Data harmonization in diverse datasets

Václav Papež, Spiros Denaxas, Harry Hemingway
Institute of Health Informatics
University College London, UK
http://denaxaslab.org

Maxim Moinat, Stefan Payrable
The Hyve, NL
https://thehyve.nl

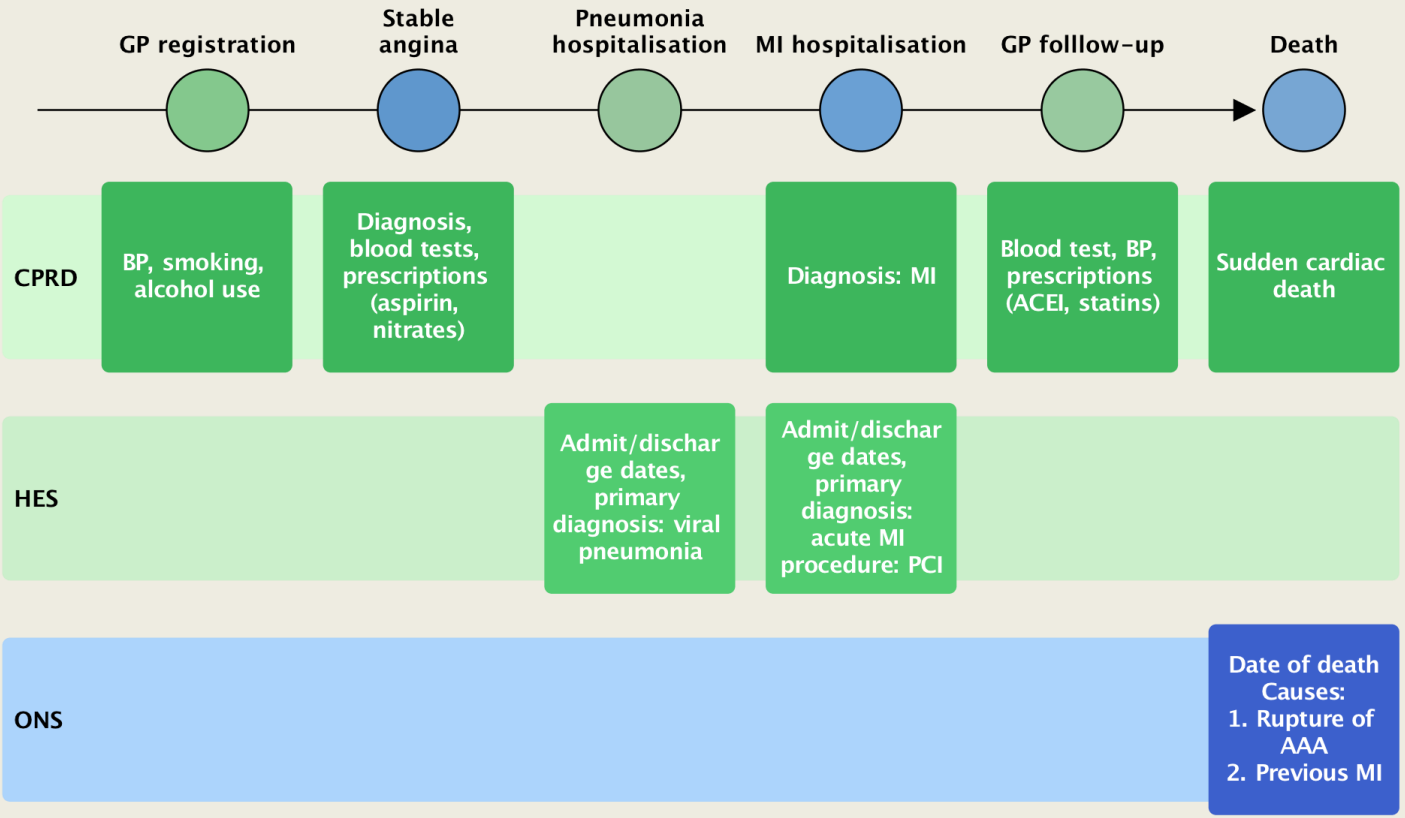
Overview

- CALIBER data resource
- Harmonization of Data Storage (OMOP CDM)
- Harmonization of Phenotyping Algorithms (Semantic Web Technologies)
- Results

CALIBER Data Resource

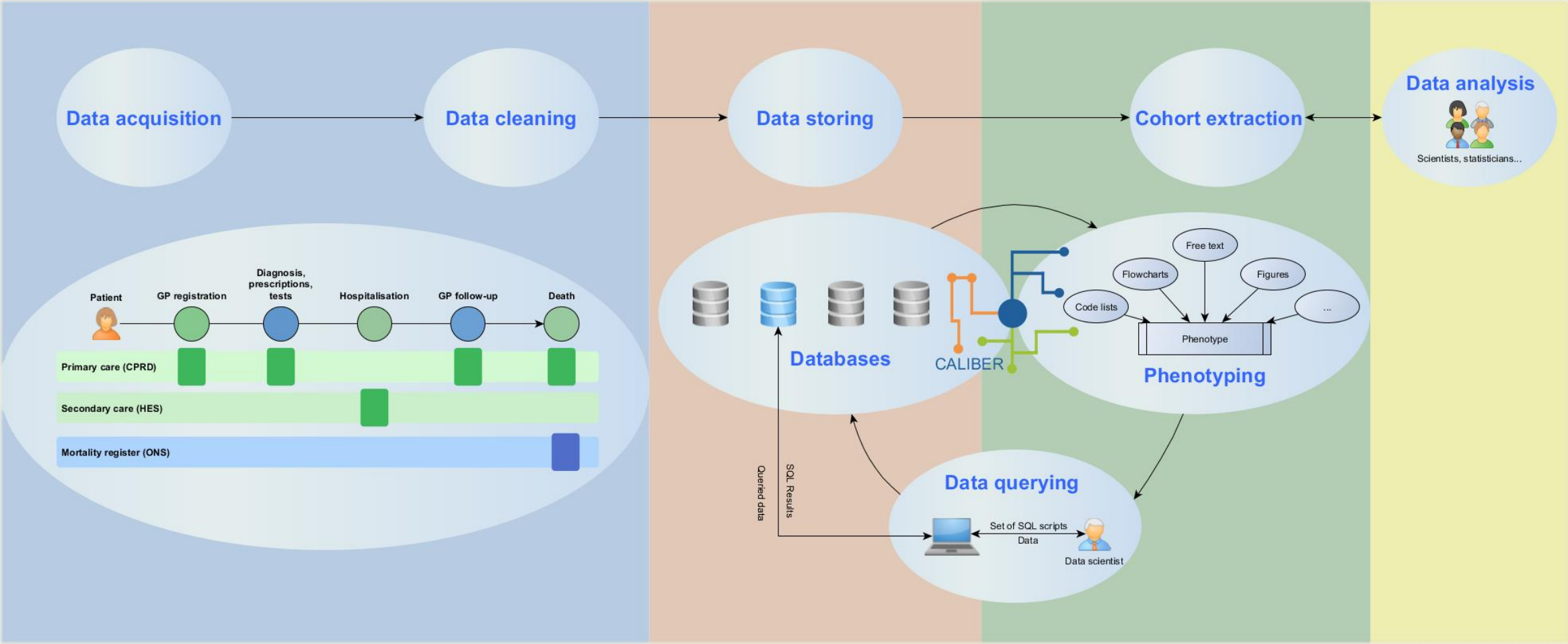
CALIBER

- Translational research platform linking national structured data and socioeconomic information from
  - primary care (CPRD)
  - hospital care (HES)
  - mortality registry (ONS)

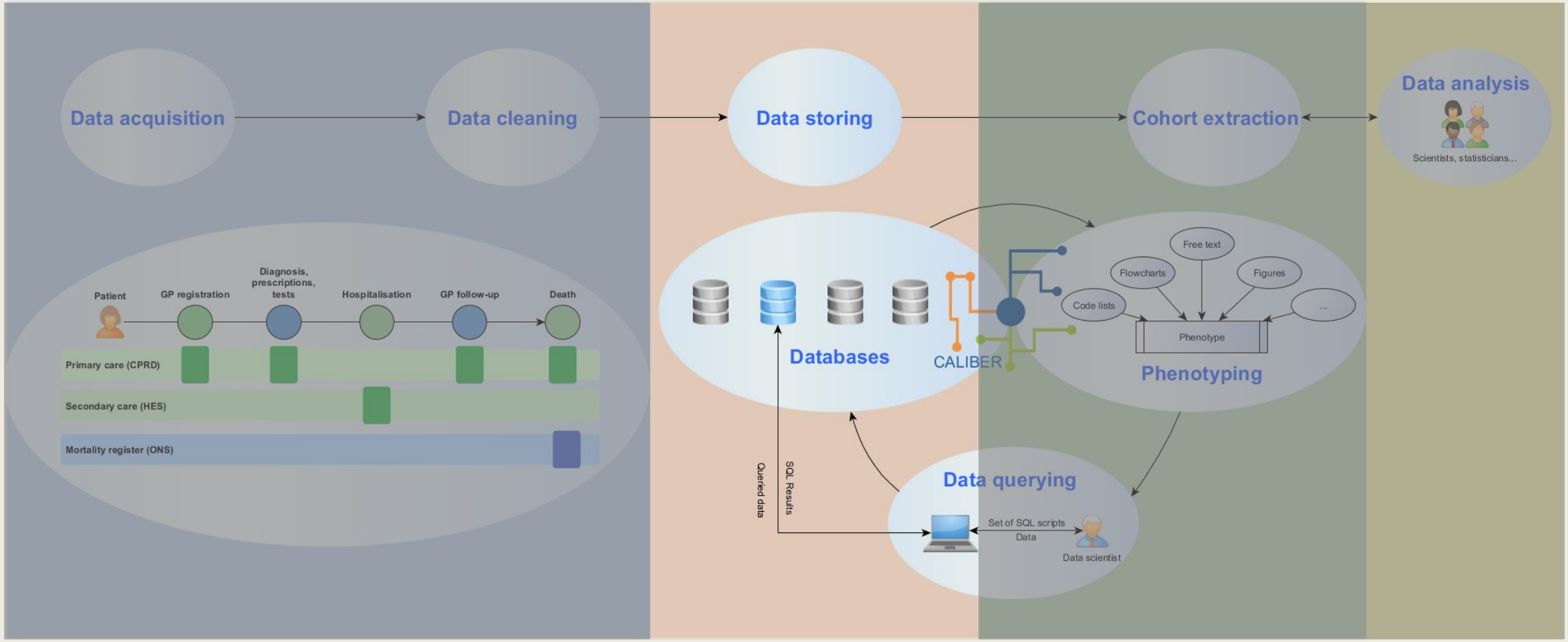


Denaxas S. et al., Int J Epidemiology, 2013, doi: 10.1093/ije/dys188

Linked EHR workflow



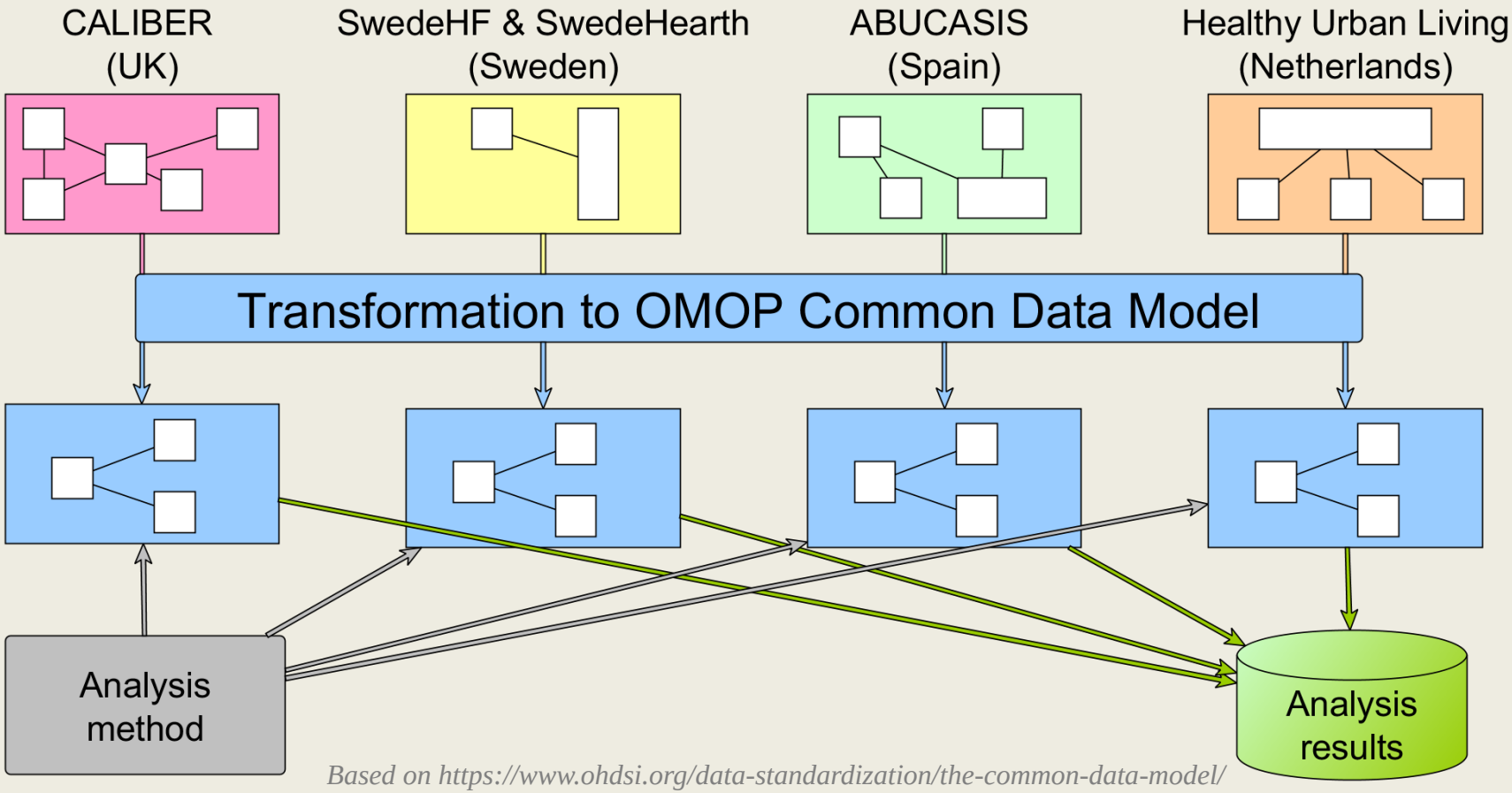
Harmonization of diverse data storages



The work was realized in cooperation with The Hyve, Utrecht, NL, https://thehyve.nl

