



# PoliMi - UCLA

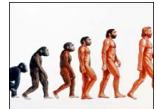
## SCHEMA EVOLUTION IN WIKIPEDIA: *toward a Web Information System Benchmark*

Carlo A. Curino  
(carlo.curino@polimi.it)

Hyun J. Moon  
Letizia Tanca  
Carlo Zaniolo



# Motivations



- **Information Systems (IS)** are subject of **continuous evolution**
  - requirements change to adapt to an evolving environment
  - waterfall development methodologies are **inadequate** (especially on the WEB)



- **Evolution in IS** is an extremely difficult and **expensive** task
  - software evolution and maintenance represents **90%** of the costs
  - **Legacy Applications** cannot be modified, but must be supported



- The **data management core**:
  - is one of the **most difficult** portion of a system to evolve
  - data evolves: need for **historical archiving** due to accountability
  - schema evolves: with **dramatic impact** on queries and applications



# Wikipedia Analyses



- **Why Wikipedia:** DB-centric Web Information System (WIS), very popular, open-source license of the software platform (MediaWiki) and data, rich schema history



- **Tools:** We developed a tool-suite to analyze WIS DB backends



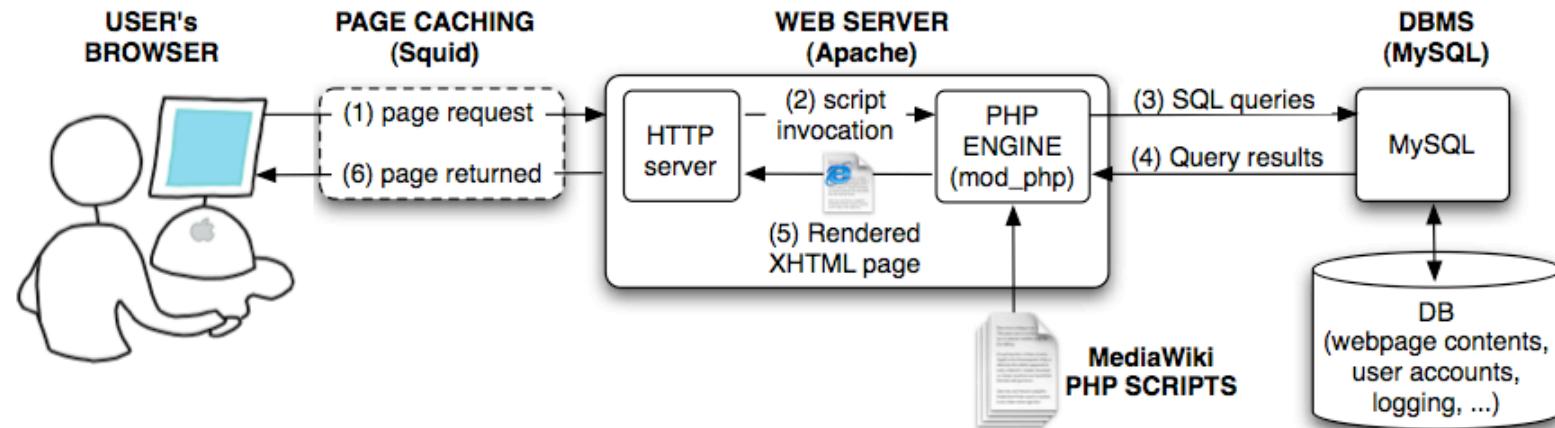
- **The analysis:** We collect and dissect MediaWiki schema history: 170+ schema versions in 4.5 years



- **The benchmark:** We publicly released tools, results and datasets as a first step: “*towards a Benchmark for Schema Evolution*”



# MediaWiki Architecture



- **Scalability issues:**

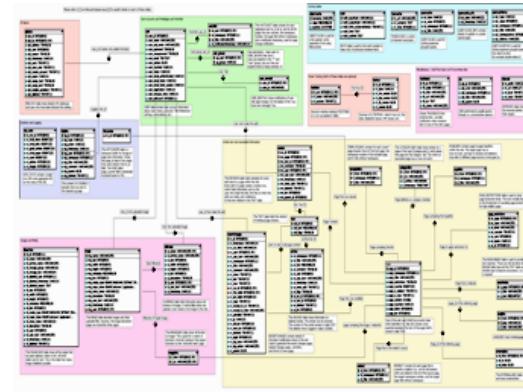
- DB size exceeds **700 Gb** (excluding multimedia content)
- Wikipedia receives an average of **29k requests/sec** (peak 85k) each producing several DB queries
- Several Layers of **caching**
- **DBMS performance is the bottleneck:** poor partitioning?



