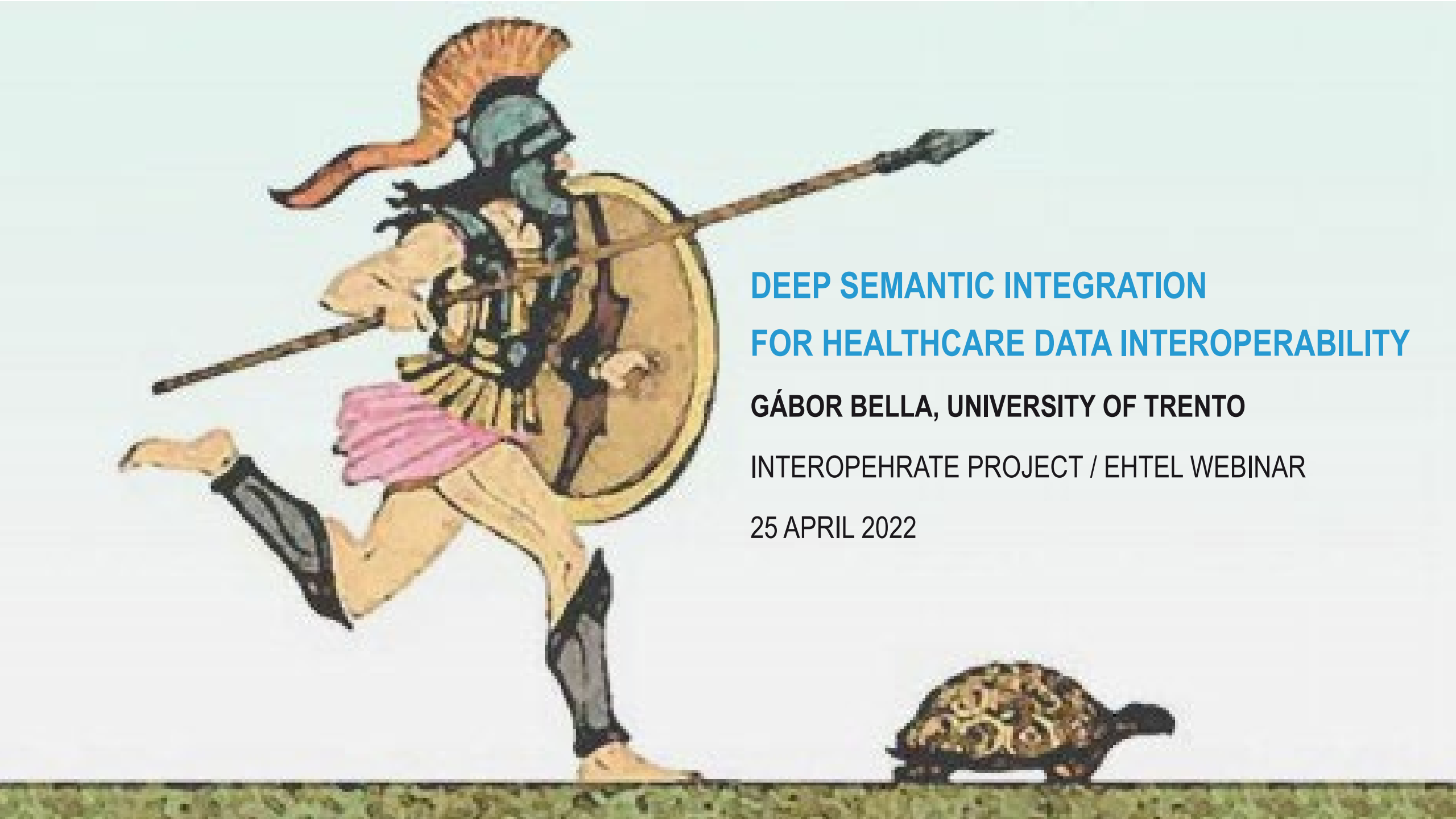


DEEP SEMANTIC INTEGRATION FOR HEALTHCARE DATA INTEROPERABILITY

GÁBOR BELLA, UNIVERSITY OF TRENTO

INTEROPEHRATE PROJECT / EHTEL WEBINAR

25 APRIL 2022

SUMMARY

1. Health data interoperability:
   a short reminder of motivations, problems, and known solutions
2. Why are we not there yet?
3. Innovation ideas for faster progress, based on experience
   from InteropEHRate and from past projects

