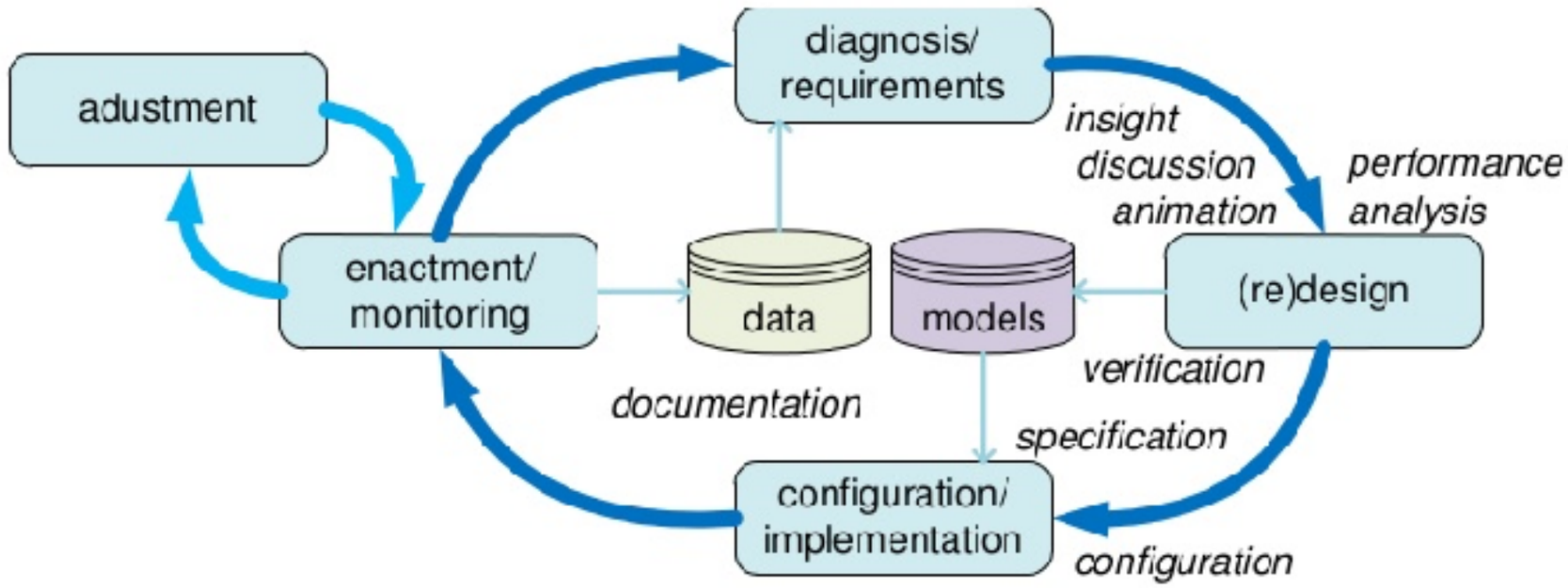
Data-aware Processes: Modeling, Mining, and Verification
Part 2: Mining

Diego Calvanese
(with material from Marco Montali and Will van der Aalst)
Research Centre for Knowledge and Data
Free University of Bozen-Bolzano

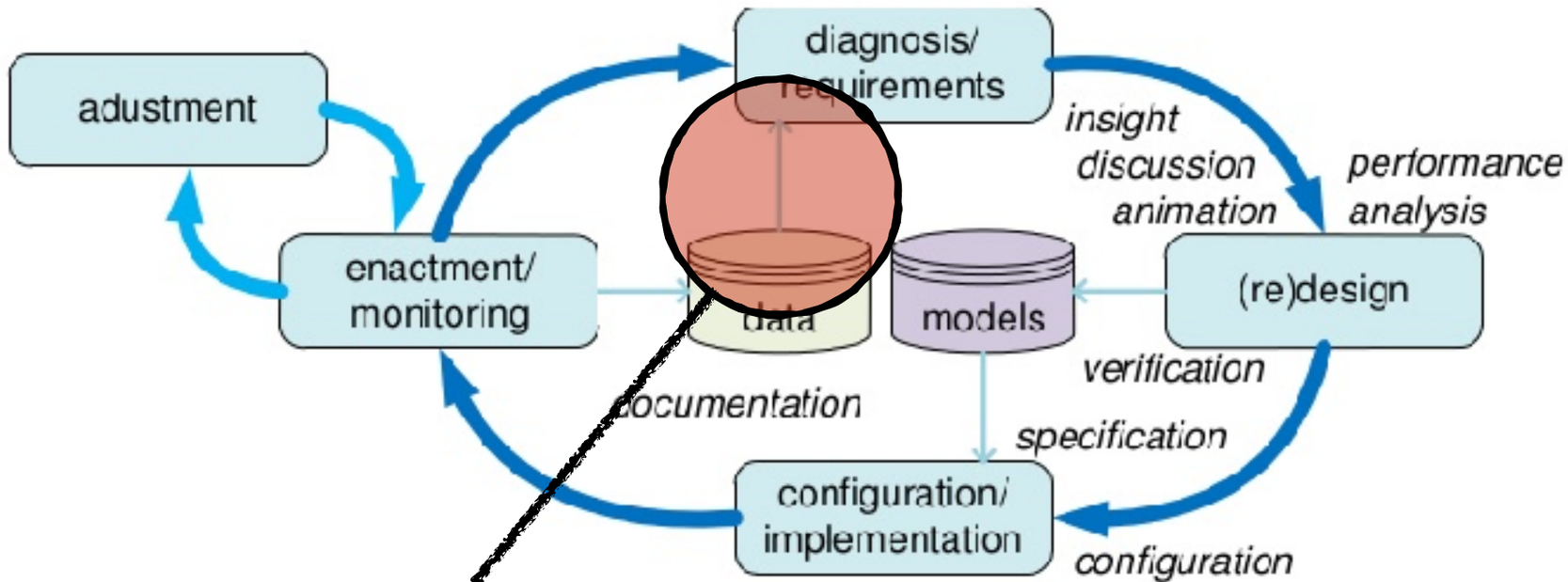
3rd International Winter School on Big Data (BigDat 2017)
13–17/2/2017 – Bari, Italy

