Understanding the Hidden Web

Pierre Senellart

17 June 2005

The Hidden Web

Definition (Hidden Web)

The set of webpages (which may or may not be dynamically generated) not accessible from the hyperlinked structure of the World Wide Web.

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)

The set of webpages (which may or may not be dynamically generated) not accessible from the hyperlinked structure of the World Wide Web.

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)

The set of webpages (which may or may not be dynamically generated) not accessible from the hyperlinked structure of the World Wide Web.

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, unsupervized, 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, unsupervized, way!

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

Understanding the Hidden Web

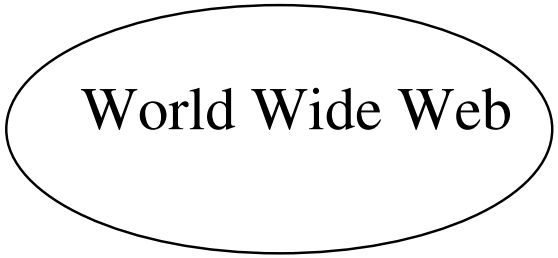
Purpose

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

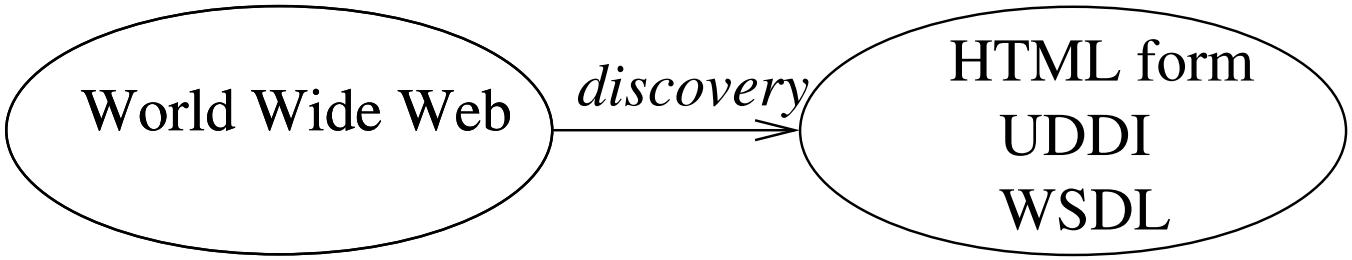
In a fully, unsupervized, way!

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

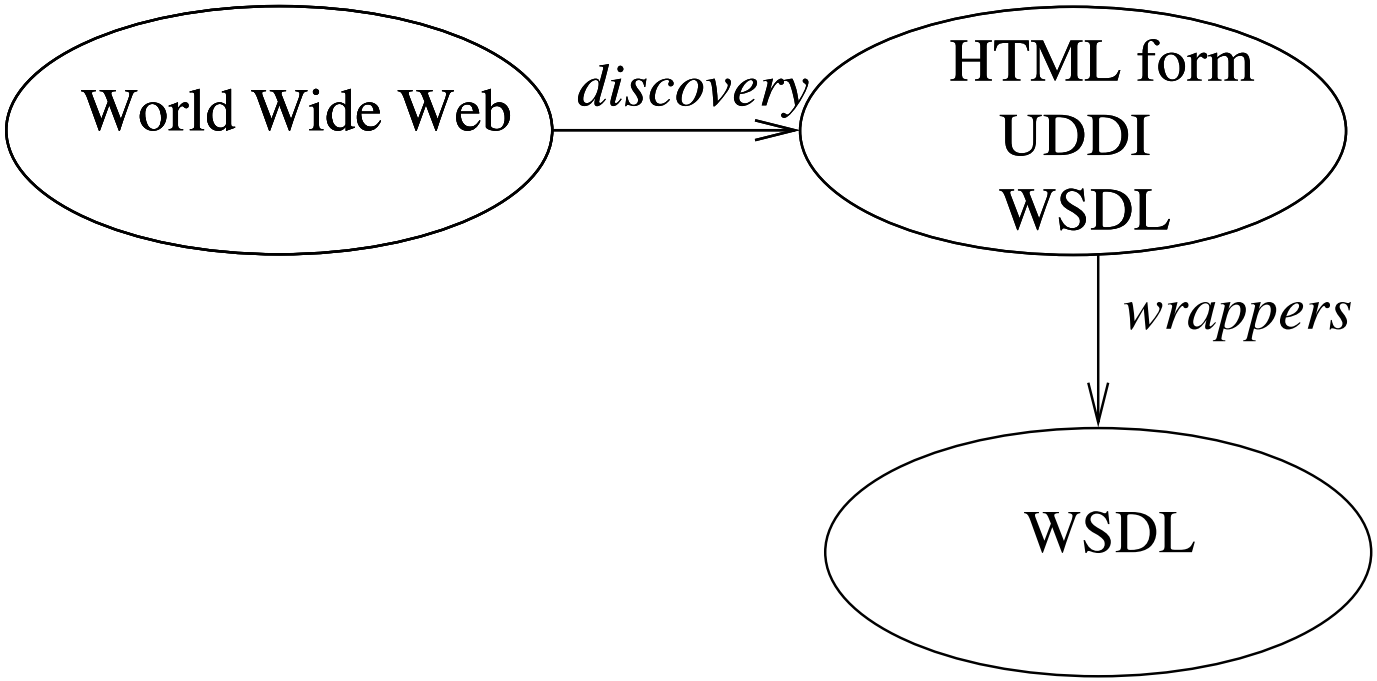
Web Service Semantic Interpretation Process



