



# Automatic Identification of Ontology Versions Using Machine Learning Techniques

Carlo Allocca

Mathieu d'Aquin and Enrico Motta

Knowledge Media Institute  
The Open University  
Milton Keynes, UK

# Why Do We Need This? 1/2



**Swoogle**  
search and metadata for the semantic web

**indice**



What is it? - [Submit URI](#) - [Website](#) - [Blog](#) - [APIs](#)

publication student

Found 112 semantic documents - [Search Options](#)

- 1- <http://oaci.ontologymatching.org/2004/Ccontest/303/onto.rc>
  - <http://www.aifb.uni-karlsruhe.de/ontology#public>
  - <http://www.aifb.uni-karlsruhe.de/ontology#Public>
  - <http://www.aifb.uni-karlsruhe.de/ontology#stude>
  - <http://www.aifb.uni-karlsruhe.de/ontology#Stude>
- 2- <http://swrc.ontoware.org/ontology/portal>
  - <http://swrc.ontoware.org/ontology/portal#publica>
  - <http://swrc.ontoware.org/ontology/portal#Publica>
  - <http://swrc.ontoware.org/ontology/portal#student>
  - <http://swrc.ontoware.org/ontology/portal#Student>
- 3- <http://www.ontoweb.org/ontology/1>
  - <http://www.ontoweb.org/ontology/1#publication>
  - <http://www.ontoweb.org/ontology/1#Publication>
  - <http://www.ontoweb.org/ontology/1#student>
  - <http://www.ontoweb.org/ontology/1#Student>
  - <http://www.ontoweb.org/ontology/1#PhDStudent>
- 4- <http://lsdis.cs.uga.edu/proj/sweto>
  - [http://lsdis.cs.uga.edu/proj/sweto#publication\\_ke](http://lsdis.cs.uga.edu/proj/sweto#publication_ke)
  - [http://lsdis.cs.uga.edu/proj/sweto#Publication\\_Cl](http://lsdis.cs.uga.edu/proj/sweto#Publication_Cl)
  - [http://lsdis.cs.uga.edu/proj/sweto#Scientific\\_Publ](http://lsdis.cs.uga.edu/proj/sweto#Scientific_Publ)
  - [http://lsdis.cs.uga.edu/proj/sweto#classified\\_with](http://lsdis.cs.uga.edu/proj/sweto#classified_with)
  - [http://lsdis.cs.uga.edu/proj/sweto#listed\\_author\\_i](http://lsdis.cs.uga.edu/proj/sweto#listed_author_i)
  - [http://lsdis.cs.uga.edu/proj/sweto#ACM\\_Third\\_le](http://lsdis.cs.uga.edu/proj/sweto#ACM_Third_le)
  - [http://lsdis.cs.uga.edu/proj/sweto#ACM\\_Top\\_lev](http://lsdis.cs.uga.edu/proj/sweto#ACM_Top_lev)
  - [http://lsdis.cs.uga.edu/proj/sweto#ACM\\_Second\\_](http://lsdis.cs.uga.edu/proj/sweto#ACM_Second_)
  - [http://lsdis.cs.uga.edu/proj/sweto#researcher\\_woi](http://lsdis.cs.uga.edu/proj/sweto#researcher_woi)
- 5- <http://www.aifb.uni-karlsruhe.de/WBS/meh/mapping/data/>
  - <http://swrc.org/swrc#publication>
  - <http://swrc.org/swrc#Publication>
  - [http://swrc.org/swrc#publication\\_\\_PhDStudent](http://swrc.org/swrc#publication__PhDStudent)
  - [http://swrc.org/swrc#publication\\_\\_AcademicSta](http://swrc.org/swrc#publication__AcademicSta)
  - [http://swrc.org/swrc#title\\_\\_Publication](http://swrc.org/swrc#title__Publication)
  - <http://swrc.org/swrc#student>
  - <http://swrc.org/swrc#Student>

# Why Do We Need This? 2/2

- **Versioning links are not made explicit** in results of SWSE systems.
- SWSE systems **do not provide support** for the users in exploring/**browsing versions of ontologies**.
- It is an important part of **supporting** the users in **ontology searching** that such systems should be able to identify ontology versions.



