Big Data and B2B Platforms: the next big opportunity for Europe

WAVESTONE

TNO

ICCS

CEPS

KOMIS

GRIMALDI STUDIO LEGALE

3nd Workshop
Infrastructure of health databases in T2D

Big Data and B2B platforms: the next big opportunity for Europe

EASME/COSME/2018/004

22 June 2020

WAVESTONE

CEPS

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

Infrastructure and application of shared (or connected) T2D databases.

TNO innovation for life

European Commission

AGENDA

9:00	9:05	5 min	Welcome	Wavestone
09:05	09:10	5 min	Goal setting Recap on the objectives and aims of this study under WP2 "Explore the feasibility of, pilot, promote and stimulate strategic investment in	TNO
09:10	09:25	15 min	Progress Update of the Pilot on high-quality, diabetes-related data repository by using the latest technologies and big data breakthroughs The status of the infrastructure development and the use cases will be presented : - Demonstration of the 1st use case - Demonstration of the 2nd use case	TNO
9:25	9:50	25 min	Demonstration of the Information and Data portal - Live demo of the search functionality and selection of database preferences and features; - Demonstration of data extraction with hypothetical business use case application	TNO
9:50	10:25	35 min	Break-outs: sessions on specific questions on the usability of the portal	TNO
10:25	10:40	15 min	Break - Bring your own tea/coffee	
10:40	11:10	30 min	Invited speaker	Sitra/Findata
11.10	11.40	30 min	Break-outs: Discussion on future solutions for data reuse (centralized/decentralized)	TNO
11:40	11:55	15 min	Break - Bring your own tea/coffee	
11.55	12.05	10 min	Summary of the feedback on portal & explanation of further feedback	TNO
12:05	12:20	15 min	Summary on the Market Analysis	TNO
12:20	12:40	20 min	Break-outs: Opportunities for Commercialisation	TNO
12:40	12:45	5 min	Thank you for the participation Satisfaction survey will be sent via email to be filled in.	All

Update

09:10	09:25	15 min	**Progress Update of the Pilot on high-quality, diabetes-related data repository by using the latest technologies and big data breakthroughs** **The status of the infrastructure development and the use cases will be presented :** **- Demonstration of the 1st use case** **- Demonstration of the 2nd use case**	**TNO**

General goal of the T2D pilot

Test and further develop a **federated infrastructure** that connects and unlocks data from different research sources (including omics data) and Real World Data

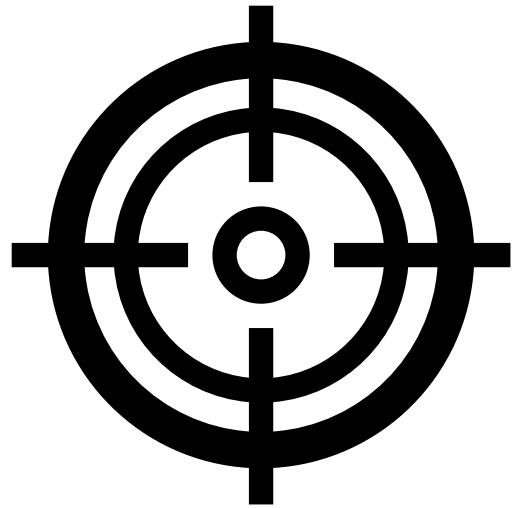
Two use cases

- 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**. These data could 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.

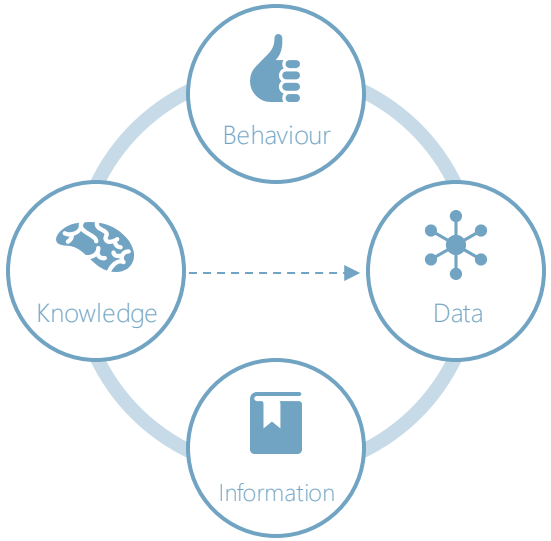
Goals today

- Review the progress and give feedback on the implementation process of the Pilot
- Discuss on the functionality of the system, and on the usability of the interface
- See how the concepts analysed in this pilot are used in practice (presentation by Sitra`- Saara Malkamäki)
- Results of the Market Analysis will be presented
- Business opportunities if data sources on Type 2 Diabetes can be accessed, will be explored with the participants.



Use case 1. Get access to several T2D diabetes dataset for development of biomarkers

- T2D Data availability & FAIRness (portal)
- Demonstration of the real-time interoperability solution as it can be used by everybody, so that a community diabetes ontology based terminology can be build
- Demo of use case 1: smart phone app that can predict your insulin based on glucose using several distributed datasets
- Federated Machine leaning model on data of distributed databases



Use case 2: Including real world data

We will present the outcomes of the scripts of:

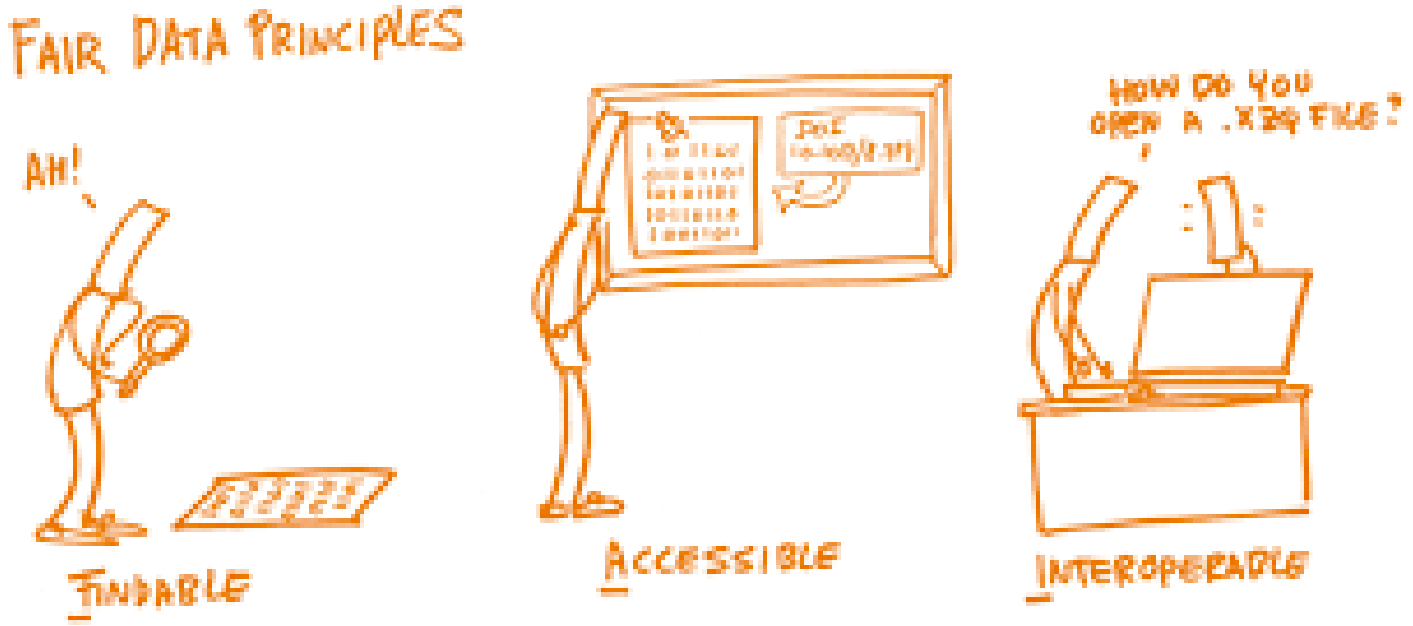
- Federated Machine leaning model on data of distributed databases including real world data