Motivation and Project Challenges

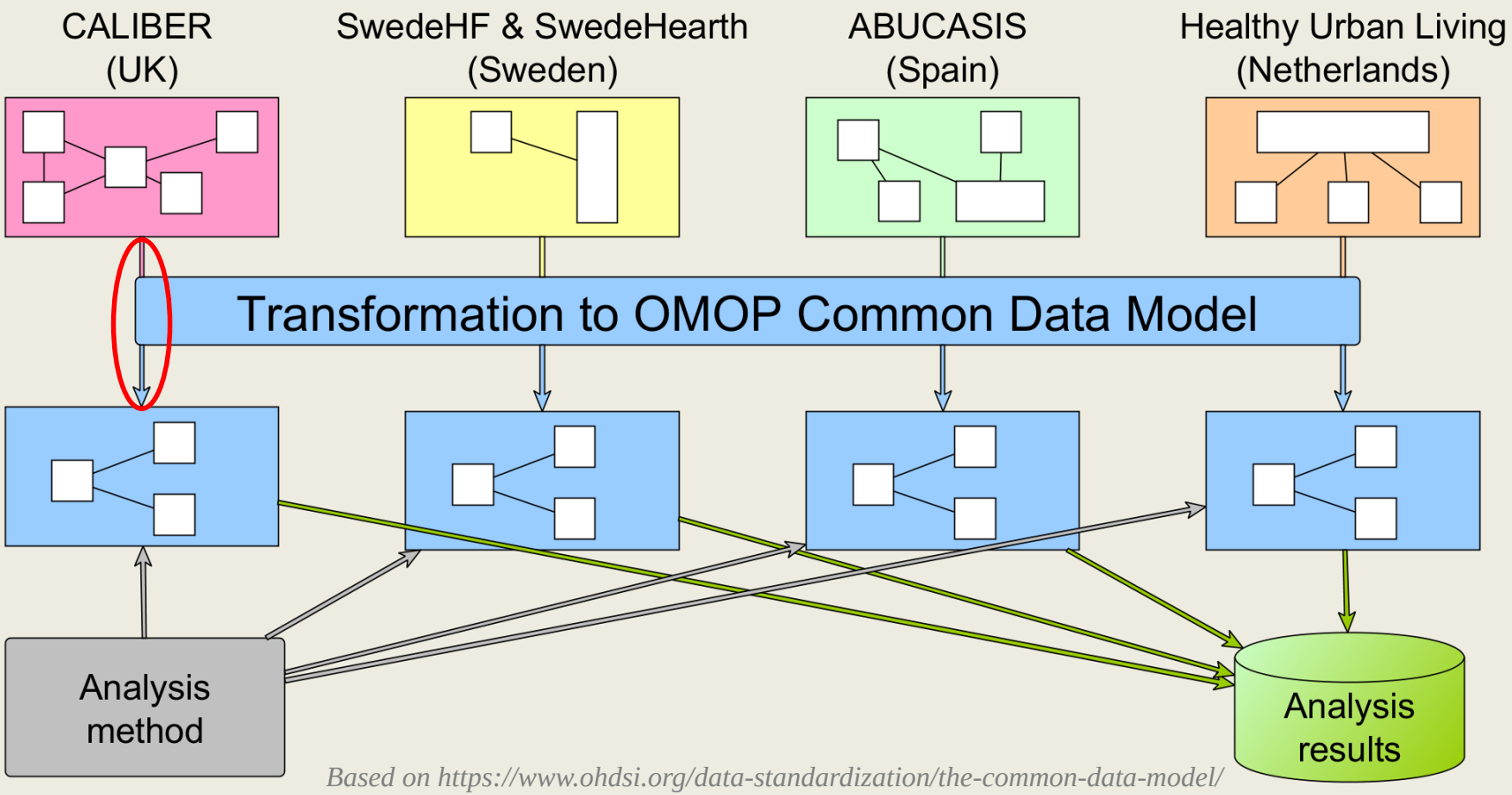
- IMI BigData@Heart project
  - Compare Heart Failure survival



Based on https://www.ohdsi.org/data-standardization/the-common-data-model/

Motivation and Project Challenges

- IMI BigData@Heart project
  - Compare Heart Failure survival



Based on https://www.ohdsi.org/data-standardization/the-common-data-model/

Goals and Objectives

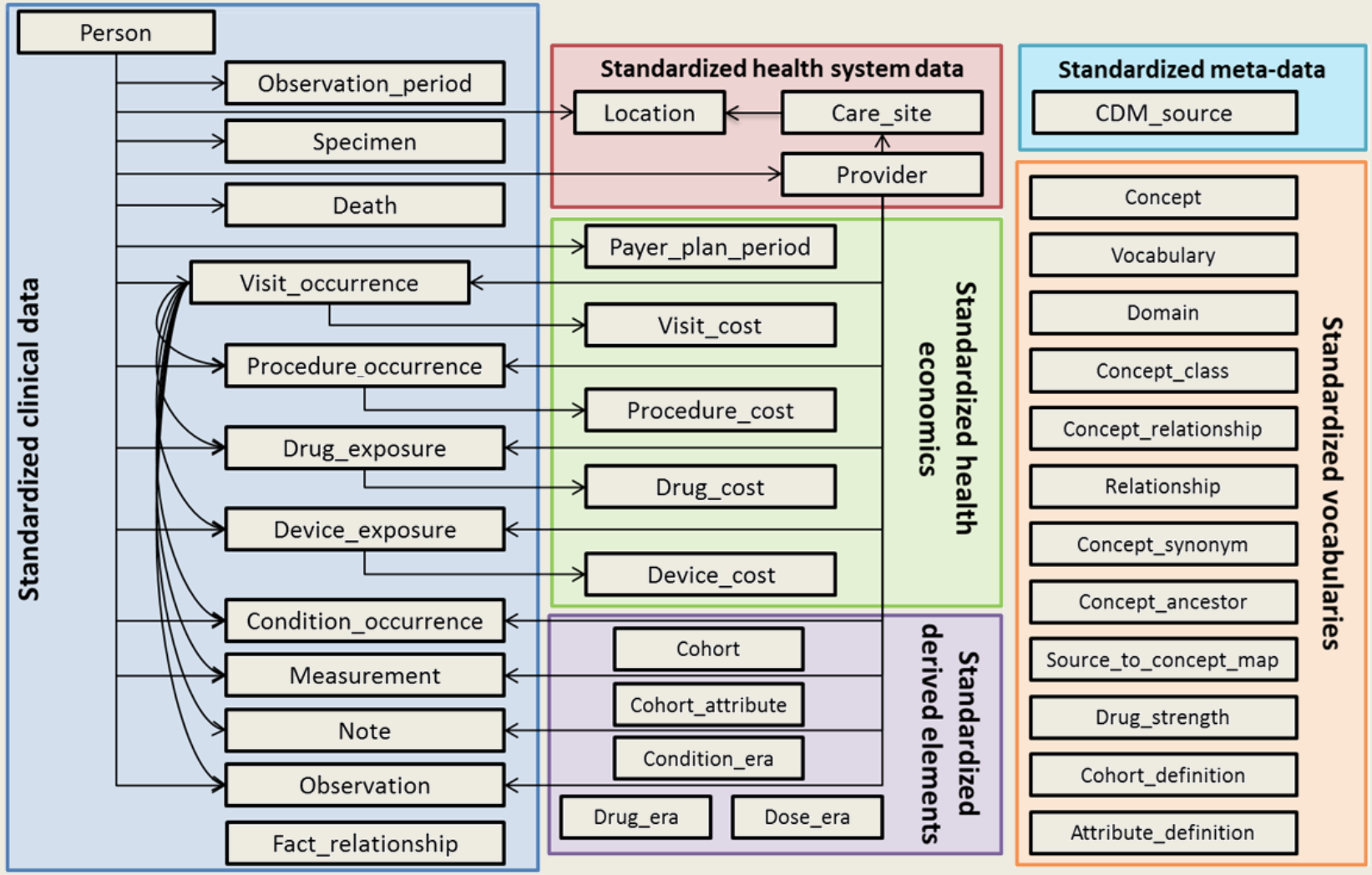
- High quality mapping of CALIBER data source into OMOP CDM
  - To develop an automatic mapping process from CALIBER to OMOP
  - To use the OHDSI tools for data quality assessment
  - To asses the vocabulary mapping quality
  - To use an ATLAS tool for data source exploration, cohort definition, etc.

CALIBER Challenges

- Diverse clinical term coding (READ codes, ICD10, ICD9, OPCS4, Product codes, etc.)
- Diverse recording practice across primary care, secondary care and ONS

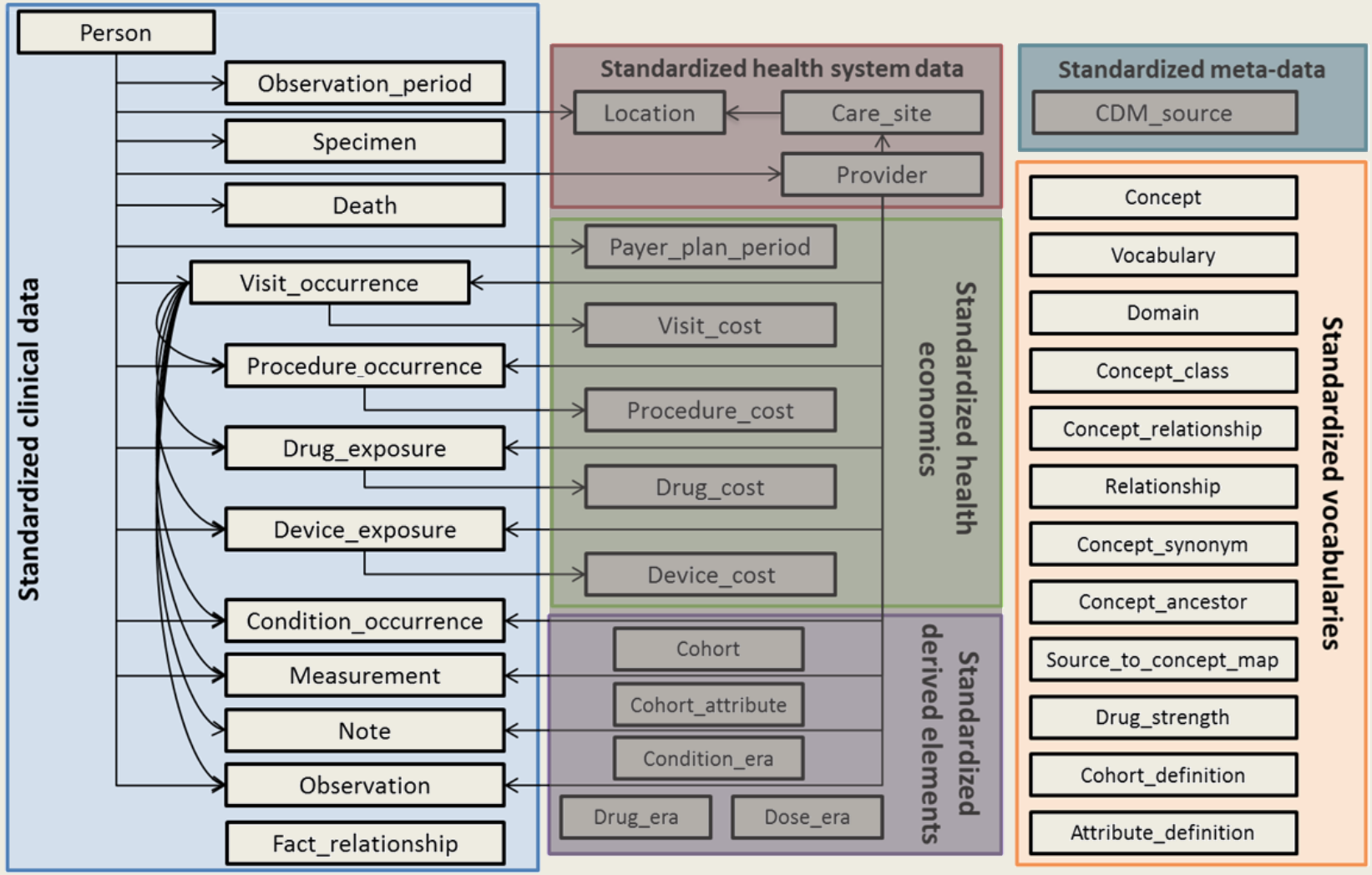
OMOP Common Data Model (v5)

