Carleton
UNIVERSITY

Contexts for Data Quality Assessment and Cleaning

Leopoldo Bertossi⋆
Carleton University
Ottawa, Canada

⋆: Faculty Fellow IBM CAS

Contexts and Data Quality

A table containing data about the temperatures of patients at a hospital

TempNoon

	Patient	Value	Time	Date
1	Tom Waits	38.5	11:45	Sep/5
2	Tom Waits	38.2	12:10	Sep/5
3	Tom Waits	38.1	11:50	Sep/6
4	Tom Waits	38.0	12:15	Sep/6
5	Tom Waits	37.9	12:15	Sep/7

Is this quality data?

If not, is there anything to clean? What?

(Join work with Flavio Rizzolo)

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

Actually the table is supposed to contain *temperature measurements for Tom taken at noon by a certified nurse with an oral thermometer*

Is this quality data? We still do not know ...

Maybe we can say something about the time

Maybe good enough for the time to be "around noon" (meaning?)

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:

- make sense of the data
- assess the data
- on that basis, support data cleaning
- etc. (see below)

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 semantic web

  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., ...)

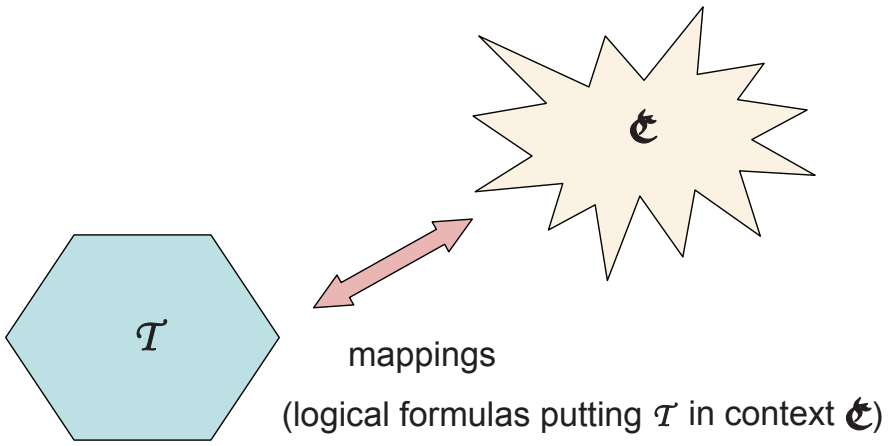
  A unifying framework seems to be missing

Contexts: A Vision

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

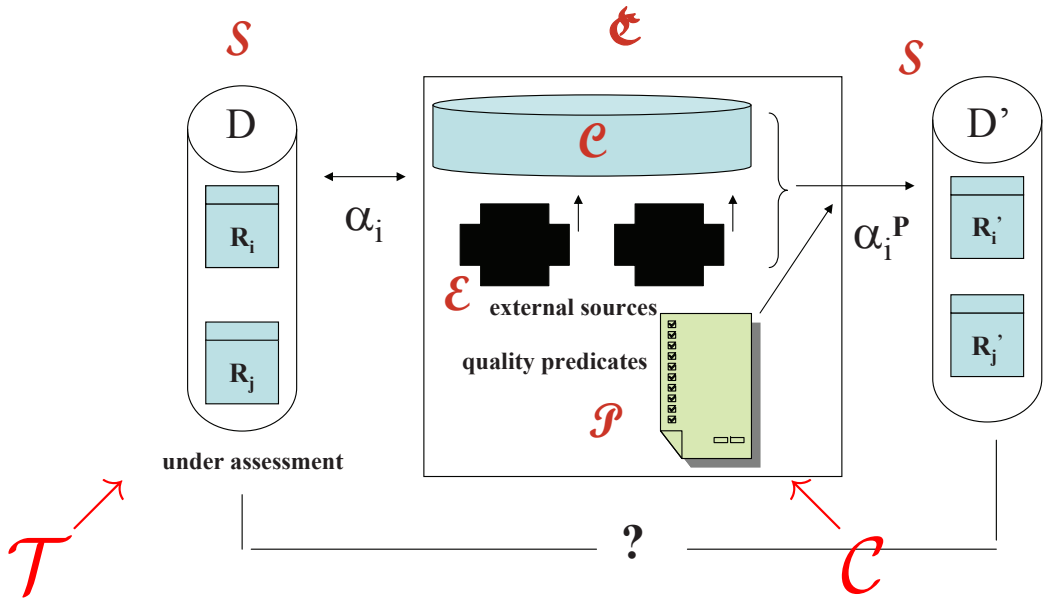
We envision it as follows:

- A logical theory $\mathcal{T}$ is the one that has to be "put in context"
  For example, a relational database can be seen as a theory
- The context is another logical theory, $\mathcal{C}$
  $\mathcal{T}$ and $\mathcal{C}$ may share some predicate symbols
- Actually, the connection between $\mathcal{T}$ and $\mathcal{C}$ is established through: connection predicates and mappings



In particular for applications in data management

In our data quality scenario: (VLDB'10 BIRTE WS, Springer LNBIP 48, 2011)



(coming ...)