Reoccurring data challenges

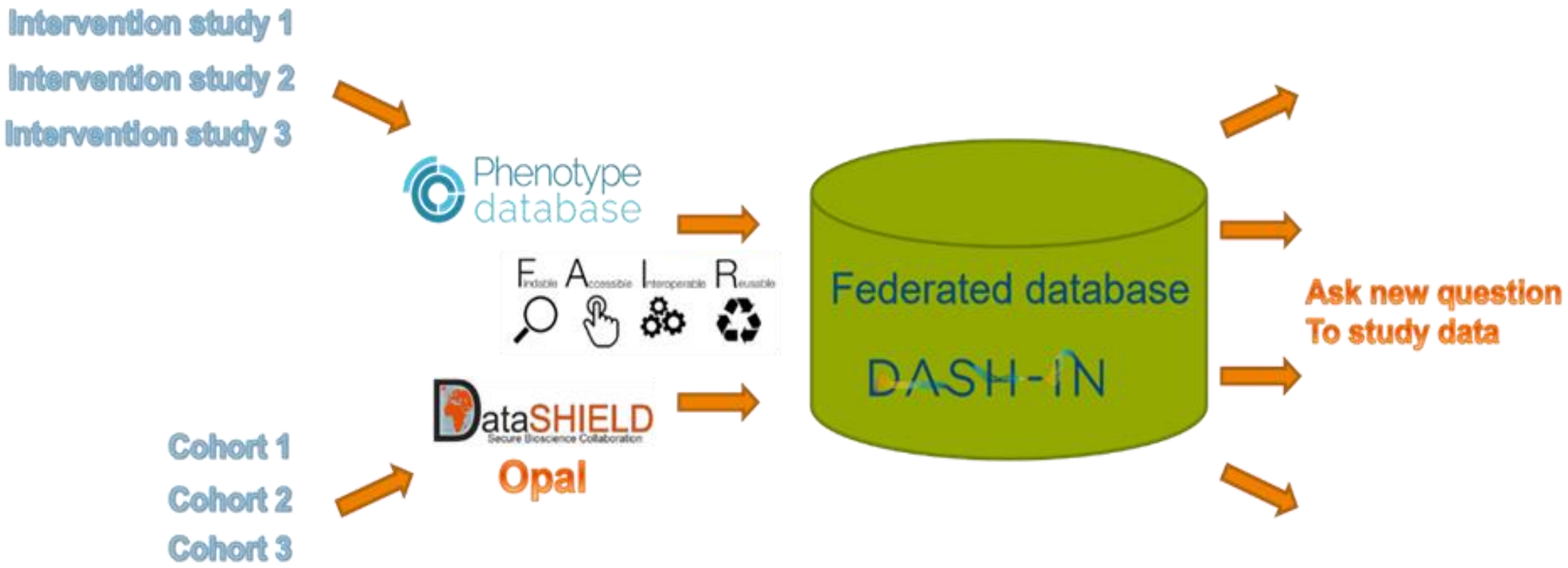
Several data sources indicate to be FAIR, however

Legal (contracts): for every data access the contract has to be negotiated and data is often proprietary

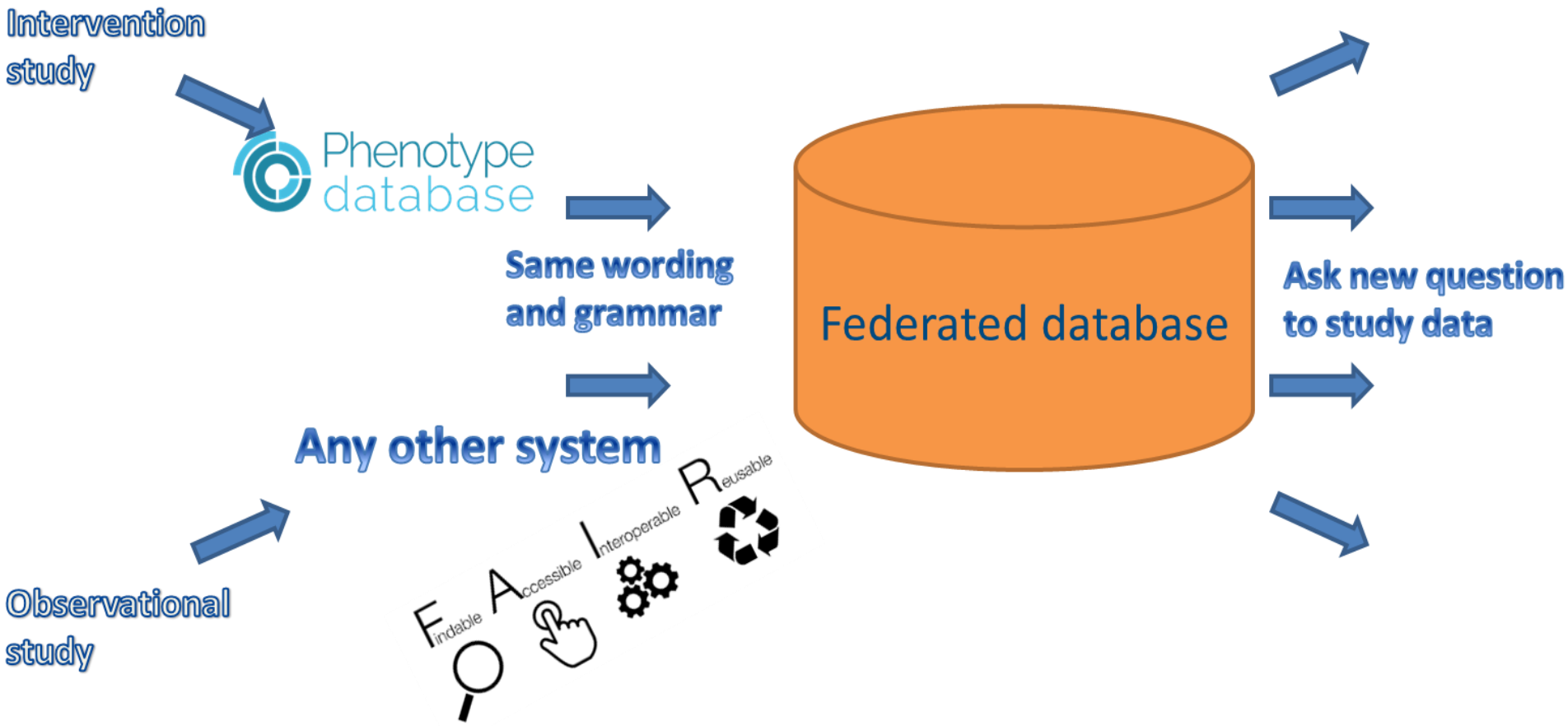
Privacy (fear of GDPR): There are no solutions yet around the GPDR, even if solution are privacy-by-design