- For systematic analysis of disparate observational databases
- OMOP CDM developed by Observational Health Data Science and Informatics community (OHDSI) together with software tools compatible with OMOP CDM
- Increasing trend in adopting OMOP Common Data Model in Europe



OMOP Common Data Model (v5)

- For systematic analysis of disparate observational databases
- OMOP CDM developed by Observational Health Data Science and Informatics community (OHDSI) together with software tools compatible with OMOP CDM
- Increasing trend in adopting OMOP Common Data Model in Europe



Conversion process

Syntactic mapping

Table → Table(s)

Source	CDMv5.2.0
patient	person
practice	observation_period
consultation	visit_occurrence
clinical	condition_occurrence
referral	procedure_occurrence
test	drug_exposure
immunisation	measurement
therapy	observation
ons_death	device_exposure
ons_imd	death
hes_patient	care_site
hes_diag_hosp	provider
hes_proc_epi	
hes_diag_epi	
hes_op_clinical	

Table → Table(s)

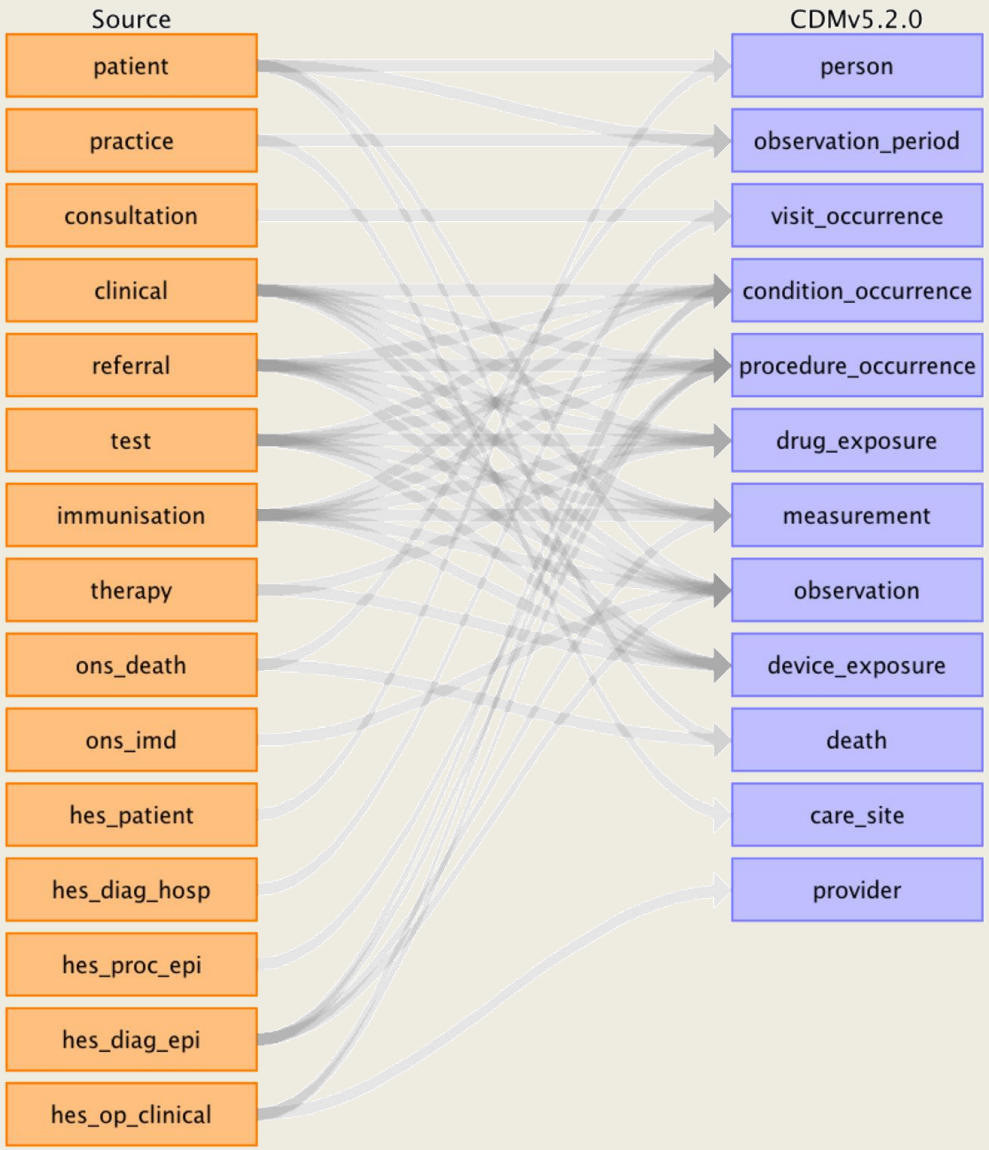


Table → Table(s)

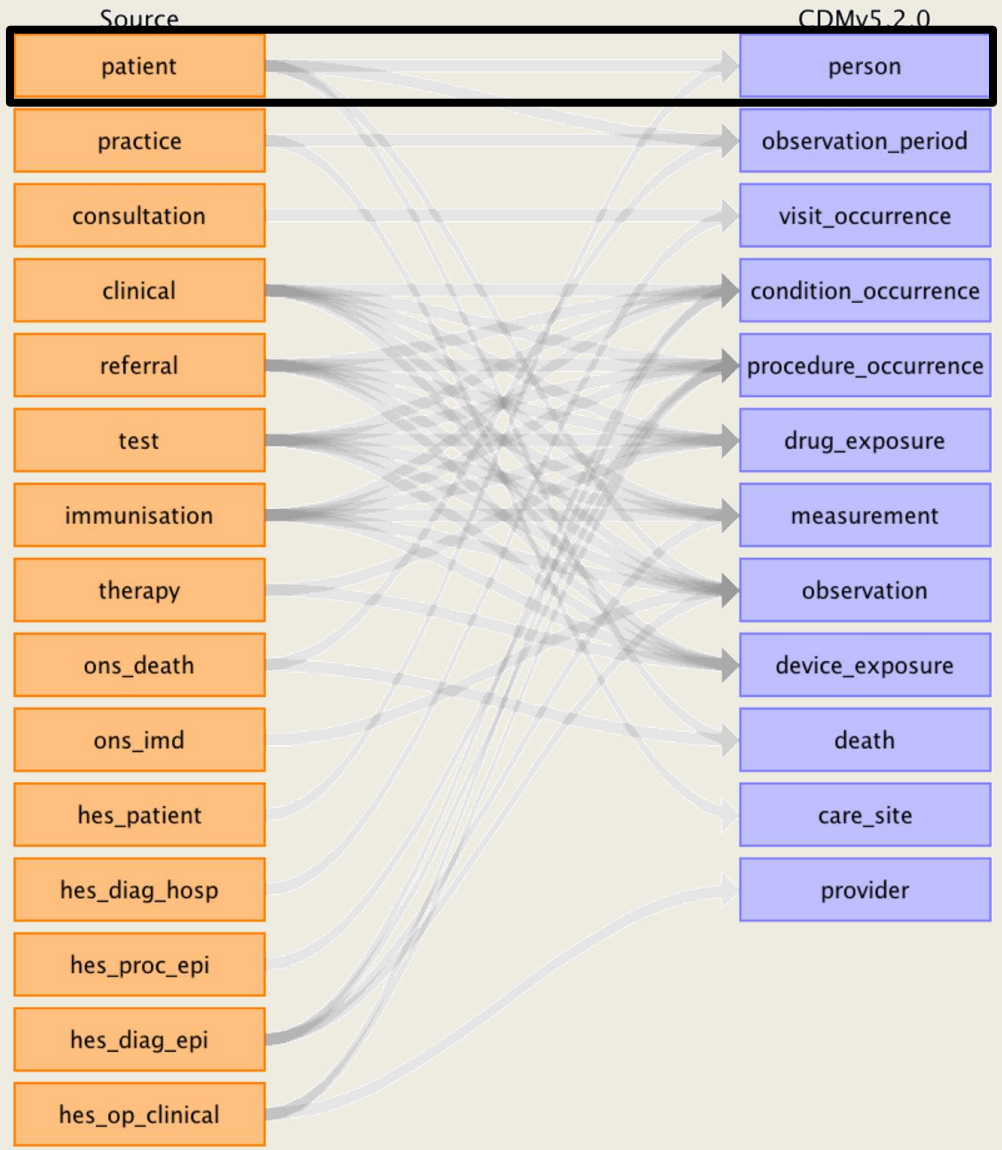
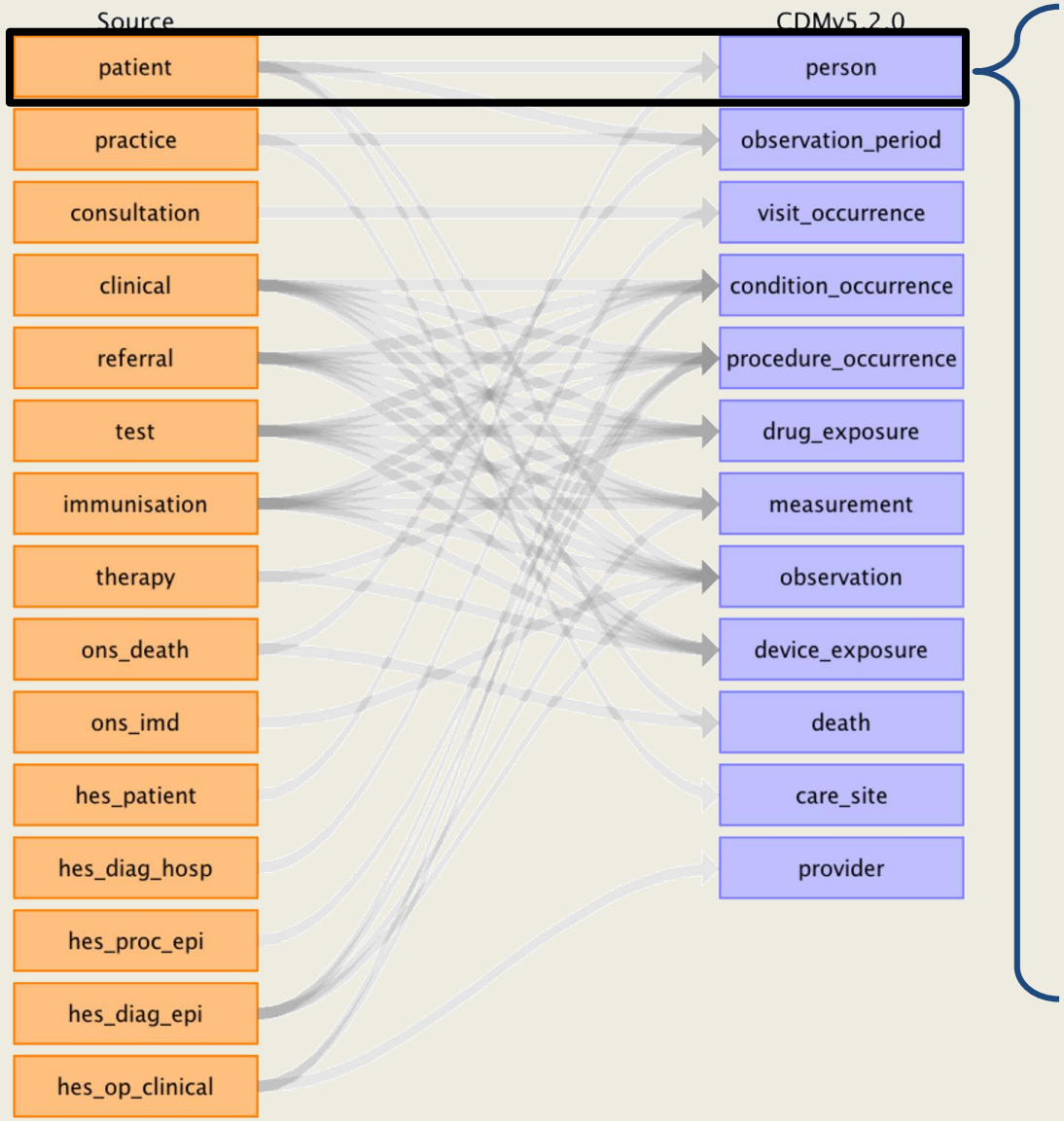
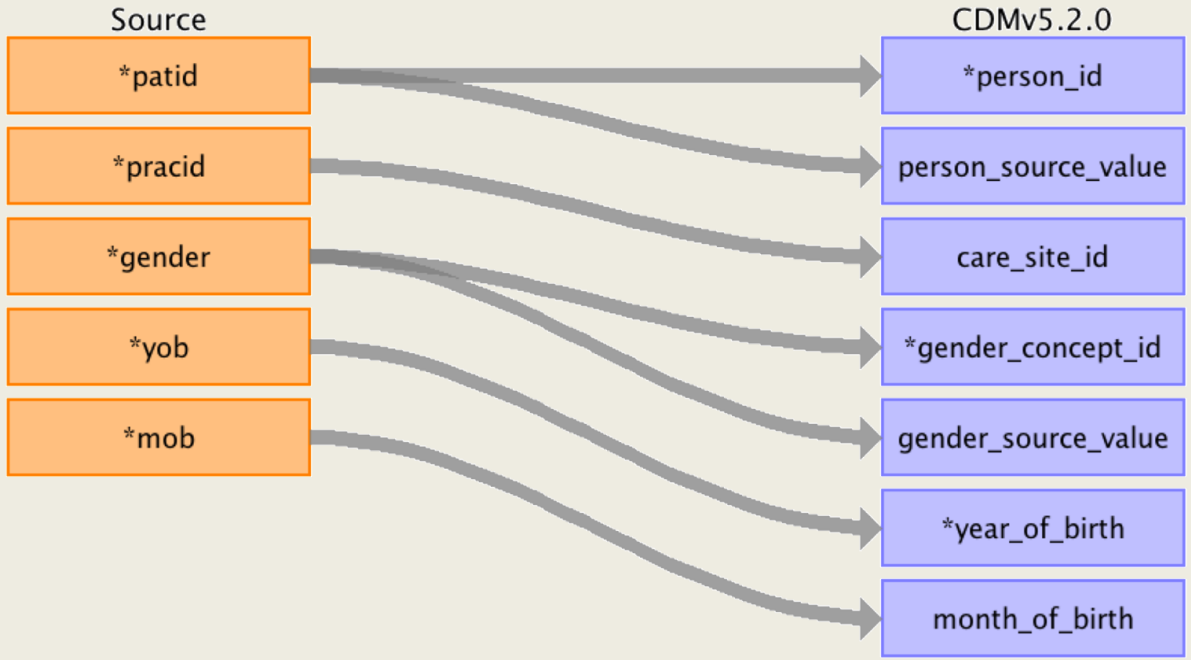


