

Objective

To assess the practicality and utility of using ontologies with real world data sets and the use of ontology-based methods to address the challenges with the conceptualization of context, mechanisms and impact, and assessing data quality of datasets of routinely collected “big data” from multiple general practice electronic health records and determine their fitness for clinical or research use.

1. Ontology development

A. Conceptualisation:

- ✓ The automation of identifying diabetes patients is conceptualized as semantic retrieval.
- ✓ The conceptualisation of diabetes assessment and management is drawn from evidence-based guidelines and local clinicians.

B. Specification:

- ✓ The SNOMED CT-AU® standard guided the specification of the data and domains in the diabetes identification ontology (DIO).
- ✓ Hierarchical conceptual modeling was used to formalise

C. Formalisation:

- ✓ The formalized ontology consists of 4 main classes (*Actor, Content, Mechanism and Impact*) and 68 subclasses with object/data properties.
- ✓ Some of the concepts are mappable to the SNOMED CT-AU Ontology (SCAO), which has more than 300,000 concepts
- ✓ Example: T2DM is a Disease under the subclass of Problem which has a superclass Context in the DIO. In the SCAO, T2DM is a disorder of glucose metabolism which is a subclass of Disease under the highest level concept of Clinical finding. Similarly, Actor class in DIO is mapped to Environment or Geographical location in SCAO.

D. Implementation:

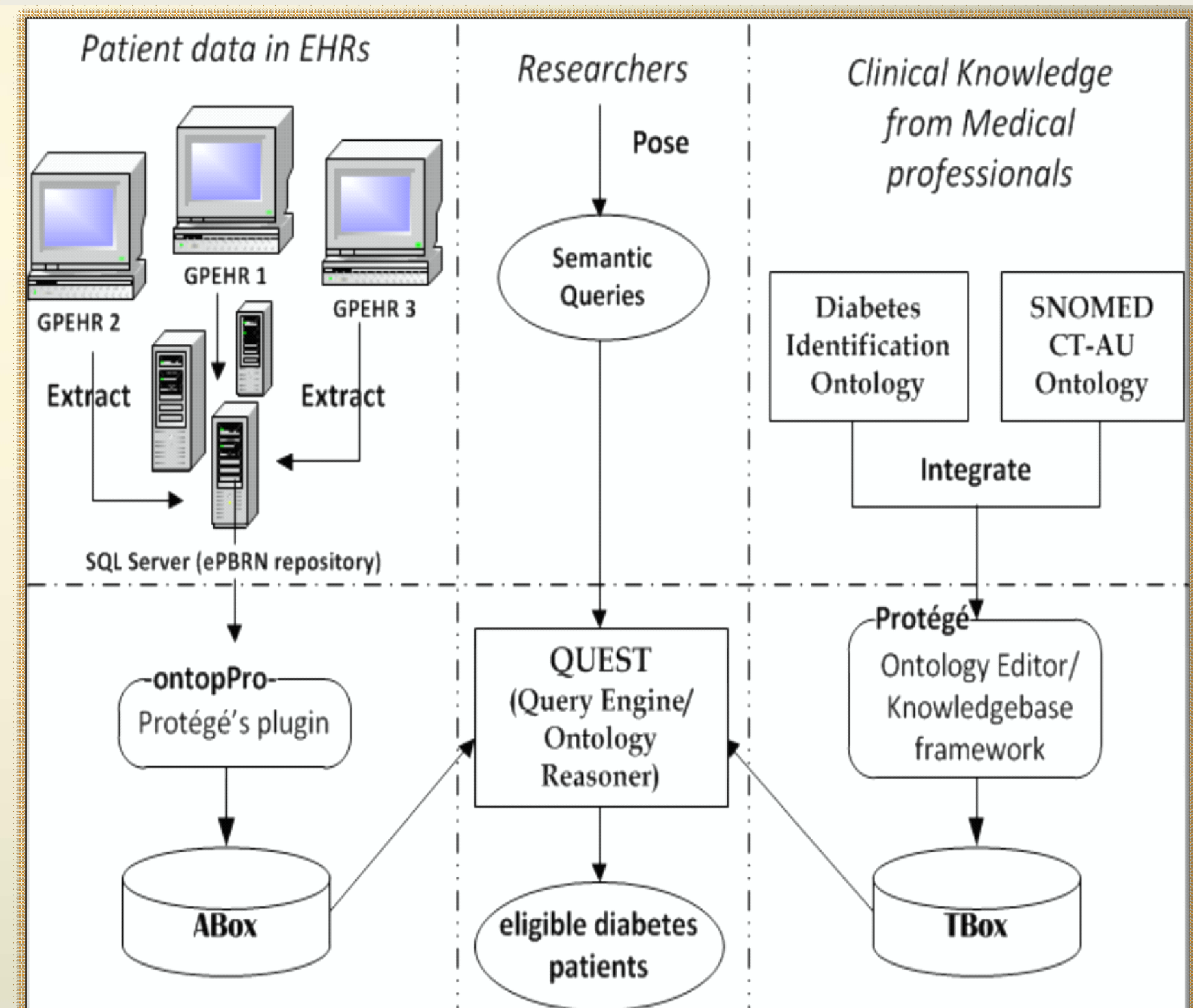
- ✓ The ontology created was implemented using T-SQL
- ✓ OBDA methods were used to map the existing data in a subset of the general practices in the electronic Practice Based Research Network (ePBRN) to the ontology.

2. Ontology Based Data Access (OBDA)

OBDA methods to integrate data and test phenotyping algorithm

- ✓ The TBox, related to conceptual terminologies defined in ontologies, was built through Protégé, a popular open source ontology editor and knowledgebase framework.
- ✓ The ABox, associated with instances of ontology classes or properties, was populated through ontopPro (an OBDA plugin for Protégé).
- ✓ Clinical selection criteria were formulated as semantic queries in SPARQL Protocol and RDF Query Language (SPARQL). The SPARQL query engine QUEST that comes with ontopPro checked the queries against the knowledgebase to retrieve matched patients.

3. Architecture and data flow



4. Evaluation

- ✓ The accuracy of the phenotyping algorithm was measured using the ePBRN subset (thirteen years of c 100,000 patient records).
- ✓ Functional evaluation: usability, interoperability and scalability.
- ✓ Architecture evaluation: modifiability with many facades/locations of data/data types, integrability and extensibility

Discussion points

Activities and Findings:

- ✓ We addressed and solved some engineering challenges around ontology creation, data access and data integration, using tested methods to map datasets to ontologies;
- ✓ We used real world patient datasets to test our solutions, which necessitated the assessment and management of data quality; and
- ✓ We confirmed that automated identification of diabetes patients can be specified systematically as a solution supported by semantic retrieval.

Discussion:

- ✓ Primary and integrated care “big data” tasks that require automating include semantic integration of clinical data from multiple EHRs; assessment and management of the provenance and quality of EHR data such as reasons for visit, chronic conditions and diagnoses, pathology tests and prescriptions; and preservation of meaning of the data, information and knowledge as they may be perceived and interpreted by clinicians and researchers.