European Commission

WAVESTONE

Institute of Communication & Computer Systems
of the National Technical University of Athens
ICCS-NTUA

TNO innovation for life

CEPS

GRIMALDI STUDIO LEGALE

EASME/COSME/2018/004
Big Data and B2B platforms: the next big opportunity for Europe

WORKSHOP BRIEF
Second Workshop on "Piloting a high-quality, diabetes-related data repository by using the latest technologies and big data breakthroughs"

September 17, 2019

Disclaimer

The information and views set out in this publication are those of the author(s) and do not necessarily reflect the official opinion of the European Commission. The European Commission does not guarantee the accuracy of the data included in this document. Neither the European Commission nor any person acting on the European Commission's behalf may be held responsible for the use which may be made of the information contained therein.

Introduction

The European Commission and the Executive Agency for Small and Medium-sized Enterprises (EASME) are conducting the study “Big Data and B2B platforms: the next big opportunity for Europe” whose objective is to analyse how to accelerate the growth of the data-based economy and to support the development of B2B platforms in Europe focusing on two sectors: automotive and healthcare.

This study gives particular attention to the opportunities offered by Big Data and B2B platforms for Small and Medium Size Enterprises (SMEs), with the purpose of maximising their innovation potential and contributing to growth and employment in Europe.

Wavestone is leading this study, in partnership with the ICCS-NTUA (National Technical University of Athens, TNO (Netherlands Organisation for Applied Scientific Research), KOMIS and CEPS. GRIMALDI Studio Legale provides legal advice transversally to all work packages.

The study started on the 20th December 2018 and will end on the 19th December 2020, for a total duration of two years.

The sector of health and healthcare is covered by **Work Package 2 (WP2).** The aim of WP2 is to design and implement a pilot to build a pan-European, high-quality, diabetes-related data repository by using the latest technologies and big data breakthroughs. The pilot will be used to study the challenges and opportunities of a type 2 diabetes health data marketplace. Part of the work is to organize a series of workshops (4 in total for WP2), all with a different focus on data driven health.

This **second workshop** will be focused on the Implementation phase of the Pilot. The work done so far will be presented: the current status of the infrastructure development, the concept of the use cases that will be executed to test the environment, and a demo of the current status of the data infrastructure. In an interactive working session, participants will debate and work towards reaching a common agreement on specific ordering of ontologies, vocabulary and on identifying potential bottlenecks related to the use cases.

This brief provides an overview of the objectives of the pilot and of the organisation of the work as a refresh for returning participants and as an onboarding for those who did not attend to the first workshop (23 May 2019, Brussels). The brief also includes a status update of the work done so far, some takeaways from the first workshop, and an explanation of the work to be done together in the second upcoming workshop, for your knowledge and preparation.

