XML Warehousing Meets Sociology

F.-X. Dudouet\textsuperscript{1} \quad I. Manolescu\textsuperscript{2} \\
B. Nguyen\textsuperscript{3} \quad P. Senellart\textsuperscript{2,4}

\textit{IADIS ICWI}, October 20th 2005
Outline

1. Introduction
   - Sociological Process
   - Standardization

2. Methodology

3. Experimentation

4. Conclusion
Sociological Process

1. **Formulate hypotheses**
2. **Validate on data**
   - Relevant sociological concepts (individuals, institutions...)
   - Data sources are: existing documents, interviews...
3. **Conclude and issue new hypotheses**

**Issue**

How to collect and manage large volumes of heterogeneous information?
Issue
How to collect and manage large volumes of heterogeneous information?

1. Formulate hypotheses
2. Validate on data
   - Relevant sociological concepts (individuals, institutions...)
   - Data sources are: existing documents, interviews...
3. Conclude and issue new hypotheses

Dudouet, Manolescu, Nguyen, Senellart
XML Warehousing Meets Sociology
Sociological Process

1. Formulate hypotheses
2. Validate on data
   - Relevant sociological concepts (individuals, institutions...)
   - Data sources are: existing documents, interviews...
3. Conclude and issue new hypotheses

Issue

How to collect and manage large volumes of heterogeneous information?
Sociological Process

1. Formulate hypotheses
2. Validate on data
   - Relevant sociological concepts (individuals, institutions...)
   - Data sources are: existing documents, interviews...
3. Conclude and issue new hypotheses

Issue
How to collect and manage large volumes of heterogeneous information?
Sociological Process

1. Formulate hypotheses
2. Validate on data
   - Relevant sociological concepts (individuals, institutions...)
   - Data sources are: existing documents, interviews...
3. Conclude and issue **new hypotheses**

**Issue**

How to collect and manage large volumes of heterogeneous information?
Sociological Process

1. Formulate hypotheses
2. Validate on data
   - Relevant sociological concepts (individuals, institutions...)
   - Data sources are: existing documents, interviews...
3. Conclude and issue new hypotheses

Issue

How to collect and manage large volumes of heterogeneous information?
Inestimable source of data

Much human activity involve Web technology

But:

Heterogeneity of sources

Not suited to classical database systems

Need of conceptual models
Case of the World Wide Web

- Inestimable source of data
- Much human activity involve Web technology

But:
- Heterogeneity of sources
- Not suited to classical database systems
- Need of conceptual models
Case of the World Wide Web

- Inestimable source of data
- Much human activity involve Web technology

But:

- Heterogeneity of sources
- Not suited to classical database systems
- Need of conceptual models
Case of the World Wide Web

- Inestimable source of data
- Much human activity involve Web technology

But:

- Heterogeneity of sources
- Not suited to classical database systems
- Need of conceptual models
Case of the World Wide Web

- Inestimable source of data
- Much human activity involve Web technology

But:

- Heterogeneity of sources
- Not suited to classical database systems
- Need of conceptual models
Standard negotiations

→ Important economic and political impact

Issue

Who? Why? How?

Example

XQuery standardization scene

- Arena quite accessible via mailing lists
- Author's acquaintance with the topic
- Process almost finished
Standard negotiations

⇒ Important economic and political impact

Issue

Who? Why? How?

Example

XQuery standardization scene

- Arena quite accessible via mailing lists
- Author’s acquaintance with the topic
- Process almost finished
Standard negotiations

⇒ Important economic and political impact

Issue

Who? Why? How?

Example

XQuery standardization scene

- Arena quite accessible via mailing lists
- Author’s acquaintance with the topic
- Process almost finished
Standard negotiations

⇒ Important economic and political impact

Issue
Who? Why? How?

Example
XQuery standardization scene

- Arena quite accessible via mailing lists
- Author’s acquaintance with the topic
- Process almost finished
Standard negociations

⇒ Important economic and political impact

**Issue**

Who? Why? How?