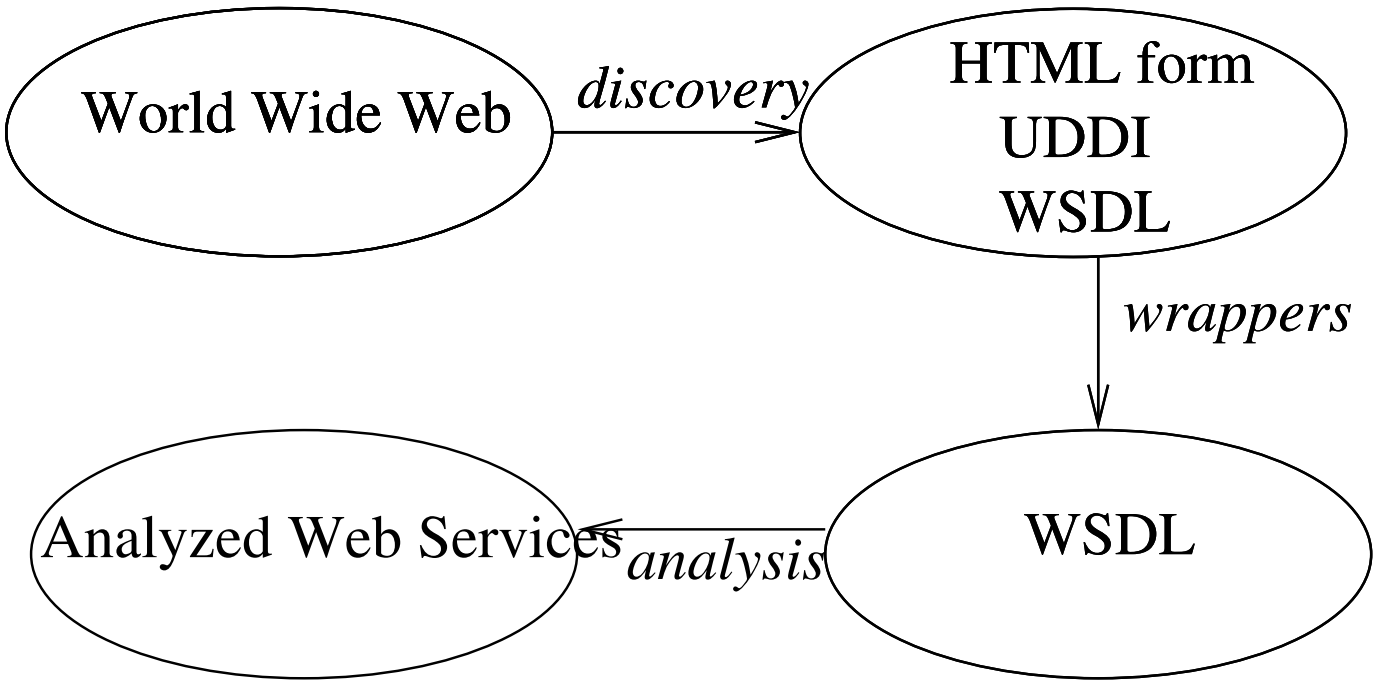
Web Service Semantic Interpretation Process



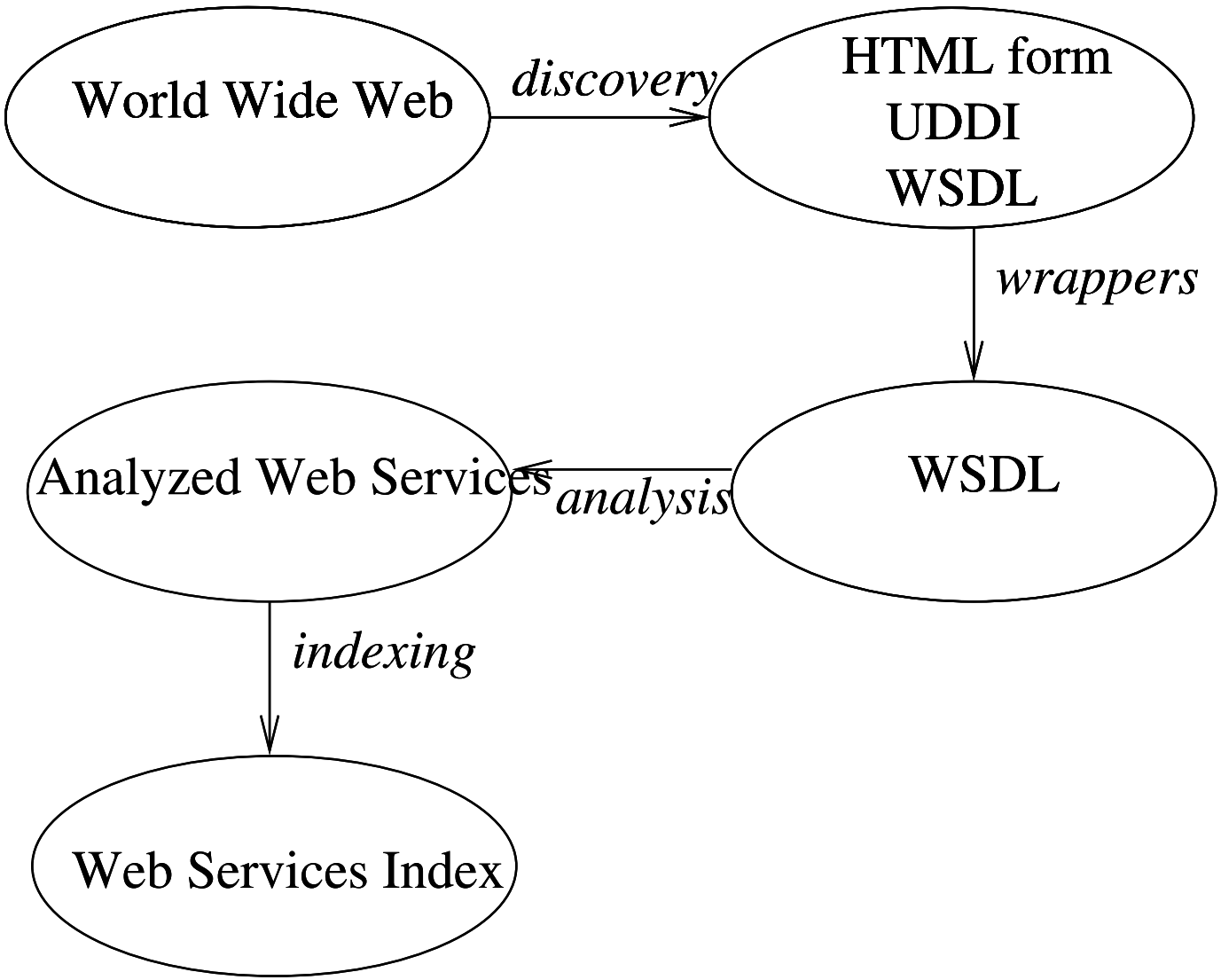
Web Service Semantic Interpretation Process