Complex Systems Lifecycle



picture by Wil van der Aalst

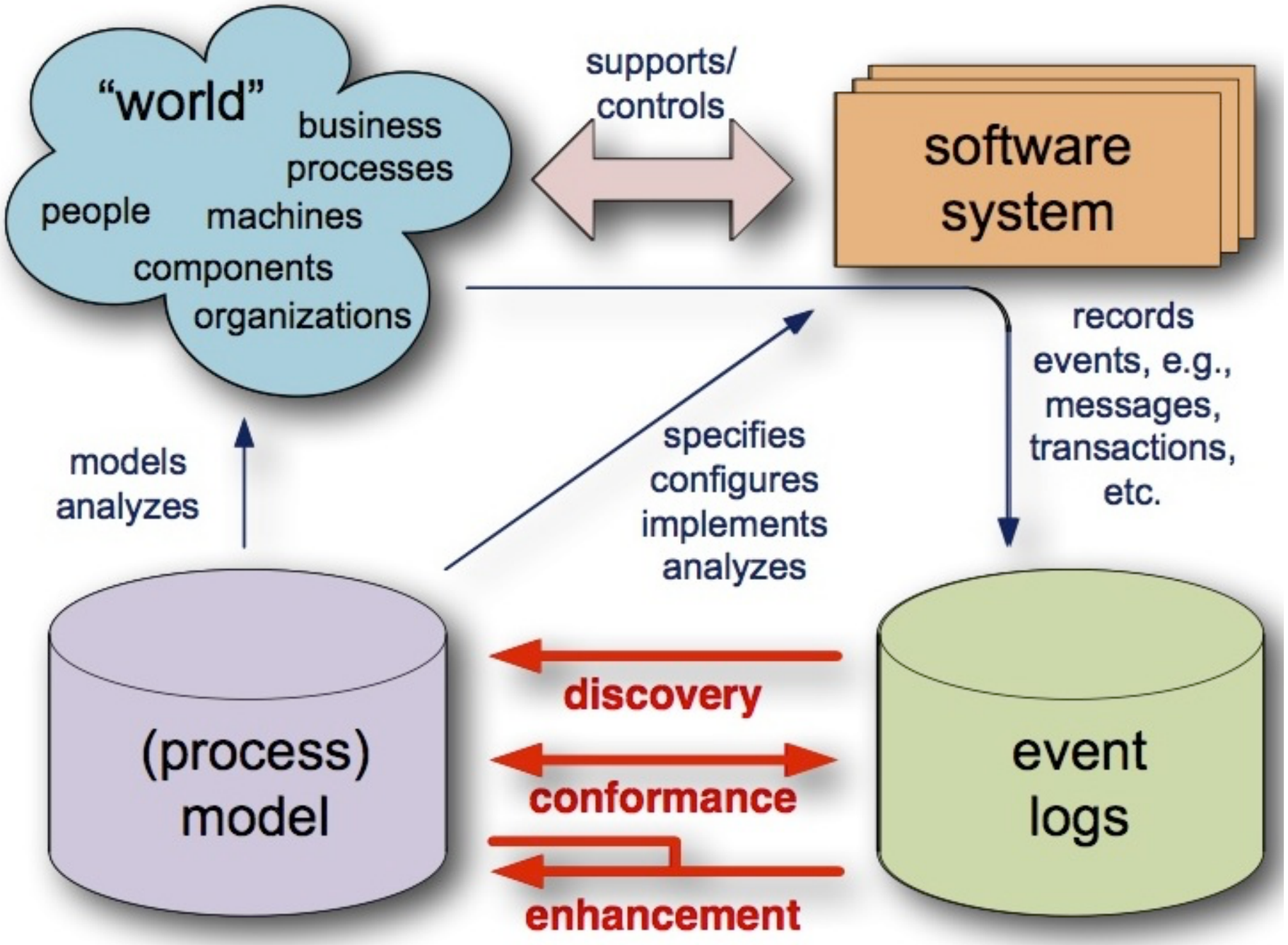
Process Mining for Diagnosis



picture by Wil van der Aalst

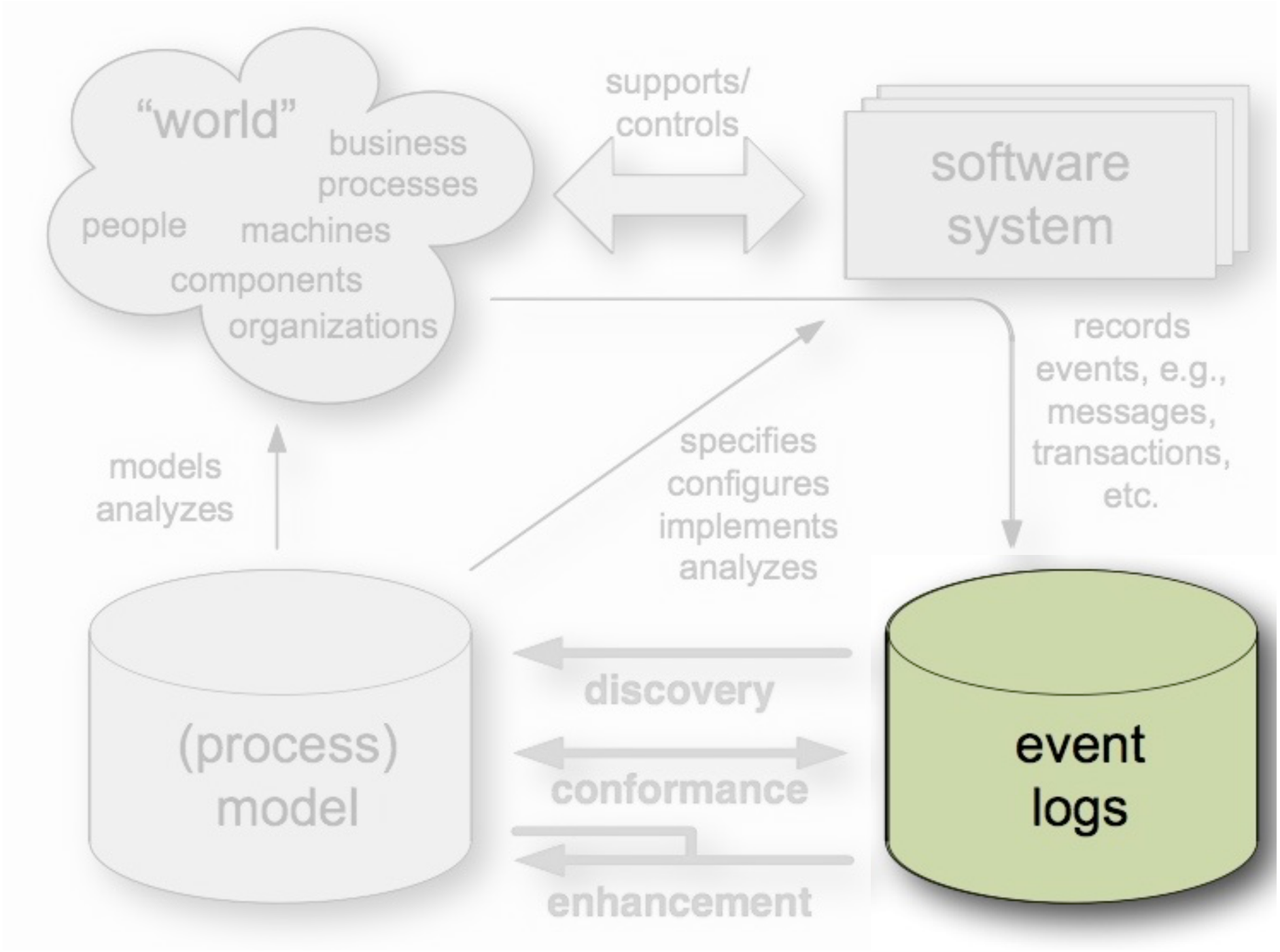
Data preparation for process mining

Process Mining



picture by Wil van der Aalst

Process Mining



picture by Wil van der Aalst

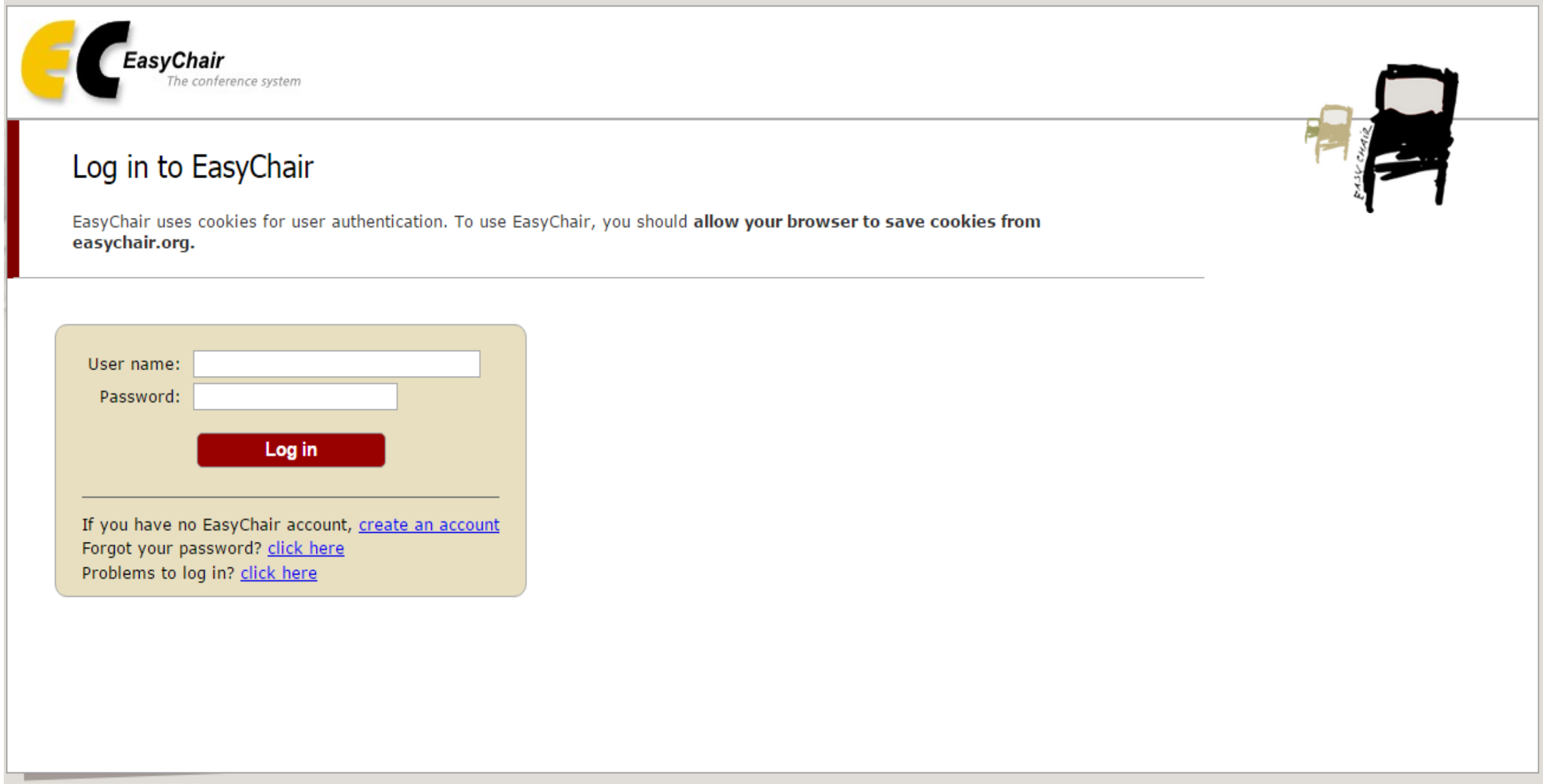
Process Mining: Getting Data

See slides by Will van der Aalst accompanying the book "*Process Mining: Discovery, Conformance and Enhancement of Business Processes*" by Springer:

http://www.processmining.org/book/start

- Chapter 1: Introduction
- Chapter 4: Getting the Data

Actual Reality



Actual Reality

LOGIN

ID	User
1	Alifah Syamsiyah
2	Marco Montali
3	Diego Calvanese
4	Wil van der Aalst

CONFERENCE

ID	Name	Organizer	Time
666	BPM 2015	2	2015-02-14 01:00:00
667	Caise 2015	4	2015-03-06 01:00:00
668	ER 2015	4	2015-03-26 01:00:00
669	EDOC 2015	2	2015-04-05 03:00:00

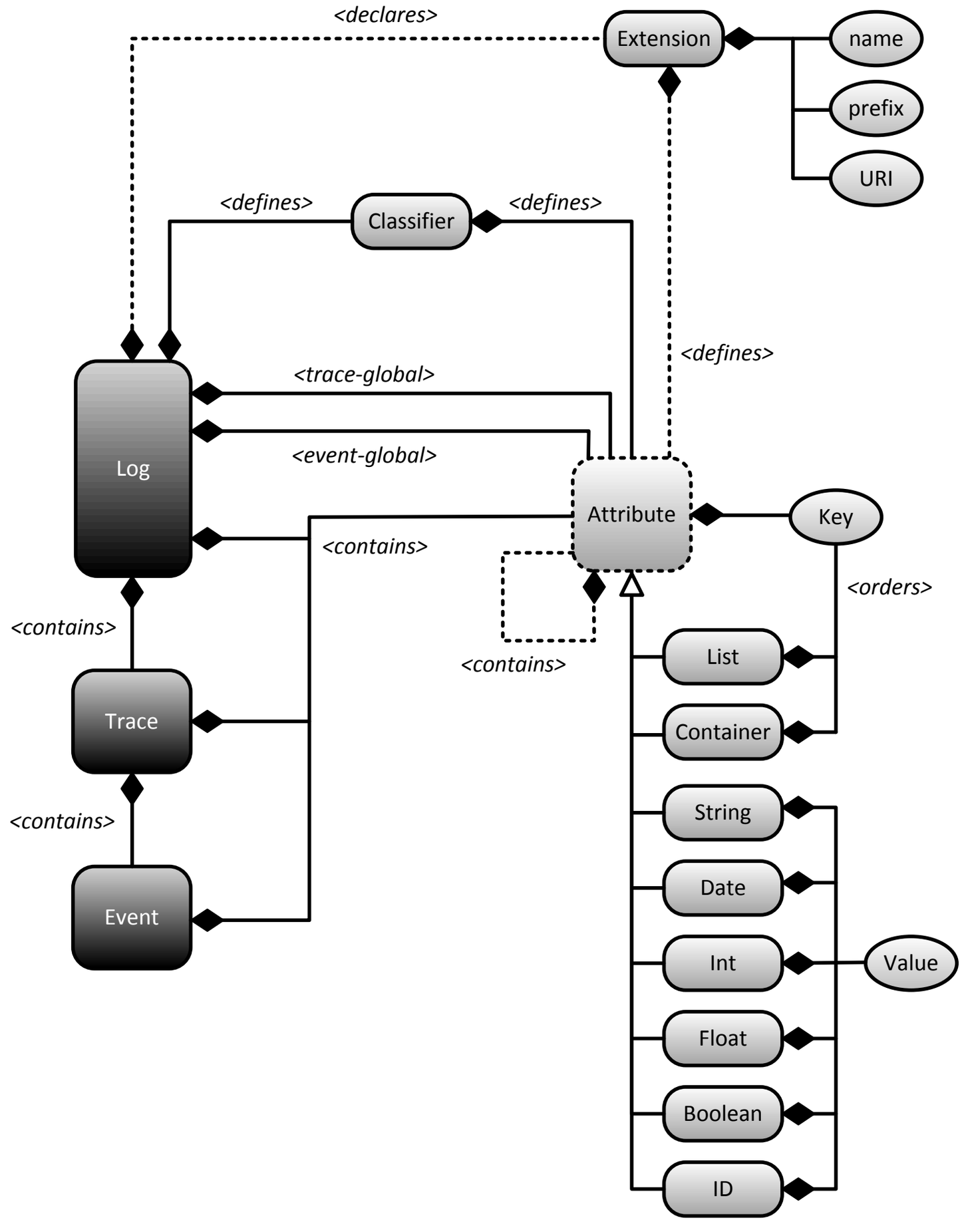
PAPERINFO

ID	Title	CT	User	Conf	Type	Status
1	Ontop at Work	2015-03-02 15:09:35	1	669	FP	RX
2	A Survey of Web Services	2015-03-02 12:36:01	3	668	SP	RX
3	The Definitive Guide for BPM	2015-03-04 13:36:20	1	666	FP	AB

Expected Reality

IEEE XES standard for event logs

- Based on XML
- Minimalistic
- Data+metadata



Expected Reality

log

Case ID	Activity	Timestamp	Resource
1	Submit paper	01-07-2015:10.05	Anna
	Invite reviewer	03-07-2015:12.00	Budi
	Get review	10-07-2015:16.10	Rudy
	Acceptance	12-07-2015:15.00	Rudy
	Submit final paper	19-07-2015:19.15	Anna
2	Submit paper	02-07-2015:17.19	John
	Invite reviewer	03-07-2015:12.00	Tiara
	Get review	11-07-2015:15.45	Clara
	Rejection	12-07-2015:15.00	Clara