**Example**

XQuery standardization scene

- Arena quite accessible via mailing lists
- Author’s acquaintance with the topic
- Process almost finished
Standard negotiations

⇒ Important economic and political impact

<table>
<thead>
<tr>
<th>Issue</th>
</tr>
</thead>
<tbody>
<tr>
<td>Who? Why? How?</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>XQuery standardization scene</td>
</tr>
</tbody>
</table>

- Arena quite accessible via mailing lists
- Author’s acquaintance with the topic
- Process almost finished
Standard negotiations

⇒ Important **economic** and **political** impact

**Issue**

Who? Why? How?

**Example**

**XQuery** standardization scene

- Arena quite **accessible** via mailing lists
- **Author’s acquaintance** with the topic
- Process **almost finished**
Outline

1. Introduction

2. Methodology
   - Conceptual process
   - XML Warehousing
   - Data filtering and enrichment
   - Complementary sociological tools

3. Experimentation

4. Conclusion
Modelling and analysis process

- Modelling the relevant **sociological entities** (actors, institutions, functions, messages, time)
- Designing a **warehouse of Web resources** relevant to the sociological analysis
- Exploiting the warehouse (feeding the warehouse, issuing queries)

Queries enable verification of the hypotheses
Modelling and analysis process

- Modelling the relevant sociological entities (actors, institutions, functions, messages, time)
- Designing a warehouse of Web resources relevant to the sociological analysis
- Exploiting the warehouse (feeding the warehouse, issuing queries)

Queries enable verification of the hypotheses
Modelling and analysis process

- Modelling the relevant **sociological entities** (actors, institutions, functions, messages, time)
- Designing a **warehouse of Web resources** relevant to the sociological analysis
- **Exploiting** the warehouse (feeding the warehouse, issuing queries)

Queries enable verification of the hypotheses
Modelling and analysis process

- Modelling the relevant **sociological entities** (actors, institutions, functions, messages, time)
- Designing a **warehouse of Web resources** relevant to the sociological analysis
- Exploiting the warehouse (feeding the warehouse, issuing queries)

Queries enable verification of the hypotheses
Warehouse construction process

- Mailing list archive
- Extraction & threading
- Messages XML Warehouse
- Queries
- Data for sociologists
- Feedback & Enrichment
- Feedback & Enrichment
- Wrappers
- Information sources
- HTML
- XML
- Actors XML Warehouse

- XML Warehousing Meets Sociology

Dudouet, Manolescu, Nguyen, Senellart
XML Warehousing

Pros

- Semi-structured information
- Tree structure of a mailing list
- Simple to understand

Queries on XML warehouses: XQuery itself!
XML Warehousing

Pros

- **Semi-structured** information
- **Tree structure** of a mailing list
- **Simple** to understand

Queries on XML warehouses: XQuery itself!
XML Warehousing

Pros

- **Semi-structured** information
- **Tree structure** of a mailing list
- **Simple** to understand

Queries on XML warehouses: **XQuery itself!**
XML Warehousing

**Pros**

- **Semi-structured** information
- **Tree structure** of a mailing list
- **Simple** to understand

Queries on XML warehouses: **XQuery** itself!
Data filtering and enrichment

- **Identify** real-world objects represented in the warehouse
  - First name, last name, institution from e-mails
  - Identifying institutions participating in the process

- **Classify** these objects according to application-driven criteria
  - Issue classification queries to populate interesting classes (iterative process)
Data filtering and enrichment

- Identify **real-world objects** represented in the warehouse
  - First name, last name, institution from e-mails
  - Identifying institutions participating in the process

- Classify these objects according to **application-driven criteria**
  Issue classification queries to populate interesting classes
  (iterative process)
Data filtering and enrichment

- **Identify** real-world objects represented in the warehouse
  - First name, last name, institution from e-mails
  - Identifying institutions participating in the process

- **Classify** these objects according to application-driven criteria
  - Issue classification queries to populate interesting classes (iterative process)
Data filtering and enrichment

- Identify **real-world objects** represented in the warehouse
  - First name, last name, institution from e-mails
  - Identifying institutions participating in the process

- **Classify** these objects according to **application-driven criteria**
  - Issue classification queries to **populate** interesting classes
    (iterative process)
## Complementary sociological tools

