Some problems and results about data incompleteness

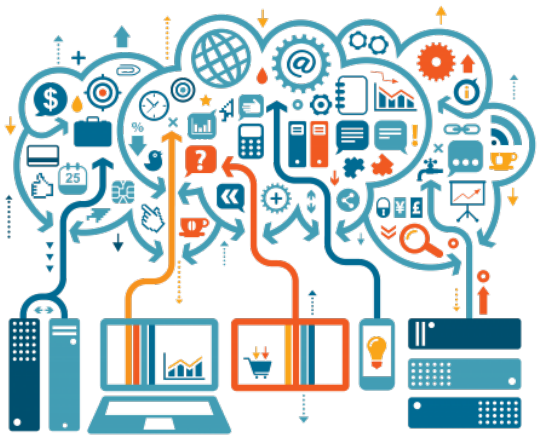
Exposé scientifique de l'équipe MOA

Claire David

joint work with **N. Francis (LIGM)**, W. Czerwiński, P. Hofman, F. Murlak, P. Parys, M. Pilipczuk (Warsaw), L. Libkin (Edinburgh), W. Martens (Bayreuth), A. Gheerbrant, L. Segoufin, C. Sirangelo (Paris), Sh. Amano (IBM Tokyo)

13 février 2019

More and more data, less and less control



Heterogeneous: many different formats

- ▷ Tables - relational, tabular
- ▷ Trees - XML, JSON
- ▷ Graphs - RDF, property graphs

and many ways to organize data within the same format.

