Multidimensional Contexts for Data Quality Assessment

Leopoldo Bertossi
Carleton University
Ottawa, Canada

Contexts and Data Quality

A table containing results of different medical tests on patients at a hospital

PatientValue

	Patient	Value	Time
1	Tom Waits	38.5	11:45/5/Sep/2011
2	Tom Waits	38.2	12:10/5/Sep/2011
3	Tom Waits	38.1	11:50/6/Sep/2011
4	Tom Waits	38.0	12:15/7/Sep/2011
5	Tom Waits	110/70	11:45/8/Sep/2011
6	Lou Reed	37.9	12:10/5/Sep/2011

Is this quality data?

If not, is there anything to clean? What?

We do not know ... It depends ...

Actually the table is supposed to contain *test results that are taken with instruments of the brand B_1*

Are these quality data? We still do not know ...

Questions about the quality of this data make sense in a broader setting

The quality of the data depends on "the context"

A context that allows us to, e.g.:

- make sense of data
- assess data quality
  (in this work wrt the expected/intended meaning or sense)
- in particular, do dimensional data assessment
- support data cleaning

For data quality assessment, an external context can provide the necessary information

The database under assessment is mapped into the context, for further data quality analysis, imposition of quality requirements, and cleaning

Contexts So Far

We find the term "context" in several places in computer science: databases, semantic web, KR, mobile applications, ...

Usually used for "*context aware* ... search, databases, applications, devices, ..."

Most of the time there is no explicit notion of context, but some mechanisms that take into account (or into computation) some contextual notions

Usually, time and geographic location, i.e. particular *dimensions*, but not much beyond

In our opinion, there is a lack of fundamental research in the area, specially for data management

Precise and formalized notions of context are rather absent

Contexts that can be implemented and used in a principled manner in data management systems

Some existing research:

- Contexts in ontologies and SW

  Lately with emphasis on using logic programs to "bridge" implicit contexts

  Impact on data management still pending

- Contexts in KR

  They are denoted at the object level and a theory specifies their properties and dynamics

  It is possible to talk about things holding in certain (named) contexts

- Contexts in data management

  Usually in connection with specific dimensions of data, like time and place

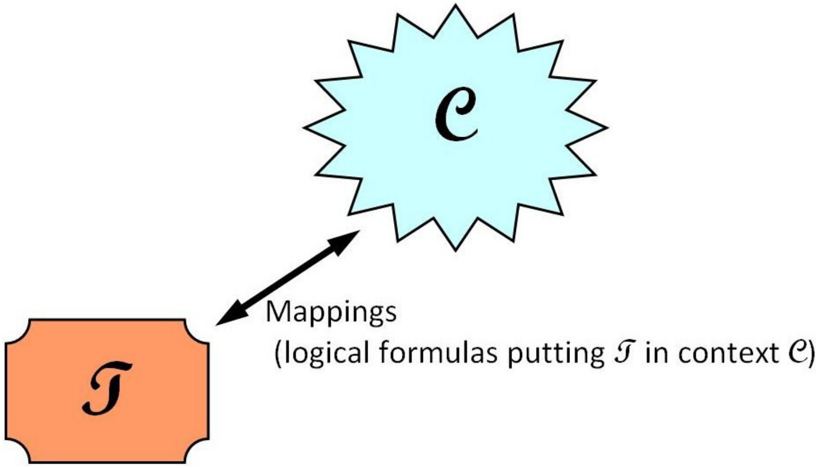
  Relevant specific research has been carried out

  (Tanca et al., Torlone-Martinenghi, Spyratos et al., ...)

  A unifying framework seems to be missing

  A general notion and theory of context have still to be developed

Contexts: A Vision


$\mathcal{C}$
$\mathcal{T}$
Mappings
(logical formulas putting $\mathcal{T}$ in context $\mathcal{C}$)

- A logical theory $\mathcal{T}$ is the one that is "put in context"
- The context is another logical theory, $\mathcal{C}$

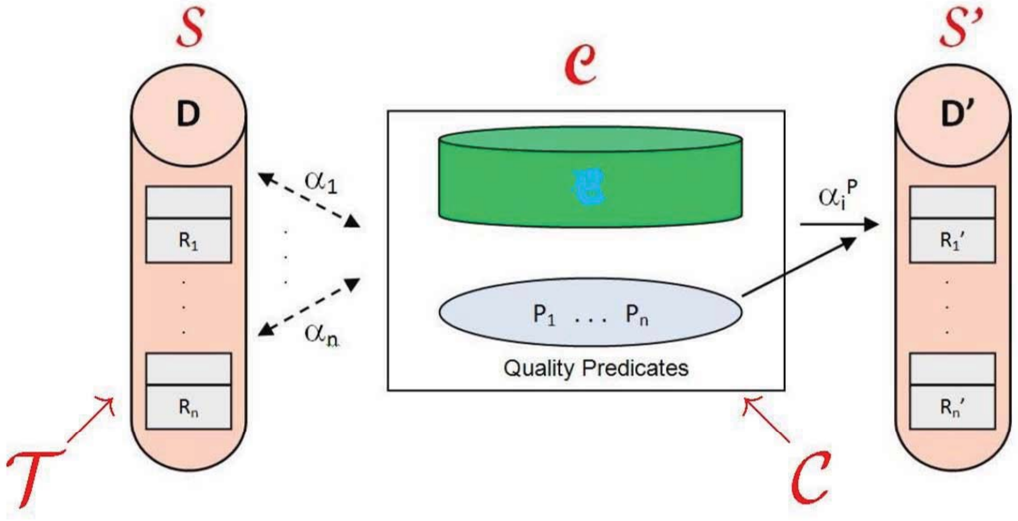
  $\mathcal{T}$ and $\mathcal{C}$ may share some predicate symbols
- Connection between $\mathcal{T}$ and $\mathcal{C}$ is established through connection predicates and mappings

