# Transforming an OWL Ontology to an OpenOffice Document Template

▶ **Massoud Toussi, MD, Msc, PhD Candidate**

▶ Medical Informatics and Bioinformatics Laboratory (LIM&BIO EA 3969), UFR SMBH, Université Paris 13, Bobigny, France
Formerly: Physician, Editor-in-Chief of Spectrum Medical Journal

# Topics

- Writing and retrieving scientific papers

- The potential advantage of Open Document Format (ODF) for semantic content management

- Paper Element Nodes (PEN): an ontology of scientific papers in the domain of biomedical science written in Web Ontology Language (OWL) using *Protégé*

- Transforming PEN to an OpenOffice Template

# Writing Scientific Papers

▶ Not a trivial task

▶ The most difficult thing is structuring the ideas

▶ In biomedical science: IMRaD:

 ▶ Introduction
 ▶ Methods
 ▶ Results
 ▶ Discussion

▶ Typing, editing, spell checking, defining styles, layout, inserting tables, figures, managing references, etc

# Retrieving Scientific Papers

▶ Today's scientific search engines are capable of searching inside the documents : Google Scholar, or PubMed or EmBase

▶ This search does not take into account the semantic structure of the articles, for example:

  ▶ When searching for Python language, there are articles about the snake
  ▶ When searching for Java language, there are articles about coffee, and the province
  ▶ When searching about pain, there are articles about bread, because pain in French means bread

# Two kinds of solutions

- Making more sophisticated search engines able to « understand » the documents

  - Dependency to native languages

  - The complexity of the solution, false negatives

- Making documents more « understandable » by integrating the semantic structure of articles into documents

  - Easy to do

  - More efficient

# Today's Office Suits

▶ Facilitate writing of scientific papers in many ways:

 ▶ Typing and formatting the text layout
 ▶ Drawing graphics, and making tables
 ▶ Spell checking
 ▶ Tracing revisions
 ▶ Making bibliographies

▶ It will be interesting to make tools for

 ▶ helping authors organize their ideas
 ▶ helping search engines find portions of articles using their semantic structure

# The Open Document Format (ODF)

- Is based on XML format
- Is an ISO/IEC International Standard
- Shared by many office suites, including:

  - OpenOffice.org
  - Koffice
  - Google Docs
  - NeoOffice
  - Zoho
  - IBM Lotus Symphony
  - Corel WordPerfect Office X4

# PEN Ontology

▶ PEN (Paper Element Nodes) is an ontology of scientific papers in biomedical domain which describes the way different parts (elements) of an article are grouped in different sections (nodes) by use of OWL language.

▶ What is an ontology?

   ▶ A formal representation of a set of concepts within a domain and the relationships between those concepts. It is used to reason about the properties of that domain, and may be used to define the domain.

# OWL (Web Ontology Language)

▶ A family of knowledge presentation languages for authoring ontologies

▶ One of the fundamental technologies of Semantic Web which is based on XML/RDF Schema
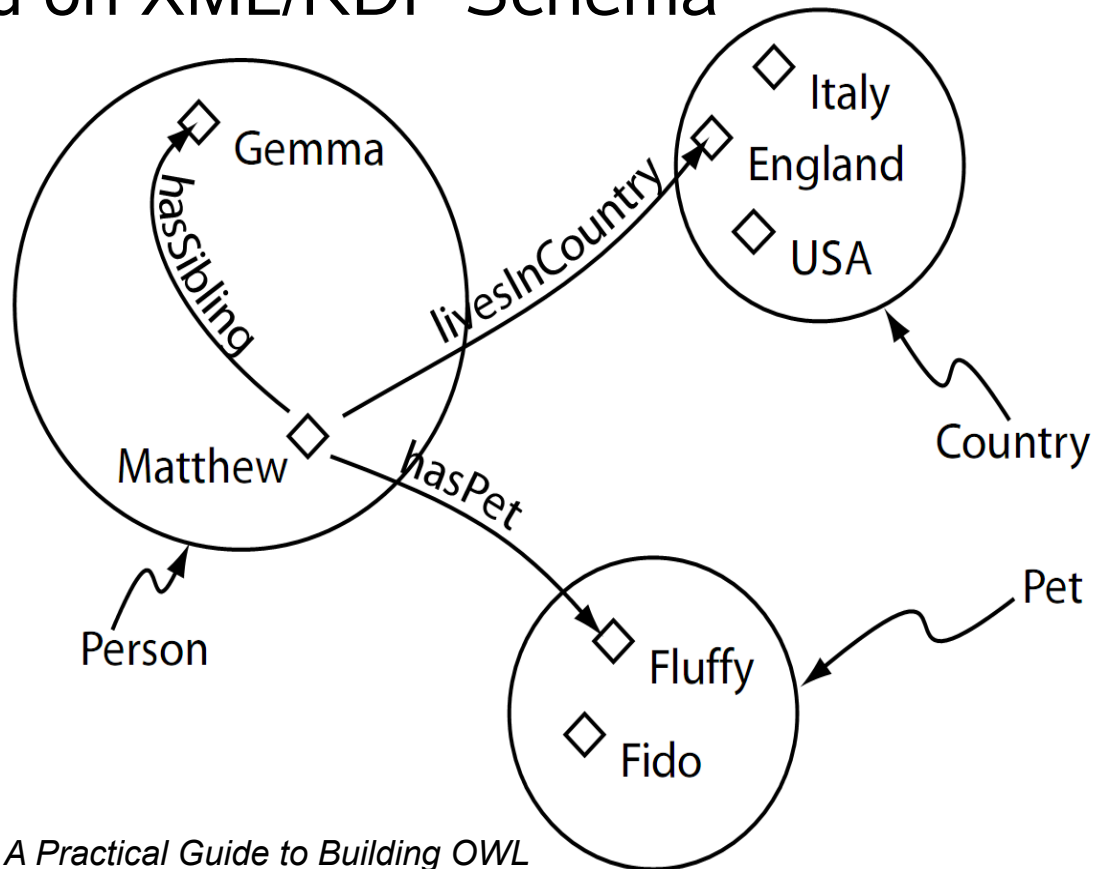
# The formation of PEN Ontology