1 Objectives and scope of the Pilot

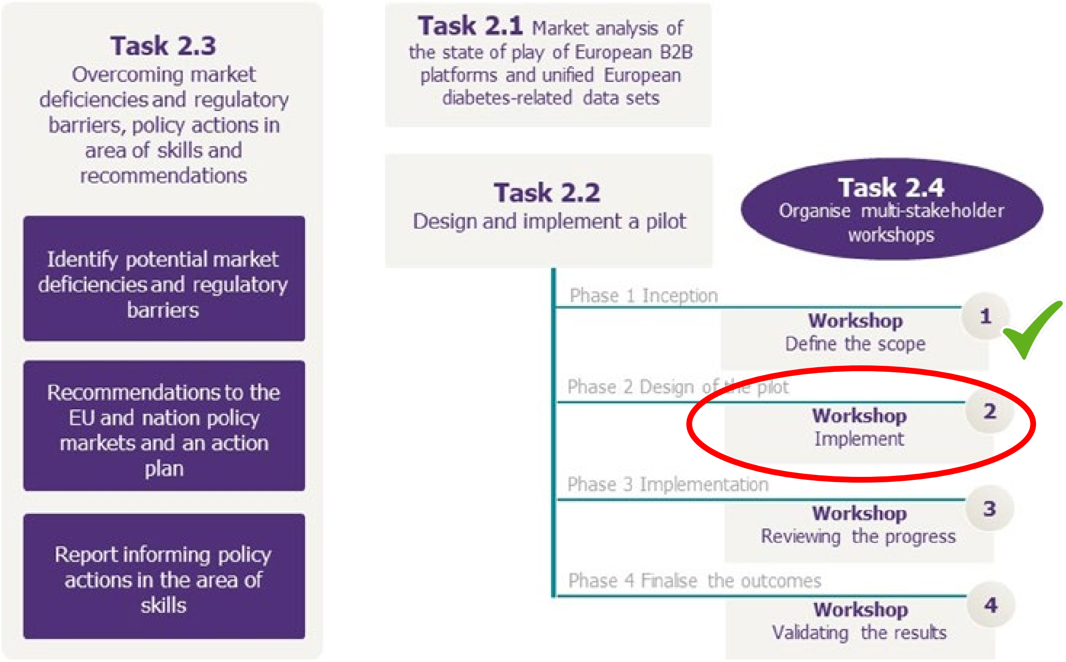
The objectives of the Type 2 Diabetes (T2D) healthcare pilot can be summarised as follows:

- To **extract and exploit data sets** from both, **‘omics’ profiling and other biomarkers** from key European and national research projects or prospective **cohorts related to T2D**. This data might be used to develop a comprehensive biomarker package that quantifies the relevant processes of the metabolic flexibility system and determines an individual’s metabolic health trajectory and predisposition to certain conditions.
- To **combine such data repository** with **Real World Evidence (RWE)** data. RWE is defined as healthcare information that is derived from multiple sources outside the typical clinical research setting.

2 Organisation of the work

Firstly, a market analysis of the state of play of European B2B platforms will be executed together with a search for unified European diabetes related data sets and FAIRification of data (**task 2.1 of WP2**). **Secondly**, the pilot will be designed, implemented and extended. The focus will be on a limited number of use cases that represents good examples and clear views of the problem. The use cases will demonstrate the weaknesses and barriers as well as the business opportunities of data sharing and new scientific findings (**task 2.2 of WP2**). **Thirdly**, a study will be conducted that will focus on overcoming market deficiencies and regulatory barriers (**task 2.3 of WP2**). The design and implementation of the pilot will be supported by a series of multi-stakeholder workshops (**task 2.4 of WP2**). **WP2** will in total organise and carry out four workshops, one for each phase of the pilot, to discuss the progress, and validate with selected experts the intermediate and final results of the work done under WP2. Figure 1 is an overview of WP2 work structure and indicates the current phase we stand in today.

Figure 1 Task organisation for WP2



Task 2.3
Overcoming market deficiencies and regulatory barriers, policy actions in area of skills and recommendations
Identify potential market deficiencies and regulatory barriers
Recommendations to the EU and nation policy markets and an action plan
Report informing policy actions in the area of skills
Task 2.1 Market analysis of the state of play of European B2B platforms and unified European diabetes-related data sets
Task 2.2
Design and implement a pilot
Task 2.4
Organise multi-stakeholder workshops
Phase 1 Inception
Workshop
Define the scope
1
Phase 2 Design of the pilot
Workshop
Implement
2
Phase 3 Implementation
Workshop
Reviewing the progress
3
Phase 4 Finalise the outcomes
Workshop
Validating the results
4

Source: Study authors

3 What we have done so far

According to the International Diabetes Federation, Type 2 diabetes (T2D) is expected to affect 592 million people worldwide by 2035 and 70 million people in Europe alone[1]. The impacts on healthcare costs will be tremendous. Big data research offers an opportunity to reverse this trend.

