Healthcare trajectories reconstruction with i2b2: issues and recommendations

C. Khnaisser[1,2], L. Lavoie[1], A. Burgun[2], J. F. Ethier[1,2]

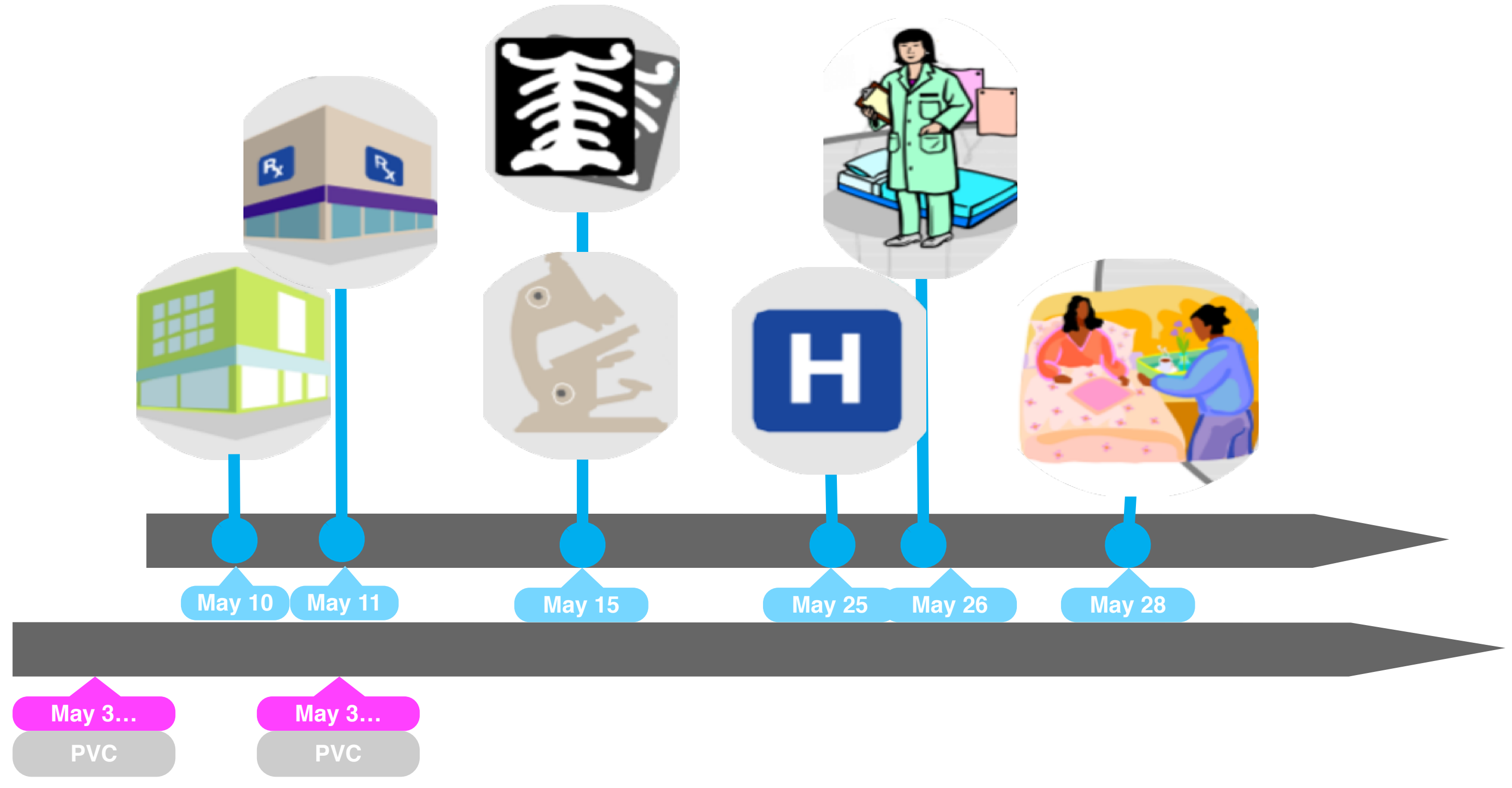
[1] GRIIS, Université de Sherbrooke, Sherbrooke, Canada; [2] INSERM, UMRS 1138, CRC, Équipe 22, Université Paris Descartes, Sorbonne Paris Cité, Paris, France;
Christina.Khnaisser@usherbrooke.ca

Context

A large volume of heavily fragmented healthcare domain data is generated from several healthcare institutions using different knowledge models, terminologies, and data models. The study of healthcare trajectories (the patient path over time and through many interactions with healthcare providers potentially from different institutions and specialties, including primary care) is one of the areas that greatly benefits from the secondary analysis of existing data in healthcare systems.
Leading clinical data warehouses (CDW) like i2b2 provides tools to integrate and query clinical data. Nevertheless, they are institution-centered and offer limited temporal query functionality. It is therefore difficult, if not impossible, to construct care trajectories in sufficient detail to enable optimal decision-making.