▶ Based ***Uniform Requirements for Articles Submitted to Biomedical Journals***

▶ Completed with other methodological articles to define a **maximum data set**

▶ Defining an ontology based on this maximum dataset

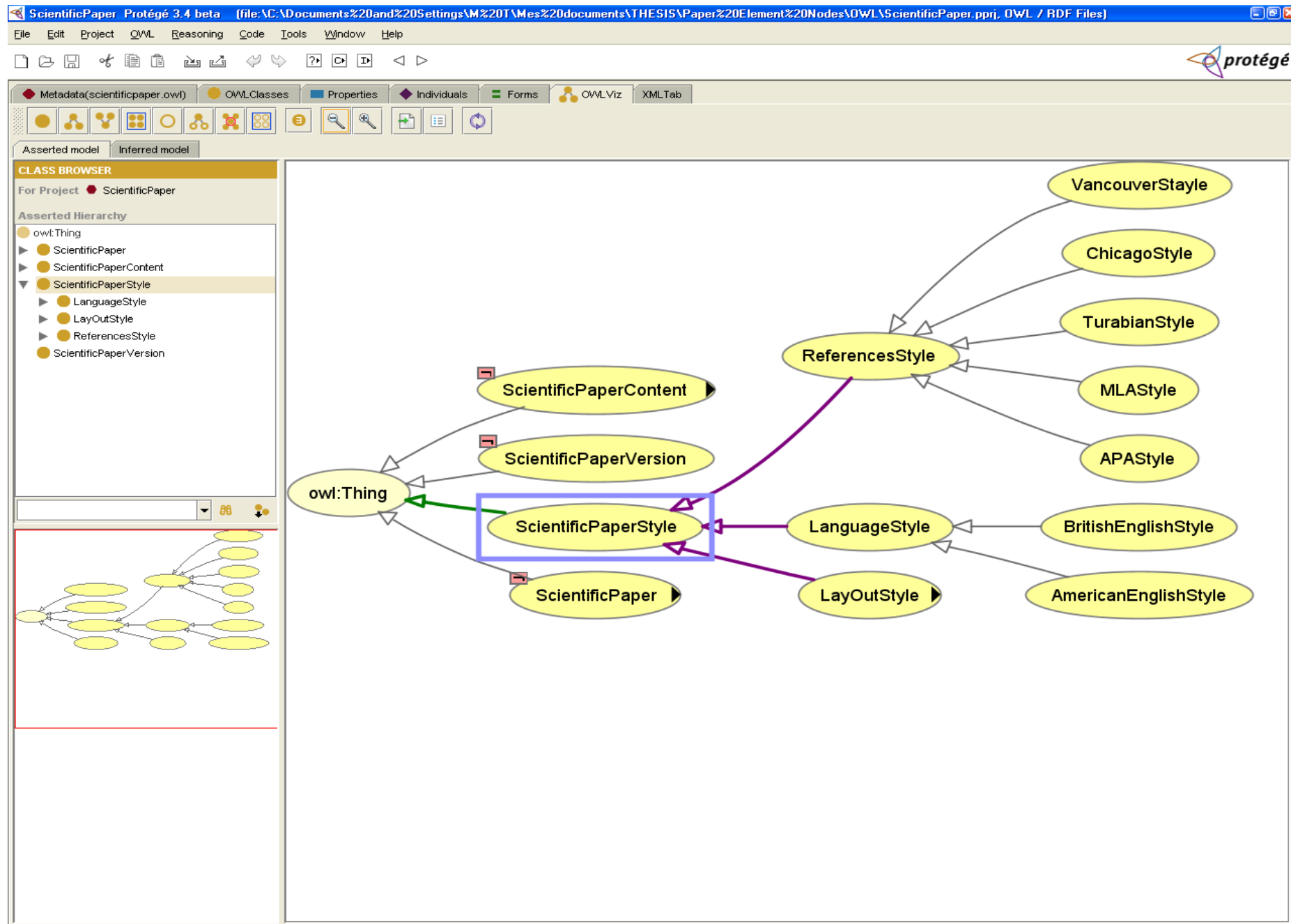   ▶ Demo: PEN in *Protégé*