What is it? - Submit URI - Website - Blog - APIs

publication student

Found 112 semantic documents - Search Options

Search Watson

3- [http://lsdis.cs.uga.edu/projects/semdis/sweto/testbed\\_v1\\_1\\_ow/](http://lsdis.cs.uga.edu/projects/semdis/sweto/testbed_v1_1_ow/) ⓘ

- Ⓞ <http://lsdis.cs.uga.edu/proj/semdis/testbed/#Conference>
- Ⓞ [http://lsdis.cs.uga.edu/proj/semdis/testbed/#Scientific\\_Public](http://lsdis.cs.uga.edu/proj/semdis/testbed/#Scientific_Public)
- Ⓞ <http://lsdis.cs.uga.edu/proj/semdis/testbed/#Event>
- ☑ [http://lsdis.cs.uga.edu/proj/semdis/testbed/#publication\\_key](http://lsdis.cs.uga.edu/proj/semdis/testbed/#publication_key)
- Ⓞ <http://lsdis.cs.uga.edu/proj/semdis/testbed/#Publication>
- Ⓞ [http://lsdis.cs.uga.edu/proj/semdis/testbed/#Publication\\_Cla](http://lsdis.cs.uga.edu/proj/semdis/testbed/#Publication_Cla)
- Ⓞ [http://lsdis.cs.uga.edu/proj/semdis/testbed/#Scientific\\_Public](http://lsdis.cs.uga.edu/proj/semdis/testbed/#Scientific_Public)
- ☑ [http://lsdis.cs.uga.edu/proj/semdis/testbed/#classified\\_with](http://lsdis.cs.uga.edu/proj/semdis/testbed/#classified_with)

⋮

⋮



What is it? - Submit URI - Website - Blog - APIs

publication student

Found 112 semantic documents - Search Options

Search Watson

15- [http://lsdis.cs.uga.edu/projects/semdis/sweto/testbed\\_v1\\_3\\_ow/](http://lsdis.cs.uga.edu/projects/semdis/sweto/testbed_v1_3_ow/) ⓘ

- ☑ [http://lsdis.cs.uga.edu/proj/semdis/testbed/#publication\\_key](http://lsdis.cs.uga.edu/proj/semdis/testbed/#publication_key)
- Ⓞ <http://lsdis.cs.uga.edu/proj/semdis/testbed/#Publication>
- Ⓞ [http://lsdis.cs.uga.edu/proj/semdis/testbed/#Publication\\_Cla](http://lsdis.cs.uga.edu/proj/semdis/testbed/#Publication_Cla)
- Ⓞ [http://lsdis.cs.uga.edu/proj/semdis/testbed/#Scientific\\_Public](http://lsdis.cs.uga.edu/proj/semdis/testbed/#Scientific_Public)
- ☑ [http://lsdis.cs.uga.edu/proj/semdis/testbed/#classified\\_with](http://lsdis.cs.uga.edu/proj/semdis/testbed/#classified_with)
- Ⓞ <http://lsdis.cs.uga.edu/proj/semdis/testbed/#Conference>

⋮

⋮

# Problem

- **How to automatically identify ontology versions in large ontology repositories?**
  
- We reformulated the problem as follows:
  - *How to select pairs of ontologies as candidate versions?*
  
  - *How to decide whether the candidate pairs are version or not?*

# How To Select Candidate Version of Ontologies?

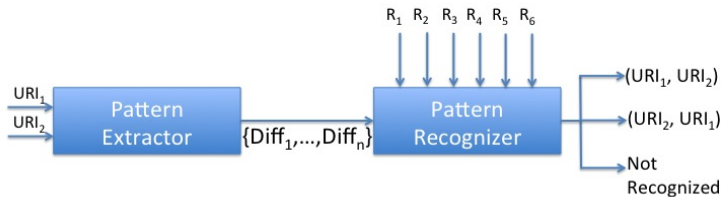
- **The starting point is:** versioning information is often encoded in the identifier of the ontologies (their URIs) using a variety of conventions and formats (**versioning information patterns**).

