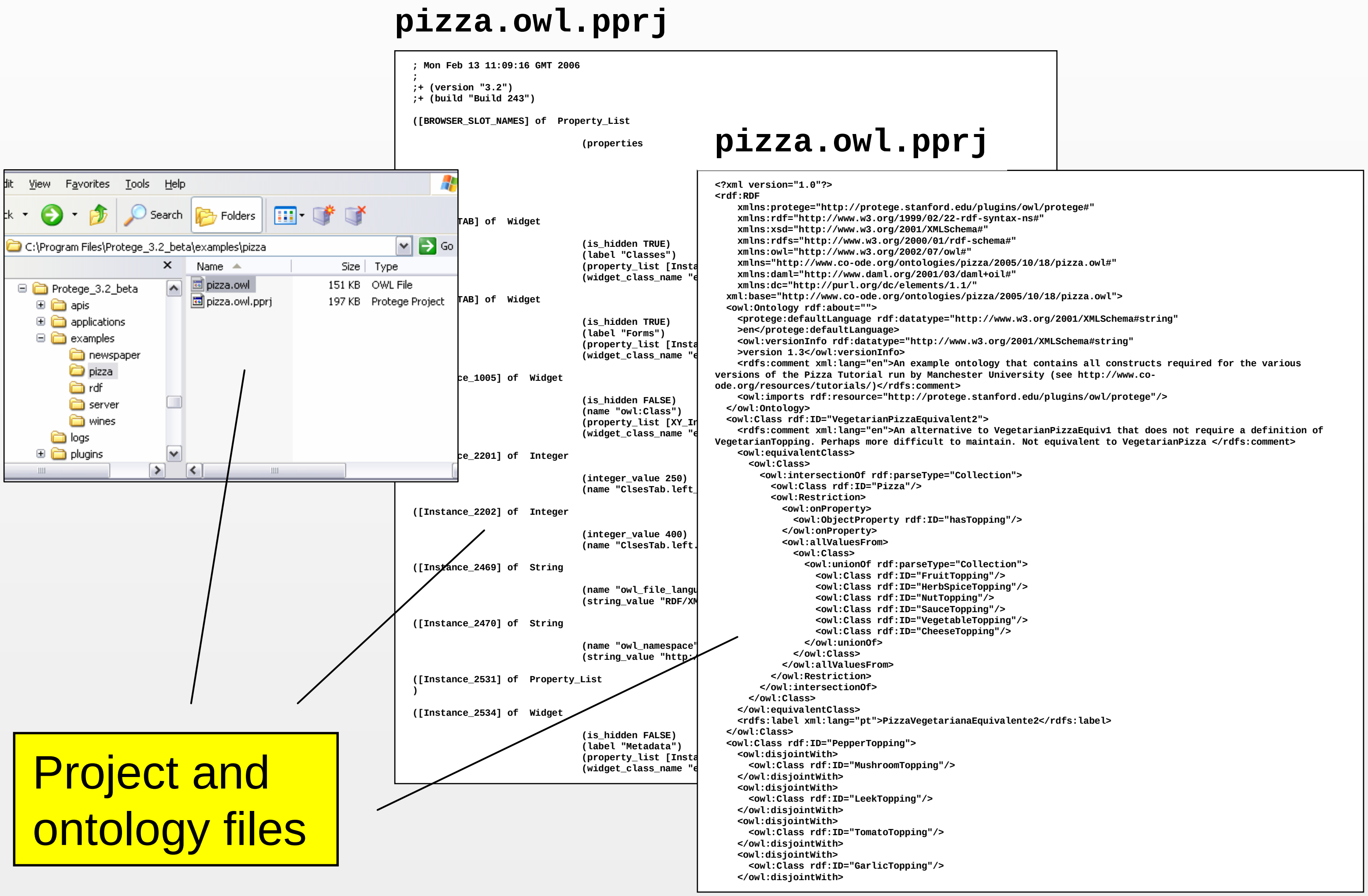
# A PDF Storage Backend for Protégé

Henrik Eriksson

Linköping University

# Storage of the Pizza example



C:\Program Files\Protege_3.2_beta\examples\pizza

| Name | Size | Type |
|---|---|---|
| pizza.owl | 151 KB | OWL File |
| pizza.owl.pprj | 197 KB | Protege Project |