trace

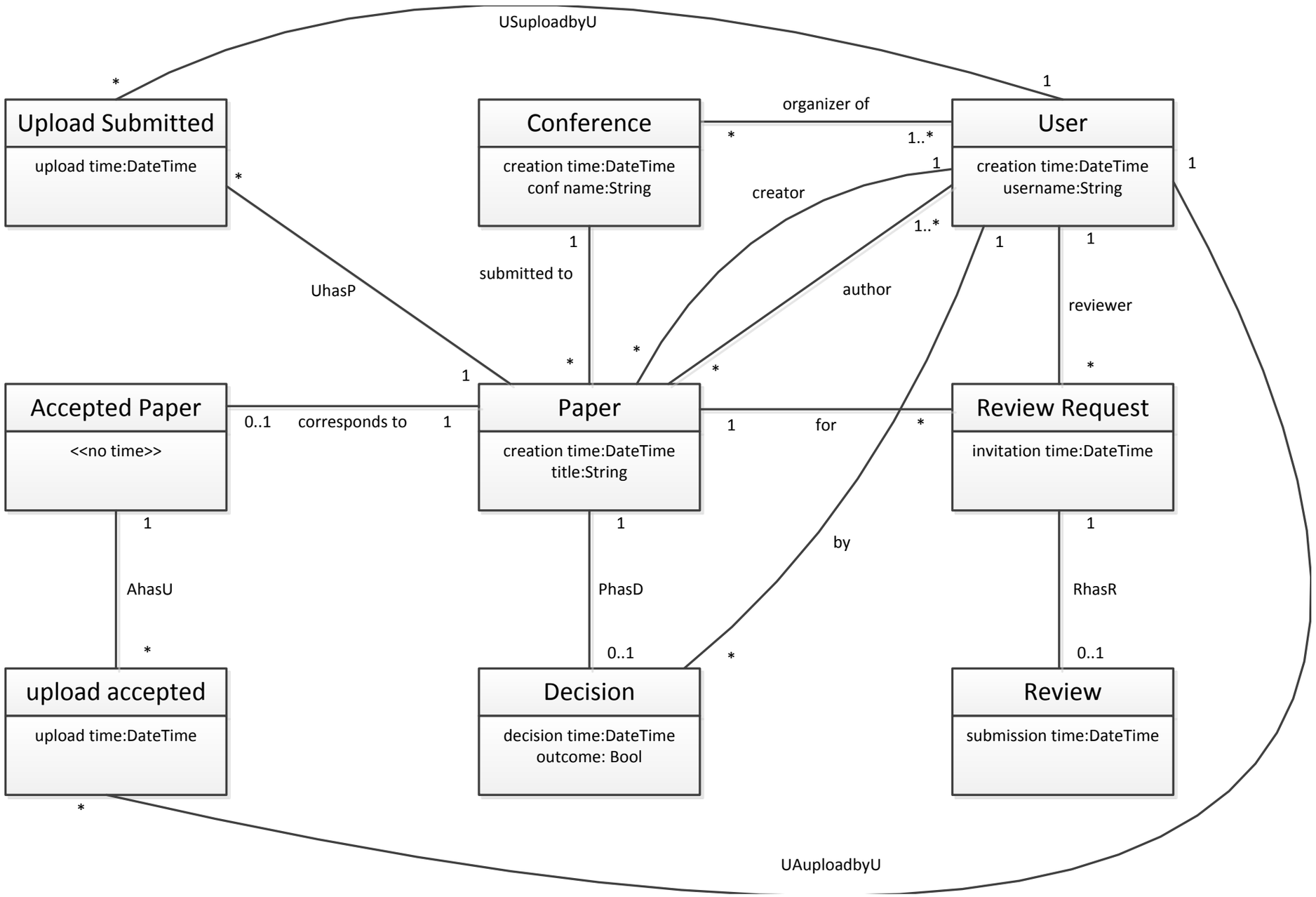
event

Expected Reality

XES standard for event logs

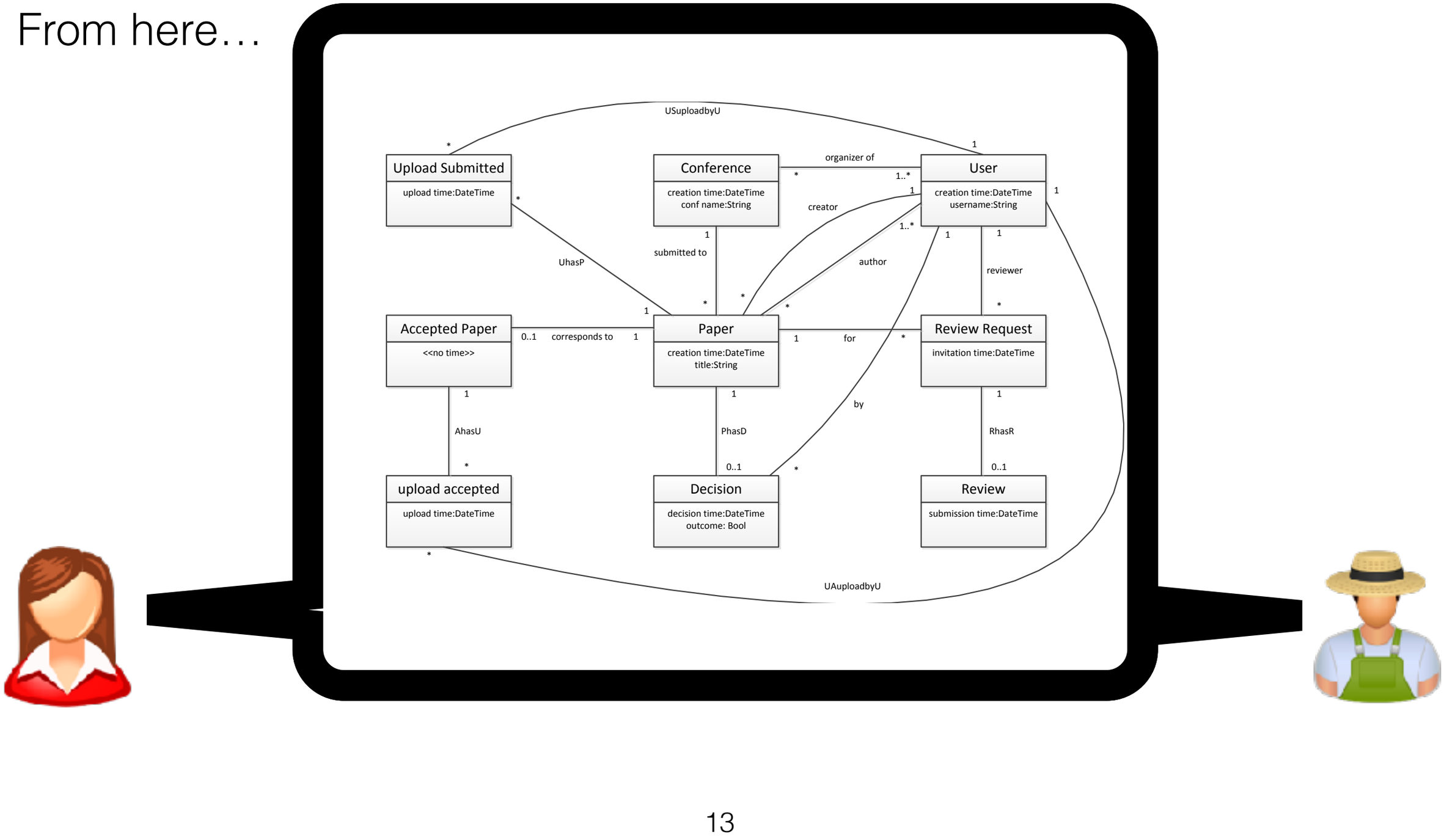
```
<log xes.version="1.0" xes.features="nested-attributes">
<trace>
   <string key=“concept:name” value=“1” />
   <event>
        <string key=“concept:name” value=“register request” />
        <date key=“time:timestamp” value=“2010-12-30T11:02:00.000+01:00” />
   </event>
</trace>
```

Understanding Reality…



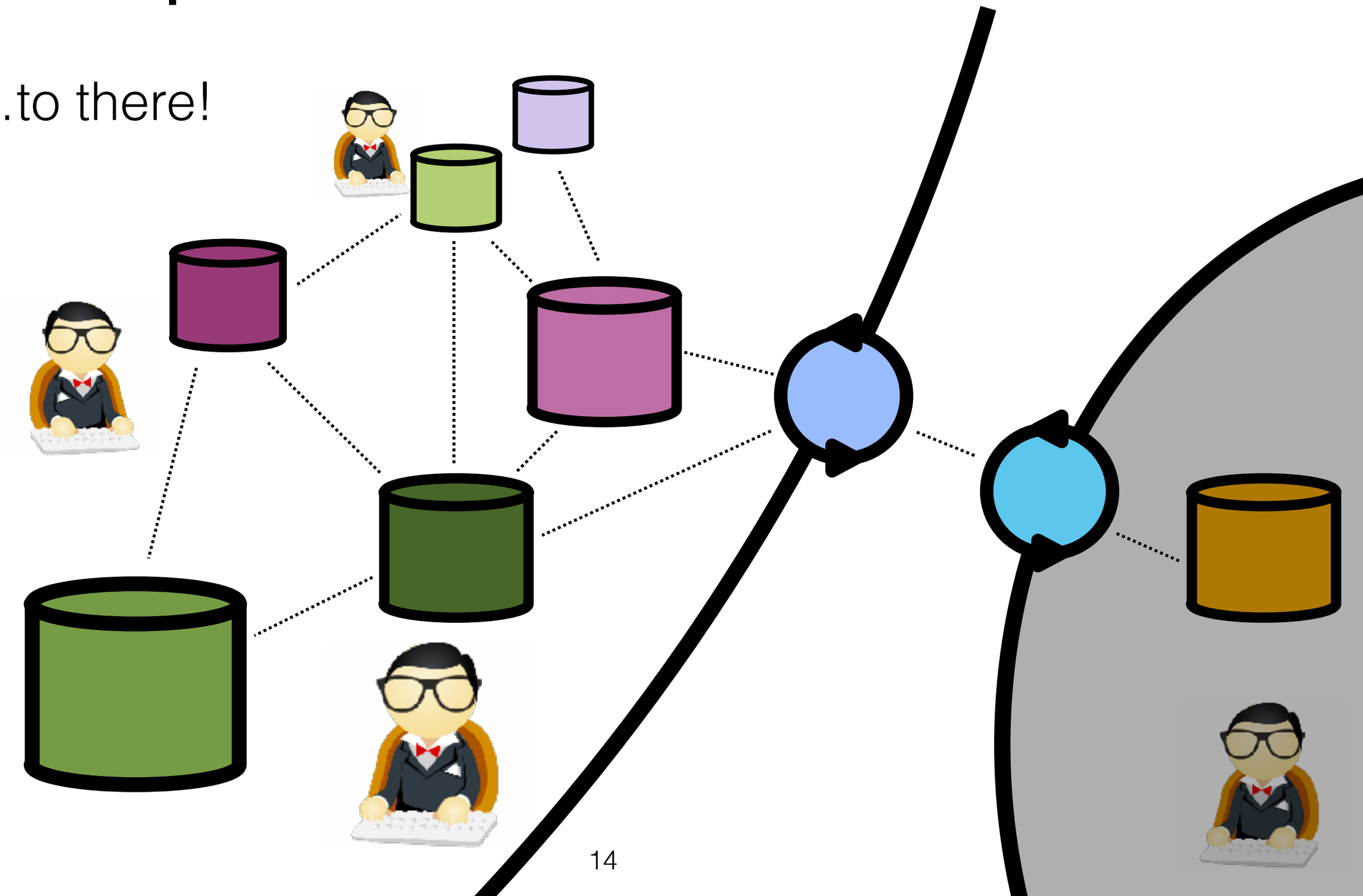
Impedance Mismatch

From here…

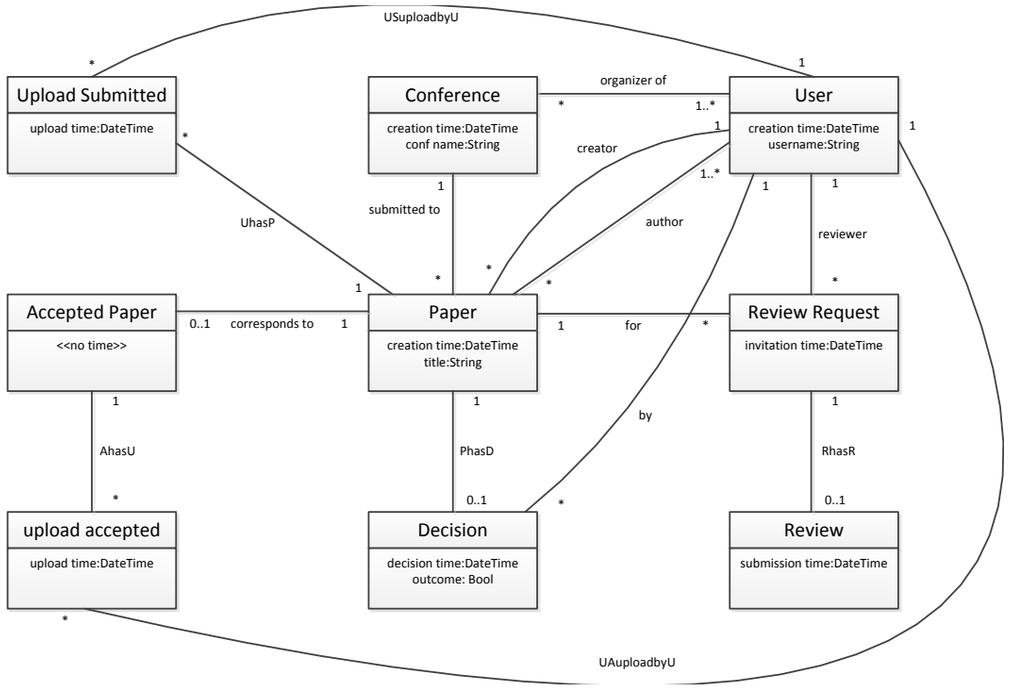


Impedance Mismatch

…to there!



Key Issues



- How to resolve the "impedance mismatch"?

- How to get a "view" of the data tailored to process mining?

