Leveraging Data Sources Exported as Forms and Web Services

Pierre Senellart

29 September 2006

The Hidden Web

Definition (Hidden Web, Deep Web, Invisible Web)

The part of Web content (which may or may not be dynamically generated) not accessible from the hyperlinked structure of the World Wide Web. Typically: HTML forms, Web Services.

Size estimate (2001) : 500 times larger than the surface Web.

How to understand it and benefit from its content?

The Hidden Web

Definition (Hidden Web, Deep Web, Invisible Web)

The part of Web content (which may or may not be dynamically generated) not accessible from the hyperlinked structure of the World Wide Web. Typically: HTML forms, Web Services.

Size estimate (2001) : 500 times larger than the surface Web.

How to understand it and benefit from its content?

The Hidden Web

Definition (Hidden Web, Deep Web, Invisible Web)

The part of Web content (which may or may not be dynamically generated) not accessible from the hyperlinked structure of the World Wide Web. Typically: HTML forms, Web Services.

Size estimate (2001) : 500 times larger than the surface Web.

How to understand it and benefit from its content?

Understanding the Hidden Web

Purpose

- Intensional indexing of the Hidden Web
- High-level queries
- $\Rightarrow$ a semantic search engine over the Hidden Web

In a fully, unsupervised, way!

Difficult and broad problem. Possible restriction to some domain (e.g. publications).

Understanding the Hidden Web

Purpose

- Intensional indexing of the Hidden Web
- High-level queries
- $\Rightarrow$ a semantic search engine over the Hidden Web

In a fully, unsupervised, way!

Difficult and broad problem. Possible restriction to some domain (e.g. publications).

Understanding the Hidden Web

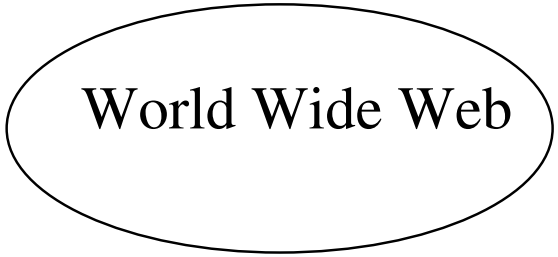
Purpose

- Intensional indexing of the Hidden Web
- High-level queries
- $\Rightarrow$ a semantic search engine over the Hidden Web

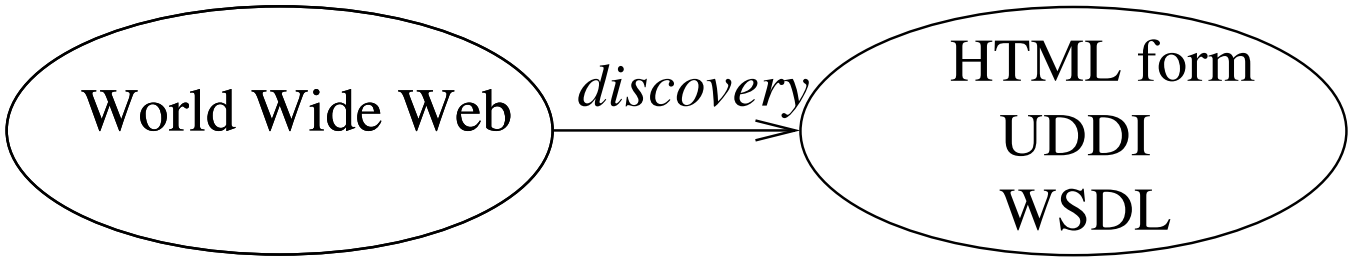
In a fully, unsupervised, way!

Difficult and broad problem. Possible restriction to some domain (e.g. publications).

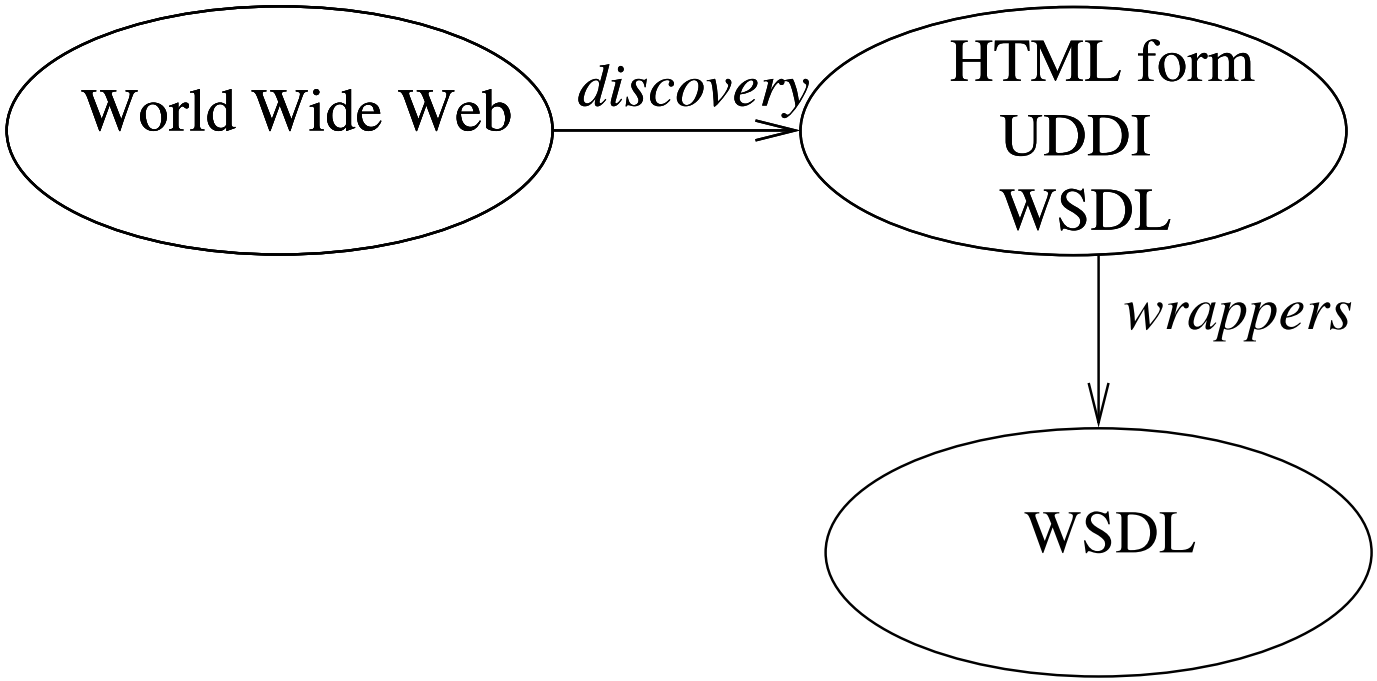
Web Service Semantic Interpretation Process



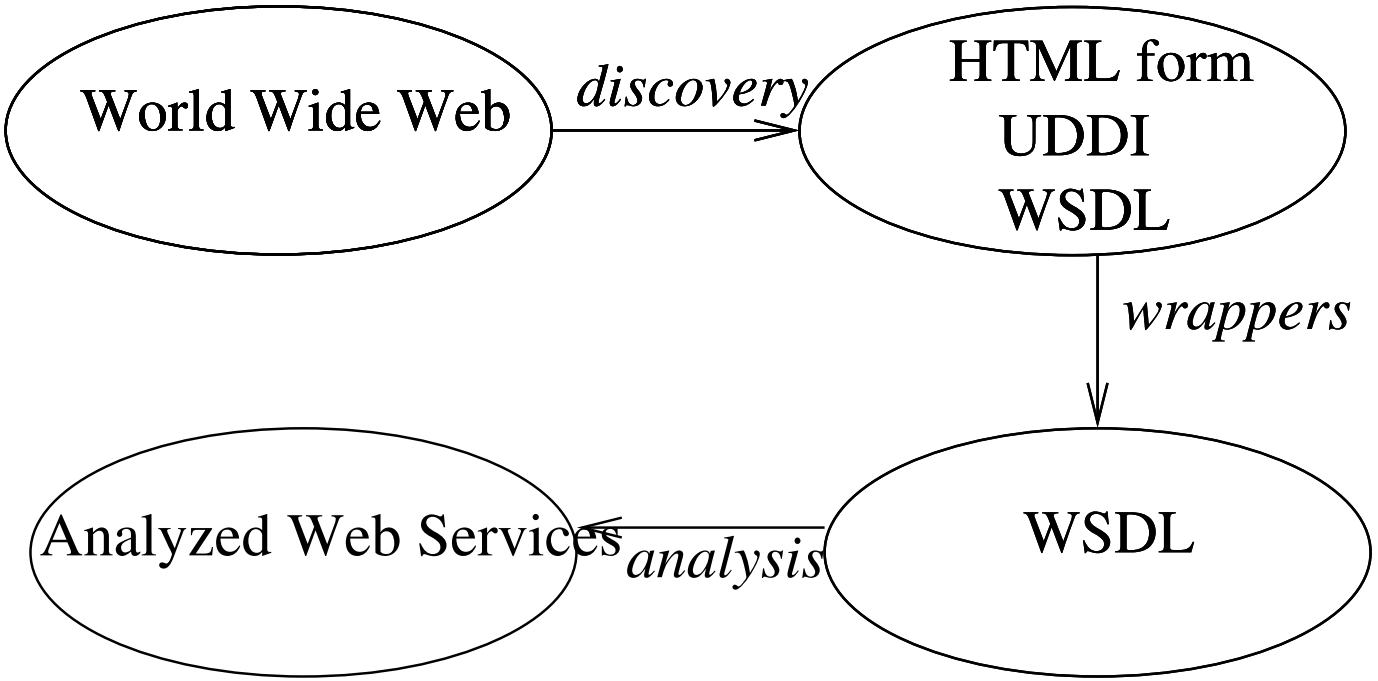
Web Service Semantic Interpretation Process



