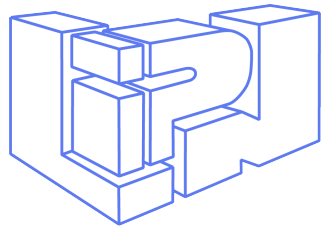


How to Assist Formalization of NL Regulations
Lessons from Business Rules Acquisition Experiments

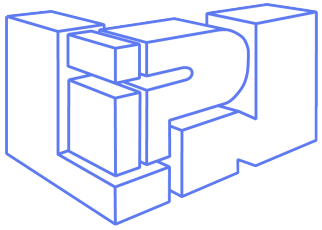
Adeline Nazarenko

adeline.nazarenko@lipn.univ-paris13.fr

LIPN

Paris13 University – Sorbonne Paris Cité
CNRS

December 13, 2013 — JURIX 2013



Acknowledgements

This is a joint work with François Lévy (LIPN, Paris 13 Univ. – SPC & CNRS)

- ONTORULE project (FP7)

ONTORULE
Ontologiesmeet Business Rules

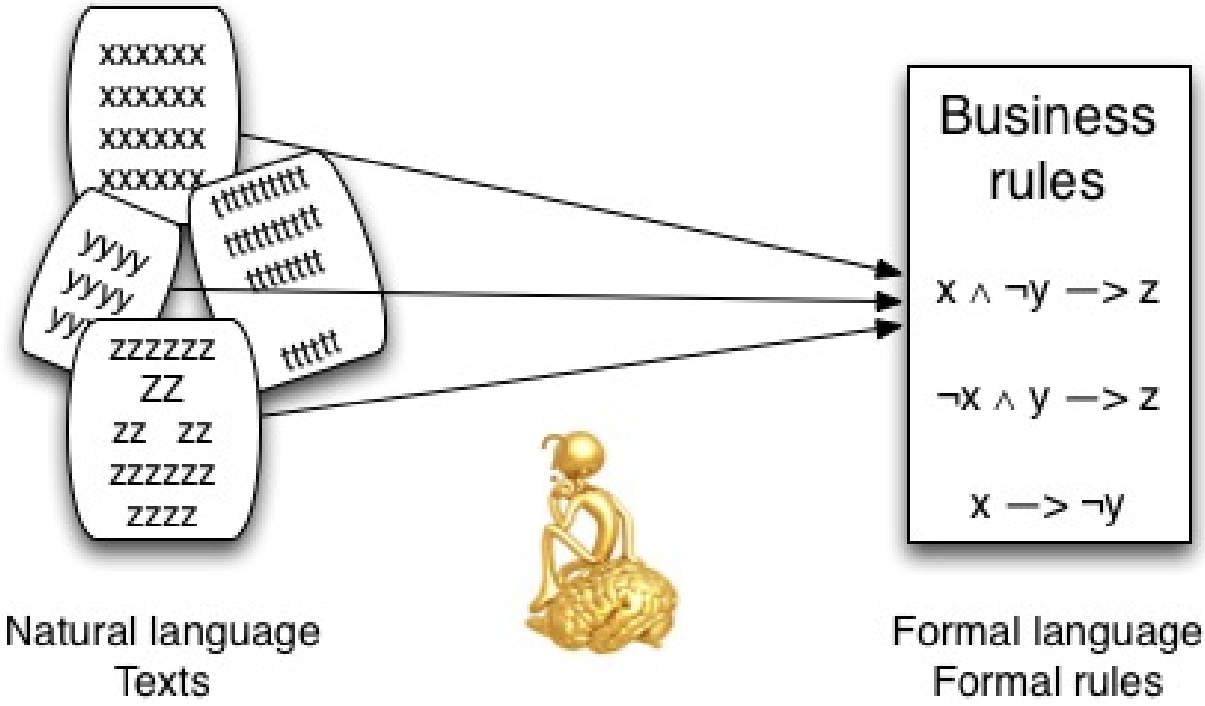
- Quaero program (Oséo)

quaero

- Labex "Empirical Foundations of Linguistics" (ANR-CGI)

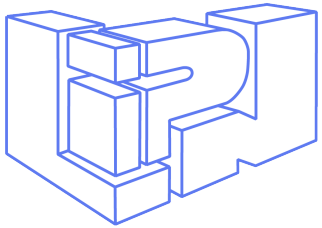
Labex EFL SORBONNE PARIS CITÉ

Deriving formal rules from NL regulations?



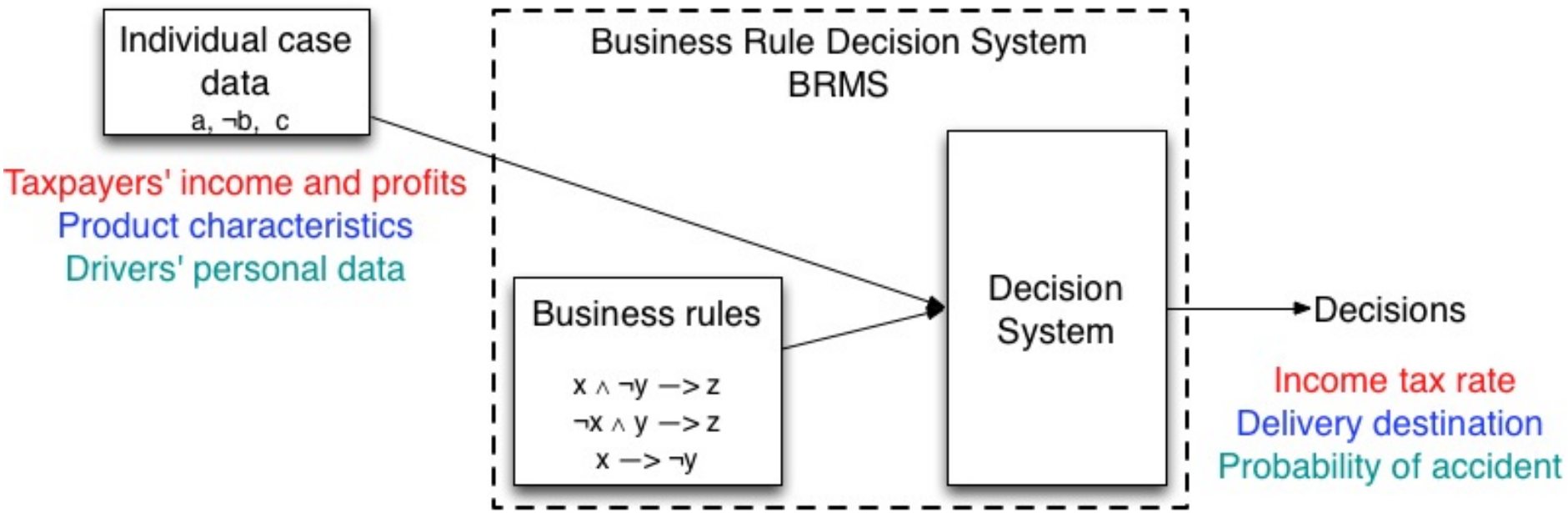
A complex task

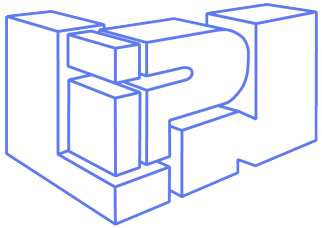
- that cannot be fully automated
- that can be guided using NLP and semantic web technologies



Context

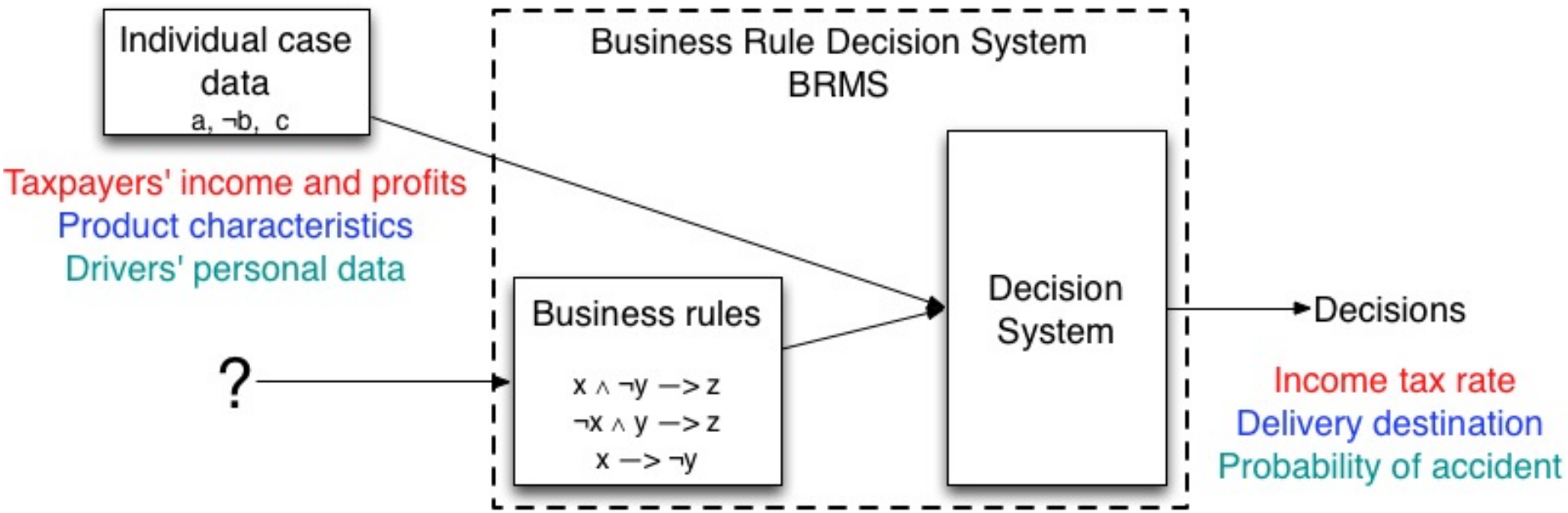
A (Business) Rule Management System takes/suggests decisions on specific cases according to a predefined set of rules.

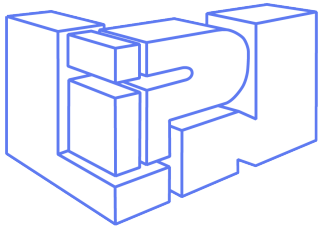




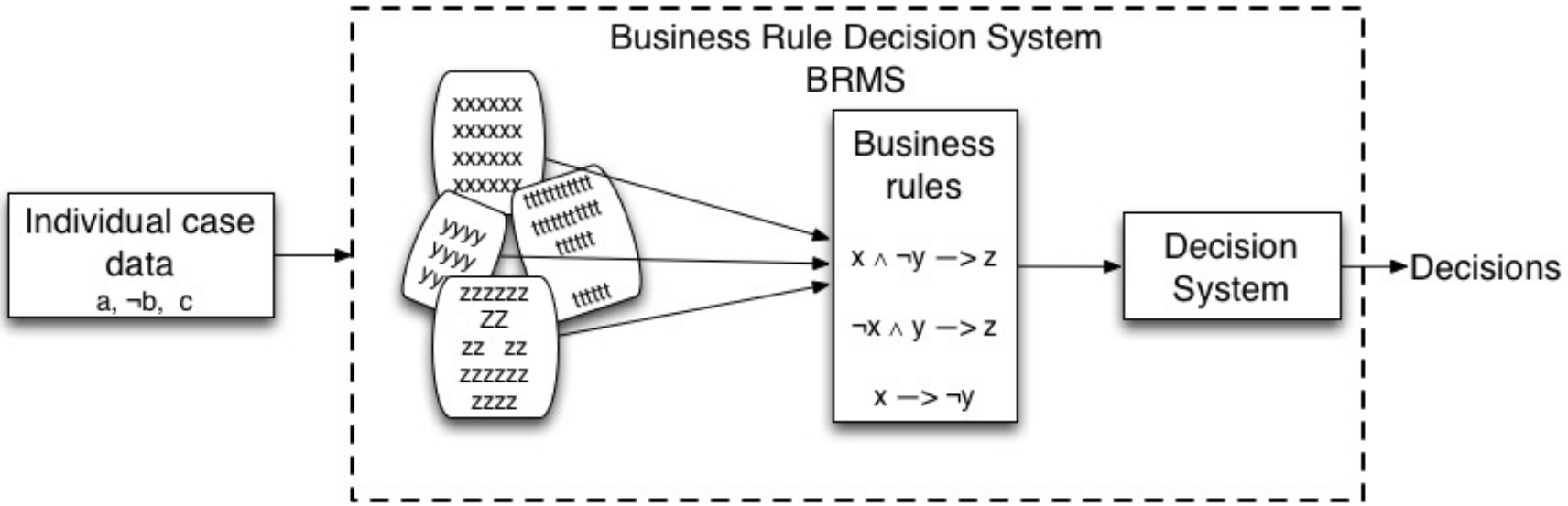
Context

A (Business) Rule Management System takes/suggests decisions on specific cases according to a predefined set of rules.