LOGIN

ID	User
1	Alifah Syamsiyah
2	Marco Montali
3	Diego Calvanese
4	Wil van der Aalst

CONFERENCE

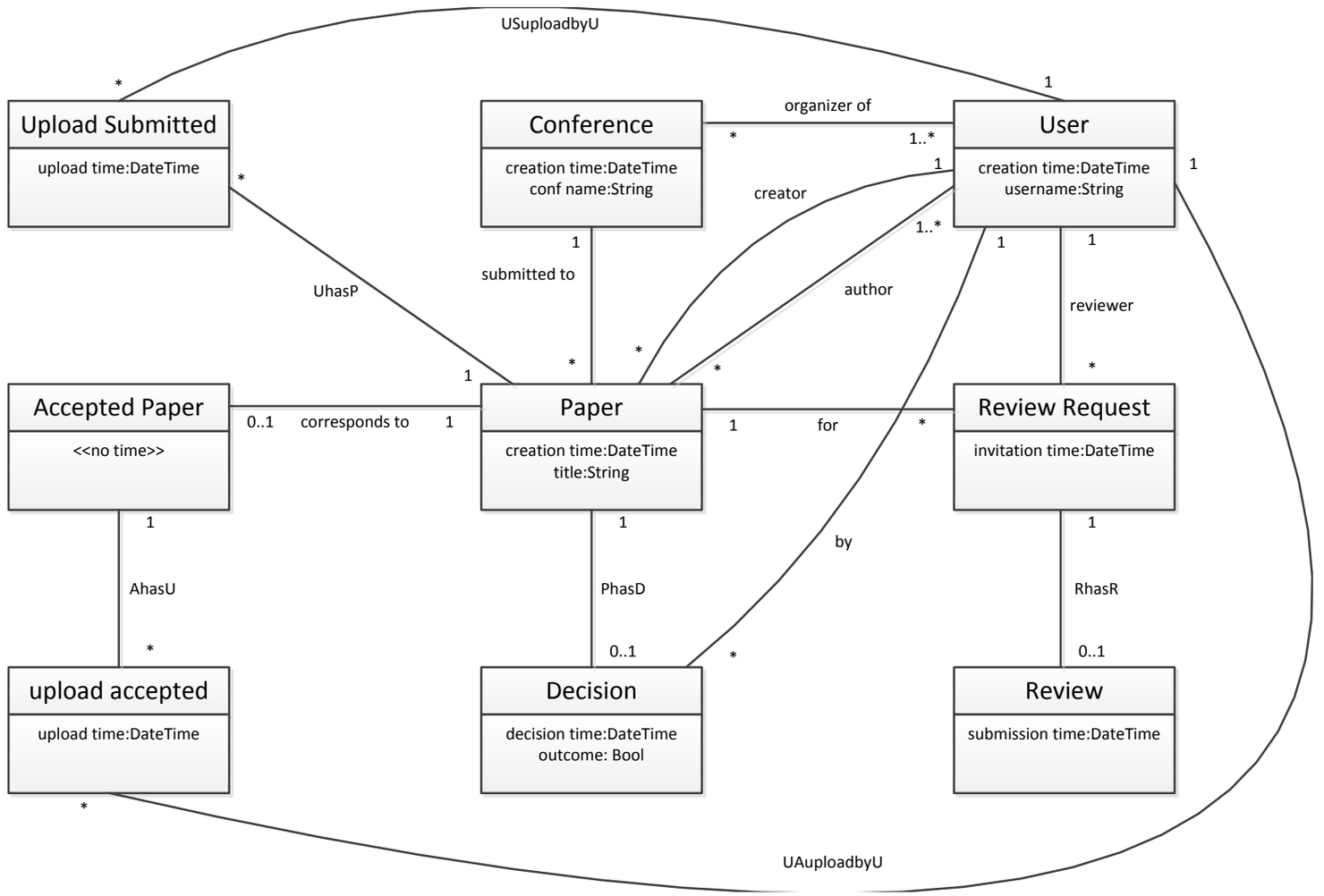
ID	Name	Organizer	Time
666	BPM 2015	2	2015-02-14 01:00:00
667	Caise 2015	4	2015-03-06 01:00:00
668	ER 2015	4	2015-03-26 01:00:00
669	EDOC 2015	2	2015-04-05 03:00:00

PAPERINFO

ID	Title	CT	User	Conf	Type	Status
1	Ontop at Work	2015-03-02 15:09:35	1	669	FP	RX
2	A Survey of Web Services	2015-03-02 12:36:01	3	668	SP	RX
3	The Definitive Guide for BPM	2015-03-04 13:36:20	1	666	FP	AB

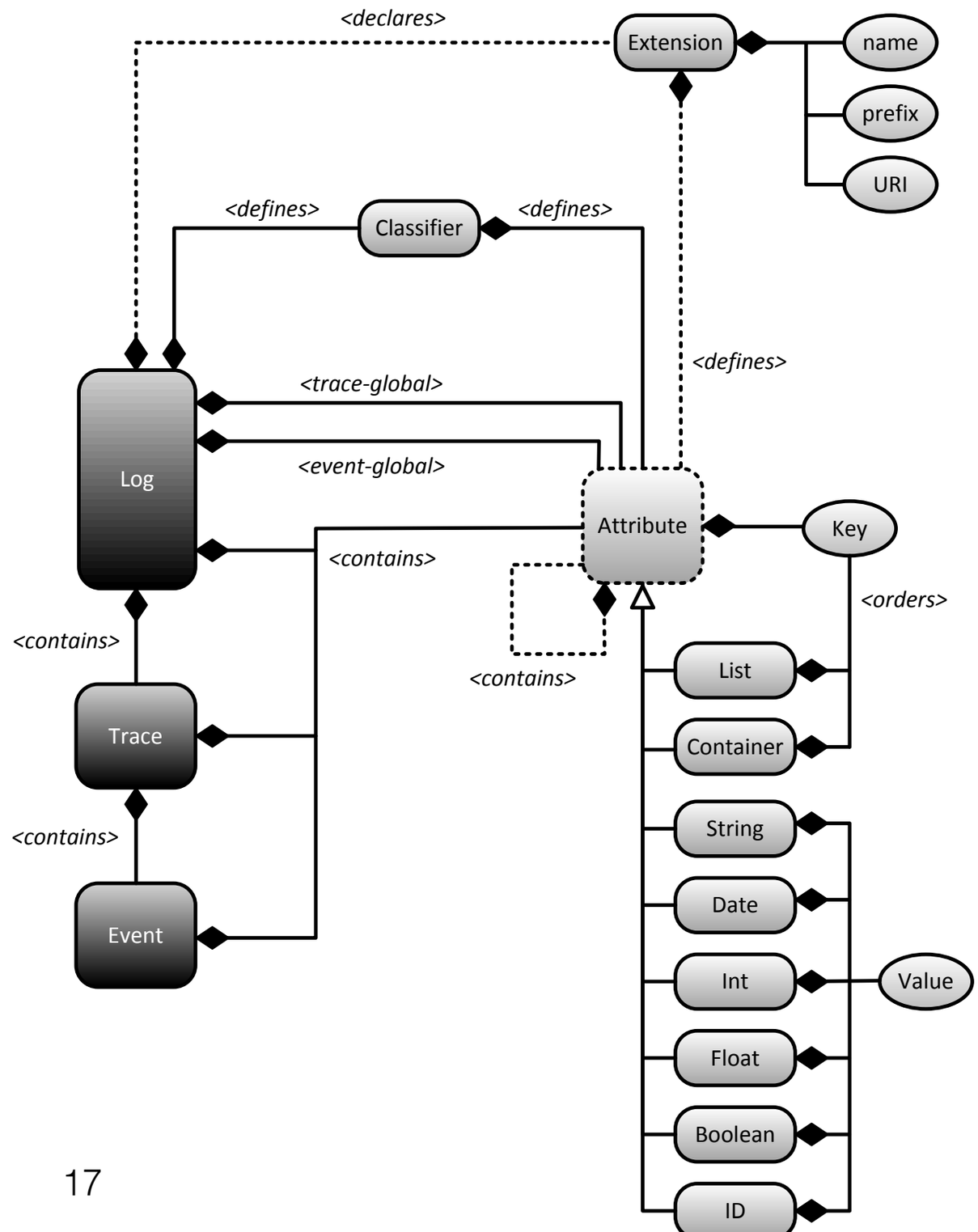
Key Issues

- Need to resolve a second impedance mismatch problem!
- From here…



Key Issues

- …To there!



Key Issues

- From here…

LOGIN

ID	User
1	Alifah Syamsiyah
2	Marco Montali
3	Diego Calvanese
4	Wil van der Aalst

CONFERENCE

ID	Name	Organizer	Time
666	BPM 2015	2	2015-02-14 01:00:00
667	Caise 2015	4	2015-03-06 01:00:00
668	ER 2015	4	2015-03-26 01:00:00
669	EDOC 2015	2	2015-04-05 03:00:00

PAPERINFO

ID	Title	CT	User	Conf	Type	Status
1	Ontop at Work	2015-03-02 15:09:35	1	669	FP	RX
2	A Survey of Web Services	2015-03-02 12:36:01	3	668	SP	RX
3	The Definitive Guide for BPM	2015-03-04 13:36:20	1	666	FP	AB

Key Issues

- …To there!

log

Case ID	Activity	Timestamp	Resource
1	Submit paper	01-07-2015:10.05	Anna
	Invite reviewer	03-07-2015:12.00	Budi
	Get review	10-07-2015:16.10	Rudy
	Acceptance	12-07-2015:15.00	Rudy
	Submit final paper	19-07-2015:19.15	Anna
2	Submit paper	02-07-2015:17.19	John
	Invite reviewer	03-07-2015:12.00	Tiara
	Get review	11-07-2015:15.45	Clara
	Rejection	12-07-2015:15.00	Clara

trace

event

Impedance Mismatch is Really an Issue

Crompton (2008): domain experts **loose too much time to dig into data** and turn them into knowledge

- Engineers in the oil/gas industry: 30-70% of their working time spent for **data searching** and **data quality**

Ontology-based Data Access

For additional details than the one given in the next slides, see separate slides on OBDA.

Optique

Scalable, End-User Access to Big Data

- http://optique-project.eu
- Goal: engineer techniques for accessing data through domain ontologies
- Case studies: Statoil, Siemens

Facts on Statoil

- **1000 TB of dati** inside relational DBMSs
- Schemas **not aligned**
- More than **2000 tables**, in a plethora of different DBs
- **900 experts** part of “Statoil Exploration”
  - Up to **4 days to formulate queries and encode them in SQL**