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

In general, a contextual theory $\mathfrak{C}$ and mappings and their logical/computational processing have to support what we expect from a context

- Capturing and narrowing down semantics
  - By defining in $\mathcal{C}$ predicates that are used in $\mathcal{T}$ (e.g. "time close to noon")
  - Contributing in $\mathcal{C}$ with additional constraints for predicates used in $\mathcal{T}$, e.g. integrity constraints for table TempNoon)
  - Term disambiguation
- Dimensions for analysis and understanding of $\mathcal{T}$'s knowledge (generalizing multidimensional DBs, DWHS)

Why not more ambitious?

- Specifying and using notions of relevance
- Explanation, diagnosis, causality
- Capturing commonsense assumptions and practices

Research has been done lately, mainly around ontologies

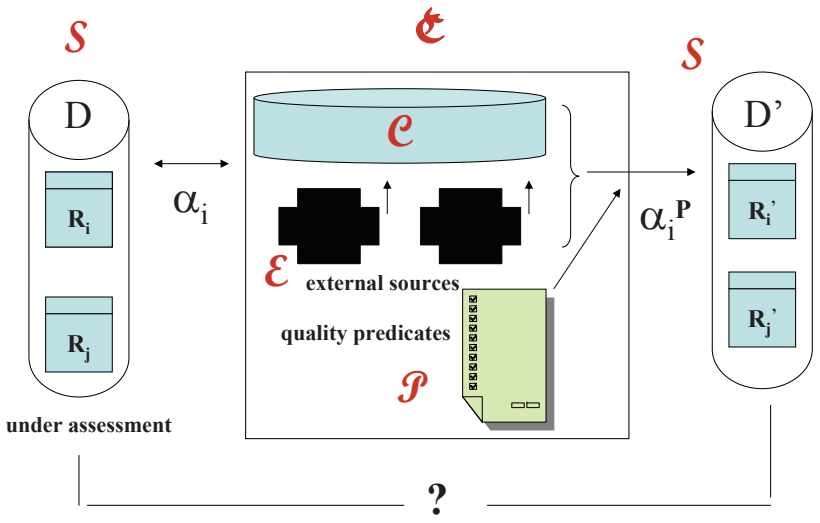
Has to be applied in data management

Making it accessible to "practical" DB people

There is interest in industry

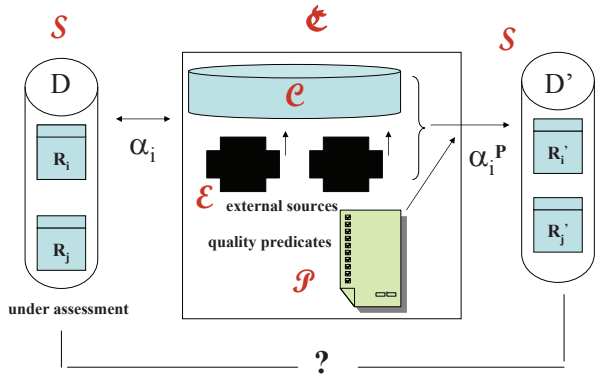
- Assessment, e.g. quality

Contexts in Data Quality Assessment



- Instance D is under assessment
- On RHS, also schema $\mathcal{S}$ (or copy $\mathcal{S}'$)
- Context $\mathfrak{C}$ is like a virtual/(semi)materialized data integration system

- The α_i are the mappings, like in VDISs or data exchange
- The C_i are contextual predicates/relations
- There are mappings to external sources E_i and quality predicates/relations P_i
- D' contains "ideal" contents for relations in D, as views



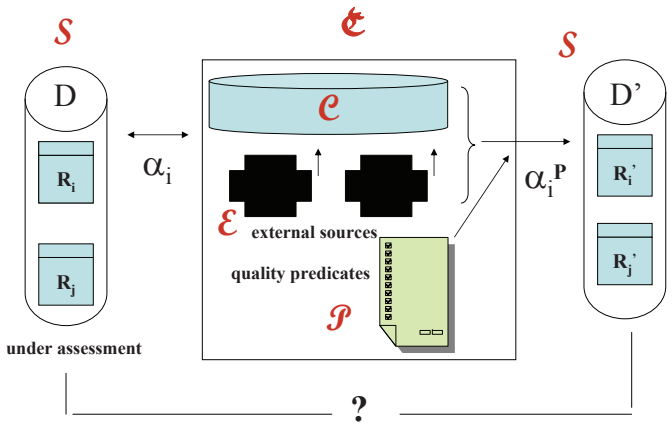
- Predicates in D' can be materialized through data in the R_i and additional massage via $\mathcal{C}$ (mapping composition at work)

- 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)

A particular case of QA query answering



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.