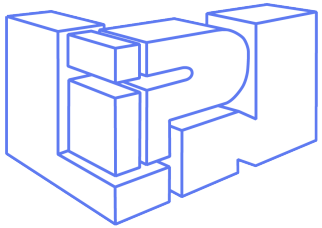


Integrating NL sources in BRMS



Benefits

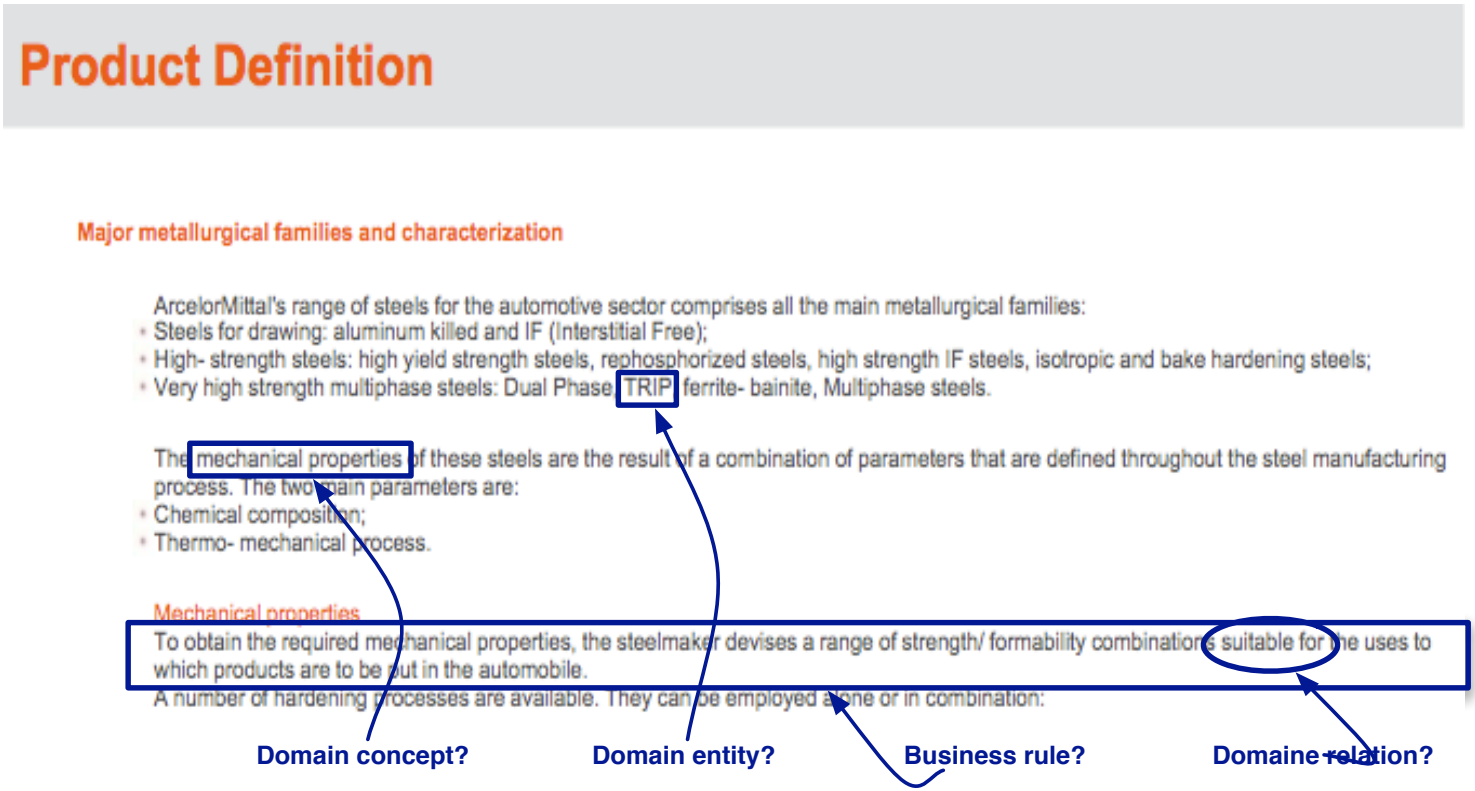
- Knowledge acquisition
- Documentation of decisions
- Knowledge base maintenance

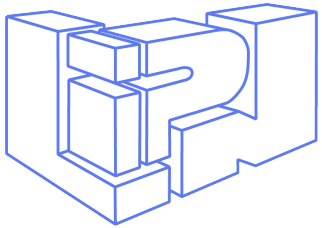


Text-based knowledge acquisition

Type of knowledge (domain model)

- Domain basic knowledge (concepts, entities, relations)
- Rules that control the decision process





Text-based knowledge acquisition

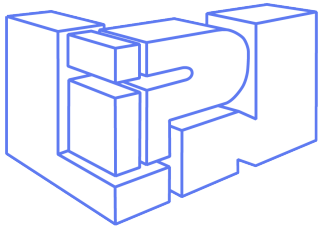
Type of knowledge (domain model)

- Domain basic knowledge (concepts, entities, relations)
- Rules that control the decision process

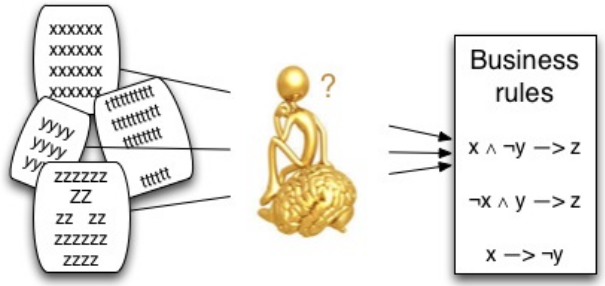
Texts are a convenient source of domain knowledge ($\neq$ Experts)

Texts are a critical source for rule information

- Legal knowledge is primarily expressed in NL texts

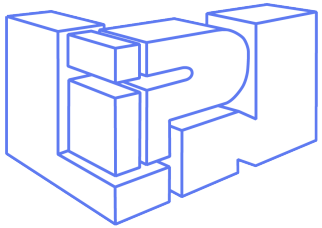


The formalization problem



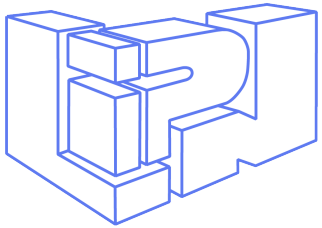
- Natural and formal rule languages stand on the opposite extremities of the formalization continuum [Baumeister *et al.*, 2011]
- Natural language is intrinsically complex
  - Factual information and rhetorical elements
  - Redundant and implicit information
  - Lexical and structural ambiguity
  - Understatement and underdetermination
- Direct and automatic translation to formal language is impossible
  - Existing approaches apply on simplified problems

    [Unger *et al.*, 2012] [Dinesh *et al.*, 2008] [Bajwa *et al.*, 2011]

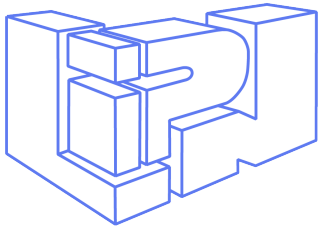


Use cases

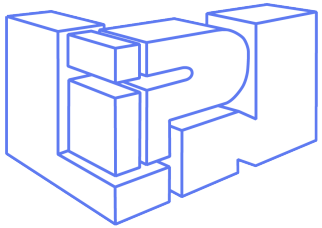
- Arcelor Mittal: assignment of coil products (internal documentation)
- Audi: certification of seat belts (UNO regulations)
- American Airlines: calculation of frequent flyer's miles & bonus (AAdvantage Frequent Flier Program)
- Car rental: terms and conditions checking (SBVR tutorial)



Outline

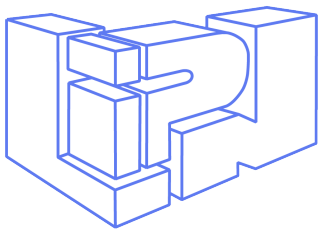


Outline

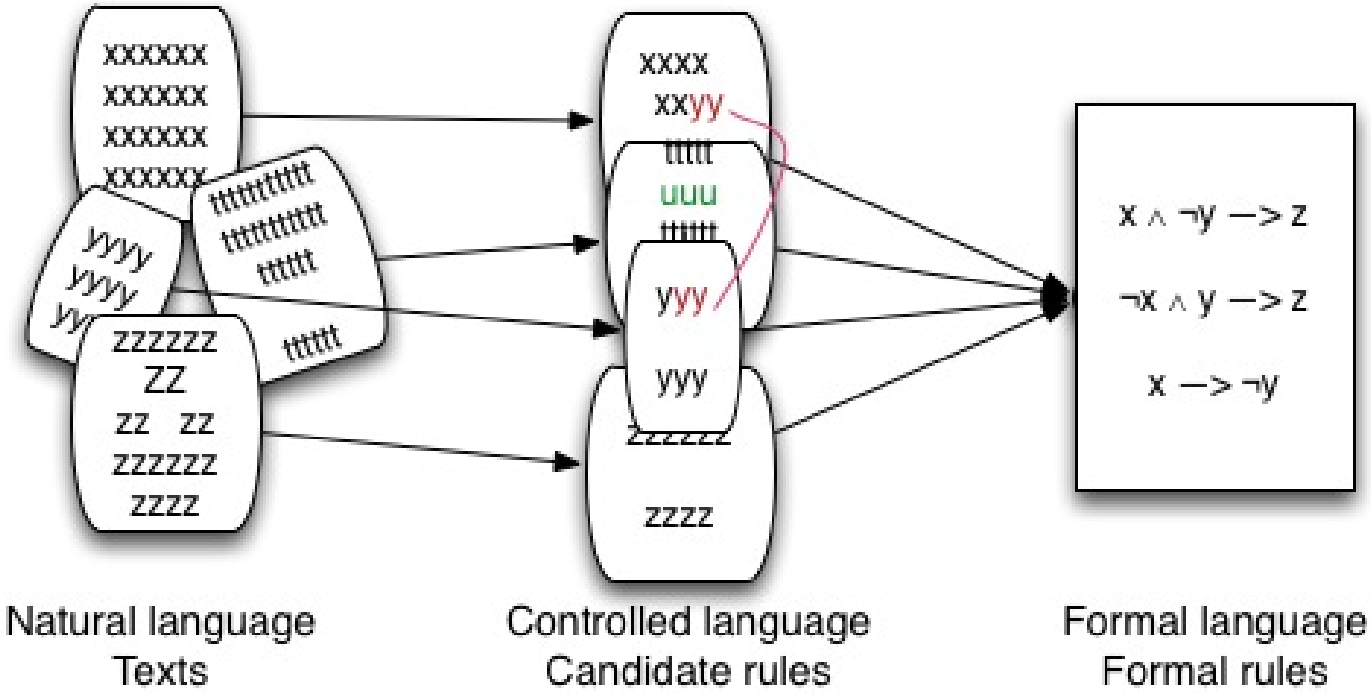


Rule extraction?

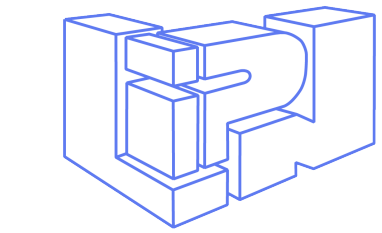
It is impossible to directly extract business rules from textual sources