Table → Table(s)



Column → Column(s)



Conversion process

Semantic mapping

Source codes mapping

- Internal mapping
  - READ codes -> SNOMED CT

Type 1 diabetes mellitus	READ	Concept ID	Concept ID	SNOMED CT
	C108.12	45420112	20125	46635009

  - ICD10 -> SNOMED CT

Dysthymia	ICD10	Concept ID	Concept ID	SNOMED CT
	F34.1	45586238	433440	78667006

  - CPRD Units -> UCUM

mmol/L	CPRD unit	Concept ID	Concept ID	UCUM
	96	2000068400	8753	mmol/L

Source codes mapping

- External mapping
  - CPRD Product codes -> RxNorm
    - Via gemscript and dm+d

Simvastatin 10mg tablets	CPRD product code	Concept ID	gemscript	dm+d	Concept ID	RxNorm
	42	2000035557	72488020	319996000	1539463	314231

  - CPRD Entity types -> LOINC
    - Via JNJ_CPRD_ET_LOINC

Examination findings -Blood pressure	CPRD Entity type	Attributes	Concept ID	JNJ_CPRD_ET_LOINC
	1	Diastolic, Systolic and 5 more	2000068426, 2000068406	1-1, 1-2
			Concept ID	**LOINC**
			3004249, 3012888	8480-6, 8462-4

Conversion process

Verification

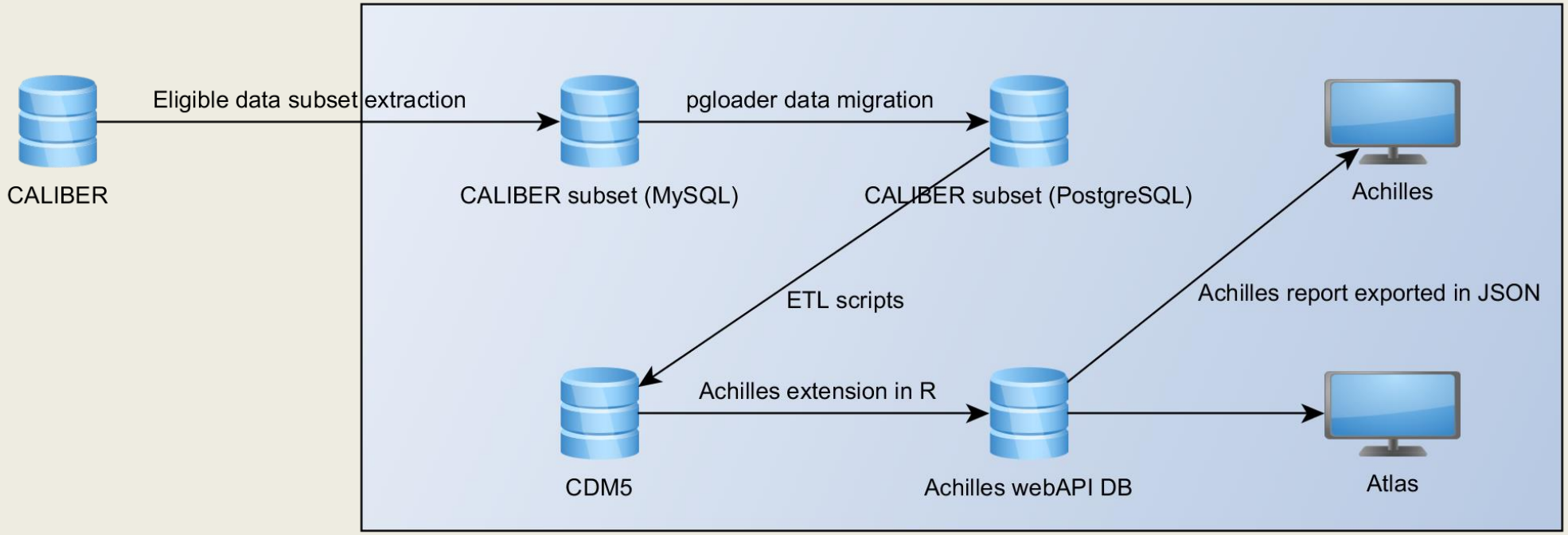
Mapping verification

- ACHILLES and ACHILLES HEEL tools
  - Quality data assessment
  - Mapping statistics
- Manual validation of top 100 mapped and unmapped terms
- Verification on predefined set of metrics
  - Direct SQL querying into CALIBER
  - Direct SQL querying into OMOP CDM
  - Desingnig of ATLAS cohorts

Results

Mapping environment

- Iterative ETL development (The Hyve) and script validation (UCL)
- Virtual Environment for processing CALIBER data (UCL)



CALIBER2OMOP Virtual Environment

Vocabulary Mapping Coverages

- 99% of the source codes mapped to a valid OMOP concept ID

Mapping	No. of source codes	No. of target concepts	Number of mapped rows	Coverage %
Condition	10889	8347	582814	100
Procedure	4252	3266	242731	100
Device	2189	2172	62743	100
Measurement unit	147	103	1455053	99.7
Observation unit	30	28	1954	98.9
Measurement	676	574	1998124	98.9
Drug	9301	5534	1708273	91
Observation	9867	7825	1949067	72.4

Metrics

Characteristics		**Derivation cohort (n=10k)**	**OMOP cohort (n=10k)**
Men / Women		4851 / 5169	4851 / 5149
Mean age (years) / Median BMI		39.32 / 26.8	39.32 / 26.8
Fasting blood glucose recorded		1700	1702
Smoking status			
	Current Smokers	1847	1847
	Ex smoker	2796	2796
	Non-smokers	5638	5657
Medical characteristics			
	Family history of diabetes	345	346
	Hypertension monitoring	930	931
	Gestional diabetes	11	11
Current drugs			
	Simvastatin	1259	1258
	Atypical antipsychotics	148	148
	Topical Corticosteroids	**1785**	**1615**

Metrics

Characteristics		**Derivation cohort (n=10k)**	**OMOP cohort (n=10k)**
Men / Women		4851 / 5169	4851 / 5149
Mean age (years) / Median BMI		39.32 / 26.8	39.32 / 26.8
Fasting blood glucose recorded		1700	1702
Smoking status			
	Current Smokers	1847	1847
	Ex smoker	2796	2796
	Non-smokers	5638	5657
Medical characteristics			
	Family history of diabetes	345	346
	Hypertension monitoring	930	931
	Gestional diabetes	11	11
Current drugs			
	Simvastatin	1259	1258
	Atypical antipsychotics	148	148
	Topical Corticosteroids	**1785**	**1615**

Smoking status

- Incompatible phenotype definitions

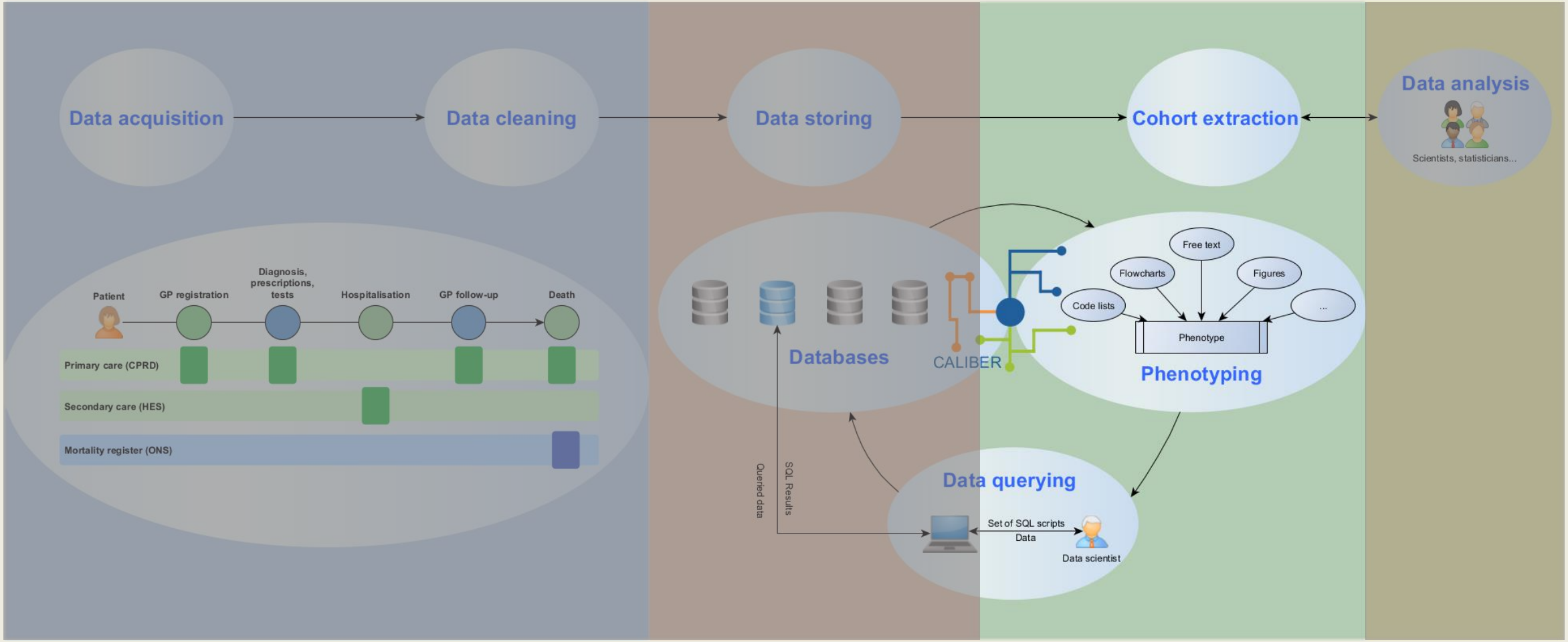
CALIBER Smoking status	SNOMED mapping in OMOP
Smoker	Smoker
Non-smoker	Non-smoker
Ex-smoker	Ex-smoker
Nicotine dependence	Cigarette smoker
Conflict: Ex and non-smoker	Current non-smoker
Conflict: Non and current smoker	Moderate cigarette smoker
Conflict: Ex and current smoker	Passive smoker
	Pipe smoker
	Aggressive ex-smoker
	…

Smoking status

- Incompatible phenotype definitions

Smoking status	CALIBER	OMOP
Current smoker	3053	2361
Non-smoker	5572	5613
Ex-smoker	2370	2316
Conflict: Ex and current smoker	1420	0
Conflict: Non and current smoker	1074	0
Ex or current smoker	4	4

Harmonization of Phenotyping Algorithms



Motivation

- No commonly-accepted machine-readable format for Computable Definitions of Electronic Health Records Phenotyping Algorithms

EHR Phenotyping

- Computational algorithms identifying patients diagnosed with particular conditions using EHR data elements (diagnosis, laboratory tests, symptoms, clinical examination findings, prescriptions etc.)
- Phenotype
  - Implementation logic
  - External data features (text, imaging, other)
  - Unstructured features (lab values, prescriptions)
  - Structured features (Controlled clinical terminologies)

Morley K. et al, PLOS ONE, doi: 10.1371/journal.pone.0110900

Challenges

- No commonly-accepted machine-readable format
- Manual translation from definition to machine code
- Reusability: difficult to share/externally validate algorithms
- Backwards compatibility due to evolving ecosystem

Computable EHR phenotyping desiderata

- Human-readable and computable representations
- Set operations/relational algebra
- Structured and temporal rules
- Standardized clinical terminologies and reusability
- Interfaces for external software algorithms
- Backwards compatibility