Initial plan for FAIRPOINT



Adapted plan

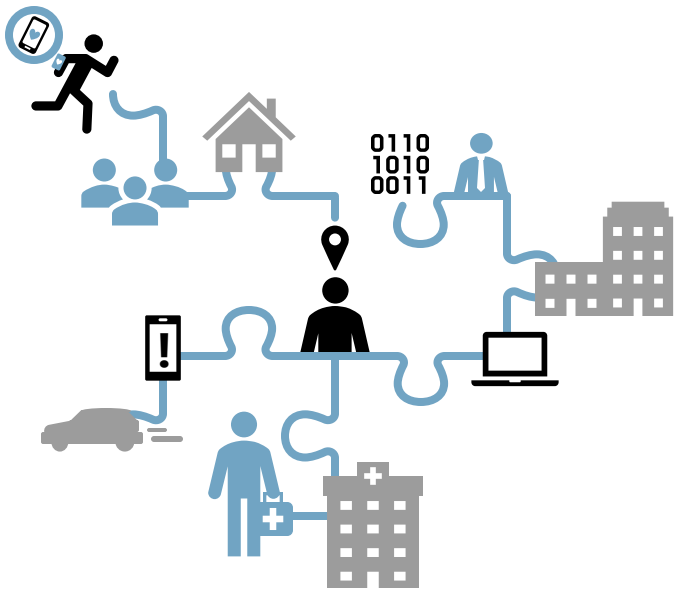


Access to the systems

Portal: https://dashin.eu/easme/login/ and please sign up on this website

Scripts (github, if licenses are OK): All the scripts are currently stored and versioned in an internal version of Gitlab

Demo-app: Only tested for 1 platform and phone-type, not yet in any store



Overall Conclusions and takeaways from WS2

- Try to involve pharmaceutical companies in sharing T2D health data
- Try to connect other organizations for sharing data by rewarding them with free access, or access for a certain period
- The overall conclusion of the Ontology working session was to start with the ICHOM standard and use the items to couple them to the DMDO ontology. Make sure that this is directly FHIR compliant
- Make use of existing ontologies whenever possible rather than trying to build yet another standard from the ground up
- Ontologies should be used to standardize data and logic should reside in the script
- Keep in mind that the main focus of the pilot should be SMEs
- The project should continue as an open source project, preferential under the Apache license and be guarded by a research infrastructure like ELIXIR.

Next steps

- Additional feedback after workshop via email (based on material send to the participants)
- Adjust portal and demo based on input from the workshop
- Report on commercialisation opportunities
- Write end report
- Report to key stakeholders (we have seen in several meetings today that the federated approach of the pilot is the goal of many initiatives but hardly any have workable examples)
- Prepare for the end meeting (with a good coverage of stakeholders)

DEMO

9:25	9:50	25 min	Demonstration of the Information and Data portal - Live demo of the search functionality and selection of database preferences and features; - Demonstration of data extraction with hypothetical business use case application	TNO

Use case 1.1: Prediction based on federated data

From the market analysis:

- Two value networks: Medical and Health value network
- The medical value network is by far the biggest market
- Pharma is the most important driver for innovation

We used hypothetical examples of companies that are comparable to actual SME companies in the medical market

Btix Technologies is a(n imaginary) company which develops mobile applications for type 2 diabetes management to help them keeping their blood glucose level under control. Insulin levels are key but are difficult to analyze at home.

Question: Can we predict insulin levels based on glucose levels, if we have enough cohort data?

Use case 1.1: Insulin prediction

Based solely on glucose level that one measures at home, would it be possible to estimate insulin levels?

- This is a challenge that requires enough data for a solid prediction model

Using EASME pilot environment, we have developed the following mobile app:

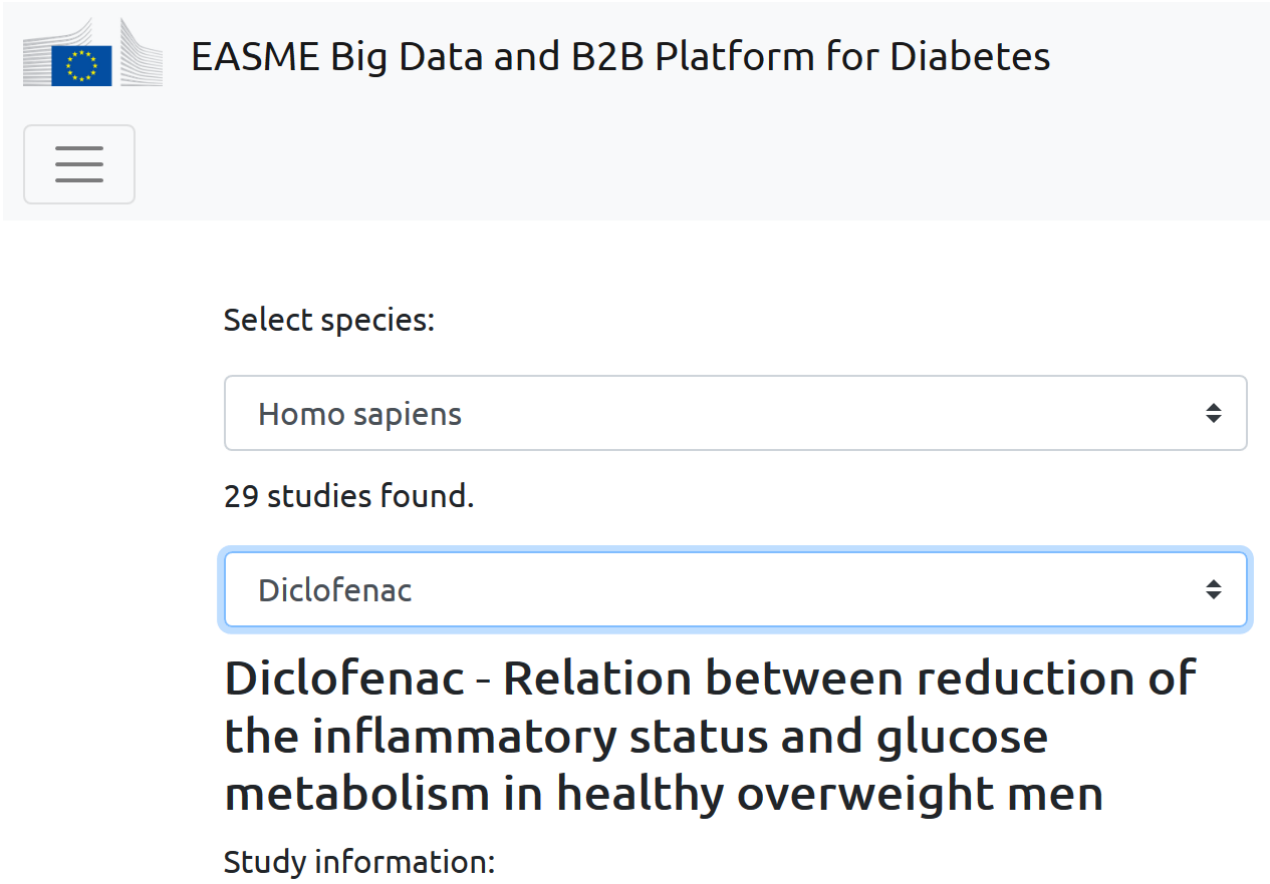
- Statistical calculations of clinically measured glucose and insulin measurements
- Machine learning models based on these clinical measurements

