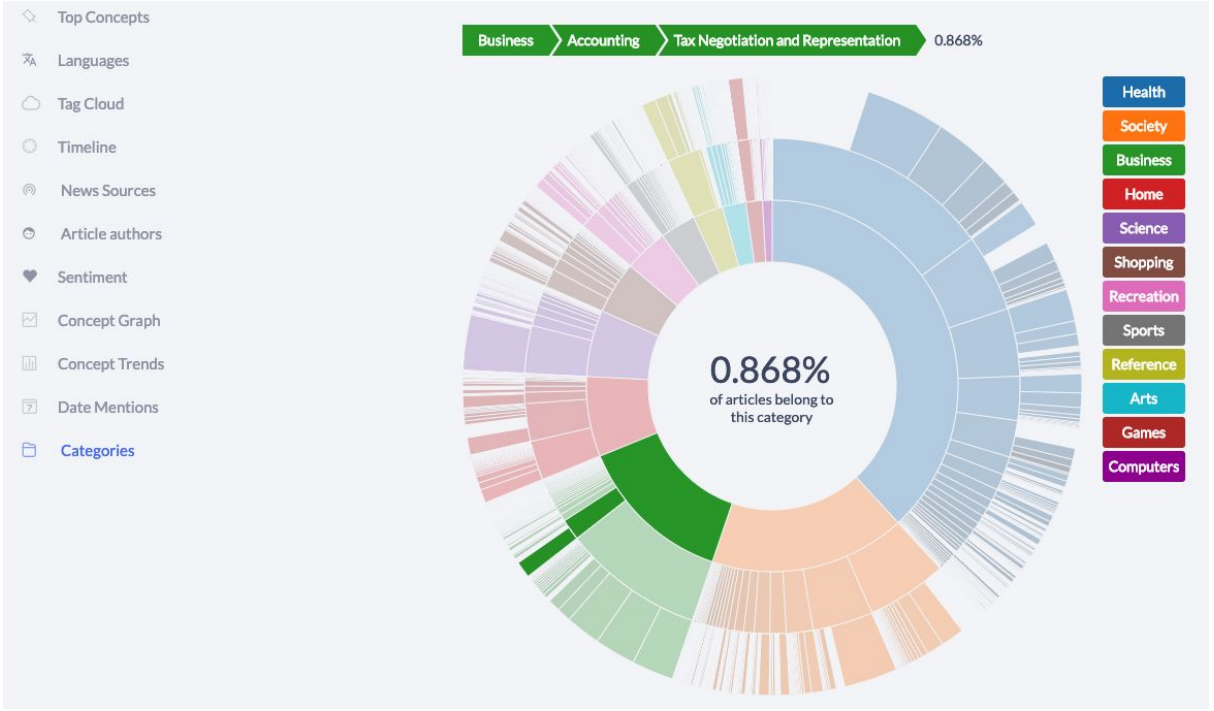


Figure 31: The visual representation of topics and subtopics that the collected and annotated that the news belong to.

Grant Agreement No: 727721

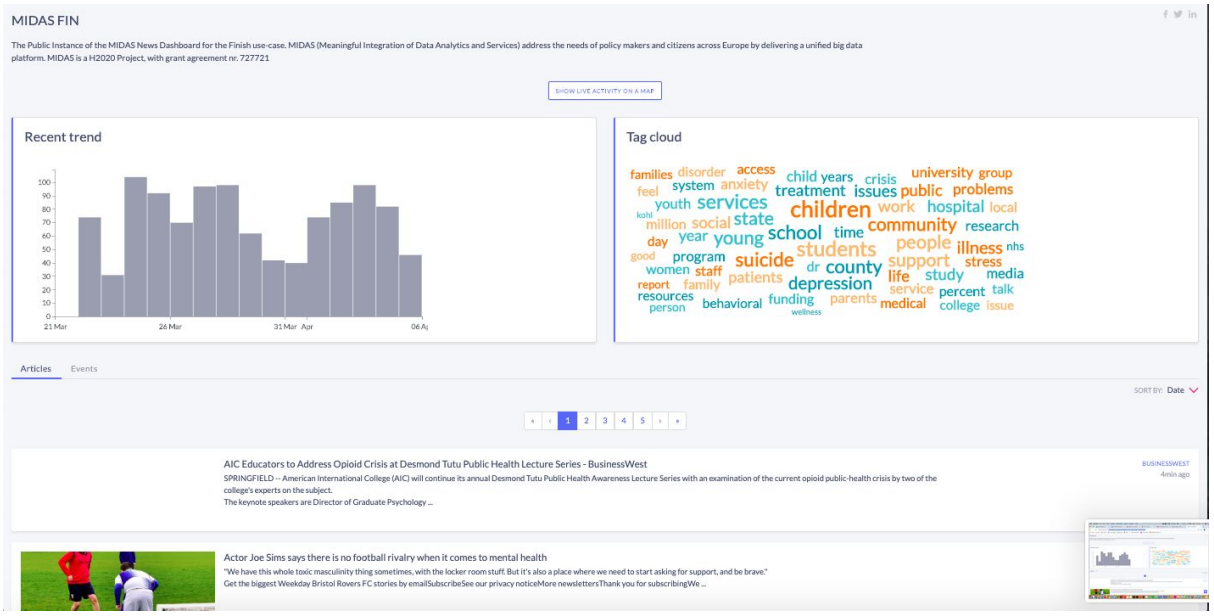


Figure 32: A public instance of the MIDAS news monitoring system embedded within iframe, showing a word cloud and a bar chart for a topic filtered and sourced for the news custom widget.

Grant Agreement No: 727721

4 Appendix 1. GYDRA

1. GYDRA - Data preparation Tool

Within a data-driven decision-making context, it is key to guarantee the quality of the data upon which decisions are made. The GYDRA tool has been developed to analyse the quality of the datasets and to prepare them for decision-making targeted analysis. GYDRA stands for “Get Your Data Ready for Analysis!”. Currently it supports datasets in CSV data format and loaded to HFDS.

1.1 Login

Access to the GYDRA tool starts with a login window. Currently, accounts are managed by GYDRA administrators for each site independently.