### Issue

Information on the Web has **holes**

- **Missing** information
- Important dimensions (e.g. time) implicitly or not at all represented
- Need to **cross** various sources to establish information

### Tools

- Interviews, inside information
- Human-readable data sources
- Statistics tools (social properties and group extraction)
- Human annotation and validation
## Issue

Information on the Web has holes

- **Missing** information
- Important dimensions (e.g. time) implicitly or not at all represented
- Need to cross various sources to establish information

## Tools

- Interviews, inside information
- Human-readable data sources
- Statistics tools (social properties and group extraction)
- Human annotation and validation
Complementary sociological tools

**Issue**

Information on the Web has **holes**
- **Missing** information
- Important dimensions (e.g. time) **implicitly** or **not at all** represented
- Need to **cross** various sources to establish information

**Tools**

- Interviews, inside information
- Human-readable data sources
- Statistics tools (social properties and group extraction)
- Human annotation and validation
Complementary sociological tools

Issue

Information on the Web has holes

- Missing information
- Important dimensions (e.g. time) implicitly or not at all represented
- Need to cross various sources to establish information

Tools

- Interviews, inside information
- Human-readable data sources
- Statistics tools (social properties and group extraction)
- Human annotation and validation
Complementary sociological tools

**Issue**

Information on the Web has **holes**

- **Missing** information
- Important dimensions (e.g. time) implicitly or not at all represented
- Need to **cross** various sources to establish information

**Tools**

- Interviews, inside information
- Human-readable data sources
- Statistics tools (social properties and group extraction)
- Human annotation and validation
Complementary sociological tools

<table>
<thead>
<tr>
<th>Issue</th>
</tr>
</thead>
<tbody>
<tr>
<td>Information on the Web has <strong>holes</strong></td>
</tr>
<tr>
<td>- <strong>Missing</strong> information</td>
</tr>
<tr>
<td>- Important dimensions (e.g. time) <strong>implicitly</strong> or not at all</td>
</tr>
<tr>
<td>represented</td>
</tr>
<tr>
<td>- Need to <strong>cross</strong> various sources to establish information</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Tools</th>
</tr>
</thead>
<tbody>
<tr>
<td>- Interviews, inside information</td>
</tr>
<tr>
<td>- Human-readable data sources</td>
</tr>
<tr>
<td>- Statistics tools (social properties and group extraction)</td>
</tr>
<tr>
<td>- Human annotation and validation</td>
</tr>
</tbody>
</table>
Complementary sociological tools

**Issue**

Information on the Web has **holes**
- **Missing** information
- Important dimensions (e.g. time) **implicitly** or **not at all** represented
- Need to **cross** various sources to establish information

**Tools**

- Interviews, inside information
- Human-readable data sources
- Statistics tools (social properties and group extraction)
- Human annotation and validation
## Complementary sociological tools

### Issue

Information on the Web has **holes**
- **Missing** information
- Important dimensions (e.g. time) **implicitly or not at all** represented
- Need to **cross** various sources to establish information

### Tools

- Interviews, inside information
- Human-readable data sources
- Statistics tools (social properties and group extraction)
- Human annotation and validation
Outline

1. Introduction
2. Methodology
3. Experimentation
   - Warehouses
   - Results
   - Sociological interpretation
4. Conclusion
Message warehouse

public-qt-comments@w3.org mailing list.

**Data**

- 5626 messages
- 2718 threads
- Maximum thread depth: 12
Message warehouse

public-qt-comments@w3.org mailing list.

Data
- 5626 messages
- 2718 threads
- Maximum thread depth: 12
Message warehouse

public-qt-comments@w3.org mailing list.

Data

- 5626 messages
- 2718 threads
- Maximum thread depth: 12
Message warehouse

public-qt-comments@w3.org mailing list.