More concretely, given the data in D and $\mathfrak{C}$, there may be a class $\mathcal{I}$ of admissible contextual instances I for schema $\mathcal{C}$

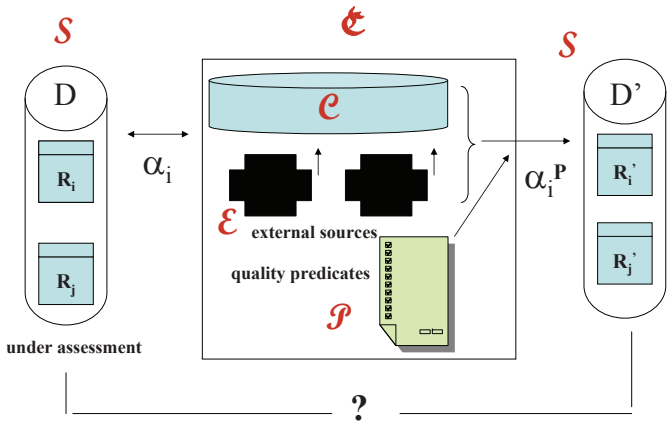
Different cases, some of them ...

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

Initial table $TempNoon$ (page 37, the R in D) is a view of $Measurements$, with mapping α

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

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



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.

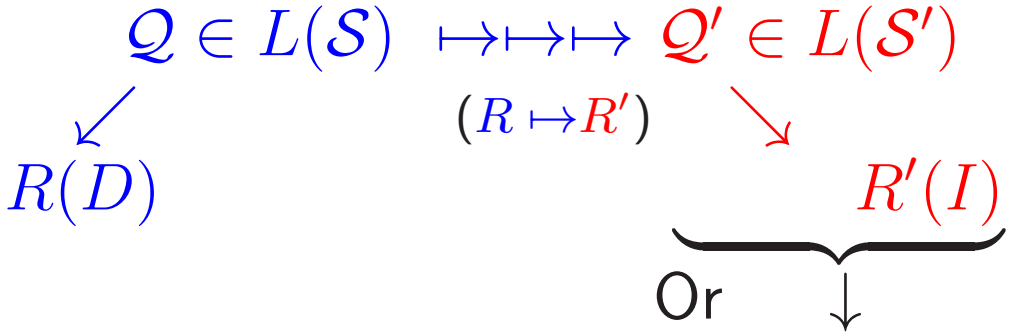
Now we impose quality requirements: (the R' and α^P above)

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

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

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

Quality query answering? (conjunctive queries)



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

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: 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 $\mathfrak{C}$)

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

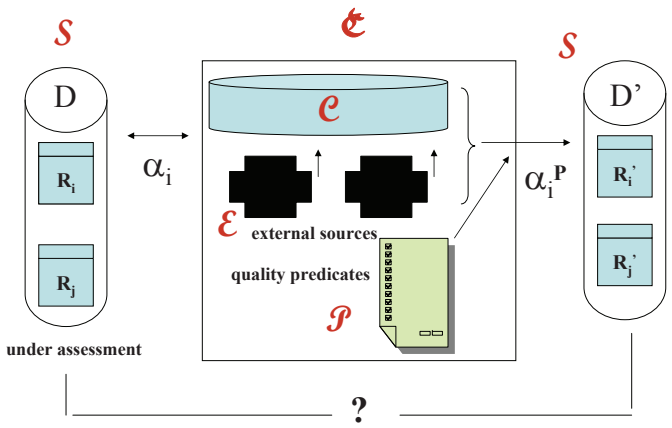
Data are in $TempNoon(D)$, no (or some) data for $Measurements$

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:

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



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)

Multidimensional Contexts

Temperature data at a hospital

Doctor requires temperatures taken with oral thermometer

Doctor expects this to be reflected in the table, but the latter does not contain the information to make this assessment