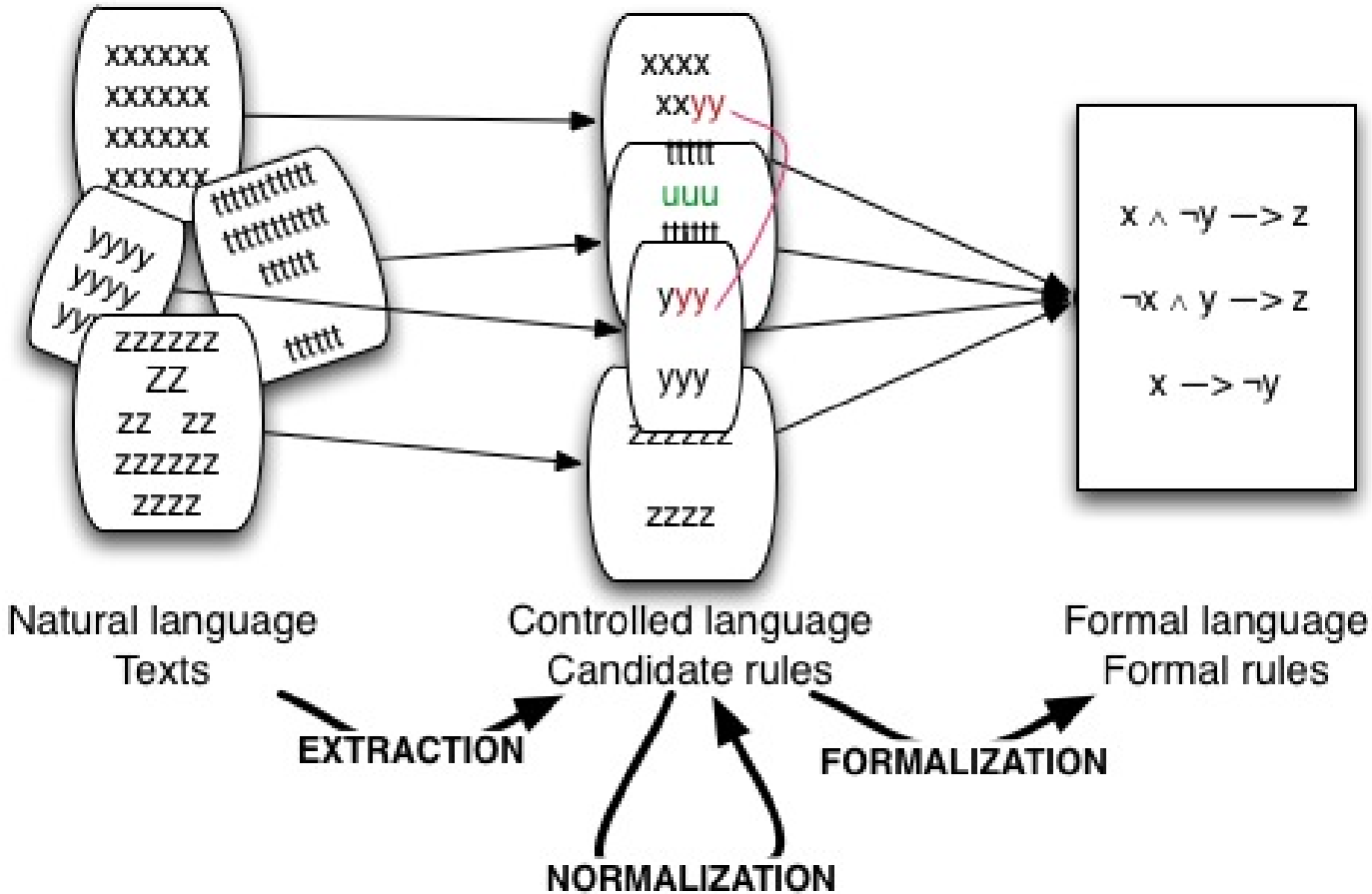
Formalization, a mediated process

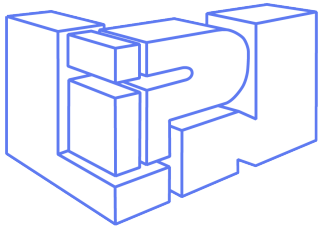


Controlled language as a mediator between NL and formal languages



Formalization, a mediated process





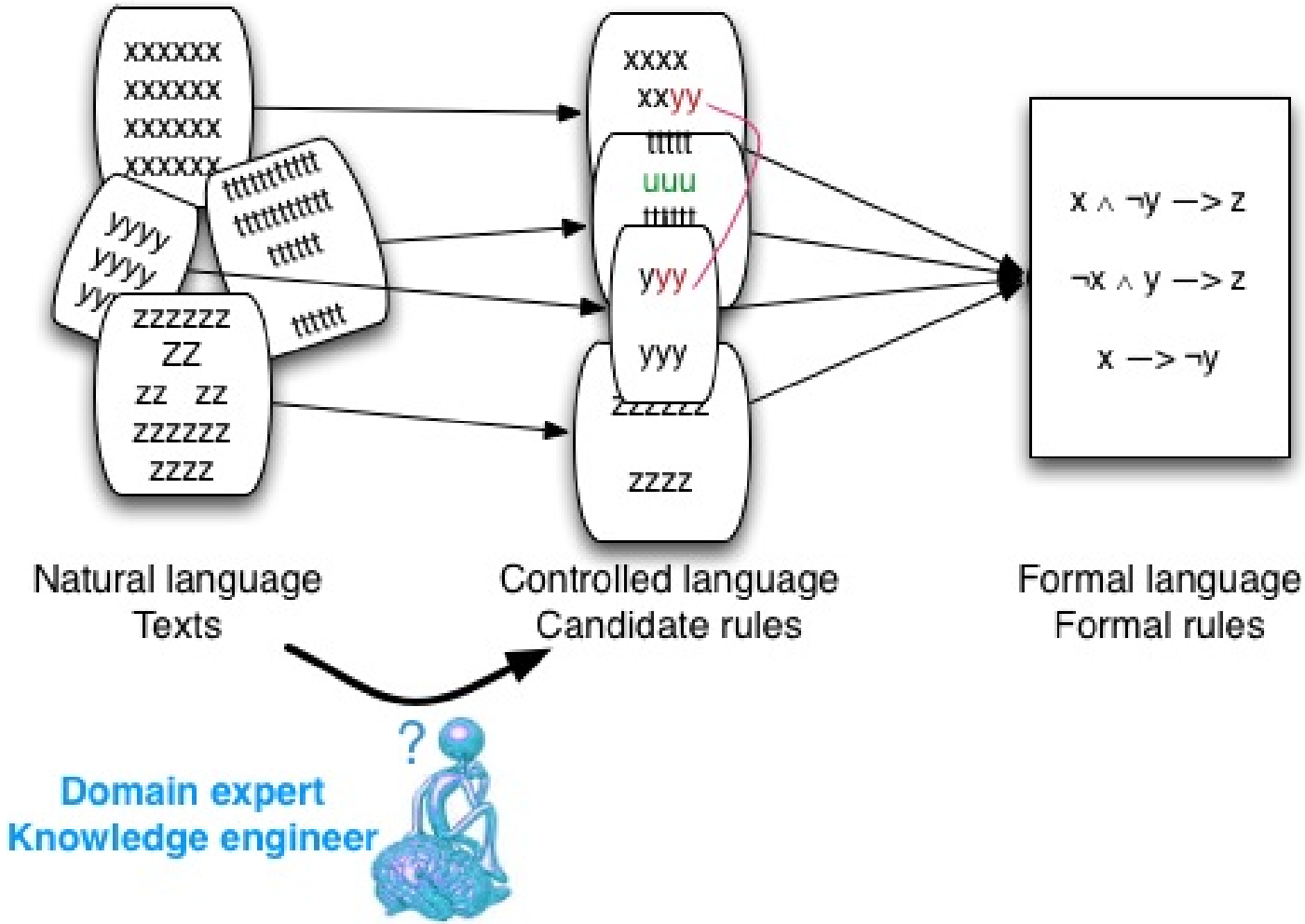
Collaboration

Several actors are usually involved in the acquisition process

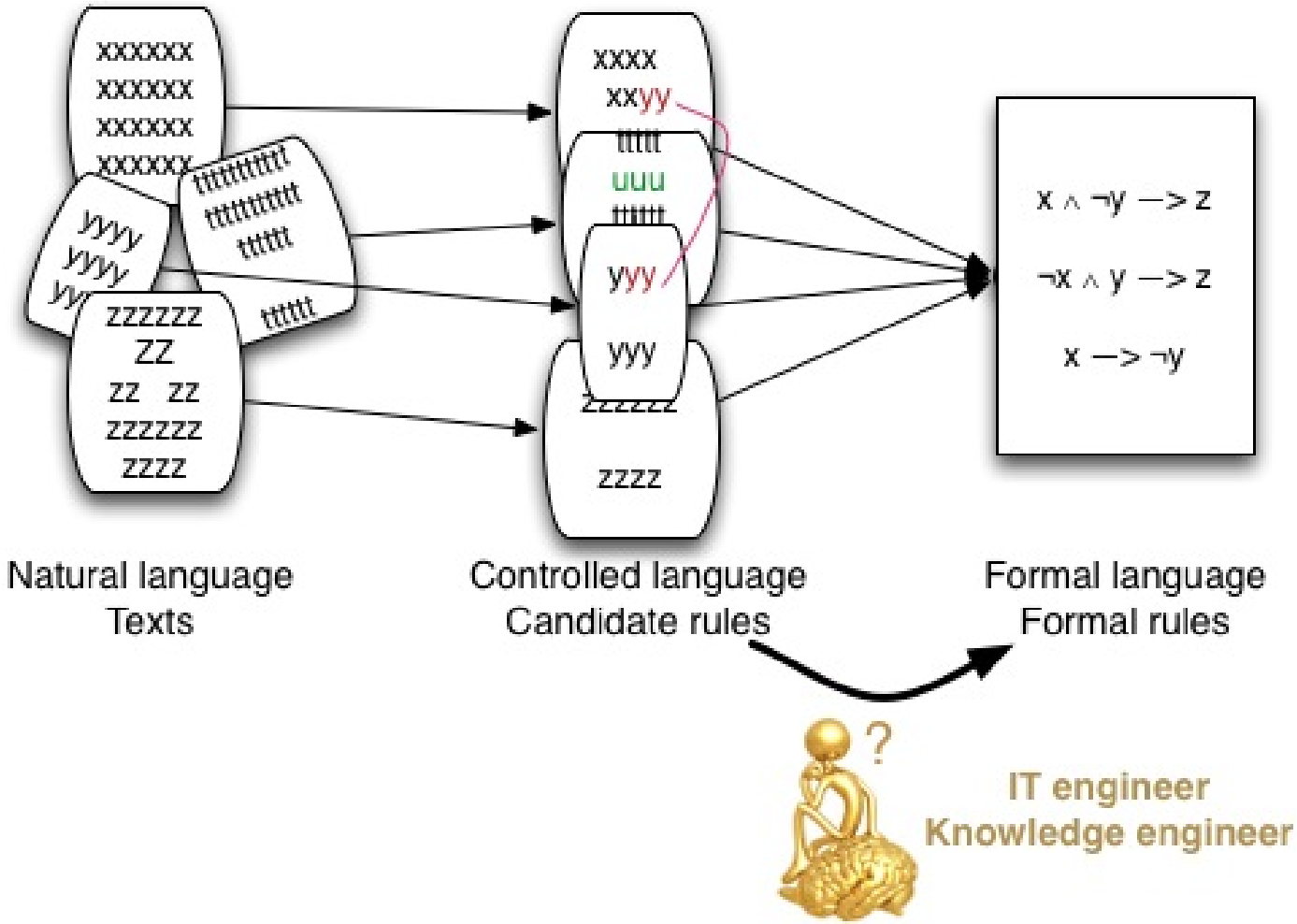
- The domain expert knows the business context and understands the written documentation
- The knowledge engineer knows how to structure and express knowledge
- The IT engineer understands how the target decision system works and how to implement the rules

None masters the whole process, from the business case to the detailed system implementation

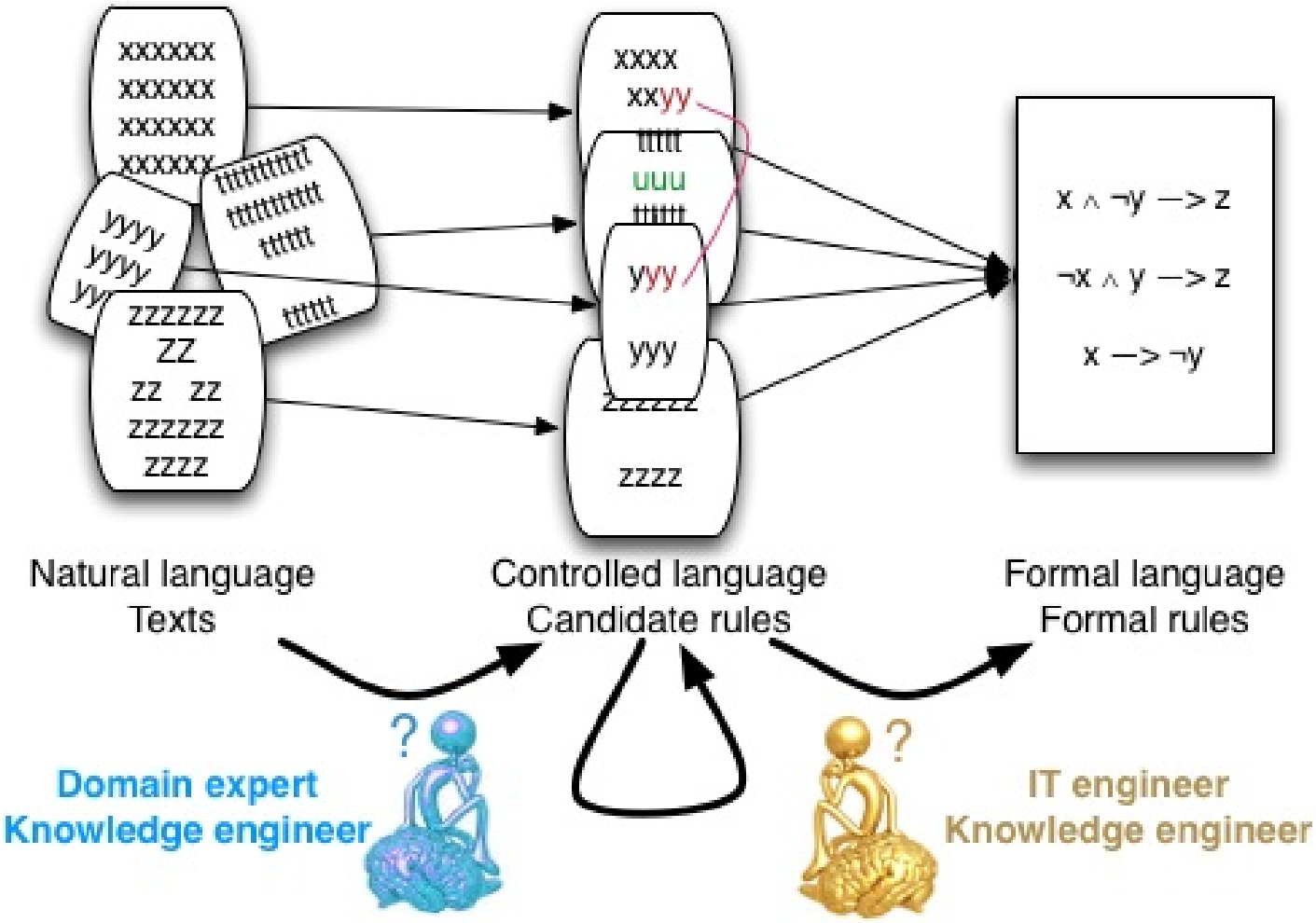
Formalization, a collaborative process

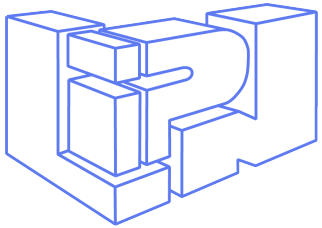


Formalization, a collaborative process



Formalization, a collaborative process





Person-Machine interaction

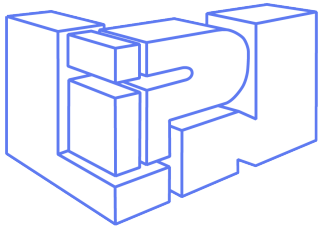
Human expertise is required for

1. Comprehensive understanding of the source regulation and the target application
   - Identifying relevant sources of information
   - Browsing complex documentations
   - Selecting the text fragments that are relevant for the target application
2. Modeling and formalization
   - Structuring domain knowledge
   - Expressing the rules in such a way that they can be properly operated to make decisions

Each rental *has exactly one* renter

↓

Each rental *has one and the same* renter *from* its begining *to* its end



Towards Person-Machine interaction

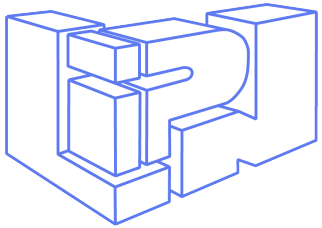
The acquisition and formalization task is

- too complex to be fully automated
- too difficult and time consuming for human experts

An interactive and collaborative approach to help the user

- exploring the source documentation
- coping with the semantic difficulties

$\longrightarrow$A platform to assist experts in rule acquisition and formalization



SemEx

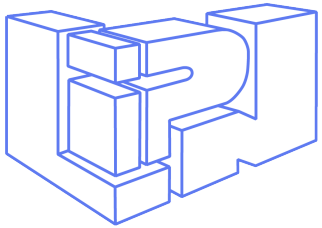
Semantic **Ex**plorer [Lévy *et al.*, 2010a] [Guissé *et al.*, 2011]

Input Domain lexicalized ontology + NL regulation

Output A documented business rule model

- A formalization methodology
- Tools to support human work
- Standard technologies
  - Eclipse application
  - W3C languages ensuring interoperability

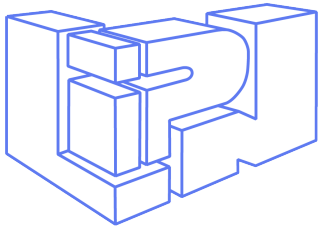
OWL, SKOS, RDF, RDAa, SPARQL



Role of controlled language

The controlled language is used to

- Describe the domain model
  - conceptual model (specialized vocabulary, ontology)
  - rules (prescribed, suggested or self-imposed rules)
- Specify the expected behavior of the rule system
- Verbalize that model in a way that is understandable to domain experts



A basic controlled language

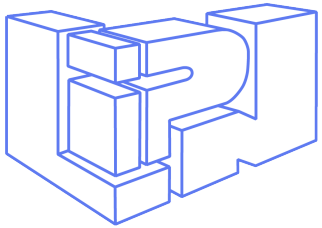
Statements

- Concept definitions: restrictions on the content of terms
- Operative rules: prescribed, suggested, self-imposed rules

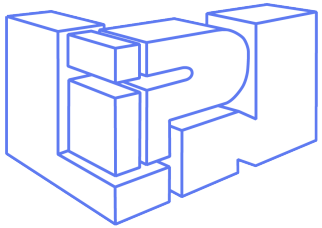
ProducedCoil : Coil that *is produced by* the Galvanization Line
It is obligatory that yield strength *is between* the upper and lower values . . . .