In particular, for applications in data management

Contexts: Data Quality Assessment

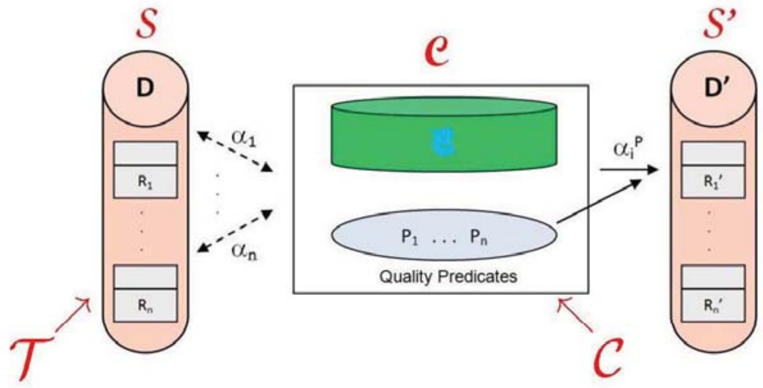
A data quality scenario:

(Bertossi, Rizzolo & Lei; VLDB'10 BIRTE WS, Springer LNBIP 48, 2011)

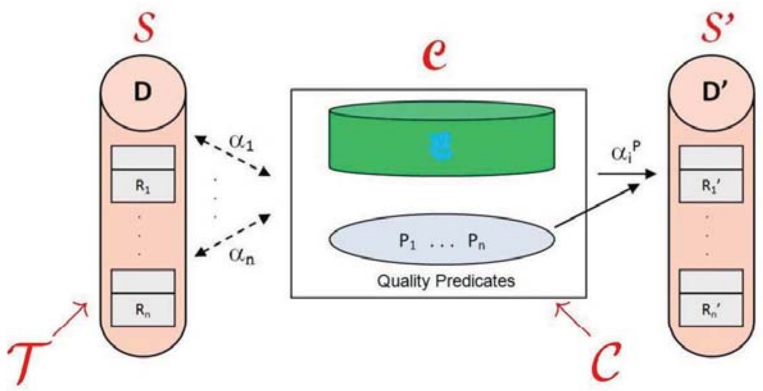


Database D can be seen as a logical theory, e.g. Reiter's logical reconstruction of a relational DB

Context $\mathcal{C}$ can be a whole ontology: Ontology-based data quality assessment and cleaning



- Instance D under assessment
- Schema $\mathcal{S}'$ a copy of $\mathcal{S}$
- Context $\mathcal{C}$: virtual/(semi)materialized data integration system
- The α_i: mappings, as in VDISs or data exchange
- In $\mathcal{C}$: Contextual predicates/relations C_j, plus quality predicates P_k
- D' contains "ideal" contents for relations in D, as views



- Predicates in D' can be materialized with data from D and $\mathcal{C}$, and "logical massage" via $\mathcal{C}$ (mapping composition)

- Quality-aware (QA) query answering about (or from) $\mathcal{S}$ can be done on top of D'

Techniques for query answering in VDISs can be applied (specially if D' is not materialized)

- Quality assessment of D can be done by comparing its contents with D' (there are some measures)

- More precisely and generally, given D and $\mathcal{C}$, there may be a class $\mathcal{I}$ of admissible contextual instances for $\mathcal{C}$'s schema, and correspondingly multiple D's

D has to be compared with the class thereof ...

There are some distance measures (see below)

A particular case of QA query answering

Different cases, some of them ...

Measurements (contextual)

	Patient	**Value**	**Time**	**Date**	**Instr**
1	T. Waits	37.8	11:00	Sep/5	Oral Therm.
2	T. Waits	38.5	11:45	Sep/5	Tympanal Therm.
3	T. Waits	38.2	12:10	Sep/5	Oral Therm.
	...	...	...	...	...
4	T. Waits	110/70	11:00	Sep/6	BPM
5	T. Waits	38.1	11:50	Sep/6	Oral Therm.
6	T. Waits	38.0	12:15	Sep/6	Oral Therm.
	...	...	...	...	...
7	T. Waits	37.6	10:50	Sep/7	Tympanal Therm.
8	T. Waits	120/70	11:30	Sep/7	BPM
9	T. Waits	37.9	12:15	Sep/7	Oral Therm.

Example: (the simple case) A contextual instance $Measurements$

Initial table $PatientValue$ (page 2, the R in D) is a view of $Measurements$, with mapping α

$$PatientValue(p, v, t, d) \longleftarrow Measurements(p, v, t, d, i)$$

Here, $\mathcal{I} = \{I\}$, a single admissible contextual instance

In this case, a single, given, materialized contextual instance

D is a "footprint" of I (kind of ... quality not considered ...)