# Identifying Versioning Information Patterns 1/3

- **Class A: version information expressed by one number:**
  - ① <http://www.vistology.com/ont/tests/student1.owl>;
  - ② <http://www.vistology.com/ont/tests/student2.owl>;
  
  - ① [http://160.45.117.10/seweb/werdf/generate\\_timestamp\\_1176978024.owl](http://160.45.117.10/seweb/werdf/generate_timestamp_1176978024.owl)
  - ② [http://160.45.117.10/seweb/werdf/generate\\_timestamp\\_1178119183.owl](http://160.45.117.10/seweb/werdf/generate_timestamp_1178119183.owl);
- **Class B: version information expressed by two numbers**
  - ① <http://lstdis.cs.uga.edu/projects/semdis/sweto/testbedv1.1.owl>
  - ② <http://lstdis.cs.uga.edu/projects/semdis/sweto/testbedv1.4.owl>
  
  - ① <http://loki.cae.drexel.edu/wbs/ontology/2003/02/iso-metadata>
  - ② <http://loki.cae.drexel.edu/wbs/ontology/2003/10/iso-metadata>
- **Class C: version information expressed by three numbers**
  - ① <http://ontobroker.semanticweb.org/ontologies/ka2-onto-2000-11-07.daml>
  - ② <http://ontobroker.semanticweb.org/ontologies/ka2-onto-2001-03-04.daml>

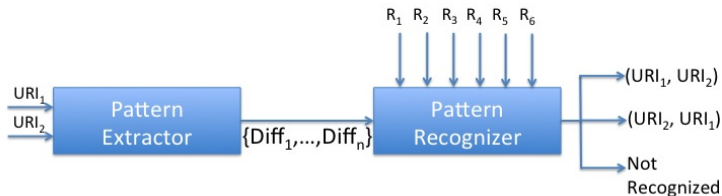
# Main Steps To Detect Candidate Versions

# Main Steps To Detect Candidate Versions





# Main Steps To Detect Candidate Versions



For example, the URIs:

- $URI_i = \text{http://loki.cae.drexel.edu/ wbs/ontology/2003/10/iso-metadata}$
- $URI_j = \text{http://loki.cae.drexel.edu/ wbs/ontology/2004/01/iso-metadata}$

are sequenced in the following way:

- **http://loki.cae.drexel.edu/ wbs/ontology/** || **2003** || / || **10** || /iso-metadata
- **http://loki.cae.drexel.edu/ wbs/ontology/** || **2004** || / || **01** || /iso-metadata

# How To Decide Whether The Candidates Are Versions Or Not?

- **The starting point is:** we define a **set of attributes** that characterised ontology versions. Based on these, we applied and compared the performances of three **Machine Learning Classifiers**:
  - *Support Vector Machine*
  - *Decision Tree*
  - *Naive Bayesian*

# Set Of Attributes: 1) Length of Chain

An example:

$\{(testbedV1.1.owl, testbedV1.3.owl), (testbedV1.3.owl, testbedV1.4.owl)\}$

the chain of ontology versions is

$\{testbedV1.1.owl, testbedV1.3.owl, testbedV1.4.owl\} \Rightarrow LC = 3$

# Set Of Attributes: 1) Length of Chain

An example:

$\{(\text{testbedV1.1.owl}, \text{testbedV1.3.owl}), (\text{testbedV1.3.owl}, \text{testbedV1.4.owl})\}$

the chain of ontology versions is

$\{\text{testbedV1.1.owl}, \text{testbedV1.3.owl}, \text{testbedV1.4.owl}\} \Rightarrow LC = 3$

**Definition 1 (Ontology Chain)** *Given set of pairs of ontologies  $\{(O_i, O_j), \dots, (O_j, O_k)\}$ , an ontology chain is defined as a sequence of ontologies,  $O_1, O_2, \dots, O_{n-1}, O_n$ , such that  $O_i$  is the previous version of  $O_{i+1}$ .*

## Set Of Attributes: 2) Similarity Measures, [Hefflin,2000]

$$\text{lexicographicSimilarity}(O_i, O_j) = \frac{|Voc(O_i) \cap Voc(O_j)|}{\max(|Voc(O_i)|, |Voc(O_j)|)}$$

where  $Voc(O_k)$  is the normalized vocabulary VOC of the ontology  $O_k$ ,  $k=i,j$ .