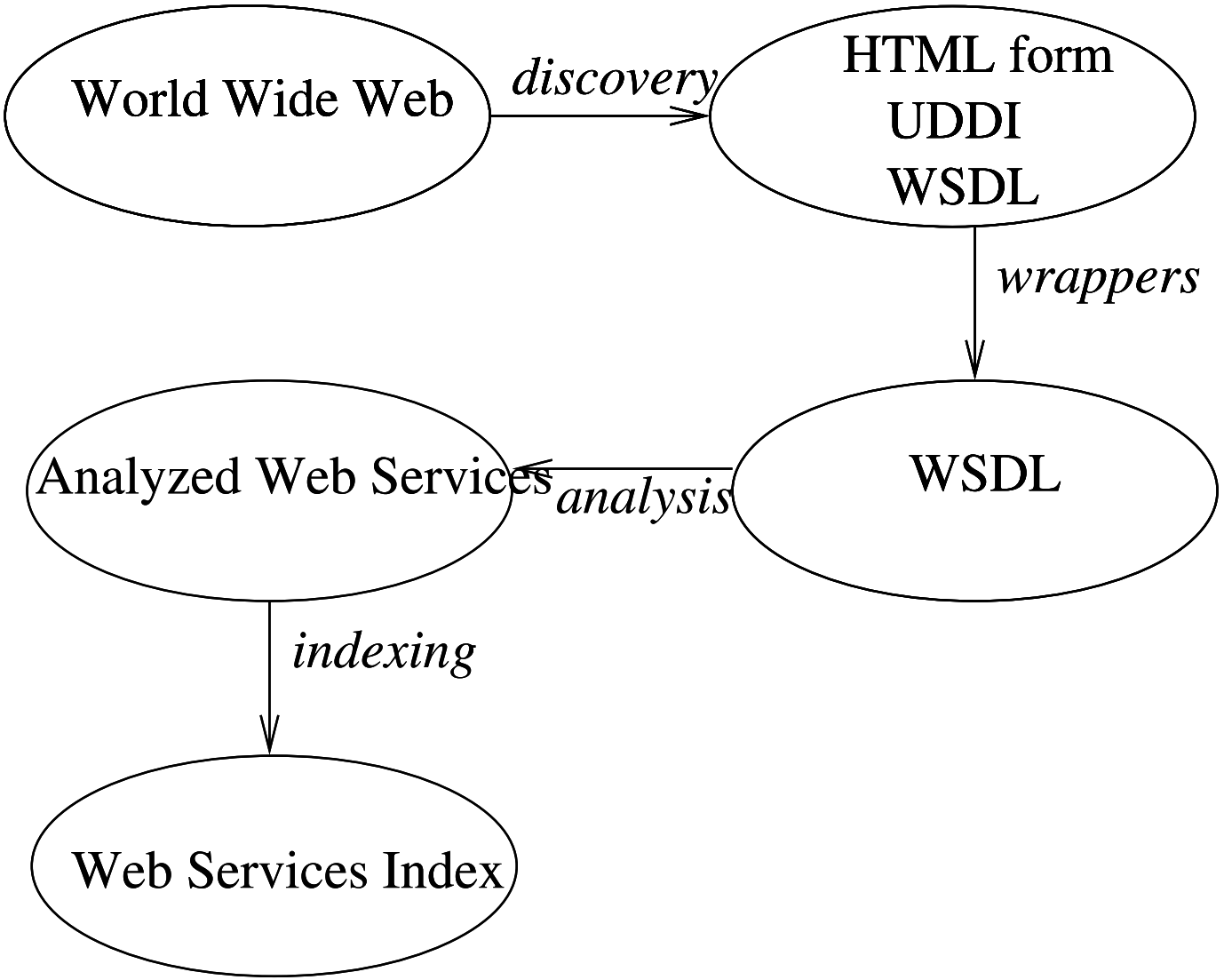
Web Service Semantic Interpretation Process



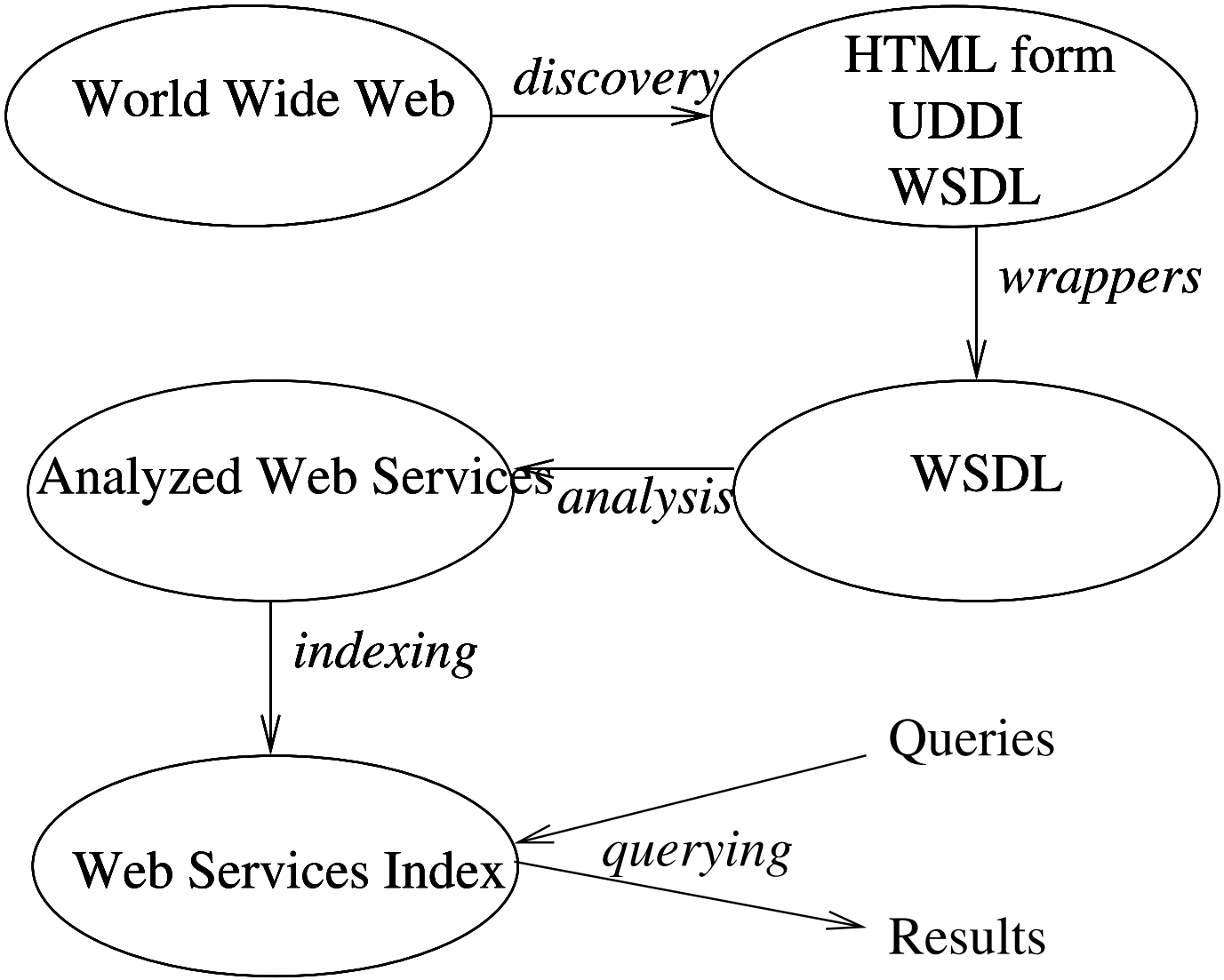
Web Service Semantic Interpretation Process



Web Service Semantic Interpretation Process



Web Service Semantic Interpretation Process



Imprecise data

- Many tasks of the process generate imprecise data, with some confidence value:
  - Information Extraction
  - Natural Language Processing
  - Data Cleaning
  - Schema Matching
  - ...
- Need for a way to manage this imprecision, to work with it throughout an entire complex process.