Now we impose quality requirements: (the R' and α^P above)

$$PatientValue'(p,v,t,d) \longleftarrow Measurements(p,v,t,d,i), \\ 11{:}30 \ \leq \ t \ \leq \ 12{:}30, \ i = oral\ therm$$

Here, $R'(I) \subseteq R(D)$, and $\Delta(R(D), R'(I))$ indicates how initial $R(D)$ departs from quality instance $R'(I)$

$$PatientValue'(I) \ \subsetneqq \ PatientValue(D)$$

Quality query answering? (conjunctive queries)

$$\begin{tikzcd} & \mathcal{Q} \in L(\mathcal{S}) \arrow[dl] \arrow[rr, "\mapsto\mapsto\mapsto", "(R \mapsto R')"'] & & \mathcal{Q}' \in L(\mathcal{S}') \arrow[dr] & \\ R(D) & & & & \underbrace{\qquad\qquad R'(I)}_{\text{Or} \qquad \downarrow} \end{tikzcd}$$

View unfolding: $\mathcal{Q}' \mapsto \mathcal{Q}'' \in L(\mathcal{C}) \ \rightarrow I$

Here: $\mathcal{Q}''(I) \subseteq \mathcal{Q}(D)$, as expected (monotone query and additional conditions)

Here, the idea is that the database at hand is a projection of an expanded, contextual database

We work with the latter, imposing on it additional quality requirements

Example: (revisited)

PatientValue

	Patient	Value	Time
1	Tom Waits	38.5	11:45/5/Sep/2011
2	Tom Waits	38.2	12:10/5/Sep/2011
3	Tom Waits	38.1	11:50/6/Sep/2011
4	Tom Waits	38.0	12:15/7/Sep/2011
5	Tom Waits	110/70	11:45/8/Sep/2011
6	Lou Reed	37.9	12:10/5/Sep/2011

Table expected to contain *test results taken with instruments of brand B_1*

Now a contextual relation with data about patients, wards and days

PatientWard

	Patient	Date	Ward
1	Tom Waits	5/Sep/2011	W_1
2	Tom Waits	6/Sep/2011	W_1
3	Tom Waits	7/Sep/2011	W_1
4	Lou Reed	5/Sep/2011	W_2

Also a contextual *Hospital Guideline 1: "Medical tests in ward W_1 are performed with instruments of brand B_1"*

A quality version of table PatientValue: Map original table into the context, joint with contextual table, select according to guideline, and project: a clean version obtained

PatientValue'

	Patient	Value	Time
1	Tom Waits	38.5	11:45/5/Sep/2011
2	Tom Waits	38.2	12:10/5/Sep/2011
3	Tom Waits	38.1	11:50/6/Sep/2011
4	Tom Waits	38.0	12:15/7/Sep/2011

Example: The difference with the previous case is that we have initial instance D, but there is an incomplete or missing contextual instance

Here the idea is to map D to the contextual schema, and impose there the quality requirements (expressed in a language associated to $\mathcal{C}$)

Again: $PatientValue(p, v, t, d) \longleftarrow Measurements(p, v, t, d, i)$

Data are in $PatientValue(D)$, no (or some) data for $Measurements$

Instrument i could be obtained (or not) from additional contextual data)

As in LAV: Possible several admissible instances I in $\mathcal{I}$

Then, with the quality requirements:

$$PatientValue'(p,v,t,d) \longleftarrow Measurements(p,v,t,d,i), \\ 11{:}30 \;\leq\; t \;\leq\; 12{:}30,\; i = oral\ therm$$

Possible several instances for schema S': $D'(I)$ with $I \in \mathcal{I}$ ($D'(I) \subseteq D$)

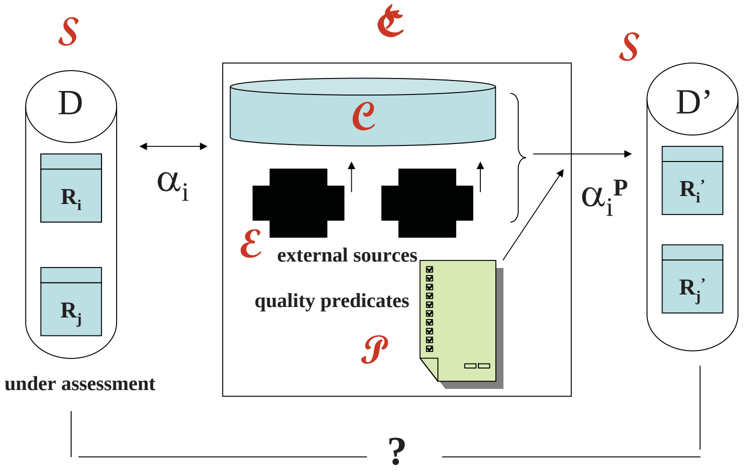
Quality of D?

Quality measure: $QM(D) := (|D| - max\{|D'(I)| : I \in \mathcal{I}\})/|D|$

Distance to a class of quality instances (computation, estimation?)

Quality query answers?: Like certain answers on $\{D'(I) \mid I \in \mathcal{I}\}$ (e.g. query rewriting via rule inversion)

Extension:

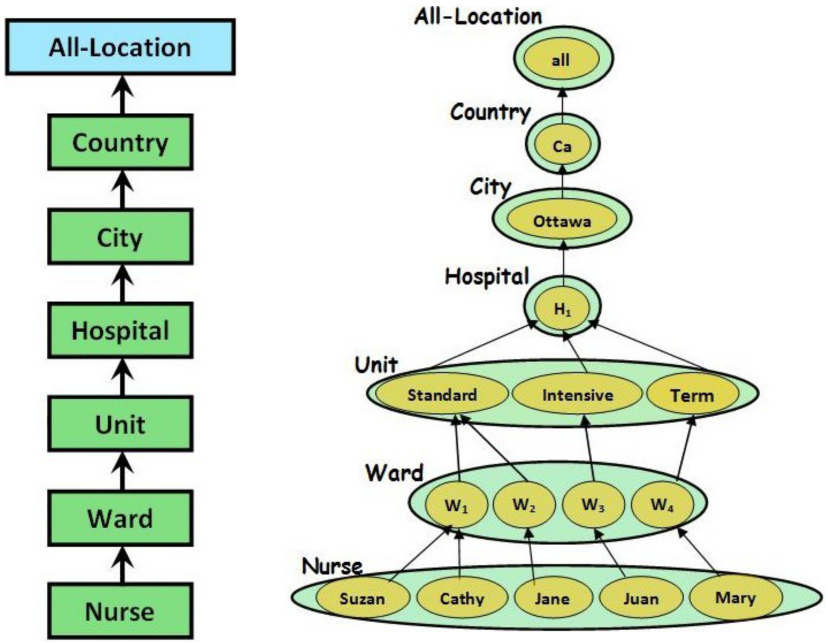


- There can also be mappings to external sources E_i

Multidimensional Contexts

Dimensions are naturally and commonly associated to contexts; and are important in data quality assessment

To bring them into contexts, we build upon the Hurtado-Mendelzon (HM) model of MDDBS



We embed an HM model into context $\mathcal{C}$

Later on we care about making the result "ontological" ...

Example: (revisited)

PatientValue

	Patient	Value	Time
1	Tom Waits	38.5	11:45/5/Sep/2011
2	Tom Waits	38.2	12:10/5/Sep/2011
3	Tom Waits	38.1	11:50/6/Sep/2011
4	Tom Waits	38.0	12:15/7/Sep/2011
5	Tom Waits	110/70	11:45/8/Sep/2011
6	Lou Reed	37.9	12:10/5/Sep/2011

Now table expected to contain *test results taken with instruments made by manufacturer* M_1

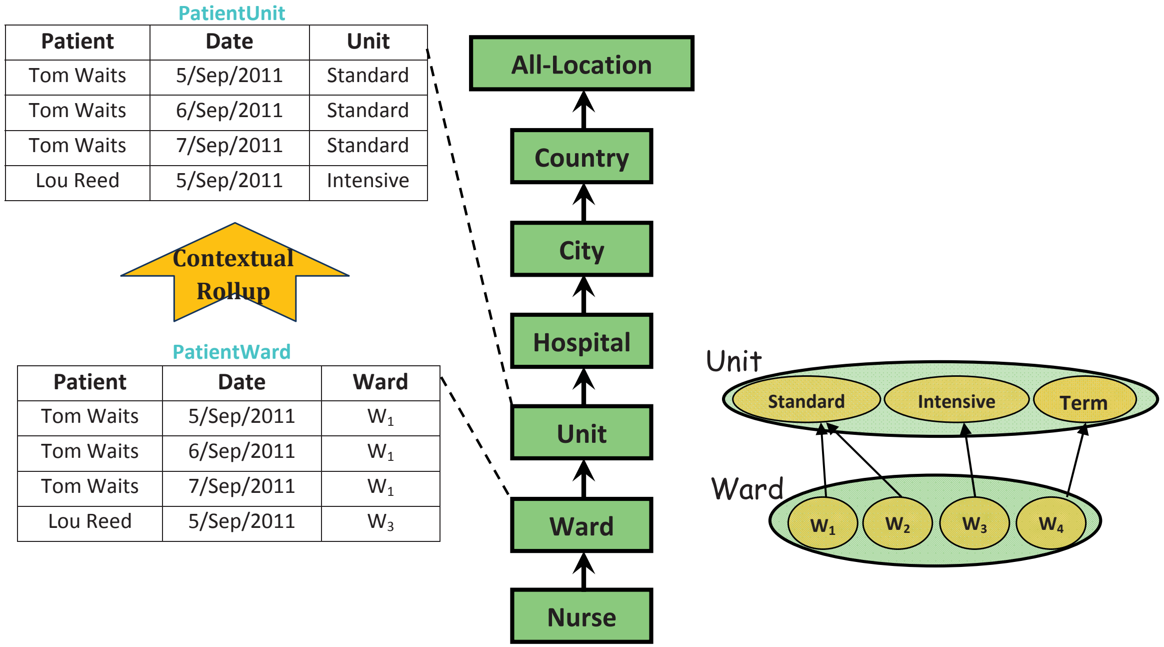
Information to make an assessment? We still have:

PatientWard

	Patient	Date	Ward
1	Tom Waits	5/Sep/2011	W_1
2	Tom Waits	6/Sep/2011	W_1
3	Tom Waits	7/Sep/2011	W_1
4	Lou Reed	5/Sep/2011	W_2

And a new contextual *Hospital Guideline 2: "Medical tests on patients in standard care unit have to be taken with instruments made by manufacturer* M_1*"*

Plus the dimensional information ...



PatientUnit

Patient	Date	Unit
Tom Waits	5/Sep/2011	Standard
Tom Waits	6/Sep/2011	Standard
Tom Waits	7/Sep/2011	Standard
Lou Reed	5/Sep/2011	Intensive

PatientWard

Patient	Date	Ward
Tom Waits	5/Sep/2011	W_1
Tom Waits	6/Sep/2011	W_1
Tom Waits	7/Sep/2011	W_1
Lou Reed	5/Sep/2011	W_3

We have data related to *Wards*, not about *Care Units* where we could apply the guideline

We roll-up via *Location* dimension from *Wards* to *Care Units*

We identify wards $W_1,\ W_2$ as belonging to the *standard CU*

Guideline 2 applies to them, inferring the tests there were taken with instruments made by manufacturer M_1

Contextual roll-up is used to access/generate missing data at certain levels, by lattice navigation

Other dimensions can be added

- generating multidimensional (MD) contextual information
- for additional and finer-granularity data quality assessment

In this direction, the HM model can be enriched

Going beyond classical applications of the DWH kind ...