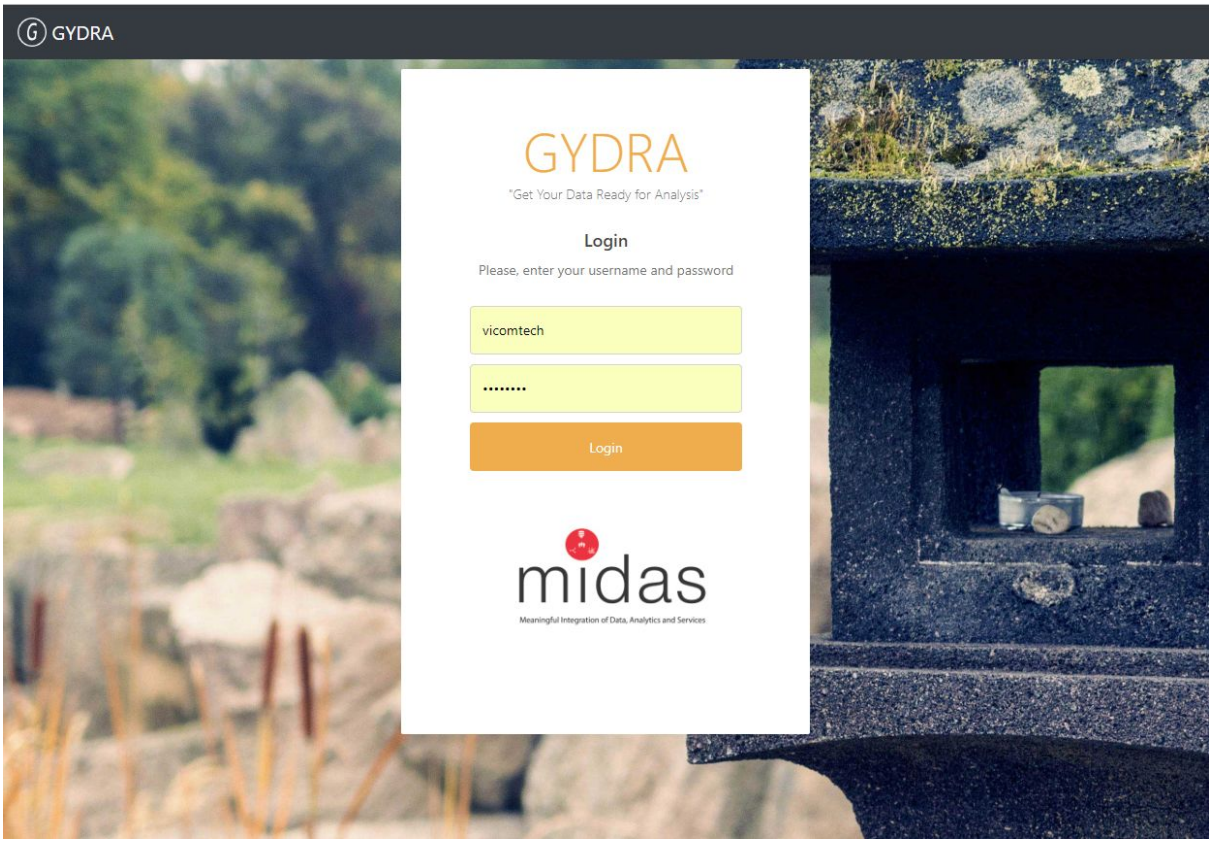


Figure 1: GYDRA tool login screen

1.2 Home

This screen appears once a user has successfully logged in. In this window, status of launched pre-processing datasets appears. If successfully preprocessed, it’s possible to start the exploratory data analysis and preparation of a dataset by clicking on the View button.

Grant Agreement No: 727721

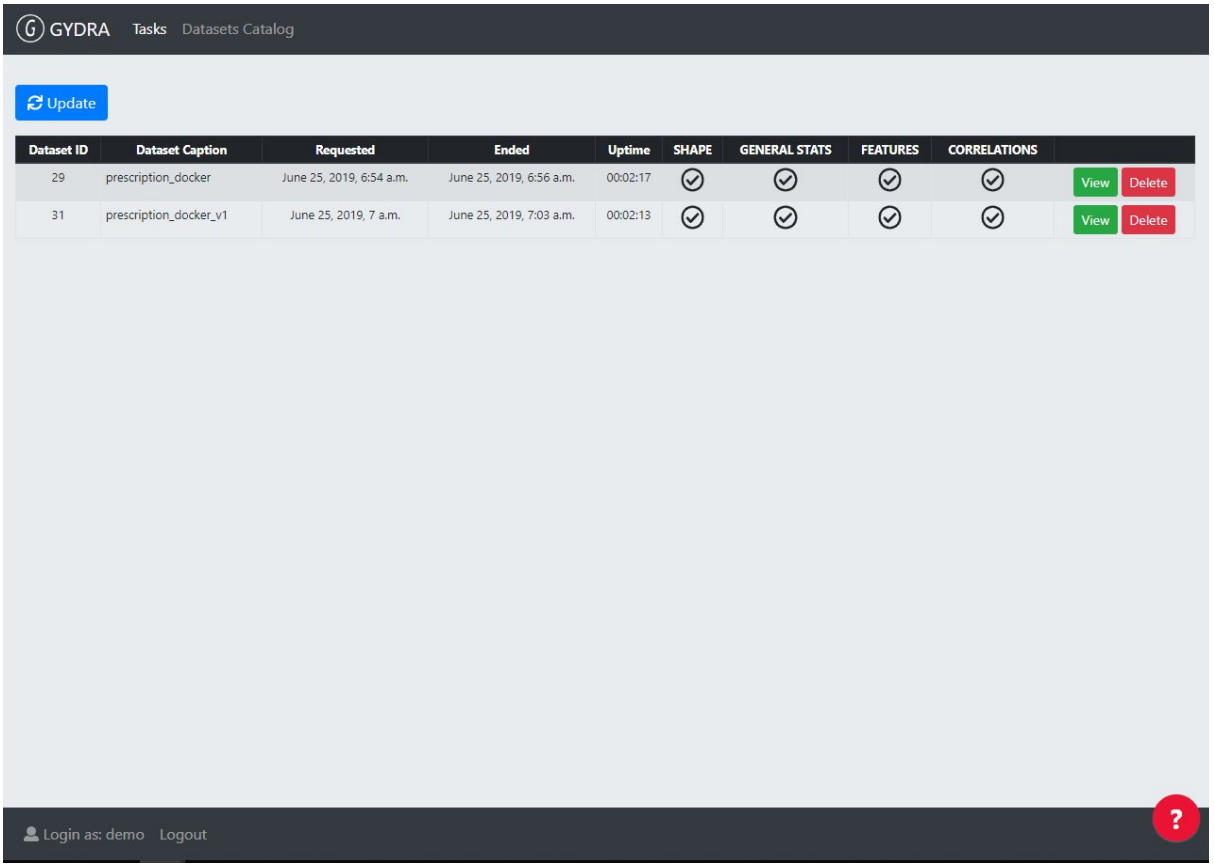


Figure 2: GYDRA tool home screen and preprocessing task status

Next, by clicking on “Dataset Catalog” in the navbar, registered datasets and their information are displayed. From this window the user can register a new dataset by clicking on the blue “Add dataset button” (or de-register an old one by clicking on the bin red icon button), launch the pre-processing of a dataset by clicking on the blue “Enqueue Preprocess” button, check the alignment of dataset metadata described by data owners (in Isaacus) with the metadata inferred by the GYDRA tool within the pre-processing of the dataset (by clicking on the blue eye icon button) or deploy the dataset to the MIDAS platform (by clicking on the black icon button).