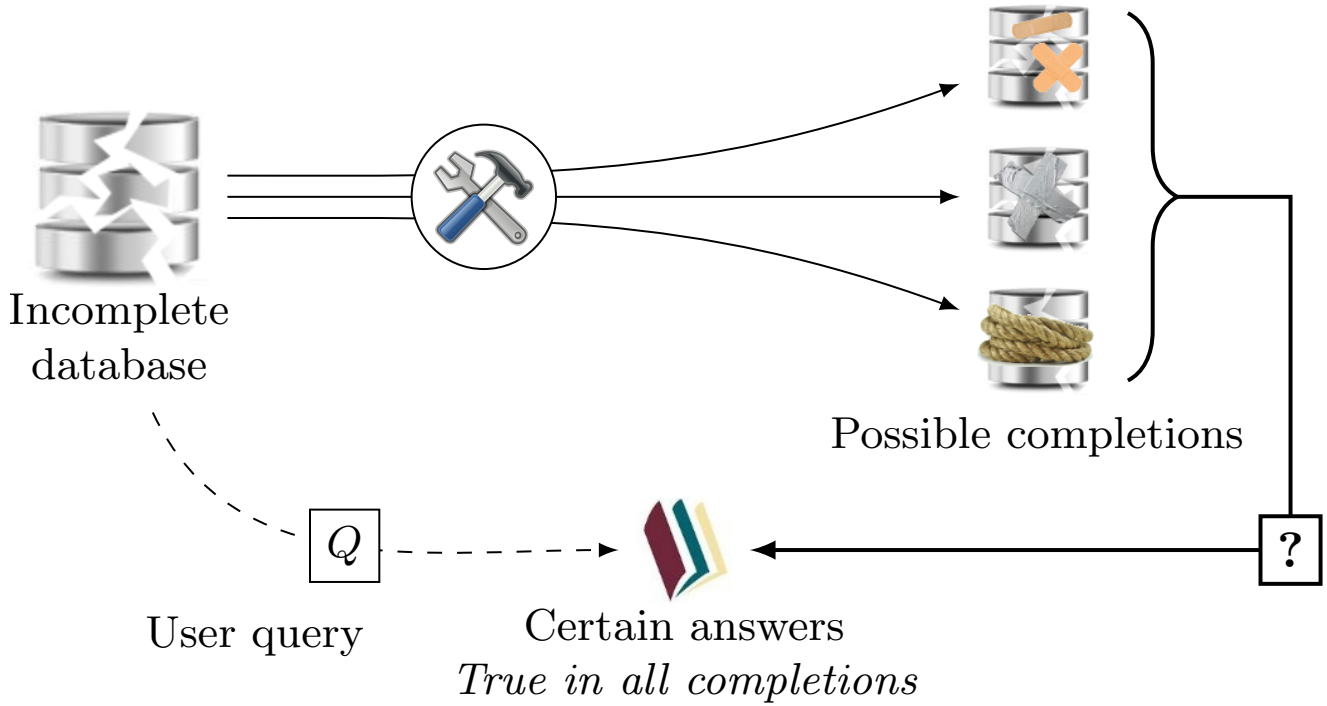
Distributed data: several sources

- ▷ Partial access to the information
- ▷ Inconsistent data
- ▷ Trustworthiness issues

How to analyze, transform, query data in the context of incomplete, inconsistent or uncertain information?

[SIGMOD Record'16]

Scenario - Incomplete data



- Is there a complete instance that conforms to the specification?

  Consistency – [FSTTCS'14,TCS'18]

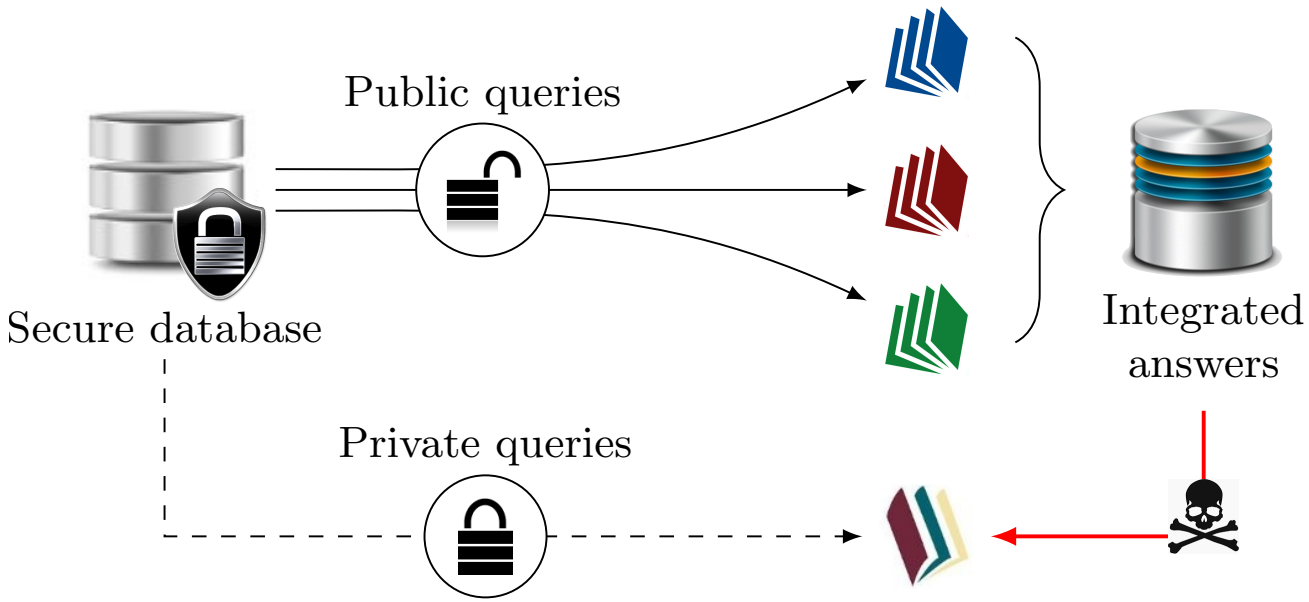
- What is the answer to a query over the incomplete instance?

  Certain answers

- What is a good completion of an incomplete instance?