Imprecise data

- Many tasks of the process generate imprecise data, with some confidence value:
  - Information Extraction
  - Natural Language Processing
  - Data Cleaning
  - Schema Matching
  - ...
- Need for a way to manage this imprecision, to work with it throughout an entire complex process.

Imprecise data

- Many tasks of the process generate imprecise data, with some confidence value:
  - Information Extraction
  - Natural Language Processing
  - Data Cleaning
  - Schema Matching
  - ...
- Need for a way to manage this imprecision, to work with it throughout an entire complex process.

Imprecise data

- Many tasks of the process generate imprecise data, with some confidence value:
  - Information Extraction
  - Natural Language Processing
  - Data Cleaning
  - Schema Matching
  - ...
- Need for a way to manage this imprecision, to work with it throughout an entire complex process.

Imprecise data

- Many tasks of the process generate imprecise data, with some confidence value:
  - Information Extraction
  - Natural Language Processing
  - Data Cleaning
  - Schema Matching
  - ...
- Need for a way to manage this imprecision, to work with it throughout an entire complex process.

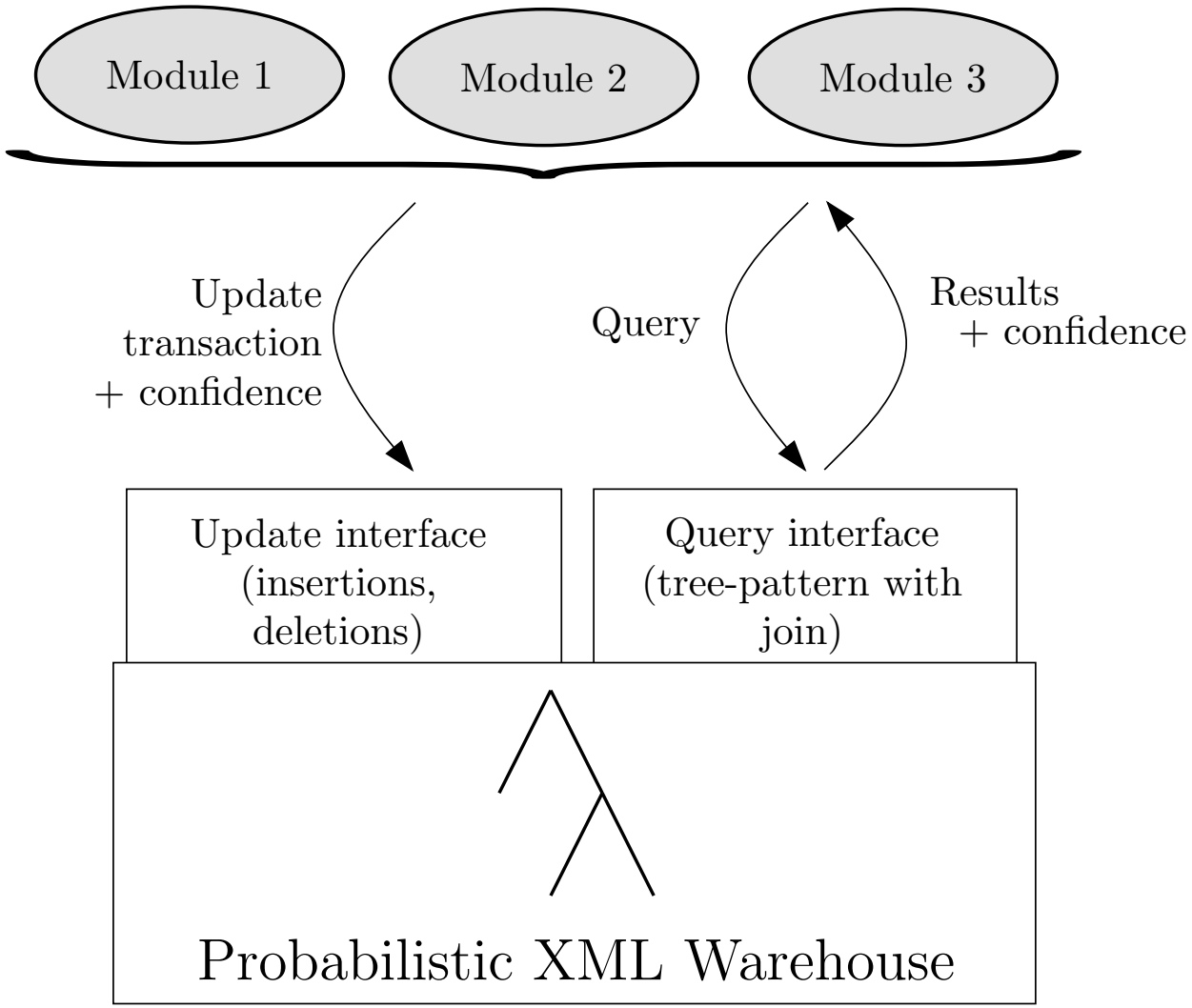
Imprecise data

- Many tasks of the process generate imprecise data, with some confidence value:
  - Information Extraction
  - Natural Language Processing
  - Data Cleaning
  - Schema Matching
  - ...
- Need for a way to manage this imprecision, to work with it throughout an entire complex process.

Imprecise data

- Many tasks of the process generate imprecise data, with some confidence value:
  - Information Extraction
  - Natural Language Processing
  - Data Cleaning
  - Schema Matching
  - ...
- Need for a way to manage this imprecision, to work with it throughout an entire complex process.

A Probabilistic XML Warehouse



Outline

1 Introduction

2 Process description
- Web Service Discovery
- Wrapping Web Service Descriptions
- Web Service Semantic Analysis
- Web Service Indexing and Querying

3 Conclusion

Web Service Discovery

Crawling the World Wide Web for:

- HTML forms implementing a Web Service
- UDDI registries
- WSDL descriptions
- Other resources (XML, HTML, Web as a full-text index. . .)

Usually, only interested in a specific domain

$\Rightarrow$ focused crawling?

Web Service Discovery

Crawling the World Wide Web for:

- HTML forms implementing a Web Service
- UDDI registries
- WSDL descriptions
- Other resources (XML, HTML, Web as a full-text index...)

Usually, only interested in a specific domain

$\Rightarrow$ focused crawling?

Web Service Filtering

Only interested in Web Services with no side effects:

OK	Not OK
• Yellow Pages	• Booking services
• Publication databases	• Mailing list management
• ...	• ...

Filtering using a number of heuristics:

- No credit card number, password, email address (?)
- HTTP GET should be side-effect free
- ...

Web Service Filtering

Only interested in Web Services with no side effects:

OK	Not OK
• Yellow Pages	• Booking services
• Publication databases	• Mailing list management
• ...	• ...

Filtering using a number of heuristics:

- No credit card number, password, email address (?)
- HTTP GET should be side-effect free
- ...

Web Service Filtering

Only interested in Web Services with no side effects:

OK	Not OK
• Yellow Pages	• Booking services
• Publication databases	• Mailing list management
• ...	• ...

Filtering using a number of heuristics:

- No credit card number, password, email address (?)
- HTTP GET should be side-effect free
- ...

Web Service Filtering

Only interested in Web Services with no side effects:

OK	Not OK
• Yellow Pages	• Booking services
• Publication databases	• Mailing list management
• ...	• ...