MOTIVATIONS

CARE	Local: **integration of hospital departments, services, and their IT systems** National: **movement of patients within the country using EHR** Cross-border: **EU-wide movement of patients, EU-wide EHR** **Citizen empowerment: retrieving one's own health records**
RESEARCH	National: **already frequent, integrating local data sources** Cross-border: **mostly ad-hoc, e.g. precision medicine, large cohorts**
NON-MEDICAL (ACCOUNTING, REIMBURSEMENT, STATISTICS, ETC.)	Local: **towards the institution itself** National: **towards the state**

THE CURRENT LANDSCAPE: PROBLEMS

In the EU, data heterogeneity is pervasive, progress towards interoperability is uneven.

- Wide use of free text in the local language, as documents (discharge reports) and in data structures (prescriptions, terms) => a problem for automation and for cross-border uses;
- local conventions for terms, codes, and data schemas, heterogeneity can also exist within institutions;
- a lot of details are left implicit as they are evident from local context and practices.

THE CURRENT LANDSCAPE: SOLUTIONS

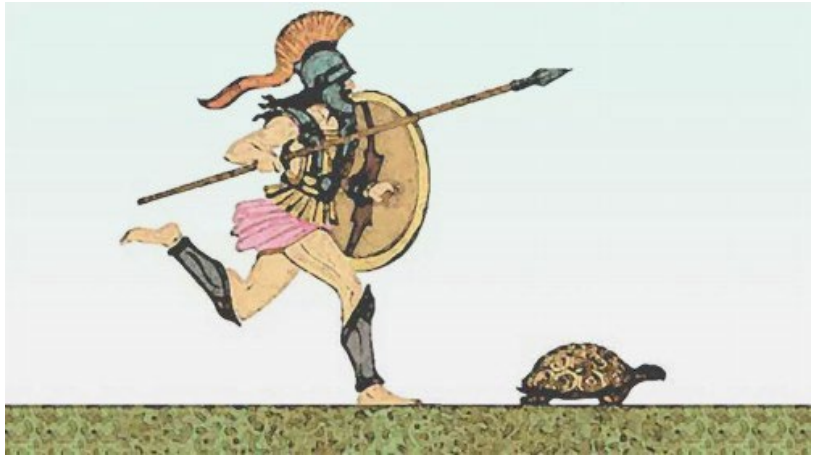
National and international standards:

- monolingual and multilingual **terminologies** (SNOMED CT, UCUM, EDQM): reasonable progress;
- international **coding systems** (ICD, LOINC, ATC): reasonable progress.
- **information models** (HL7, CDA, FHIR, OMOP): slow progress, only for communication.

Supporting technology:

- **terminology servers**: increasingly used;
- **Extract-Transform-Load (ETL) tools**: graphical UI for file conversions, schema mappings, data transformations, sometimes code mappings: widely used;
- **information extraction tools**: never in care, somewhat in research, more often for accounting.

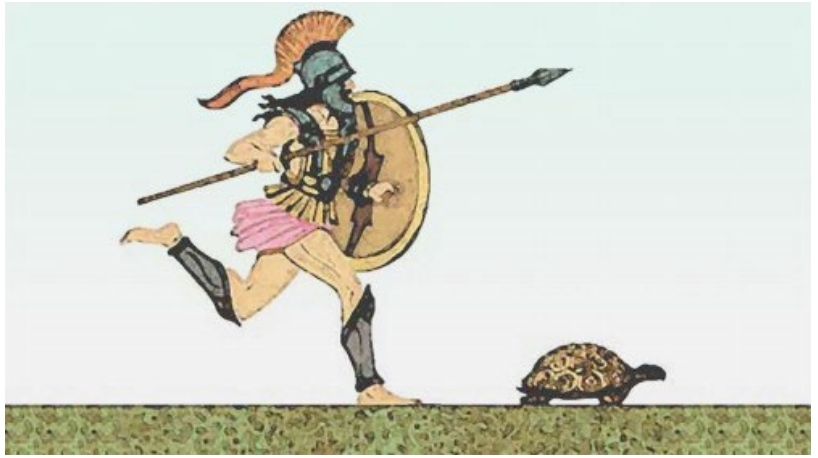
WHY ARE WE NOT THERE YET?