# PEN Ontology Main Classes
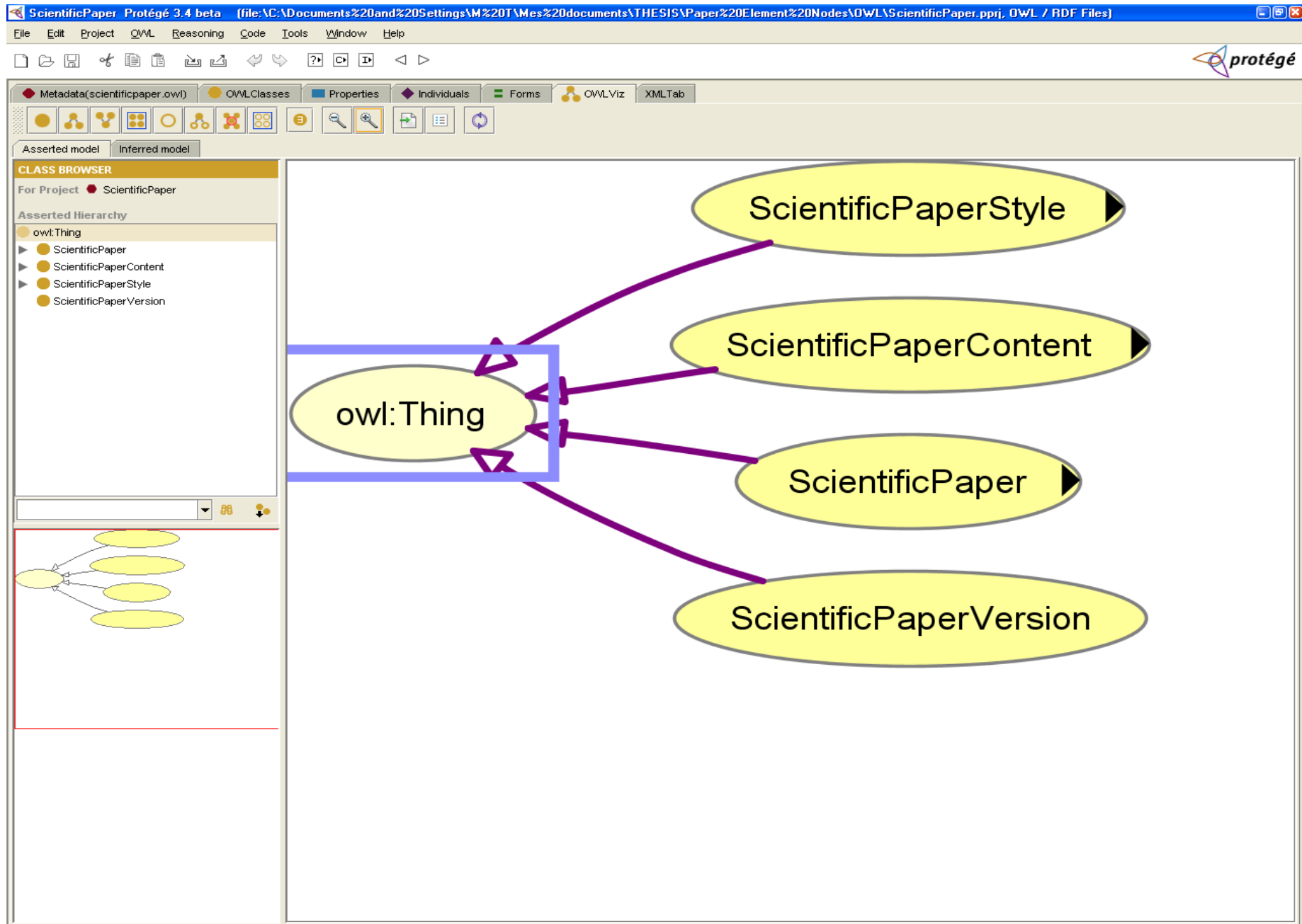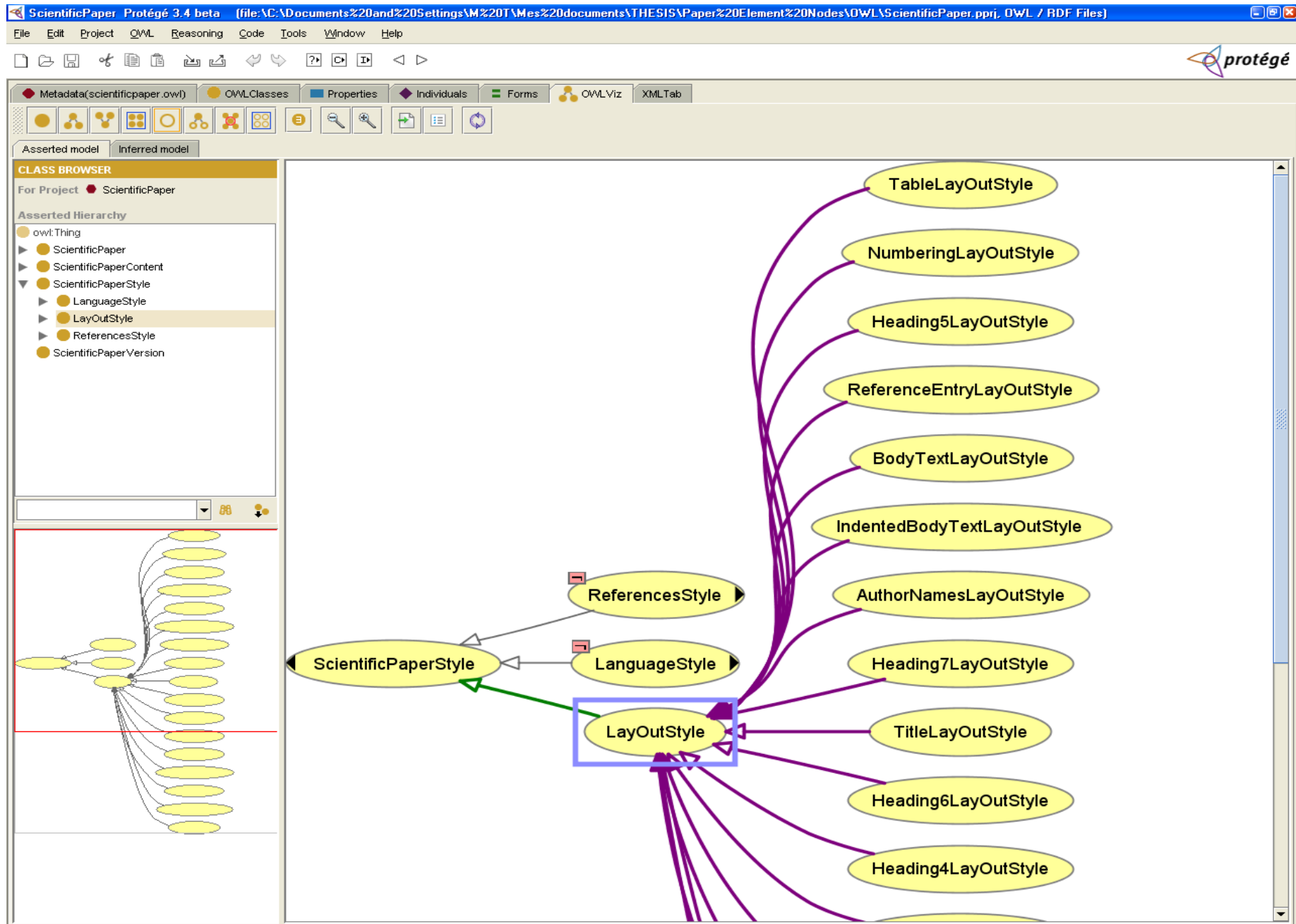
# PEN Ontology Properties