Filtering using a number of heuristics:

- No credit card number, password, email address (?)
- HTTP GET should be side-effect free
- ...

Web Service Filtering

Only interested in Web Services with no side effects:

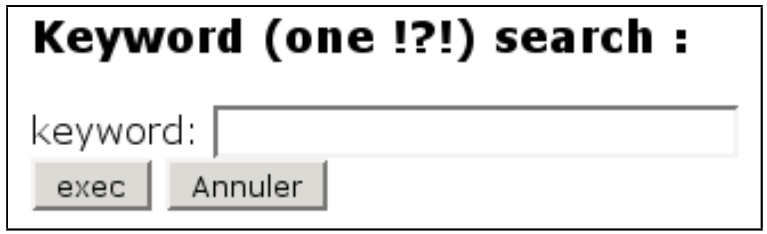
OK	Not OK
• Yellow Pages	• Booking services
• Publication databases	• Mailing list management
• ...	• ...

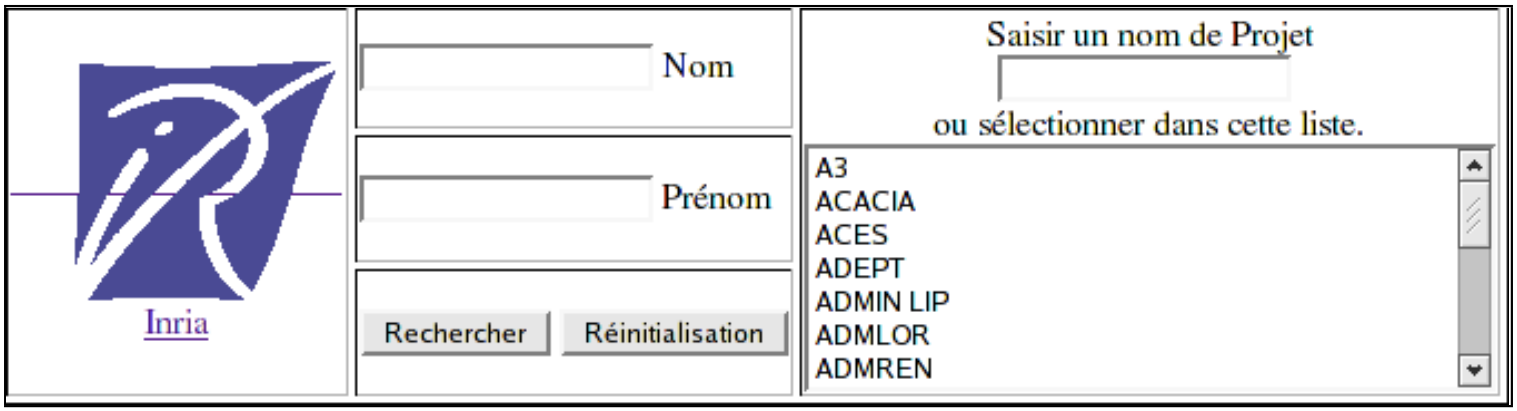
Filtering using a number of heuristics:

- No credit card number, password, email address (?)
- HTTP GET should be side-effect free
- ...

Analyzing HTML forms

Analyzing the structure of HTML forms.



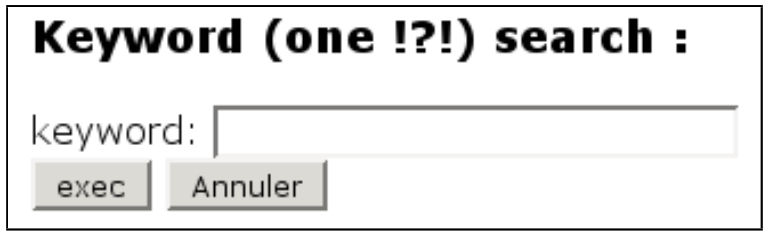


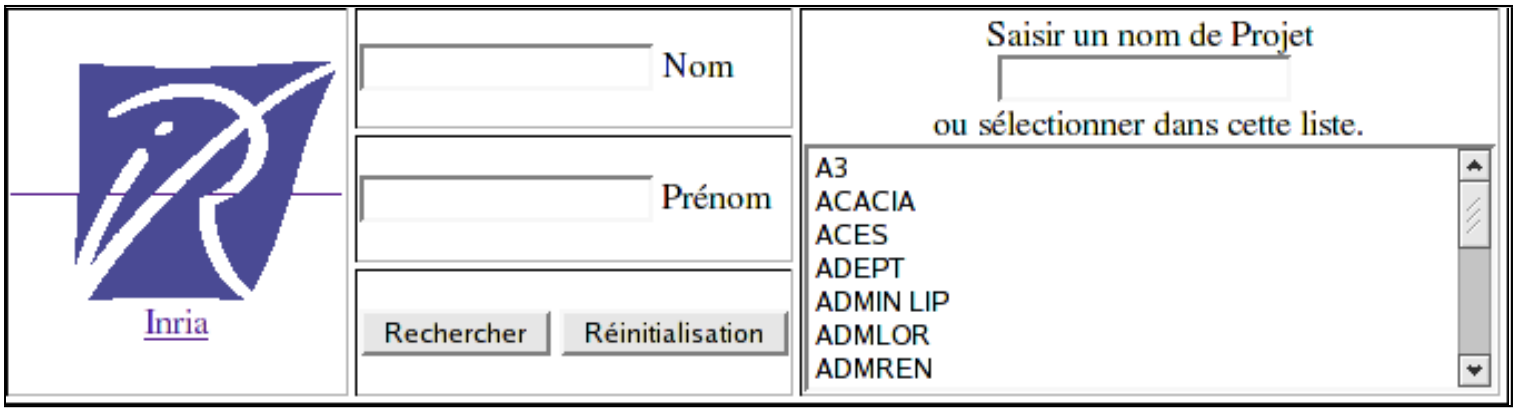
Issues

- What are the relevant form fields?
- What is the concrete type of each field?
- What is the label of each field?

Analyzing HTML forms

Analyzing the structure of HTML forms.



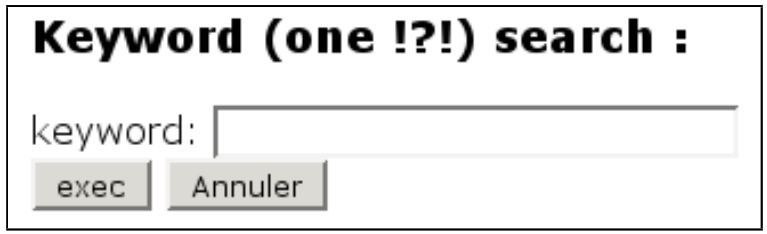


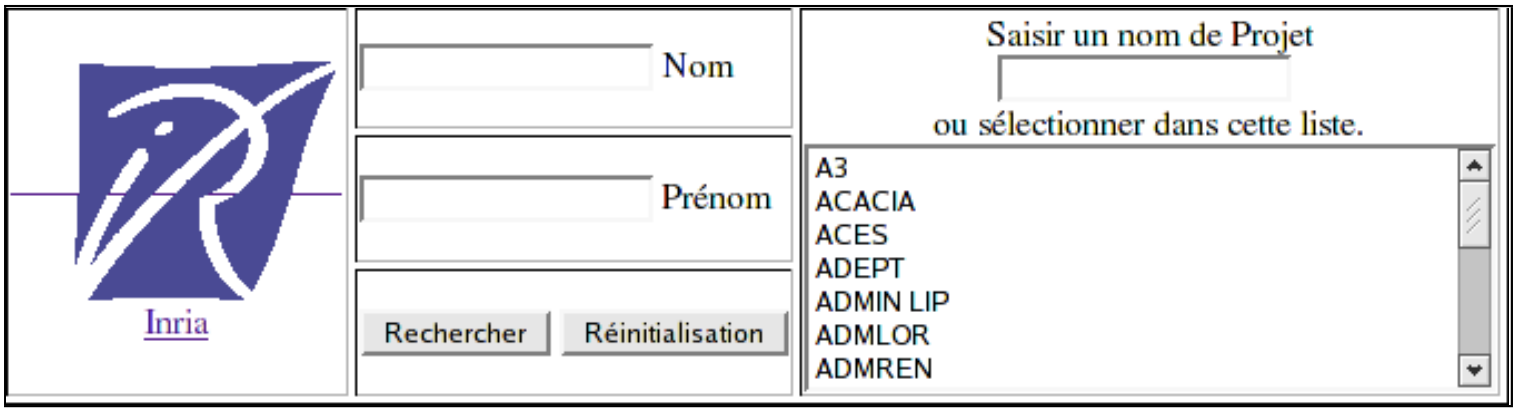
Issues

- What are the relevant form fields?
- What is the concrete type of each field?
- What is the label of each field?