ADOPTING STANDARDS IS HARD

- Standards are an absolute necessity;
- supporting health standards is simply **a lot of work** involving top experts;
- and even if all parties “play the game” with good intentions,
- the complexity of standards ALWAYS leads to diverging implementations (based on differing priorities for implementors).

WHY ARE WE NOT THERE YET?

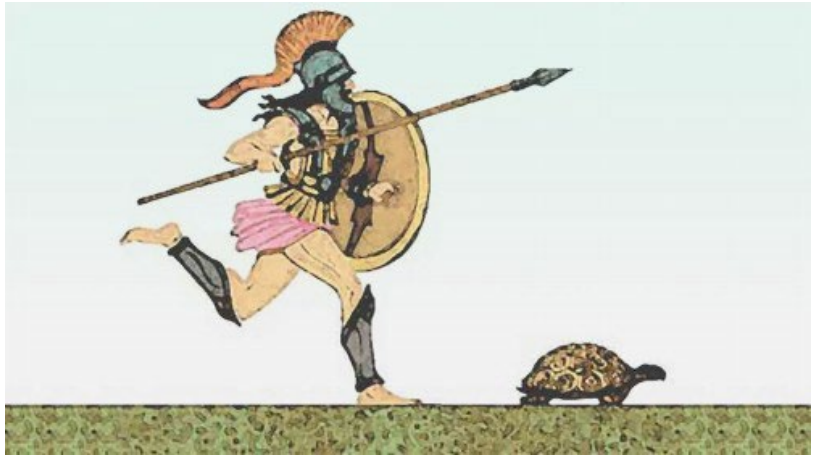


TECHNOLOGY IS NOT ENOUGH

- Some technological promises of the last 20 years:
  - (multilingual) **terminology servers**, e.g. UMLS => increasingly used,
  - **ontologies** to automate data exchange => occasionally used,
  - **artificial intelligence** => used in specific contexts (image analysis), not widely adopted (quality control, the “black box” problem, the “garbage-in-garbage-out” problem).
- **ETL tools** for data transformations:

  usually no support to understand natural-language labels, nor for cross-lingual transformations.
- The health sector is not an early adopter due to **quality and traceability requirements.**

WHY ARE WE NOT THERE YET?



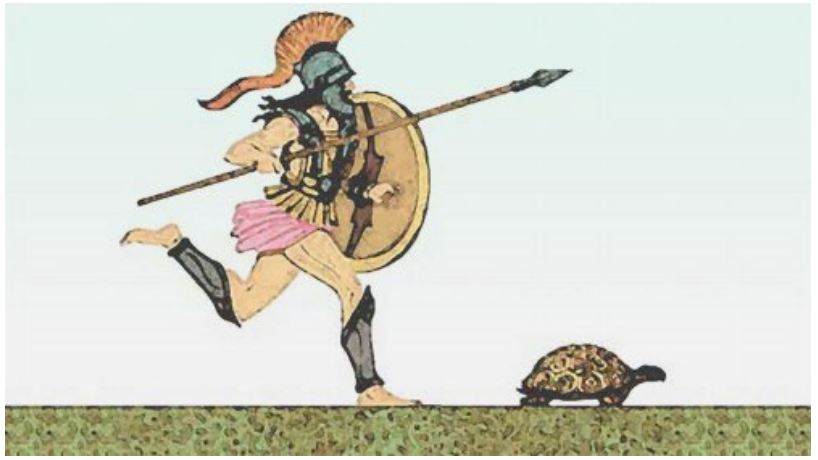
HEALTHCARE IS EVOLVING

- A constant evolution of:
  - pathologies,
  - our knowledge of pathologies, treatments, etc.,
  - local practices => of local data models,
  - standards.