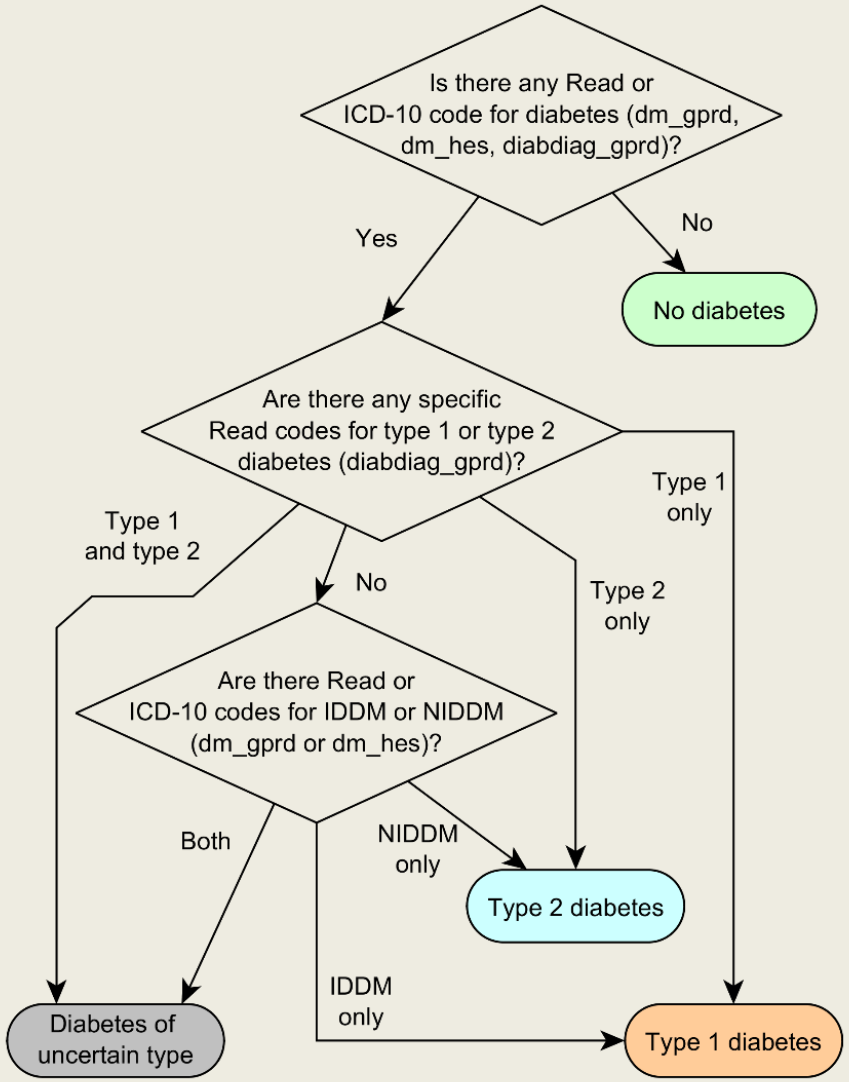
Mo H. et al, JAMIA, doi: 10.1093/jamia/ocv112

Goals and Objectives

- Investigate how Semantic Web Technologies can address these challenges
- Explore RDF and OWL for storing machine-readable EHR phenotyping algorithms
- Evaluate against desiderata developed by Mo at al.

Case study: diabetes

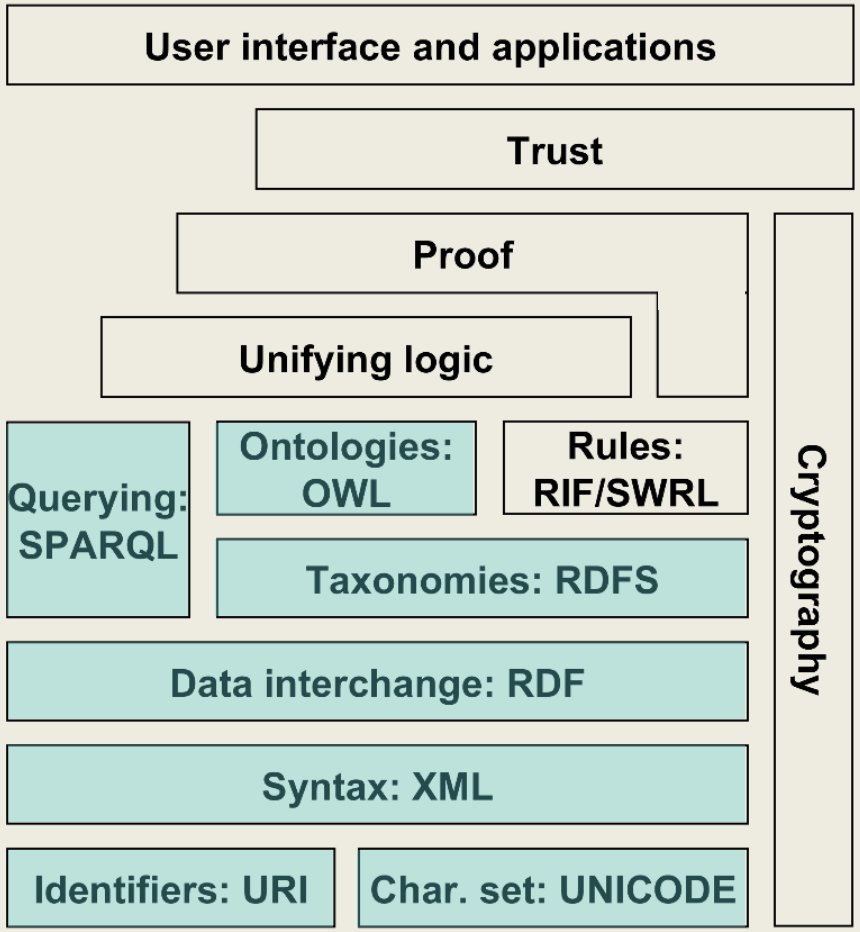
- Patients classification
  - type 1 diabetes
  - type 2 diabetes
  - diabetes unspecified
  - diabetes excluded
- Algorithm components
  - specific diagnostic codes for T1D and T2D
  - less specific codes for insulin/non-insulin dependent diabetes



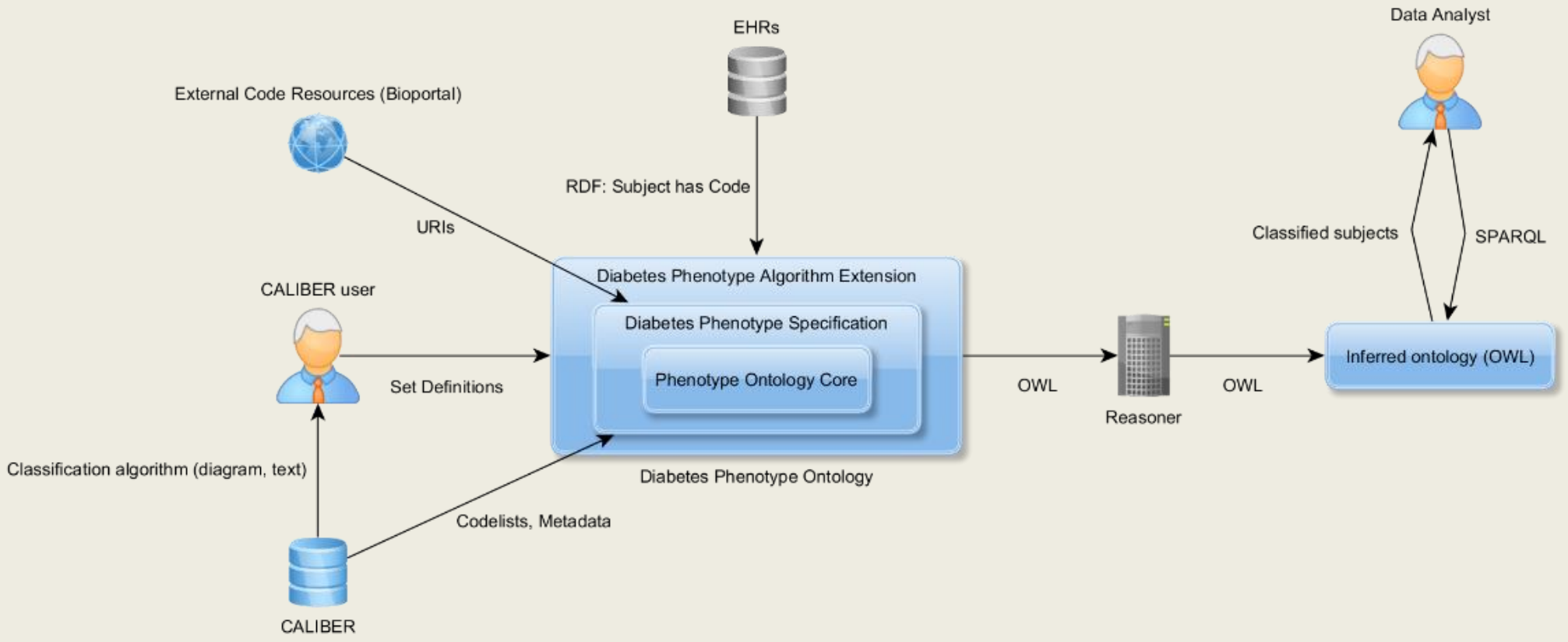
Shah A. et al., Lancet Diab Endocrinol, doi: 10.1016/S2213-8587(14)70219-0

Semantic Web Technologies

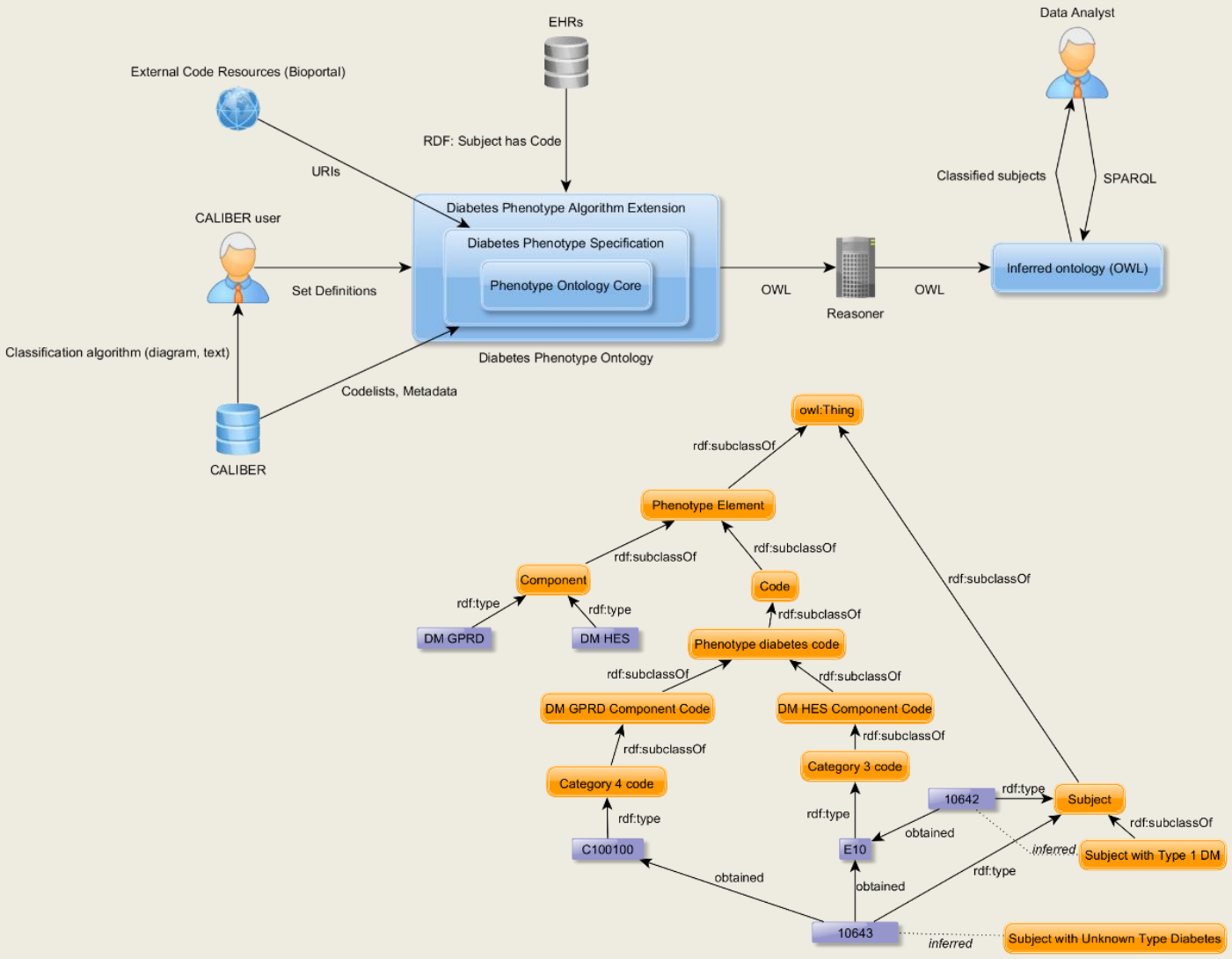
- Annotating and sharing data using Web protocols
- Automated data integration and reuse in a machine-readable manner
- Automatic reasoning