Grant Agreement No: 727721

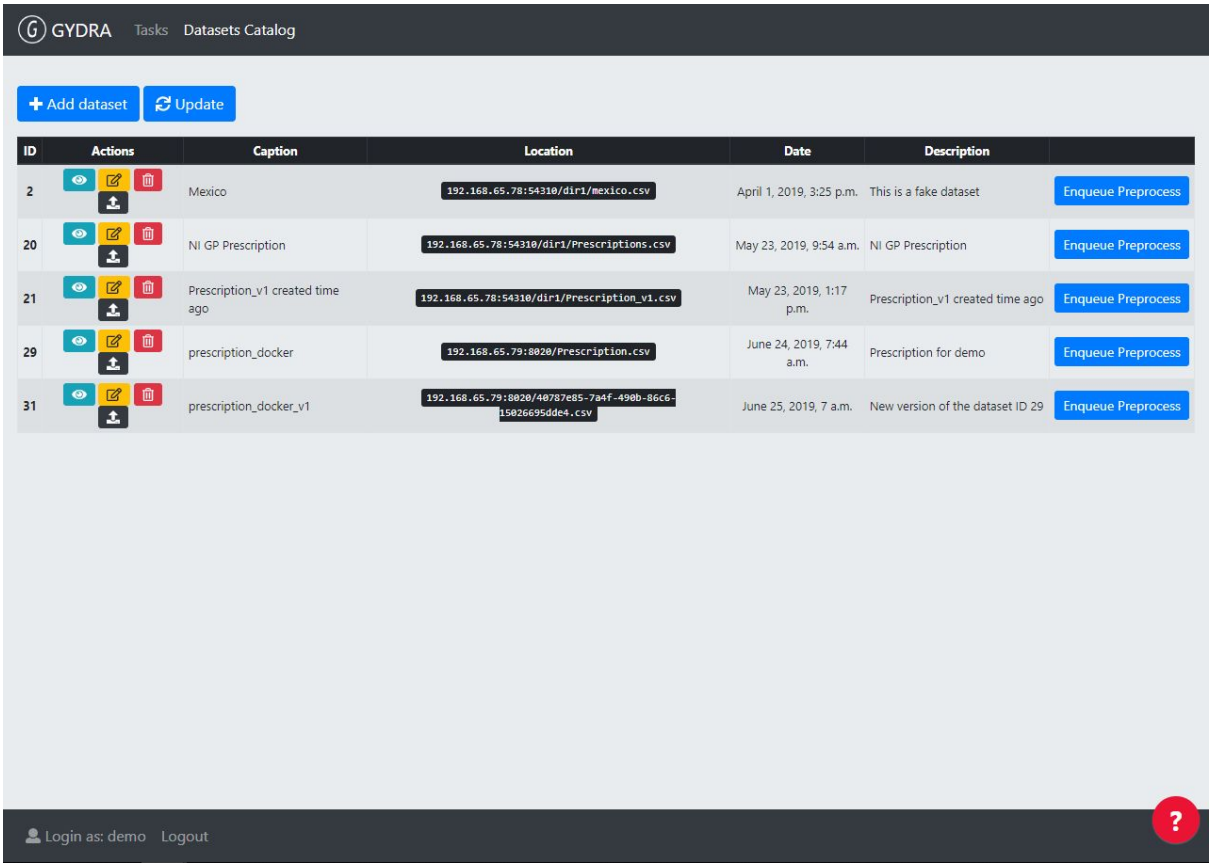


Figure 3: GYDRA tool dataset catalog GUI

GYDRA tool users can register new (Blue “Add dataset button”) or edit datasets (Yellow per dataset button) by providing the description (caption + description), HDFS configuration (ip + port + username + password) and file descriptors (location path, encoding and separator used). If later automatic deployment of the dataset to MIDAS Platform is planned it’s also important to set the corresponding ISAACUS metadata description, by configuring ISAACUS study and dataset in the provided interface (combo boxes automatically show available studies and dataset descriptions). ISAACUS metadata description steps have been detailed in Section 2.3 of deliverable D3.6.

Grant Agreement No: 727721

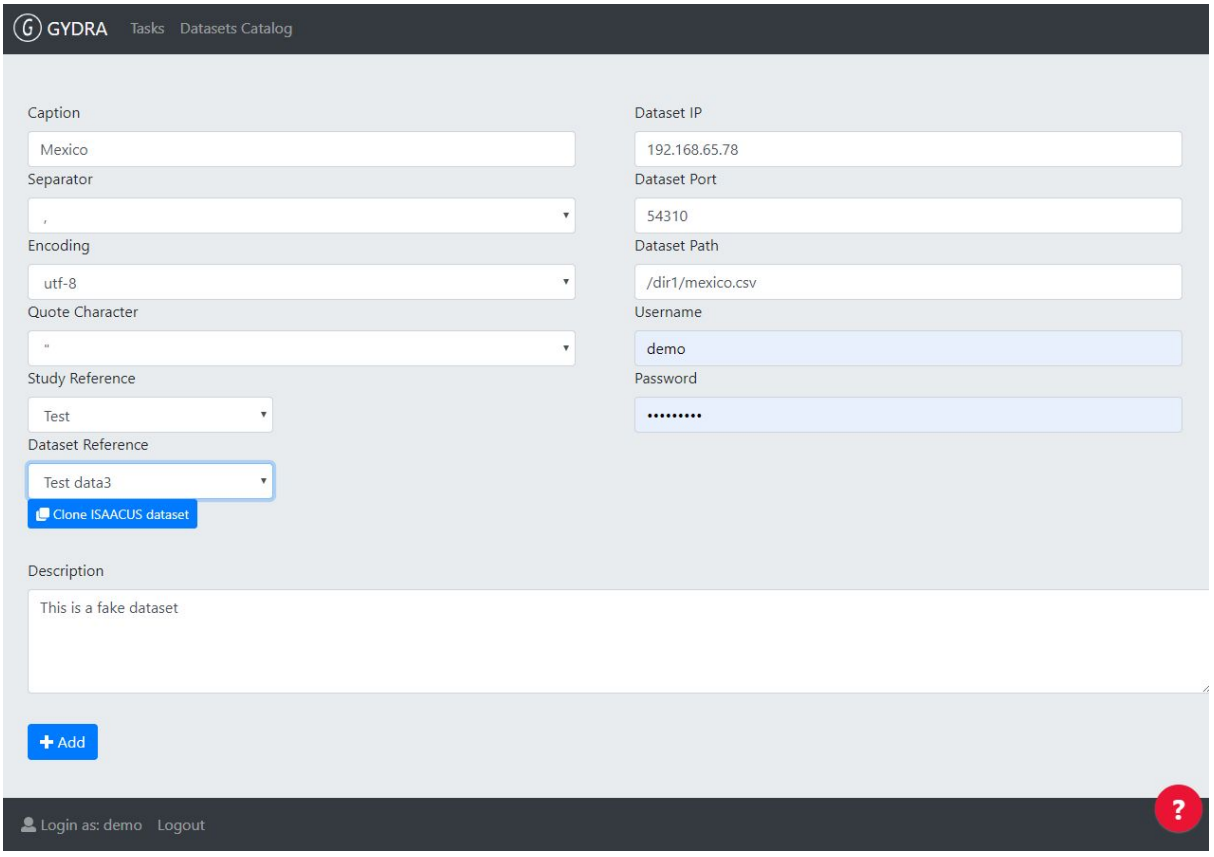


Figure 4: GYDRA tool home screen's add dataset option

1.3 General Description of the Data

A general overview is presented in this section, allowing users to have an initial understanding of the dataset status. The main depicted characteristics are total number of values, samples and features, inferred types of the features and summary of amount of missing data, highly correlated features and outliers.