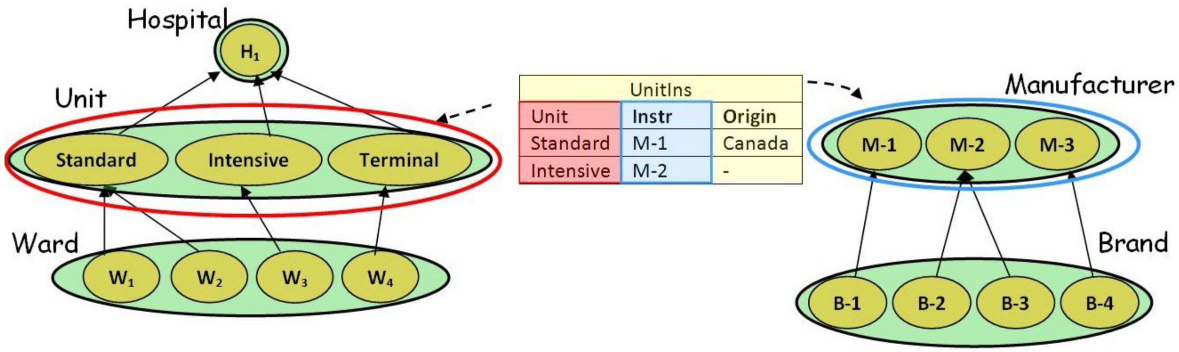
We concentrate mostly on the extension and representation of multidimensional contexts (as opposed to data quality assessment)

Extending the HM Model

In the spirit of enriching contexts with multidimensional representations, we extend the HM model

We associate predicates/relations to categories at different levels (or groups thereof) of a hierarchy (or several of them)

Categorical Relations:



UnitIns		
Unit	Instr	Origin
Standard	M-1	Canada
Intensive	M-2	-

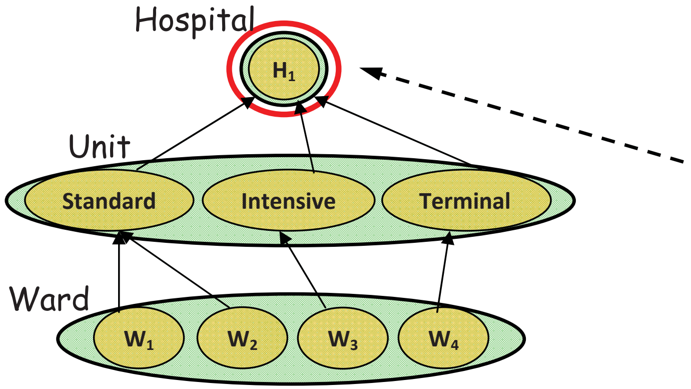
- Some attributes of a categorical relation (CR) share the domain with a dimension category

- Possibly several dimensions/categories/levels involved

- Connection between the attribute and its category via a schema mapping, e.g. $\forall u \forall i \forall o (UnitIns(u,i,o) \rightarrow Manufacturer(i))$

Attributive Relations: (particular case)

- ARs are CRs which are connected to a single category, in a single dimension schema



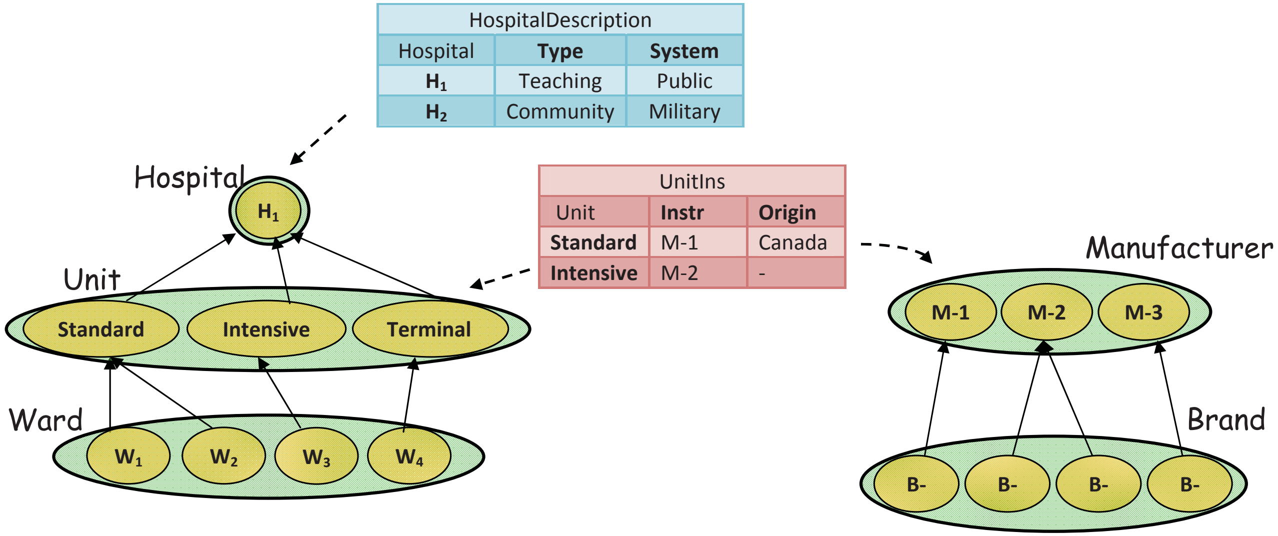
HospitalDescription		
Hospital	**Type**	**System**
H_1	Teaching	Public
H_2	Community	Military

