HAL
open science

Storing biological data in i2b2:Extending the representation of laboratory exams beyond the sheer result

Vincent Looten, Chistina Khnaisser, Anita Burgun, Bastien Rance

To cite this version:

Vincent Looten, Chistina Khnaisser, Anita Burgun, Bastien Rance. Storing biological data in i2b2:Extending the representation of laboratory exams beyond the sheer result. i2b2 and tranSMART european meeting, Oct 2017, Paris, France. hal-01702475

HAL Id: hal-01702475

https://hal.science/hal-01702475v1

Submitted on 6 Feb 2018

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Storing biological data in i2b2:Extending the representation of laboratory exams beyond the sheer result

V. Looten[1,2,3], C. Khnaisser[1,2,4], A. Burgun[1,2,3], B. Rance[1,2,3]

[1] Université Paris Descartes, Sorbonne Paris Cité, Paris, France; [2] INSERM, UMRS 1138, CRC, Équipe 22, Paris, France; [3] Hôpital Européen Georges Pompidou, APHP, Paris, France; [4] GRIIS, Université de Sherbrooke, Sherbrooke, Canada;

Introduction

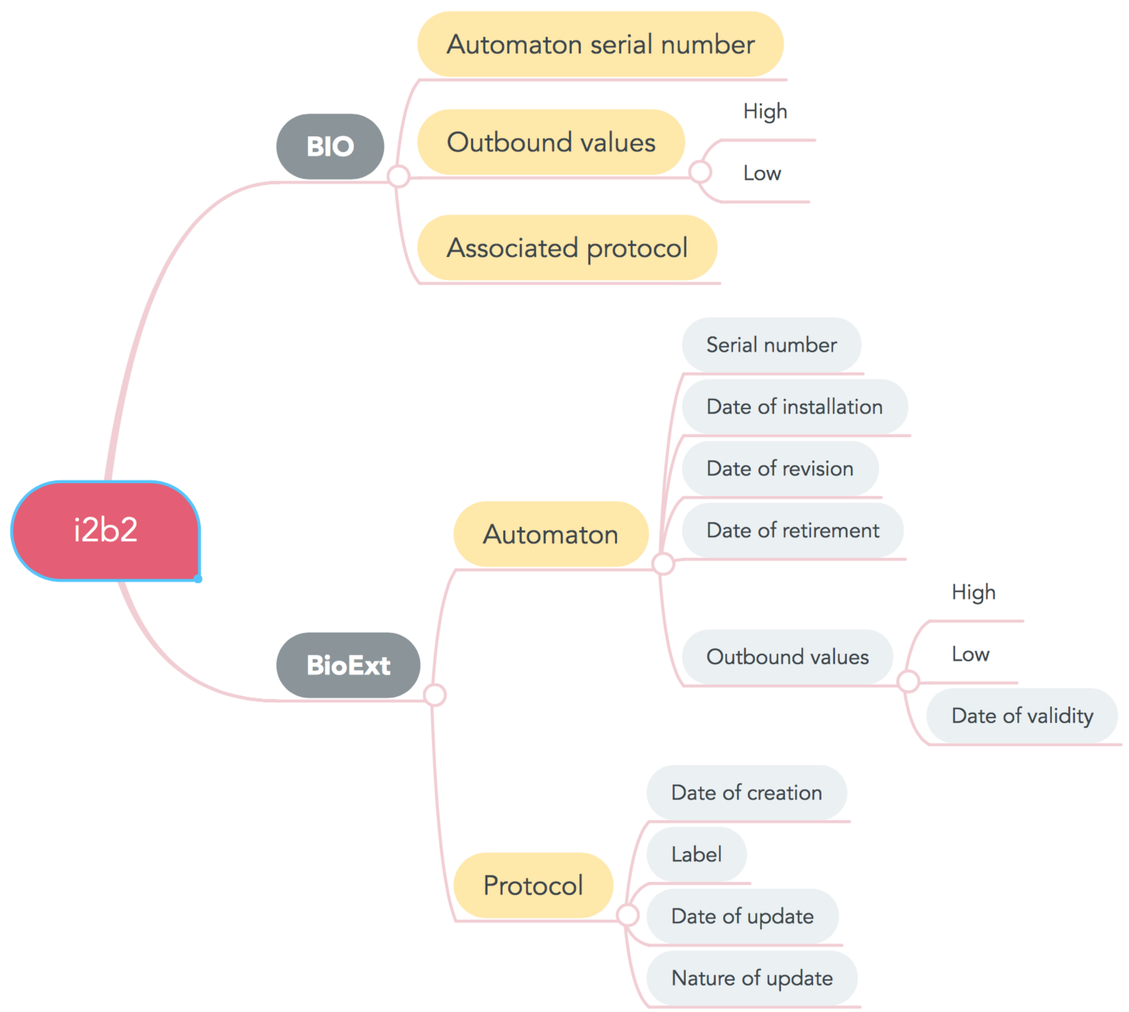
Among all the data stored worldwide in i2b2 clinical data warehouses (CDW), laboratory data are probably the most numerous. At the European Hospital Georges Pompidou (HEGP), over 17 years of activity have generated close to 150 million laboratory results integrated in our local CDW [1]. The exploitation of such a resource is a great opportunity for the clinicians, and biologists. We studies 200+ exams to track evolutions in the global profile of the data and identified a series of information required to make sense for a retrospective use, with regards to the context of generation.

