



Uncertain Version Control in Open Collaborative Editing of Tree-Structured Documents

M. Lamine BA, Talel Abdessalem & Pierre Senellart

<http://dbweb.enst.fr/>

13th ACM Symposium on DocEng – Sept 10-13, Florence (Italy)





Version Control of Uncertain Data

- Data in large-scale, open and collaborative editing platforms, such as Wikipedia, are inherently uncertain
 - contributors with different reliability, conflicts, malicious edits, ...
 - need version control to maintain the quality of document versions
- Existing version control approaches are all deterministic
 - no room to uncertainty handling in the versioning process
- A version control model aware of uncertain data may be helpful

13th ACM Symposium on DocEng – Sept 10-13, Florence (Italy)

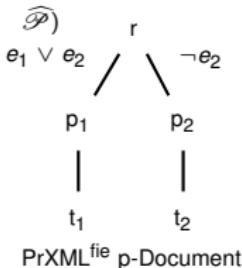




Uncertain Tree-Structured Data

Probabilistic XML [Kimelfeld & Senellart.(2013)]

- Unordered, unranked, and labeled **XML trees** with annotated edges
 - annotations are **propositional formulas** of random Boolean variables



$d_1)$	r	$d_2)$	r	$Pr(e_1) = 0.2$
	/ \			$Pr(e_2) = 0.8$
	p_1	p_1		$Pr(d_1) = Pr(e_1) \times Pr(\neg e_2)$
			t_1	$Pr(d_2) = (Pr(\neg e_1) \times Pr(e_2)) + (Pr(e_1) \times Pr(e_2))$
	t_1	t_2	$\mathcal{F}_{21} = \{e_2\}$	
				$\mathcal{F}_{11} = \{e_1\}$ $\mathcal{F}_{22} = \{e_1, e_2\}$

Possible worlds and their probabilities

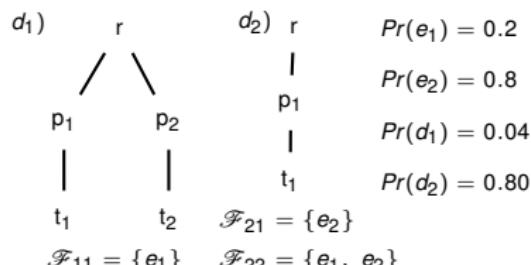
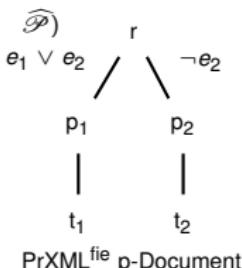
13th ACM Symposium on DocEng – Sept 10-13, Florence (Italy)



Uncertain Tree-Structured Data

Probabilistic XML [Kimelfeld & Senellart.(2013)]

- Unordered, unranked, and labeled **XML trees** with annotated edges
 - annotations are **propositional formulas** of random Boolean variables



Possible worlds and their probabilities

13th ACM Symposium on DocEng – Sept 10-13, Florence (Italy)

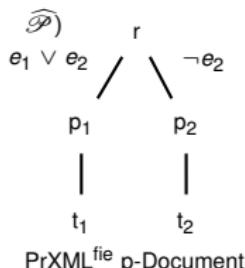




Uncertain Tree-Structured Data

Probabilistic XML [Kimelfeld & Senellart.(2013)]

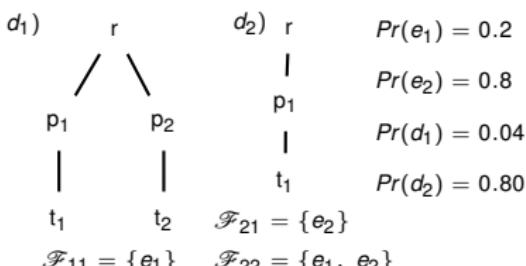
- Unordered, unranked, and labeled **XML trees** with annotated edges
 - annotations are **propositional formulas** of random Boolean variables



- Enumerating all possible worlds and their probabilities

- Enable also to model uncertain updates on (uncertain) nodes [Kharlamov et al.(2010)]