→ DEMO

Use case 1: Insulin prediction

The client can requests an account and we provide access and API token for them to use our platform and get access to data that are accessible for them

The client searches through the studies and finds an interesting reference study to use for the application



Use case 1: Insulin prediction

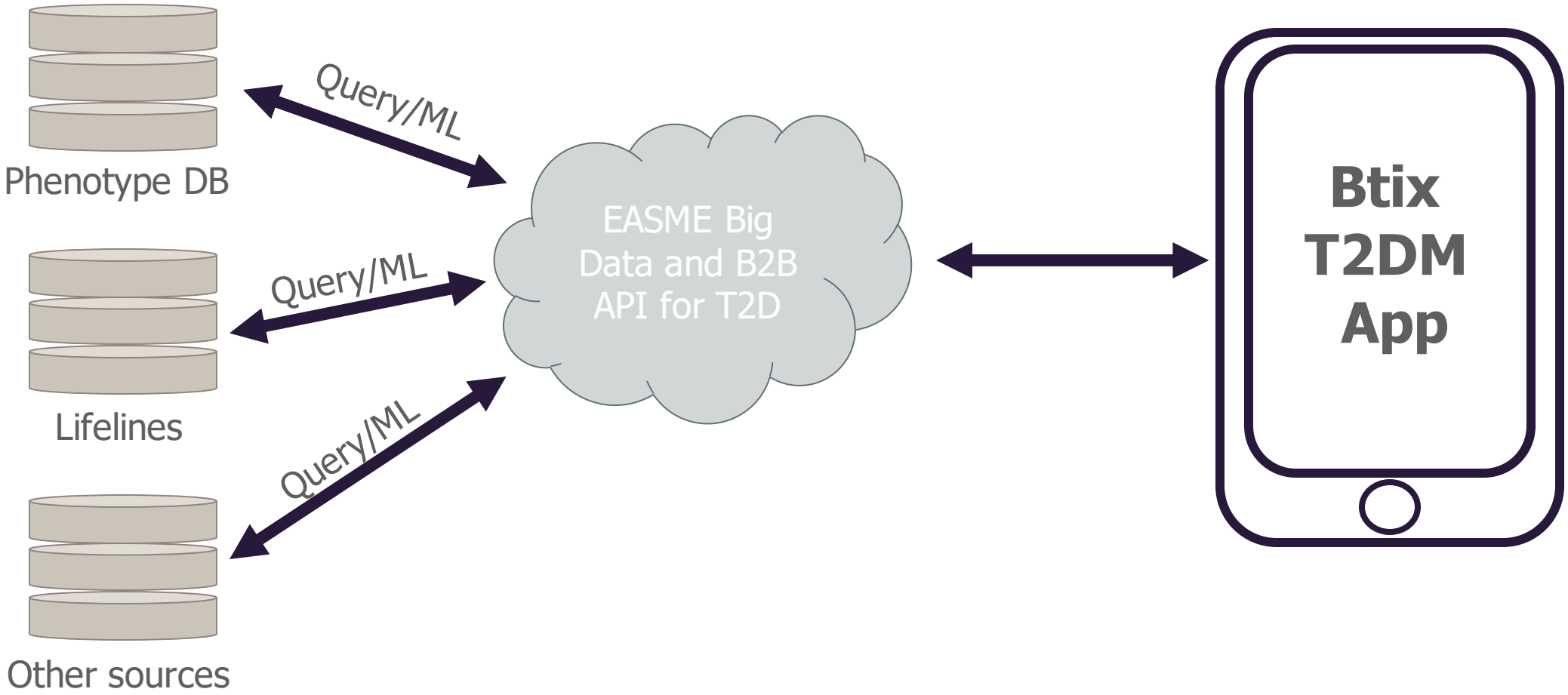
Using the following code, they embed the pilot API to their mobile application

```
getRequest = (args: Object): Promise<Response> => {
  let requestArguments = [];
  for(let [key,value] of Object.entries(args)){
    requestArguments.push(`${key}=${value}`);
  }
  let allArguments = requestArguments.join("&");
  return  fetch(`https://dashin.eu/easme/api/calculation?study_code=Diclofenac&${allArguments}`,{
    method: 'GET',
    headers: {
      authorization: 'Token [redacted]',
    }
  })
}
```