pizza.owl.pprj

```
; Mon Feb 13 11:09:16 GMT 2006
;
;+ (version "3.2")
;+ (build "Build 243")

([BROWSER_SLOT_NAMES] of  Property_List

                              (properties



TAB] of  Widget

                              (is_hidden TRUE)
                              (label "Classes")
                              (property_list [Insta
                              (widget_class_name "e

TAB] of  Widget

                              (is_hidden TRUE)
                              (label "Forms")
                              (property_list [Insta
                              (widget_class_name "e

ce_1005] of  Widget

                              (is_hidden FALSE)
                              (name "owl:Class")
                              (property_list [XY_In
                              (widget_class_name "e

ce_2201] of  Integer

                              (integer_value 250)
                              (name "ClsesTab.left_

([Instance_2202] of  Integer

                              (integer_value 400)
                              (name "ClsesTab.left.

([Instance_2469] of  String

                              (name "owl_file_langu
                              (string_value "RDF/XM

([Instance_2470] of  String

                              (name "owl_namespace"
                              (string_value "http:/

([Instance_2531] of  Property_List
)

([Instance_2534] of  Widget

                              (is_hidden FALSE)
                              (label "Metadata")
                              (property_list [Insta
                              (widget_class_name "e
```

pizza.owl.pprj

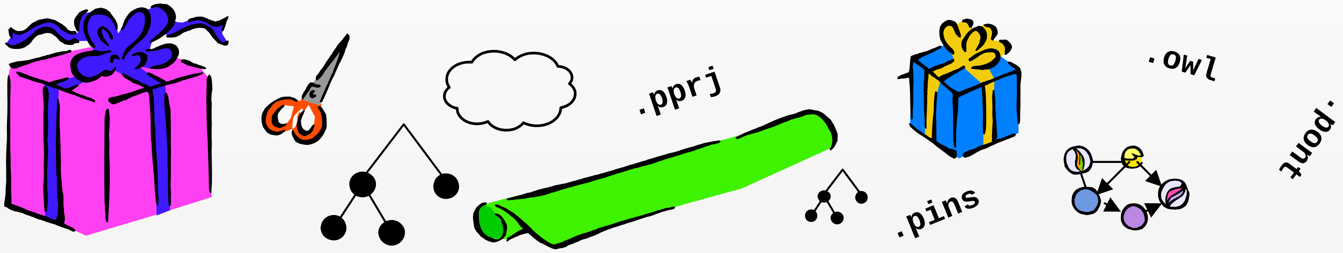
```
<?xml version="1.0"?>
<rdf:RDF
    xmlns:protege="http://protege.stanford.edu/plugins/owl/protege#"
    xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
    xmlns:xsd="http://www.w3.org/2001/XMLSchema#"
    xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
    xmlns:owl="http://www.w3.org/2002/07/owl#"
    xmlns="http://www.co-ode.org/ontologies/pizza/2005/10/18/pizza.owl#"
    xmlns:daml="http://www.daml.org/2001/03/daml+oil#"
    xmlns:dc="http://purl.org/dc/elements/1.1/"
  xml:base="http://www.co-ode.org/ontologies/pizza/2005/10/18/pizza.owl">
  <owl:Ontology rdf:about="">
    <protege:defaultLanguage rdf:datatype="http://www.w3.org/2001/XMLSchema#string"
    >en</protege:defaultLanguage>
    <owl:versionInfo rdf:datatype="http://www.w3.org/2001/XMLSchema#string"
    >version 1.3</owl:versionInfo>
    <rdfs:comment xml:lang="en">An example ontology that contains all constructs required for the various
versions of the Pizza Tutorial run by Manchester University (see http://www.co-
ode.org/resources/tutorials/)</rdfs:comment>
    <owl:imports rdf:resource="http://protege.stanford.edu/plugins/owl/protege"/>
  </owl:Ontology>
  <owl:Class rdf:ID="VegetarianPizzaEquivalent2">
    <rdfs:comment xml:lang="en">An alternative to VegetarianPizzaEquiv1 that does not require a definition of
VegetarianTopping. Perhaps more difficult to maintain. Not equivalent to VegetarianPizza </rdfs:comment>
    <owl:equivalentClass>
      <owl:Class>
        <owl:intersectionOf rdf:parseType="Collection">
          <owl:Class rdf:ID="Pizza"/>
          <owl:Restriction>
            <owl:onProperty>
              <owl:ObjectProperty rdf:ID="hasTopping"/>
            </owl:onProperty>
            <owl:allValuesFrom>
              <owl:Class>
                <owl:unionOf rdf:parseType="Collection">
                  <owl:Class rdf:ID="FruitTopping"/>
                  <owl:Class rdf:ID="HerbSpiceTopping"/>
                  <owl:Class rdf:ID="NutTopping"/>
                  <owl:Class rdf:ID="SauceTopping"/>
                  <owl:Class rdf:ID="VegetableTopping"/>
                  <owl:Class rdf:ID="CheeseTopping"/>
                </owl:unionOf>
              </owl:Class>
            </owl:allValuesFrom>
          </owl:Restriction>
        </owl:intersectionOf>
      </owl:Class>
    </owl:equivalentClass>
    <rdfs:label xml:lang="pt">PizzaVegetarianaEquivalente2</rdfs:label>
  </owl:Class>
  <owl:Class rdf:ID="PepperTopping">
    <owl:disjointWith>
      <owl:Class rdf:ID="MushroomTopping"/>
    </owl:disjointWith>
    <owl:disjointWith>
      <owl:Class rdf:ID="LeekTopping"/>
    </owl:disjointWith>
    <owl:disjointWith>
      <owl:Class rdf:ID="TomatoTopping"/>
    </owl:disjointWith>
    <owl:disjointWith>
      <owl:Class rdf:ID="GarlicTopping"/>
    </owl:disjointWith>
```