  Canonical solution

Scenario - Data leak



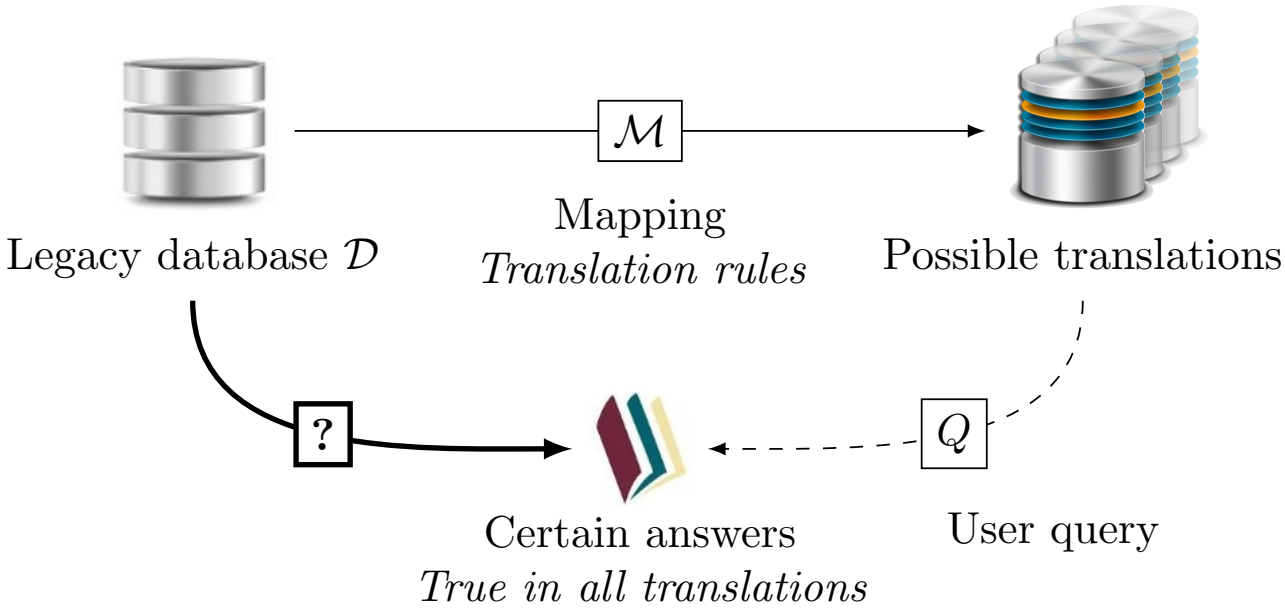
- Can we rewrite a private query using only the public queries?

  Query rewriting – [LMCS'15,ToCS'17]

- Can we ensure that public data do not disclose enough information for a potential attacker to efficiently compute private data?

  Determinacy – [LMCS'15,ToCS'17]