API link

Secret token provided to the client

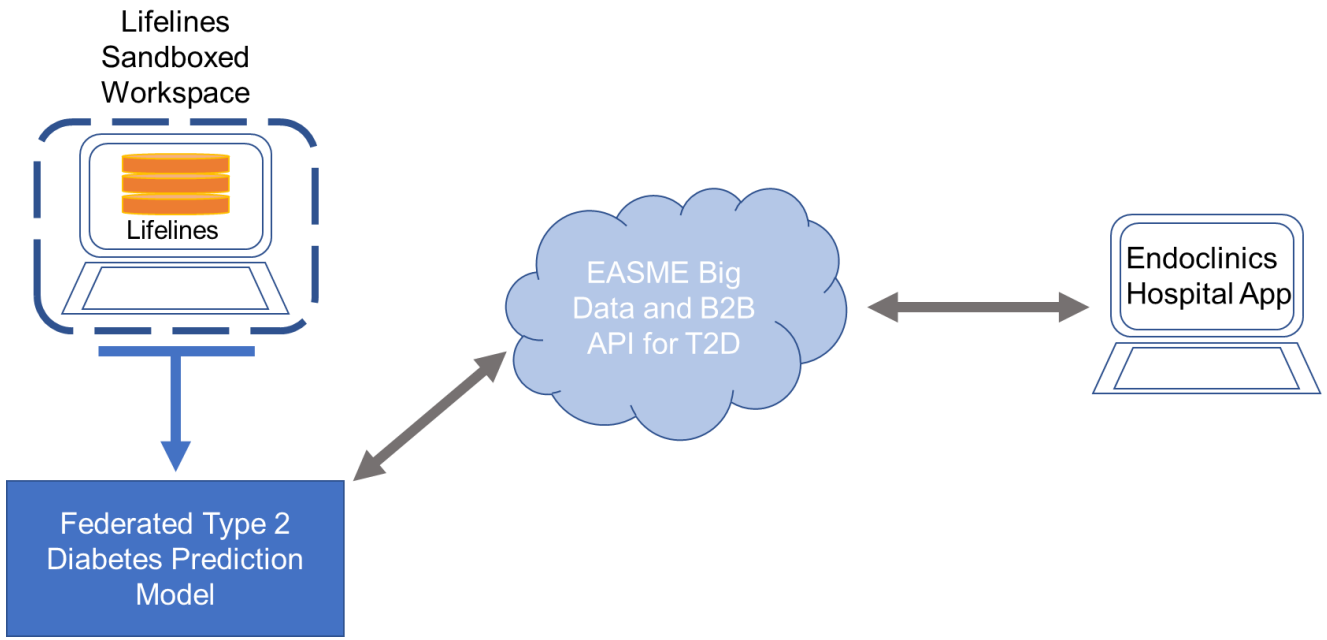
Use case 1: Insulin prediction



Use case 1.2: Prediction of diabetes onset

Endoclinics is a small (imaginary) hospital that specializes in endocrinological diseases. They would like to know if the patients are likely to develop type 2 diabetes.

In order to improve diagnosis precision, the client can make use of federated type 2 diabetes prediction model provided by this pilot. To do this, first they will obtain access to the data in the pilot. Afterwards, using their API token, they can request pretrained model parameters to make the prediction. In this way, no patient data is exposed to outside.



Data

- lifelines

  Rich cohort study, diverse information about 167k subjects from North of the Netherlands

Measurements: Anthropometrics, different proteins (lab measurements), lifestyle (questionnaires)

Diagnosis of Diabetes II: Based on questionnaire, plus levels of fasting glucose and HbA1c

2 assessments with 2-11 years in between, most around 4

Subjects used:

1. Consistently healthy

2. Healthy in the 1st assessment, Diabetic in the 2nd

Resulting clean dataset: 106k subjects, 74 measurements for each

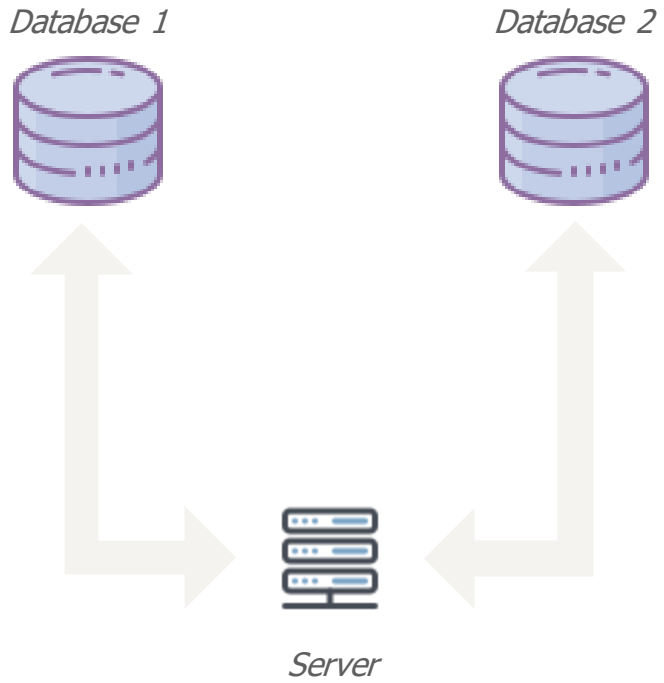
Federated model

- **Task:** Early prediction of Diabetes II onset
- **Model:** Logistic regression – probabilistic model, commonly used for Diabetes II prediction

  › Federated learning

  Branch of Machine Learning that assumes the data to be split among data bases – not centralized.

  It works by applying the model locally and combining the local models.



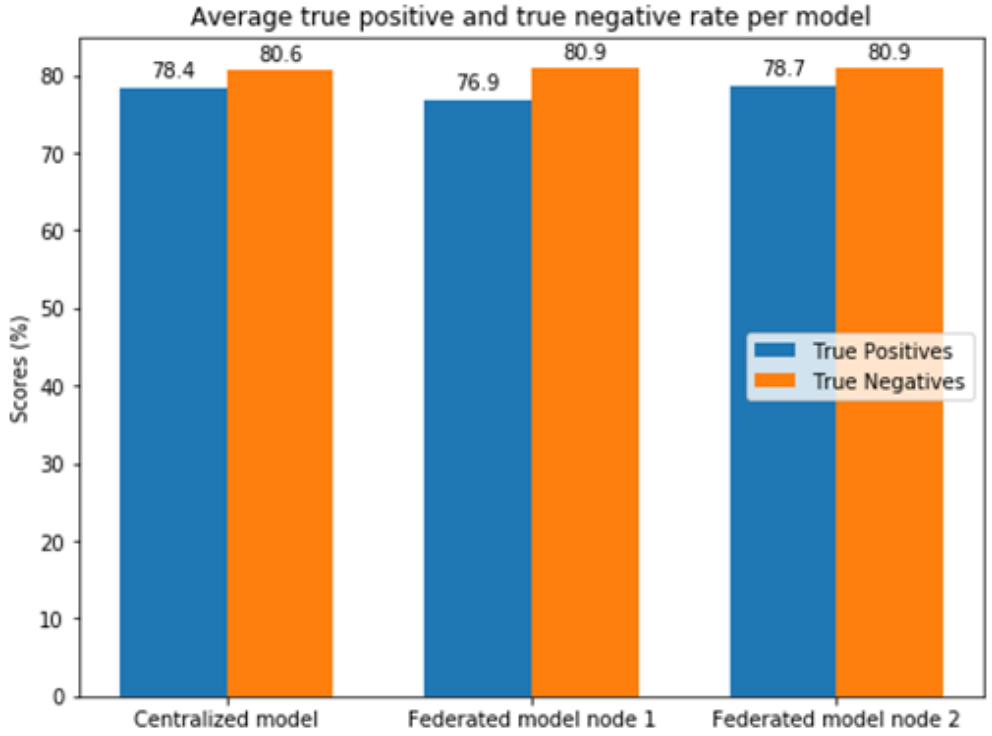
Properties:

- **Efficiency:** Removes need for centralization of increasing amount of data.
- **Privacy preservation:** No local data leaves the premises.