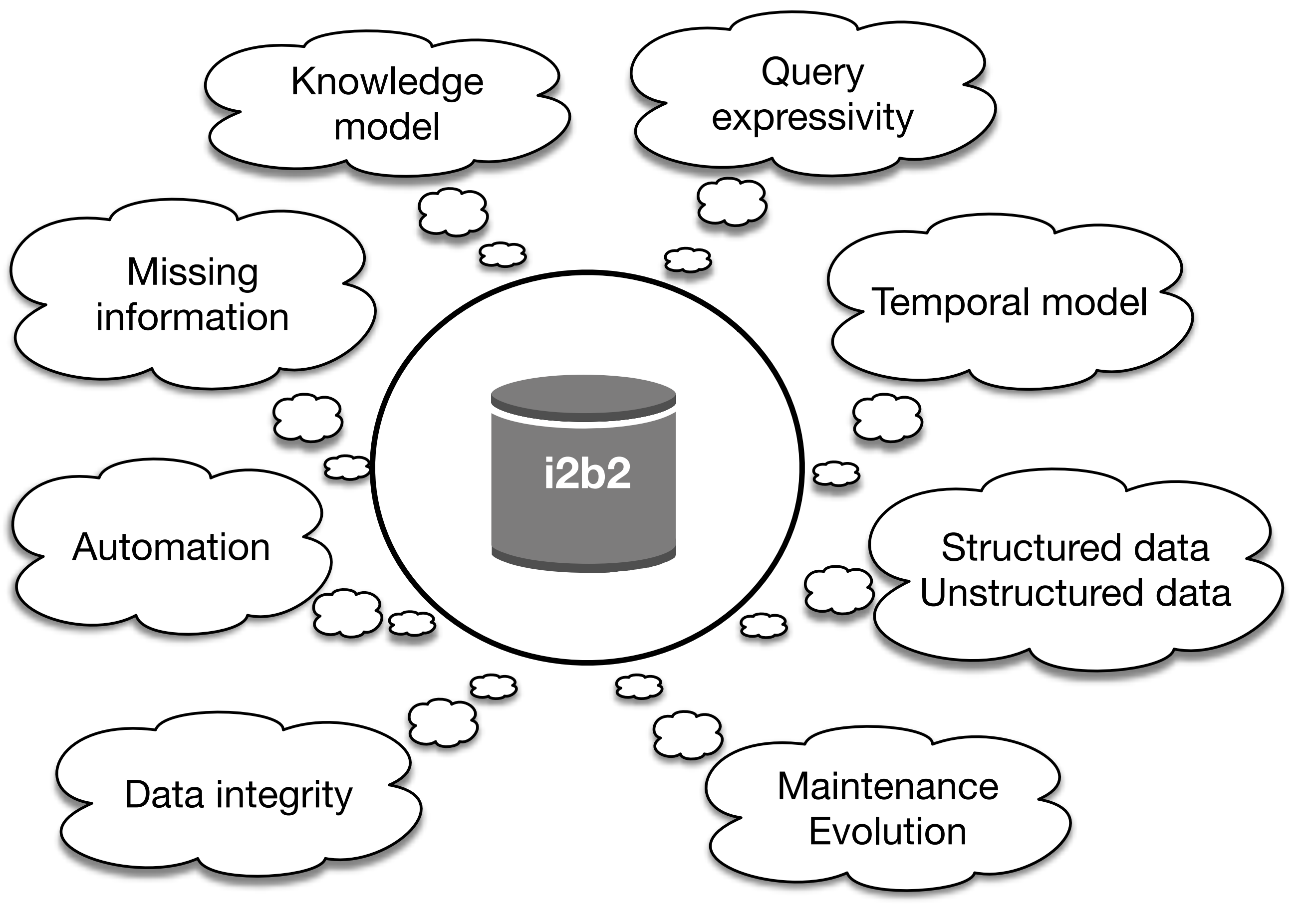
Goal

The aim of this work is to study i2b2 data warehouse schemes to identify the structure and the modelling techniques needed to go beyond patient count and build coherent and complete healthcare trajectories.



Challenges

The episode of care of a patient often occurs in interaction with several institutions at different points in time and the data can be in one or multiple CDW (i2b2 or not). The aggregation and reconstruction and patient healthcare trajectories remain a big challenge. Several challenges limit the optimal use of i2b2.