# The MediaWiki Schema



- Tables can be grouped in:
  - article and content (6 tables)
  - links and structure (9 tables)
  - users and permissions (5 tables)
  - performance and caching (7 tables)
  - statistics and special features (3 tables)
  - history and archival (4 tables)
- **NOTE: History management represents about 30% of the schema (4 tables + several attributes in the other tables)**

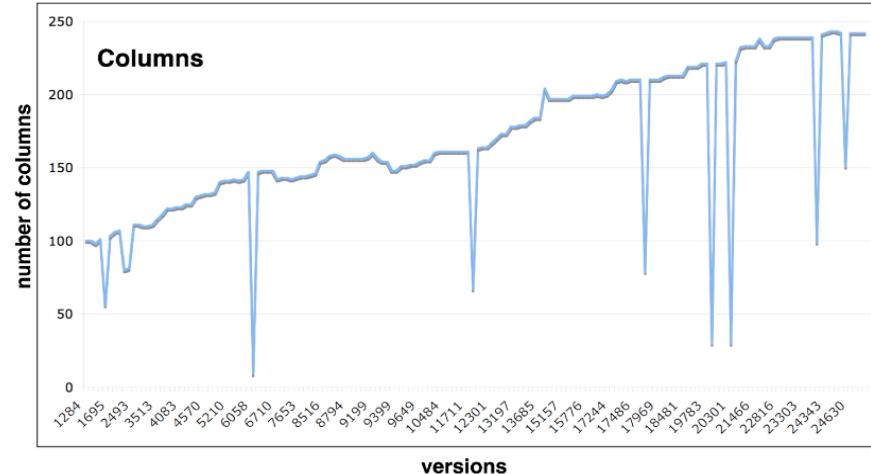




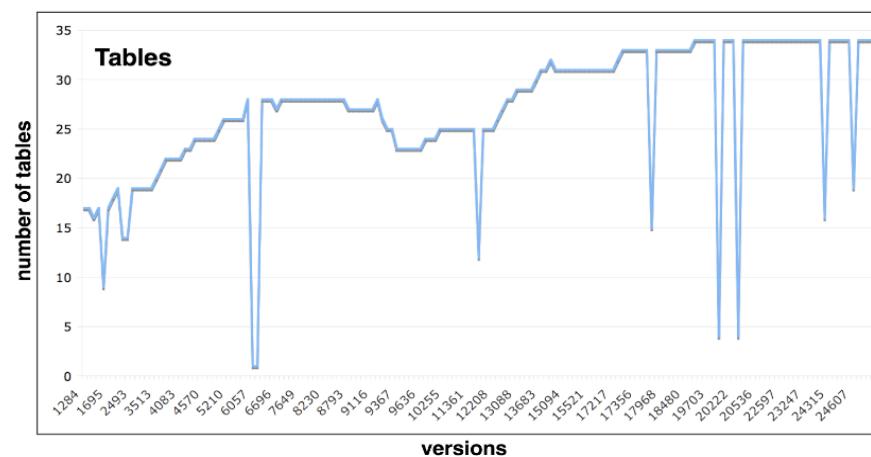
# Basic Statistics: schema growth

- Schema Evolution:

- 170+ versions in 4.5 years
- almost 250% increase



- WIS evolve faster than Traditional IS
  - 38% w.r.t. [Sjoberg93]
  - 539% w.r.t. [Marche93]

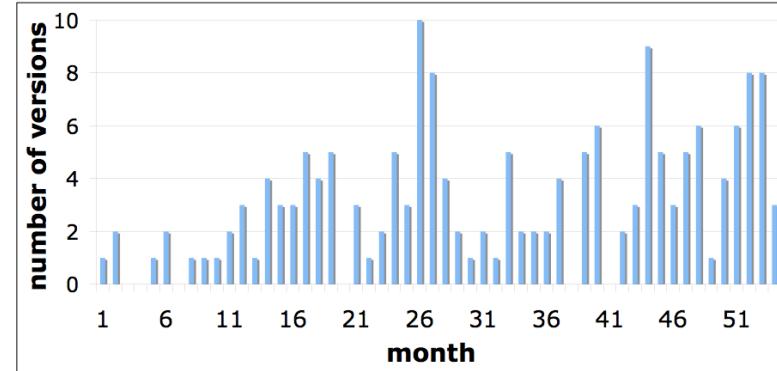


- Note: Collaborative WISs embrace information sharing (better data to study!)