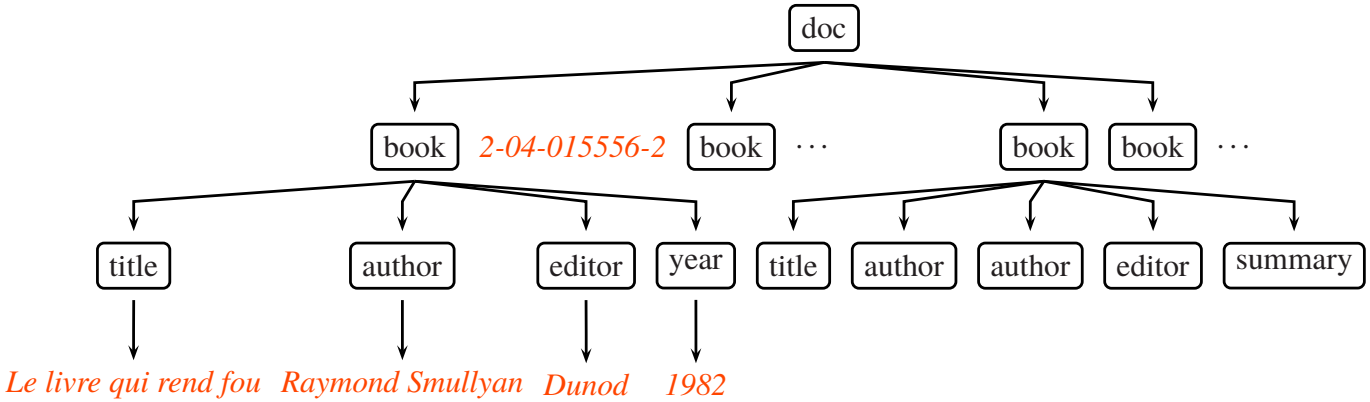
Scenario - Data exchange



- ▷ Does there exist at least one translation? Consistency – [JACM'14]
- ▷ Can we synthesize an effective implementation of the mapping? Synthesis – [ICDT'14]
- ▷ How to define and compute the answer to a query Q from $\mathcal{M}$ and $\mathcal{D}$ without computing the entire set of translations? Certain answers – [JACM'14][PODS'17]

Two contributions - Consistency for tree-structured data

- ▷ Tree-structured data - XML, JSON... **Data trees**



- ▷ Schema - DTD, XML schema **Regular tree language**
- ▷ Constraints **FO formula over the structure and the data**

Consistency: Is there a data tree that conforms to a specification?

- Specification on structure only [FSTTCS'14]
- Specification on structure and data [TCS'18]

Consistency problem - tree pattern with injective semantics

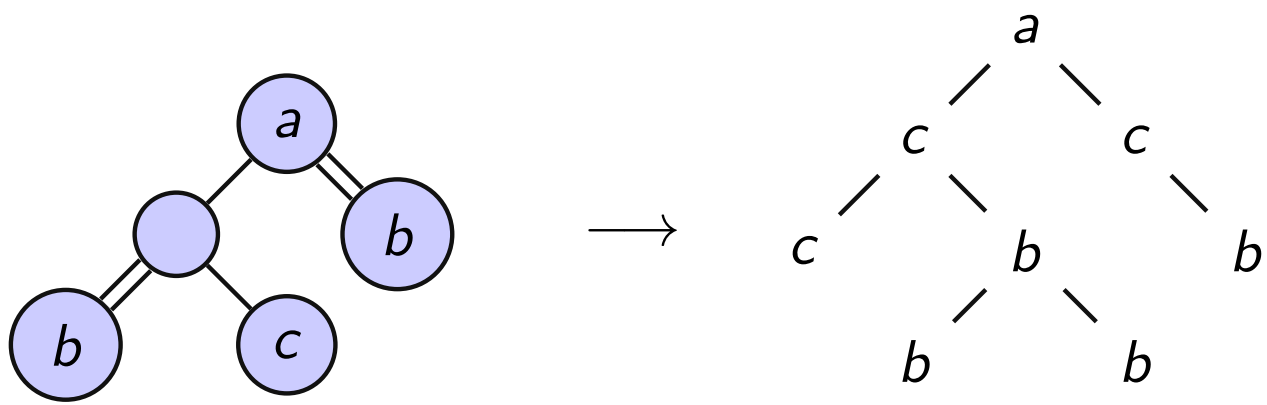
Joint work with N. Francis, F. Murlak

Given an incomplete document, can we find a completion that conforms to some formatting rules?

Formal abstraction

Incomplete XML document $\rightsquigarrow$ tree pattern π
Document schema $\rightsquigarrow$ regular tree language $\mathcal{L}$

Consistency: Can π be matched in some tree from $\mathcal{L}$?



Consistency problem - tree pattern with injective semantics

Joint work with N. Francis, F. Murlak

Given an incomplete document, can we find a completion that conforms to some formatting rules?