For our case: LifeLines data artificially split in several parts to simulate different databases, starting with two.

Performance & problems

- The federated model is able to trace around **76-79%** of the future patients in both "databases", equivalently to the centralized model
- Work in progress, but already promising given the task's novelty
  - **Obstacles**



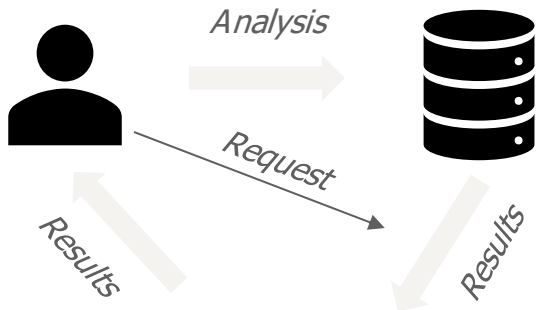
No internet connection in the workspace

Not possible to connect to other databases

Human readable data, but impossible to automate communication for real Federated Learning applications

IDEAL SET UP

LIFELINES SET UP



lifelines

Next steps

Steps planned within this project:

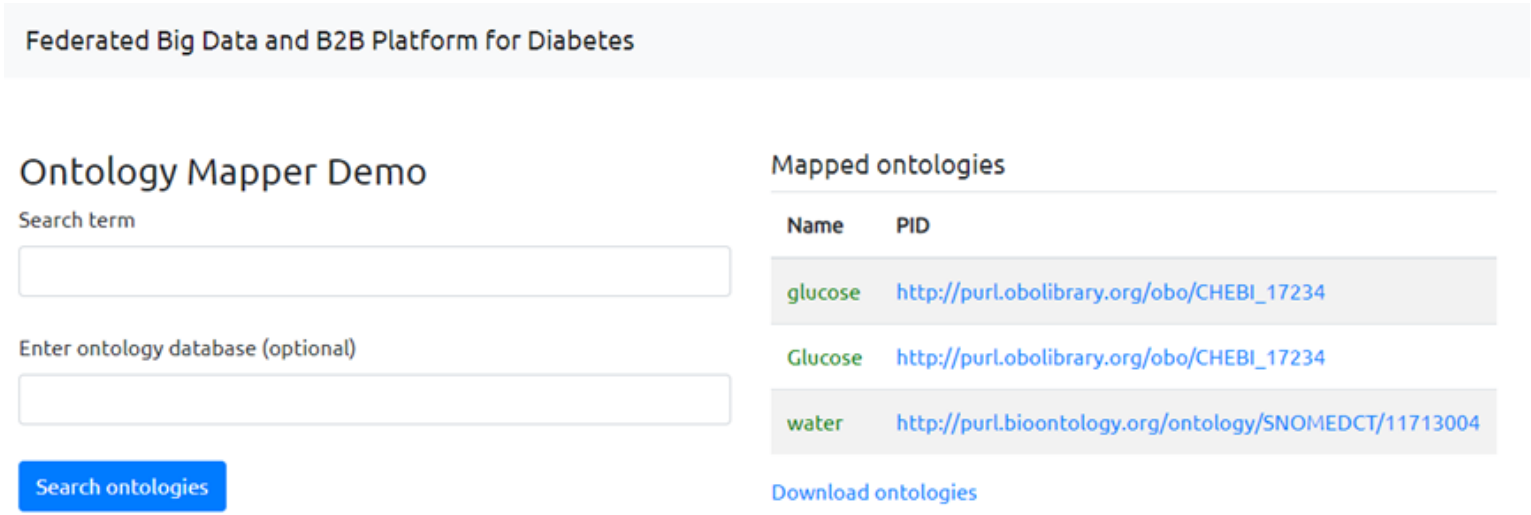
- Model development
- Application of more advanced models
- Different data divisions: more databases, imbalanced data
- Inclusion of cryptographic analysis (in the aggregation part) to minimize information leakage

Ensuring interoperability

Interoperability is a persistent challenge whenever multiple data sources are connected, regardless of how well curated they may be.

One of the takeaways from WS2 was that existing ontologies should be utilized when possible, however, this presents a challenge as all relevant terms must then be mapped to these existing ontologies, of which there are often many relevant choices.

The ontology mapper has been built to facilitate the creation of an 'ontology dictionary' that provides data terminology standardization functionality for between database interoperability.