Grant Agreement No: 727721

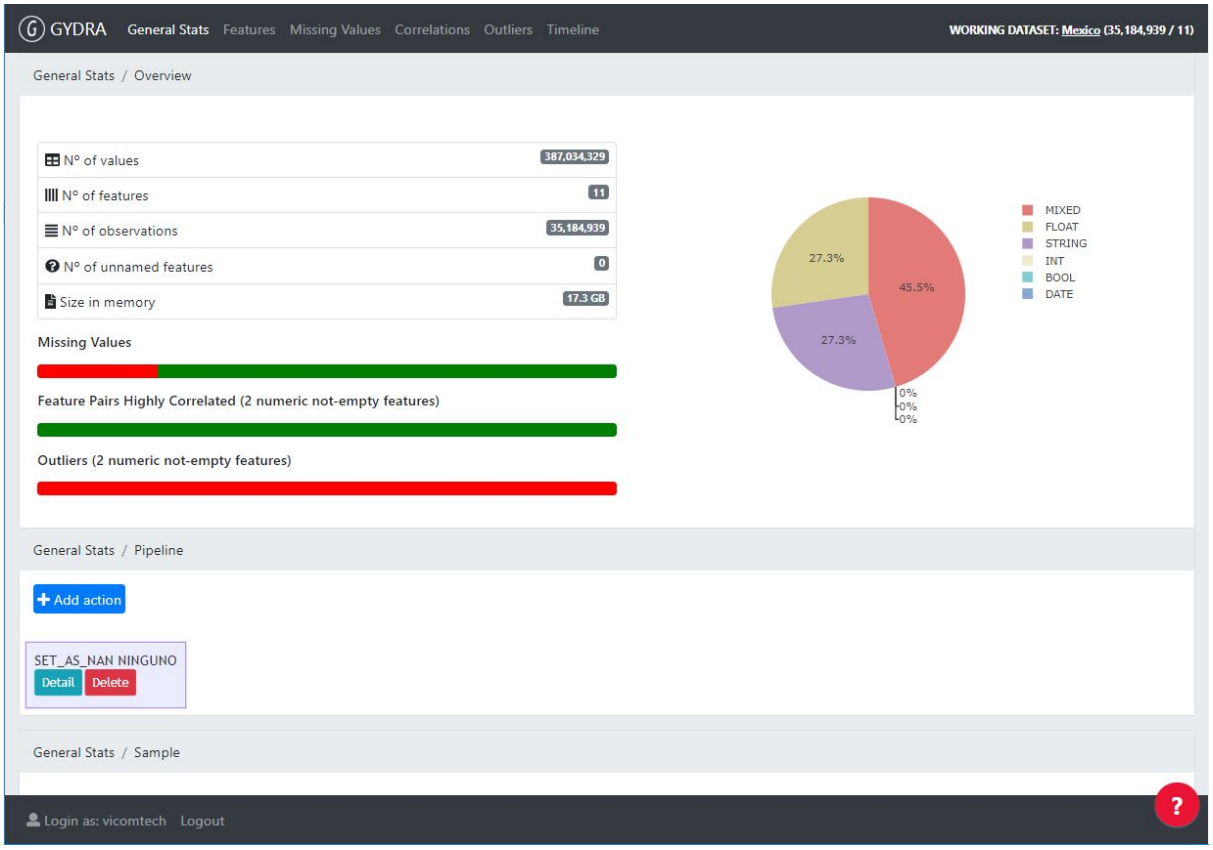
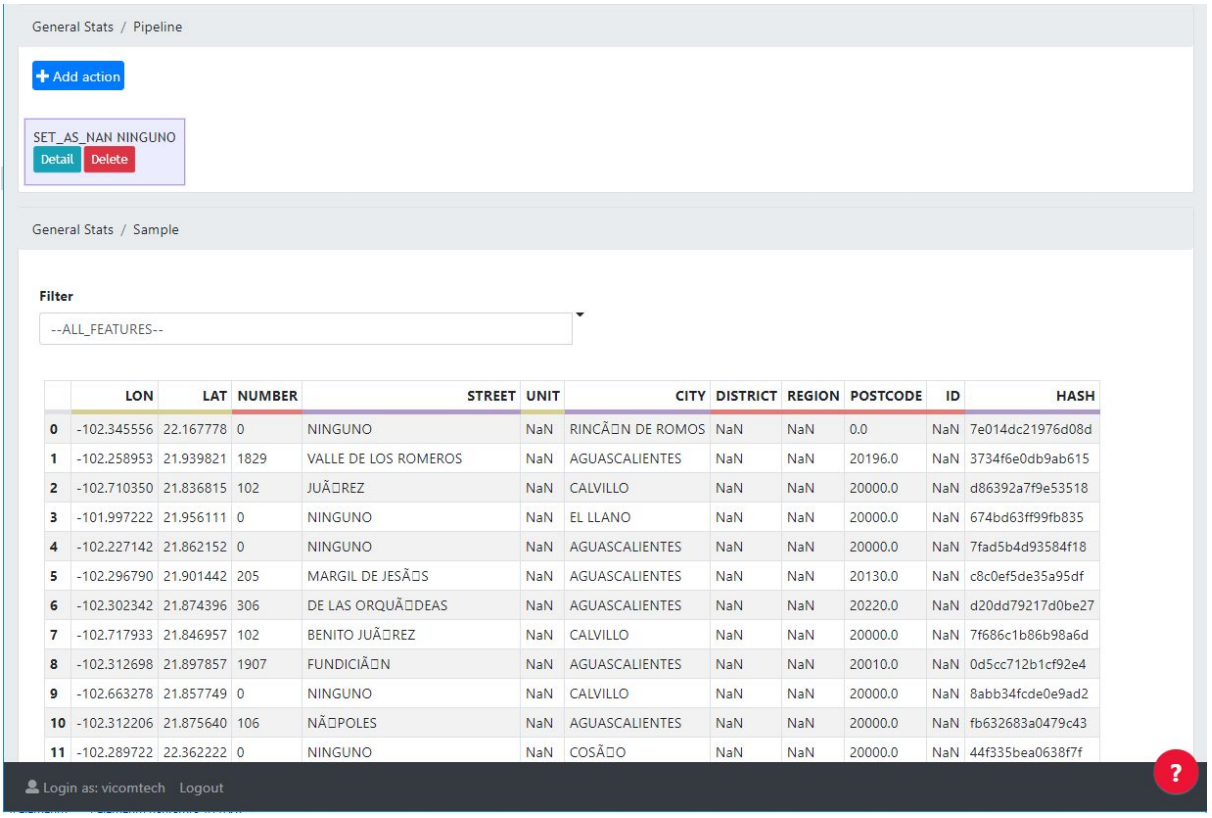


Figure 5: General data overview

Additionally, transformation pipeline and data example windows are shown below, to demonstrate how the user can understand the dataset and add dataset transformation actions as they are identified (see Figure below).

Grant Agreement No: 727721



	LON	LAT	NUMBER	STREET	UNIT	CITY	DISTRICT	REGION	POSTCODE	ID	HASH
0	-102.345556	22.167778	0	NINGUNO	NaN	RINCÃ□N DE ROMOS	NaN	NaN	0.0	NaN	7e014dc21976d08d
1	-102.258953	21.939821	1829	VALLE DE LOS ROMEROS	NaN	AGUASCALIENTES	NaN	NaN	20196.0	NaN	3734f6e0db9ab615
2	-102.710350	21.836815	102	JUÃ□REZ	NaN	CALVILLO	NaN	NaN	20000.0	NaN	d86392a7f9e53518
3	-101.997222	21.956111	0	NINGUNO	NaN	EL LLANO	NaN	NaN	20000.0	NaN	674bd63ff99fb835
4	-102.227142	21.862152	0	NINGUNO	NaN	AGUASCALIENTES	NaN	NaN	20000.0	NaN	7fad5b4d93584f18
5	-102.296790	21.901442	205	MARGIL DE JESÃ□S	NaN	AGUASCALIENTES	NaN	NaN	20130.0	NaN	c8c0ef5de35a95df
6	-102.302342	21.874396	306	DE LAS ORQUÃ□DEAS	NaN	AGUASCALIENTES	NaN	NaN	20220.0	NaN	d20dd79217d0be27
7	-102.717933	21.846957	102	BENITO JUÃ□REZ	NaN	CALVILLO	NaN	NaN	20000.0	NaN	7f686c1b86b98a6d
8	-102.312698	21.897857	1907	FUNDICIÃ□N	NaN	AGUASCALIENTES	NaN	NaN	20010.0	NaN	0d5cc712b1cf92e4
9	-102.663278	21.857749	0	NINGUNO	NaN	CALVILLO	NaN	NaN	20000.0	NaN	8abb34fcde0e9ad2
10	-102.312206	21.875640	106	NÃ□POLES	NaN	AGUASCALIENTES	NaN	NaN	20000.0	NaN	fb632683a0479c43
11	-102.289722	22.362222	0	NINGUNO	NaN	COSÃ□O	NaN	NaN	20000.0	NaN	44f335bea0638f7f

Figure 6: GYDRA tool common transformation pipeline and data example windows across application tabs

1.4 Features

The next section, Features, presents a report for every variable in the dataset. It automatically presents its name, type and the number of distinct values (and distribution).