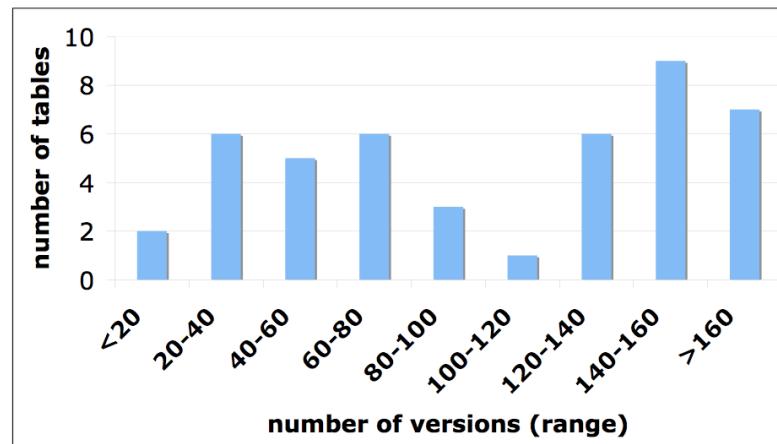


# Basic Statistics: evolution and lifetime

- More frequent schema changes far away from releases



- Schema Elements Lifetime:
  - a group of stable relations
  - young tables and columns





# Type of Changes

Type of Change	# of evolution steps	% of evolution steps
<i>Actual Schema</i>	94	54.9%
<i>Index/Key</i>	69	40.3%
<i>Data Type</i>	22	12.8%
<i>Syntax Fix</i>	20	11.7%
<i>Rollback</i>	15	8.8%
<i>Doc Only</i>	13	7.6%
<i>Engine</i>	6	3.5%

- **NOTE:** sum exceeds 100%, since several changes might coexist in a single evolution step
- Total lack of integrity constraints (except for primary keys)!!



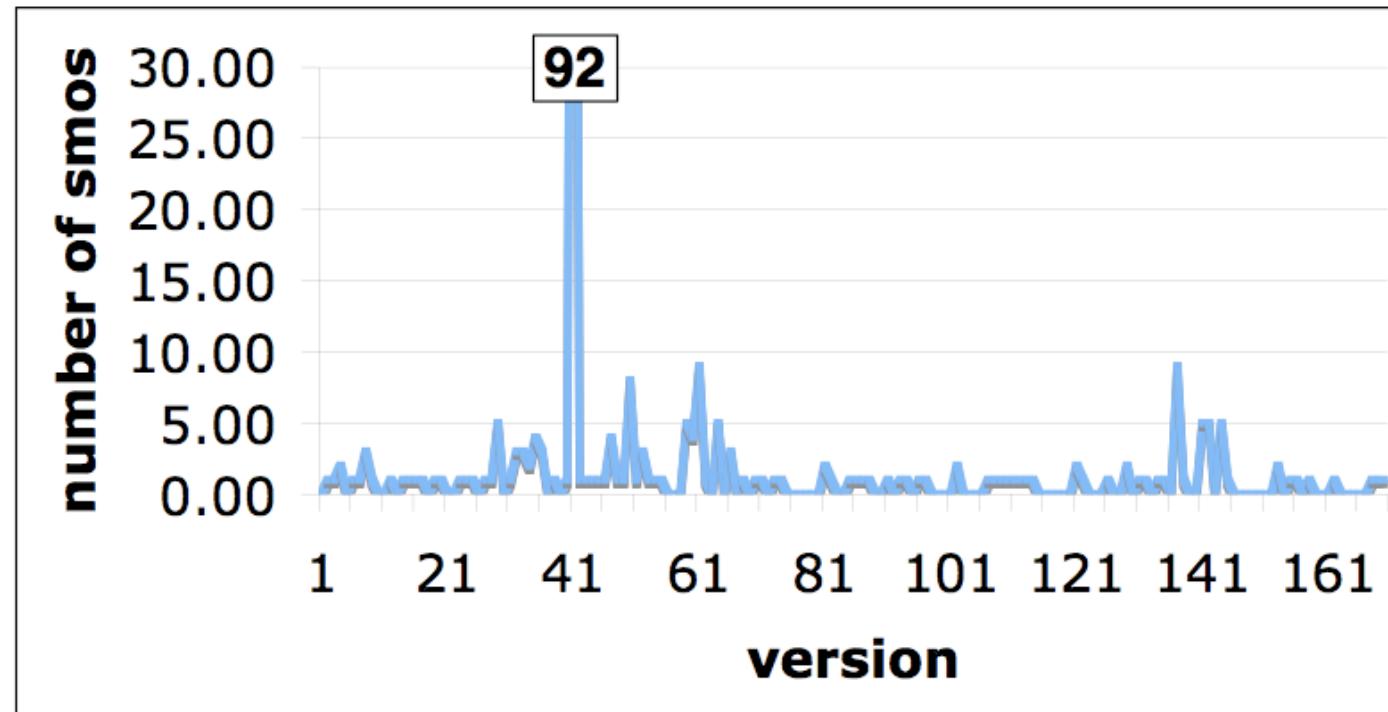
# Changes: by Schema Modification Operators