Issue, consequences and recommandations

Data model:

Issues:

- Absence of sound axiomatically based temporal model. Temporal queries are not always reproducible.
- Absence of uniform approach to handle or ignore missing information.
- Post-processing i2b2 query results (pivoting process) often required for data analytics tools [7, 1].
- Absence of a uniform structure for facts: two facts with the same concept code (proposition) can have different modifier code (attributes).

Consequences:

- The traceability of the data evolution is inadequate and insufficient for reconstructing a coherent and complete timeline view of episodes of care [6].
- Temporal queries are often not reproducible because the bitemporal model is not handled adequately. Moreover, most used temporal models are not compatible with EAV design.
- Missing data is not recorded or stored using default values increases query complexity and post-processing.
- Data integrity is compromised by data redundancy and the absence of constraints to validate data. Constraints are managed externally by the ETL process where the semantics are encapsulated in the code and not accessible outside.
- Not normalized schema does not benefit from new performance relational engines: in-memory database (Vertica, Hanna, VoltDB) or parallel processing database (GreenPlum).

Recommendations:

- Use a sound axiomatically based temporal models using well-known temporal semantics such as valid and transaction time. Adding two temporal dimensions for each attribute ensures a coherent evolution of each value independently and ensures queries reproducibility in different points in time [5].
- Increase data integrity by defining formally attributes (modifiers) constraints and their functional dependencies.
- Handle missing information uniformly by design techniques. Represent known and unknown data explicitly [8].
- Normalize the data schema for easy validation of the data, extension of the model and track data evolution over multiple temporal dimensions.

Knowledge model:

Issues:

- Absence of complete and systematic distinction between terminologies, taxonomies and knowledge model.
- Absence of a common well-documented knowledge model and uniformly applicable. I2b2 metadata is not a knowledge model. It's a data dictionary that presents simple syntactic definition of the structure and hierarchy of the data repository [3].
- Absence of a mechanism allowing the definition of links between events recorded in i2b2 repository [2].

Consequences:

- Mappings and synchronization between data and metadata require a considerable effort made manually outside the database.
- Semantic links between events or attributes (modifiers) cannot be coherently preserved extended or upgraded.
- Reasoning done manually by text mining (in concept path) is neither reliable nor effective.
- Hierarchical representation in a relational database of the i2b2 metadata is very strict and hard to query (text search), extend and upgrade.

Recommendations:

- Use a knowledge model based on realism paradigm [4] allows formal definition between concepts using axioms thus increasing semantic interoperability.
- Use properly relational modelling techniques thus the data catalogue can be generated and controlled by the database management system.
- Use OWL and RDF representation of the ontology to benefit from the reasoning tools. OWL entity is more reliable than text mining.

Conclusion

I2b2 is a leading CDW but query complexity remains high, historical data preservation is minimal and data validation is hard. All the effort of semantic-data correspondence, verification, validation and evolution is carried out outside the database in an ad hoc manner, therefore, it is hard to consume outside the i2b2 ecosystem. Constructing healthcare trajectories across different facilities over episodes of care requires a sound modelling techniques unifying the knowledge model and the temporal data model.

[1] Sebastian, M., Ixchel, C., Thomas, G., Hans-Ulrich, P., and Stefan, K. 2017. Standards-Based Procedural Phenotyping: The Arden Syntax on i2b2. Studies in Health Technology and Informatics, 37–41.

[2] Haarbrandt, B., Tute, E., and Marschollek, M. 2016. Automated population of an i2b2 clinical data warehouse from an openEHR-based data repository. Journal of Biomedical Informatics 63, 277–294.

[3] Westra, B.L., Christie, B., Johnson, S.G., et al. 2016. Expanding Interprofessional EHR Data in i2b2. AMIA Summits on Translational Science Proceedings 2016, 260–268.

[4] Smith, B. 2015. Basic Formal Ontology 2.0: Specification and user's guide.

[5] Date, C.J., Darwen, H., and Lorentzos, N. 2014. Time and Relational Theory: Temporal Databases in the Relational Model and SQL. Morgan Kaufmann.

[6] Defossez, G., Rollet, A., Dameron, O., and Ingrand, P. 2014. Temporal representation of care trajectories of cancer patients using data from a regional information system: an application in breast cancer. BMC Medical Informatics and Decision Making 14, 1, 24.

[7] Ganslandt, T., Mate, S., Helbing, K., Sax, U., and Prokosch, H.U. 2011. Unlocking Data for Clinical Research – The German i2b2 Experience. Applied Clinical Informatics 2, 1, 116–127.

[8] Date, C.J. and Darwen, H. 2010. Database Explorations: essays on the Third Manifesto and related topics. Trafford Publishing.