# Scientific Paper Class

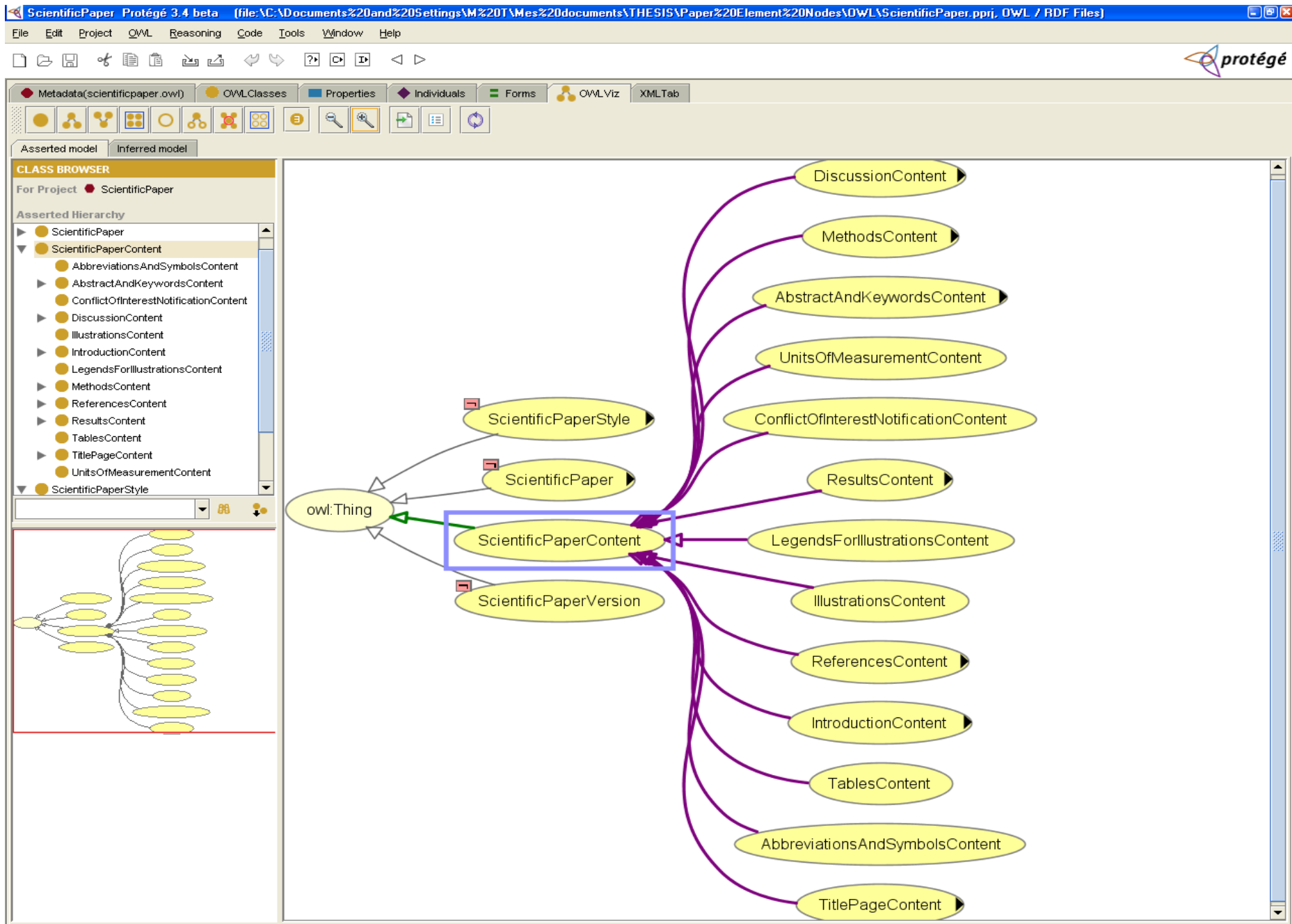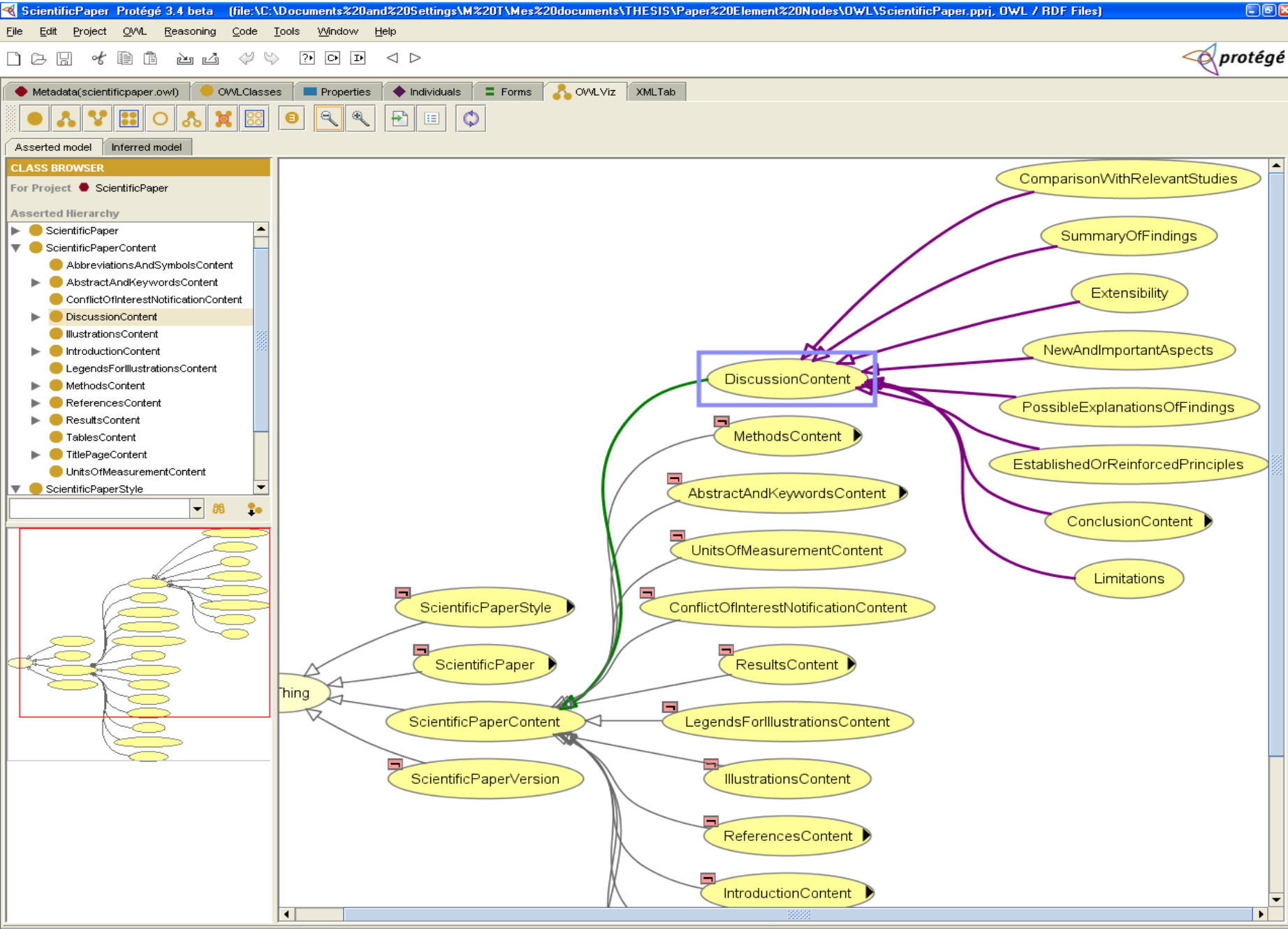