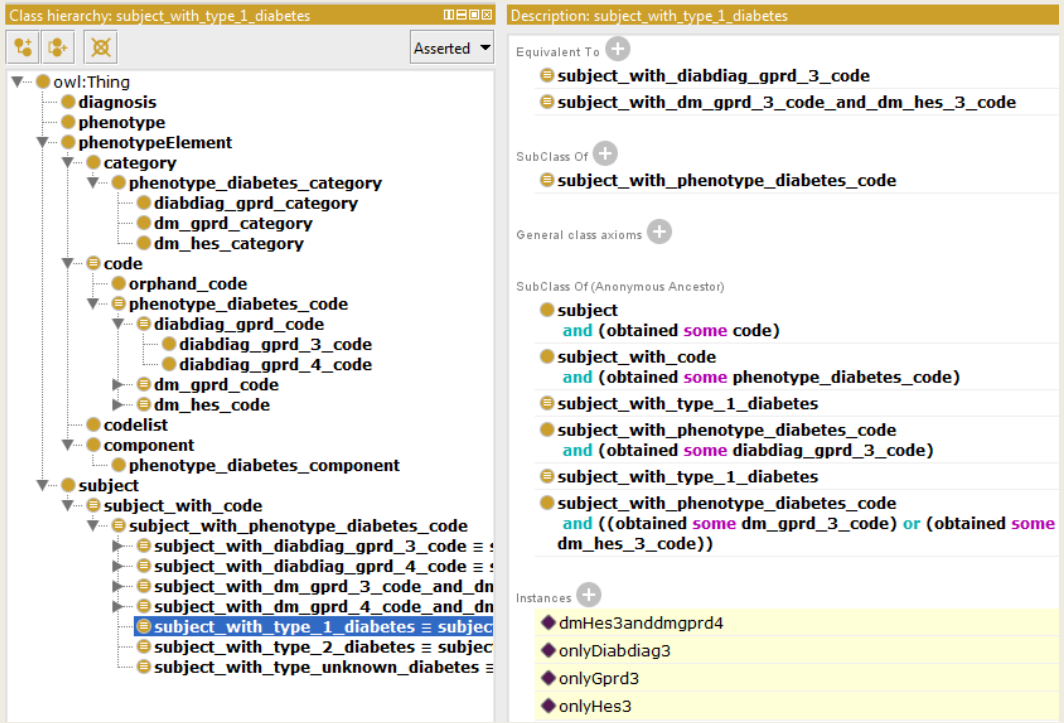


System architecture overview

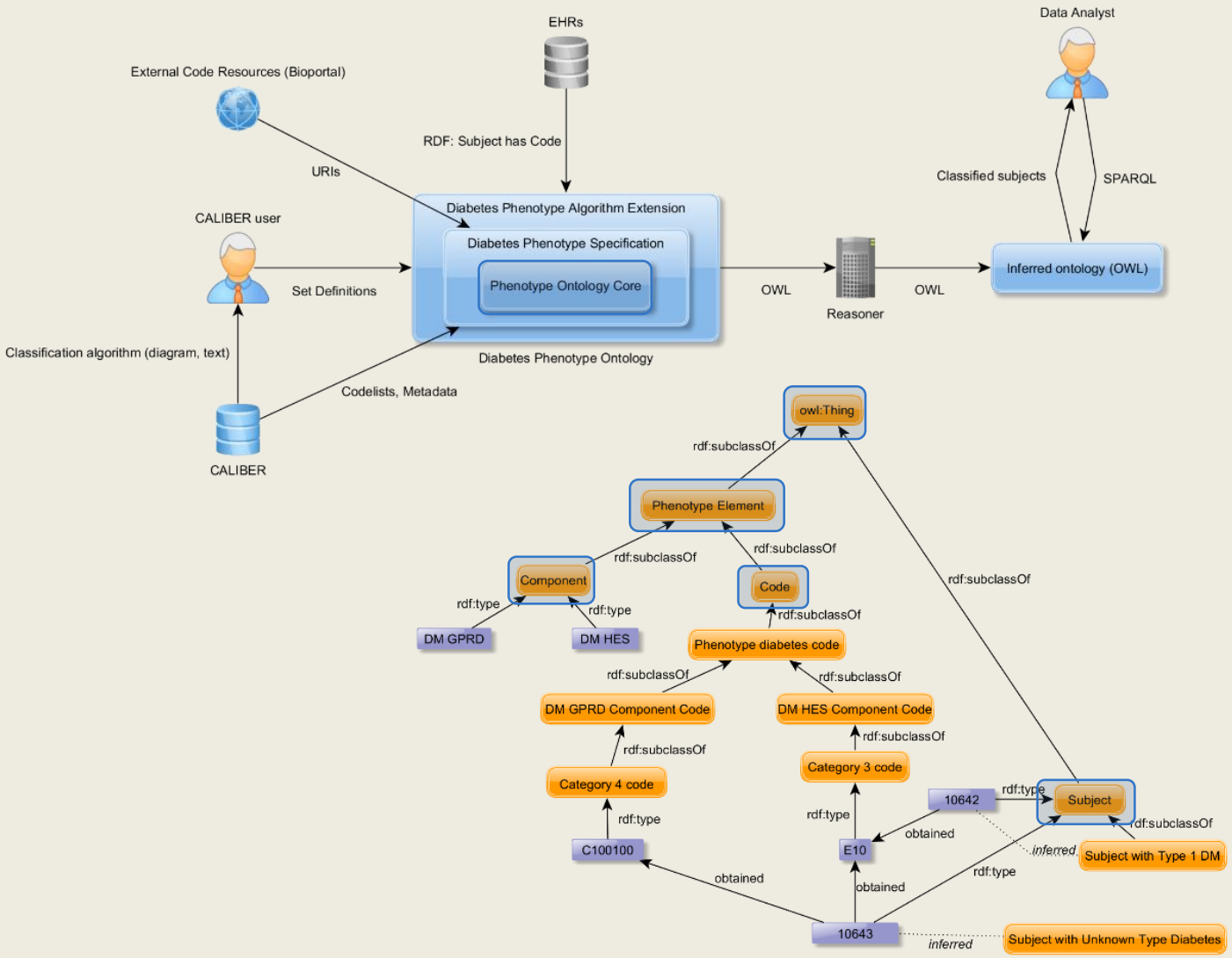


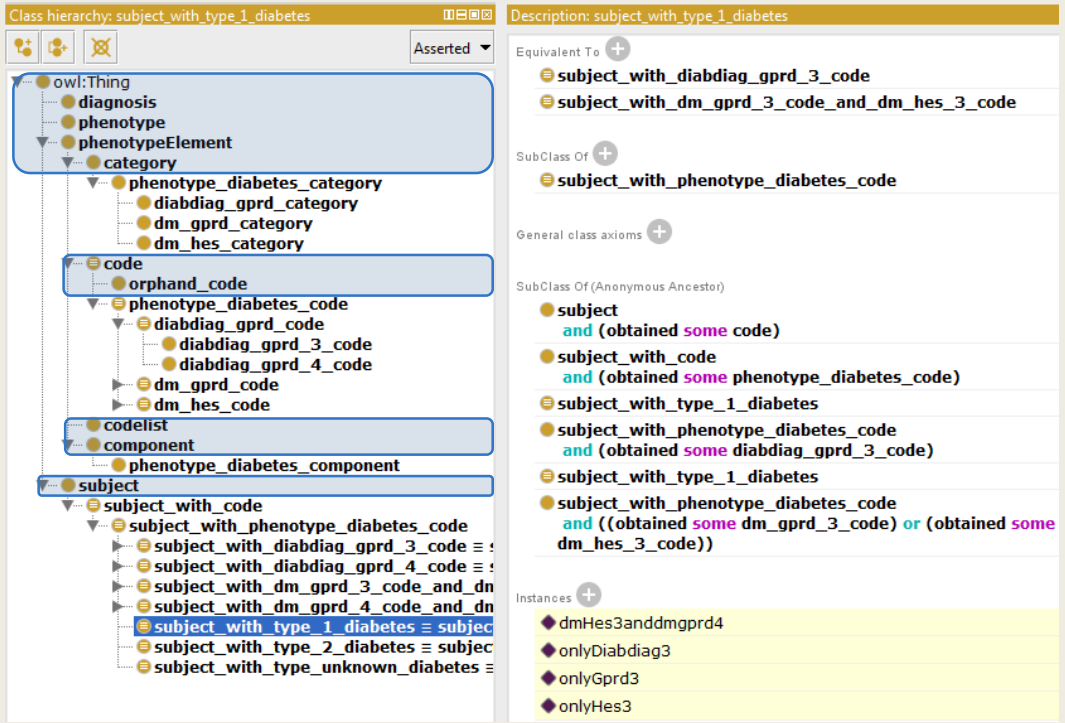
Incremental building





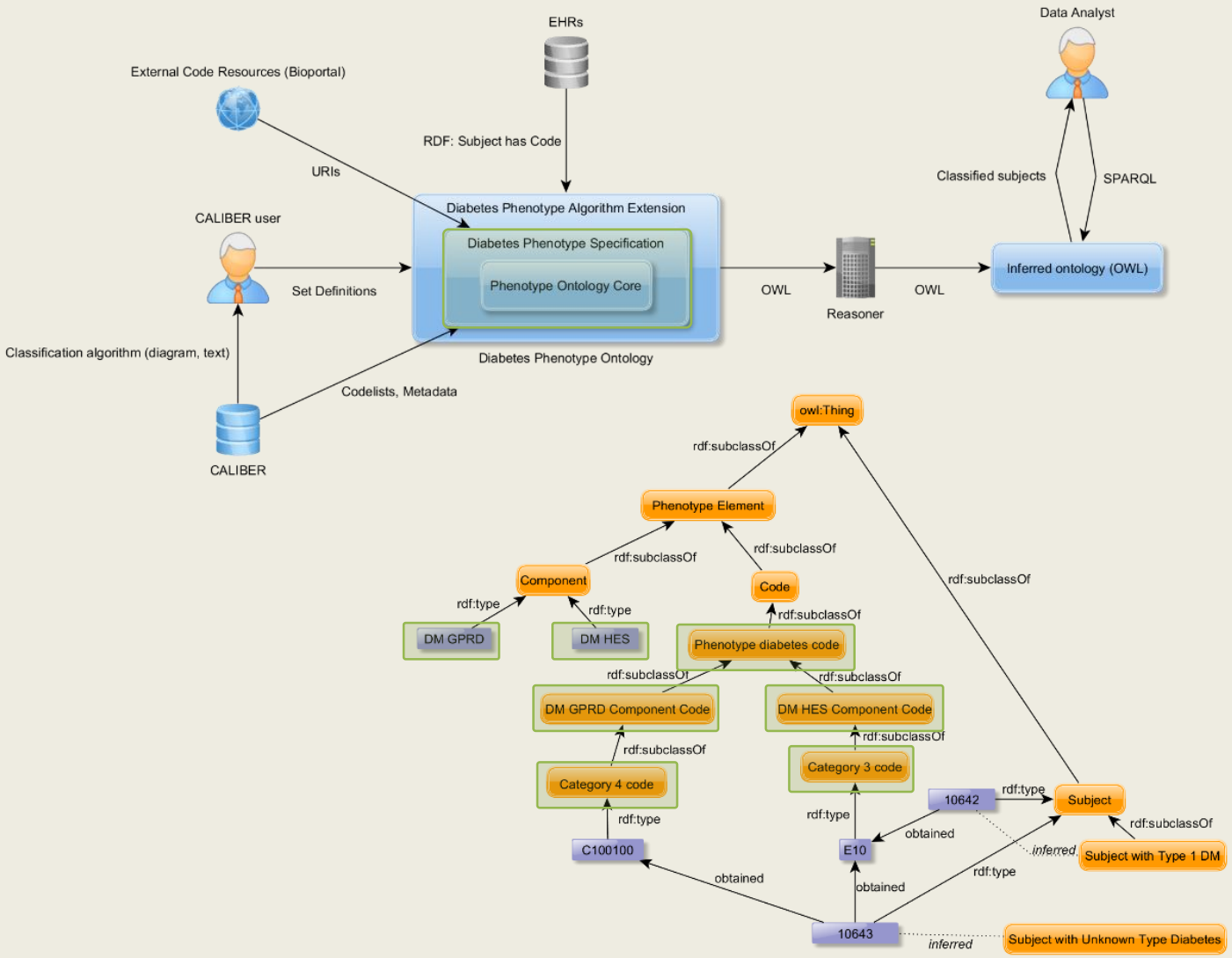
Incremental building

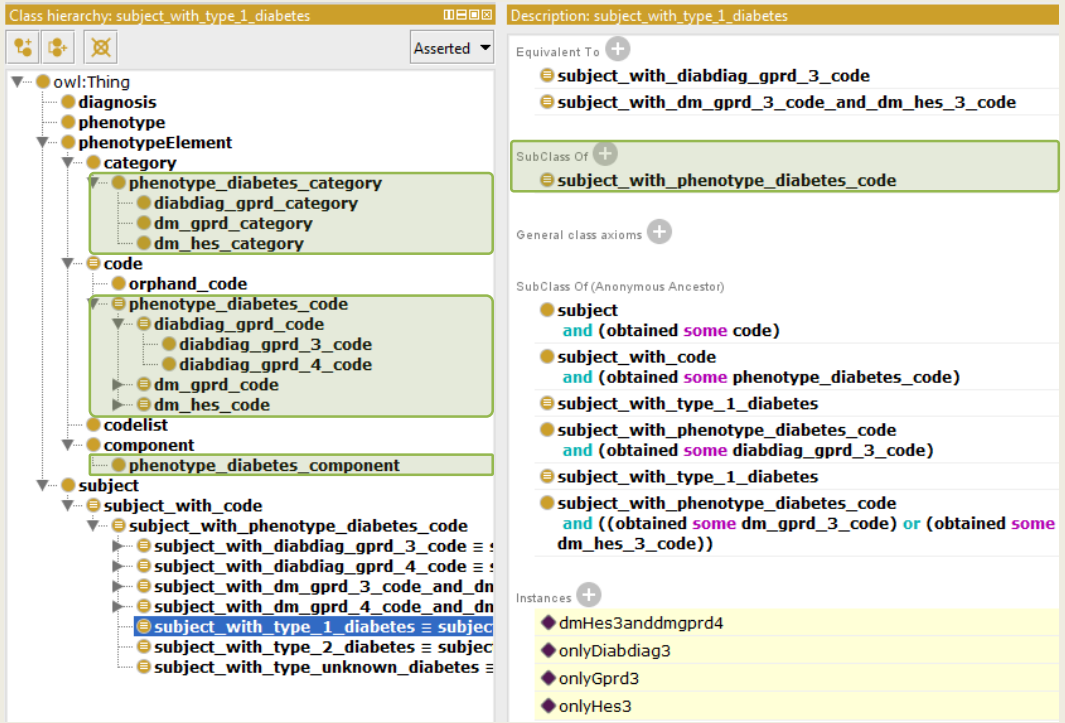




- Predefined ontology core