TempNoon@Ward

Patient	Value	Time	Date	Ward
Tom Waits	38.5	11:45	Sep/5	1
Tom Waits	38.2	12:10	Sep/5	1
Tom Waits	38.1	11:50	Sep/6	1
Tom Waits	38.0	12:15	Sep/6	1
Tom Waits	37.9	12:15	Sep/7	1
Lou Reed	37.9	12:10	Sep/5	2
Lou Reed	37.6	12:05	Sep/6	2
Lou Reed	37.6	12:05	Sep/7	2

An external context can provide that information, making it possible to assess the given data

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

We can see the context as an ontology

- Hospital guideline:

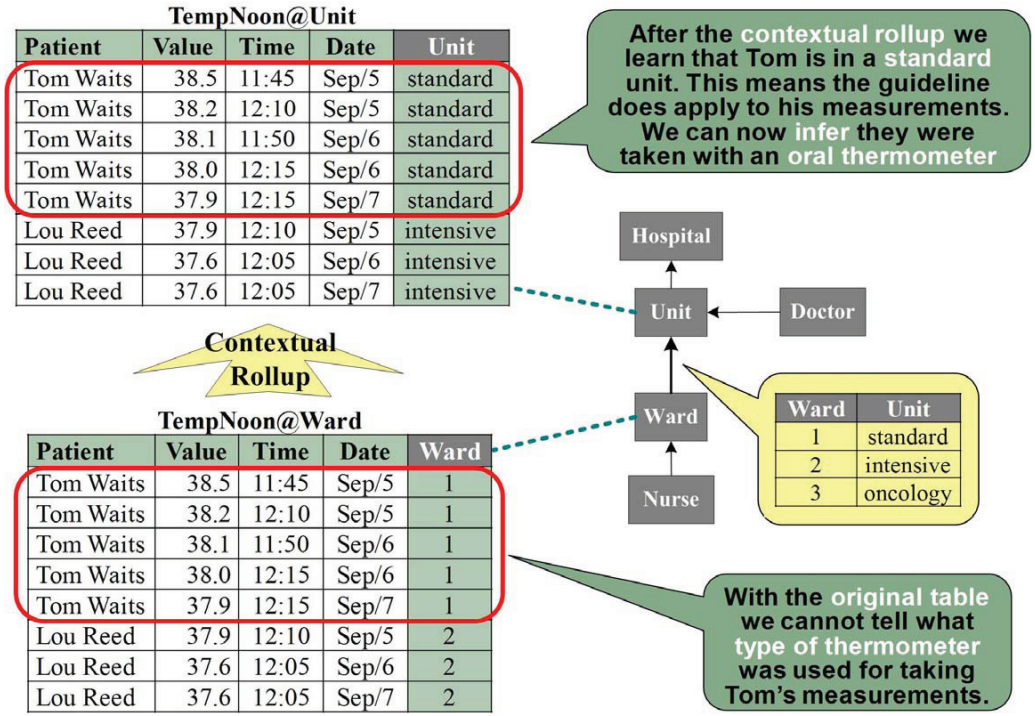
"The temperature of patients in standard care units have to be taken with an oral thermometer"

Captured by means of a rule (hard, or possibly, default rule)

Or a hard constraint

- The information in the context is commonly of a multidimensional nature

We embed (an extension of) the Hurtado-Mendelzon model for MDDBs into our ontological context



A specification of the hierarchical/dimensional hospital structure

Other dimensions could be easily considered, generating multidimensional (MD) contextual information, for additional and finer-granularity data quality assessment

Contextual roll-up can be used to access missing information at certain level, by lattice navigation

Mechanisms for querying database with taxonomies could be applied/embedded (Martinenghi & Torlone; ER10)

Many interesting issues open ...

Look Ahead

The general formalization and computational use of contexts is still an open problem

Many aspects of contexts have to be taken into account and modeled

Ours is a long term general research

Also in terms of applications to data quality assessment and cleaning

We have sketched some first steps in this direction

Next steps have to do with:

- Use of quality predicates (among those in $\mathcal{P}$ on page 44) Possibly of the kind specifically defined for capturing data quality concerns [Borgida, Mylopoulos, Lei; ER'08]
- Related to previous item, specification of *sense* (of data items) by imposing additional semantics
- Techniques for QA query answering

Final Remarks

In (database centered, lower-level) data management, data quality assessment usually deals with problems arising from the acquisition and integration of data: typos, inaccuracy, incompleteness, inconsistency, etc.

At the other end, BI applications require data quality assessment at higher levels of abstraction, where subjectiveness, usefulness, sense, and interpretation play a central role

From a BI perspective, the meaning of the data, in a broad sense, and therefore its quality, are context dependent

In our broad and long term research we are investigating the role and use of contexts in data quality assessment and cleaning

With flexible, adaptive and generic data quality frameworks, solutions and tools in mind