Although a large number of diabetes-related data repositories is available within Europe, the full potential of this data still cannot be exploited. Several issues underlie this overall problem, which is not only limited to diabetes data. For instance, 1) data are stored in data silos that cannot be connected, for example because the datasets are not findable (no persistent identifier); 2) it is not clear for what purpose the data are accessible; they are fully closed (even the study description); and 3) the terms

[1] Data refer to the year 2017. See also: https://www.diabetesatlas.org/

used are not mapped to commonly used terms (ontologies or vocabularies). These issues are hampering the re-use of health data in general and type 2 diabetes in specific. Applying the **FAIR (Findable, Accessible, Interoperable, Reusable)** data principles resolves those issues (Wilkinson et al., 2016).

To fully understand the complexity of T2D and to be able develop interventions that help to prevent the disease these data need to be enriched with medical data of individuals and with their lifestyle data (real world data). Obviously, such data are very sensitive and require the full execution of the GDPR, which requires secure data solutions and full consent by the individual. Unified EU-wide data sets for diabetes, with a view to pave the way for pan-European high-quality aggregation of data, are only possible if data are well standardized and privacy is preserved.

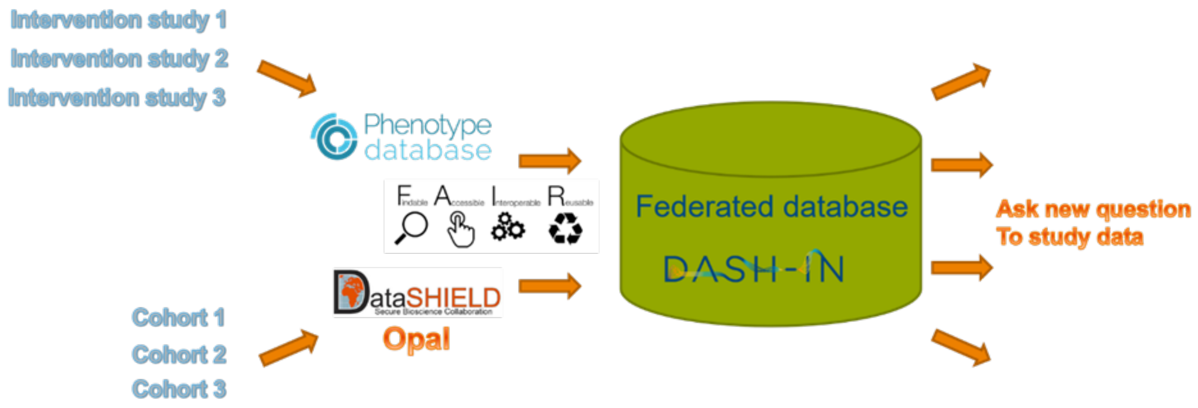
This study will pilot a scenario for research databased on the **data shield Federated data** architecture.

In this study, in order to implement the GDPR we want to connect to data in a privacy-by-design way. Therefore, we will set-up a pilot that is able to test and further develop a federated infrastructure that connects data from different research sources (including omics data) and real-world data. The pilot will address specific use cases that will show the added value of such an infrastructure.

For this pilot we will make use of the Datashield solution that is implemented in ENPADASI for nutritional study data, connect to other diabetes cohorts and we will fully incorporate the FAIR principles. For diabetic databases that can be adapted, we will:

- facilitate **interoperability** of the system's parameters,
- make the API **findable**, and
- add the **accessibility** details to the system (make them FAIR points).

Figure 2. Schematic representation of the data shield federated data architecture



Source: Study authors

In addition to research data, patient-centred health data also referred to as Real World Data (RWD) is needed**.** RWD is described as healthcare information that is derived from sources found outside standard clinical research data. Examples of Real World Data (RWD)[2] are data from Electronic Health Records (EHRs) and data gathered from self-tests, wearables and other connected devices and eHealth applications. RWD can enrich research-based data. This is especially important for chronic diseases such as T2D where lifestyle and nutrition are important causes of the disease. RWD are personal and therefore the patient/citizen is crucial to the gathering and consent regarding the use of this data. The GDPR allows for more control over health data; the patient is entitled to receive copies of health data,

[2] https://www.mckinsey.com/industries/pharmaceuticals-and-medical-products/our-insights/real-world-evidence-from-activity-to-impact-in-healthcare-decision-making

and data providers should make such data portable. The study team implements an application that will support retrieval and reuse of RWD in a citizen-controlled way.