☞ Integrate such a representation in a typical version control process



Possible worlds and their probabilities

13th ACM Symposium on DocEng – Sept 10-13, Florence (Italy)



Uncertain Multi-Version XML Document

Uncertain Version Control Model

Semantics of Updates

Evaluation of the model

Performance Analysis

Filtering capabilities





Uncertain Multi-Version XML Document

Uncertain Version Control Model

Defines two equivalent views over any uncertain multi-version XML tree

- set \mathcal{V} of random variables $e_0, e_1 \dots e_n$ modeling the tree states
- infinite set \mathcal{D} of all (unordered) XML trees including the versions

 (\mathcal{G}, Ω) : Logical View

 $(\mathcal{G}, \widehat{\mathcal{P}})$: Probabilistic XML Encoding

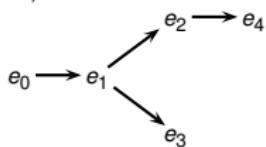
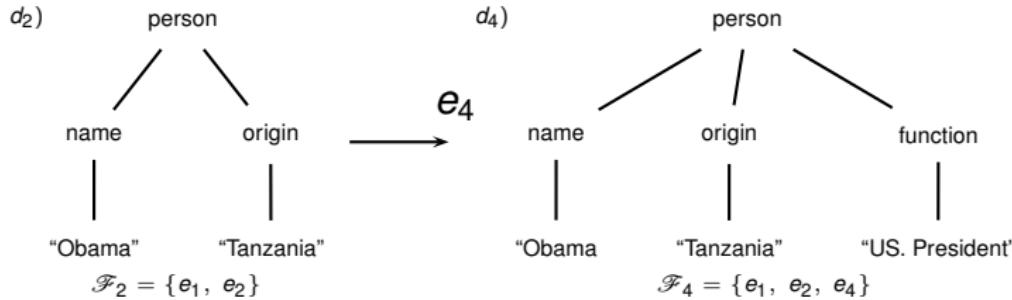
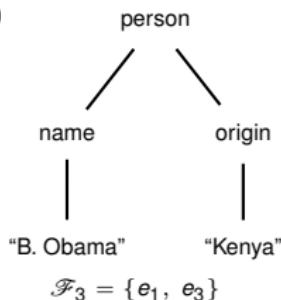
- DAG \mathcal{G} built on variables in \mathcal{V}
- Similar DAG \mathcal{G} of random variables in \mathcal{V}
- Mapping $\Omega : 2^{\mathcal{V} \setminus \{e_0\}} \rightarrow \mathcal{D}$ which computes the possible versions according to sets of valid events
- Probabilistic XML tree $\widehat{\mathcal{P}}$ which defines the same probability distribution as Ω mapping





Uncertain Multi-Version XML Document

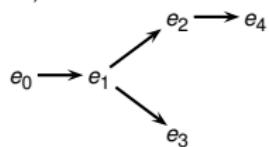
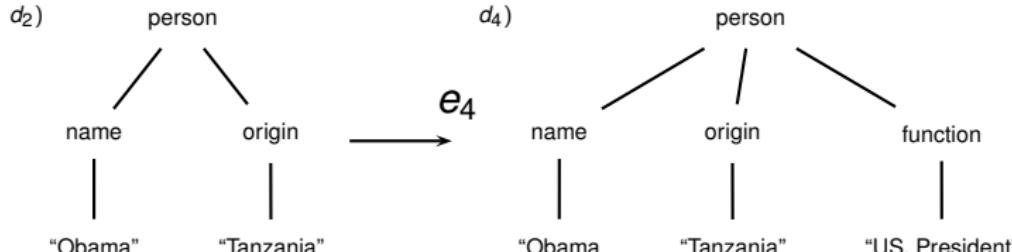
Uncertain Version Control Model (Example)

 $\mathcal{G})$  $d_2)$  e_2 $d_1)$ person $d_3)$ 



Uncertain Multi-Version XML Document

Uncertain Version Control Model (Example)