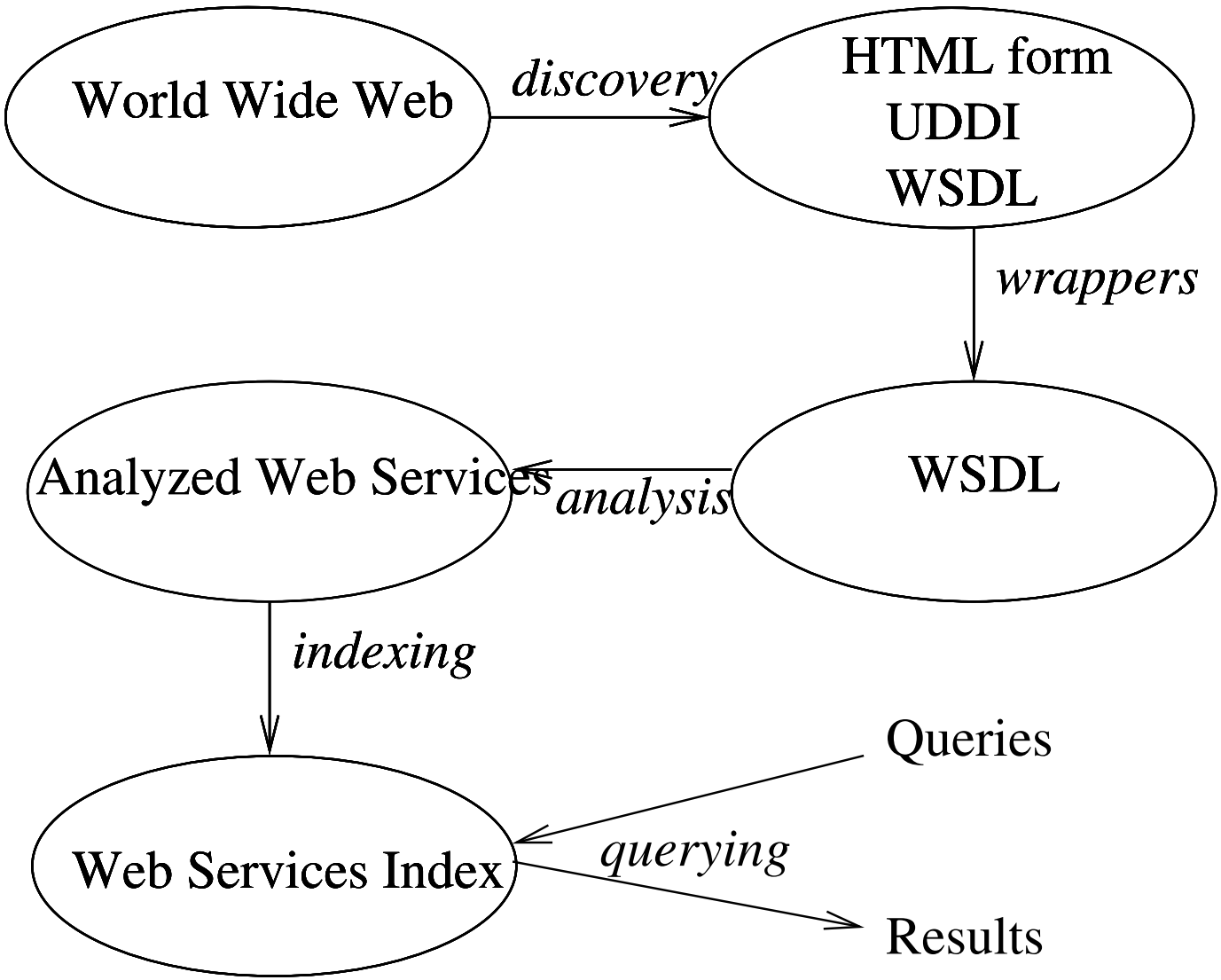
Web Service Semantic Interpretation Process



Web Service Semantic Interpretation Process



Web Service Semantic Interpretation Process



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…)

Only interested in Web Services with no side effects:

Ok	Not Ok
Yellow Pages	Booking services
Publication databases	Mailing list management
…	…

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...)