Both solutions will lead to a **prototype of a T2D health data marketplace** that will respect privacy of citizens/patients and at the same time allow for valorisation of health data. Thereby we will be able to find a fair description of the health data available without having to connect directly to the owners of the data, which can be research institutes (anonymised research data) as well as patients. For this, the governance of the study is of the utmost importance; a sophisticated consent system shall be in place to arrange consent over the use of the data. The governance will also provide guidelines on how data needs to be processed in compliance with the GDPR. Application areas are that might utilise such a health data marketplace are for example biomarker discovery, A.I. modelling, patient subtyping, personalised advice model development; and connection of clinical trials to patients (or patients to clinical trials).

Importantly, several **risks need to be identified and addressed**. The most important risk is data quality. Use within clinical applications, especially, requires high quality data. Standardisation of research data can at least help to identify the level of quality of the data. Provenance of data is therefore key. Another risk is the lack of context of data. This is especially the case for self-reporting, or data gathered by individuals through wearables.

Applying the outcomes of the first workshop (see also section 4 of this document) we started developing an infrastructure using available datasets.

3.1 Use cases: biomarkers for T2D sub-typing and development of a prognostic T2D model

We propose two **use cases** to test the functionality of the infrastructure and show the added value of combining different types of T2D related data. The **first use case** will focus on the **development of biomarkers for the sub-typing of type 2 diabetes**. This use case will apply data from research studies of different sources. The **second use case** will be focused on the **development of a prognostic model for the development of type 2 diabetes** and will apply personal health data.

Biomarkers for sub-typing of type 2 diabetes. Type 2 diabetes is currently diagnosed and treated as one disease. Recent research has shown that type 2 diabetes has various underlying mechanistic causes that may need different interventions for optimal treatment of the disease[3]. The first use case will target studies that contain data from which the subtypes can be deduced. The studies that in addition contain omics data will be used to discover new biomarkers for the sub-typing of type 2 diabetes (from the data recourses MOLGENIS/ENPADASI). Insights derived from this use case might be exploited by various SMEs, particularly those active in the Health Care sector as a new opportunity for pharmaceutical and Consumer Health segment (this is currently being investigated in the market analysis for this study).

The other use case explores the possibility of Personal Health Data for the development of **prognostic (AI-based) models** that can be applied for the prediction of (early) onset of type 2 diabetes. Data will be derived from the available health data sets within the T2D data infrastructure. This kind of prognostic models could be exploited by SMEs that operate in the Health Care and Consumer Health segment.

In addition, both use cases need standardisation of data according to the FAIR principles. The 'FAIRness' of data or enabling and supporting the data infrastructure is regarded as an important asset for SMEs from the Academic Research segment. The necessity of creating a system according to the FAIR

[3] Wopereis, S. (2017), Multi-parameter comparison of a standardized mixed meal tolerance test in healthy and type 2 diabetic subjects: the PhenFlex challenge, Genes Nutr. 2017 Aug 29;12:21. doi: 10.1186/s12263-017-0570-6. eCollection 2017, available at: https://www.ncbi.nlm.nih.gov/pubmed/28861127

(findable, accessible, interoperable and reusable) principles was discussed so far and validated as a building block of the pilot's development (see also findings from the first workshop in section 4 of this document). If all systems adhere to these principles the data could easily be accessed, even across databases. Data is considered accessible if machine readability can identify for whom, when and for what the data can be used.