Grant Agreement No: 727721

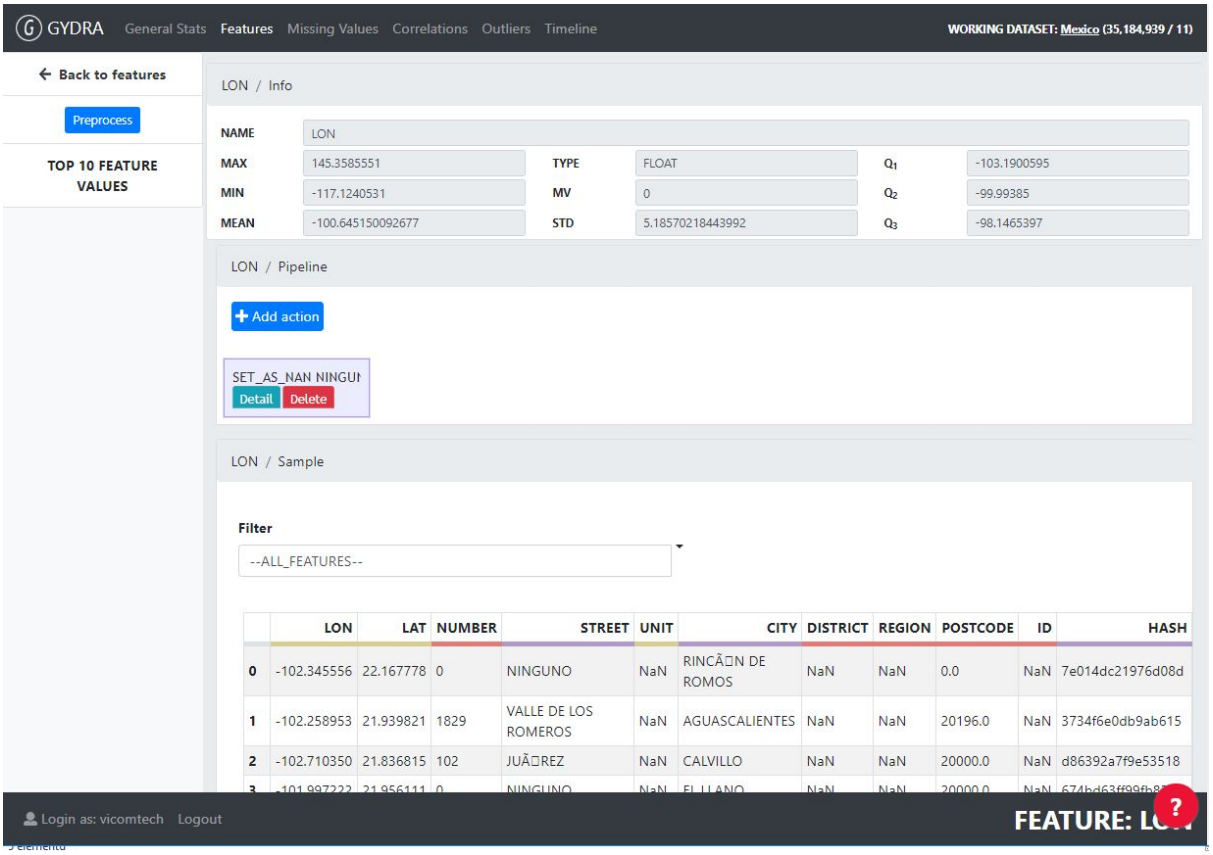


Figure 7: GYDRA tool per feature description analysis tab

1.5 Missing Values

The presence of missing data is quite common within datasets and they usually have a significant effect on the conclusions that can be drawn from the data. This section provides users with a set of tools and visualizations to deal with missing values.

An overview of the amount of missing data per values, samples and features is provided in this section. Additionally, bar chart is drawn to help users in the identification of features containing a meaningful percentage of values missing.

Grant Agreement No: 727721

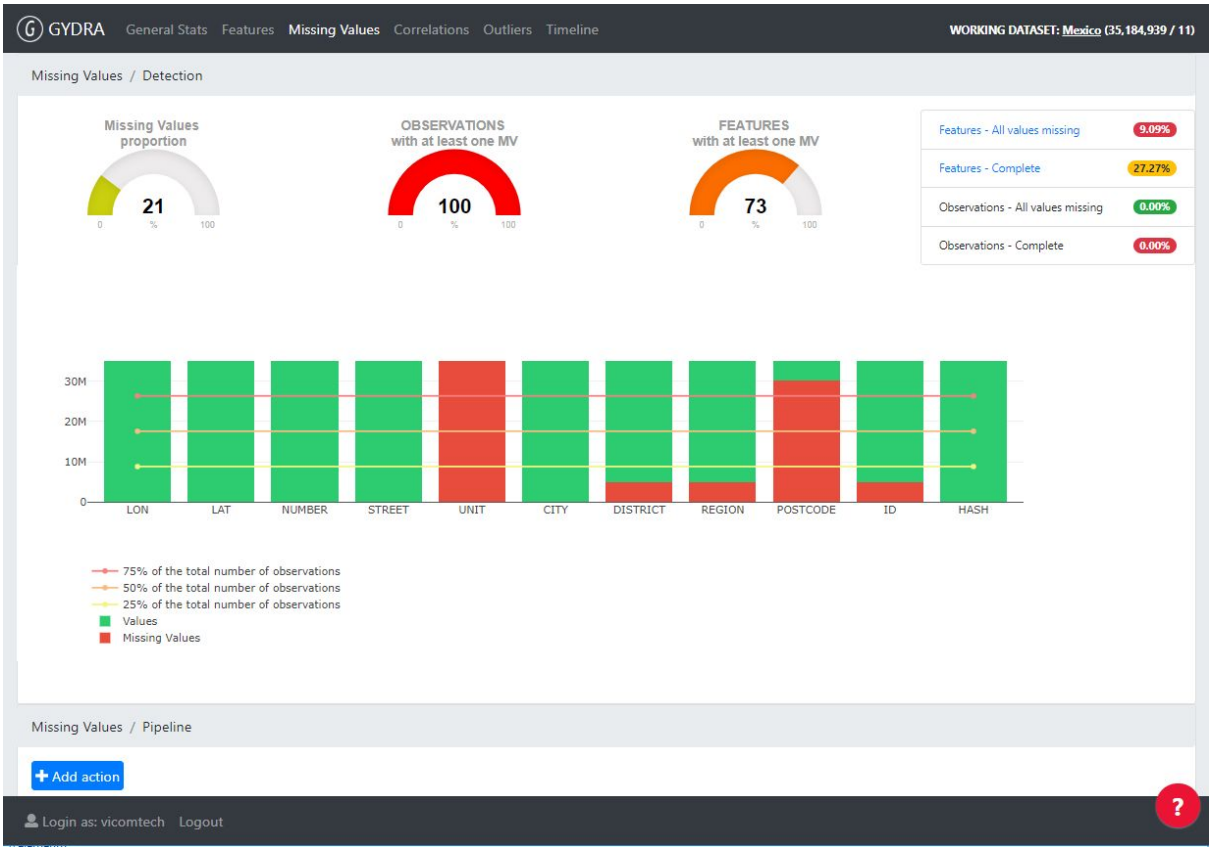


Figure 8: GYDRA tool Missing Values analysis tab

In case there are features with all values missing, a clickable link is shown and a modal window is opened by clicking on it.