Only interested in Web Services with no side effects:

Ok	Not Ok
Yellow Pages	Booking services
Publication databases	Mailing list management
...	...

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…)

Only interested in Web Services with no side effects:

Ok	Not Ok
Yellow Pages	Booking services
Publication databases	Mailing list management
…	…

Web Service Discovery

Crawling the World Wide Web for:

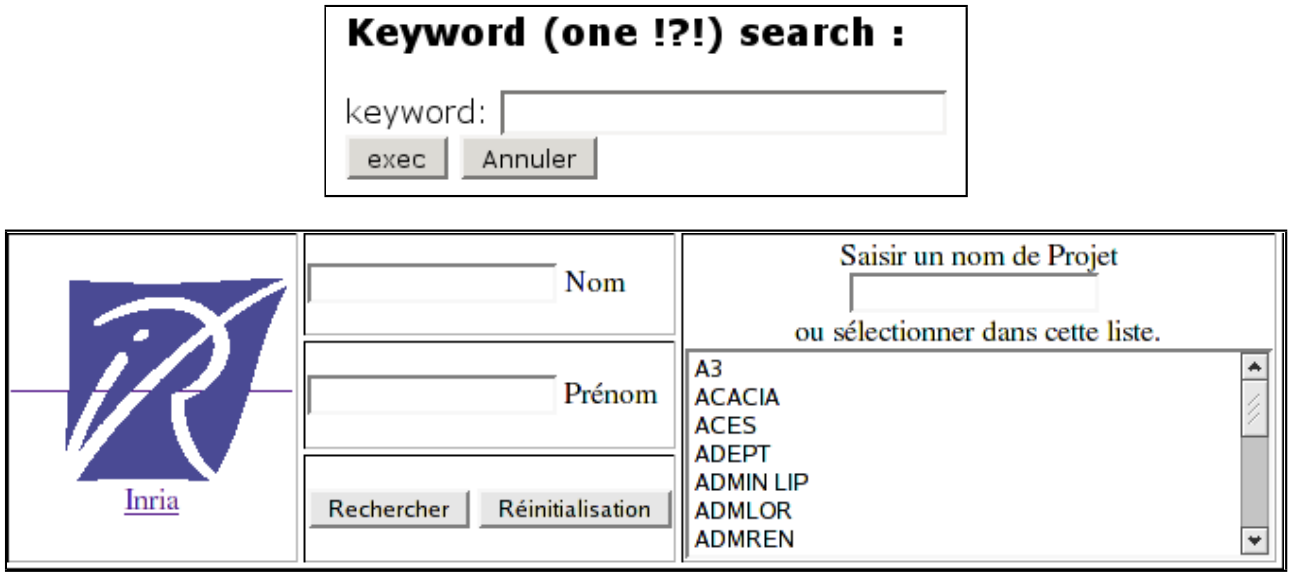
- HTML forms implementing a Web Service
- UDDI registries
- WSDL descriptions
- Other resources (XML, HTML, Web as a full-text index…)

Only interested in Web Services with no side effects:

Ok	Not Ok
Yellow Pages	Booking services
Publication databases	Mailing list management
…	…

Analyzing HTML forms

Analyzing the structure of HTML forms.

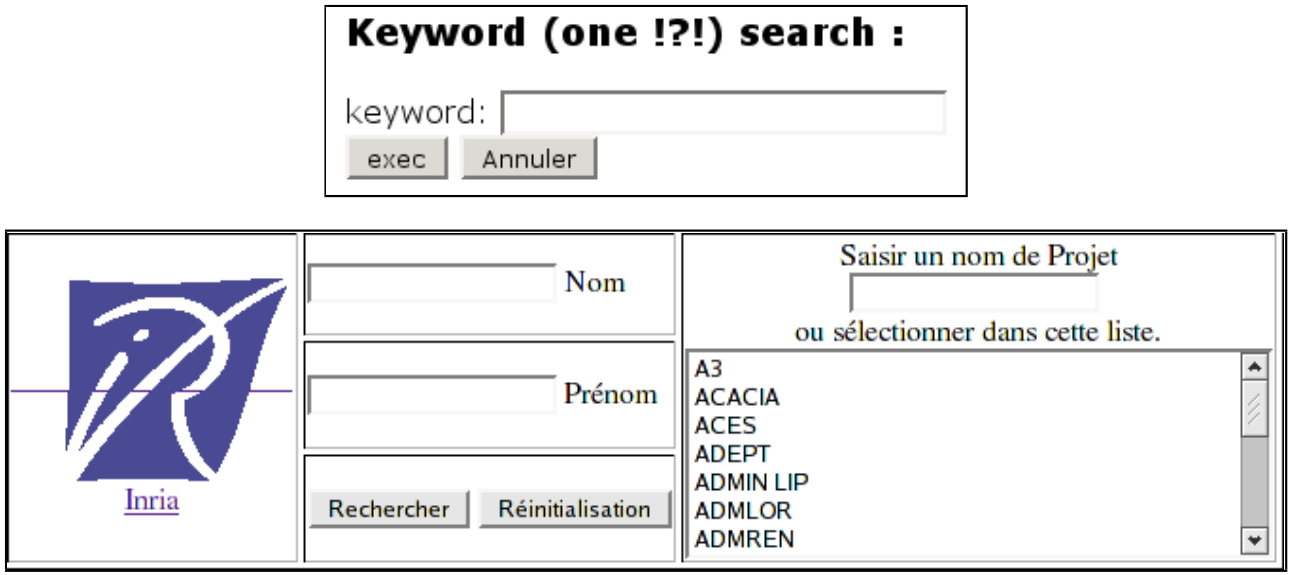


Issues

- 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.