SMO type	# of usages	% of usage	% per version
CREATE TABLE	24	8.9%	14%
DROP TABLE	9	3.3%	5.2%
RENAME TABLE	3	1.1%	1.75%
DISTRIBUTE TABLE	0	0.0%	0%
MERGE TABLE	4	1.5%	2.33%
COPY TABLE	6	2.2%	3.5%
ADD COLUMN	104	38.7%	60.4%
DROP COLUMN	71	26.4%	41.5 %
RENAME COLUMN	43	16.0%	25.1 %
MOVE COLUMN	1	0.4%	0.58%
COPY COLUMN	4	1.5%	2.33%
<b>Total</b>	<b>269</b>	<b>100%</b>	<b>-</b>

- NOTE: These operators have been used in PRISM [vdlb2008a] and PRIMA [vdlb2008b] systems



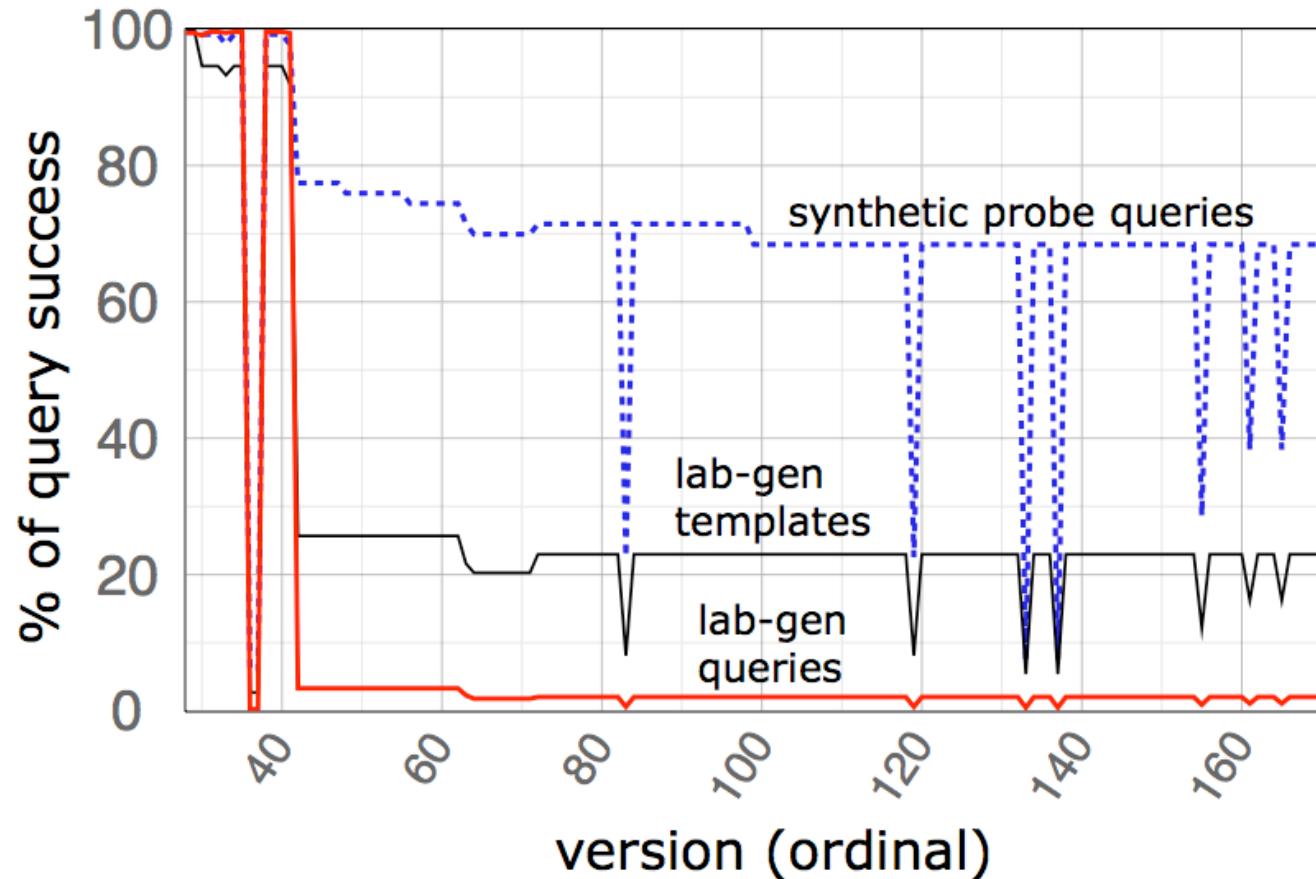
# Schema Changes per Version



- NOTE: version 41-42 represents a MAJOR evolution step where article versioning management is heavily modified!!