Elements

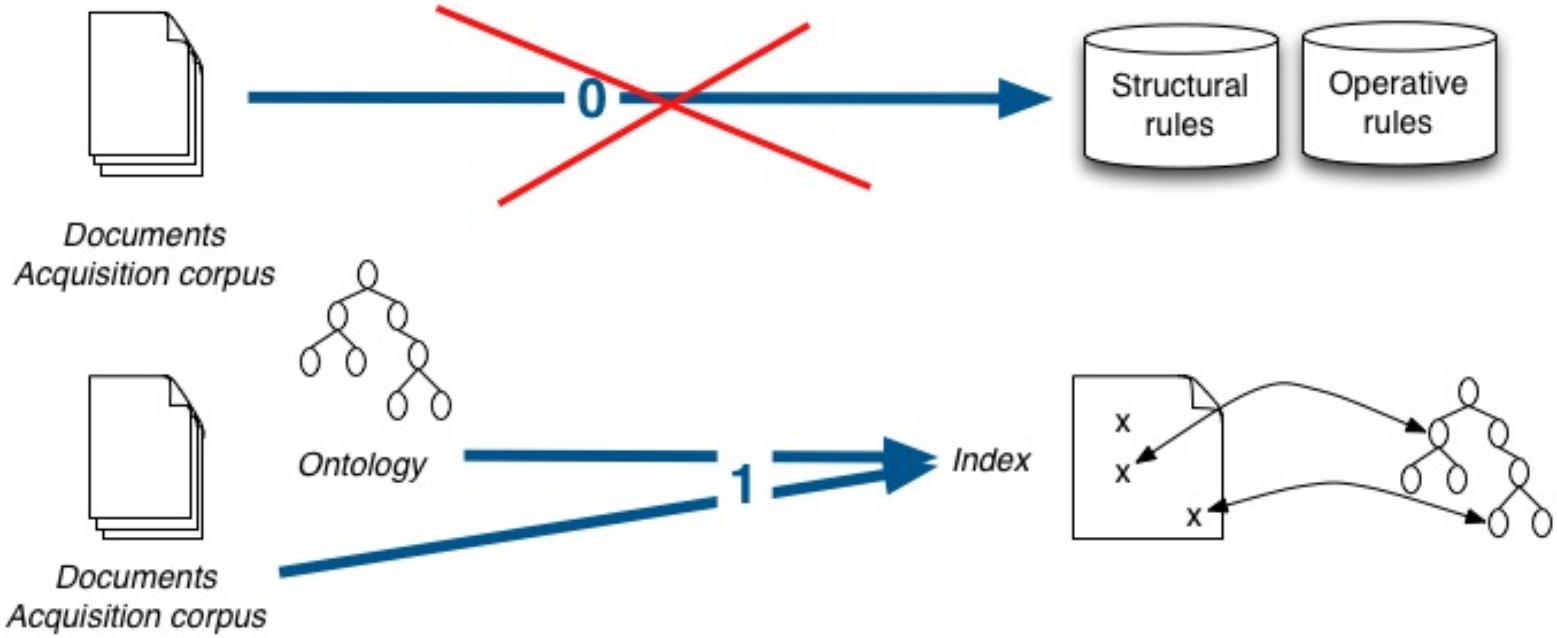
- Conceptual terms: Width Thickness Company Coil
- Individual terms: ArcelorMittal Coil #13 Galvanization Line
- Relational terms: *belongs to*
- Keywords: that
- Modal operators: It is obligatory that must



Outline

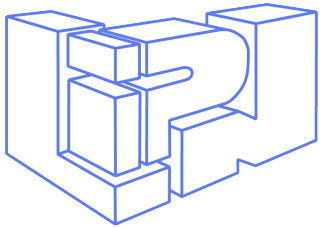


1st step: annotation wrt. an ontology



~~0 Rule extraction~~

1 Semantic annotation wrt. ontology



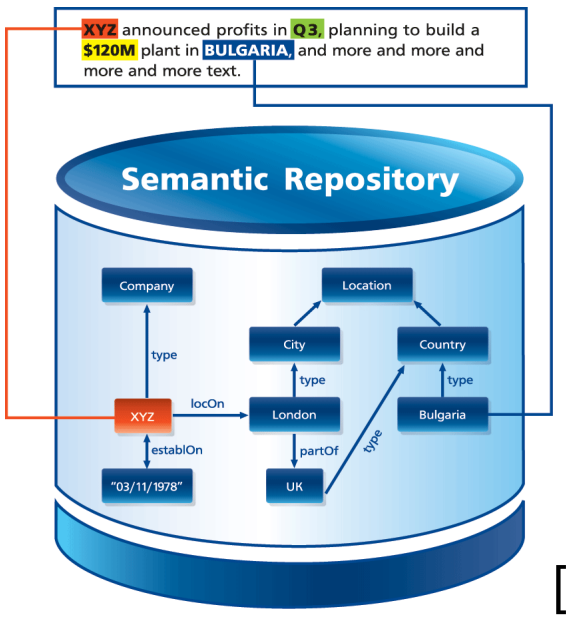
What is semantic annotation?

- Text annotation: metada attached to fragments of a text

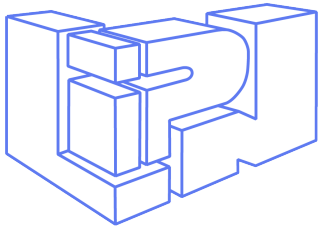
  grade, comment, explanation, presentational markups

- Semantic annotation: the metadata belong to a specific resource

  controlled vocabulary, terminology, gazetteer, thesaurus, ontology



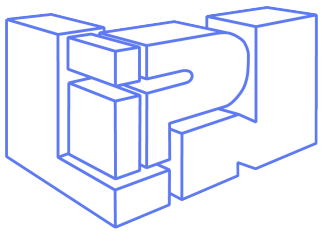
[Popov *et al.*, 2004]



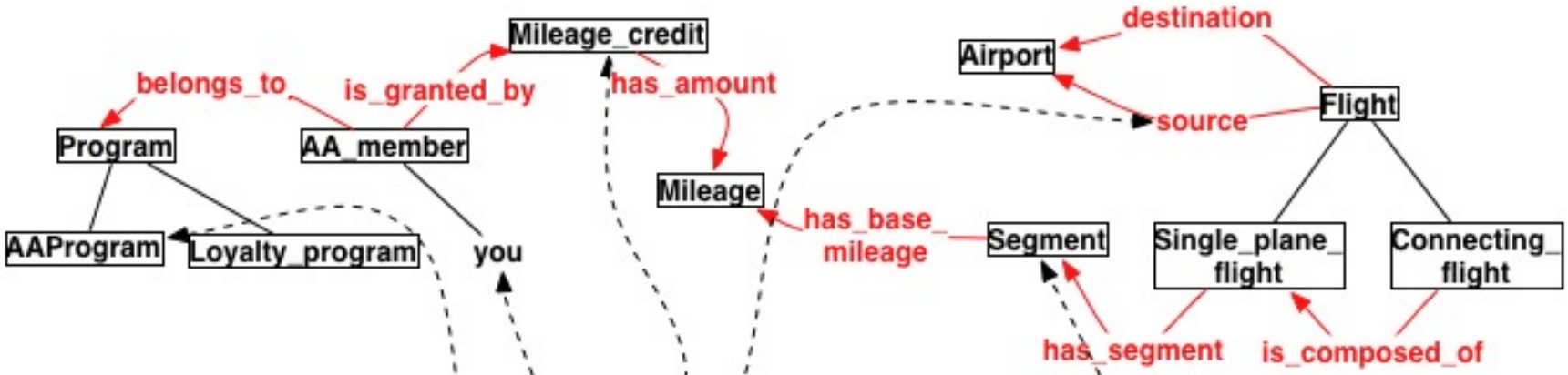
Types of semantic annotation

- Initial focus on instance annotation and ontology population [Vargas-Vera *et al.*, 2002] [Popov *et al.*, 2004] [Amardeilh *et al.*, 2005][Magnini *et al.*, 2006]
- Towards a richer semantic annotation [Ma *et al.*, 2010]
  - Fine-grained annotation
  - Exploitation of all the full semantics of ontologies [Lévy *et al.*, 2010b]

    individuals, concepts, roles/relations, rules

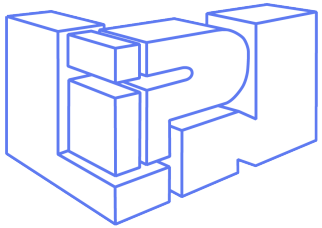


Annotation wrt. an ontology



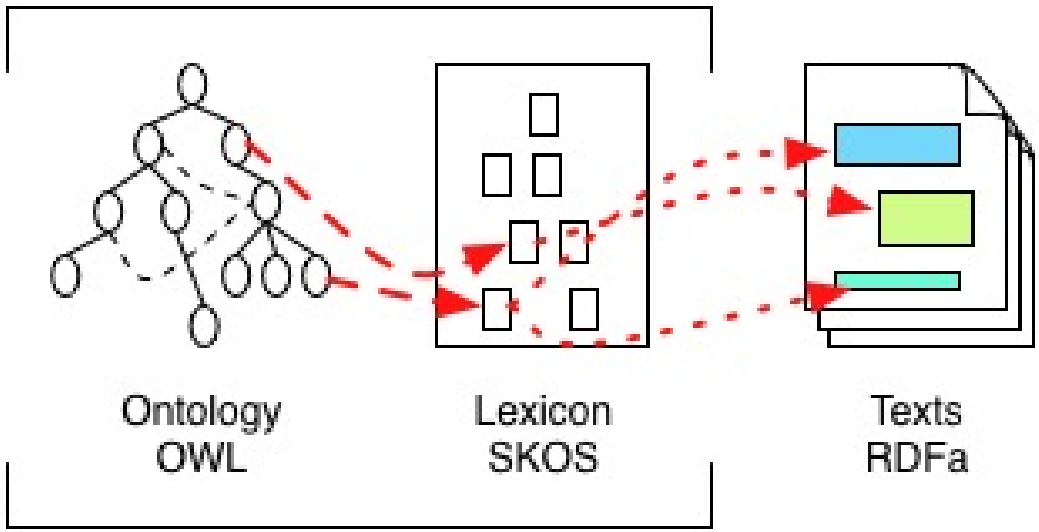
AAdvantage **flight mileage credit** is determined on the basis of nonstop distances between the **airports** where your **flight** originates and terminates.
On **connecting flights**, **you**'ll receive **mileage credit** for each **segment** of your **trip** ;
on **single-plane flights**, **you**'ll receive the nonstop origin-destination **mileage**.

Mileage credit cannot be earned for the same **flight** in more than one of the following **programs**:
the **AAdvantage program** or any other **loyalty program** in which **American Airlines** participates.



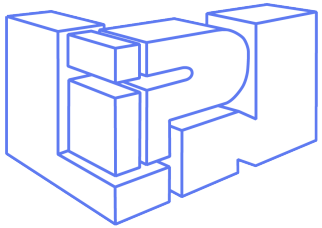
Lexicalized ontology

[Omrane *et al.*, 2011a]



Lexicalized ontology
OWL + SKOS

```
<rdf:Description rdf:about="http://lipn.univ-paris13.fr/RCLN/
   terminae/Audi#SeatBelt">
  <skos:prefLabel>seat belt</skos:prefLabel>
  <skos:altLabel>belt</skos:altLabel>
  <rdf:type rdf:resource="http://www.w3.org/2004/02/skos/core#Concept"/>
</rdf:Description>
```



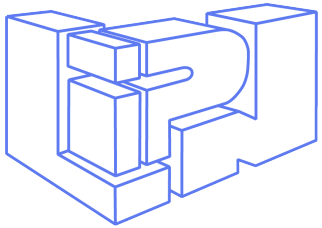
Semantic annotator

Java module

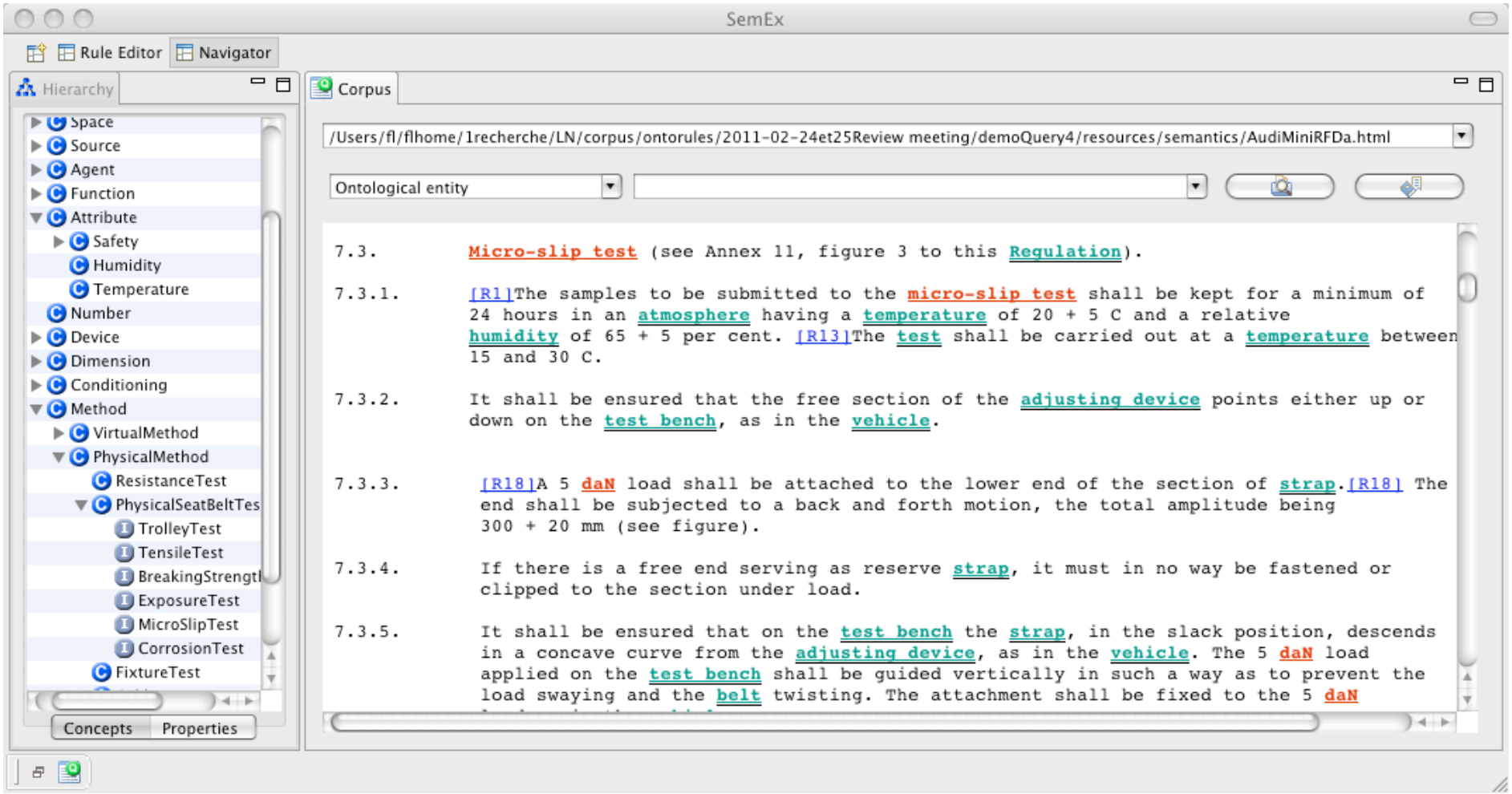
- Originality
  - Can take any ontology as input
  - Can process pre-annotated corpora
- Current version