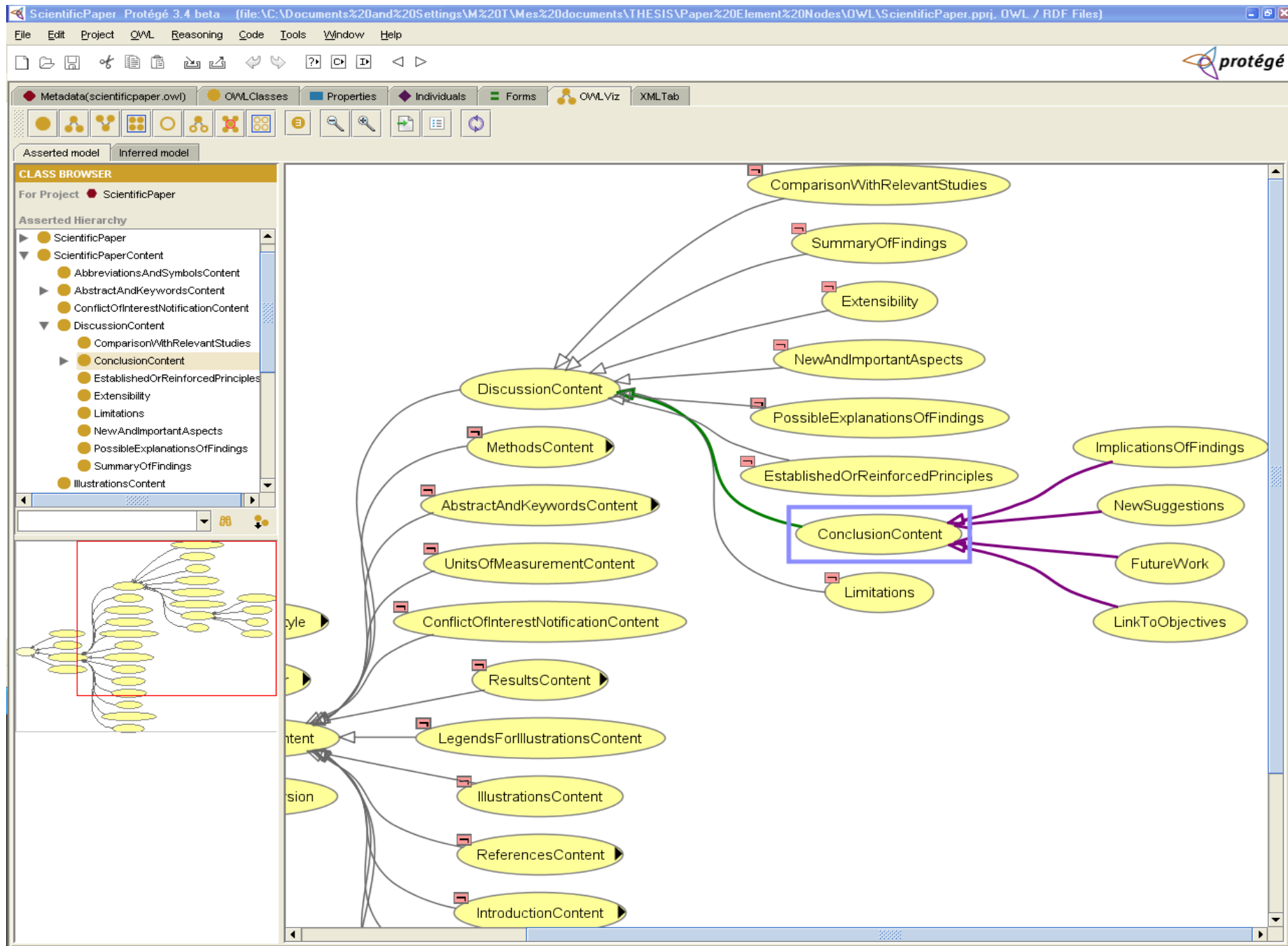# Scientific Article Styles Class