Data

- 5626 messages
- 2718 threads
- Maximum thread depth: 12
Akers warehouse

Dudouet, Manolescu, Nguyen, Senellart

XML Warehousing Meets Sociology
Simple results

Actors repartition and volume of interaction by affiliation profile

<table>
<thead>
<tr>
<th>Profile</th>
<th># actors</th>
<th># messages</th>
</tr>
</thead>
<tbody>
<tr>
<td>Companies</td>
<td>135</td>
<td>2689</td>
</tr>
<tr>
<td>Universities</td>
<td>39</td>
<td>112</td>
</tr>
<tr>
<td>Organizations</td>
<td>33</td>
<td>197</td>
</tr>
<tr>
<td>Companies &amp; Universities</td>
<td>3</td>
<td>532</td>
</tr>
<tr>
<td>Companies &amp; Organizations</td>
<td>22</td>
<td>1052</td>
</tr>
<tr>
<td>Universities &amp; Organizations</td>
<td>6</td>
<td>36</td>
</tr>
<tr>
<td>Non specified</td>
<td>65</td>
<td>681</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>303</strong></td>
<td><strong>5299</strong></td>
</tr>
</tbody>
</table>
Companies dominate XQuery standardization

Key actors tend to have multiple affiliation

Different profiles of participation in the list, even for key actors.
Sociological interpretation

- Companies **dominate** XQuery standardization
- **Key actors** tend to have **multiple affiliation**
- Different **profiles** of participation in the list, even for key actors.
Sociological interpretation

- Companies **dominate** XQuery standardization
- **Key actors** tend to have **multiple affiliation**
- Different **profiles** of participation in the list, even for key actors.
Outline

1. Introduction
2. Methodology
3. Experimentation
4. Conclusion
   - Summary
   - Perspectives
Summary

- **Interdisciplinary** approach
- Use of semi-structured technology for sociological study
- Built an XML warehouse based on XQuery public W3C information
- Preliminary analysis of the warehouse data
- Companies seem to be first in standardization process
Summary

- Interdisciplinary approach
- Use of semi-structured technology for sociological study
- Built an XML warehouse based on XQuery public W3C information
- Preliminary analysis of the warehouse data
- Companies seem to be first in standardization process
Summary

- Interdisciplinary approach
- Use of semi-structured technology for sociological study
- Built an XML warehouse based on XQuery public W3C information
- Preliminary analysis of the warehouse data
- Companies seem to be first in standardization process
Summary

- Interdisciplinary approach
- Use of semi-structured technology for sociological study
- Built an XML warehouse based on XQuery public W3C information
- Preliminary analysis of the warehouse data
- Companies seem to be first in standardization process
Summary

- **Interdisciplinary** approach
- Use of *semi-structured* technology for *sociological* study
- Built an **XML warehouse** based on XQuery public W3C information
- **Preliminary analysis** of the warehouse data
- Companies seem to be **first in standardization process**
Generic Framework for the Social Scientist

Concept definition interface

Visual XML query interface

XML visualization tool

Statistica tools

Data management tool

Sociological concepts

XML filtering and cleaning functions

XML database

XQuery processor

Web data acquisition modules

produced document
Future Work

- **Textual analysis** of message contents (e.g. agree/disagree)
- Proper management of temporal dimension
- Enriched actor warehouse with more sources (WWW in particular)
- Similar work on larger/other/private mailing lists
- More complex queries
Future Work

- **Textual analysis** of message contents (e.g. agree/disagree)
- Proper management of **temporal dimension**
- Enriched actor warehouse with more sources (WWW in particular)
- Similar work on larger/other/private mailing lists
- More complex queries
Future Work

- **Textual analysis** of message contents (e.g. agree/disagree)
- Proper management of **temporal dimension**
- **Enriched** actor warehouse with more sources (WWW in particular)
- Similar work on larger/other/private mailing lists
- More **complex** queries
Future Work

- **Textual analysis** of message contents (e.g. agree/disagree)
- Proper management of **temporal dimension**
- **Enriched** actor warehouse with more sources (WWW in particular)
- Similar work on **larger/other/private** mailing lists
- More **complex** queries
Future Work

- **Textual analysis** of message contents (e.g. agree/disagree)
- Proper management of **temporal dimension**
- **Enriched** actor warehouse with more sources (WWW in particular)
- Similar work on **larger/other/private** mailing lists
- More **complex** queries