Grant Agreement No: 727721

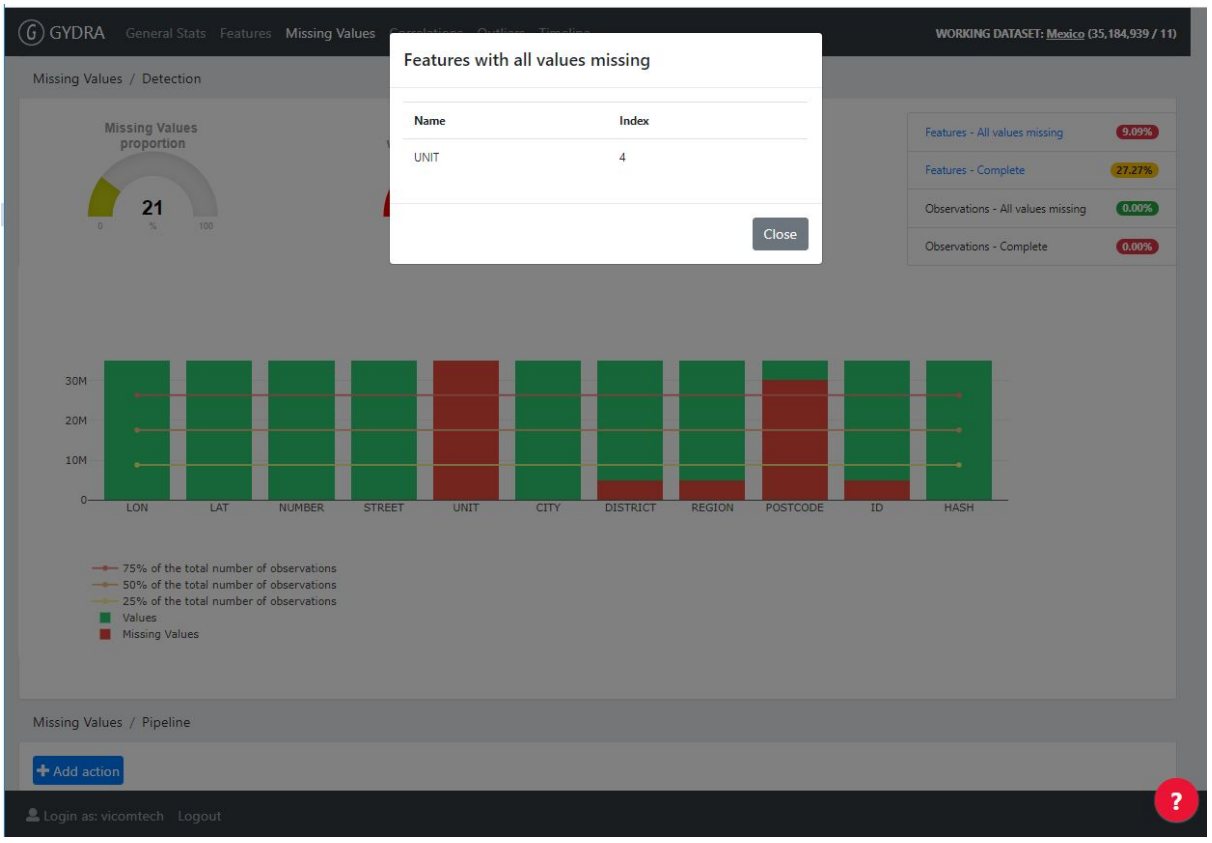


Figure 9: GYDRA tool Missing Values showing a Modal for features with all values missing

1.6 Correlations

As a first step in dimensionality reduction, a section has been developed to detect correlations among variables using a visual representation of the correlation matrix.

Grant Agreement No: 727721

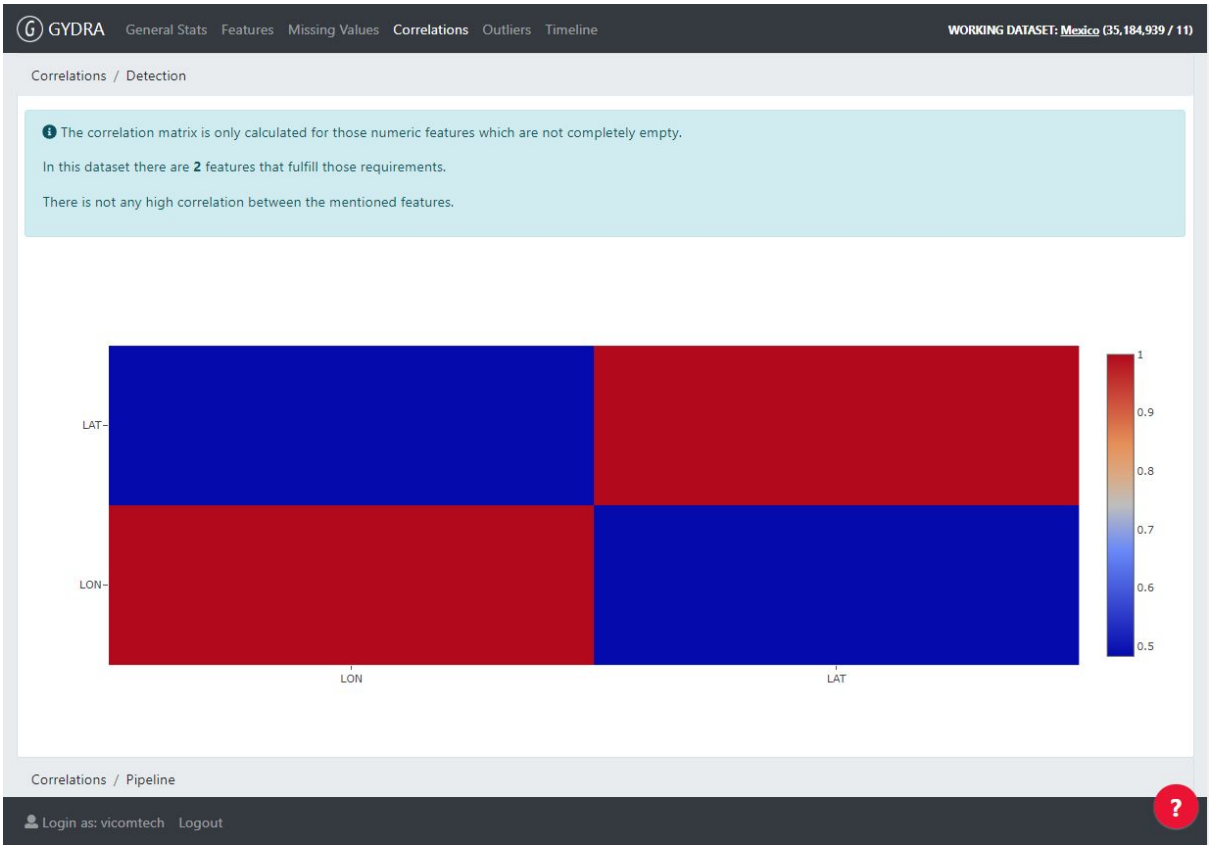


Figure 10: GYDRA tool Correlation analysis tab

1.7 Outliers

Tukey's method has been implemented to detect outliers within the dataset at feature level, considering values from each feature individually to detect the outliers.