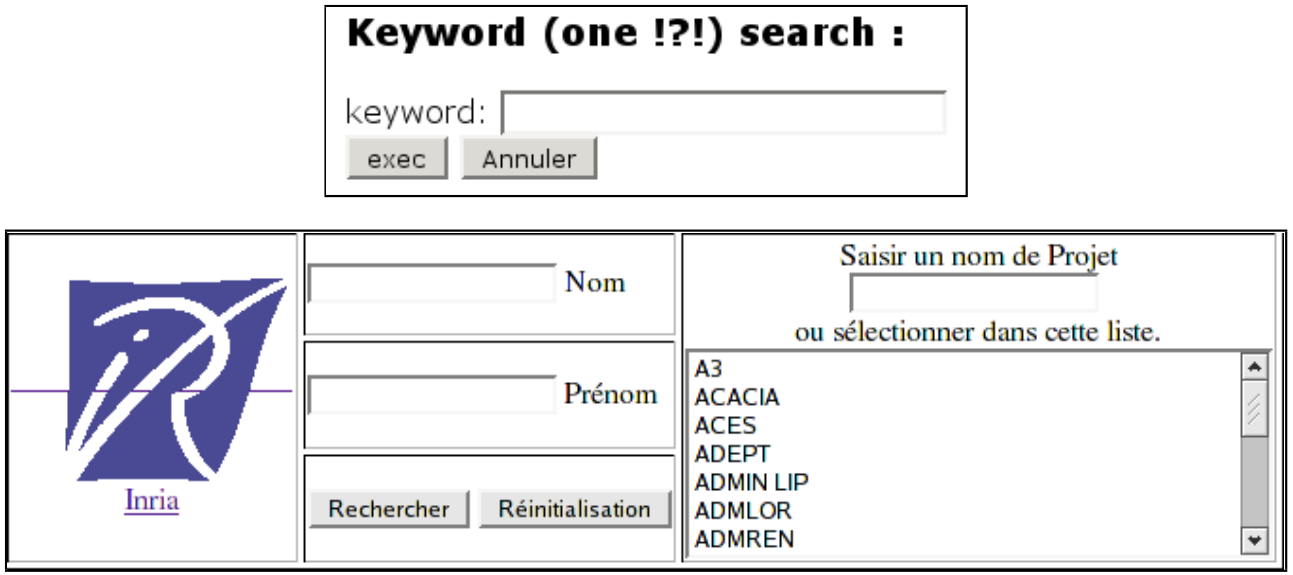


Issues

- 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.



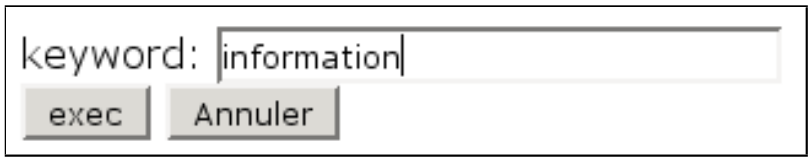
Issues

- 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



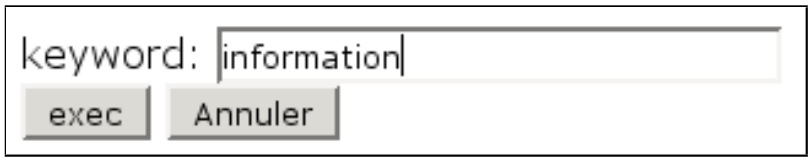
- With nonsense words

- With domain words

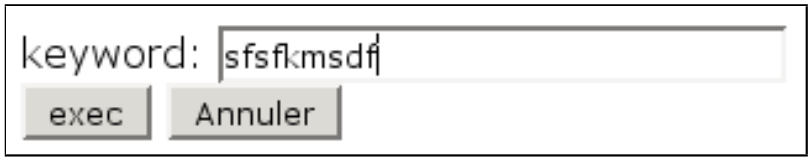
Probing HTML forms

Probing HTML forms to retrieve sample HTML answer pages:

- With dictionary words



- With nonsense words

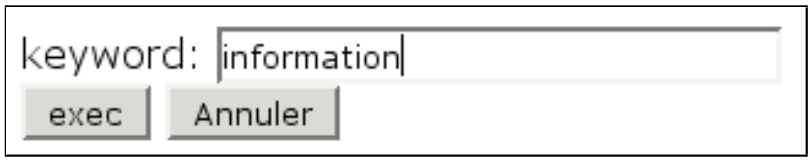


- With domain words

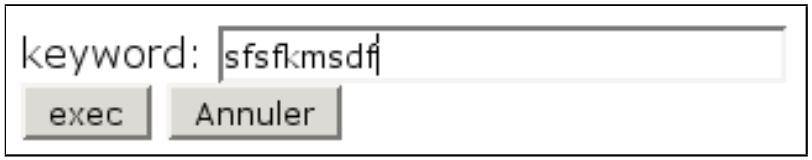
Probing HTML forms

Probing HTML forms to retrieve sample HTML answer pages:

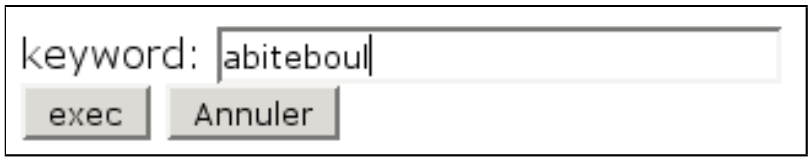
- With dictionary words



- With nonsense words

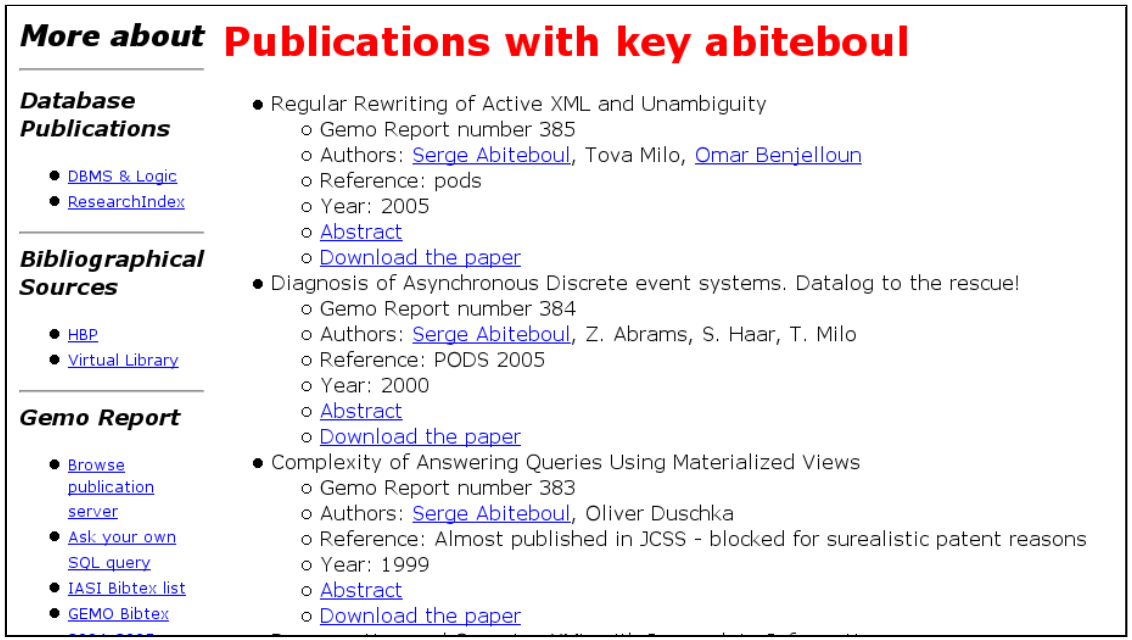


- With domain words



Query-answer webpages

Extract data from query-answer webpages.



More about

Database Publications

- DBMS & Logic
- ResearchIndex

Bibliographical Sources

- HBP
- Virtual Library

Gemo Report

- Browse publication server
- Ask your own SQL query
- IASI Bibtex list
- GEMO Bibtex

Publications with key abiteboul

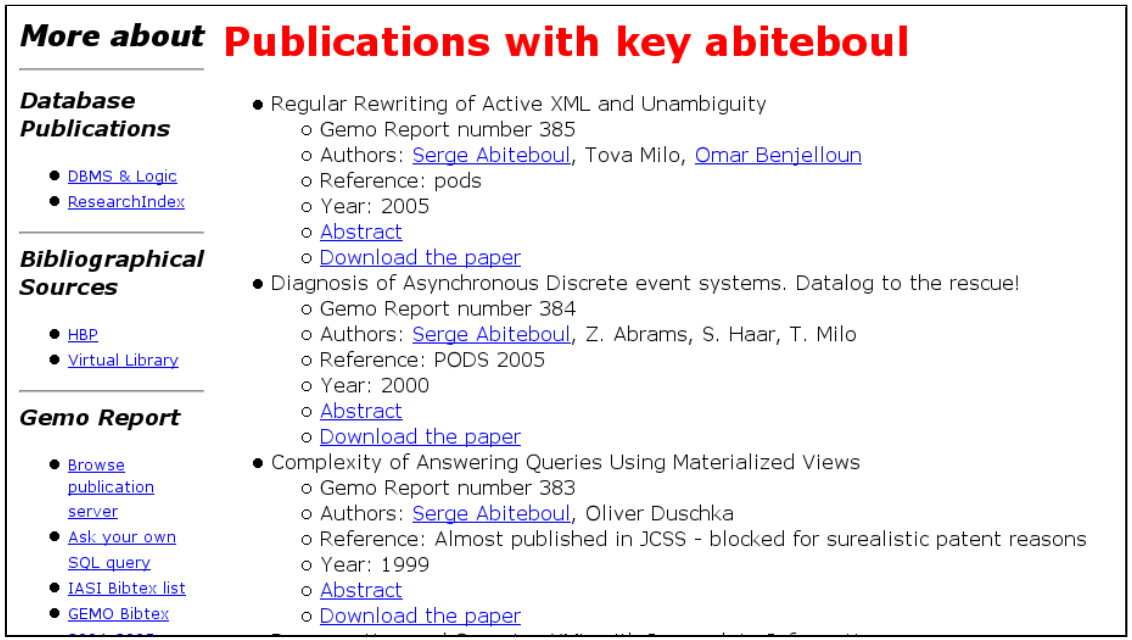
- Regular Rewriting of Active XML and Unambiguity
  - Gemo Report number 385
  - Authors: Serge Abiteboul, Tova Milo, Omar Benjelloun
  - Reference: pods
  - Year: 2005
  - Abstract
  - Download the paper
- Diagnosis of Asynchronous Discrete event systems. Datalog to the rescue!
  - Gemo Report number 384
  - Authors: Serge Abiteboul, Z. Abrams, S. Haar, T. Milo
  - Reference: PODS 2005
  - Year: 2000
  - Abstract
  - Download the paper
- Complexity of Answering Queries Using Materialized Views
  - Gemo Report number 383
  - Authors: Serge Abiteboul, Oliver Duschka
  - Reference: Almost published in JCSS - blocked for surealistic patent reasons
  - Year: 1999
  - Abstract
  - Download the paper