$$\text{syntacticSimilarity}(O_i, O_j) = \frac{|Axioms(O_i) \cap Axioms(O_j)|}{\max(|Axioms(O_i)|, |Axioms(O_j)|)}$$

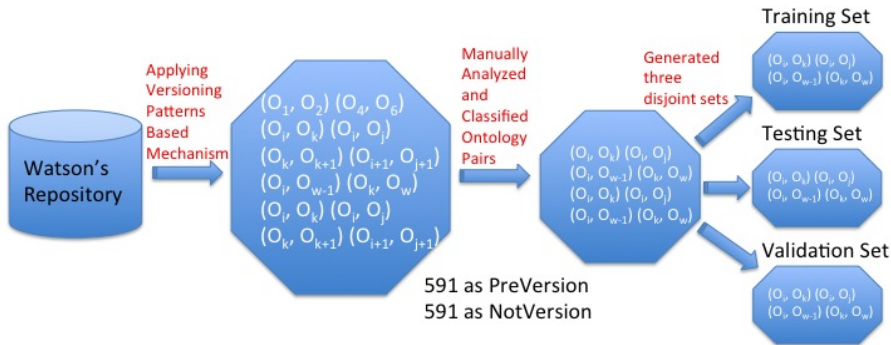
where  $Axioms(O_k)$  is the set of normalized axioms of the ontology  $O_k$ ,  $k=i,j$ .

# Set Of Attributes: 3) Ontology Versioning Pattern Rules

- R1** corresponds to the most straightforward case where there is only one numerical difference between two URIs.
- R2** corresponds to those cases where two numbers differ from one URI to the other. The version information corresponds to a version number, in which case the number on the left is more significant.
- R3,R4** correspond to those cases where two numbers differ from one URI to the other. The version information corresponds to a date including the year and month only, in which case, the year is more significant.
- R5,R6** correspond to the cases where three numerical differences exist between the considered URIs, which is the representation of dates using 3 numbers. Therefore, we define the rules corresponding to dates, either in big or little endian form.

# Applying ML Classifiers: Generating Dataset

# Applying ML Classifiers: Generating Dataset

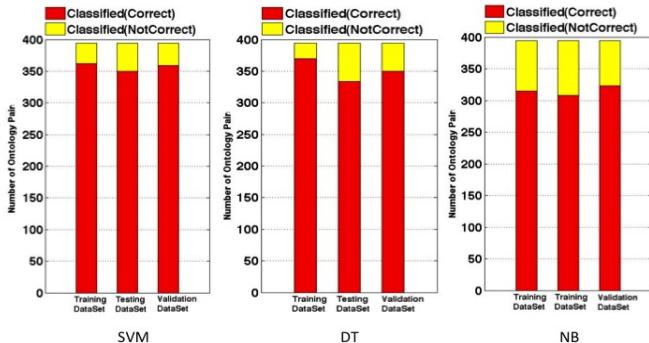




# Applying ML Classifiers: Generating Dataset

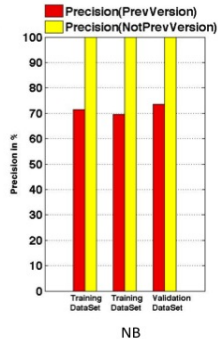
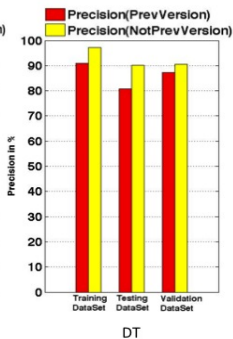
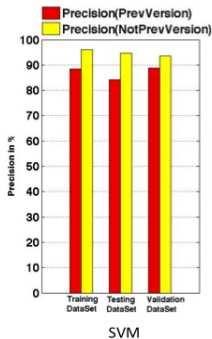
- **Training Set.** It is used during the *Training* phase that is responsible for building the classifier model  $M_T$ .
- **Testing Set.** It is used during the *Testing* phase that is responsible for checking the accuracy of the model  $M_T$ .
- **Validation Set.** It is used during the *Validation* phase that is in charge of evaluating the stability of the classifier.