 \mathcal{G}  $d_2)$  e_2

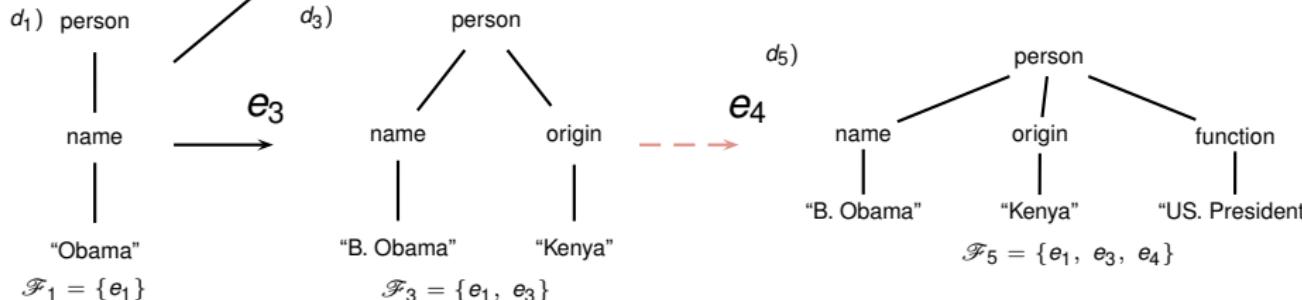
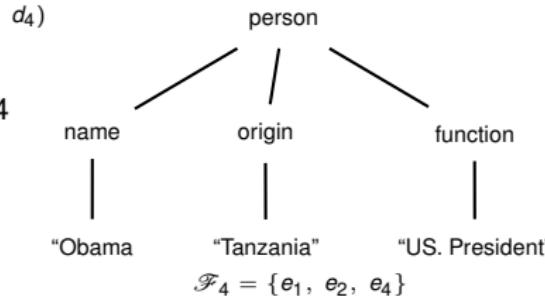
"Obama"
 $\mathcal{F}_2 = \{e_1, e_2\}$