  Input Text + Lexicalized ontology (lexical items)

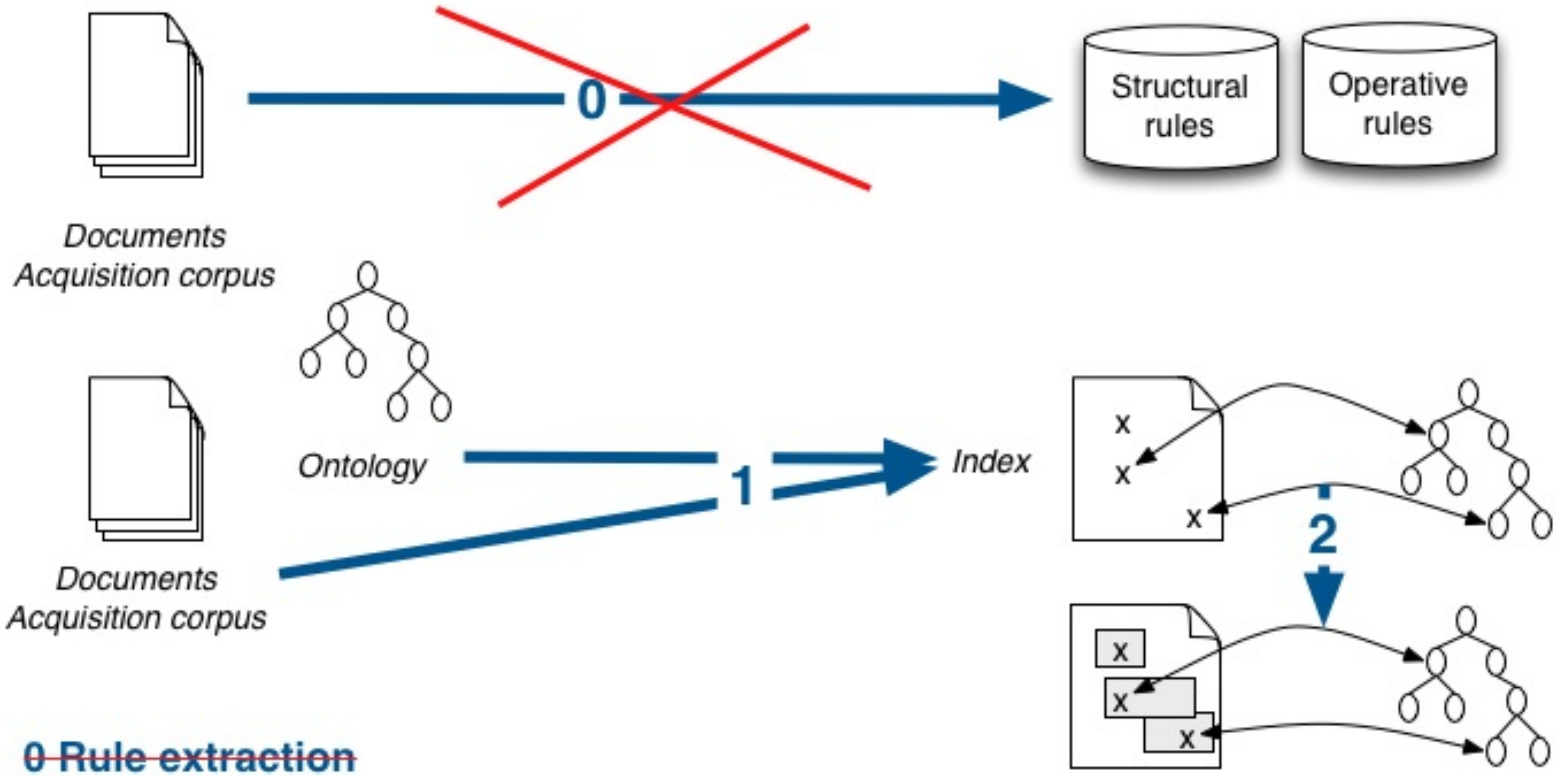
  Output Text with individual and concept mentions annotated
- Future version
  - Role annotations
  - Lexicalized ontology with lexico-syntactic patterns
  - UIMA module



SemEx navigation perspective



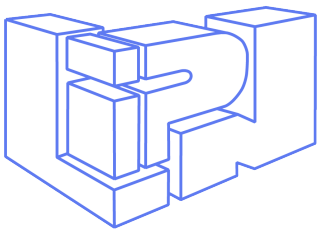
2nd step: Rule selection



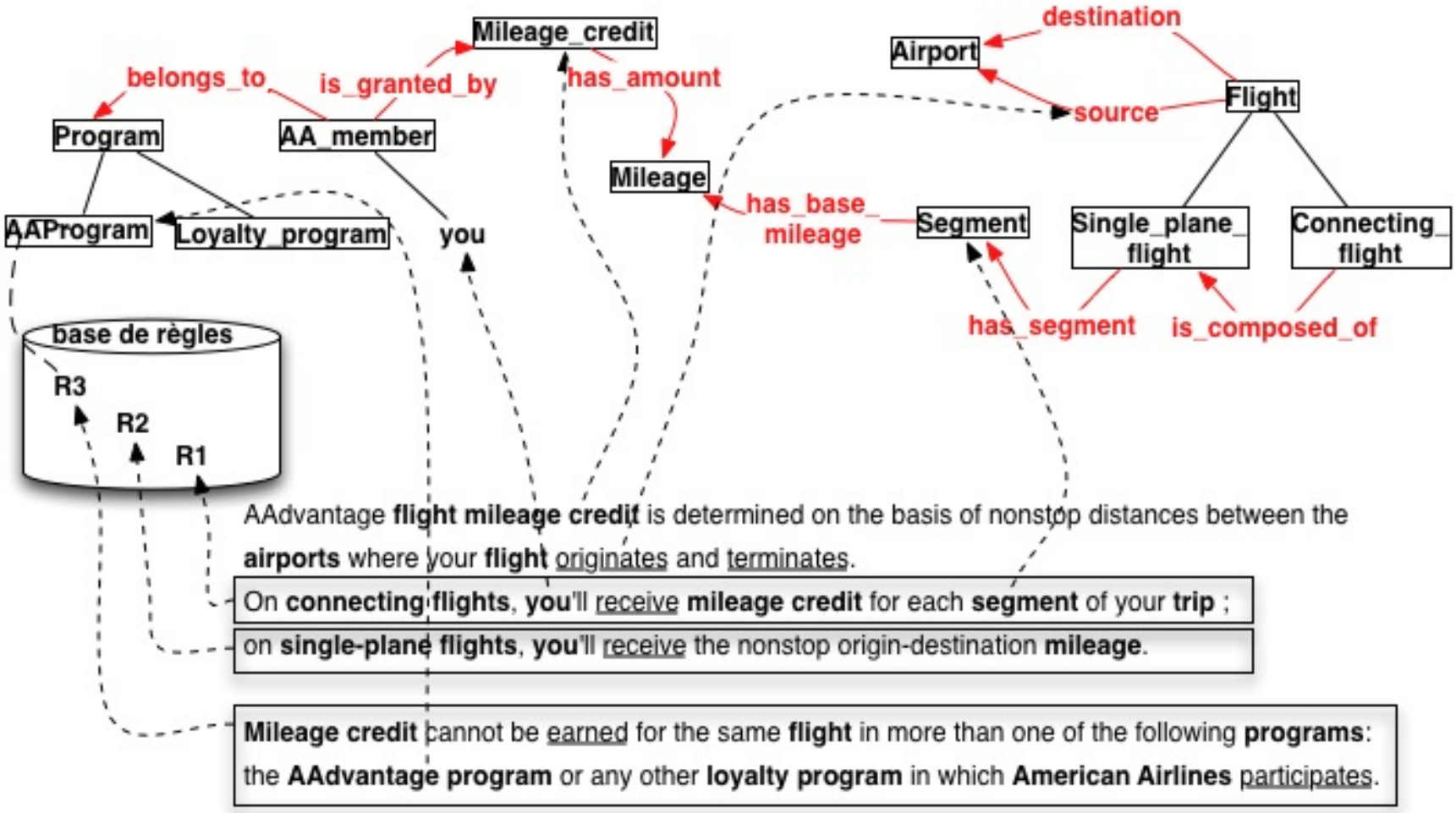
~~0 Rule extraction~~

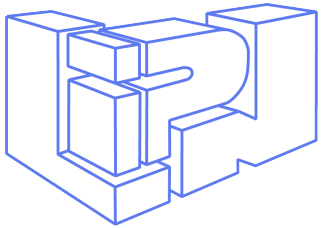
1 Semantic annotation wrt. ontology

2 Extraction of relevant textual fragments



Annotation wrt. an ontology and a rule base



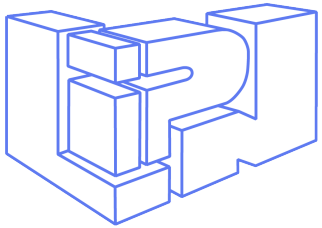


How to guide rule selection?

Rule selection requires good knowledge of the business context and thorough understanding of the source documentation

It relies on human expertise but

- Semantic annotation enhances text browsing

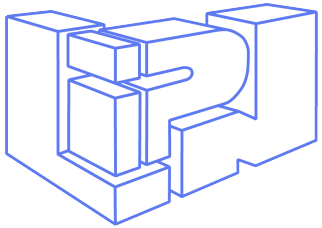


How to guide rule selection?

Rule selection requires good knowledge of the business context and thorough understanding of the source documentation

It relies on human expertise but

- Semantic annotation enhances text browsing
  - Sentences with at least one annotation
    - American Airlines: recall = 39%, low precision
    - Audi: recall = 72%, low precision

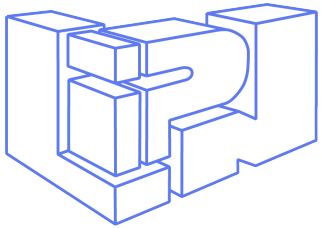


How to guide rule selection?

Rule selection requires good knowledge of the business context and thorough understanding of the source documentation

It relies on human expertise but

- Semantic annotation enhances text browsing
- Keywords help identifying relevant text fragments

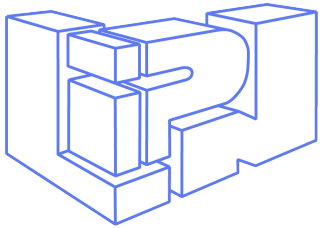


How to guide rule selection?

Rule selection requires good knowledge of the business context and thorough understanding of the source documentation

It relies on human expertise but

- Semantic annotation enhances text browsing
- Keywords help identifying relevant text fragments
  - Audi use case
    - Single keyword: 100% of recall, 80% of precision
    - Combination of keywords (*shall + if*): 97% of recall, 95% of precision

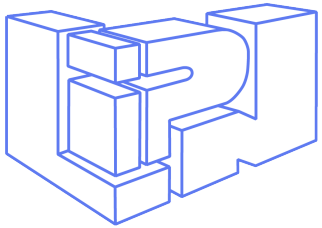


How to guide rule selection?

Rule selection requires good knowledge of the business context and thorough understanding of the source documentation

It relies on human expertise but

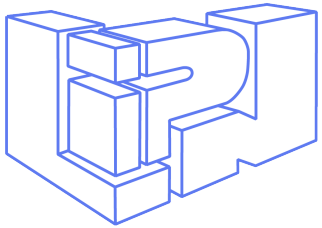
- Semantic annotation enhances text browsing
- Keywords help identifying relevant text fragments
- The expert can run Sparql queries combining regular expressions and semantic tags



SPARQL queries

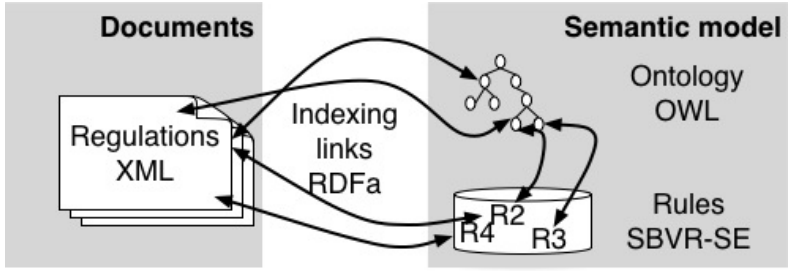
Find the sentences containing at least one of the keywords shall or if and annotated by the concept Upgrade

```
PREFIX schema: <http://lipn.univ-paris13.fr/RCLN/schema#>
PREFIX AA: <http://lipn.univ-paris13.fr/RCLN/terminae/AAdvantage#>
select DISTINCT ?sentence ?text
where{
?sentence rdf:type schema:Sentence.
?sentence schema:content ?text.
filter (?text ~ "␣shall␣" || ?text ~ "␣if␣")
?sentence schema:annoted ?textlink.
?textlink schema:defineResource ?resource2.
?resource2 schema:realizeConcept AA:Upgrade
}
```

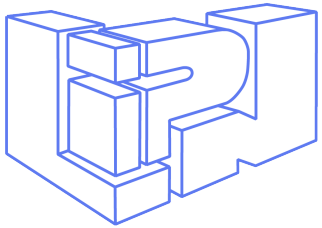


Rule selection, a complex but crucial task

- Problems
  - Corpus variability: extraction patterns are not equally relevant for all corpora
  - Focus must be put on recall rather than on precision but overloading the text with annotations hinders browsing
- A crucial step for bootstrapping the annotations of rules