Analyzing HTML forms

Analyzing the structure of HTML forms.





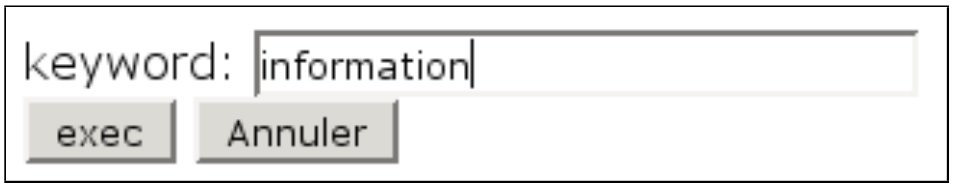
Issues

- What are the relevant form fields?
- What is the concrete type of each field?
- What is the label of each field?

Probing HTML forms

Probing HTML forms to retrieve sample HTML answer pages:

- With dictionary words



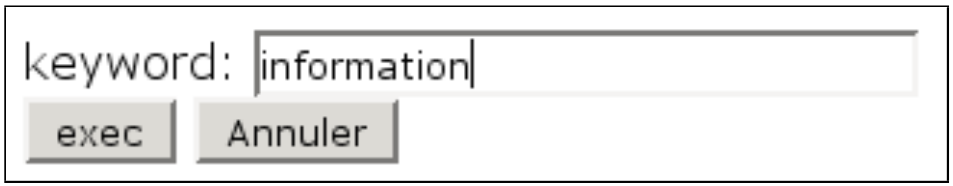
- With nonsense words

- With domain words

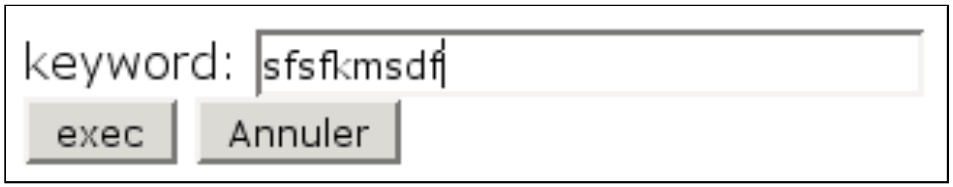
Probing HTML forms

Probing HTML forms to retrieve sample HTML answer pages:

- With dictionary words



- With nonsense words

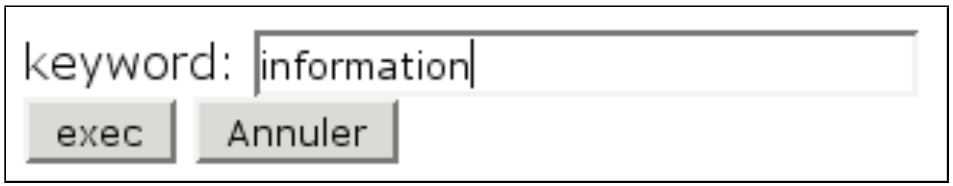


- With domain words

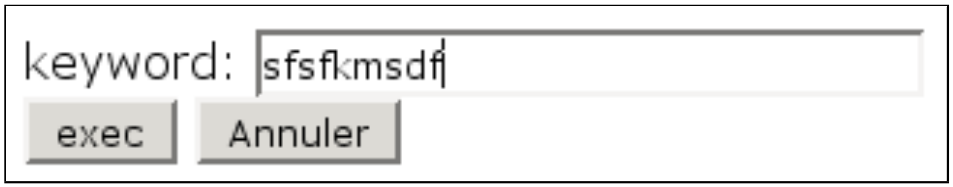
Probing HTML forms

Probing HTML forms to retrieve sample HTML answer pages:

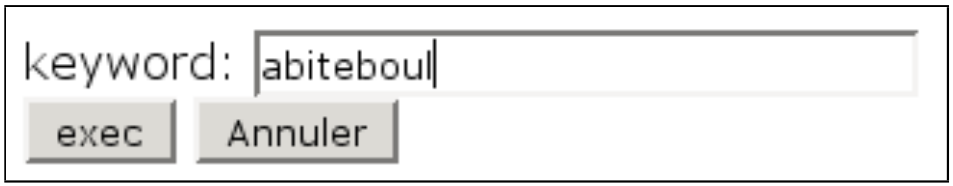
- With dictionary words



- With nonsense words



- With domain words



Query-answer Web pages

Extract data from query-answer Web pages.

Showing results 1 through 25 (of 94 total) for all:xml

1. cs.LO/0601085 [abs, ps, pdf, other] :
Title: **A Formal Foundation for ODRL**
Authors: Riccardo Pucella, Vicky Weissman
Comments: 30 pgs, preliminary version presented at WITS-04 (Workshop on Issues in the Theory of Security), 2004
Subj-class: Logic in Computer Science; Cryptography and Security
ACM-class: H.2.7; K.4.4

2. astro-ph/0512493 [abs, pdf] :
Title: **VOFilter, Bridging Virtual Observatory and Industrial Office Applications**
Authors: Chen-zhou Cui (1), Markus Dolensky (2), Peter Quinn (2), Yong-heng Zhao (1), Francoise Genova (3) ((1)NAO China, (2) ESO, (3) CDS)
Comments: Accepted for publication in ChJAA (9 pages, 2 figures, 185KB)

3. cs.DS/0512061 [abs, ps, pdf, other] :
Title: **Matching Subsequences in Trees**
Authors: Philip Bille, Inge Li Goertz
Subj-class: Data Structures and Algorithms

4. cs.IR/0510025 [abs, ps, pdf, other] :
Title: **Practical Semantic Analysis of Web Sites and Documents**
Authors: Thierry Despeyroux (INRIA Rocquencourt / INRIA Sophia Antipolis)
Subj-class: Information Retrieval

5. cs.CR/0510013 [abs, pdf] :
Title: **Safe Data Sharing and Data Dissemination on Smart Devices**
Authors: Luc Bouganim (INRIA Rocquencourt), Cosmin Cremarenco (INRIA Rocquencourt), François Dang Ngoc (INRIA Rocquencourt, PRISM - UVSQ), Nicolas Dieu (INRIA Rocquencourt), Philippe Pucheral (INRIA Rocquencourt, PRISM - UVSQ)
Subj-class: Cryptography and Security; Databases

Issues

- What part of the Web page contains the answer?
- How to extract structured content?
- How to label this structured content?

Query-answer Web pages

Extract data from query-answer Web pages.

Showing results 1 through 25 (of 94 total) for all:xml

1. cs.LO/0601085 [abs, ps, pdf, other] :
Title: **A Formal Foundation for ODRL**
Authors: Riccardo Pucella, Vicky Weissman
Comments: 30 pgs, preliminary version presented at WITS-04 (Workshop on Issues in the Theory of Security), 2004
Subj-class: Logic in Computer Science; Cryptography and Security
ACM-class: H.2.7; K.4.4

2. astro-ph/0512493 [abs, pdf] :
Title: **VOFilter, Bridging Virtual Observatory and Industrial Office Applications**
Authors: Chen-zhou Cui (1), Markus Dolensky (2), Peter Quinn (2), Yong-heng Zhao (1), Francoise Genova (3) ((1)NAO China, (2) ESO, (3) CDS)
Comments: Accepted for publication in ChJAA (9 pages, 2 figures, 185KB)

3. cs.DS/0512061 [abs, ps, pdf, other] :
Title: **Matching Subsequences in Trees**
Authors: Philip Bille, Inge Li Goertz
Subj-class: Data Structures and Algorithms

4. cs.IR/0510025 [abs, ps, pdf, other] :
Title: **Practical Semantic Analysis of Web Sites and Documents**
Authors: Thierry Despeyroux (INRIA Rocquencourt / INRIA Sophia Antipolis)
Subj-class: Information Retrieval

5. cs.CR/0510013 [abs, pdf] :
Title: **Safe Data Sharing and Data Dissemination on Smart Devices**
Authors: Luc Bouganim (INRIA Rocquencourt), Cosmin Cremarenco (INRIA Rocquencourt), François Dang Ngoc (INRIA Rocquencourt, PRISM - UVSQ), Nicolas Dieu (INRIA Rocquencourt), Philippe Pucheral (INRIA Rocquencourt, PRISM - UVSQ)
Subj-class: Cryptography and Security; Databases

Issues

- What part of the Web page contains the answer?
- How to extract structured content?
- How to label this structured content?