Use Cases

In i2b2 CDW, each biological exam is a concept cd. When a physician prescribe an exam for a patient, hospital laboratory produce the result using a specific automaton. For a same concept, different automaton can be used. Moreover, we can group different biological concepts into a set called "protocol". A same concept can be associated with different protocol. An exam for a patient can be associated with protocol or not.

Methods

We identified: metadata related to the exam (identification of the automaton on which the exam was performed, associated normal and outbound values), metadata related to the automatons themselves (data of installation, date of retirement), and protocols in which the exam is present (date of creation and subsequent date of enrichment or simplification).



We propose two modelling approach : i2b2 modeling design [2] and a normalized design.

Conclusion

Intuitively, entity-attribute-value design (EAV) facilitates data storage or even queries, but this structure increases the number of unnecessary redundant rows. Schema normalization is required to reduce redundancy of a large volume of data such as lab results, to foster model extension and data evolution traceability over multiple hierarchical dimensions.

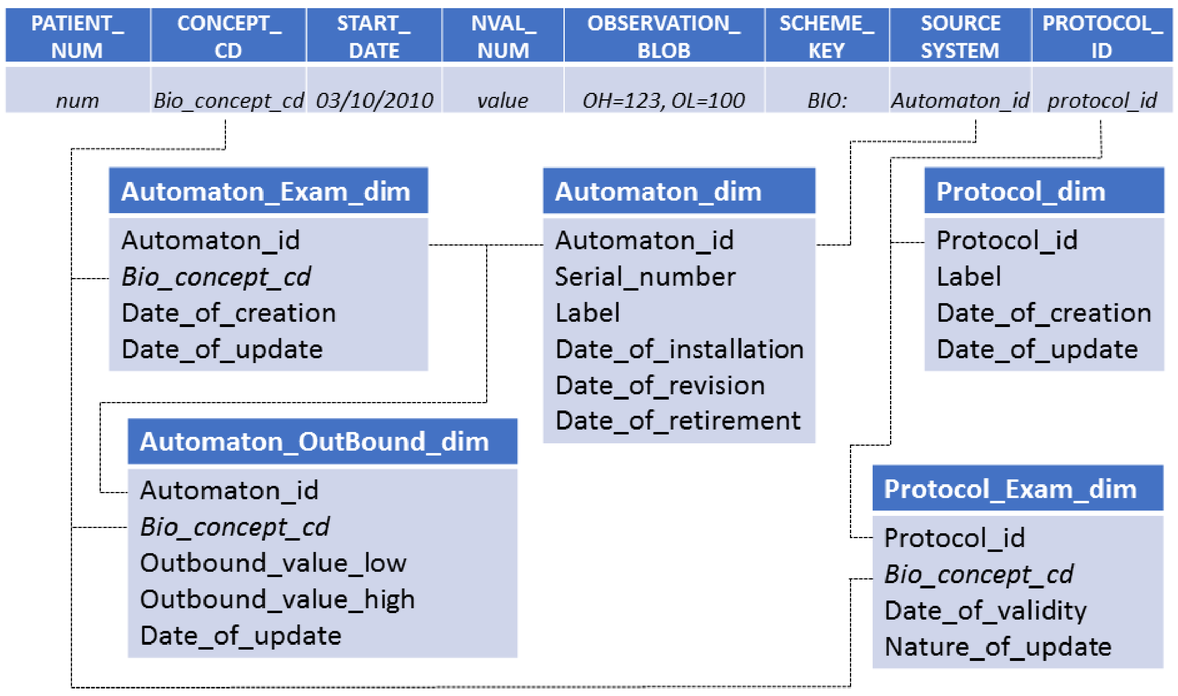
Results

I2b2 modeling design: redundancy generates high volume of data

PATIENT_NUM	ENCOUNTER_NUM	INSTANCE_NUM	CONCEPT_CD	START_DATE	MODIFIER_CD	VALTYPE_CD	TVAL_CHAR	NVAL_NUM
10001	100001	1	BIOEXA:1	01/01/2017	@	N	E	12
10001	100001	1	BIOEXA:1	01/01/2017	@	N	E	1234
10001	100001	1	BIOEXT:PROTOCOL	01/01/2017	PROTOCOL:LABEL	T	PROTOX	
10001	100001	1	BIOEXT:PROTOCOL	01/01/2017	PROTOCOL:CRE	T	01/01/2017	
10001	100001	2	BIOEXA:1	01/01/2017	@	N	E	12345
10001	100001	2	BIOEXA:1	01/01/2017	EXA:AUTOSN	T	123-123-123	