Consequence: implementations must constantly follow the evolution:

the work is never fully done.

WHY ARE WE NOT THERE YET?



IN SUMMARY: CONTRADICTING NEEDS

- **scalability:** automate data mappings and transformations as much as possible;
- **correctness:** be very precise (close to 100% for care, above 90% for research).

Our attempt at reconciling these needs is called DEEP SEMANTIC INTEGRATION.

It is based on innovation in both technology and methodology.

APPLICATIONS SO FAR

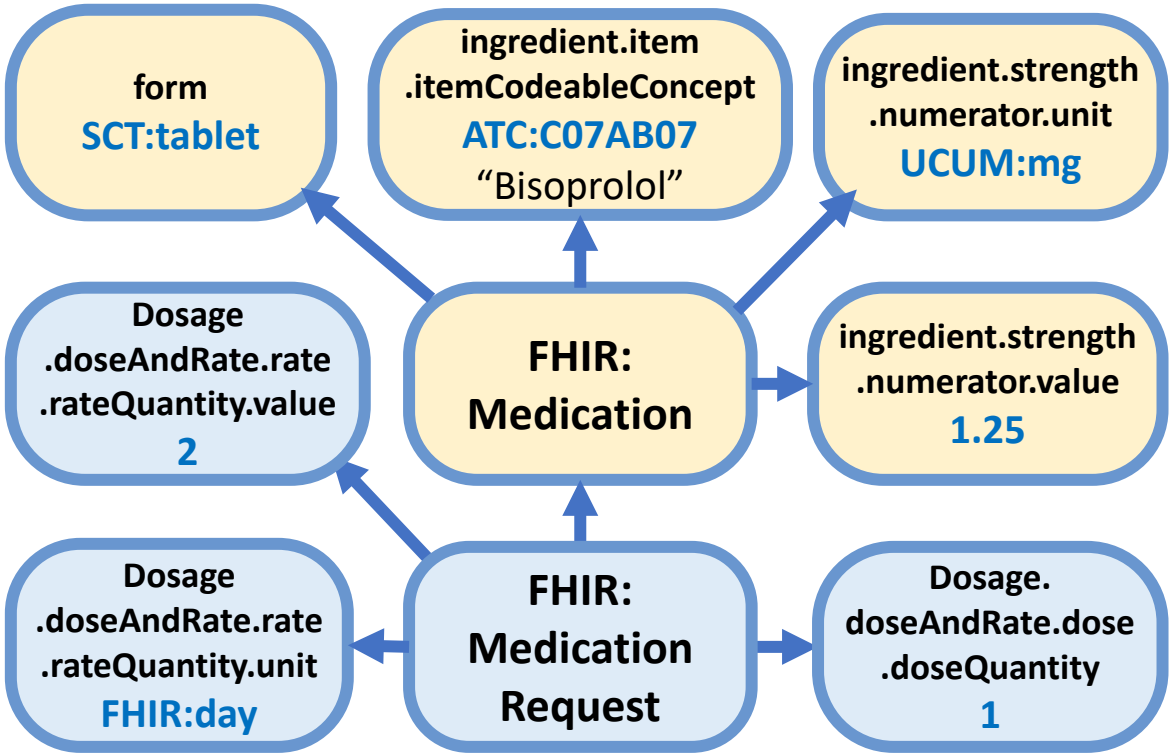
- 2016-2017: Cross-border research, Italy-Scotland, Healthcare Data Safe Havens project, EIT
- 2018: National health data integration, Scotland, HDR UK SPRINT
- 2019-ongoing: Cross-border care and research, InteropEHRate H2020 project
- 2021-ongoing: master-level courses on data integration:
  - University of Trento, Italy
  - Jilin University, China
  - Dublin City University, Ireland

DEEP SEMANTIC INTEGRATION

"Deep and semantic" means that every single data value (relevant to the task) is understood, made explicit, and represented in a formal, language-independent manner.