Query Example

Show all norwegian wellbores with some aditional attributes (wellbore id, completion date, oldest penetrated age,result). Limit to all wellbores with a core and show attributes like (wellbore id, core number, top core depth, base core depth, intersecting stratigraphy). Limit to all wellbores with core in Brentgruppen and show key atributes in a table. After connecting to EPDS (slegge) we could for instance limit futher to cores in Brent with measured permeability and where it is larger than a given value, for instance 1 mD. We could also find out whether there are cores in Brent which are not stored in EPDS (based on NPD info) and where there could be permeability values. Some of the missing data we possibly own, other not.

```
SELECT [...]
FROM
db_name.table1 table1,
db_name.table2 table2a,
db_name.table2 table2b,
db_name.table3 table3a,
db_name.table3 table3b,
db_name.table3 table3c,
db_name.table3 table3d,
db_name.table4 table4a,
db_name.table4 table4b,
db_name.table4 table4c,
db_name.table4 table4d,
db_name.table4 table4e,
db_name.table4 table4f,
db_name.table5 table5a,
db_name.table5 table5b,
db_name.table6 table6a,
db_name.table6 table6b,
db_name.table7 table7a,
db_name.table7 table7b,
db_name.table8 table8,
db_name.table9 table9,
db_name.table10 table10a,
db_name.table10 table10b,
db_name.table10 table10c,
db_name.table11 table11,
db_name.table12 table12,
db_name.table13 table13,
db_name.table14 table14,
db_name.table15 table15,
db_name.table16 table16
WHERE [...]
```

```
table2a.attr1='keyword' AND
table3a.attr2=table10c.attr1 AND
table3a.attr6=table6a.attr3 AND
table3a.attr9='keyword' AND
table4a.attr10 IN ('keyword') AND
table4a.attr1 IN ('keyword') AND
table5a.kinds=table4a.attr13 AND
table5b.kinds=table4c.attr74 AND
table5b.name='keyword' AND
(table6a.attr19=table10c.attr17 OR
(table6a.attr2 IS NULL AND
table10c.attr4 IS NULL)) AND
table6a.attr14=table5b.attr14 AND
table6a.attr2='keyword' AND
(table6b.attr14=table10c.attr8 OR
(table6b.attr4 IS NULL AND
table10c.attr7 IS NULL)) AND
table6b.attr19=table5a.attr55 AND
table6b.attr2='keyword' AND
table7a.attr19=table2b.attr19 AND
table7a.attr17=table15.attr19 AND
table4b.attr11='keyword' AND
table8.attr19=table7a.attr80 AND
table8.attr19=table13.attr20 AND
table8.attr4='keyword' AND
table9.attr10=table16.attr11 AND
table3b.attr19=table10c.attr18 AND
table3b.attr22=table12.attr63 AND
table3b.attr66='keyword' AND
table10a.attr54=table7a.attr8 AND
table10a.attr70=table10c.attr10 AND
table10a.attr16=table4d.attr11 AND
table4c.attr99='keyword' AND
table4c.attr1='keyword' AND
```

```
table11.attr10=table5a.attr10 AND
table11.attr40='keyword' AND
table11.attr50='keyword' AND
table2b.attr1=table1.attr8 AND
table2b.attr9 IN ('keyword') AND
table2b.attr2 LIKE 'keyword'% AND
table12.attr9 IN ('keyword') AND
table7b.attr1=table2a.attr10 AND
table3c.attr13=table10c.attr1 AND
table3c.attr10=table6b.attr20 AND
table3c.attr13='keyword' AND
table10b.attr16=table10a.attr7 AND
table10b.attr11=table7b.attr8 AND
table10b.attr13=table4b.attr89 AND
table13.attr1=table2b.attr10 AND
table13.attr20=''keyword'' AND
table13.attr15='keyword' AND
table3d.attr49=table12.attr18 AND
table3d.attr18=table10c.attr11 AND
table3d.attr14='keyword' AND
table4d.attr17 IN ('keyword') AND
table4d.attr19 IN ('keyword') AND
table16.attr28=table11.attr56 AND
table16.attr16=table10b.attr78 AND
table16.attr5=table14.attr56 AND
table4e.attr34 IN ('keyword') AND
table4e.attr48 IN ('keyword') AND
table4f.attr89=table5b.attr7 AND
table4f.attr45 IN ('keyword') AND
table4f.attr1='keyword' AND
table10c.attr2=table4e.attr19 AND
(table10c.attr78=table12.attr56 OR
(table10c.attr55 IS NULL AND
table12.attr17 IS NULL))
```

```
SELECT [...]
FROM
db_name.table1 table1,
db_name.table2 table2a,
db_name.table2 table2b,
db_name.table3 table3a,
db_name.table3 table3b,
db_name.table3 table3c,
db_name.table3 table3d,
db_name.table4 table4a,
db_name.table4 table4b,
db_name.table10 table10b,
db_name.table10 table10c,
db_name.table11 table11,
db_name.table12 table12,
db_name.table13 table13,
db_name.table14 table14,
db_name.table15 table15,
db_name.table16 table16
WHERE [...]
```

```
table2a.attr1='keyword' AND
table3a.attr2=table10c.attr1 AND
table3a.attr6=table6a.attr3 AND
table3a.attr9='keyword' AND
table4a.attr10 IN ('keyword') AND
table4a.attr1 IN ('keyword') AND
table5a.kinds=table4a.attr13 AND
table5b.kinds=table4c.attr74 AND
table5b.name='keyword' AND
(table6a.attr19=table10c.attr17 OR
(table6a.attr2 IS NULL AND
table8.attr4='keyword' AND
table9.attr10=table16.attr11 AND
table3b.attr19=table10c.attr18 AND
table3b.attr22=table12.attr63 AND
table3b.attr66='keyword' AND
table10a.attr54=table7a.attr8 AND
table10a.attr70=table10c.attr10 AND
table10a.attr16=table4d.attr11 AND
table4c.attr99='keyword' AND
table4c.attr1='keyword' AND
```

```
table11.attr10=table5a.attr10 AND
table11.attr40='keyword' AND
table11.attr50='keyword' AND
table2b.attr1=table1.attr8 AND
table2b.attr9 IN ('keyword') AND
table2b.attr2 LIKE 'keyword'% AND
table12.attr9 IN ('keyword') AND
table7b.attr1=table2a.attr10 AND
table3c.attr13=table10c.attr1 AND
table3c.attr10=table6b.attr20 AND
table3c.attr13='keyword' AND
table16.attr5=table14.attr56 AND
table4e.attr34 IN ('keyword') AND
table4e.attr48 IN ('keyword') AND
table4f.attr89=table5b.attr7 AND
table4f.attr45 IN ('keyword') AND
table4f.attr1='keyword' AND
table10c.attr2=table4e.attr19 AND
(table10c.attr78=table12.attr56 OR
(table10c.attr55 IS NULL AND
table12.attr17 IS NULL))
```

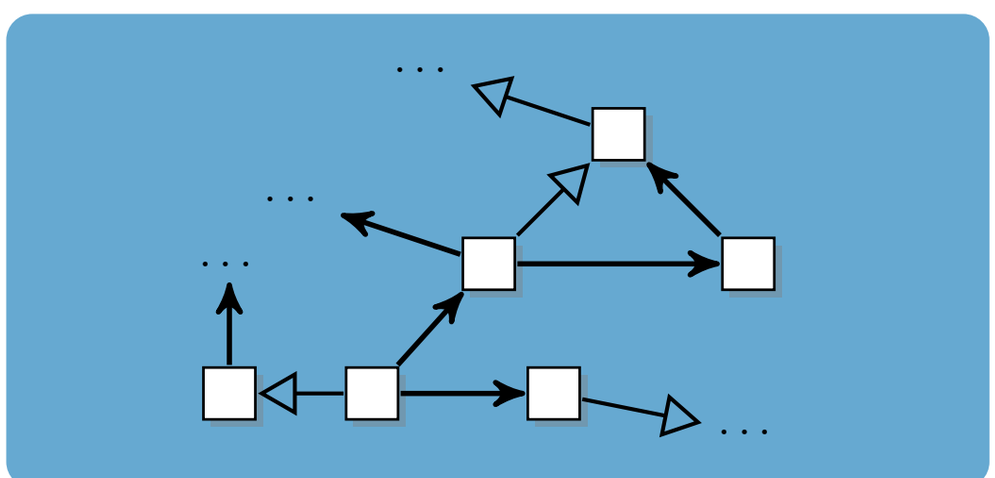
50.000.000

€/year

Ontology-Based Data Access

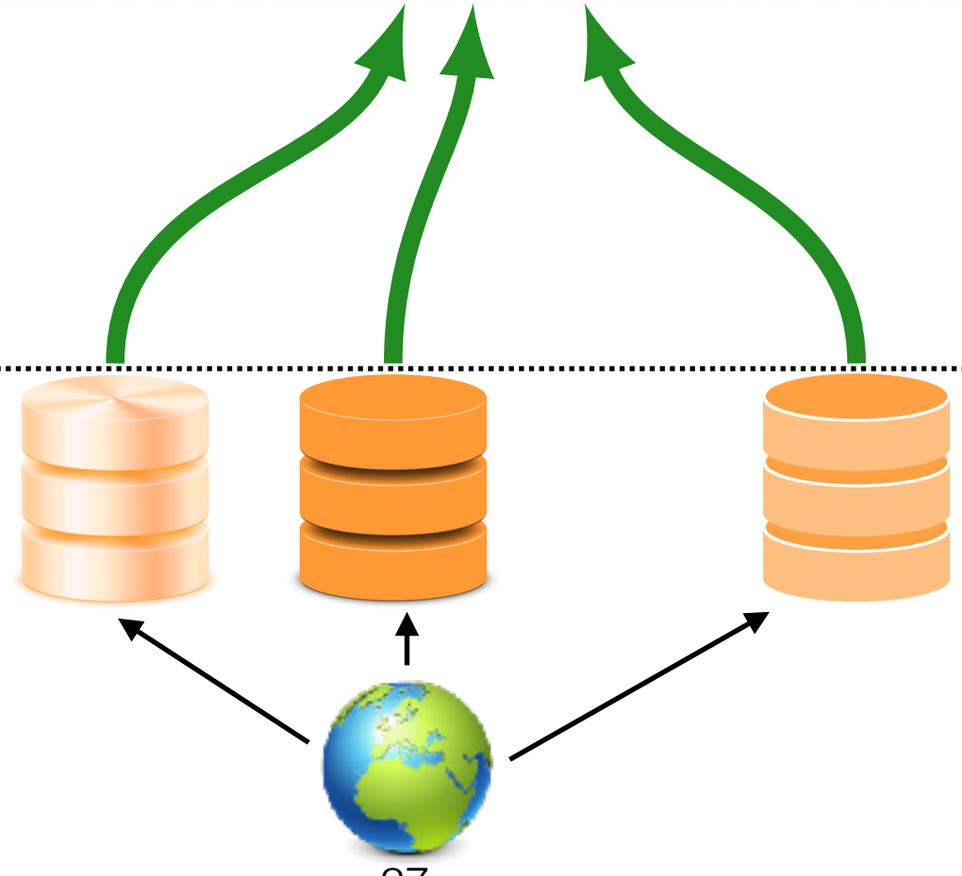
Query

Result



"Lightweight" conceptual model

Mapping

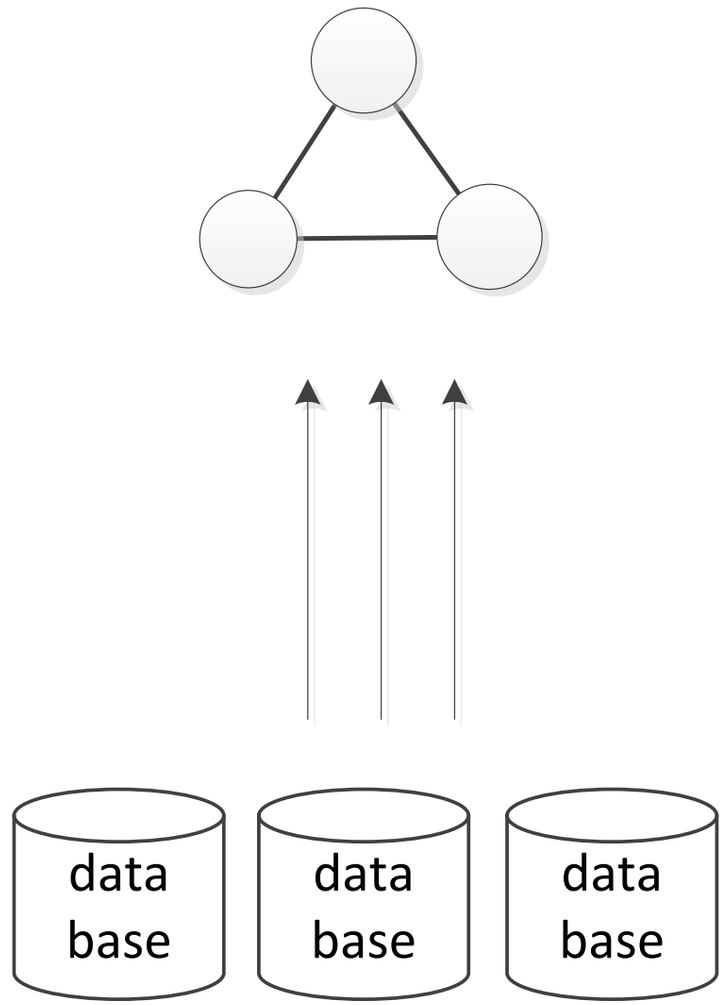


Data sources

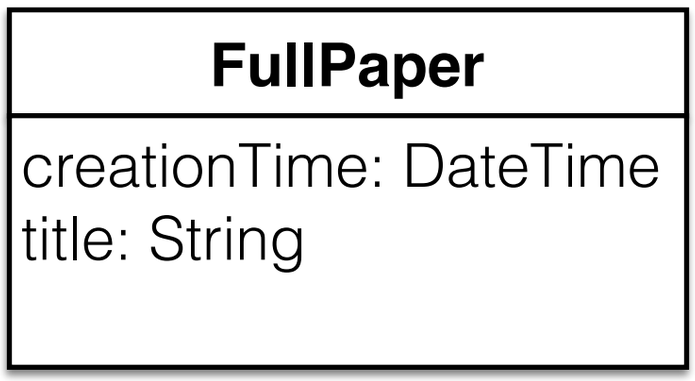
ontop

- Open-source OBDA technology developed here at UNIBZ
- Fully supports semantic web standards (OWL/ SPARQL)
- Integrates with many different relational DBMSs
- Apache 2 open license
- http://ontop.inf.unibz.it

Resolving the Impedance Mismatch



Resolving the Impedance Mismatch



mappingId	**fp-mapping**
target	**paper{ID} a :FullPaper; :title {Title}; :creationTime{CT}**
source	**select I.ID, I.Title, I.CT** **from PaperInfo I** **where I.Type = "FP"**