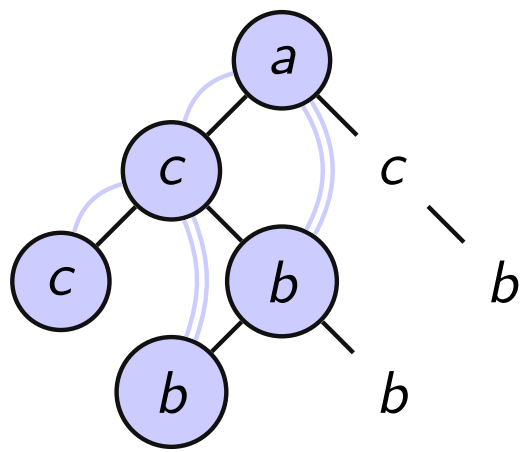
Formal abstraction

Incomplete XML document $\rightsquigarrow$ tree pattern π
Document schema $\rightsquigarrow$ regular tree language $\mathcal{L}$

Consistency: Can π be matched in some tree from $\mathcal{L}$?

Pattern π is matched in tree $\mathcal{T}$ if there is an injective homomorphism $h\colon \pi \to T$.

(injective semantics $\rightsquigarrow$ XML DOM model)

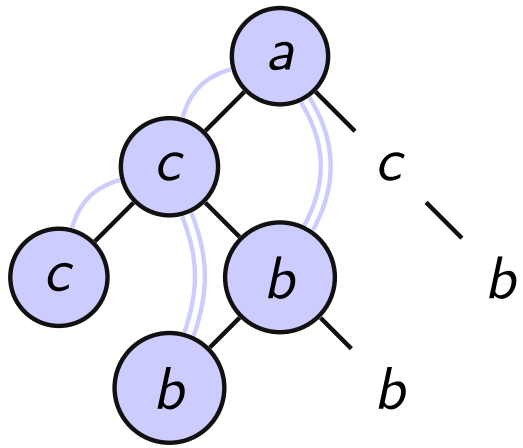


Consistency problem - tree pattern with injective semantics

Joint work with N. Francis, F. Murlak

Consistency: Can π be matched in some tree from $\mathcal{L}$?

Pattern π is matched in tree $\mathcal{T}$ if there is an injective homomorphism $h\colon \pi \to T$.



Theorem [FSTTCS'14]

Consistency of tree patterns under injective semantics is

- ▷ NP-complete
- ▷ PTIME with descendant only (without child).

Consistency problem - integrity constraints

Joint work with W. Czerwiński, F. Murlak and P. Parys

Ingredients:

- ▷ Tree-structured data: data trees (XML, JSON...)
- ▷ Schema: regular tree langage (DTD, XML schema...)
- ▷ Integrity constraints:

$$\forall \bar{x}\ q(\bar{x}) \to \alpha(\bar{x})$$

selects tuples of nodes — $UCQ(\downarrow, \downarrow^+, \to, \to^+, lab)$

checks data (in)equalities — $UCQ(\sim, \not\sim)$

Consistency of constraints:

Given a schema $\mathcal{S}$ and a set of constraints $\mathcal{C}$, is there a data tree satisfying $\mathcal{S}$ and $\mathcal{C}$?

Consistency problem - integrity constraints

Joint work with W. Czerwiński, F. Murlak and P. Parys

Each constraint in $\mathcal{C}$ is of the form:

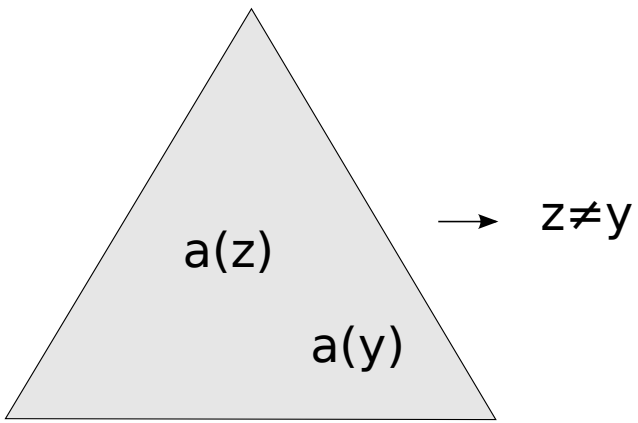
$$\forall \bar{x}\ q(\bar{x}) \rightarrow \alpha(\bar{x})$$

selects tuples of nodes
$UCQ(\downarrow, \downarrow^+, \rightarrow, \rightarrow^+, lab)$