10001	100001	2	BIOEXA:1	01/01/2017	EXA:VALUE	N	E	120
10001	100001	3	BIOEXT:AUTO	01/01/2017	@	T	123	
10001	100001	3	BIOEXT:AUTO	01/01/2017	AUTO:INST	T	01/01/2007	
10001	100001	3	BIOEXT:AUTO	01/01/2017	AUTO:RETIR	T	01/01/2017	

C_FULLNAME	M_APPLIED_PATH	C_NAME	C_BASECODE
/i2b2/bio/exam/	@	Lab exam	BIOEXA:1
/i2b2/bioext/exam/automaton	@	Automaton	BIOEXT:AUTO
/i2b2/bioext/protocol/	@	Protocol	BIOEXT:PROTOCOL
/Label/	/i2b2/bioext/protocol/	Label	PROTOCOL:LABEL
/Date of creation/	/i2b2/bioext/protocol/	Date of creation	PROTOCOL:CRE
/Date of update/	/i2b2/bioext/protocol/	Date of update	PROTOCOL:UPD
/Nature of update/	/i2b2/bioext/protocol/	Nature of update	PROTOCOL:NATURE
/Automaton serial number/	/i2b2/bioext/exam/ %	Automaton serial number	EXA:AUTOSN
/Automaton outbound value/	/i2b2/bioext/exam/ %	Automaton outbound value	EXA:VALUE
/Associated Protocol/	/i2b2/bioext/exam/ %	Associated Protocol	EXA:ASSPRO
/Automaton serial number/	/i2b2/bioext/exam/automaton/	Automaton serial number	AUTO:SN
/Date of installation/	/i2b2/bioext/exam/automaton/	Date of installation	AUTO:INST
/Date of retirement/	/i2b2/bioext/exam/automaton/	Date of retirement	AUTO:RETIR
/Date of revision/	/i2b2/bioext/exam/automaton/	Date of revision	AUTO:REV

Normalization: increase structure complexity



[1] Zapletal E, Rodon N, Grabar N, Degoulet P. Methodology of integration of a clinical data warehouse with a clinical information system: the HEGP case. Stud Health Technol Inform. 2010;160(Pt 1):193–7.

[2] Murphy SN, Mendis M, Hackett K, Kuttan R, Pan W, Phillips LC, Gainer V, Berkowicz D, Glaser JP, Kohane I, and Chueh HC. Architecture of the open-source clinical research chart from Informatics for Integrating Biology and the Bedside. AMIA Annu Symp Proc 2007; 548-52.

Email : vincent.looten@aphp.fr