# Applying ML Classifiers: Performances



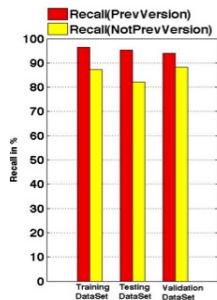
- **Classified(Correct)**. It indicates the number of tuples correctly classified.
- **Classified(Incorrect)**. It indicates the number of tuples incorrectly classified.

# Applying ML Classifiers: Performances

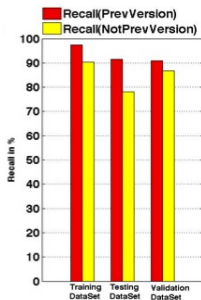


- **Precision(PrevVersion):**  $\frac{CPV \cap EPV}{CPV}$
- **Precision(NotPrevVersion).** similarly, it indicates the proportion of tuples correctly classified as *NotPrevVersion* among all those that were classified as *NotPrevVersion*

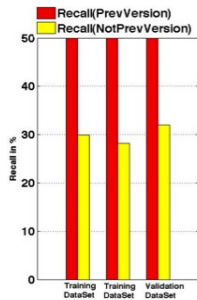
# Applying ML Classifiers: Performances



SVM



DT

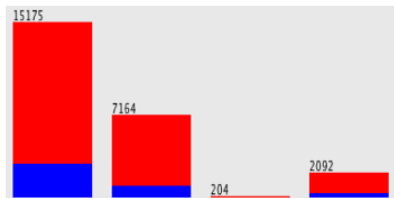


NB

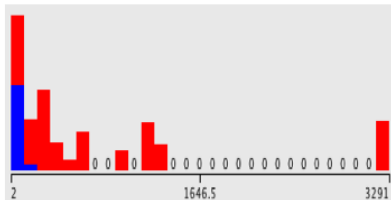
- **Recall(PrevVersion):**  $\frac{CPV \cap EPV}{EPV}$ .
- **Recall(NotPrevVersion).** It similarly, indicates the proportion of tuples classified as *NotPrevVersion*, among those that were identified as *NotPrevVersion* in our evaluation.

# Analysis Of The Attributes

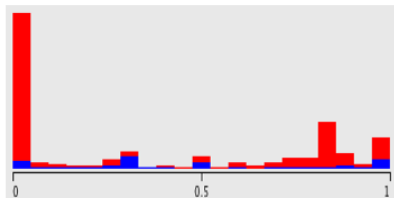
# Analysis Of The Attributes



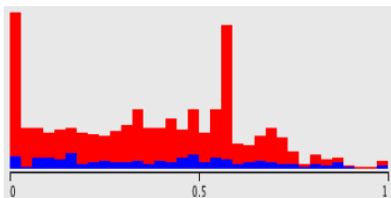
(a) Selected Attribute: *Rn*



(b) Selected Attribute: *LC*



(c) Selected Attribute: *VocSim*



(d) Selected Attribute: *SynSim*

**Figure:** Blue colour represents pairs classified as versions by the Decision Tree classifier and red colour are pairs which are not recognised as versions.

# Conclusion and Future Work

- We proposed an approach to automatic identification of ontology versions based on two steps.
  - **The first step** tackles the issue of selecting pair of ontologies as candidate versions: **we applied an ontology URIs based mechanism**
  - **The second step** deals with the issue of deciding whether the selected pair of ontologies are versions or not: **we applied machine learning classifiers (SVM, precision 87%)**
- Future
  - **Empirical point of view:** discovering relevant *patterns* in ontology evolution;
  - **Practical point of view:** supporting the users in browsing the SWSE's page of results.