Query-answer Web pages

Extract data from query-answer Web pages.

Showing results 1 through 25 (of 94 total) for all:xml

1. cs.LO/0601085 [abs, ps, pdf, other] :
Title: **A Formal Foundation for ODRL**
Authors: Riccardo Pucella, Vicky Weissman
Comments: 30 pgs, preliminary version presented at WITS-04 (Workshop on Issues in the Theory of Security), 2004
Subj-class: Logic in Computer Science; Cryptography and Security
ACM-class: H.2.7; K.4.4

2. astro-ph/0512493 [abs, pdf] :
Title: **VOFilter, Bridging Virtual Observatory and Industrial Office Applications**
Authors: Chen-zhou Cui (1), Markus Dolensky (2), Peter Quinn (2), Yong-heng Zhao (1), Francoise Genova (3) ((1)NAO China, (2) ESO, (3) CDS)
Comments: Accepted for publication in ChJAA (9 pages, 2 figures, 185KB)

3. cs.DS/0512061 [abs, ps, pdf, other] :
Title: **Matching Subsequences in Trees**
Authors: Philip Bille, Inge Li Goertz
Subj-class: Data Structures and Algorithms

4. cs.IR/0510025 [abs, ps, pdf, other] :
Title: **Practical Semantic Analysis of Web Sites and Documents**
Authors: Thierry Despeyroux (INRIA Rocquencourt / INRIA Sophia Antipolis)
Subj-class: Information Retrieval

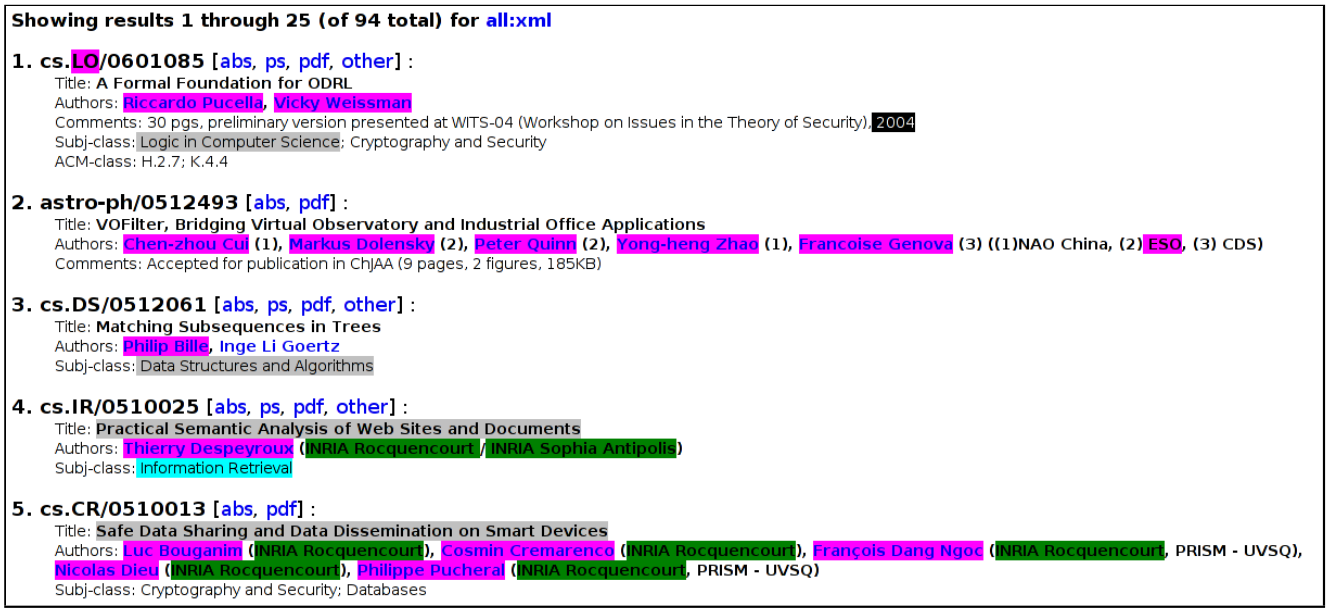
5. cs.CR/0510013 [abs, pdf] :
Title: **Safe Data Sharing and Data Dissemination on Smart Devices**
Authors: Luc Bouganim (INRIA Rocquencourt), Cosmin Cremarenco (INRIA Rocquencourt), François Dang Ngoc (INRIA Rocquencourt, PRISM - UVSQ), Nicolas Dieu (INRIA Rocquencourt), Philippe Pucheral (INRIA Rocquencourt, PRISM - UVSQ)
Subj-class: Cryptography and Security; Databases

Issues

- What part of the Web page contains the answer?
- How to extract structured content?
- How to label this structured content?

Automatic wrapper induction using domain knowledge

- Annotate pages with knowledge domain (finite automata techniques)



Showing results 1 through 25 (of 94 total) for all:xml

1. cs.LO/0601085 [abs, ps, pdf, other] :
Title: **A Formal Foundation for ODRL**
Authors: Riccardo Pucella, Vicky Weissman
Comments: 30 pgs, preliminary version presented at WITS-04 (Workshop on Issues in the Theory of Security), 2004
Subj-class: Logic in Computer Science; Cryptography and Security
ACM-class: H.2.7; K.4.4

2. astro-ph/0512493 [abs, pdf] :
Title: **VOFilter, Bridging Virtual Observatory and Industrial Office Applications**
Authors: Chen-zhou Cui (1), Markus Dolensky (2), Peter Quinn (2), Yong-heng Zhao (1), Francoise Genova (3) ((1)NAO China, (2) ESO, (3) CDS)
Comments: Accepted for publication in ChJAA (9 pages, 2 figures, 185KB)

3. cs.DS/0512061 [abs, ps, pdf, other] :
Title: **Matching Subsequences in Trees**
Authors: Philip Bille, Inge Li Goertz
Subj-class: Data Structures and Algorithms

4. cs.IR/0510025 [abs, ps, pdf, other] :
Title: **Practical Semantic Analysis of Web Sites and Documents**
Authors: Thierry Despeyroux (INRIA Rocquencourt / INRIA Sophia Antipolis)
Subj-class: Information Retrieval

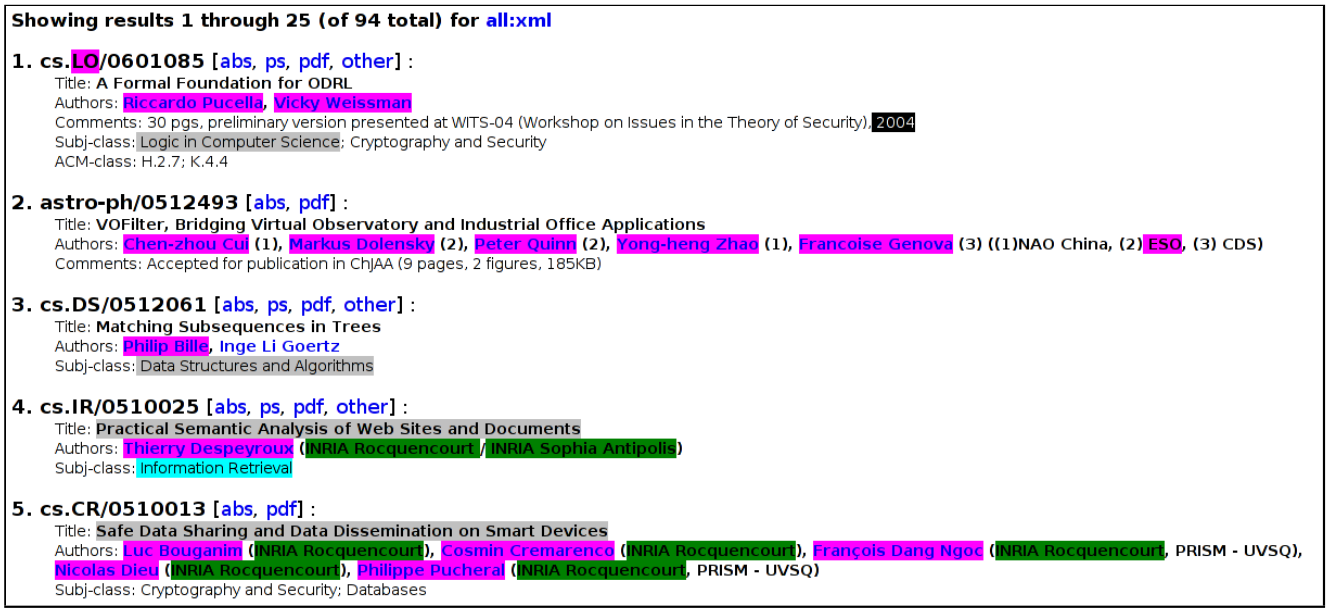
5. cs.CR/0510013 [abs, pdf] :
Title: **Safe Data Sharing and Data Dissemination on Smart Devices**
Authors: Luc Bouganim (INRIA Rocquencourt), Cosmin Cremarenco (INRIA Rocquencourt), François Dang Ngoc (INRIA Rocquencourt, PRISM - UVSQ), Nicolas Dieu (INRIA Rocquencourt), Philippe Pucheral (INRIA Rocquencourt, PRISM - UVSQ)
Subj-class: Cryptography and Security; Databases