- ARs provide descriptions for elements of a category

- Mappings as before, e.g.

$$\forall h \forall t \forall s (HospitalDescription(h, t, s) \rightarrow Hospital(h))$$

$$\forall h \exists t \exists s (Hospital(h) \rightarrow HospitalDescription(h, t, s))$$



HospitalDescription

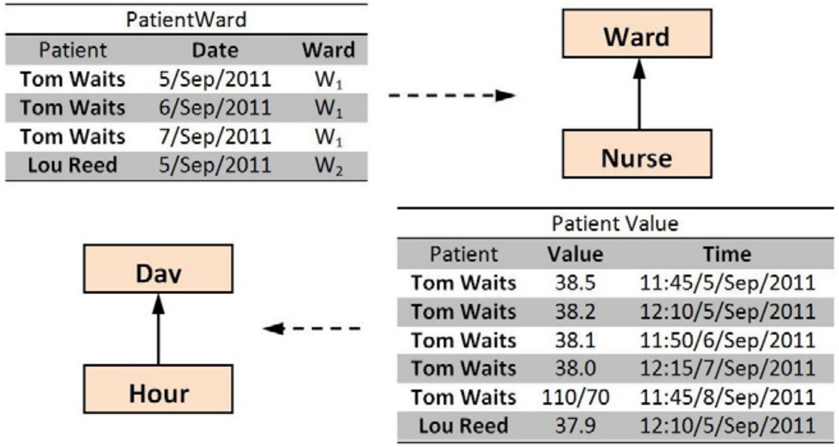
Hospital	Type	System
H_1	Teaching	Public
H_2	Community	Military

UnitIns

Unit	Instr	Origin
Standard	M-1	Canada
Intensive	M-2	-

Inter-dimensional Constraints:

Some combinations of values may not be semantically allowed in CRs and ARs



PatientWard

Patient	Date	Ward
Tom Waits	5/Sep/2011	W_1
Tom Waits	6/Sep/2011	W_1
Tom Waits	7/Sep/2011	W_1
Lou Reed	5/Sep/2011	W_2

Patient Value

Patient	Value	Time
Tom Waits	38.5	11:45/5/Sep/2011
Tom Waits	38.2	12:10/5/Sep/2011
Tom Waits	38.1	11:50/6/Sep/2011
Tom Waits	38.0	12:15/7/Sep/2011
Tom Waits	110/70	11:45/8/Sep/2011
Lou Reed	37.9	12:10/5/Sep/2011

Inter-DCs prohibit combinations of values from different dimensions involved in CRs, e.g.

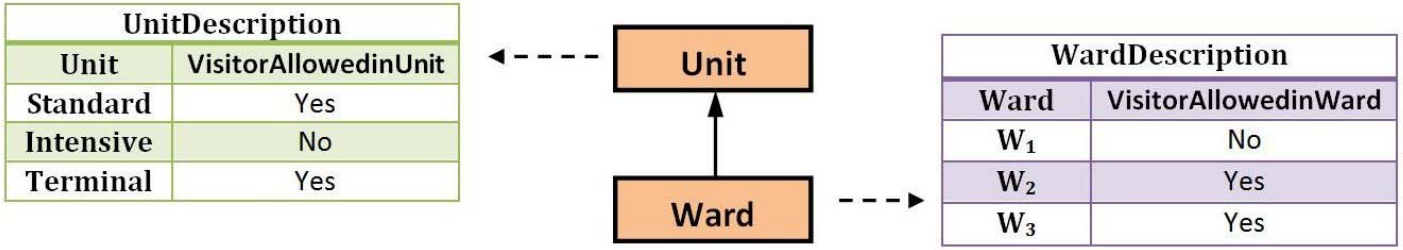
"No single measurement can be taken by more than one nurse"

As a denial constraint:

$$\neg\exists p\ v\ t\ d\ w\ n_1 n_2 (\mathit{PatientValue}(p,v,t) \wedge \mathit{PatientWard}(p,d,w) \wedge T(t,d) \wedge L(n_1, w) \wedge L(n_2, w) \wedge n_1 \neq n_2\ )$$

Involving different dimensions: *"No Czech Republic before 1989"* (Time and Geo Location)

Intra-dimensional Constraints:



UnitDescription

Unit	VisitorAllowedinUnit
Standard	Yes
Intensive	No
Terminal	Yes

WardDescription

Ward	VisitorAllowedinWard
W_1	No
W_2	Yes
W_3	Yes

Intra-DCs restrict certain combinations of descriptive values in ARs, e.g.