3.2 Connecting data sources for the use cases

Important for the use cases is to start with the selection and collection of a suitable set of T2D data. There are many systems that collect research data and several initiatives have worked towards catalogues of data or integrated systems. Ontologies can help to map the relation between the structure of the data from the different resourses by defining the logic between the different data structures and make them interoperable. It is possible to connect data from different studies if terms are the same (i.e. it is interoperable) and can be reused if the data are accompanied by rich meta-data (i.e. information on how the data are collected and details about the study are available). In this pilot different systems that contain T2D data will be harmonised to be able to learn from the combination of the data.

We have segmented the different types of data available for T2D into two categories: 1) **Research data** (Observational study (cohort) data, Interventional study data) and 2) **Personal health data** (Medical data and real world health data).

3.2.1 Research data

The biggest portion of relevant data available for T2D are research data. Research data for T2D can be split in data from **interventional** and **observational studies**. In interventional studies participants receive an intervention in the form of a treatment. This might be medication or lifestyle related, but participants might also get coaching or behavioral strategies as intervention. Clinical trials are an example of intervention studies that evaluate the efficacy of an intervention. Participants in observational studies do not receive an intervention. These are typically cohorts, observational, exploratory, case-control and survey studies. Data from observational studies can be used to build predictive models. Observational studies are often rich in the number of participants but poor on the number of data points whereas intervention studies might be rich of individual data points (for example by using omics data; genome, epigenome, metabolome, microbiome) but poor in the number of participants involved in the study.

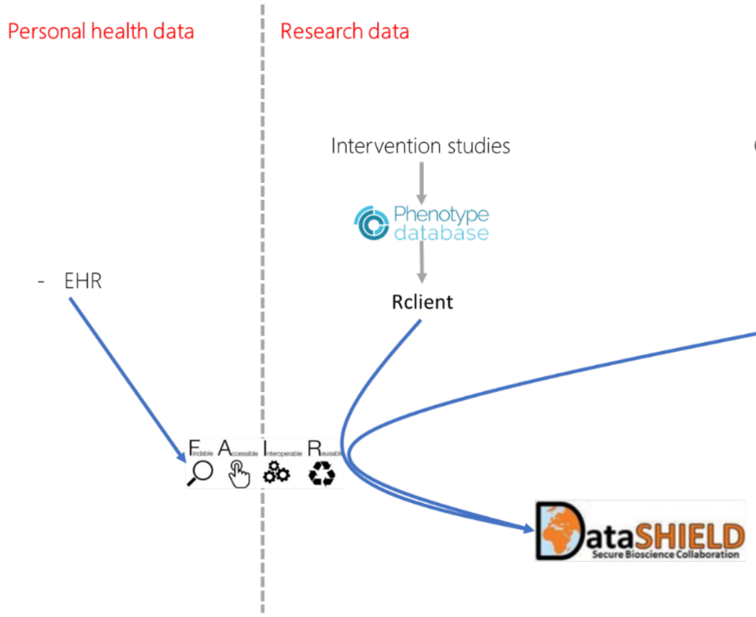
Implementation: next steps for research data

In the current pilot, the solution developed in ENPADASI (DASH-IN), that contains over 100 nutritional intervention studies, is connected to MOLGENIS[4]. MOLGENIS is a scalable software infrastructure that enables to capture, exchange, and exploit large amounts of research data, mainly focusing on cohort and observational studies. Coupling DASH-IN and MOLGENIS results in an infrastructure were different types of research questions related to type 2 diabetes can be answered. MOLGENIS is used in the biobanking consortium BBMRI. In addition, we will use Datashield[5], a privacy preserving system to connect DASH-IN and MOLGENIS. Datashield will make it possible to learn from (summarised) data without ever needing access to the individual data, this resolves the GDPR and legal issues of data reuse.