- Generic phenotype elements
- Domain independent

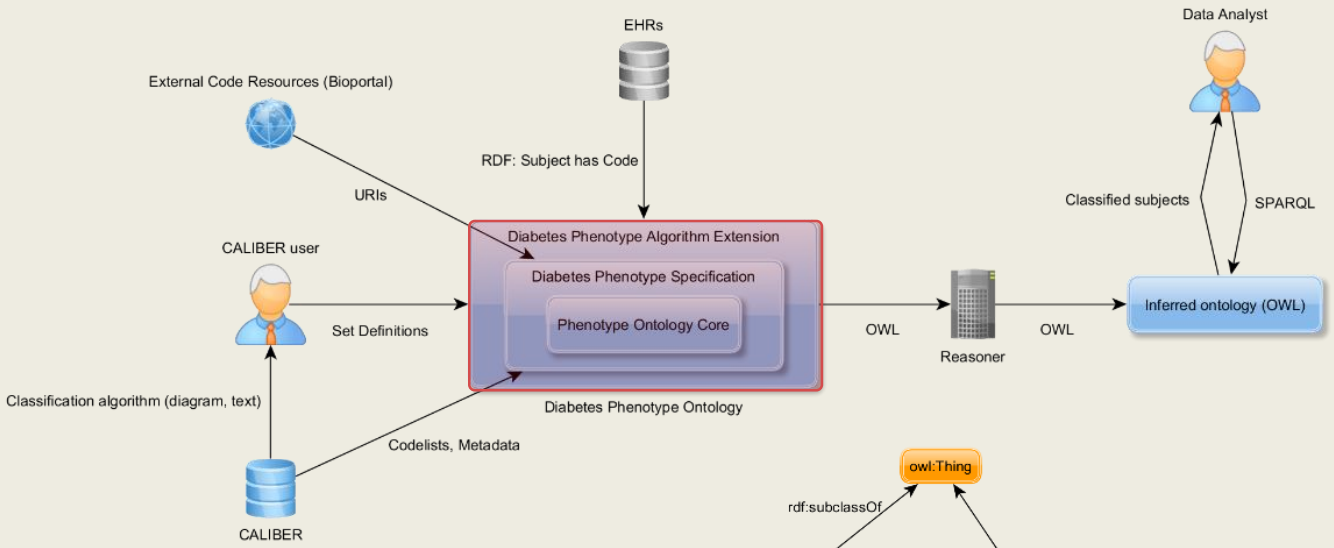
Incremental building

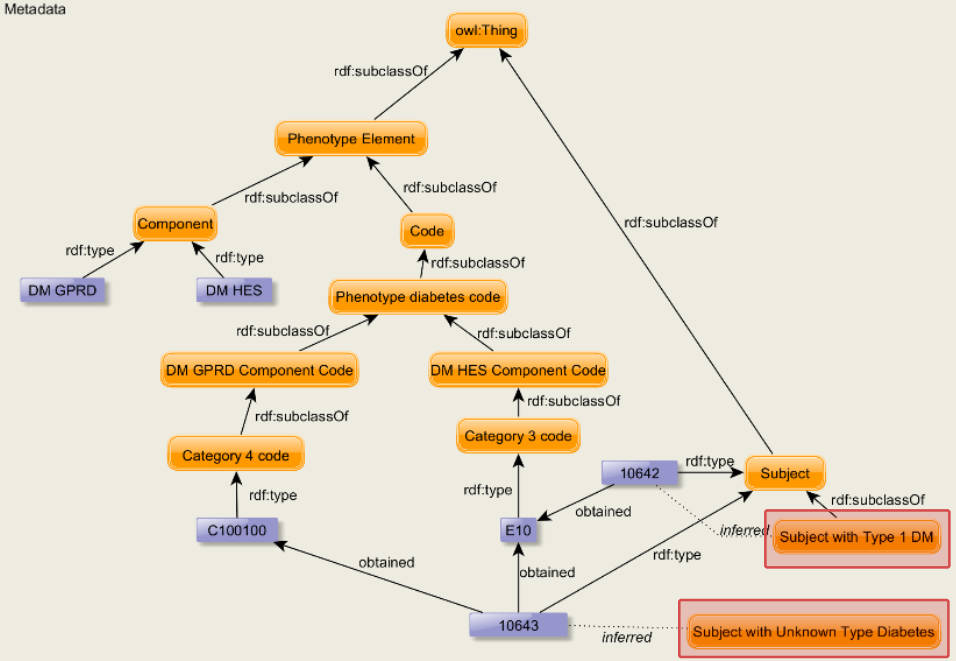


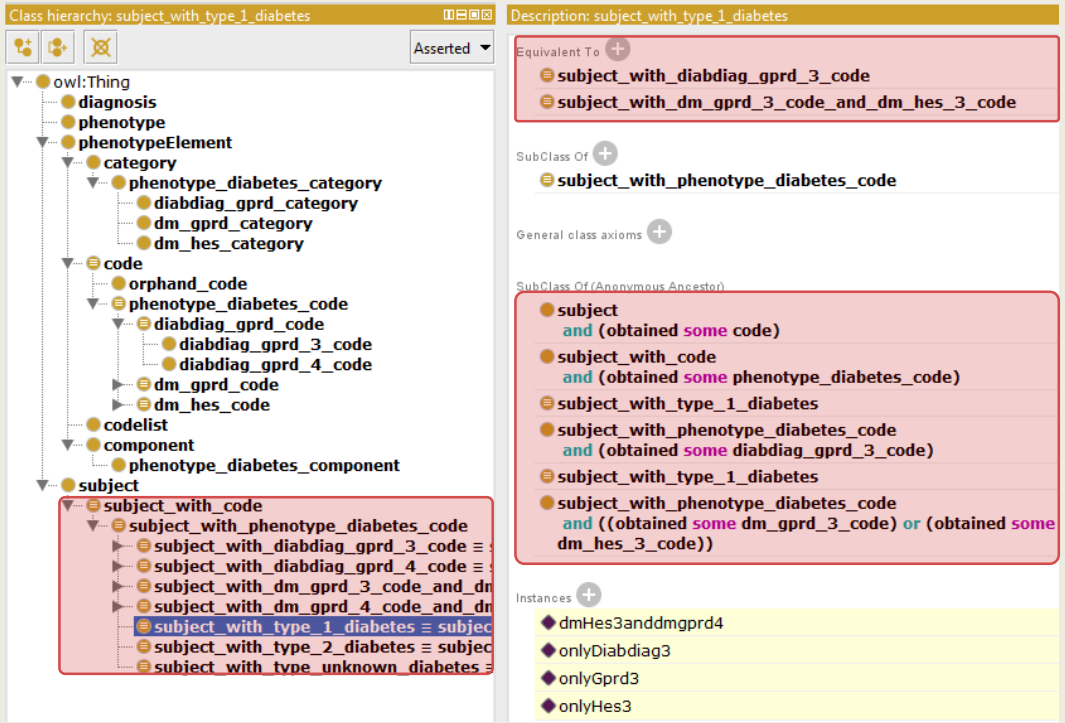


- Automatically imported structured components
- Disease/phenotype specific code lists
- Domain dependent