Issues

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

Query-answer webpages

Extract data from query-answer webpages.



More about

Database Publications

- DBMS & Logic
- ResearchIndex

Bibliographical Sources

- HBP
- Virtual Library

Gemo Report

- Browse publication server
- Ask your own SQL query
- IASI Bibtex list
- GEMO Bibtex

Publications with key abiteboul

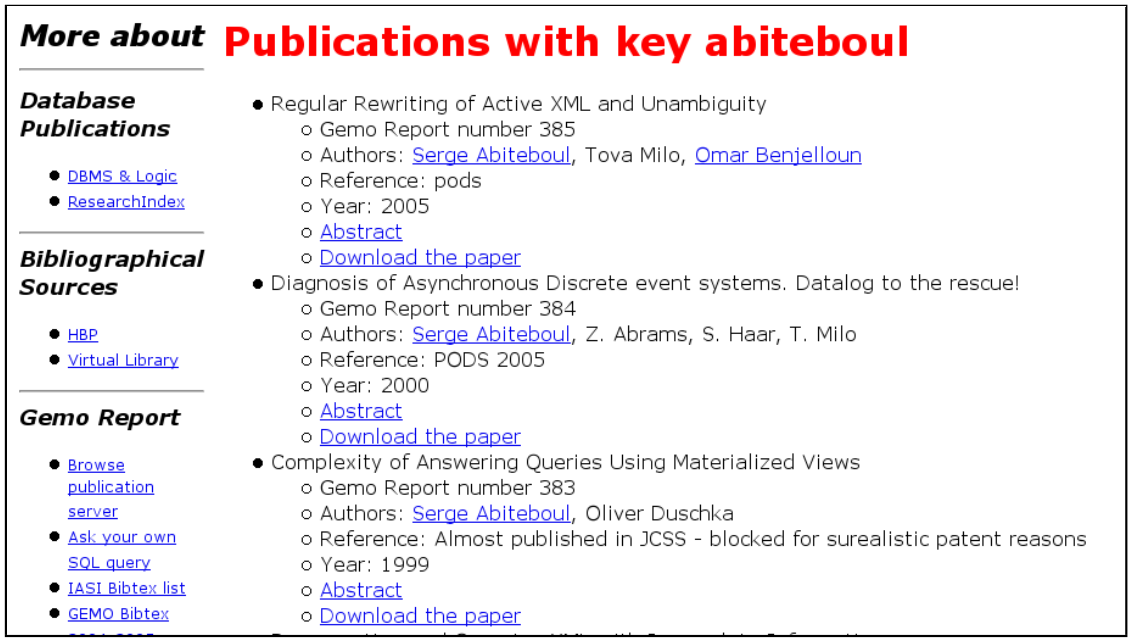
- Regular Rewriting of Active XML and Unambiguity
  - Gemo Report number 385
  - Authors: Serge Abiteboul, Tova Milo, Omar Benjelloun
  - Reference: pods
  - Year: 2005
  - Abstract
  - Download the paper
- Diagnosis of Asynchronous Discrete event systems. Datalog to the rescue!
  - Gemo Report number 384
  - Authors: Serge Abiteboul, Z. Abrams, S. Haar, T. Milo
  - Reference: PODS 2005
  - Year: 2000
  - Abstract
  - Download the paper
- Complexity of Answering Queries Using Materialized Views
  - Gemo Report number 383
  - Authors: Serge Abiteboul, Oliver Duschka
  - Reference: Almost published in JCSS - blocked for surealistic patent reasons
  - Year: 1999
  - Abstract
  - Download the paper

Issues

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

Query-answer webpages

Extract data from query-answer webpages.