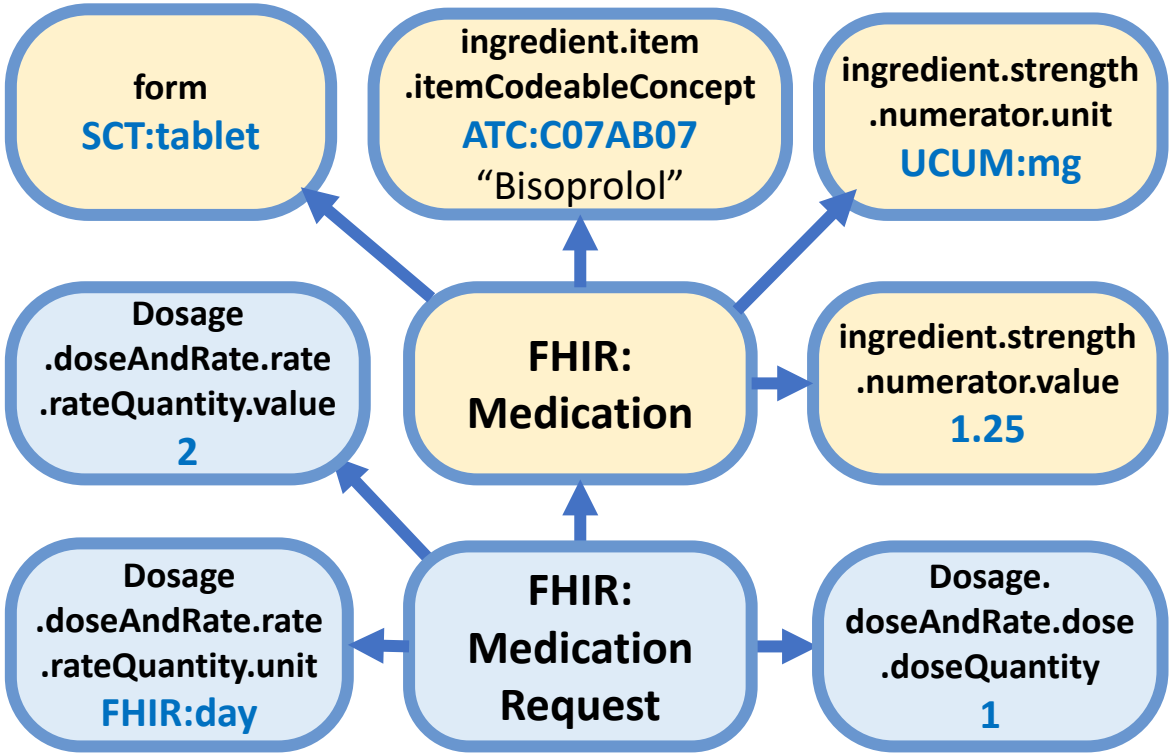
Prescrizione
Bisoprololo (Bisoprololo san) 1,25 mg 1 cpr x 2/die

In-depth understanding



DEEP SEMANTIC INTEGRATION

Starting from a correct and complete knowledge graph, **in-depth adaptation** (conversion, translation) can be automated in a robust way.