 $d_1)$

person

 $d_3)$

person

 $\mathcal{F}_1 = \{e_1\}$ $\mathcal{F}_3 = \{e_1, e_3\}$ $d_4)$ 

"Tanzania"
 $\mathcal{F}_4 = \{e_1, e_2, e_4\}$

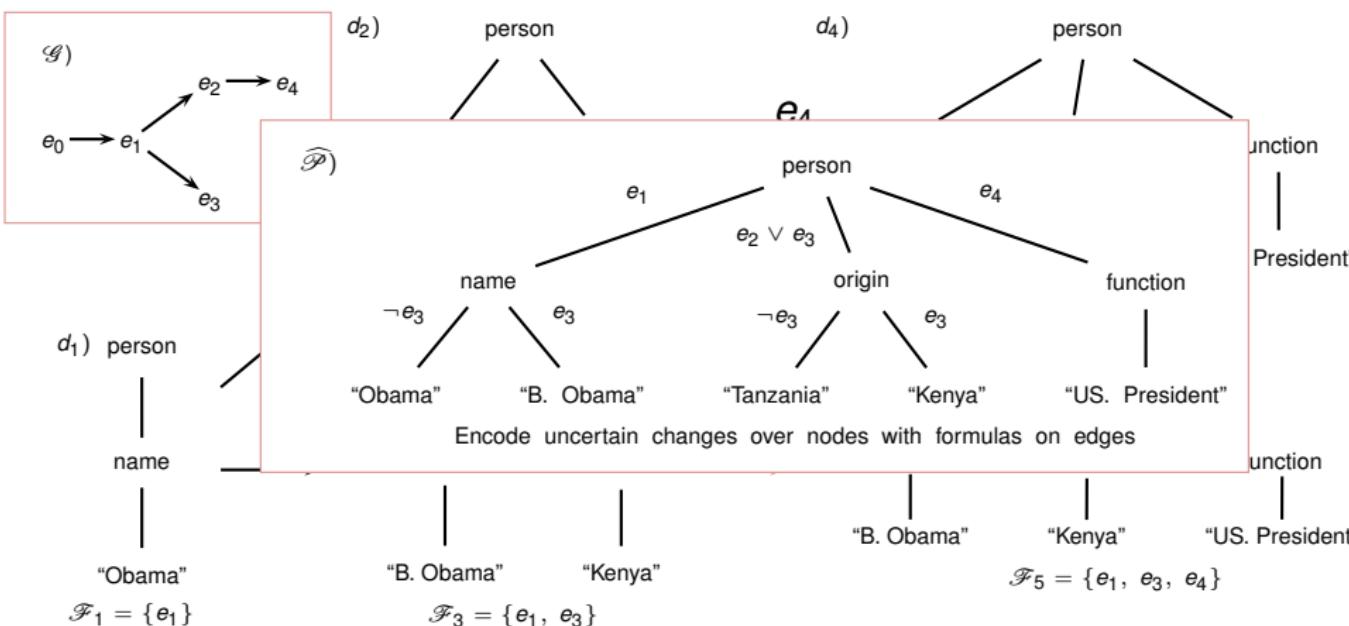
 $d_5)$ 

"US. President"
 $\mathcal{F}_5 = \{e_1, e_3, e_4\}$



Uncertain Multi-Version XML Document

Uncertain Version Control Model (Example)





Uncertain Multi-Version XML Document

Semantics of Updates

- Assume an event e_i in \mathcal{G} pointing to the edited version
- Come with a new event e_j *not in \mathcal{G}* and an edit script Δ

Logical definition

Input: (\mathcal{G}, Ω) , e_i , e_j , Δ

- $\mathcal{G} := \mathcal{G} \cup (\{e_j\}, \{(e_i, e_j)\})$
- Extension of Ω to a Ω' mapping

\iff

For each event set $\mathcal{F} \in 2^{(\mathcal{V} \setminus \{e_0\}) \cup \{e_j\}}$:

- $\Omega'(\mathcal{F}) = [\Omega(\mathcal{F} \setminus \{e_j\})]^\Delta$ if $e_j \in \mathcal{F}$
- $\Omega'(\mathcal{F}) = [\Omega(\mathcal{F})]$ if $e_j \notin \mathcal{F}$

Probabilistic XML Update

Input: $(\mathcal{G}, \widehat{\mathcal{P}})$, e_i , e_j , Δ

- $\mathcal{G} := \mathcal{G} \cup (\{e_j\}, \{(e_i, e_j)\})$
- Updating $\widehat{\mathcal{P}}$ with operations in Δ

For an insert of x and a delete of y :

- $fie(x) := fie(x) \vee (e_j)$ if $x \in \widehat{\mathcal{P}}$ or
insert x in $\widehat{\mathcal{P}}$ with $fie(x) := (e_j)$
- $fie(y) := fie(y) \wedge \neg(e_j)$



Uncertain Multi-Version XML Document

Uncertain Version Control Model

Semantics of Updates

Evaluation of the model

Performance Analysis

Filtering capabilities





Evaluation of the model

Performance Analysis

- ▶ Estimation of two main metrics: commit time and checkout cost

Baseline Systems

- ☛ Versioning tools SubVersion and Git
 - Use of their Java implementations based on the APIs SvnKit and JGit

Real Datasets

- History of commits over two large file systems (shared tree-structured data)
 - ☛ Linux kernel development
 - ☛ Cassandra project

- Set up our system (PrXML) in Java language
- Measures are obtained with all accesses in RAM Disk





Evaluation of the model

Performance Analysis (Results)