More about

Database Publications

- DBMS & Logic
- ResearchIndex

Bibliographical Sources

- HBP
- Virtual Library

Gemo Report

- Browse publication server
- Ask your own SQL query
- IASI Bibtex list
- GEMO Bibtex

Publications with key abiteboul

- Regular Rewriting of Active XML and Unambiguity
  - Gemo Report number 385
  - Authors: Serge Abiteboul, Tova Milo, Omar Benjelloun
  - Reference: pods
  - Year: 2005
  - Abstract
  - Download the paper
- Diagnosis of Asynchronous Discrete event systems. Datalog to the rescue!
  - Gemo Report number 384
  - Authors: Serge Abiteboul, Z. Abrams, S. Haar, T. Milo
  - Reference: PODS 2005
  - Year: 2000
  - Abstract
  - Download the paper
- Complexity of Answering Queries Using Materialized Views
  - Gemo Report number 383
  - Authors: Serge Abiteboul, Oliver Duschka
  - Reference: Almost published in JCSS - blocked for surealistic patent reasons
  - Year: 1999
  - Abstract
  - Download the paper

Issues

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

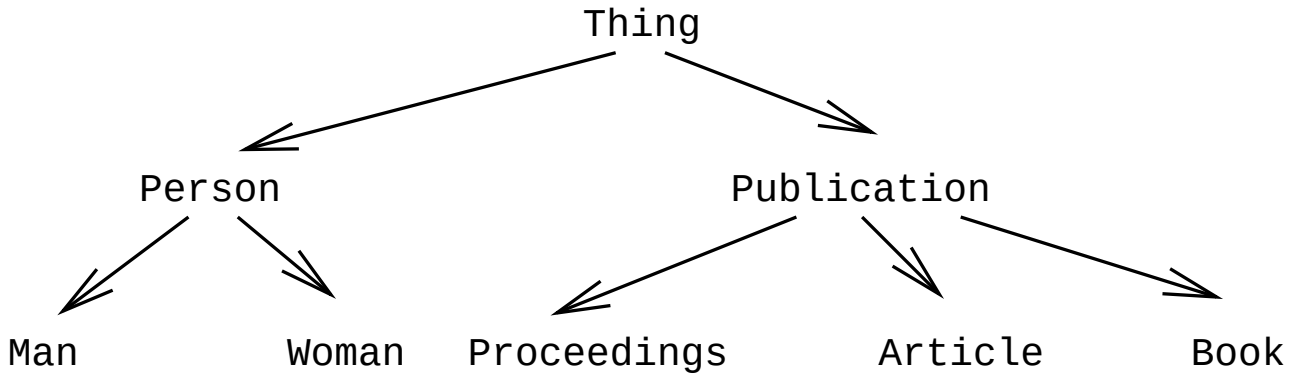
RoadRunner

Example: RoadRunner information extraction engine

Publication of Gemo with key	Publications with key			
Publication of Gemo with key abiteboul	Publications with key abiteboul	_C_	Gemo Report number	
		Regular Rewriting of Active XML and Unambiguity	Gemo Report number 385	• Authors: Serge Abiteboul, Tova Milo, Omar Benjelloun
		Diagnosis of Asynchronous Discrete event systems. Datalog to the rescue!	Gemo Report number 384	• Authors: Serge Abiteboul, Z. Abrams, S. Haar, T. Milo
		Complexity of Answering Queries Using Materialized Views	Gemo Report number 383	• Authors: Serge Abiteboul, Oliver Duschka
		Representing and Querying XML with Incomplete Information	Gemo Report number 382	• Authors: Serge Abiteboul, Luc Segoufin, Victor Vianu
		Active XML: A Data-Centric Perspective on Web Services	Gemo Report number 381	• Authors: Serge Abiteboul, Omar Benjelloun, Ioana Manolescu, Tova Milo, Roger Weber

Conceptual Model

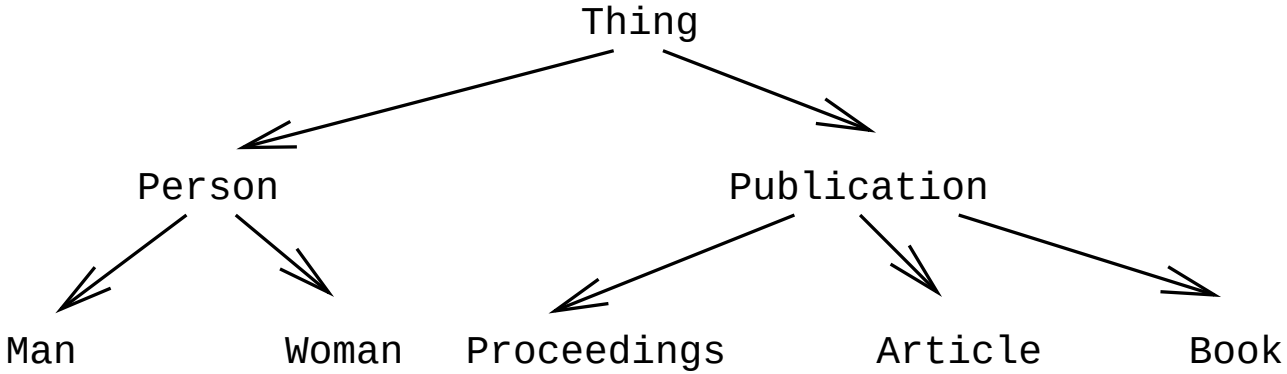
- IsA ontology of concepts (simple DAG)



- *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

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

Example

Query:

$$\langle A, T^* \rangle^* \leftarrow \texttt{WrittenBy}(P,A), \texttt{Article}(P), \texttt{HasTitle}(P,T), \texttt{KeywordOf}(\text{“xml”}, P)$$

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.

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 webservices

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 webservices

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 webservices

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 webservices

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!

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!

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!

Web Service Semantic Interpretation Process