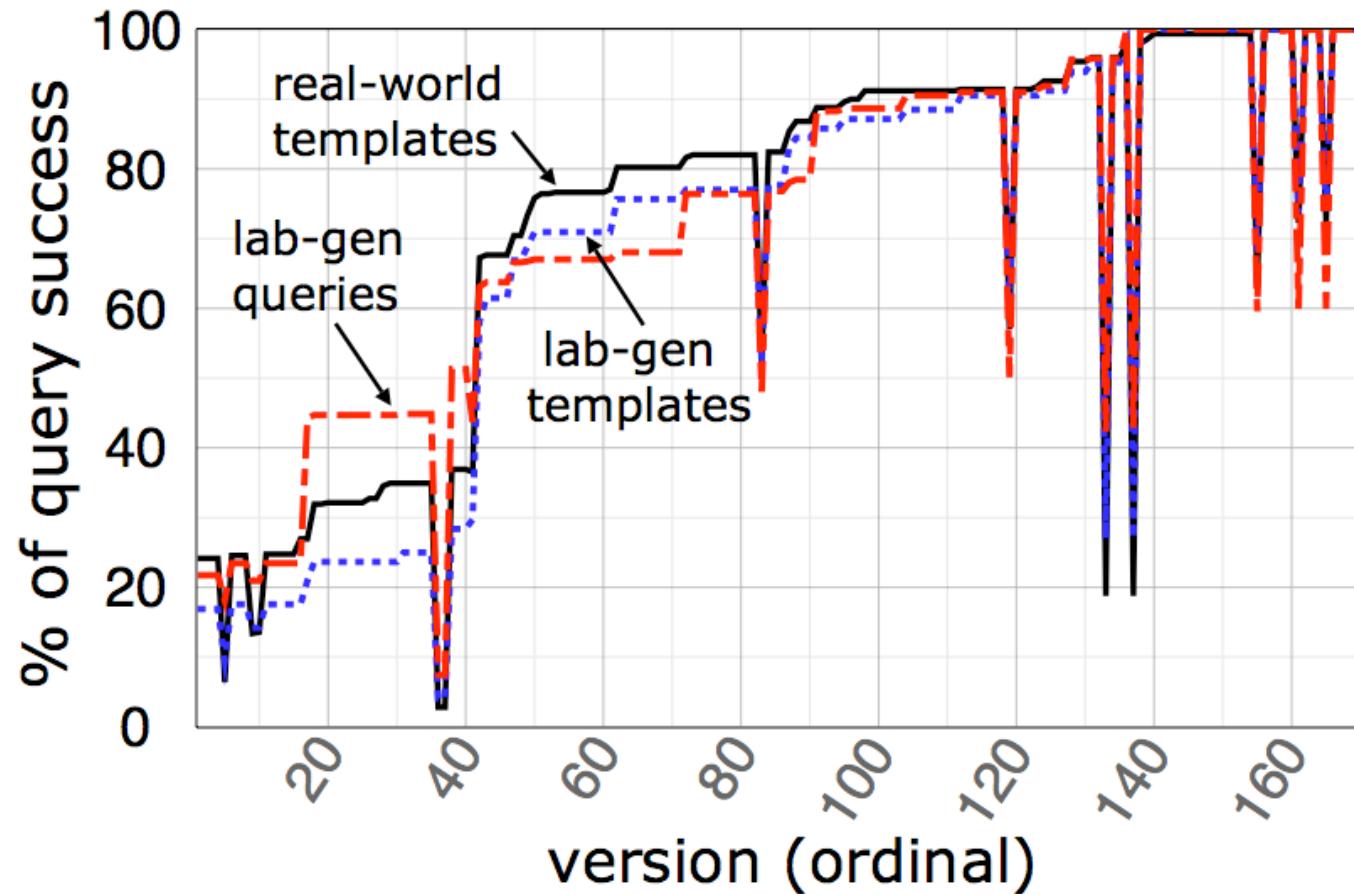
# Impact on the Applications



- NOTE: over ~4000 queries (version 28) from which we extract 75 templates



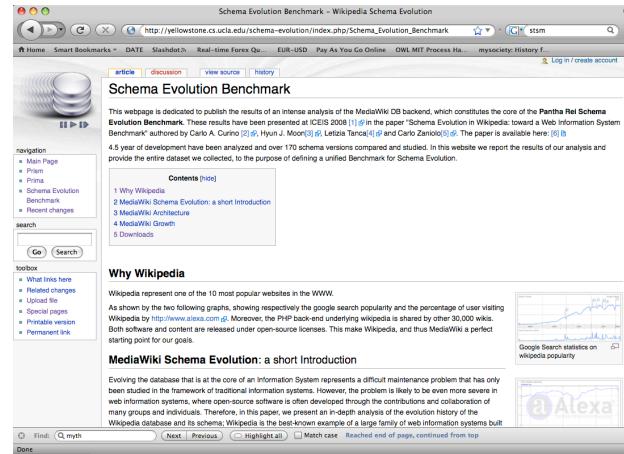
# Real Wikipedia Queries



- NOTE: 500 most common templates out of 2k extracted from over 780 millions query instances from the On-Line Wikipedia Profiler  
<http://noc.wikimedia.org/cgi-bin/report.py>



# Conclusions (1/2)



- *Strong evidence:* need for better Schema Evolution and Data Archiving
- Schema history, analysis data (raw and aggregated), queries and tool-suite are available at:

<http://yellowstone.cs.ucla.edu/schema-evolution/index.php>



# Conclusions (2/2)



**Goal: create a benchmark for schema evolution (and in general a standard relational DB dataset).**

- Extend the analysis to several other Open-Source WIS (*Joomla!*, *TikiWiki*, *Slashcode*, *Zen-Cart*, *Wordpress*)
- Extend the analysis towards Public Scientific DB (*Genome*, *HGVS*)
- This work is part of the bigger project "*Panta Rhei*" tackling:
  - Schema Evolution: *PRISM* [vldb2008a]
  - Transaction Time DB under schema evolution: *PRIMA* [VLDB-2008b]
  - History Metadata Management [STSM-2008], [ECDM-2008]
  - for more information visit:  
<http://yellowstone.cs.ucla.edu/schema-evolution/>



# Bibliography

- **[VLDB-2008a]** "*Graceful database schema evolution: the prism workbench*" Carlo A. Curino, Hyun J. Moon, and Carlo Zaniolo. VLDB, 2008
- **[VLDB-2008b]** "*Managing and querying transaction-time databases under schema evolution*" Hyun J. Moon, Carlo A. Curino, Alin Deutsch, C.-Y. Hou, and Carlo Zaniolo. VLDB, 2008.
- **[ECDM-2008]** "*Managing the History of Metadata in support for DB Archiving and Schema Evolution*", Carlo A. Curino, Hyun J. Moon, Carlo Zaniolo, ER International Workshop on Evolution and Change in Data Management (ECDM) 2008
- **[STSM-2008]** "*Information Systems Integration and Evolution: Ontologies at Rescue*" Carlo A. Curino, Letizia Tanca, Carlo Zaniolo, STSM, 2008