[4] MOLGENIS database: https://www.molgenis.org

[5] For more information: http://www.datashield.ac.uk/

Figure 3 Specification of the T2D data infrastructure



Source: Study authors

The systems will be made FAIR (findable, accessible, interoperable and reusable) to the current standards, so that the wording and grammar of the systems adhere to each other (standardization). In the first workshop, the need for interoperability vocabularies and ontologies, which are specific to certain areas, was extensively discussed. It was agreed that for the T2D pilot specific choices for ontologies/vocabularies need to be made and that some extensions of ontologies/vocabularies are needed. A basic set-up for these vocabularies/ ontologies have been created. Several relevant ontologies were identified in Bioportal[6], such as the Diabetes mellitus diagnosis and Diabetes mellitus treatment ontology. In addition, we have mapped the relevant clinical marker to a set of ontologies and have started in mapping the work of the expert group from Obedis to ontologies. This last expert group developed 'European expert guidelines on a minimal core set of variables to include in randomized, controlled clinical trials of obesity interventions'.

These datasets will be used in the project for two use cases (biomarkers for sub-typing T2D and prognostic models for T2D, see below), to find new biomarkers we will focus on datasets that are rich (e.g. containing mainly parameters per individual, such as in the case of omics data).

The following actions are required to achieve the objectives mentioned above:

- Make R-client Phenotype database FAIR (the R client is a connection between the database and R so that the data from the database can be directly analysed in the statistical package R);
- Connect to the developers of Molgenis to adapt Datashield client (to make a direct connection between the Phenotype database and MOLGENIS with Datashield);
- Add T2D additional data sets to the system;
- Normalize data so that studies are comparable;

Find a virtual cohort in existing studies of T2D and healthy subjects.

These steps have been partially implemented. The R-client was adapted to get access to individual parameters multiple studies. Regarding interoperability, several ontologies have been selected. During the second workshop the current status of the pilot will be presented.

[6] See also: https://bioportal.bioontology.org/

More information about the outcomes of the first workshop can be found in the first workshop report.

3.2.2 Personal Health Data

A second rich source of health data for T2D are Electronic Health Records (EHR), that contain medical data. This data is typically collected in a hospital or general practitioner (GP) setting. EHR generally contain longitudinal data of T2D patients. Dependent on the health care setting, data will be collected by GPs or in hospitals. EHRs contain both structured data, for example clinical measurements but also unstructured data like doctors' notes.

An interesting new data source is real world health data. Since T2D is a disease partially caused by lifestyle, real world health data is important to fully understand the disease and its treatment. For example, wearables that measure activity, sleep or heart-rate are sources of these data**.**

Implementation: next steps for personal data

The actions planned for the next phase in the area of personal data are:

- Focus on structured data and based on the use case we will define which unstructured data should be made structured (by making the data FAIR).
- Connect personal health data (both Electronic Health Records and real-world health data) with the research data. This will make it possible to understand the influence of lifestyle on T2D even better. We will start collecting medical and real-world data with small groups of volunteers in a federated system and connect the data to the research data by FAIRification of the personal data API.

The actions for the implementation of personal health data have not been fully carried out. Then current issues with implementing personal health data will be addressed during the plenary discussion of the second workshop.

4 The fist workshop ('Define the scope')

This first workshop held on 23 May 2019, covered two content blocks: the **infrastructure of the data sharing**, and the **opportunities and barriers for the health data market and commercial use** in the Type2 diabetes (T2D) domain. The workshop's main objective was to introduce the participants to the objectives of the health data pilot, present the work done so far on the infrastructure, and most importantly to gather the participants' views on infrastructure and business opportunities for health data in the T2D domain.

During the workshop, the criteria for an infrastructure were presented. It was agreed that in order to connect the different data sources, standards are needed. The forementioned FAIR principles and the need for specific type 2 diabetes targeted ontologies, the FHIR (Fast Healthcare Interoperability Resources) standard for interoperbalility of Health Data and the specific Dutch MedMij standard were discussed. The **overall Conclusions and takeaways for the next phases of the Pilot** can be summarised as follows:

- For the majority of the participants, federated solutions are preferable to one overarching IT-infrastructure; therefore, we will continue further implementing the DataShield and connecting the Phenotype database to Molgenis, the open source data-platform for research data.
- The structure should apply FAIR principles for better health data management.
- "If you don't have to share, do not share", which means that many times it is not necessary to share the full database/dataset, but only what is valuable for the specific request.