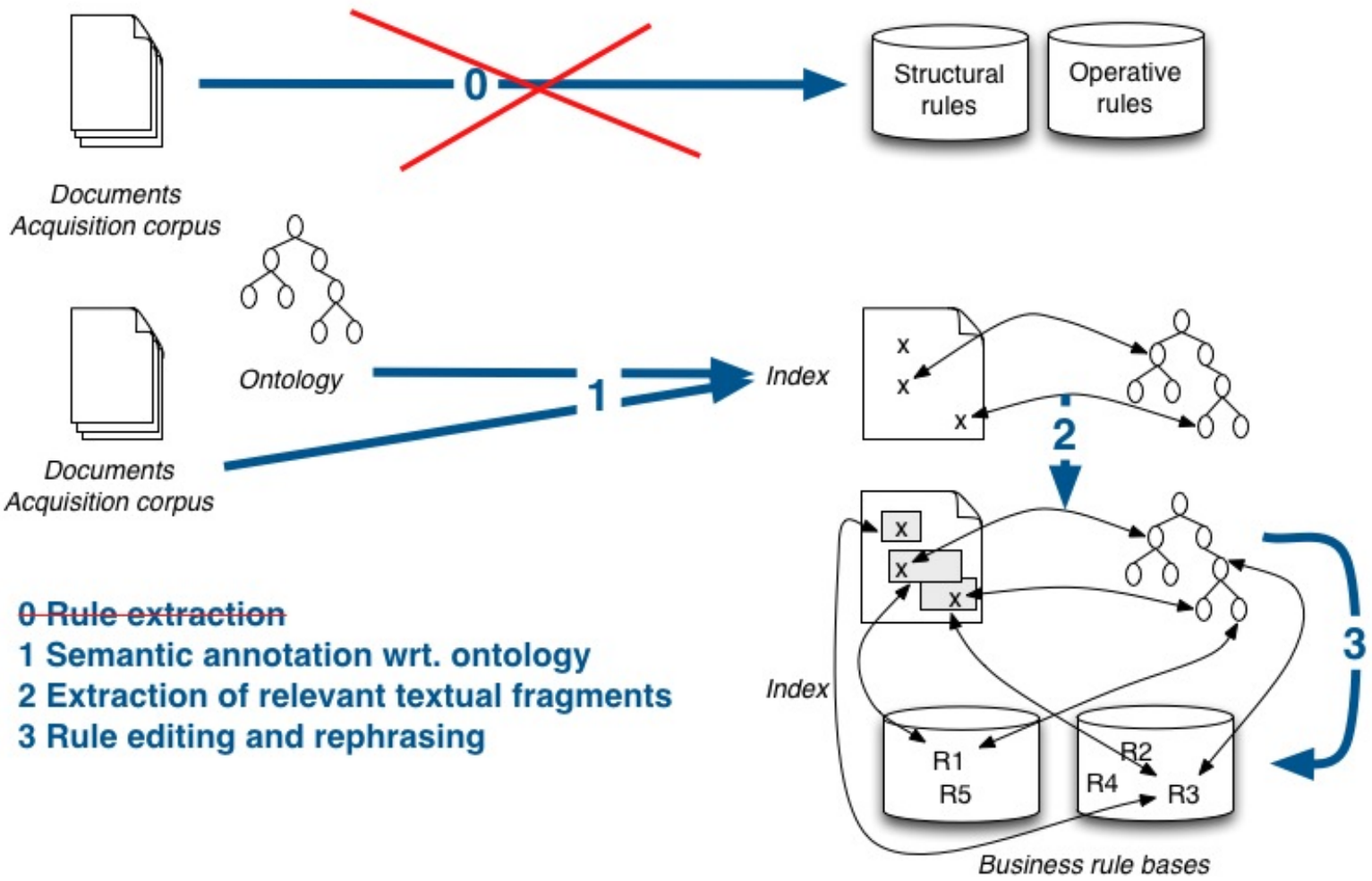


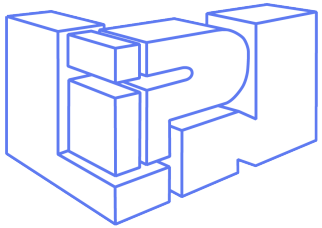
- Future work: the interactive learning of selection patterns



Outline

Goal


0
Structural rules
Operative rules
Documents Acquisition corpus
Ontology
1
Index
Documents Acquisition corpus
x
x
x
2
x
x
x
3
Index
R1
R5
R4
R2
R3
Business rule bases
0 Rule extraction
1 Semantic annotation wrt. ontology
2 Extraction of relevant textual fragments
3 Rule editing and rephrasing



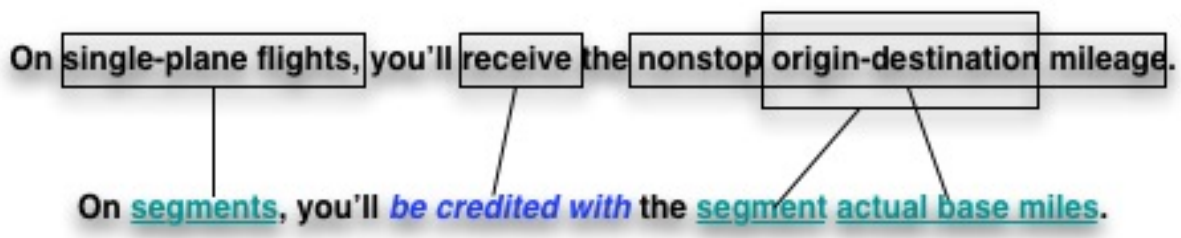
Lexical normalization

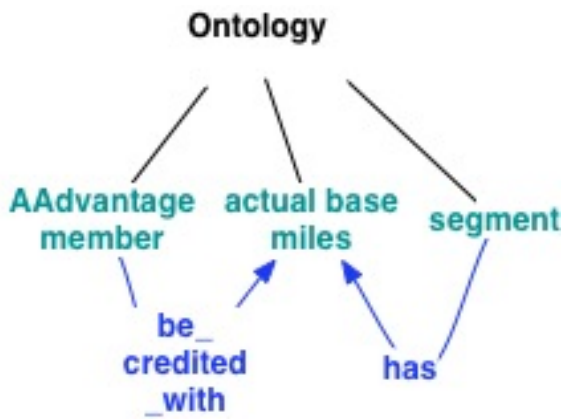
Goal

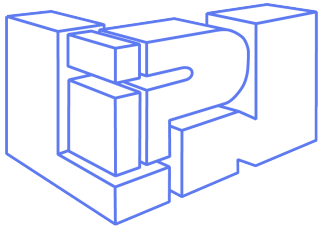
- Checking the domain vocabulary of a candidate rule
- Replacing all the mentioned terms by their preferred forms
- Disambiguating the ambiguous terms

Method

- Automatic semantic annotation
- Manual revision







Decontextualization

Co-reference links must be broken and the actual referent be made explicit so that the rules can be understood independently of their context

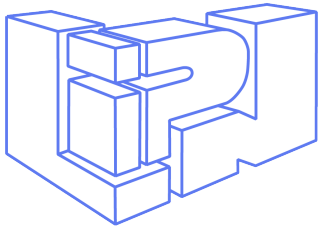
- Pronouns

  *All the adjustment devices shall undergo a strength test [. . .]. **They** must not break [. . .].*

- Generic business terms

  *The samples to be submitted to the micro-slip test [. . .]. **The test** shall be carried out at a temperature [. . .].*

- Reference keys

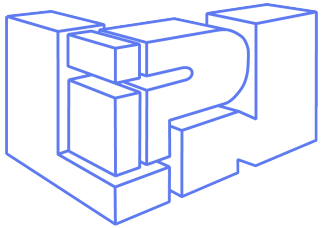


Syntactic normalization

- Sentence reordering

  *Upgrades **are void if** sold for cash or other consideration.*
  ***If** upgrades are sold for cash or other consideration, these upgrades **are void**.*

- Splitting enumerations
- Splitting rules

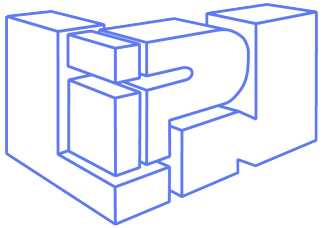


Semantic restoration

Due to decontextualization or syntactic normalization, some implicit discourse entities have to be restored

- Restoring an entity to solve a reference
- Restoring an interval to express constraints

> *The breaking load shall be determined within 5 minutes* ***after*** *the strap is removed from [. . .].*
> ***The determination time*** *is the time when the breaking load is determined. The* ***removing time*** *is the time when the strap is removed from [. . .]. The* ***delay between the removing time and the determination time*** *will be less than 5 minutes.*



Use cases

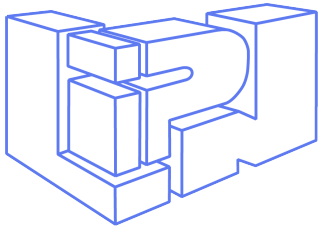
Detailed analysis on 2 candidate rule samples

- American Airlines: 95 candidate rules

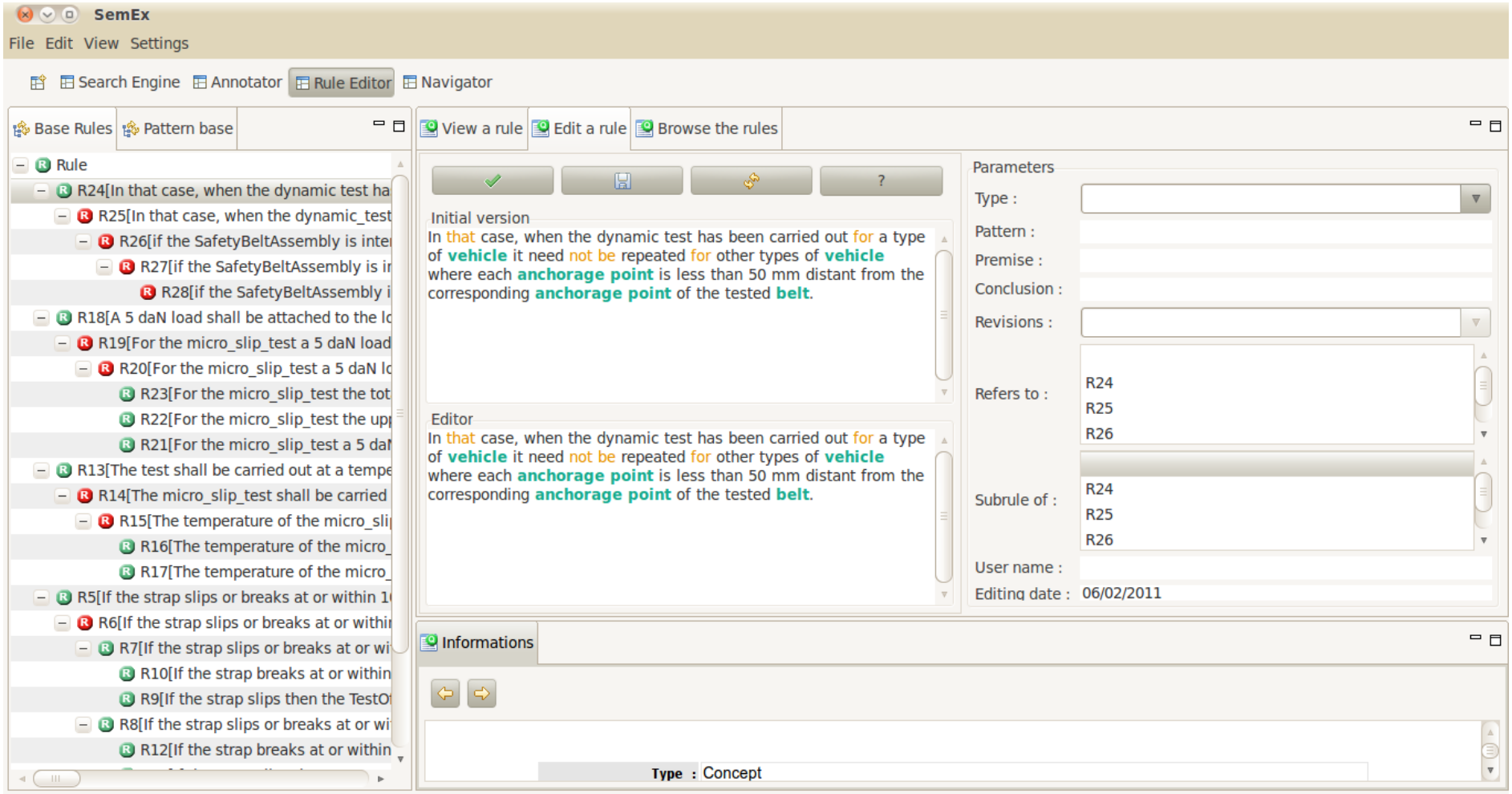
Normalization operations	% of normalized candidate rules
Lexical	65%
Contextual	64%
Syntactic	100%
$\rightarrow$*Decomposition*	*30%*
$\rightarrow$*Restructuration*	*68%*

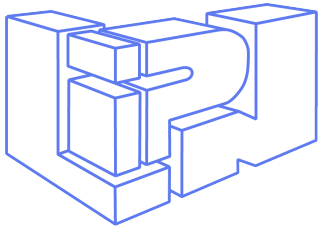
- Audi : 100 candidate rules

Normalization operations	% of normalized candidate rules
Lexical	61%
Contextual	57%
Syntactic	100%
$\rightarrow$*Decomposition*	*40%*
$\rightarrow$*Restructuration*	*32%*

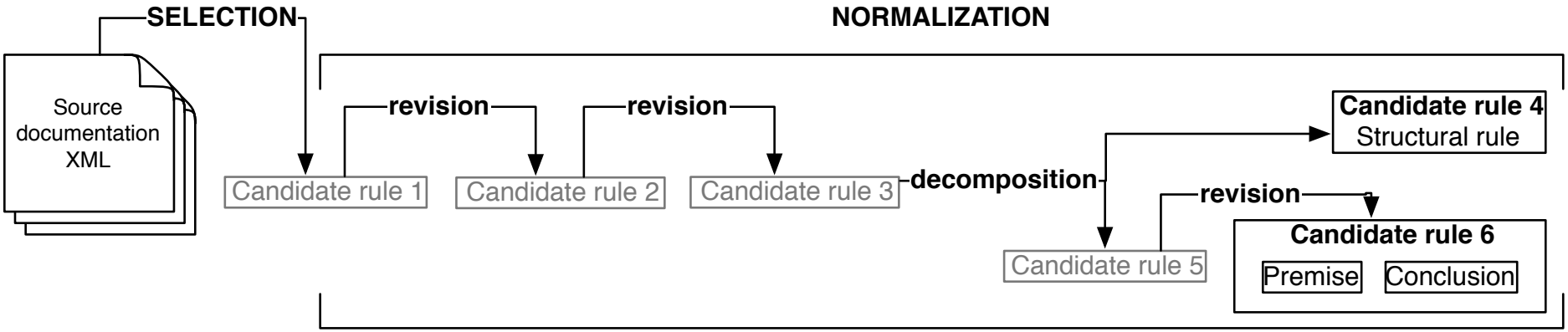


SemEx rule editor perspective





An iterative process



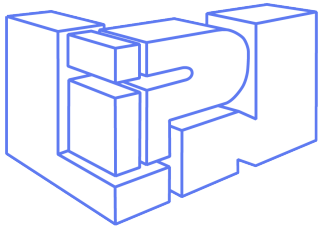
Input A set of textual fragments extracted from the source text (NL)

Process A sequence of normalization operations applied on each fragment (Human control)