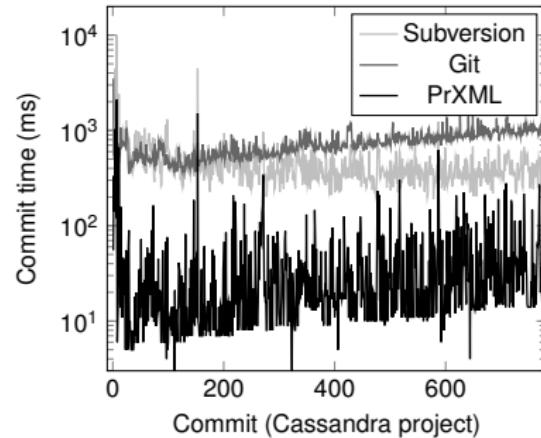
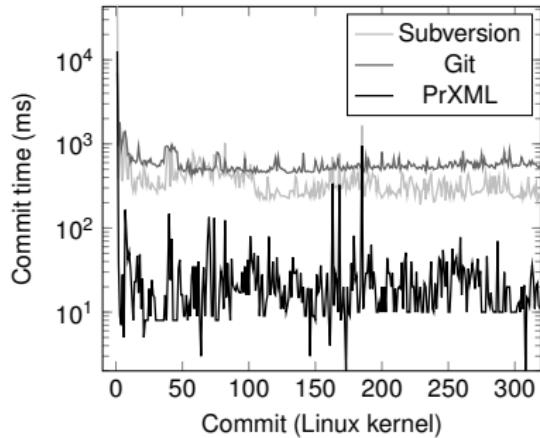


Figure : commit time over real-world datasets (logarithmic y-axis)





Evaluation of the model

Performance Analysis (Results)

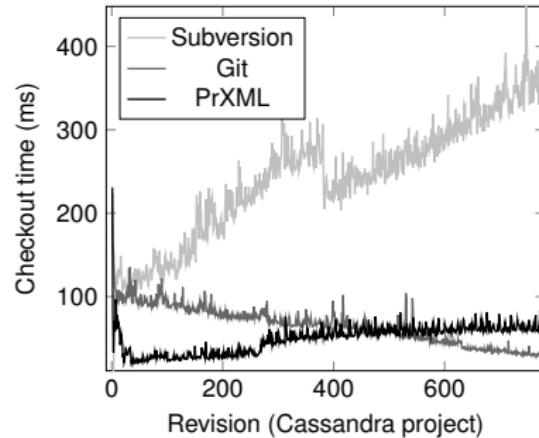
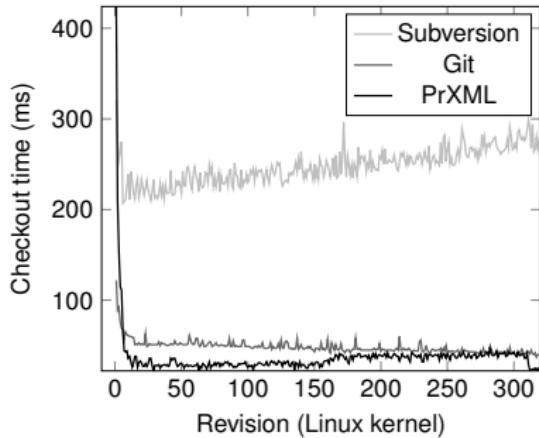


Figure : checkout time over real-world datasets (linear axes)





Evaluation of the model

Filtering capabilities

- ▶ Tests are run over a sample of articles from the Wikipedia dump
 - Automatic filtering of unreliable content, e.g. spams, in versions of articles
 - Generate arbitrary versions that fit user preference
 - ▶ versions from trustworthy authors
 - Test more advanced operations over critical articles such as vandalized pages
 - ▶ e.g. study the impact of considering as reliable some versions affected by vandalism in the history of the edition of a given article
 - Detection of vandalism as well as Wikipedia robots do, automatically manage it while keeping all uncertain versions available for checkout.
 - etc.