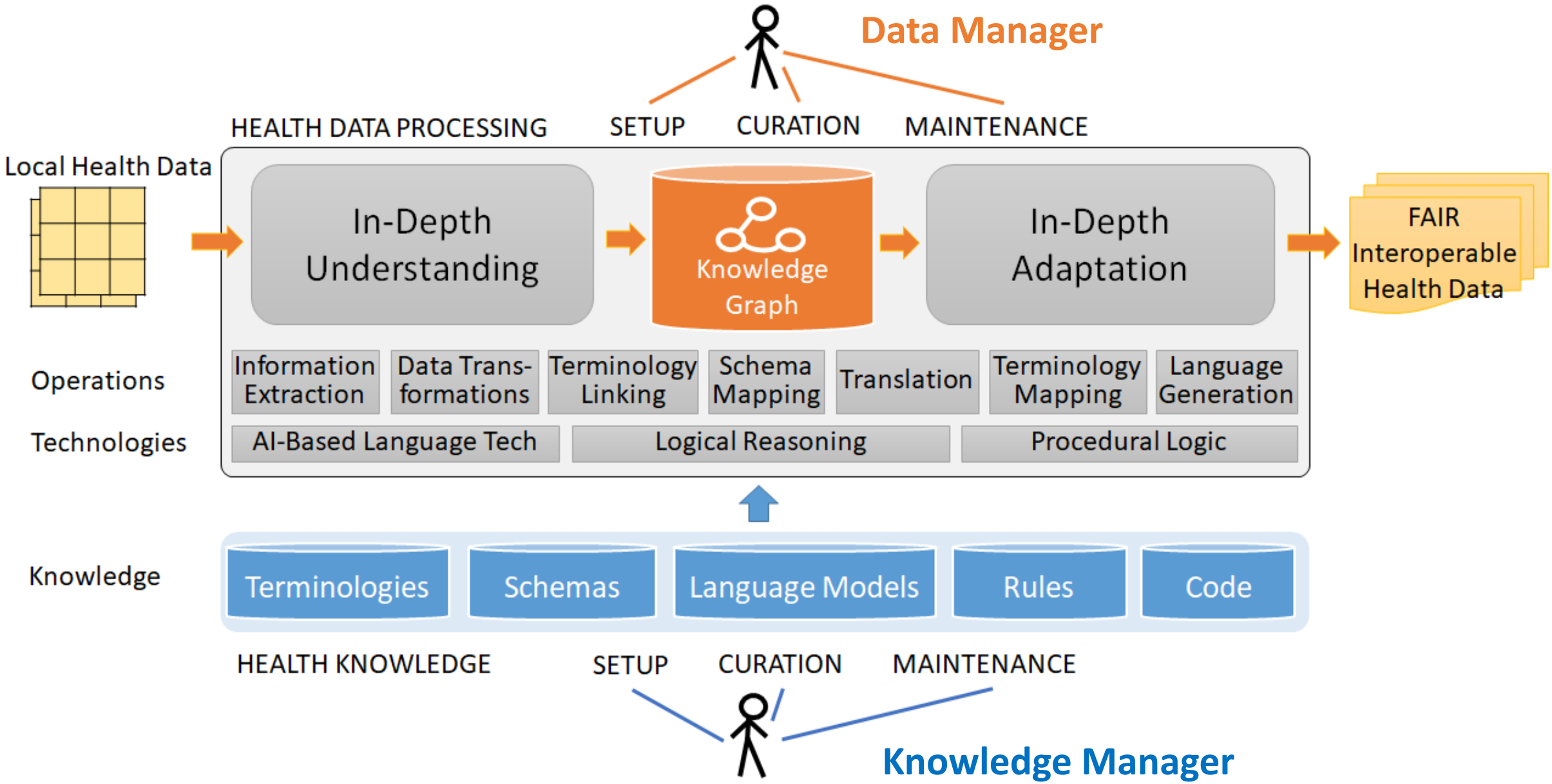


In-depth adaptation

Principe actif	Produit	Dosage	Unité de mesure
Bisoprolol	Bisoprololo san	1,25	mg

Quantité	Forme pharmaceutique	Fréquence	Période
1	comprimé	2	jour

COMPONENTS AND HIGH-LEVEL METHODOLOGY



A HUMAN-CENTRED, YET AUTOMATED METHODOLOGY

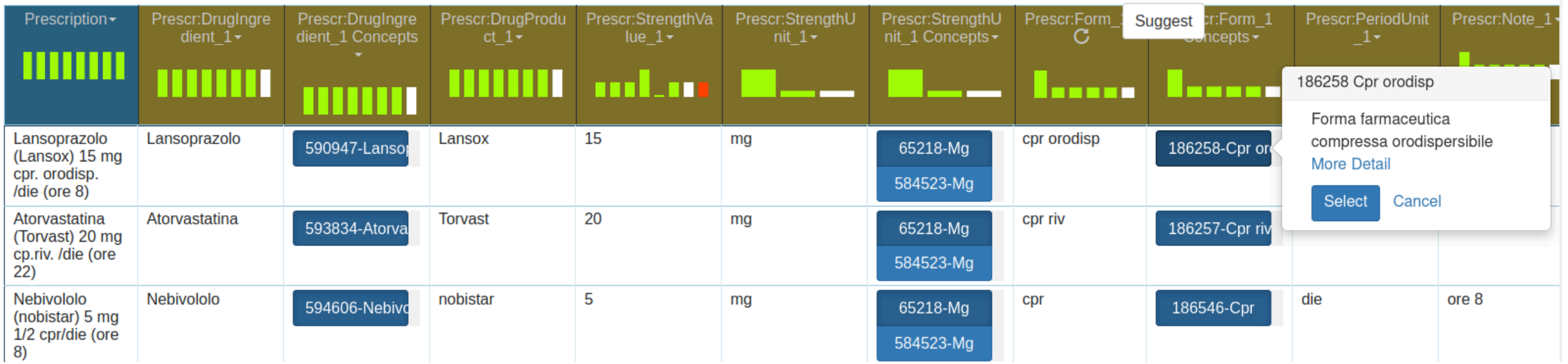
For the correctness of in-depth understanding, **human supervision is essential**.

- A human **knowledge manager** is charged with the setup, curation, and maintenance of knowledge (term bases, schemas, language models, rules, code snippets).
- A human **data manager** is charged with the setup, curation, and maintenance of the data mapping and transformation process.

How to reconcile human supervision and keep the scalability of automation?

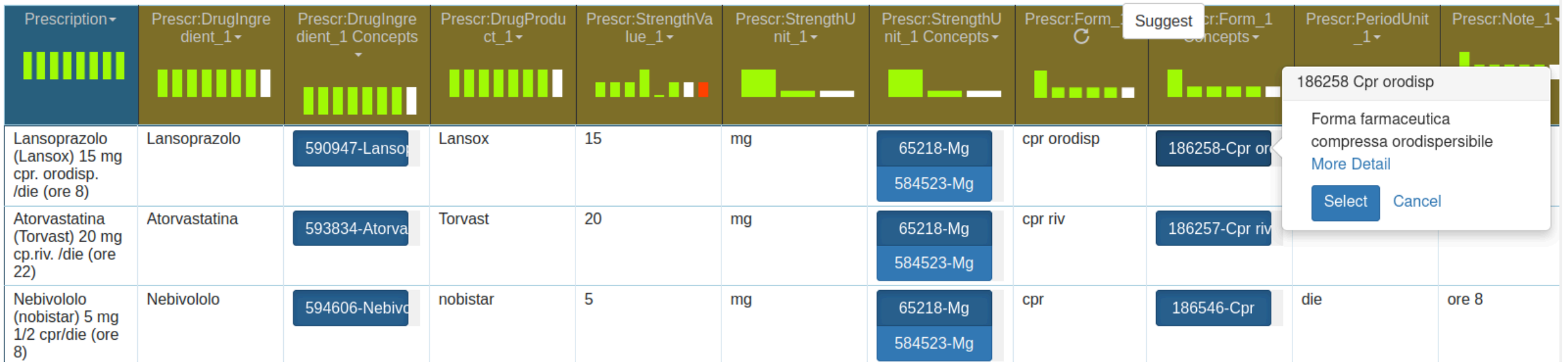
We propose a **set of knowledge and data management tools**, integrated into a single graphical interface.

A HUMAN-CENTRED, YET AUTOMATED METHODOLOGY



1. The example-driven ETL paradigm is extended by native multilingual support, semantic operations, and knowledge graph building.
2. All operations of knowledge management and data processing (setup, curation, maintenance) are executed through an integrated graphical interface.

A HUMAN-CENTRED, YET AUTOMATED METHODOLOGY



3. All automated results (e.g. of information extraction) can be manually curated.
4. Input from the data manager is recorded and automatically replayed on large datasets.
5. Ongoing research: the system learns from human curation and gradually improves performance.

CONCLUSIONS

- Full interoperability over health data is hard and will always be.
- Standards and technological innovation are necessary but not sufficient.
- Methodological innovation is needed to reconcile needs for precision, traceability, and automation.