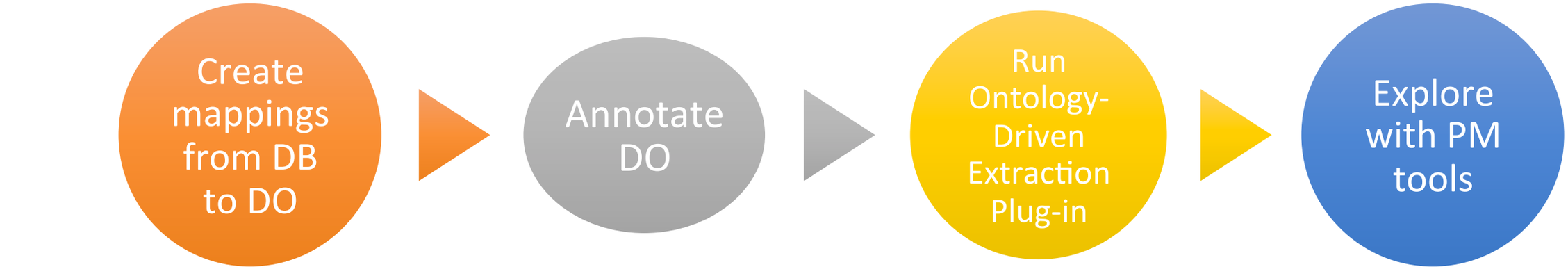
PaperInfo

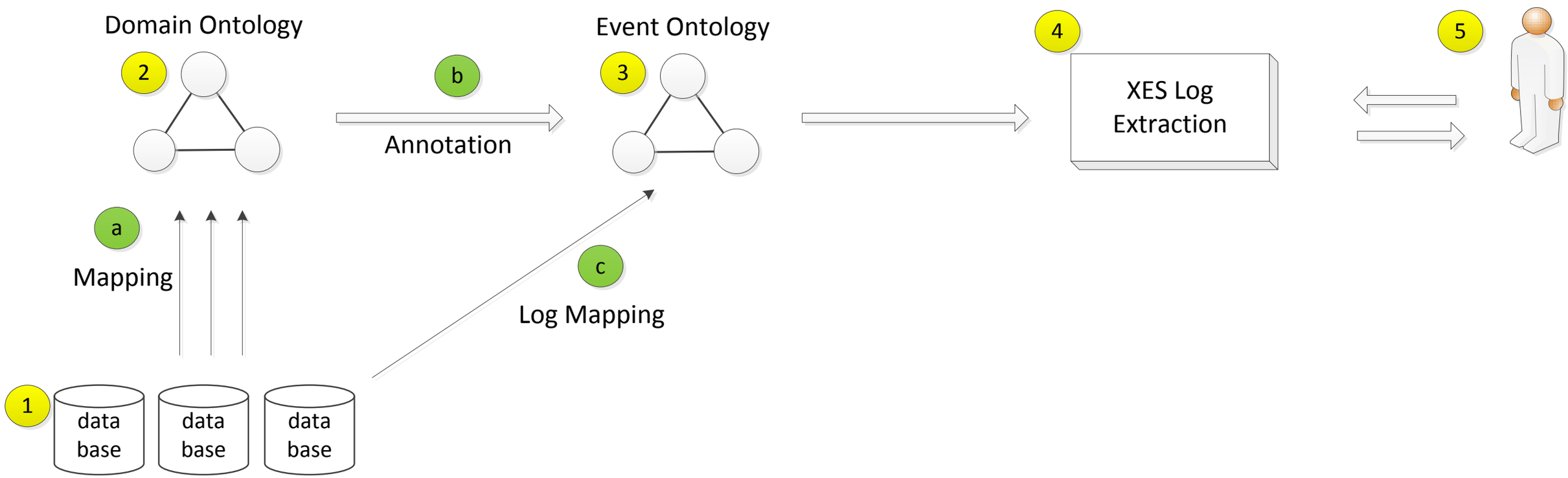
ID	Title	CT	User	Conf	Type	Status
1	Ontop at Work	2015-03-02 15:09:35	1	669	FP	RX
2	A Survey of Web Services	2015-03-02 12:36:01	3	668	SP	RX
3	The Definitive Guide for BPM	2015-03-04 13:36:20	1	666	FP	AB

My DB May Be Very Nice

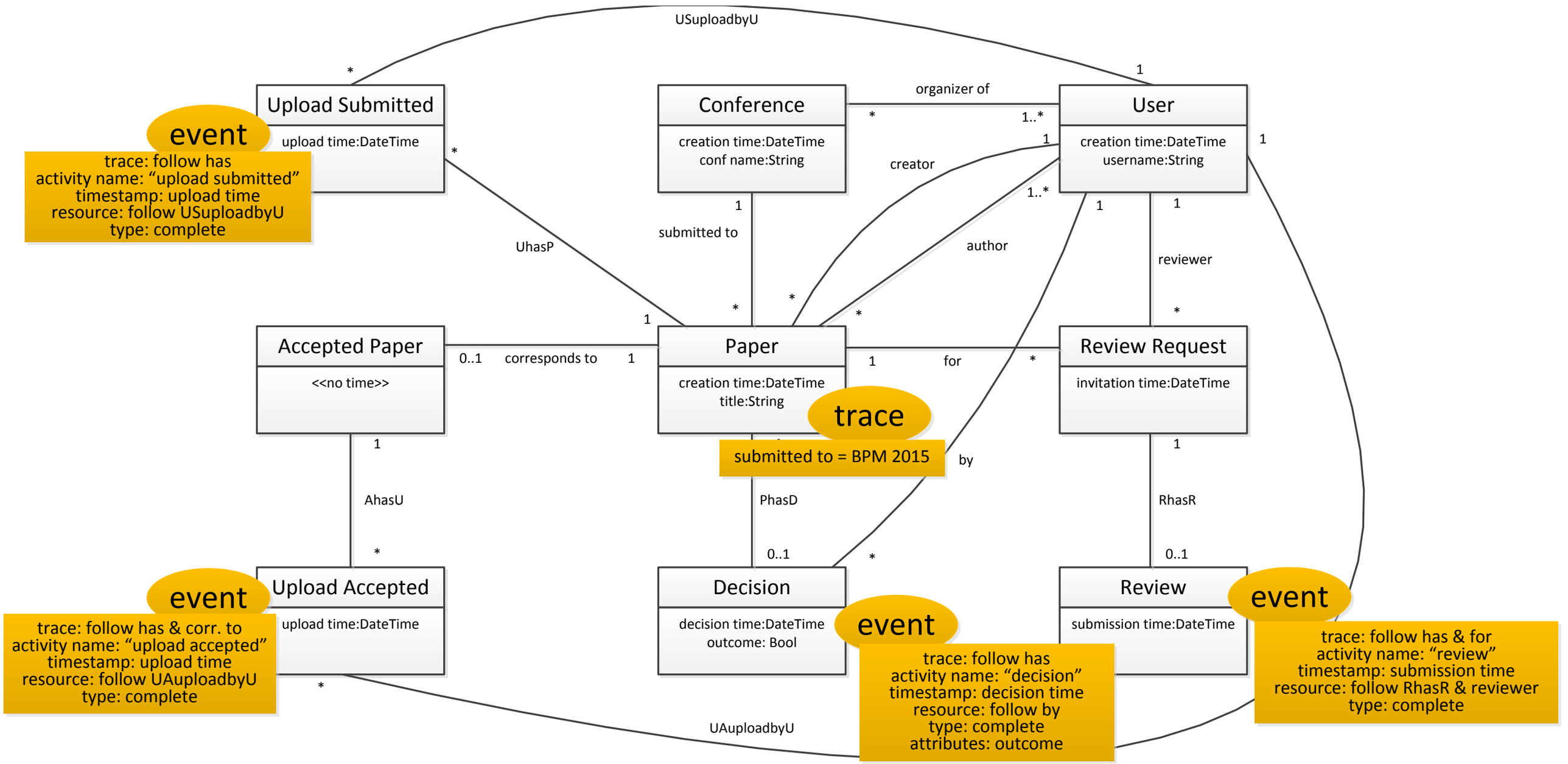
- We can use ontology bootstrapping to automatically create:
  - a conceptual model that mirrors 1-1 the relational DB
  - identity mappings
- The bootstrapped ontology and mappings need to be manually refined
- Still useful for “small” case studies