"No visitors allowed in wards where visitors in their units are prohibited"

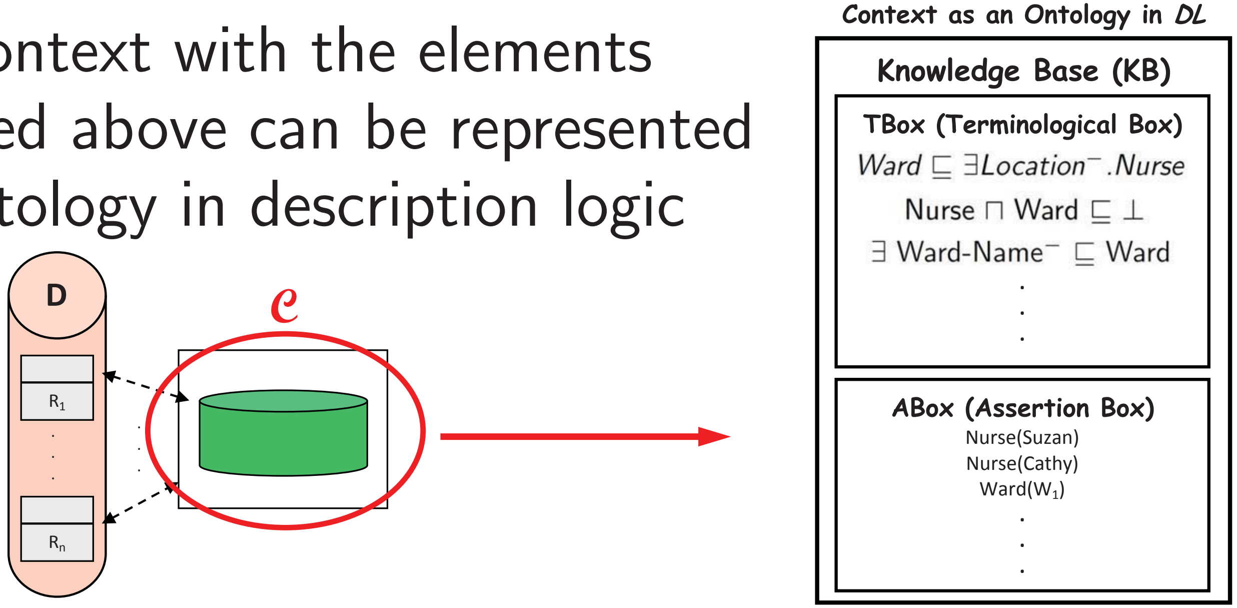
As a denial:

$$\neg \exists u\ vu\ w\ vw\ (UnitDescription(u, vu) \wedge WardDescription(w, vw) \wedge Location(w, u) \wedge vu ='\ NO' \wedge vu \neq vw)$$

“If there is an operation in a year, it must appear on a particular day of (associated to) that year”

MD Contexts as DL Ontologies

A MD context with the elements introduced above can be represented as an ontology in description logic



Context as an Ontology in *DL*

Knowledge Base (KB)

TBox (Terminological Box)

$Ward \sqsubseteq \exists Location^{-}.Nurse$

$\text{Nurse} \sqcap \text{Ward} \sqsubseteq \bot$

$\exists\ \text{Ward-Name}^{-} \sqsubseteq \text{Ward}$

.
.
.

ABox (Assertion Box)

Nurse(Suzan)
Nurse(Cathy)
Ward(W_1)
.
.
.

- Context becomes a knowledge base, an ontology, a theory in DL, containing explicit data, metadata, and rules

- Can be used to extract and generate (implicit) data

- In principle, logical reasoning becomes possible

- Choice of the DL becomes and issue

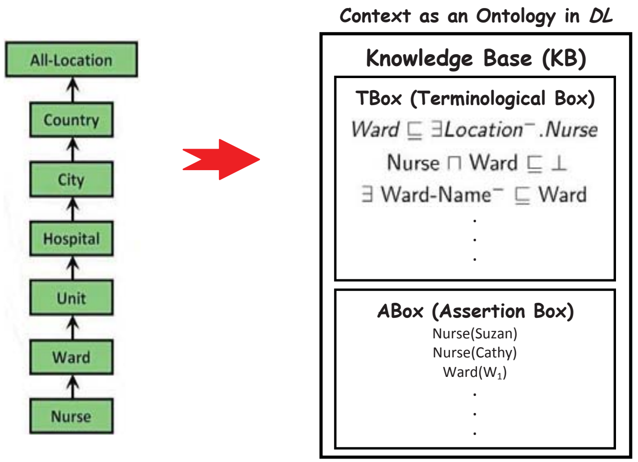
We sketch a DL-based representation of the extended HM model in one of the members of *DL-Lite* family of DLs

(Calvanese et al. JAR 2007)

Representing MD Contextual Schemas:

- Categories and attribute domains for ARs are represented as concepts: *Nurse*, *Ward*, . . . *String*, ...

- The empty concept property is for disjointness, in particular of categories: $Nurse \sqcap Ward \sqsubseteq \bot$



Context as an Ontology in *DL*

Knowledge Base (KB)

TBox (Terminological Box)

$Ward \sqsubseteq \exists Location^{-}.Nurse$

$\text{Nurse} \sqcap \text{Ward} \sqsubseteq \bot$

$\exists\ \text{Ward-Name}^{-} \sqsubseteq \text{Ward}$

. . .

ABox (Assertion Box)

Nurse(Suzan)

Nurse(Cathy)

Ward(W_1)

. . .