Incremental building

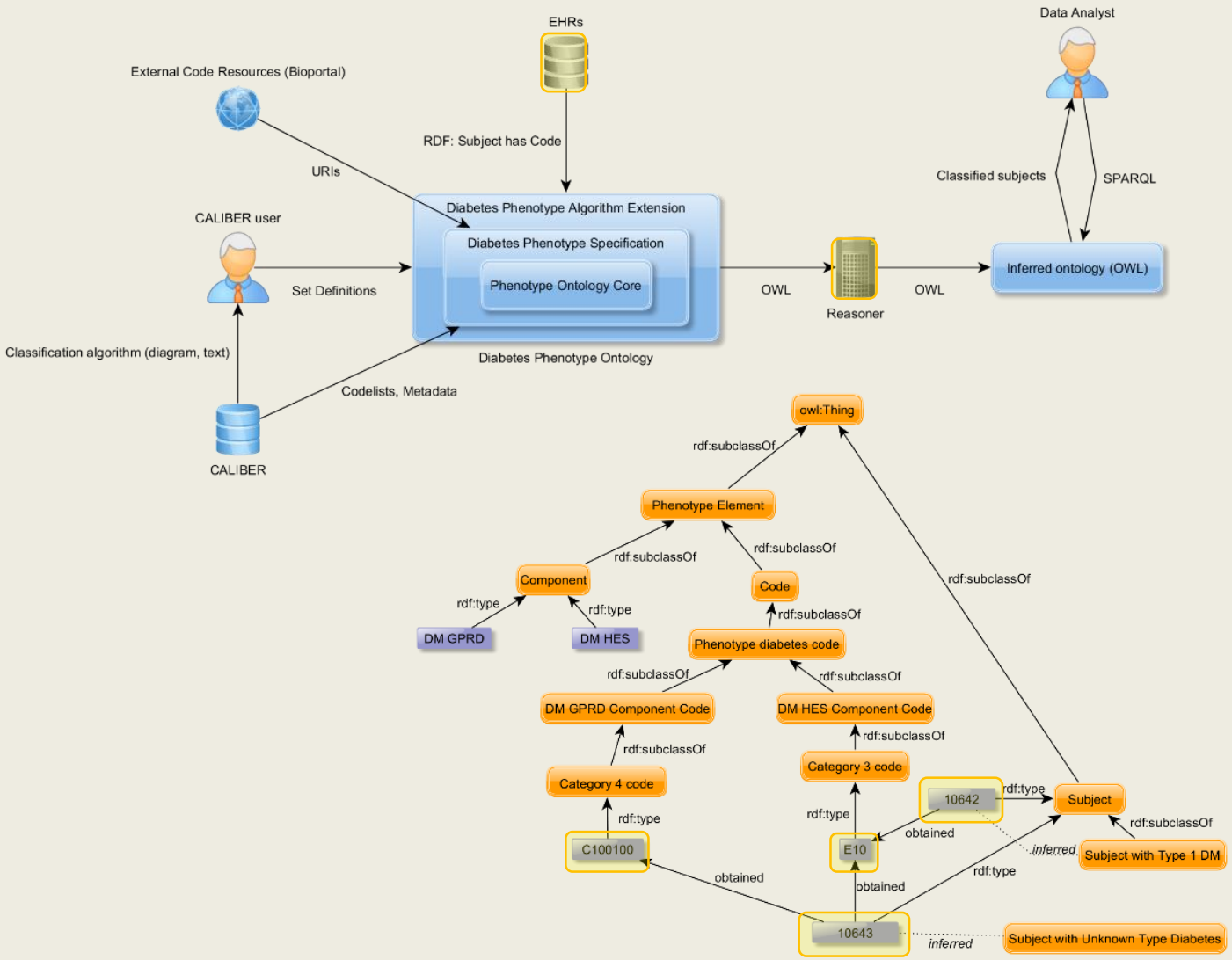


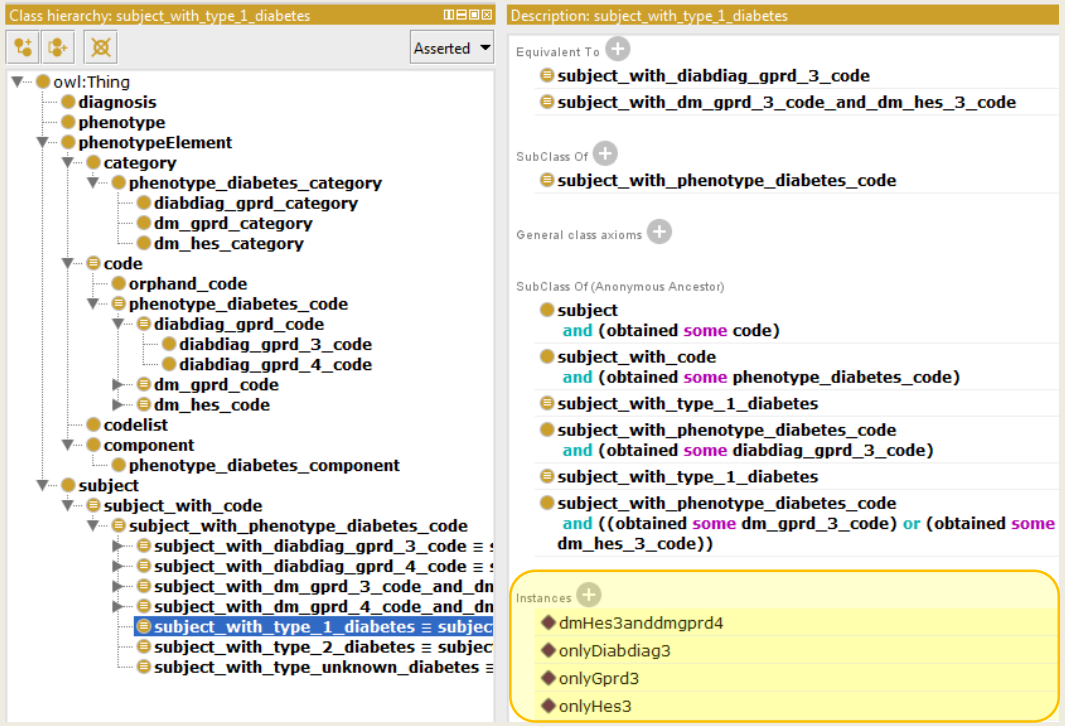




- Manually defined algorithmic logic
- Classification groups
- Domain dependent

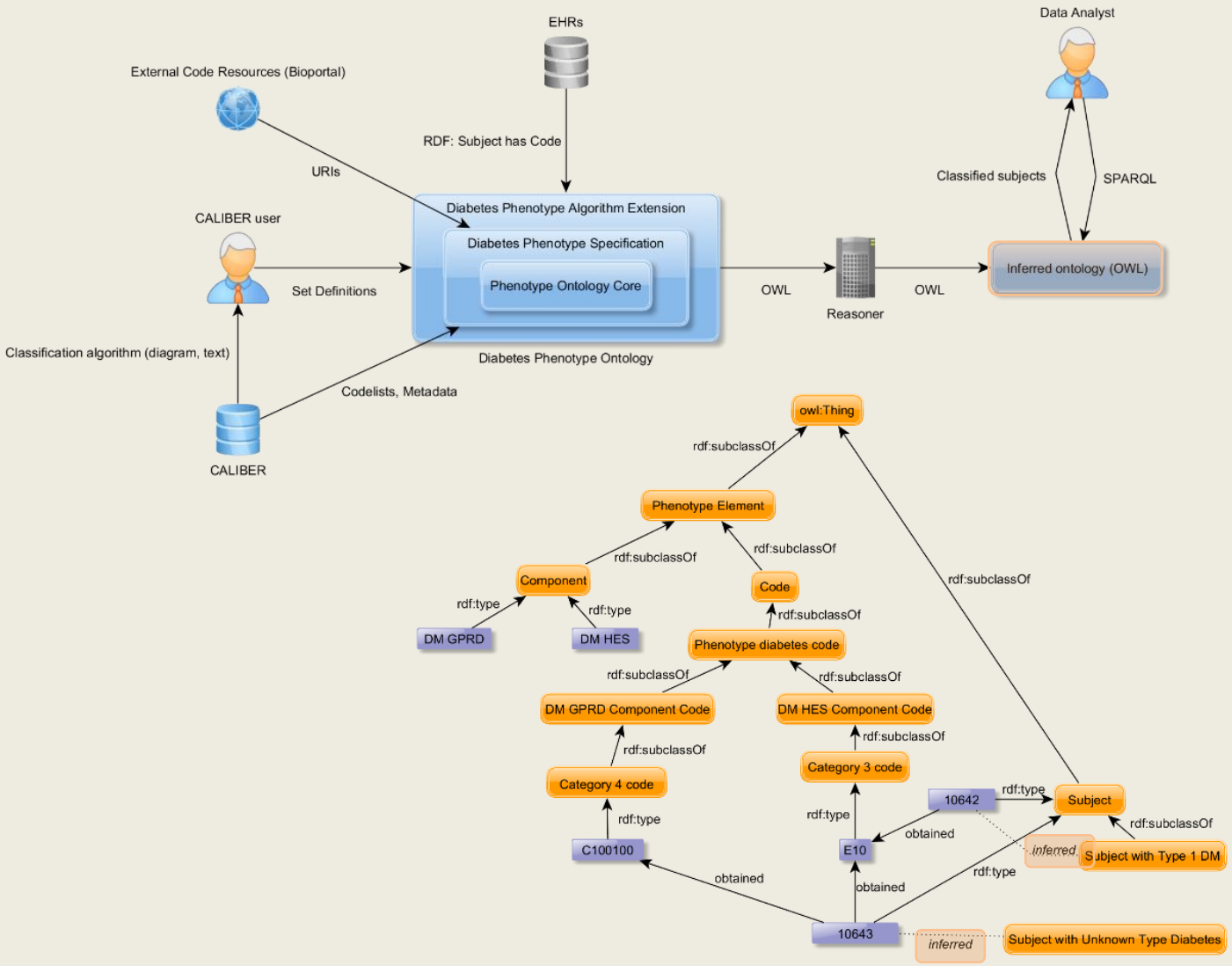
Incremental building

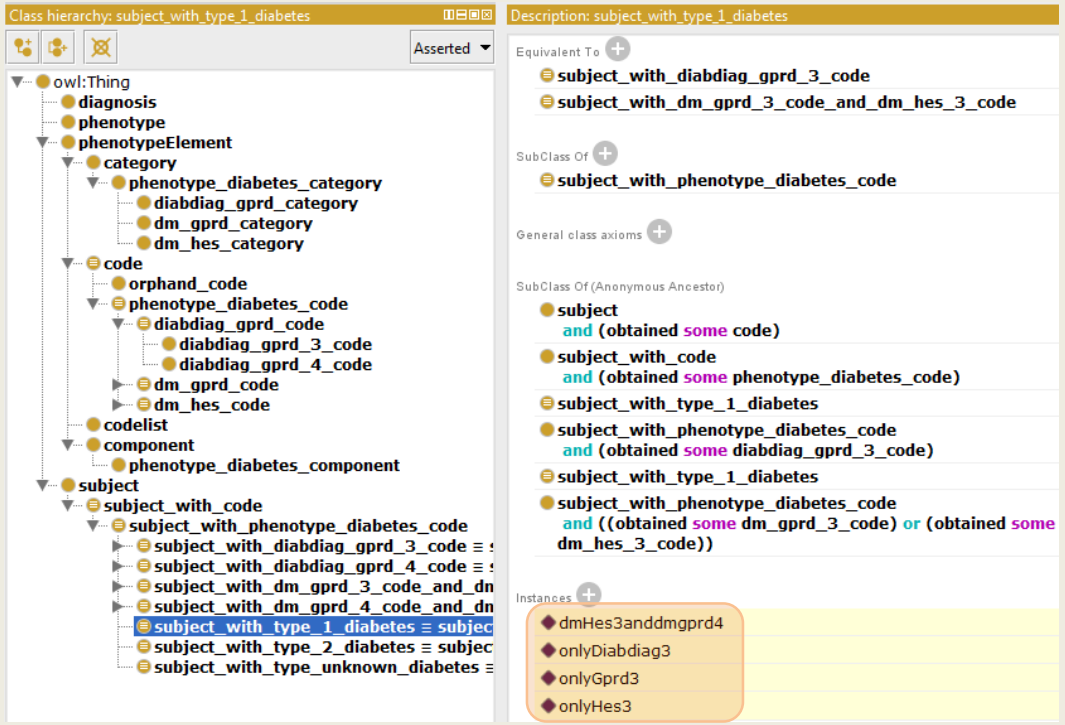




- EHRs appended to RDF graph
- Reasoner executed in order to infer classification
- Domain independent

Incremental building





- Inferred ontology stored
- Cohort extracted by SPARQL
- Domain dependent

Meeting the desiderata

Desiderata	Evaluation
Human-readable and computable representation	✓ Serialization into RDF/XML, OWL/XML, Turtle… ✓ RDF visualized as graphs
Set operations, relational algebra	✓ Set operations natively supported by OWL ✓ SPARQL algebra
Structure rules, temporal relations	✓ Complex structures by merging RDF graphs ○ Limited temporal rules via SPARQL logic
Standardized nomenclature	✓ Main ontology driving principle ✓ Linkages with existing controlled clinical terminology URIs
External interfacing, NLP	✓ OWL API for building interfaces ○ No native NLP support
Backward compatibility	✓ OWL2 backwards compatible with OWL ✓ Revision control systems/ontology portals

Meeting the desiderata

Desiderata	Evaluation
Human-readable and computable representation	✓ Serialization into RDF/XML, OWL/XML, Turtle… ✓ RDF visualized as graphs
Set operations, relational algebra	✓ Set operations natively supported by OWL ✓ SPARQL algebra
Structure rules, temporal relations	✓ Complex structures by merging RDF graphs ○ Limited temporal rules via SPARQL logic
Standardized nomenclature	✓ Main ontology driving principle ✓ Linkages with existing controlled clinical terminology URIs
External interfacing, NLP	✓ OWL API for building interfaces ○ No native NLP support
Backward compatibility	✓ OWL2 backwards compatible with OWL ✓ Revision control systems/ontology portals

Meeting the desiderata

Desiderata	Evaluation
Human-readable and computable representation	✓ Serialization into RDF/XML, OWL/XML, Turtle… ✓ RDF visualized as graphs
Set operations, relational algebra	✓ Set operations natively supported by OWL ✓ SPARQL algebra
Structure rules, temporal relations	✓ Complex structures by merging RDF graphs ○ Limited temporal rules via SPARQL logic
Standardized nomenclature	✓ Main ontology driving principle ✓ Linkages with existing controlled clinical terminology URIs
External interfacing, NLP	✓ OWL API for building interfaces ○ No native NLP support
Backward compatibility	✓ OWL2 backwards compatible with OWL ✓ Revision control systems/ontology portals

Meeting the desiderata

Desiderata	Evaluation
Human-readable and computable representation	✓ Serialization into RDF/XML, OWL/XML, Turtle… ✓ RDF visualized as graphs
Set operations, relational algebra	✓ Set operations natively supported by OWL ✓ SPARQL algebra
Structure rules, temporal relations	✓ Complex structures by merging RDF graphs ○ Limited temporal rules via SPARQL logic
Standardized nomenclature	✓ Main ontology driving principle ✓ Linkages with existing controlled clinical terminology URIs
External interfacing, NLP	✓ OWL API for building interfaces ○ No native NLP support
Backward compatibility	✓ OWL2 backwards compatible with OWL ✓ Revision control systems/ontology portals

Meeting the desiderata

Desiderata	Evaluation
Human-readable and computable representation	✓ Serialization into RDF/XML, OWL/XML, Turtle… ✓ RDF visualized as graphs
Set operations, relational algebra	✓ Set operations natively supported by OWL ✓ SPARQL algebra
Structure rules, temporal relations	✓ Complex structures by merging RDF graphs ○ Limited temporal rules via SPARQL logic
Standardized nomenclature	✓ Main ontology driving principle ✓ Linkages with existing controlled clinical terminology URIs
External interfacing, NLP	✓ OWL API for building interfaces ○ No native NLP support
Backward compatibility	✓ OWL2 backwards compatible with OWL ✓ Revision control systems/ontology portals

Meeting the desiderata

Desiderata	Evaluation
Human-readable and computable representation	✓ Serialization into RDF/XML, OWL/XML, Turtle… ✓ RDF visualized as graphs
Set operations, relational algebra	✓ Set operations natively supported by OWL ✓ SPARQL algebra
Structure rules, temporal relations	✓ Complex structures by merging RDF graphs ○ Limited temporal rules via SPARQL logic
Standardized nomenclature	✓ Main ontology driving principle ✓ Linkages with existing controlled clinical terminology URIs
External interfacing, NLP	✓ OWL API for building interfaces ○ No native NLP support
Backward compatibility	✓ OWL2 backwards compatible with OWL ✓ Revision control systems/ontology portals

Results

Advantages

- OWL and RDF can create machine-readable phenotyping algorithms, satisfy most desiderata
- Automatic patient classification by semantic reasoners
- Independent on underlying EHR storage solution

Challenges

- Serious performance issues for advance OWL DL constructs (e.g. disjointWith, equivalentClass, intersectionOf)
- Limited support for external algorithms e.g. NLP
- Temporal relationships
- Interface

Next steps

- Solve the performance issues
- Evaluate more complex phenotypes
- Integration of various sources of clinical information
- Formal evaluation of the algorithm implementation and cohort accuracy
- Human-friendly interface

Acknowledgement

- Institute of Health Informatics, University College London, GB
  - Spiros Denaxas
  - Richard Dobson
  - Arturo González-Izquierdo
  - Kenan Direk
- The Hyve, Utrecht, NL
  - Maxim Moinat
  - Stefan Payralbe
  - Marinel Cavelaars
- University Medical Center Utrecht, Utrecht University, NL
  - Stefan Koudstaal
  - Alicia Uijl

Discussion