Grant Agreement No: 727721

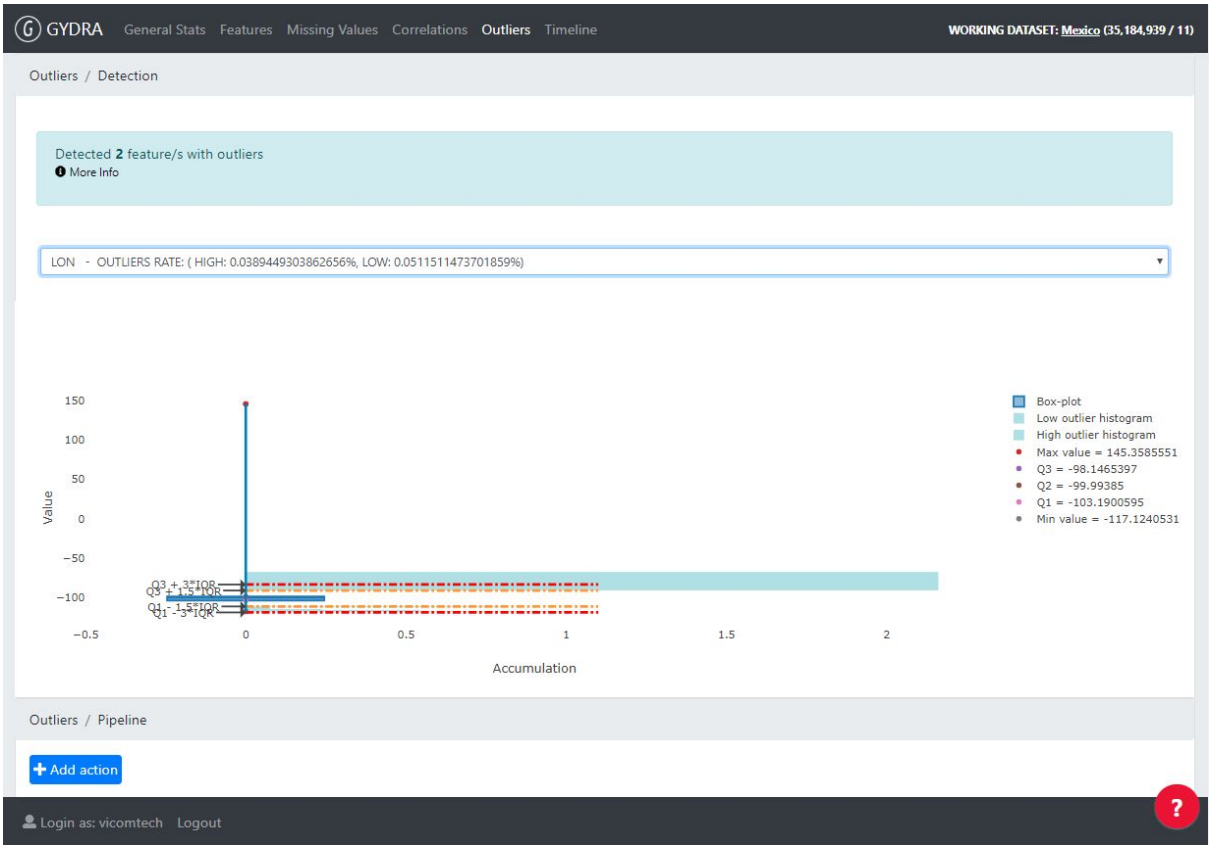


Figure 11: GYDRA tool Outliers analysis tab

At the top of the window there is a “More info” link that shows and hides outliers diagram interpretation information. The diagram represents the outliers of a feature, by grouping outliers above and below normal values (Quartile 3 + 1.5 Interquartile range, and Quartile 1 - 1.5 Interquartile range consecutively) into ten bins each. With Big Data it is not feasible to store, send and show all outlier values.

Grant Agreement No: 727721

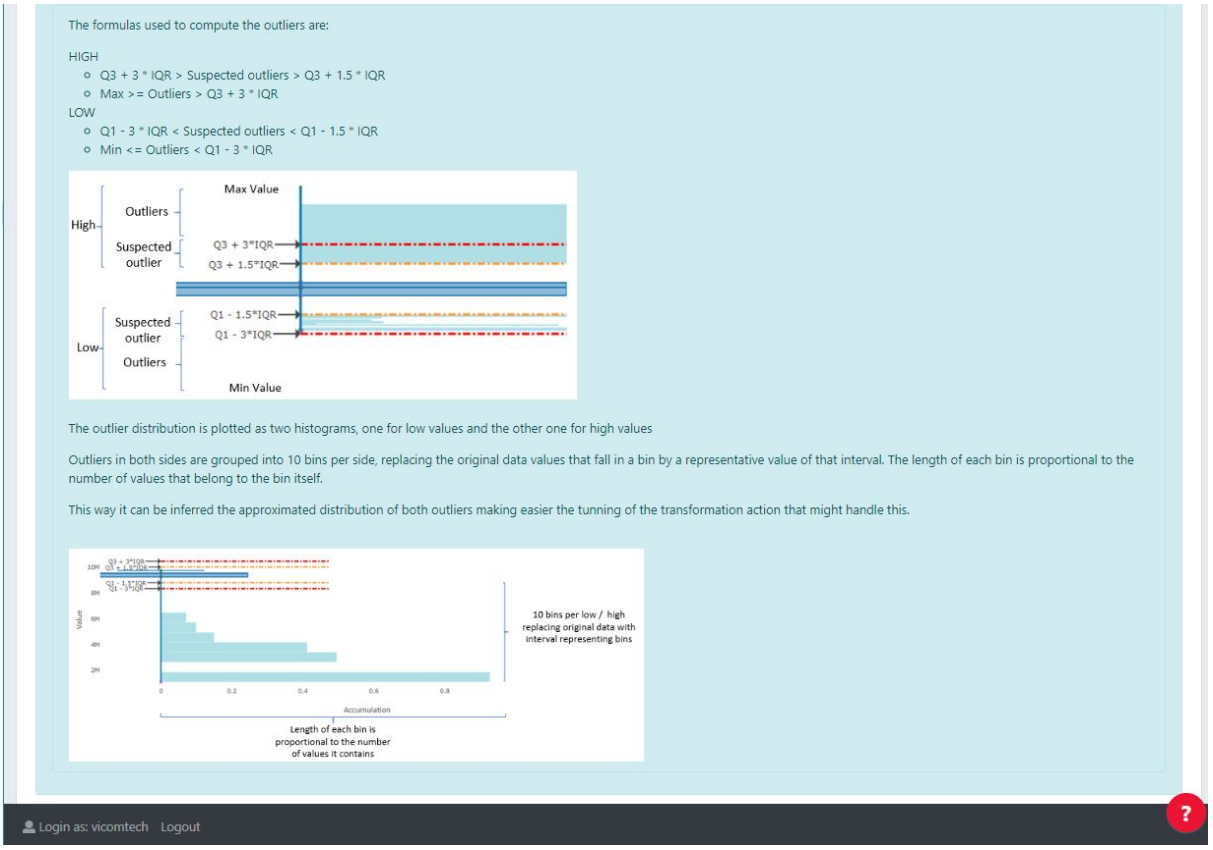


Figure 12: GYDRA tool Outliers diagram explanation content

1.8 Transformation pipeline

Actions can be added / edited in any of the first five results revision GUIs (i.e. General Stats, Features (once a feature is selected), Missing Values, Correlations or Outliers), by clicking on the “Add Action” button. An example is shown in the three figures shown below. The first image shows the transformation pipeline where “Add action” triggers the opening of the transformation configuration modal window. The second image therefore shows the transformation configuration modal window (which updates based on the transformation action selected). Starting at the top in this screen, the transformation action, the transformation target (feature / observation) and the transformation specific details are configured. The third and final figure shows a successfully configured transformation action. If further actions are configured, they’re placed one after the other.

Grant Agreement No: 727721

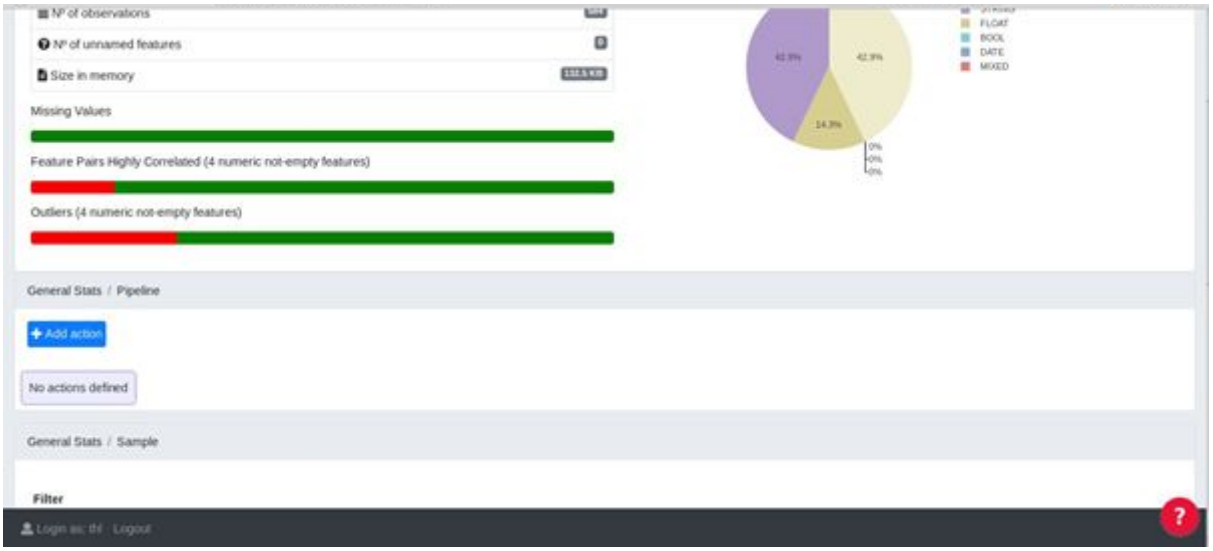


Figure 13: GYDRA Add transformation step 1

Figure 14: GYDRA Add transformation step 2

Grant Agreement No: 727721

Figure 15: GYDRA showing successfully configured transformation pipeline

Pipeline execution can be requested from the Pipeline tab in the navigation bar, by clicking on “Apply actions”.