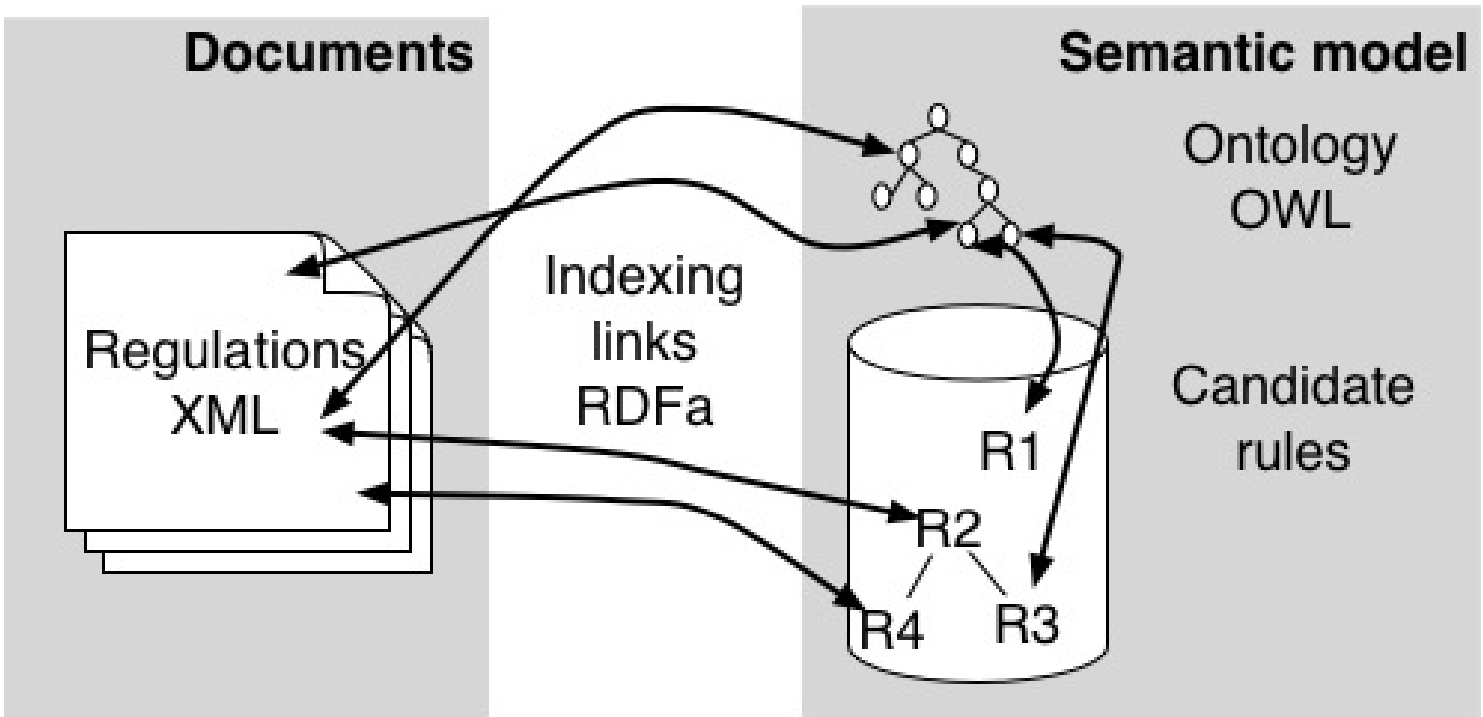
Output A set of rule statements that are independent, decontextualized, unambiguous, (possibly) structured into premise and conclusion

A derivation tree of candidate rules (Controlled Language)

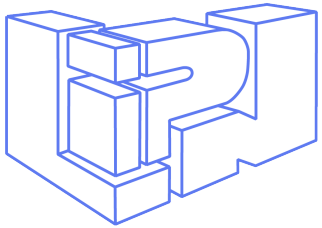
IF a test has a duration less than 6 hours, THEN the test is InvalidTest



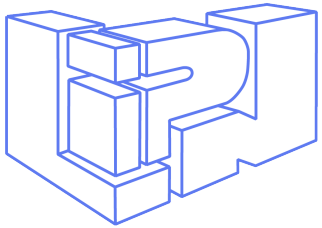
Underlying index structure



A documented rule model



Outline

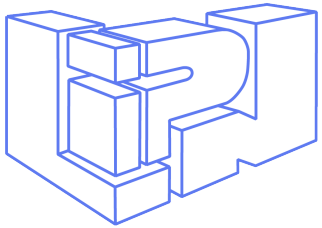


Further semantic transformation

The business expert in charge of the normalization of rules often cannot achieve their semantic transformation

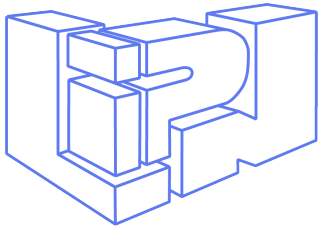
- His/her job is to clarify the business knowledge
- He/She is usually not aware of the details of
  - the implementation language
  - the way the rules must be finally encoded to be machine processable

New semantic operations are required to describe the real business processes (expected *vs.* deviant)



Decision rules

- Taking a decision = choosing an action in a set of possible actions
- A decision rule
  1. the set of actions among which one must be chosen
  2. the triggering conditions of the decision rule
  3. the action to be undertaken



Decision formalization method

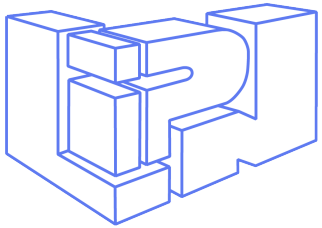
- Approach
  - Splitting problems
  - Relying on decision variables
- A two-step process
  1. Create a **detection rule** stating that a decision has to be taken

     Premise: same conditions as a standard rule

     Conclusion: a decision variable
  2. Create a **decision rule** associating a decision to a decision variable

     Premise: a decision variable + (specific sub-conditions)

     Conclusion: an action



Example

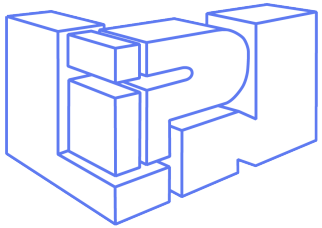
Candidate rule The car must *be returned* at the return branch

Detection rule If the drop-off location of a rental is not the return branch of the rental , the rental is *elsewhere-returned*

Decision variable the rental is *elsewhere-returned*

Decision rule If the rental is *elsewhere-returned* then ACTION

- Breaking: “cancel the rental” or “end the rental”
- Re-trying: ask the renter to drop-off in the right place
- Repairing: “charge a penalty”



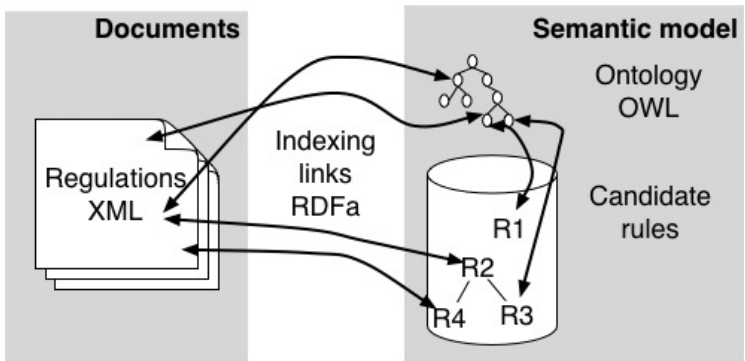
Consistency checking

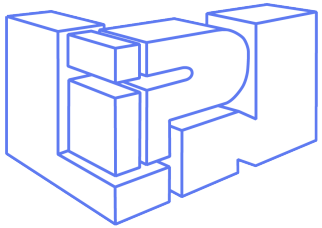
Non process is error prone

- Extraction: long-distance dependencies
- Normalization: interpretation incosistencies
- Formalization: formal inconsistencies [Fink *et al.*, 2012]

The index structure supports consistency analysis [Nazarenko and Lévy, 2013]

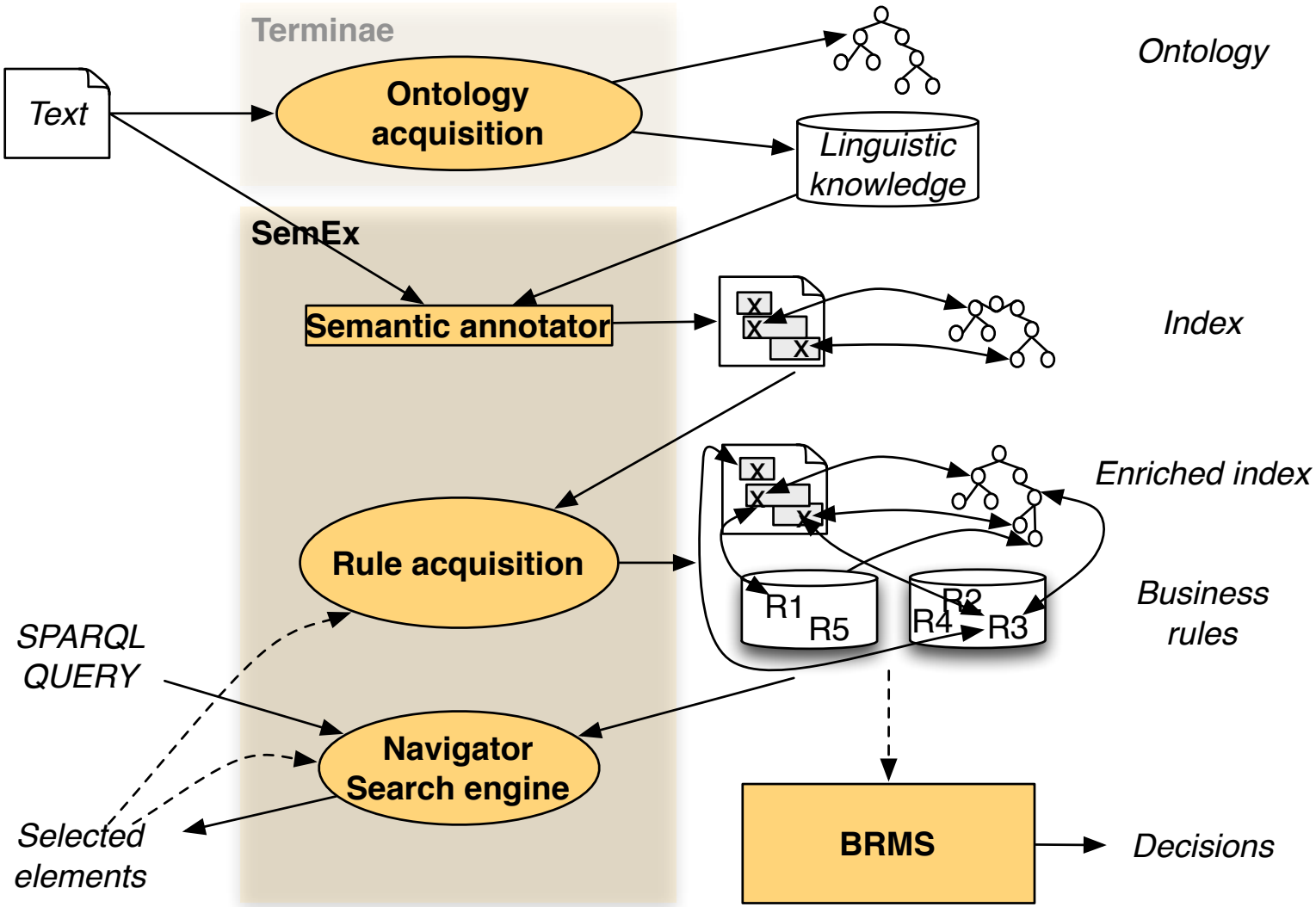
- Select all the candidate rules with the concept C in the premise
- Select all the candidate rules derived form sentences with the word W
- Select all the candidate rules derived from the candidate rule CR

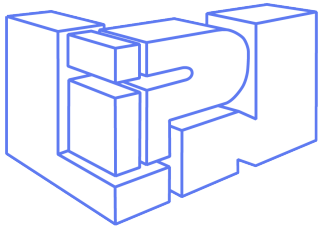




Outline

Overall architecture





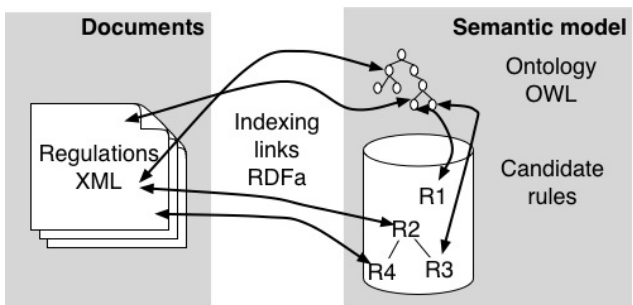
Integrating LN sources in decision systems

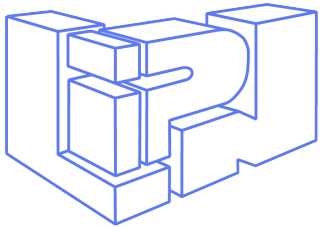
Of course, human expertise is required for

- Understanding the business case (documentation & target application)
- Building the relevant domain ontology [Omrane *et al.*, 2011b]
- Selecting the relevant rule fragments in the source documentation
- Rephrasing those NL fragments into CL and formal statements
- Modeling and formalizing the candidate rules wrt. the target application

but the Documented rule model integrates NL sources in rule systems

- Acquisition of rule that are anchored in source NL regulations
- Traceability of the rule base and system decisions
- Joint maintenance of the NL regulations and the knowledge base