# Practical point of view



What is it? - Submit URI - Website - Blog - APIs

student

versions

Search Watson

Found 1080 semantic documents

- 1- <http://www.vistology.com/ont/tests/st>  
<http://www.vistology.com/ont/>  
<http://www.vistology.com/ont/>  
<http://www.vistology.com/ont/>

  - Ⓞ <http://www.vistology.com/ont/>
  - ◆ <http://www.vistology.com/ont/>

3 similar results 3 other versions
- 2- <http://www.vistology.com/ont/tests/stu>  
<http://www.vistology.com/ont/>  
<http://www.vistology.com/ont/tests/student2.owl>  
<http://www.vistology.com/ont/tests/student3.owl>

  - Ⓞ <http://www.vistology.com/ont/tests/student4.owl#Student>
  - ◆ <http://www.vistology.com/ont/tests/student4.owl>

3 similar results 3 other versions 57 results from the same domain in
- 3- <http://www.cs.vu.nl/~kubbe/webkr/model.daml> [B]

  - Ⓞ <http://www.cs.vu.nl/~kubbe/webkr/model.daml#Student>
  - ▣ [http://www.cs.vu.nl/~kubbe/webkr/model.daml#has\\_student](http://www.cs.vu.nl/~kubbe/webkr/model.daml#has_student)
  - ▣ [http://www.cs.vu.nl/~kubbe/webkr/model.daml#has\\_audience](http://www.cs.vu.nl/~kubbe/webkr/model.daml#has_audience)
  - ▣ [http://www.cs.vu.nl/~kubbe/webkr/model.daml#follows\\_study](http://www.cs.vu.nl/~kubbe/webkr/model.daml#follows_study)
  - ▣ [http://www.cs.vu.nl/~kubbe/webkr/model.daml#Examination\\_m\\_1](http://www.cs.vu.nl/~kubbe/webkr/model.daml#Examination_m_1)
  - ▣ [http://www.cs.vu.nl/~kubbe/webkr/model.daml#has\\_course\\_code](http://www.cs.vu.nl/~kubbe/webkr/model.daml#has_course_code)
  - ▣ [http://www.cs.vu.nl/~kubbe/webkr/model.daml#has\\_to\\_complete](http://www.cs.vu.nl/~kubbe/webkr/model.daml#has_to_complete)
  - ▣ <http://www.cs.vu.nl/~kubbe/webkr/model.daml#Studies>
  - ▣ <http://www.cs.vu.nl/~kubbe/webkr/model.daml#Coursecode>
  - ▣ [http://www.cs.vu.nl/~kubbe/webkr/model.daml#has\\_completed](http://www.cs.vu.nl/~kubbe/webkr/model.daml#has_completed)

More...

47 results from the same domain
- 4- <http://www.cs.vu.nl/~jkieviet/web/model.daml> [B]

  - Ⓞ <file:///H:/www/web/ideaal.daml#Student>
  - ▣ [file:///H:/www/web/ideaal.daml/krijgt\\_les\\_in](file:///H:/www/web/ideaal.daml/krijgt_les_in)
  - ▣ [file:///H:/www/web/ideaal.daml/krijgt\\_les\\_van](file:///H:/www/web/ideaal.daml/krijgt_les_van)
  - ▣ [file:///H:/www/web/ideaal.daml/geeft\\_les\\_aan](file:///H:/www/web/ideaal.daml/geeft_les_aan)
  - ▣ [file:///H:/www/web/ideaal.daml/info\\_uitgever](file:///H:/www/web/ideaal.daml/info_uitgever)
  - ▣ <file:///H:/www/web/proefje.daml#nummer>
  - ▣ <file:///H:/www/web/ideaal.daml#letter>
  - ▣ <file:///H:/www/web/ideaal.daml#literatuur>
  - ▣ [file:///H:/www/web/ideaal.daml#is\\_schrijver\\_van](file:///H:/www/web/ideaal.daml#is_schrijver_van)
  - ▣ <file:///H:/www/web/ideaal.daml#bepaalt>

More...

1 similar results 47 results from the same domain
- 5- <http://www.ontoweb.org/ontology/1> [B]

  - ▣ <http://www.ontoweb.org/ontology/1#student>
  - Ⓞ <http://www.ontoweb.org/ontology/1#Student>
  - Ⓞ <http://www.ontoweb.org/ontology/1#PhDStudent>

9 similar results





# Thank You!

[c.allocca@open.ac.uk](mailto:c.allocca@open.ac.uk)

Automatic Identification of Ontology Versions Using Machine Learning Techniques