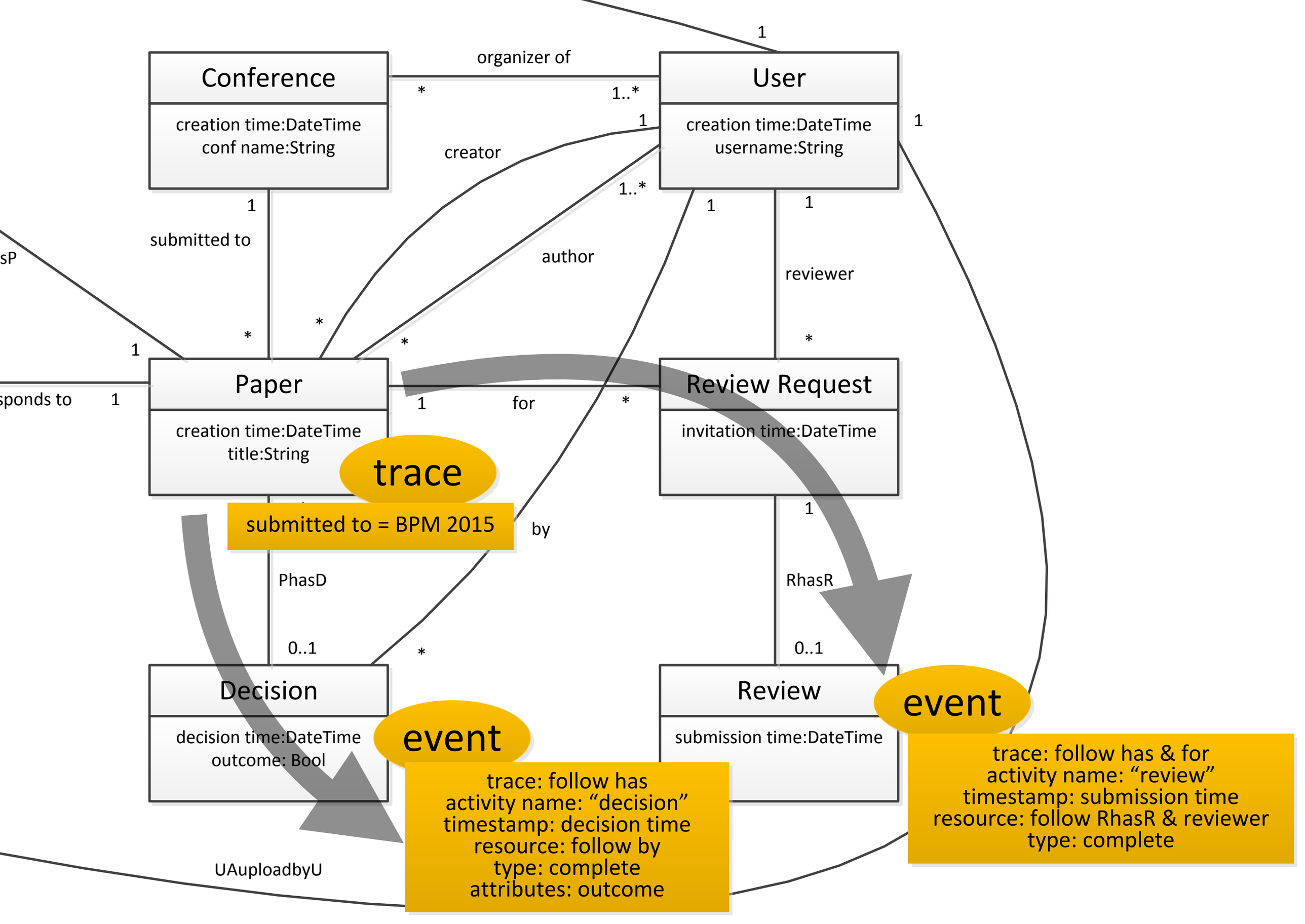
Our Framework



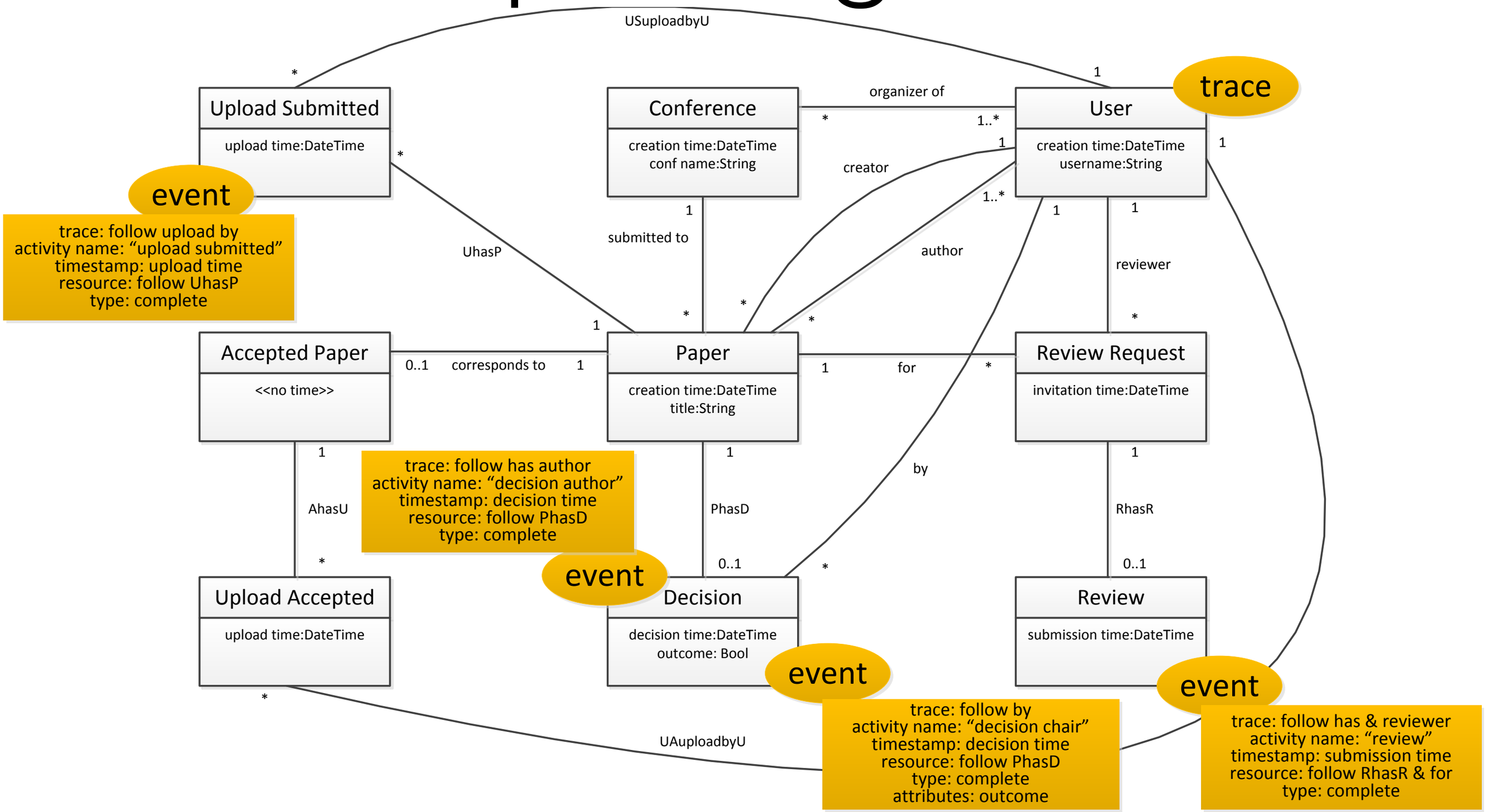


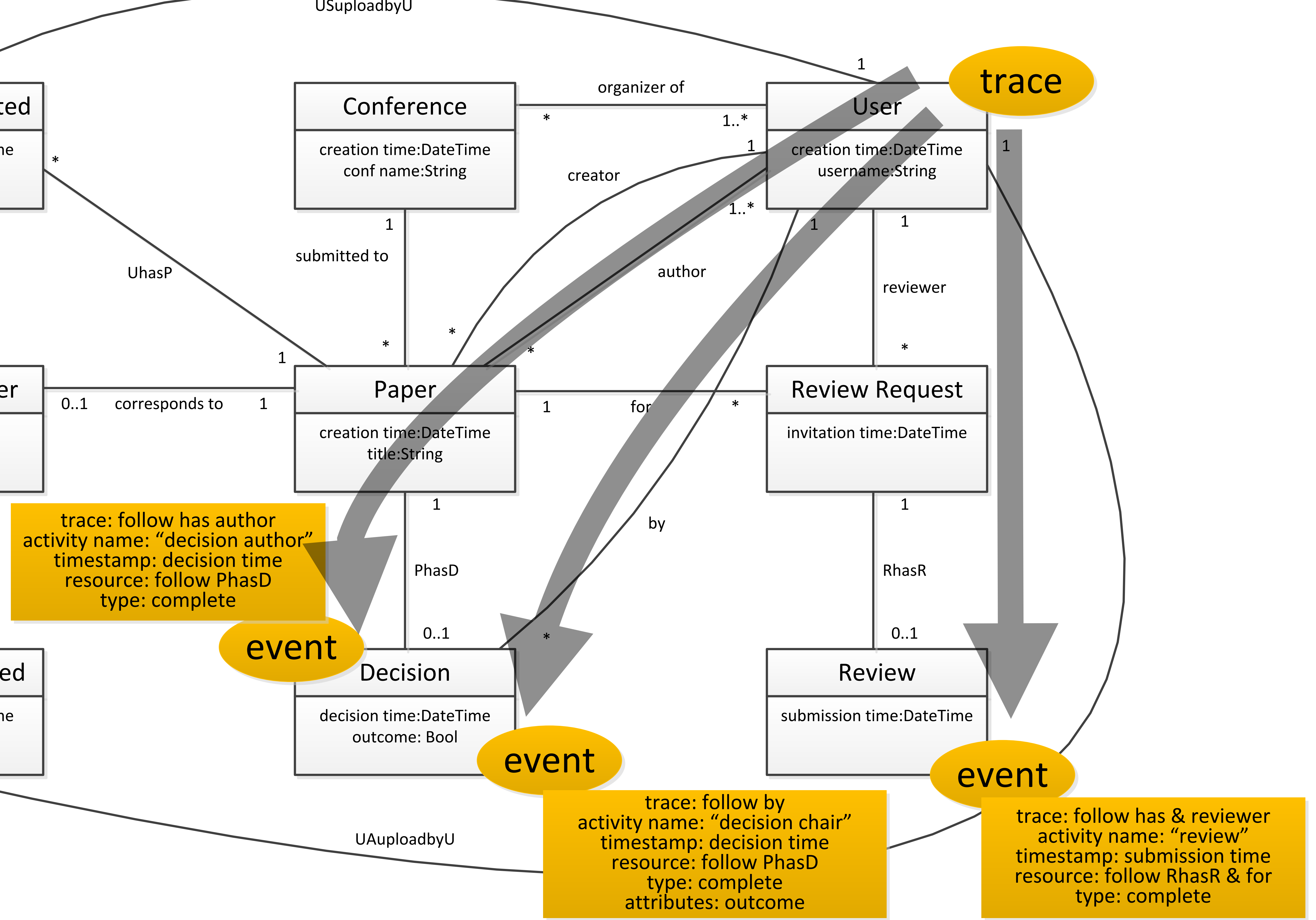
Log Annotations



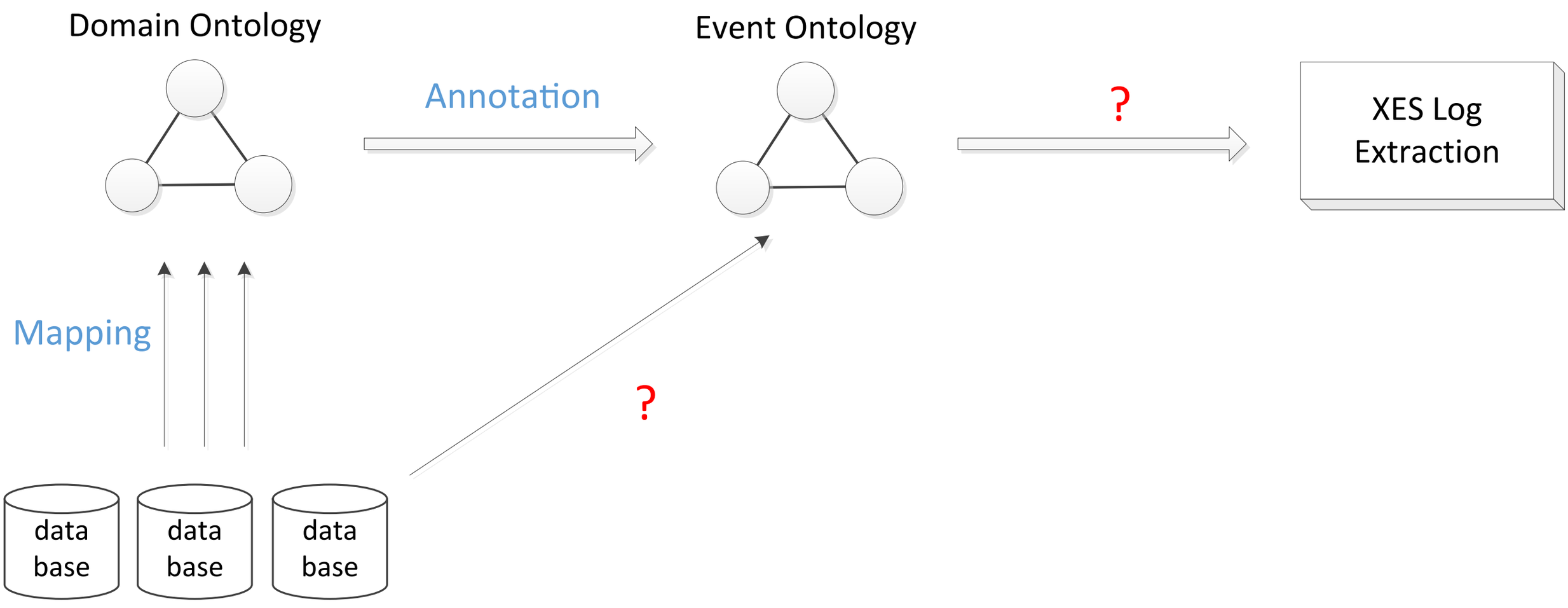

1
organizer of
Conference
creation time:DateTime
conf name:String
*
1..*
User
creation time:DateTime
username:String
1
creator
1..*
1
1
1
submitted to
author
reviewer
*
*
*
*
1
sP
sponds to
1
Paper
creation time:DateTime
title:String
1
for
*
Review Request
invitation time:DateTime
trace
submitted to = BPM 2015
by
1
PhasD
RhasR
0..1
0..1
*
Decision
decision time:DateTime
outcome: Bool
Review
submission time:DateTime
event
event
trace: follow has
activity name: “decision”
timestamp: decision time
resource: follow by
type: complete
attributes: outcome
trace: follow has & for
activity name: “review”
timestamp: submission time
resource: follow RhasR & reviewer
type: complete
UAuploadbyU

Multiple Log Views

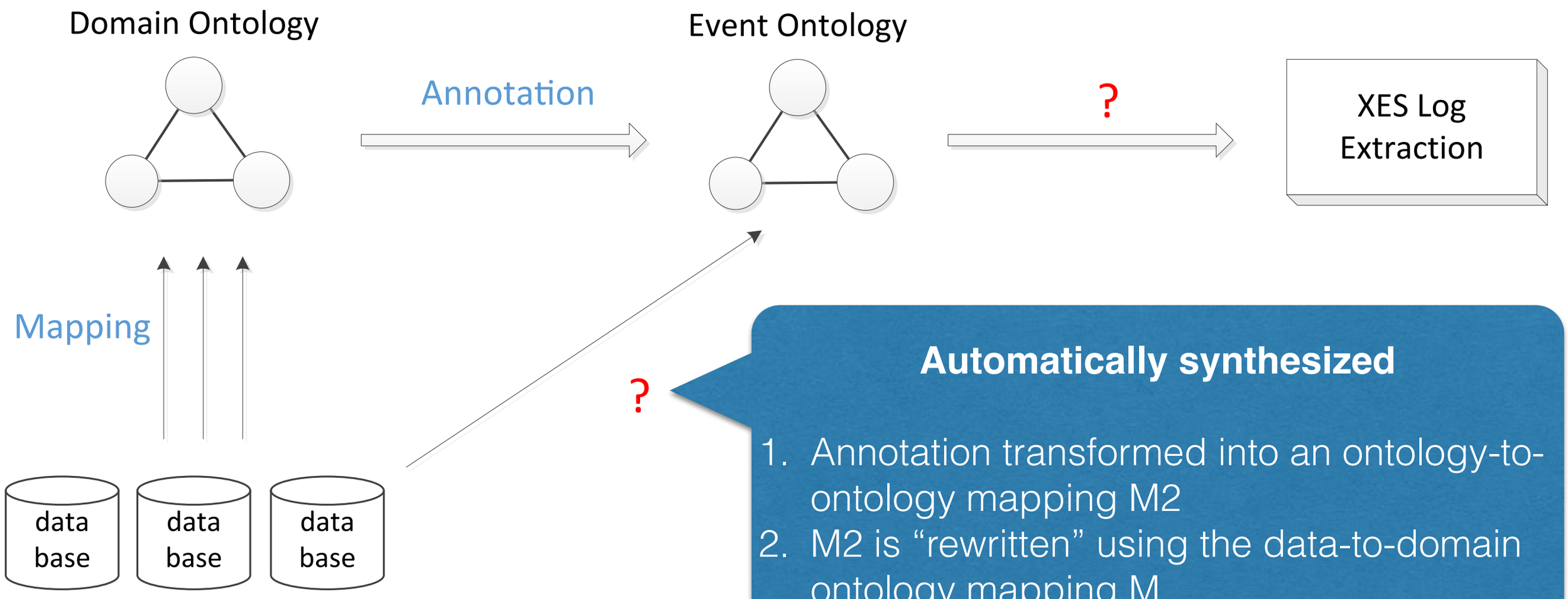



USuploadbyU
Conference
creation time:DateTime
conf name:String
organizer of
User
creation time:DateTime
username:String
trace
creator
author
submitted to
UhasP
reviewer
Paper
creation time:DateTime
title:String
corresponds to
for
Review Request
invitation time:DateTime
by
PhasD
RhasR
trace: follow has author
activity name: “decision author”
timestamp: decision time
resource: follow PhasD
type: complete
event
Decision
decision time:DateTime
outcome: Bool
Review
submission time:DateTime
event
event
UAuploadbyU
trace: follow by
activity name: “decision chair”
timestamp: decision time
resource: follow PhasD
type: complete
attributes: outcome
trace: follow has & reviewer
activity name: “review”
timestamp: submission time
resource: follow RhasR & for
type: complete

And Now?