- The development of the pilot should be fully compliant to GDPR and thereby privacy will be preserved. The Datashield solution is built for this purpose.
- Dynamic consent has been proposed as a way to guarantee a more effective handling of sharing health data; for the real-world data, we will pilot the digi.me type of environment.
- Standardisation is an important issue to be addressed: it is important to come up with a minimum set of semantic standards (e.g. controlled vocabularies, data dictionaries), therefore we will use one of the next workshops to define standards for the diabetes community.
- Quality of the data needs to be taken into account.
- The key is to have clear metadata so to be able to judge the quality of the underlying data.
- There should be openness on what is feasible and what not, we will/cannot resolve all but take it step by step.

5 The second workshop ('Implement')

5.1 Objectives of the second workshop

This second health data workshop will cover: 1) the current **status of the infrastructure of the data** sharing and 2) an interactive session on the **use cases and needed ontologies/vocabularies**. Hence, the workshop's **main objective** is to get input on the work done so far on the infrastructure, and most importantly to define the terms and structure of the diabetes FAIRport.

5.2 Working together during the second workshop

The workshop will start with a **presentation session** were the current status of the pilot will be presented, and a demo on the data infrastructure carried out. More specifically:

- The **design choices that were implemented** as result of the first workshop will be presented and discussed in a Q&A. Especially the choice for the combination of dbNP, Molgenis and Datashield will be discussed (e.g. Are the data sufficient for the proposed use cases? Are the use cases relevant to test the type 2 diabetes database?).
- The **data infrastructure will be demonstrated**, including how we will perform complex analysis within Datashield. This includes the selected databases and how data is integrated. Part of the demo is the research queries. The possibilities and the limitations of the queries will be discussed in a Q&A (e.g. Is the chosen infrastructure too limited for complex queries and calculations or are there workarounds? Are these workarounds good enough for the purpose of this project?).

The **interactive session** will be dedicated to the definitions of ontologies that are needed for the pilot. The ontologies for data logics (ONS and ONE), diabetes diagnostics, diabetes treatment, etc. will be shown. The participants will be asked to actively contribute by filling the blanks. The **interactive session** is organised in two subgroups:

- **Ontology structure:** Type 2 Diabetes FAIRport developed within this pilot will require a specific ontology structure to be able to answer the relevant questions. Experts with a technical background are requested to give input on this part of the workshop. The main emphasis will be to identify the missing ontologies that are essential for the use cases. We will discuss how the current ontologies can support the use cases within the limitations of Datashield, e.g. how the data logics ontologies can help to prepare the data in the local data instance.

- **Vocabulary definition:** ontologies and vocabularies need to be selected and prioritized to be able to make the datasets interoperable. The selections and blank spots will be discussed in this subgroup. Experts with a biomedical or analytical background are requested to join this part of the workshop. This group will work on definitions around the data; for example, which ontologies are already in place and useable, which one are missing, and which terms are missing and should be chosen for certain purposes.

A report identifying potential market deficiencies and regulatory barriers is being prepared in the scope of this study.

The report will look at the entire ‘regulatory stack’ for the healthcare sector as well as at horizontal areas that affect more generally the development of data analytics and data-driven services. At the same time, it will consider how emerging technological solutions could mitigate existing barriers in the short- to medium term. Finally, policy recommendations to overcome the detected barriers will be prepared.

We would like to kindly invite you to take part to an expert interview to help us assess the barriers identified so far, and potentially suggest additional barriers as well as policy and technological solutions to overcome such obstacles.

In case you have not yet received the invite to these interviews and you would like to contribute, we would greatly appreciate if you could inform Felice Simonelli felice.simonelli@ceps.eu, and Nadina Iacob nadina.iacob@ceps.eu.

We would like to take this opportunity to thank you for your participation in this important study.