XML Warehousing Meets Sociology

F.-X. Dudouet[1] I. Manolescu[2]
B. Nguyen[3] P. Senellart[2,4]

1 2

3 4

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?

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?

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?

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

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

Standardization

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

Standardization

Standard negociations
$\Longrightarrow$ 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

Standardization

Standard negociations
$\Longrightarrow$ 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

Standardization

Standard negociations
$\Longrightarrow$ 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

Standardization

Standard negociations

$\Longrightarrow$ 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

Standardization

Standard negociations

$\Longrightarrow$ 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

Standardization

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

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)

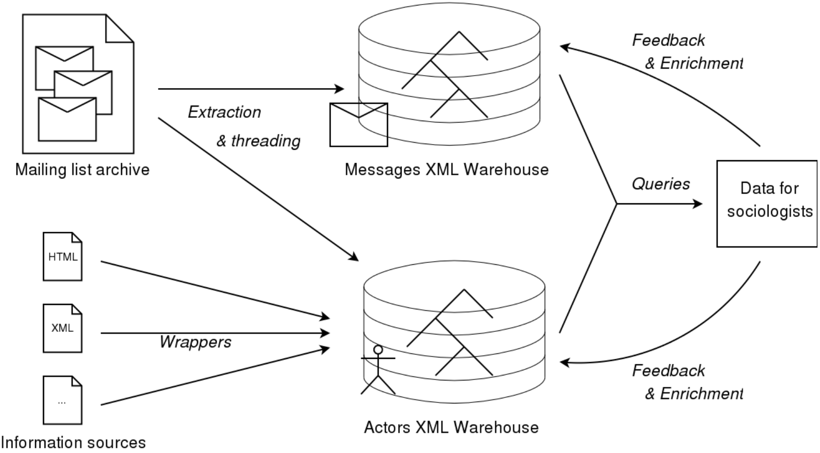
Queries enable verification of the hypotheses

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



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

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

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

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

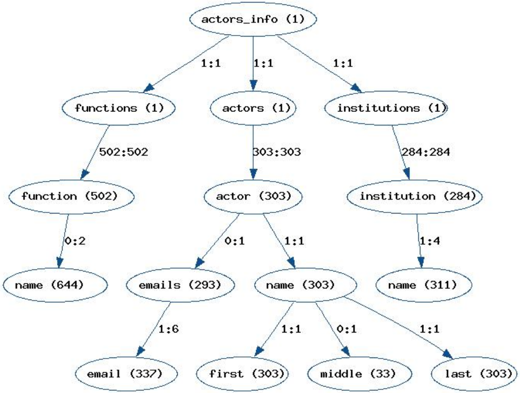
Message warehouse

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

Data

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

Actors warehouse

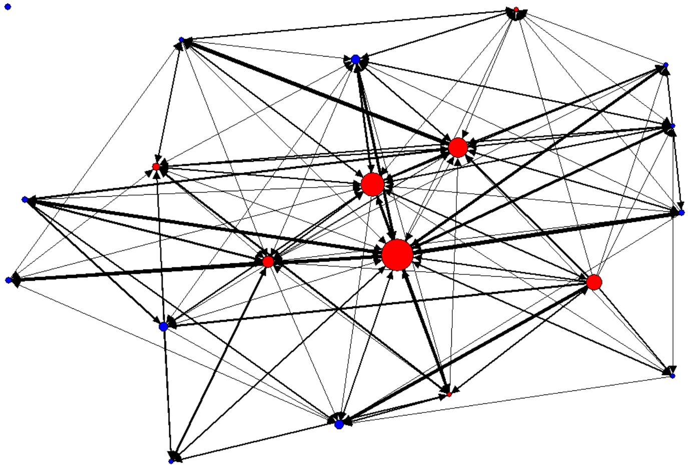


Simple results

Actor repartition and volume of interaction by affiliation profile

Profile	# actors	# messages
Companies	135	2689
Universities	39	112
Organizations	33	197
Companies & Universities	3	532
Companies & Organizations	22	1052
Universities & Organizations	6	36
Non specified	65	681
Total	**303**	**5299**

Answer network



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.

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

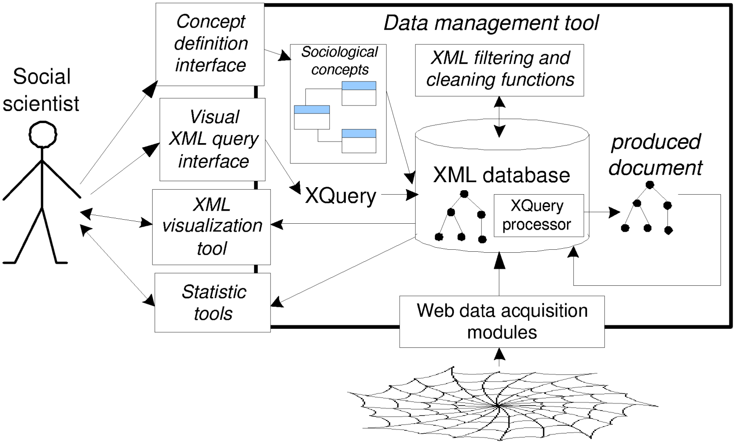
Summary

- 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



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