# Scientific Article Layout Styles

# Scientific Paper Content Class

# Discussion Content Class

# Conclusion Content Class

# Title Page Content Class

# Exporting PEN in RDF

# Details of PEN's OWL / RDF File

```
</owl:Class>
-<owl:Class rdf:about="#Limitations">
  -<owl:disjointWith>
      <owl:Class rdf:about="#SummaryOfFindings"/>
  </owl:disjointWith>
  -<owl:disjointWith>
      <owl:Class rdf:about="#NewAndImportantAspects"/>
  </owl:disjointWith>
  -<owl:disjointWith>
      <owl:Class rdf:about="#EstablishedOrReinforcedPrinciples"/>
  </owl:disjointWith>
      <owl:disjointWith rdf:resource="#ComparisonWithRelevantStudies"/>
      <owl:disjointWith rdf:resource="#PossibleExplanationsOfFindings"/>
      <owl:disjointWith rdf:resource="#Extensibility"/>
  +<owl:disjointWith></owl:disjointWith>
      <rdfs:subClassOf rdf:resource="#DiscussionContent"/>
</owl:Class>
+<owl:Class rdf:about="#TableLayOutStyle"></owl:Class>
-<owl:Class rdf:about="#AuthorNamesLayOutStyle">
  -<owl:disjointWith>
      <owl:Class rdf:about="#Heading3LayOutStyle"/>
  </owl:disjointWith>
  -<owl:disjointWith>
      <owl:Class rdf:about="#IndentedBodyTextLayOutStyle"/>
  </owl:disjointWith>
  -<owl:disjointWith>
      <owl:Class rdf:about="#Heading1LayOutStyle"/>
  </owl:disjointWith>
  -<owl:disjointWith>
      <owl:Class rdf:about="#BodyTextLayOutStyle"/>
  </owl:disjointWith>
      <owl:disjointWith rdf:resource="#TitleLayOutStyle"/>
      <owl:disjointWith rdf:resource="#AuthorAffiliationsLayOutStyle"/>
  -<owl:disjointWith>
      <owl:Class rdf:about="#Heading7LayOutStyle"/>
  </owl:disjointWith>
      <owl:disjointWith rdf:resource="#Heading6LayOutStyle"/>
      <owl:disjointWith rdf:resource="#NumberingLayOutStyle"/>
  -<owl:disjointWith>
      <owl:Class rdf:about="#BulletLayOutStyle"/>
  </owl:disjointWith>
  -<rdfs:subClassOf>
      <owl:Class rdf:about="#LayOutStyle"/>
```
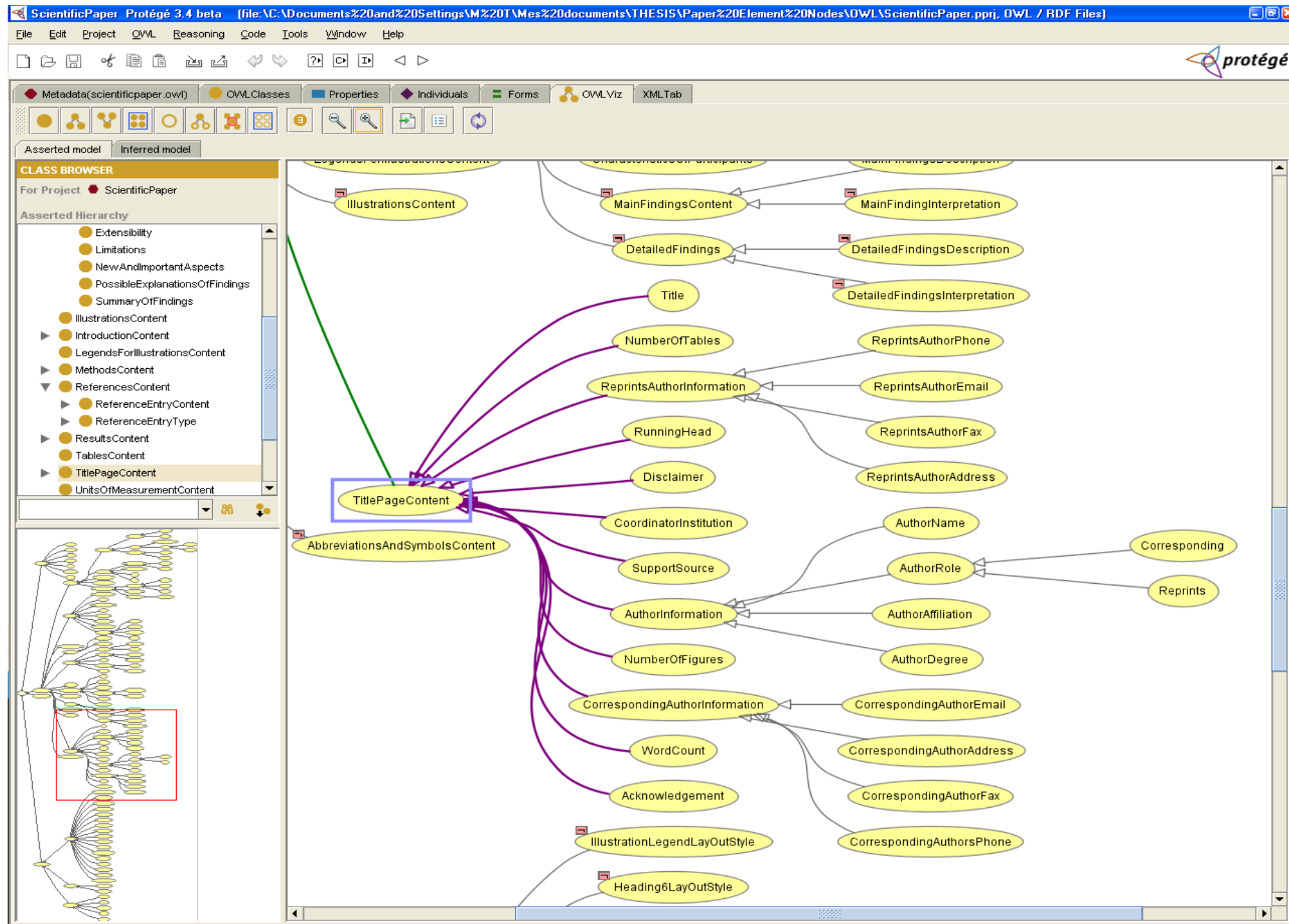
This format can be parsed by search engines but is not readable by office suits