(Both imperfect and incomplete)

- Use machine learning to generalize the result into a structural extraction wrapper (work in progress with MOSTRARE).

Automatic wrapper induction using domain knowledge

- Annotate pages with knowledge domain (finite automata techniques)



Showing results 1 through 25 (of 94 total) for all:xml

1. cs.LO/0601085 [abs, ps, pdf, other] :
Title: **A Formal Foundation for ODRL**
Authors: Riccardo Pucella, Vicky Weissman
Comments: 30 pgs, preliminary version presented at WITS-04 (Workshop on Issues in the Theory of Security), 2004
Subj-class: Logic in Computer Science; Cryptography and Security
ACM-class: H.2.7; K.4.4

2. astro-ph/0512493 [abs, pdf] :
Title: **VOFilter, Bridging Virtual Observatory and Industrial Office Applications**
Authors: Chen-zhou Cui (1), Markus Dolensky (2), Peter Quinn (2), Yong-heng Zhao (1), Francoise Genova (3) ((1)NAO China, (2) ESO, (3) CDS)
Comments: Accepted for publication in ChJAA (9 pages, 2 figures, 185KB)

3. cs.DS/0512061 [abs, ps, pdf, other] :
Title: **Matching Subsequences in Trees**
Authors: Philip Bille, Inge Li Goertz
Subj-class: Data Structures and Algorithms

4. cs.IR/0510025 [abs, ps, pdf, other] :
Title: **Practical Semantic Analysis of Web Sites and Documents**
Authors: Thierry Despeyroux (INRIA Rocquencourt / INRIA Sophia Antipolis)
Subj-class: Information Retrieval

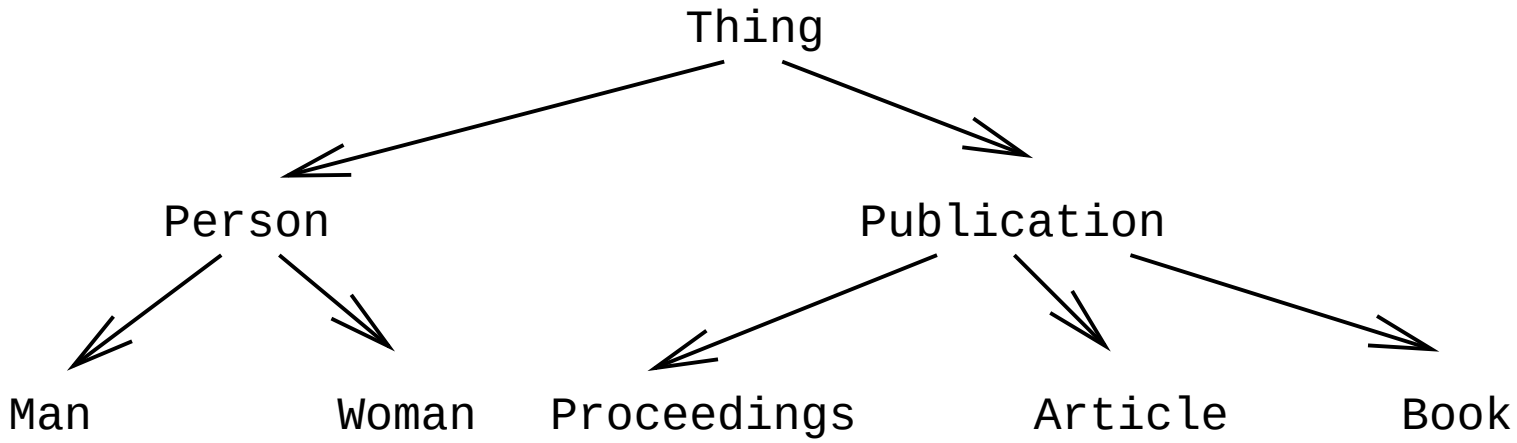
5. cs.CR/0510013 [abs, pdf] :
Title: **Safe Data Sharing and Data Dissemination on Smart Devices**
Authors: Luc Bouganim (INRIA Rocquencourt), Cosmin Cremarenco (INRIA Rocquencourt), François Dang Ngoc (INRIA Rocquencourt, PRISM - UVSQ), Nicolas Dieu (INRIA Rocquencourt), Philippe Pucheral (INRIA Rocquencourt, PRISM - UVSQ)
Subj-class: Cryptography and Security; Databases

(Both imperfect and incomplete)

- Use machine learning to generalize the result into a structural extraction wrapper (work in progress with MOSTRARE).

Conceptual Model

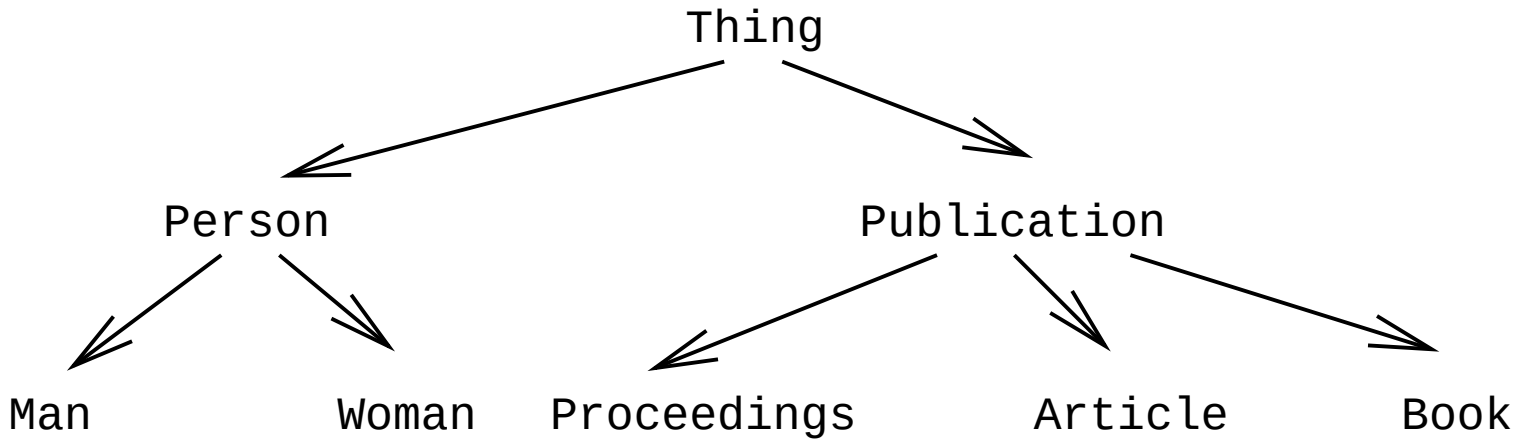
- IsA ontology of concepts (simple DAG)



- *n*-ary typed roles
  - `AuthorOf(Person,Publication)`
  - `HasName(Person,Name)`

Conceptual Model

- IsA ontology of concepts (simple DAG)



- *n*-ary typed roles
  - `AuthorOf(Person,Publication)`
  - `HasName(Person,Name)`

Semantic representation of a service

What is a service described by?

- A *n*-uple of typed input parameters
- A complex (= nested) type of its output
- Semantic relations between inputs and outputs

Definition (Complex types)

S: set of concepts

$$T \longleftarrow S | < T, \ldots, T > | T*$$

Semantic representation of a service

What is a service described by?