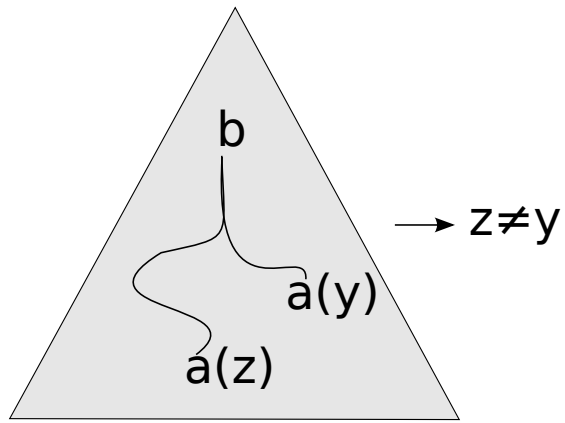
checks data (in)equalities
$UCQ(\sim, \not\sim)$

Examples of constraints:

- **keys**:
  all a-nodes use different values



- **relative keys**:
  key constraint in all b-subtrees



Consistency problem - integrity constraints

Joint work with W. Czerwiński, F. Murlak and P. Parys

Non-mixing constraints:

Each constraint in $\mathcal{C}$ is of the one of the following forms:

$$q(\bar{x}) \to \alpha_{\sim}(\bar{x}) \qquad\qquad q(\bar{x}) \to \alpha_{\not\sim}(\bar{x})$$

Theorem

Consistency with respect to schema is:

- ▷ undecidable for mixed $\sim$ and $\not\sim$ [BMS'08]
- ▷ **2EXP-complete for non-mixing constraints** **[TCS'18]**
- ▷ 2EXP-complete for $\sim$ only and $\not\sim$ only [BMS'08]

Perspectives - ANR-blanc QUID 2019-2023

Principal Investigator:

Claire David 

Scientific programme:

- ▷ Determinacy, privacy and data leakage
- ▷ Repairs of inconsistent databases
- ▷ Data incompleteness
- ▷ Connections with Constraint Satisfaction Problems

Collaborators: 

- ▷ Partners: IRIF, LaBRI, ENS-Ulm
- ▷ LIGM collaborators: N. Francis, V. Marsault, J. Thapper

Exciting research collaborations to come!

More details...

- **Consistency and Entailment**
  [TCS'18] Reasoning About Integrity Constraints for Tree-Structured Data.
  W. Czerwiński, C. David, F. Murlak, P. Parys
  [FSTTCS'14] Consistency of injective tree patterns.
  C. David, N. Francis, F. Murlak
  [ICDT'13] Containment of Pattern-Based Queries over Data Trees.
  C. David, A. Gheerbrant, L. Libkin, W. Martens
- **Data exchange**
  [PODS'17] Schema mappings for Data Graphs.
  N. Francis, L. Libkin
  [JACM'14] XML Schema Mappings: Data Exchange and Metadata Management.
  S. Amano, C. David, L Libkin, F. Murlak
  [ICDT'14] Synthesizing transformations from XML schema mappings.
  C. David, P. Hofman, F. Murlak, M. Pilipczuk
- **Determinacy**
  [ToCS'17] Asymptotic Determinacy of Path Queries Using Union-of-Paths Views.
  N. Francis
  [LMCS'15] Datalog Rewritings of Regular Path Queries using Views.
  N. Francis, L. Segoufin, C. Sirangelo