# Making a PEN OpenOffice Template

# Content.xml file of the .ott file

```xml
-<office:document-content office:version="1.2">
   <office:scripts/>
  +<office:font-face-decls></office:font-face-decls>
  +<office:automatic-styles></office:automatic-styles>
  -<office:body>
     -<office:text text:use-soft-page-breaks="true">
        +<text:sequence-decls></text:sequence-decls>
        -<text:p text:style-name="Title">
           <text:placeholder text:placeholder-type="text" text:description="Click to insert the title of your article here. It may have the form of
           "Statement : the design""><Title [Statement: Desing]></text:placeholder>
        </text:p>
        -<text:p text:style-name="Subtitle">
           <text:placeholder text:placeholder-type="text" text:description="Click to write authors' names and degrees"><Autho's names, degrees>
           </text:placeholder>
        </text:p>
        -<text:p text:style-name="Subtitle">
           <text:placeholder text:placeholder-type="text" text:description="Click to insert authors' affiliations"><Author's affiliations>
           </text:placeholder>
        </text:p>
        <text:h text:style-name="Heading_20_5" text:outline-level="5">Corresponding author</text:h>
        -<text:p text:style-name="Text_20_body">
           <text:placeholder text:placeholder-type="text" text:description="Click to enter the correspondig author's address"><Corresponding
           author's name and address></text:placeholder>
        </text:p>
        <text:h text:style-name="Heading_20_5" text:outline-level="5">Acknowledgments</text:h>
        -<text:p text:style-name="Text_20_body">
           <text:placeholder text:placeholder-type="text" text:description="Write the name of anyone who worked on the study but not fulfilled
           the criteria of authors."><Acknowledgments></text:placeholder>
        </text:p>
        <text:h text:style-name="Heading_20_5" text:outline-level="5">Keywords</text:h>
        -<text:p text:style-name="Text_20_body">
           <text:placeholder text:placeholder-type="text" text:description="Enter up to 10 keywords. It is preferred to use Medical Subject
           Headings (MeSH) keywords."><Keywords></text:placeholder>
        </text:p>
        -<text:h text:style-name="Heading_20_5" text:outline-level="5">
           Word count:
           <text:placeholder text:placeholder-type="text" text:description="Click to enter the word count (from introduction to
           conclusion)"><Word count></text:placeholder>
        </text:h>
        -<text:h text:style-name="Heading_20_5" text:outline-level="5">
```

# Discussion

▶ We created an OpenOffice template which keeps both the layout and the semantic structure of a scientific paper

▶ This method as well as the PEN ontology can be generalized to domains other than biomedical sciences

▶ Critics:

   ▶ The process of transforming OWL to a template is not automatic, but it is possible to do so.
   ▶ The order is not specified in OWL ontologies. (adding *follows / isFollowedBy* property or add hasOrder / is OrderOf ?)

# Conclusions

▶ Similar ontologies can be created for other types of documents, such as hospitalization reports, film scripts, essays,...(and why not love letters?).

▶ Perspectives:

  ▶ A new generation of office suits can be imagined which will help organising ideas more efficiently

  ▶ Search engines can be enabled to parse this file for advanced semantic research in scientific papers, so that:

    ▷ Searching articles about an OWL ontology of birds (including owls) would no longer be messy!

# Thank you

massoudtoussi@gmail.com
CloseClinical.org