- Roles represent ARs, e.g. *HospitalType*, and also relationships between (elements of) two adjacent categories in a dimension (cf. page 25)

- Restrictions on attribute values in ARs, as concept inclusions

In using the number restriction $(\geq qR)$ of $\mathit{DL\text{-}Lite}^{\mathcal{N}}$ (R a role)

$\exists \mathit{HospitalType}^{-} \sqsubseteq \mathit{Hospital}, \quad \exists \mathit{HospitalType} \sqsubseteq \mathit{String}, \quad \geq 2\,\mathit{HospitalType} \sqsubseteq \bot$

- Child/Parent relationships between elements of categories ($<$) is represented by a role

E.g. the $\mathit{Location}$ (dimension) becomes a role

The $<$ relation between the elements of each two categories, e.g. Ward and Unit, is represented by: $\mathit{Unit} \sqsubseteq \exists \mathit{Location}^{-}.\mathit{Ward}$

We use role hierarchies of $\mathit{DL\text{-}Lite}^{\mathcal{HN}}$ as a basis for defining role transitive, in the extension $\mathit{DL\text{-}Lite}^{\mathcal{HN}+}$ (Artale et al., JAIR 2009)

(Role) $\mathit{Location}$ is made transitive with the axiom $\mathit{Tra(Location)}$

This is all allowed by $DL\text{-}Lite^{(HN)^+}_{Horn}$

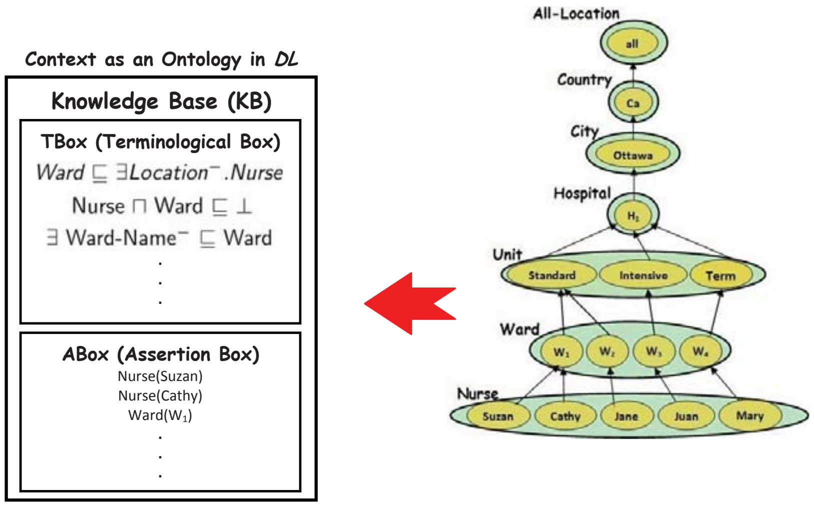
The combined complexity of $DL\text{-}Lite^{\mathcal{HN}}$ is P-complete

Representing MD Contextual Instances:

ABox not explicitly represented

TBox collects "assertive data" (facts) from the data sources via mappings

"Putting" them into concepts and roles



Context as an Ontology in *DL*

Knowledge Base (KB)

TBox (Terminological Box)

$Ward \sqsubseteq \exists Location^{-}.Nurse$

$\text{Nurse} \sqcap \text{Ward} \sqsubseteq \bot$

$\exists\ \text{Ward-Name}^{-} \sqsubseteq \text{Ward}$

⋮

ABox (Assertion Box)

Nurse(Suzan)

Nurse(Cathy)

Ward(W_1)

⋮

For example, consider $locationIns(Ward, Unit)$, the subrelation of the $Location$ dimension instance (external to TBox, a source)

Mappings building the (virtual) instances for concept $Ward$, resp. for role $Location$:

$$\forall w \forall u (locationIns(w,u) \ \rightarrow \ Ward(f_{ward}(w))$$

$$\forall w \forall u (locationIns(w,u) \ \rightarrow \ Location(f_{ward}(w), f_{unit}(u))$$

Instances of concepts, e.g. for $Wards$ and $Units$, become abstract representations of data values at the sources

Attributive relation instances?

E.g. attribute $Type$ in $HospitalDescription$ is mapped to the role $HospitalType$ through the mapping:

$$\forall h \forall t (HospitalDescription(h, t) \rightarrow HospitalType(f_{hospital}(h), t))$$

Hospital type (a string) is mapped as it is as a value at the ontological level

Representing Guidelines:

As axioms in the TBox, e.g. the *Hospital Guideline 2:*

"Medical tests on patients in standard care units have to be taken with instruments made by manufacturer M_1"

New concepts:

- $\texttt{standard}Con$: Consisting of element $\texttt{standard}$ from category $Unit$, with $\texttt{standard}Con \sqsubseteq Unit$

- $\texttt{M}_1Con$: Consisting of element $\texttt{M}_1$ from category $Manufacturer$, with $\texttt{M}_1Con \sqsubseteq Manufacturer$

- $\texttt{standard}Relate$ consisting of all locations that have $\texttt{standard}$ as ancestor in the $Unit$ category

$$\texttt{standard}Relate \equiv \exists Location.\texttt{standard}Con$$

- $\mathtt{M}_1\mathit{Relate}$ consisting of instruments with $\mathtt{M}_1$ as an ancestor in the $\mathit{Manufacturer}$ category, with

$$\mathtt{M}_1\mathit{Relate} \;\equiv\; \exists \mathit{Instrument}.\mathtt{M}_1\mathit{Con}$$

Finally, *Guideline 2* is expressed in TBox using the role $\mathit{UnitInst}$ (used for the CR on page 24):

$$\exists \mathit{UnitIns}^{-}.\mathtt{standard}\mathit{Relate} \sqsubseteq \mathtt{M}_1\mathit{Relate}$$

Conclusions

We have concentrated on the developments on MD contexts

We extended the MD model of data

We represented the resulting MD contexts in DL

In parallel we are also investigating $Datalog^{+}_{-}$ to represent the ontology

And to extend it, for generating the implicit data through a chase procedure; navigating towards the required data ...

Next step is about using MD contexts for data quality assessment, data cleaning, and quality query answering

Ours is a long term general research, about Ontology-Based Data Quality