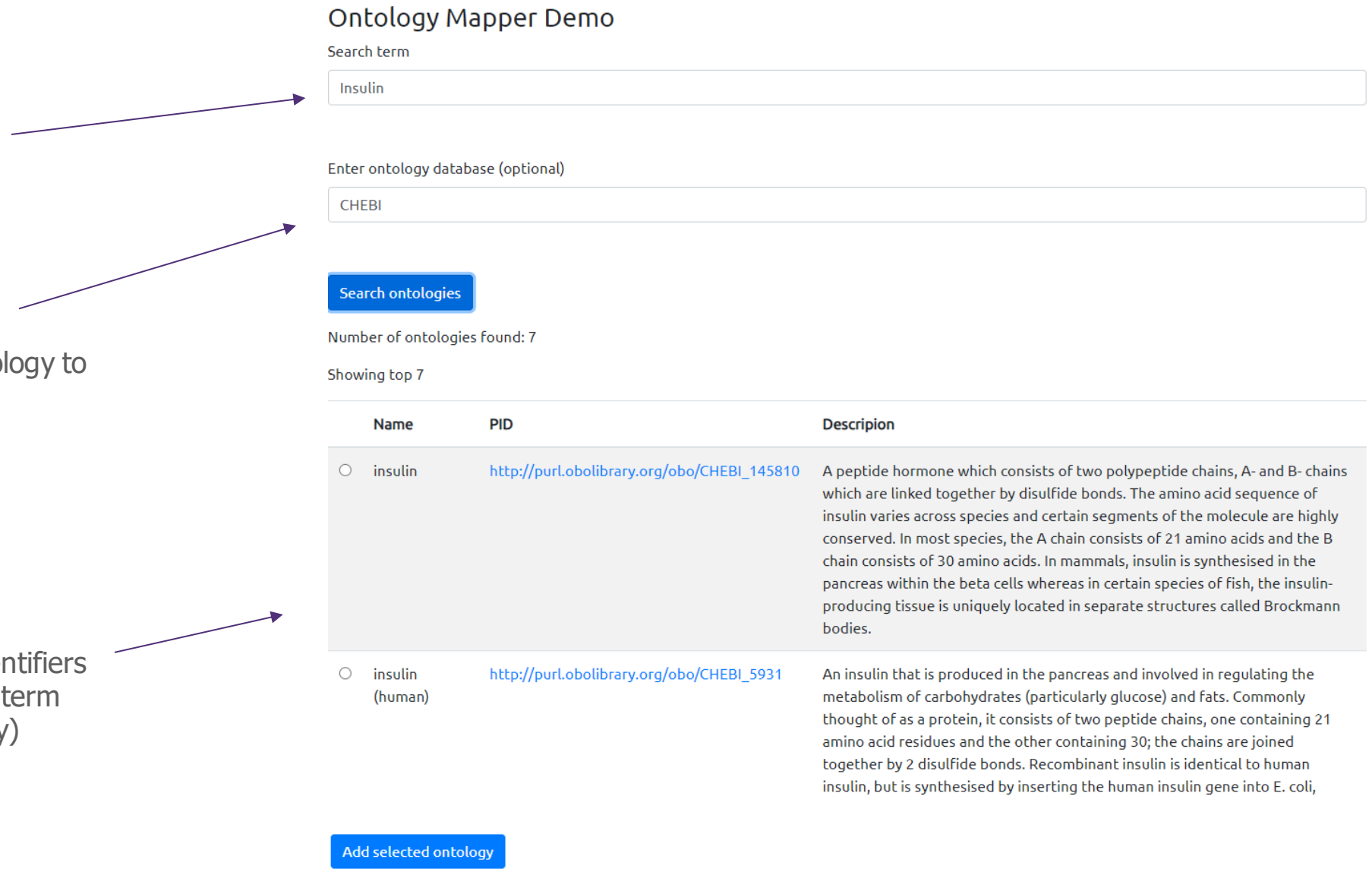


Ensuring interoperability

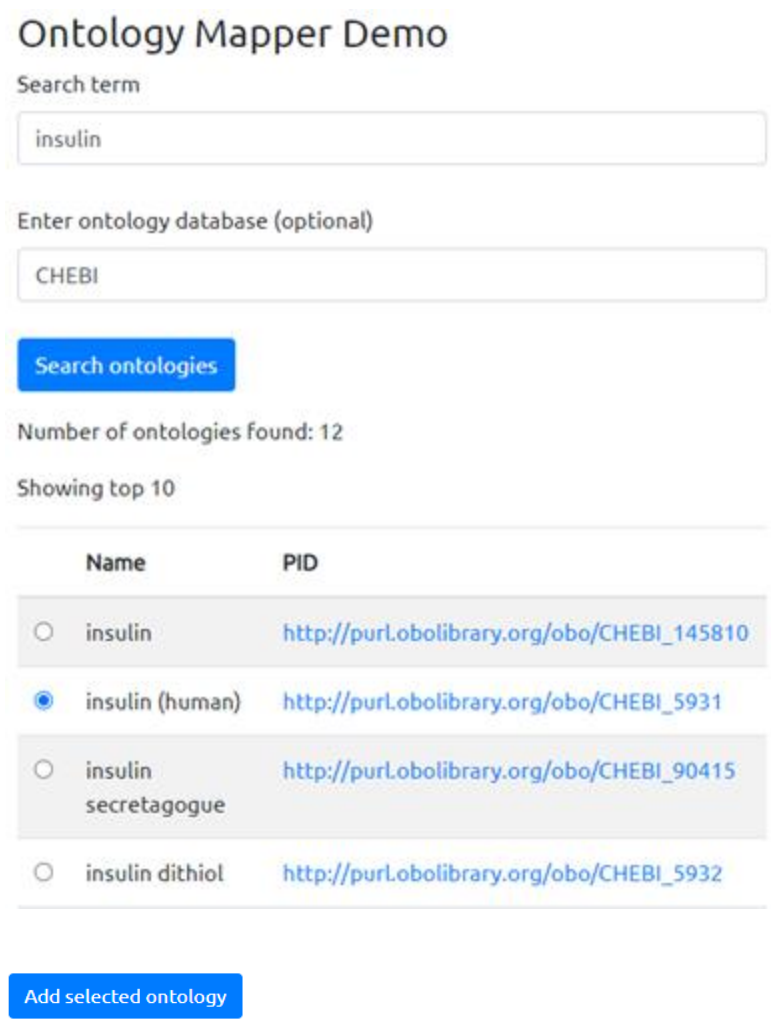
Search term to map to ontology

Optional preferred ontology to be mapped to

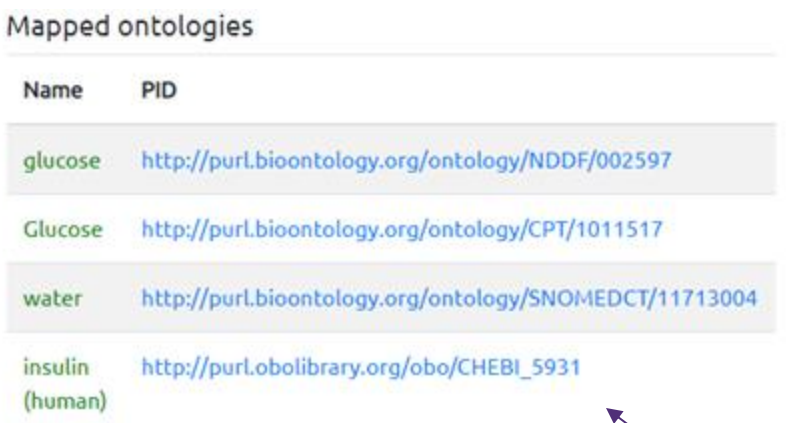
Resulting persistent identifiers associated with search term (and preferred ontology)



Ensuring interoperability



Preferred term selected from list



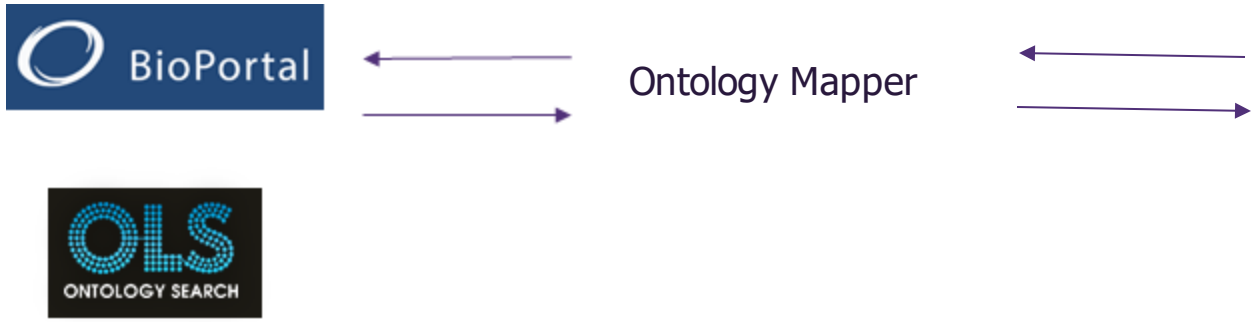
Term is then saved to Mapped Ontologies

CSV of terms to persistent identifiers is available to download directly

Ontology Mapper Architecture Overview

1. Term is sent via API to ontology service

3. Based on ontology mapper logic or user selection, term is inserted into database



Name	PID	Ontology _id	Created_ date	Modified_ date

Id	Ontology	Created_d ate

2. Response is parsed, and filtered by ontology mapper

- String matching
- Filtering based on completeness of response
- Automatic selection of ontologies by frequency