- A *n*-uple of typed input parameters
- A complex (= nested) type of its output
- Semantic relations between inputs and outputs

Definition (Complex types)

S: set of concepts

$$T \longleftarrow S | < T, \ldots, T > | T*$$

Semantic representation of a service

What is a service described by?

- A n-uple of typed input parameters
- A complex (= nested) type of its output
- Semantic relations between inputs and outputs

Definition (Complex types)

S: set of concepts

$$T \longleftarrow S | < T, \ldots, T > | T*$$

Semantic representation of a service

What is a service described by?

- A n-uple of typed input parameters
- A complex (= nested) type of its output
- Semantic relations between inputs and outputs

Definition (Complex types)

S: set of concepts

$$T \longleftarrow S|<T, \ldots, T>|T*$$

Services and queries

Example

Service giving authors from publication titles

$$\mathrm{A}^{*} \leftarrow \texttt{WrittenBy}(\mathrm{P},\mathrm{A}),\texttt{HasTitle}(\mathrm{P},\mathrm{T}),\mathrm{Input}(\mathrm{T})$$

Example

Query:

```
<A,T*>*  ←  WrittenBy(P,A), Article(P),
            HasTitle(P,T), KeywordOf("xml",P)
```

Services and queries

Example

Service giving authors from publication titles

$$\mathrm{A}^{*} \leftarrow \texttt{WrittenBy}(\mathrm{P},\mathrm{A}),\texttt{HasTitle}(\mathrm{P},\mathrm{T}),\mathrm{Input}(\mathrm{T})$$

Example

Query:

<A,T*>* ← WrittenBy(P,A), Article(P),
HasTitle(P,T), KeywordOf("xml",P)

Managing extensional information

How to represent extensional information (i.e. documents) in this formalism?

Definition

A document is a service with no input.

Complex types: natural representation of a DTD.

(Disjunctions `a|b` simulated by `(a?,b?)`).

Semantic Interpretation of a Service

How to analyze a Web Service?

- Field labels, variable names, tag names
- Concrete type descriptions
  (e.g. `\d{4}-\d{2}-\d{2}` is a date)
- Linguistic analysis of plain text descriptions and pages linking to the service

Note: Extracting concepts is easier than extracting relations between concepts.

Semantic Interpretation of a Service

How to analyze a Web Service?

- Field labels, variable names, tag names
- Concrete type descriptions
  (e.g. `\d{4}-\d{2}-\d{2}` is a date)
- Linguistic analysis of plain text descriptions and pages linking to the service

Note: Extracting concepts is easier than extracting relations between concepts.

Semantic Interpretation of a Service

How to analyze a Web Service?

- Field labels, variable names, tag names
- Concrete type descriptions
  (e.g. `\d{4}-\d{2}-\d{2}` is a date)
- Linguistic analysis of plain text descriptions and pages linking to the service

Note: Extracting concepts is easier than extracting relations between concepts.

Semantic Interpretation of a Service

How to analyze a Web Service?

- Field labels, variable names, tag names
- Concrete type descriptions
  (e.g. `\d{4}-\d{2}-\d{2}` is a date)
- Linguistic analysis of plain text descriptions and pages linking to the service

Note: Extracting concepts is easier than extracting relations between concepts.

Web Service Indexing and Querying

Given a query, represented as an Analyzed Web Service, how to know which known web services to query?

Issues

- Subsumption of input/output parameters
- Missing input parameters
- Composition of Web Services

Web Service Indexing and Querying

Given a query, represented as an Analyzed Web Service, how to know which known web services to query?

Issues

- Subsumption of input/output parameters
- Missing input parameters
- Composition of Web Services

Web Service Indexing and Querying

Given a query, represented as an Analyzed Web Service, how to know which known web services to query?

Issues

- Subsumption of input/output parameters
- Missing input parameters
- Composition of Web Services

Web Service Indexing and Querying

Given a query, represented as an Analyzed Web Service, how to know which known web services to query?

Issues

- Subsumption of input/output parameters
- Missing input parameters
- Composition of Web Services

Differences with classical database querying

Three main differences:

- Information can be queried only through *views* (Local As View)
- Nested types
- Incomplete information

Three sources of complexity!

Differences with classical database querying

Three main differences:

- Information can be queried only through *views* (Local As View)
- Nested types
- Incomplete information

Three sources of complexity!

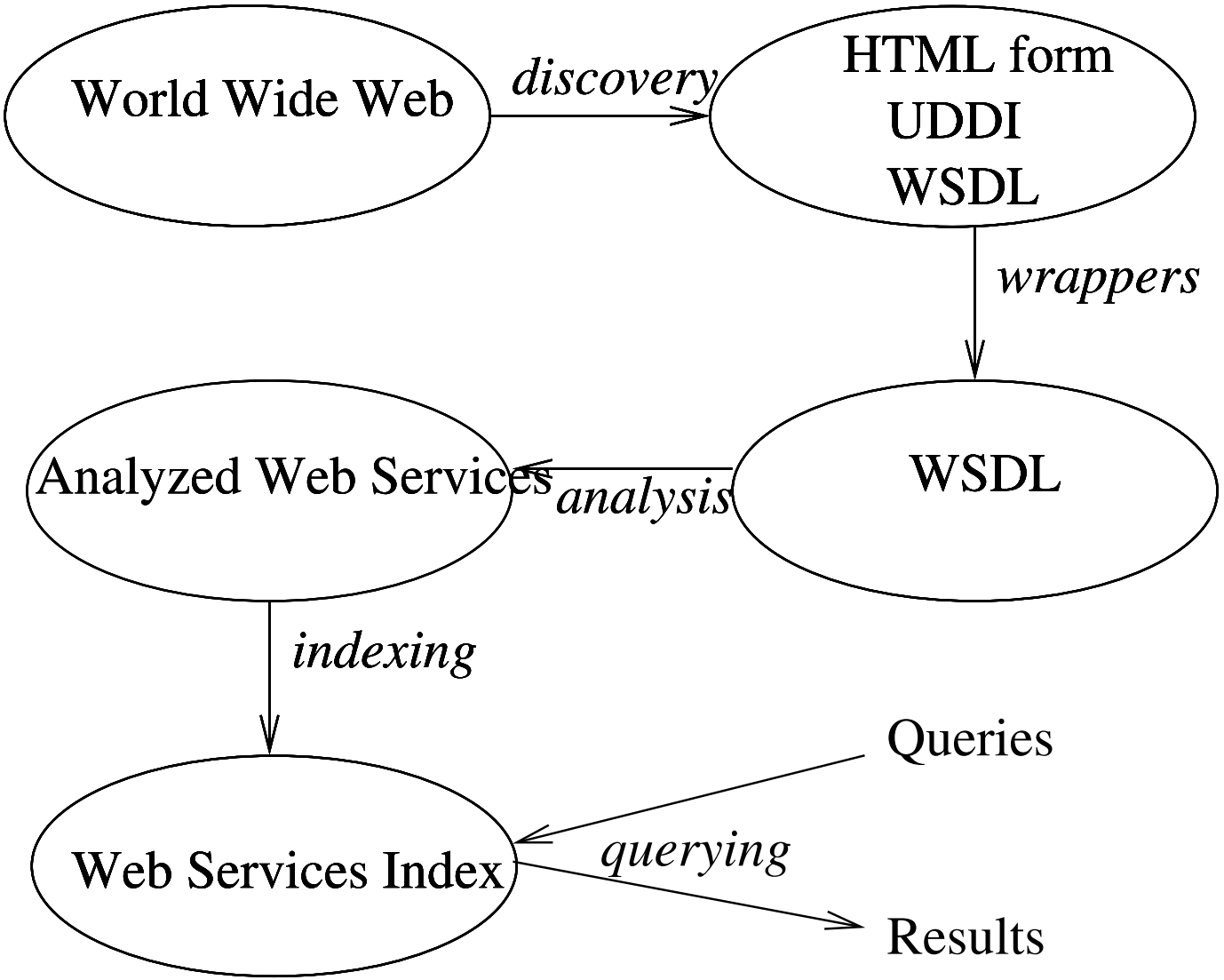
Outline

1 Introduction

2 Process description

3 Conclusion

Web Service Semantic Interpretation Process



Perspectives

Still a lot to do... In particular:

- Answering queries using views on the semantic model.
- Continue work on automatic wrapper induction.
- Form probing.
- System implementation of the whole process on a specific domain.