Domain Ontology
Annotation
Event Ontology
?
XES Log Extraction
Mapping
?
data base
data base
data base

Mapping Synthesis



Mapping Synthesis

To synthesize the data-to-event ontology (D2EO) mapping M3:

1. Each query Q1 associated to an annotation (for EO element E) is rewritten wrt domain ontology into a query Q2

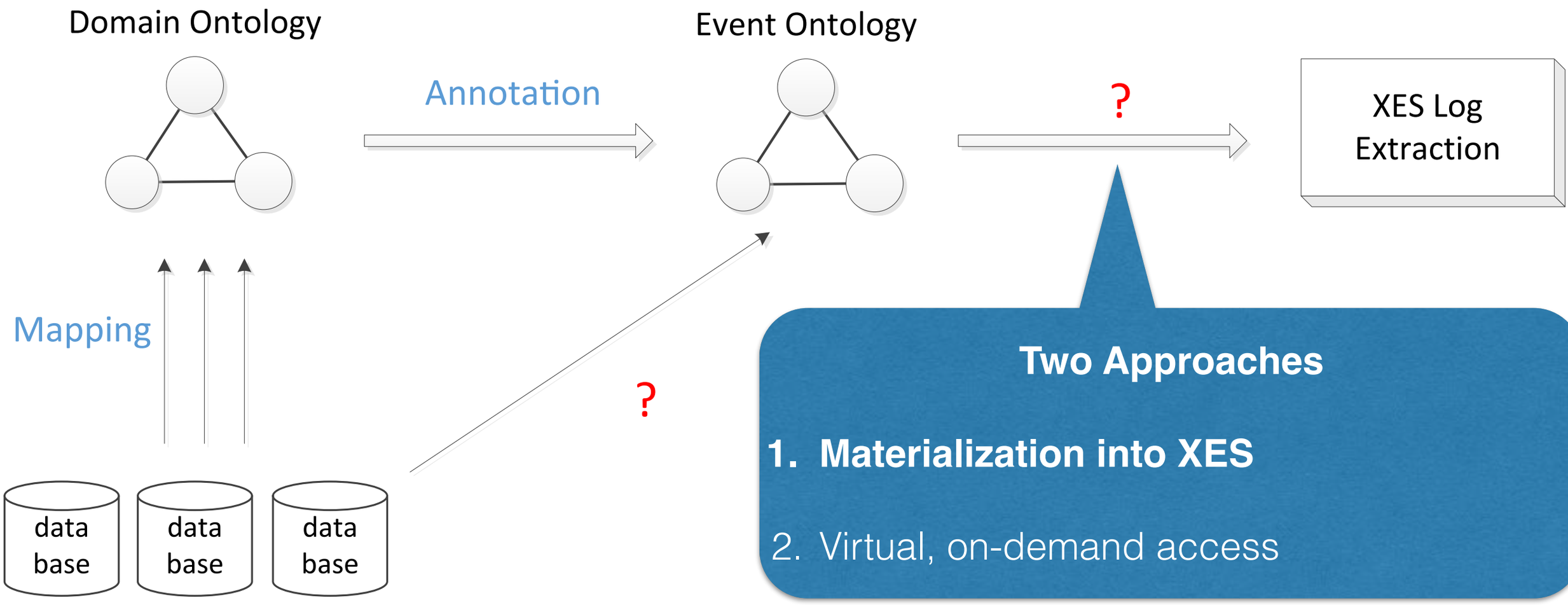
2. Each rewritten query Q2 is unfolded wrt D2DO mapping M into a query Q3 over the data

3. Each rewritten and unfolded query Q3 becomes the source query of a D2EO mapping assertion in M3, which has E as target part

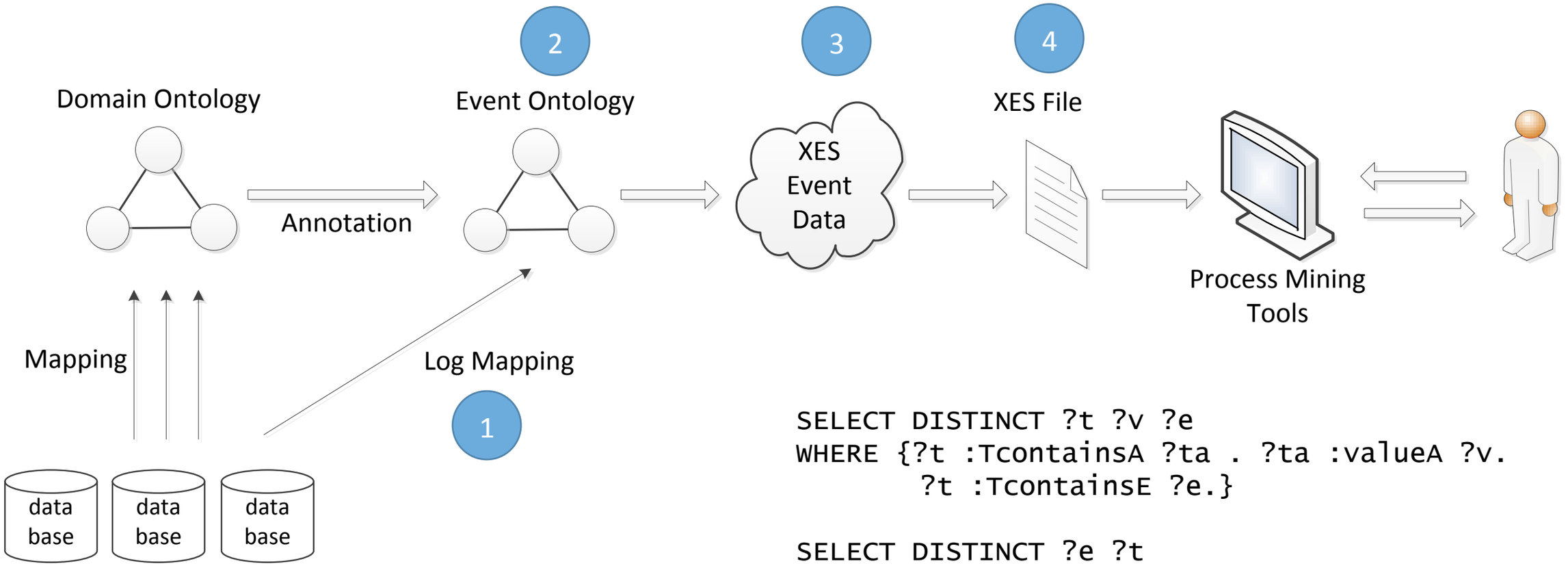
Steps 1 and 2 performed using query rewriting algorithm of ontop.

For Step 3, we need to push the URI-construction part inserted by ontop in Q3, from the source part of the mapping to the target part.

Use of Synthesized Mapping



Log Materialization



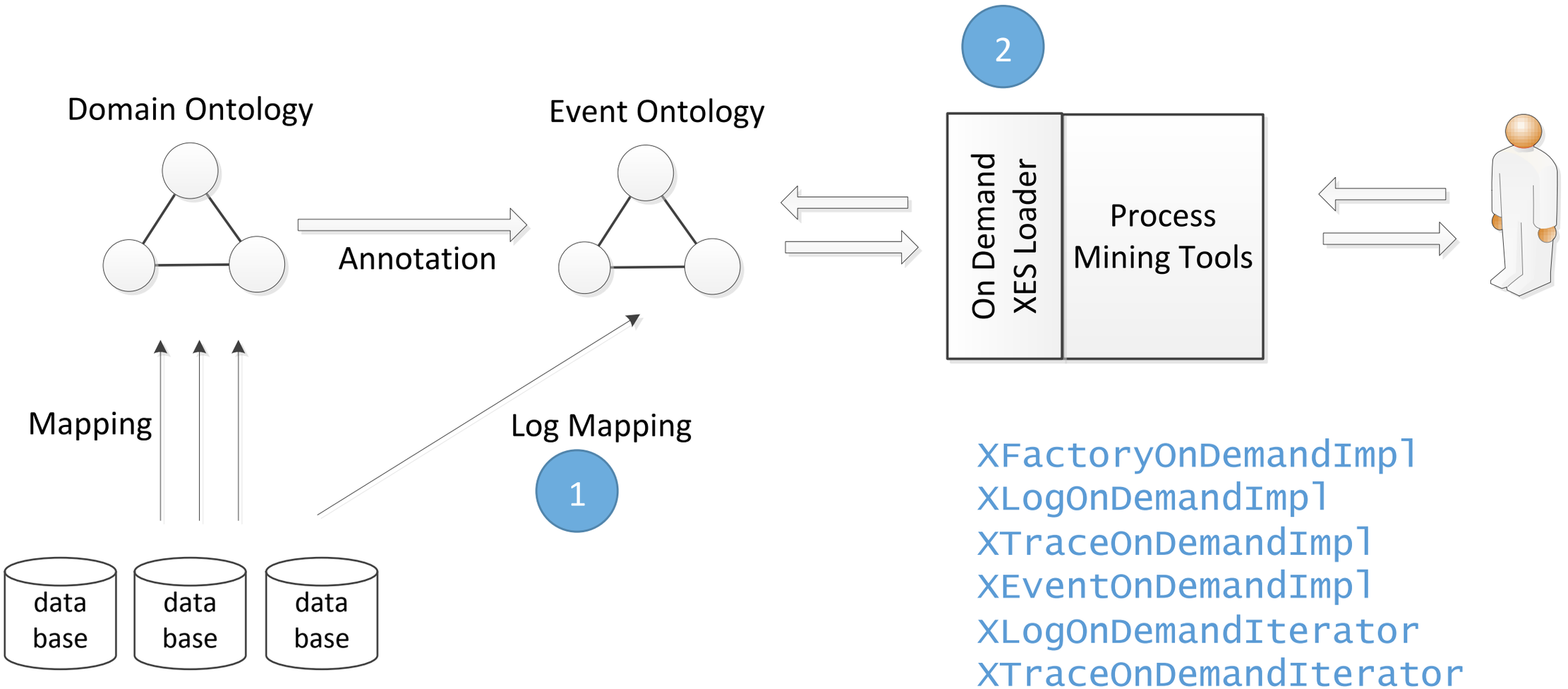

```
SELECT DISTINCT ?t ?v ?e
WHERE {?t :TcontainsA ?ta . ?ta :valueA ?v.
       ?t :TcontainsE ?e.}

SELECT DISTINCT ?e ?t
WHERE {?e :EcontainsA ?a . ?a :typeA ?t.}

SELECT DISTINCT ?e ?t
WHERE {?e :EcontainsA ?a . ?a :keyA ?t.}

SELECT DISTINCT ?e ?t
WHERE {?e :EcontainsA ?a . ?a :valueA ?t.}
```

Log Virtualization




```
XFactoryOnDemandImpl
XLogOnDemandImpl
XTraceOnDemandImpl
XEventOnDemandImpl
XLogOnDemandIterator
XTraceOnDemandIterator
```

`xlog.get(7).get(90)` to retrieve te event in index 7th inside the 90th trace in a log

Prototype Implementation

1. Editor for lightweight ontology
    - ontology is represented as UML class diagram
    - exports a standard OWL 2 QL ontology
    - proprietary format for layout information
2. Annotation editor
    - operates on UML representation of ontology
    - exports annotation in proprietary JSON format
3. Log extractor
    - design time component: generates data-to-event ontology mappings
    - run time component: extract XES event log using materialized approach

Ongoing Work

- We are optimizing and testing the scalability of the materialized approach. Fine-tuning is a must!

- We still have to integrate the “virtual” approach with process mining algorithms, to provide them access to the data.

- We are looking for interesting case studies!

Acknowledgments

All coauthors of this research, in particular

Marco Montali (UNIBZ)
Emre Kalayci (UNIBZ)
Ario Santoso (UNIBZ)
Riccardo De Masellis (FBK-Trento)
Chiara Difrancescomarino (FBK-Trento)
Chiara Ghidini (FBK-Trento)
Sergio Tessaris (UNIBZ)
Alifah Syamsiyah (TU/e)
Wil van der Aalst (TU/e)

KAOS Project (funded by Euregio)