Ontology Mapper Architecture Overview

Phenotype database

Name
Creatinine
creatinine_Creatinine.v001
HDL-cholesterol.v004
HDL-cholesterol.v005
HbA1c.v000
HbA1c.v003
HbA1c.v005
BMI (body mass index).v000
BMI (kg/m2)
BodyLength.v000
BodyWeight.v002
Body_Weight

lifelines

Name	Identifier
Creatinine	Lifelines.LABDATABL.BKR.5.1.1002
HDL Cholesterol	Lifelines.LABDATABL.HDC.5.1.1002
HbA1c	Lifelines.LABDATABL.HB1C.5.1.1002
HbA1c	Lifelines.LABDATABL.HBAC.5.1.1002
Body Mass Index	Lifelines.BEZOEK1.BMI.5.1.1002
Length	Lifelines.BEZOEK1.LENGTE.5.1.1002
Weight	Lifelines.BEZOEK1.GEWICHT.5.1.1002

Ontology Mapper

Name	PID
Creatinine	http://purl.bioontology.org/ontology/MEDLINEPLUS/C0010294
Creatinine	http://purl.bioontology.org/ontology/MEDLINEPLUS/C0010294
creatinine_Creatinine.v001	http://purl.bioontology.org/ontology/MEDLINEPLUS/C0010294
HDL Cholesterol	http://purl.obolibrary.org/obo/ddo.owl#ddo_0000276
HDL-cholesterol.v004	http://purl.obolibrary.org/obo/ddo.owl#ddo_0000276
HDL-cholesterol.v005	http://purl.obolibrary.org/obo/ddo.owl#ddo_0000276
HbA1c	http://purl.obolibrary.org/obo/DDO.owl#DDO_0000243
HbA1c	http://purl.obolibrary.org/obo/DDO.owl#DDO_0000243
HbA1c.v000	http://purl.obolibrary.org/obo/DDO.owl#DDO_0000243
HbA1c.v003	http://purl.obolibrary.org/obo/DDO.owl#DDO_0000243
HbA1c.v005	http://purl.obolibrary.org/obo/DDO.owl#DDO_0000243
Body Mass Index	http://purl.obolibrary.org/obo/CMO_0000105
BMI (body mass index).v000	http://purl.obolibrary.org/obo/CMO_0000105
BMI (kg/m2)	http://purl.obolibrary.org/obo/CMO_0000105
Length	http://purl.obolibrary.org/obo/PATO_0000122
BodyLength.v000	http://purl.obolibrary.org/obo/PATO_0000122
Weight	http://purl.obolibrary.org/obo/PATO_0000128
BodyWeight.v002	http://purl.obolibrary.org/obo/PATO_0000128
Body_Weight	http://purl.obolibrary.org/obo/PATO_0000128

Breakout 1

9:50	10:25	35 min	Break-outs: sessions on specific questions on the usability of the portal	TNO

Links to the mentimeter questionnaires (www.menti.com) session1

Group1: 29 40 32

https://www.menti.com/qh1gw7t3xt

Group2: 41 15 24

https://www.menti.com/rgr4i3gv3b

Group3: 14 73 8

https://www.menti.com/gufanosec3

Questions (work in progress) '**usability of the portal**'

Q1 What type of end-users are you? (scientist, developer, policy maker, other) more than one answer is allowed

Q2 Can you describe the way you would reuse data? (access via a portal, mobile phone, API, other) more than one answer is allowed

Q3 Would you like access to the data via a specific analysis tool? (R, Python, SAS, SPSS, excel, other) more than one answer is allowed

Q4 What health data are you searching for/ are valuable to you? (cohort study data, intervention study data, real-world medical data, personal real-world data (incl. wearables), other)

Q5 What type of quality checks do you want to have provided? (study description meta-data, contextual social data, number of missing data, ethical reuse possible, other) more than one answer is allowed

Q6 Do you miss data repositories? (Yes, no) if yes, please specify

Q7 Did you find new data repositories? (Yes, no) if yes, please specify

Q8 What is lacking (open question)?

Breakout 2

11.10	**11.40**	**30 min**	**Break-outs: Discussion on future solutions for data reuse (centralized/decentralized)**	**TNO**

Links to the mentimeter questionnaires (www.menti.com) session 2

Group1: 86 19 75

https://www.menti.com/hkgb3sfwha/0

Group2: 37 57 00

https://www.menti.com/wh5y3r26ji/0

Group3: 31 67 58

https://www.menti.com/xxvywjfcmv/0

Questions (work in progress) '**future solutions for data reuse (centralized/decentralized)**'

Q1 Would you use the solution presented by Sitra? (yes, scientifically; yes, for business, yes, other, no) ANDRÉ

Q2 Is there already an initiative started on a sharing health data solution in your country? (Yes, no) if yes, please specify

Q3 At what level should data be integrated in a federated way? (not at all, on country level, at database level, at an individual level)

Q4 Who should take responsibility for full implementation of a health data solution for science in Europe? (EU, world business, EU business, national politics, other)

Q5 Which policies need to change for full exploitation of health data? (removal of the GDPR, extension of the GDPR, adjustments of the GDPR, other)

Q6 Can we afford to do nothing and leave it up to Google, Amazon, Facebook, Palantir? (Yes, no) if yes, please specify

Q7 Could anonymization of data be the solution to circumvent the GDPR? (No, that will not allow me to do full analysis; No, then you still have the chance of reidentification; Yes, this is the way forward; Yes, we need better guidance on legal options; Yes, we need better guidance on the technical solutions (you can choice more than one))

Q8 What roles in the market place?

Q9 What would be the ultimate solution for data reuse in the future? (open question)

Breakout 3

12:20	12:40	20 min	Break-outs: Opportunities for Commercialisation	TNO

Links to the mentimeter questionnaires (www.menti.com) session 3

Group1: 13 81 37

https://www.menti.com/jnh437q97v

Group2: 85 13 43

https://www.menti.com/zdtixniy4r

Group3: 41 22 49

https://www.menti.com/mjxqdte3oh

Questions (work in progress) '**Opportunities for Commercialisation**' – 20 min/

Q1 Encourage national health authorities to develop reimbursement models for health apps

(Not important – Fairly unimportant – Neither important nor unimportant – Fairly important - Very important)

Q2 Research demonstrating the benefits of T2D lifestyle apps in partnership with patent advocacy organisations should be supported.

(Not important – Fairly unimportant - Neither important nor unimportant – Fairly important – Very important)

Q3 What is the importance of a privacy label for apps and services?

(Not important – Fairly unimportant - Neither important nor unimportant – Fairly important - Very important)

Q4 Who should take responsibility for implementation of a health data solution for business in Europe? (EU, world business, EU business, national politics, other)

Thank you