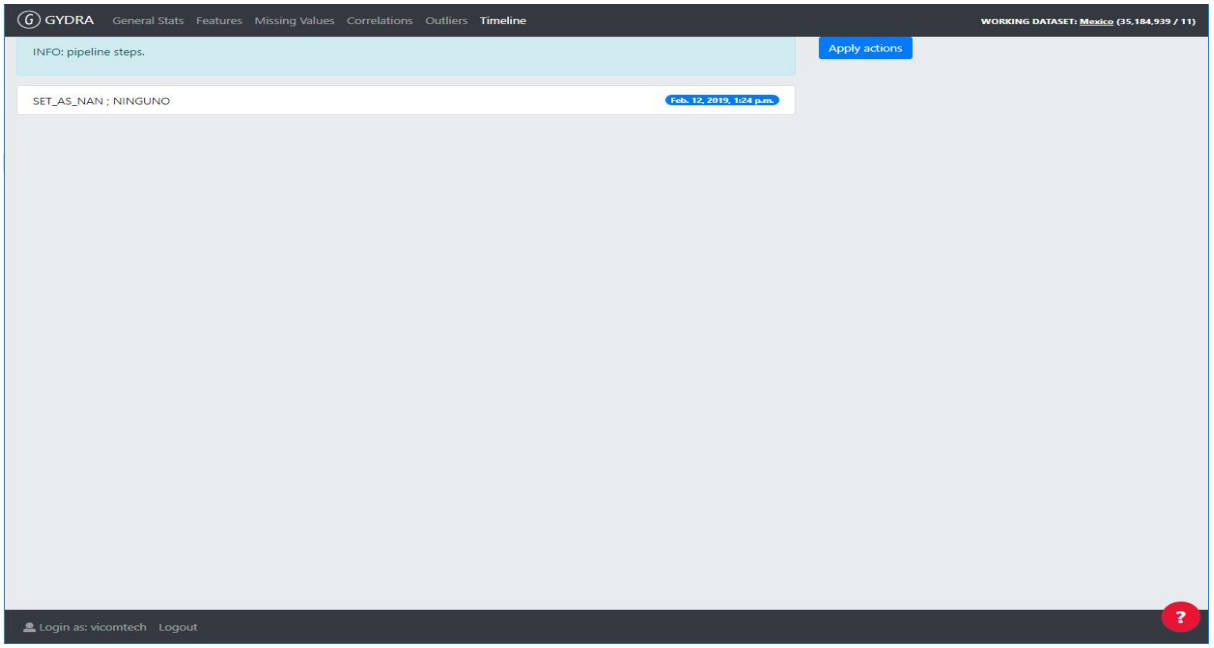


Figure 15: GYDRA tool transformation pipeline resume view and run trigger

If the action is successful a new dataset entry is created in the GYDRA home tab, renaming the current dataset name with v1 (or increased version number) and a reference to the source dataset.

Grant Agreement No: 727721

5 Appendix 2. Maelstrom Classification: Domains and subdomains

Source: https://doi.org/10.1371/journal.pone.0200926.s001

Socio-demographic and economic characteristics

Age/birth date; Sex/gender; Twin; Marital/partner status; Family and household structure; Education; Residence; Birthplace; Citizenship and immigrant status; Ethnicity, race and religion; Language; Labour force and retirement; Income, possessions, and benefits; Other socio-demographic and economic characteristics

Lifestyle and behaviours

Tobacco; Alcohol; Drugs; Nutrition; Breastfeeding; Physical activity; Transportation; Personal hygiene; Sleep; Sexual behaviours and orientation; Leisure activities; Misbehaviour and criminality; Technological devices; Other and unspecified lifestyle information

Birth, pregnancy and reproductive health history

Puberty, menstruation, menopause and andropause; Contraception; Pregnancy, delivery, and birth; Fertility and sexual health; Other reproductive health-related information

Perception of health, quality of life, development and functional limitations

Perception of health; Quality of life; Life course development; Functional limitations; Use of assistive devices; Other perception of health, quality of life and functional limitation-related information

Grant Agreement No: 727721

Diseases (ICD-10)

Certain infectious and parasitic diseases (A00-B99); Neoplasms (C00-D48); Diseases of the blood and blood-forming organs and certain disorders involving the immune mechanism (D50-D89); Endocrine, nutritional and metabolic diseases (E00-E90); Mental and behavioural disorders (F00-F99); Diseases of the nervous system (G00-G99); Diseases of the eye and adnexa (H00-H59); Diseases of the ear and mastoid process (H60-H95); Diseases of the circulatory system (I00-I99); Diseases of the respiratory system (J00-J99); Diseases of the digestive system (K00-K93); Diseases of the skin and subcutaneous tissue (L00-L99); Diseases of the musculoskeletal system and connective tissue (M00-M99); Diseases of the genitourinary system (N00-N99); Pregnancy, childbirth and the puerperium (O00-O9A); Certain conditions originating in the perinatal period (P00-P96); Congenital malformations, deformations and chromosomal abnormalities (Q00-Q99); Injury, poisoning and certain other consequences of external causes (S00-T98); External causes of morbidity and mortality (V01-Y98); Diseases without precise specification or falling into multiple categories

Symptoms and signs (ICD-10)

Symptoms and signs involving the circulatory and respiratory systems (R00-R09); Symptoms and signs involving the digestive system and abdomen (R10-R19); Symptoms and signs involving the skin and subcutaneous tissue (R20-R23); Symptoms and signs involving nervous and musculoskeletal systems (R25-R29); Symptoms and signs involving the urinary system (R30-R39); Symptoms and signs involving cognition, perception, emotional state and behaviour (R40-R46); Symptoms and signs involving speech and voice (R47-R49); General symptoms and signs (R50-R69); Symptoms related to multiple categories

Medication and supplements

Medication and supplement intake; Posology and protocol of administration; Other and unspecified pharmacological interventions

Non-pharmacological interventions

Surgical interventions; Radiological interventions; Physical therapy interventions; Cognitive, psychological and sensory interventions; Educational and health promotion interventions; Laboratory diagnosis interventions; Other and unspecified non-pharmacological interventions

Grant Agreement No: 727721

Health and community care services utilization

Visits to health professionals; Hospitalizations; Community and social care; Other health and community care

Death

Vital status; Cause of death; Other end of life or death-related information

Physical measures and assessments

Physical characteristics; Anthropometry; Circulation and respiration; Muscles, skeleton and mobility; Sensory and pain; Brain and nerves; Skin and subcutaneous tissue; Speech and voice; Digestion; Reproduction; Other physical measures and assessments

Laboratory measures

Hematology; Biochemistry; Microbiology; Virology; Immunology; Toxicology; Histology; Genomics; Other laboratory measures

Cognition, personality and psychological measures and assessments

Cognitive functioning; Personality; Psychological distress and emotions; Other psychological measures and assessments

Life events, life plans, beliefs and values

Life events; Life plans; Beliefs and values; Other life events, plans and beliefs

Preschool, school and work life

Preschool life; School life; Work life; Other preschool, school or work life-related information

Social environment and relationships

Social network; Social participation; Social support; Parenting and familial environment; Other social environment characteristics

Physical environment

Housing characteristics; Built environment/neighbourhood characteristics; Workplace characteristics; Radiation exposure; Chemical exposure; Biological exposure; Other physical environment characteristics

Administrative information

Identifiers; Date and time-related information; Questionnaire and interview-related information; Physical and cognitive measures and bio sample-related information; Data and sample collection center-related information; Other administrative information