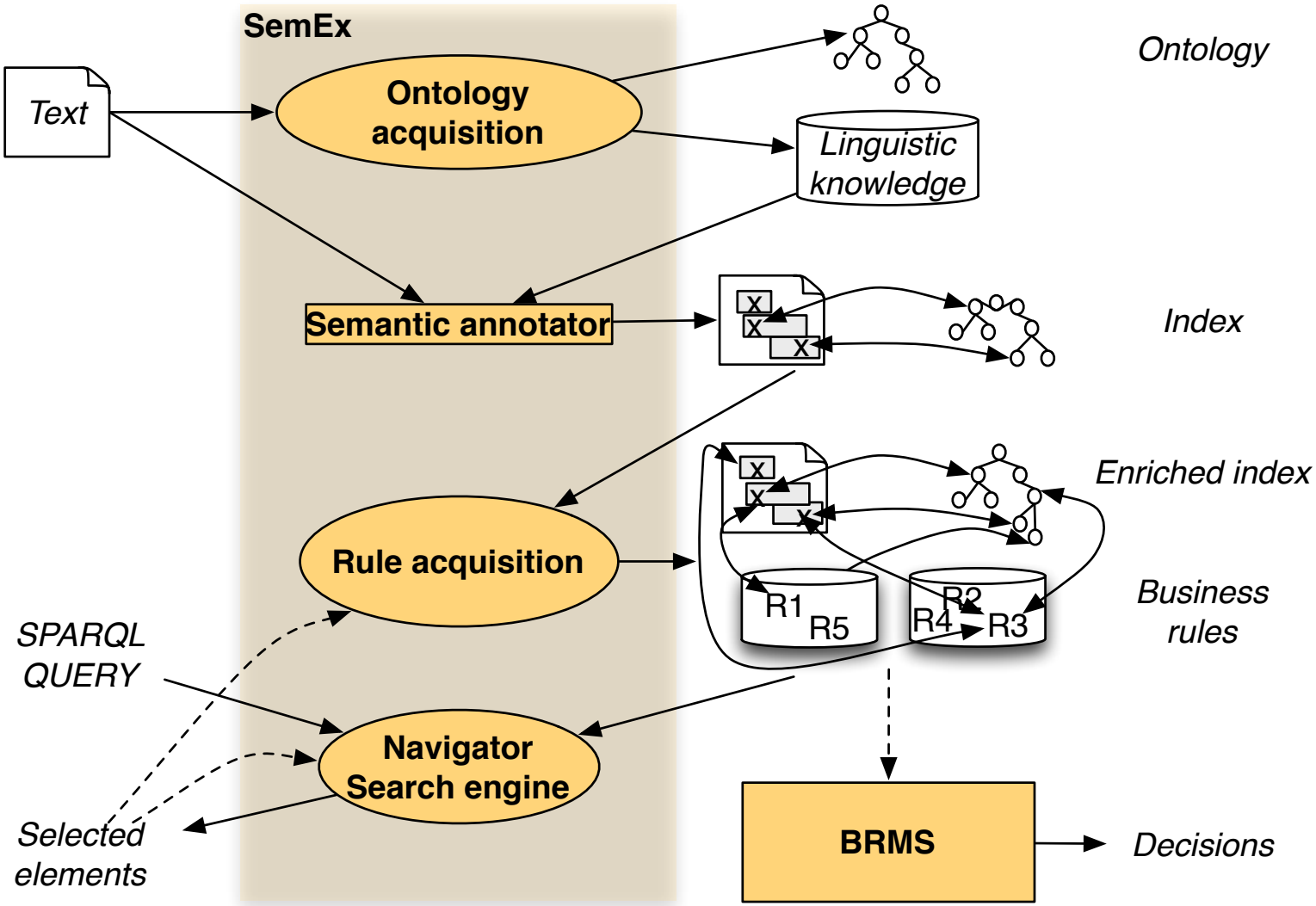


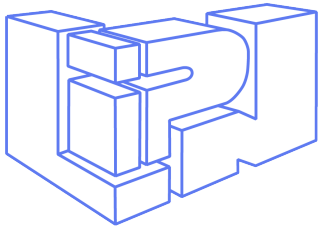


Further assistance to formalization

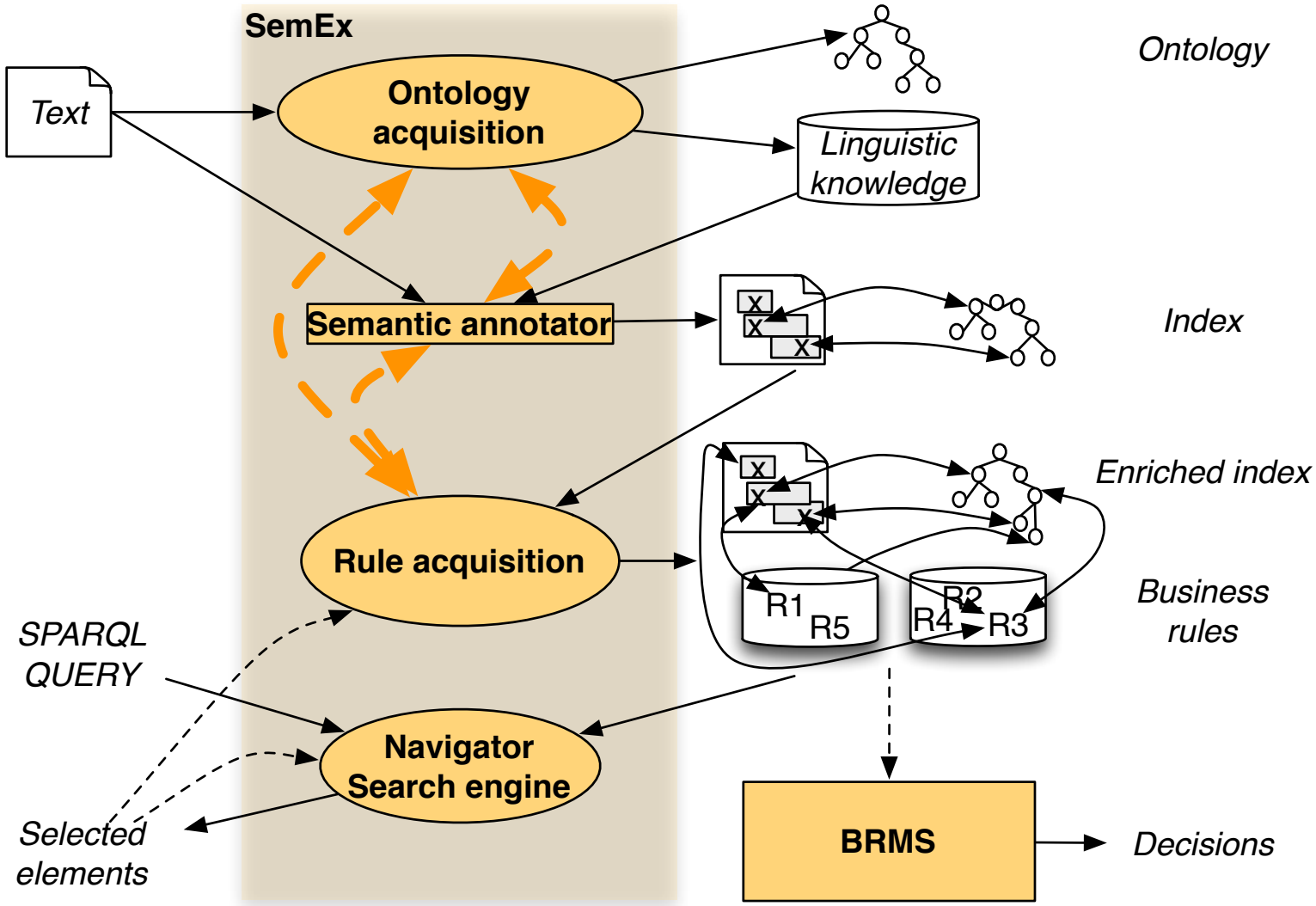
- Current SemEx platform
  - Overall methodology
  - Set of normalization operations
  - Semantic and keyword annotation
  - Interfaces
- Future work
  - Integration of NLP tools
    - Anaphora detection and resolution
    - Syntactic pattern recognition
    - Syntactic transformation
  - Stronger controlled language
    - Syntactic validation of the resulting candidate rules
    - Semantic conformance wrt. the underlying semantic model
  - Dynamic updating of the ontology
  - Machine learning of patterns

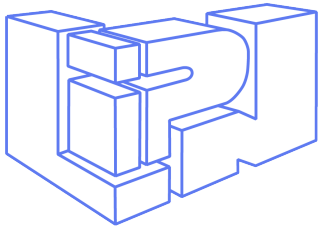
Dynamic updating of the ontology





Dynamic updating of the ontology



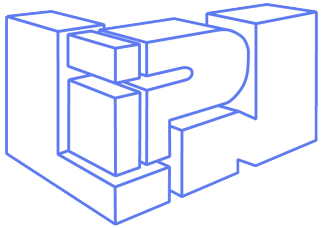


Machine learning of patterns

- Rule detection
  - It relies mainly on combinations of indices
  - Those indices and their importance vary from one corpus to another
- Rule normalization
  - The same patterns of rule are transformed in the same way
- Consistency checking

$\longrightarrow$ Towards interactive and incremental machine learning

Thank you for your attention



References I

Florence Amardeilh, Philippe Laublet, and Jean-Luc Minel.

Document annotation and ontology population from linguistic extractions.

In *Proceedings of the 3rd international conference on Knowledge capture (K-CAP '05)*, pages 161–168, New York, NY, USA, 2005. ACM.

Imran Sarwar Bajwa, Mark G. Lee, and Behzad Bordbar.

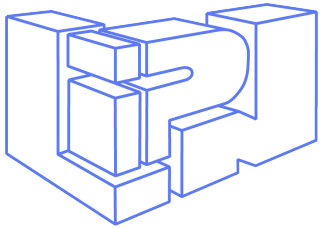
Sbvr business rules generation from natural language specification.

In *AAAI Spring Symposium 2011 Artificial Intelligence 4 Business Agility*, pages 541–545, San Francisco, USA, 2011. AAAI.

Joachim Baumeister, Jochen Reutelshoefer, and Frank Puppe.

Engineering intelligent systems on the knowledge formalization continuum.

International Journal of Applied Mathematics and Computer Science (AMCS), 21(1), 2011.



References II

Nikhil Dinesh, Aravind Joshi, Insup Lee, and Oleg Sokolski.

Reasoning about conditions and exceptions to laws in regulatory conformance checking,.

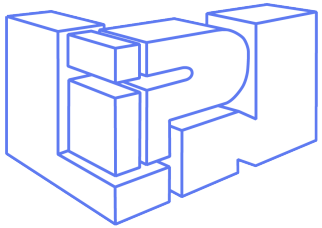
In *Proc. of the 9th Int. Conf. on Deontic Logic in Computer Science*, page 16 pages, 2008.

Michael Fink, Adil El Ghali, Amina Chniti, Roman Korf, Antonia Schwichtenberg, François Lévy, Jörg Pührе, and Thomas Eiter.

D2.6 consistency maintenance.

Final report, ONTORULE IST-2009-231875 Project, Dec. 2012.

http://ontorule-project.eu/deliverables.



References III

Abdoulaye Guissé, François Lévy, and Adeline Nazarenko.

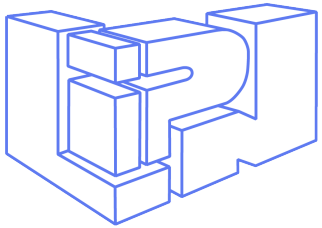
Un moteur sémantique pour explorer des textes réglementaires.

In *Actes des 22èmes journées francophones d'Ingénierie des Connaissances*, page (electronic medium), Chambéry, 2011.

François Lévy, Adeline Nazarenko, Abdoulaye Guissé, Nouha Omrane, and Sylvie Szulman.

An environment for the joint management of written policies and business rules.

In *Proceedings of the International Conference on Tools with Artificial Intelligence (IEEE-ICTAI 10)*, pages 142–149, 2010.



References IV

François Lévy, Adeline Nazarenko, and Abdoulaye Guissé.

Annotation, indexation et parcours de documents numériques.

Revue des Sciences et Technologies de l'Information (Série IDocument Numérique), 13(3/2010):121–152, December 2010.

Yue Ma, Adeline Nazarenko, and Laurent Audibert.

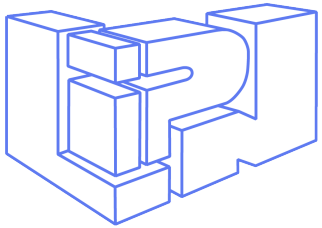
Formal description of resources for ontology-based semantic annotation.

In *Proceedings of the International Conference on Language Resources and Evaluation*, Malta, May 2010.

Yue Ma, François Lévy, and Sudeep Ghimire.

Reasoning with Annotations of Texts.

In *The 24th Florida Artificial Intelligence Research Society Conference (FLAIRS-24)*, pages 192–197, États-Unis, May 2011.



References V

Bernardo Magnini, Emanuele Pianta, Octavian Popescu, and Manuela Speranza.

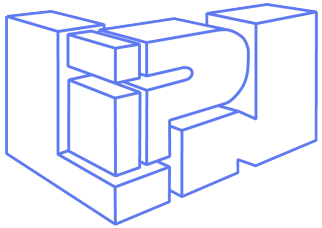
Ontology population from textual mentions: Task definition and benchmark.

In *Proceedings of the 2nd Workshop on Ontology Learning and Population: Bridging the Gap between Text and Knowledge*, pages 26–32, Sydney, Australia, July 2006. Association for Computational Linguistics.

Adeline Nazarenko and François Lévy.

Combining acquisition and debugging of business rule models.

In *Procedings of the 7th International Symposium On Rules (RuleML 2013)*, Seattle, MA, Seattle, 2013.



References VI

Nouha Omrane, Adeline Nazarenko, Peter Rosina, Sylvie Szulman, and Christoph Westphal.

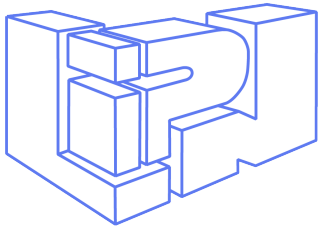
Lexicalized ontology for a business rules management platform: An automotive use case.

In *Proceedings of the 5th International Symposium on Rules, International Business Rules Forum (RuleMF@BRF)*, Ft Lauderdale, Florida, USA, November 2011.

Nouha Omrane, Adeline Nazarenko, and Sylvie Szulman.

From linguistics to ontologies : The role of named entities in the conceptualisation process.

In *Proceedings of the International Conference on Knowledge Engineering (KEOD 2011)*, Paris, Octover 2011.



References VII

- Borislav Popov, Atanas Kiryakov, Damyan Ognyanoff, Dimitar Manov, and Angel Kirilov.

  Kim – a semantic platform for information extraction and retrieval.

  Nat. Lang. Eng., 10(3-4):375–392, 2004.

- Sylvie Szulman, François Lévy, Adeline Nazarenko, and Nouha Omrane.

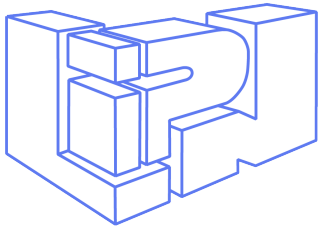
  Text-based interactive ontology acquisition tool.

  Project deliverable 1.1, ONTORULE project (FP7 2009-231875), 2009.

- Christina Unger, Lorenz Bühmann, Jens Lehmann, Axel-Cyrille Ngonga Ngomo, Daniel Gerber, and Philipp Cimiano.

  Sparql template based question answering.

  In *21st International World Wide Web Conference (WWW2012)*, April 2012.



References VIII

Maria Vargas-Vera, Enrico Motta, John Domingue, Mattia Lanzoni, Arthur Scutt, and Fabio Ciravegna.

Mnm: Ontology-driven tool for semantic markup.

In Siegfried Handschuh, Niegel Collier, Rose Dieng, and Steffen Staab, editors, *Proceedings Workshop on Semantic Authoring, Annotation & Knowledge Markup (SAAKM 2002)*, pages 43–47, Lyon, France, 2002.