Evaluation of the model

Filtering capabilities (Demo [Ba et al.(2011)])

Filtering options on revisions

Extracting the revisions of a given page

Selecting list of revisions to merge

Revision ID	Author of revision	Date of edit
Codic:Villani1	117.211.28.173	2010-08-07
Codic:Villani2	115.39.223.207	2010-08-07
Codic:Villani3	115.39.223.207	2010-08-07
Codic:Villani4	Myself.	2010-08-06
Codic:Villani5	Administrator	2010-08-06
Codic:Villani6	Administrator	2010-08-06
Codic:Villani7	LockeBot	2010-11-27
Codic:Villani8	Administrator	2010-12-10
Codic:Villani9	115.39.223.207	2010-12-10
Codic:Villani10	Llocl	2010-12-29
Codic:Villani11	Llocl	2010-12-29
Codic:Villani12	Managerpage	2011-01-06

CODIC: VILLANI: revisions are currently listed.

(a) keyword-based search engine

The screenshot shows the Mediawiki Data Merging Tool interface. At the top, there's a navigation bar with 'File', 'Options', and a search bar. Below it is a sidebar with 'Recent changes', 'Uncategorized articles', and 'Searching options'. The main area has tabs for 'Sample articles' and 'Online articles'. A large central window is titled 'Merge of revisions' with a sub-section 'Process to merging...'. It contains a list of revisions from 'Edit 1' to 'Edit 8', each with a 'Merge' button. The revisions are as follows:

Revision	User	Page	Date
Edit 1	Villainy	Lao	2010-12-29
Edit 2	Villainy	Lao	2010-12-30
Edit 3	Villainy	Lao	2010-12-30
Edit 4	Villainy	Mongope	2011-01-01
Edit 5	Villainy	Lao	2011-01-01
Edit 6	Villainy	Lao	2011-01-01
Edit 7	Villainy	Lao	2011-01-01
Edit 8	Villainy	Mongope	2011-01-01

Below the merge dialog, a message says 'Building one's own Wikipedia article'. At the bottom right, there's a section titled 'Selecting list of revisions to merge'.

(b) generation of arbitrary versions

CÉDRIC VILLANI

Cédric Villani (French) received the Fields Medal in 2010 as a French mathematician working primarily on partial differential equations and mathematical physics. He was awarded the Fields Medal at the International Congress of Mathematicians held in Hyderabad, India, in August 2010.

Cédric Villani (French) received the Fields Medal in 2010 as a French mathematician working primarily on partial differential equations and mathematical physics. He was awarded the Fields Medal at the International Congress of Mathematicians held in Hyderabad, India, in August 2010.

Contents

- [Title](#)
- [Biographical outline](#)
- [Research interests](#)
- [Prizes](#)
- [Selected publications](#)

Work

Villani worked on the theory of [partial differential equations](#), located in statistical mechanics, and on the theory of [kinetic theory](#). In 2003 he was one of the three to receive the first fast consequences award for joint work with Cédric Villani on the Boltzmann equation. In 2006 he was awarded the [Prix de l'Institut universitaire de France](#) for his work on nonlinear Landau damping. In 2008 he was awarded the [Fields medal](#) at the International Congress of Mathematicians held in Hyderabad, India, in August 2010. In 2010 he was awarded the [Crafoord Prize](#) for his work on the Boltzmann equation, for general measure-valued spaces. In December 2010 he was awarded the [Cantabrian Prize for Scientific Research](#).

View more and Query system

View Tree

Cédric Villani

- [Title](#)
- [Biographical outline](#)
- [Research interests](#)
- [Prizes](#)
- [Selected publications](#)

Spanish query

External query

(c) visualization features

13th ACM Symposium on DocEng – Sept 10-13, Florence (Italy)





Thank for your attention !





References

-  Talel Abdessalem, M. Lamine Ba, and Pierre Senellart, *A probabilistic XML merging tool*, EDBT, 2011, Demonstration.
-  M. Lamine Ba, Talel Abdessalem, and Pierre Senellart, *Towards a version control model with uncertain data*, PIKM, 2011.
-  Evgeny Kharlamov, Werner Nutt, and Pierre Senellart, *Updating Probabilistic XML*, Updates in XML, 2010.
-  Benny Kimelfeld and Pierre Senellart, *Probabilistic XML: Models and complexity*, Advances in Probabilistic Databases for Uncertain Information Management (Zongmin Ma and Li Yan, eds.), Springer-Verlag, 2013.