Project and ontology files

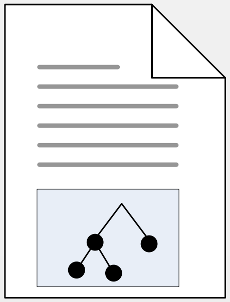
Linköpings universitet

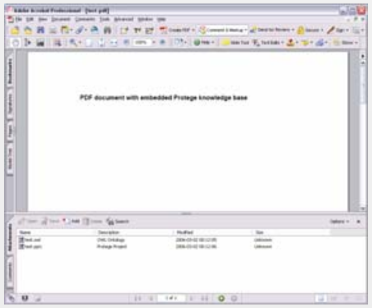
# How do you package an ontology?

- Gift wrapping?



- Document packaging



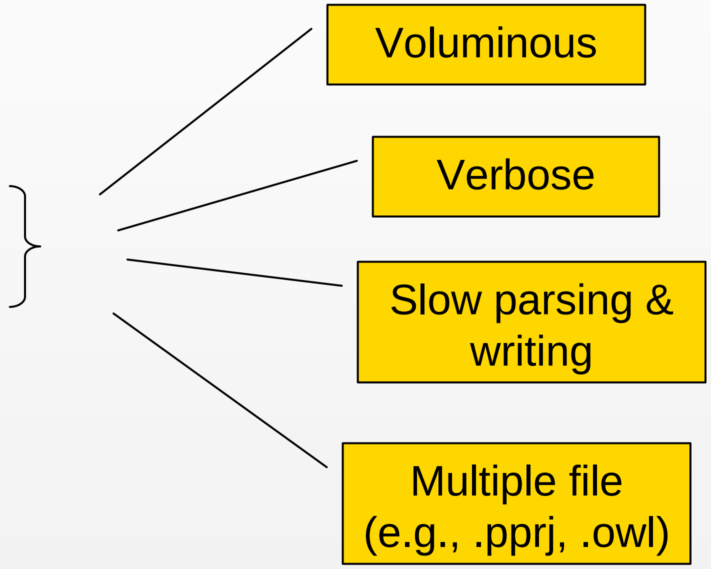


PDF document with embedded Protege knowledge base

| Name | Description | Modified | Size |
|---|---|---|---|
| test.owl | OWL Ontology | 2006-03-02 08:12:05 | Unknown |
| test.pprj | Protege Project | 2006-03-02 08:12:06 | Unknown |

# Persistent storage in Protégé

- Files
  - Serialization
  - Protégé Frames: CLIPS-like/XML
  - Protégé OWL: XML-based
- Databases



Voluminous

Verbose

Slow parsing & writing

Multiple file (e.g., .pprj, .owl)

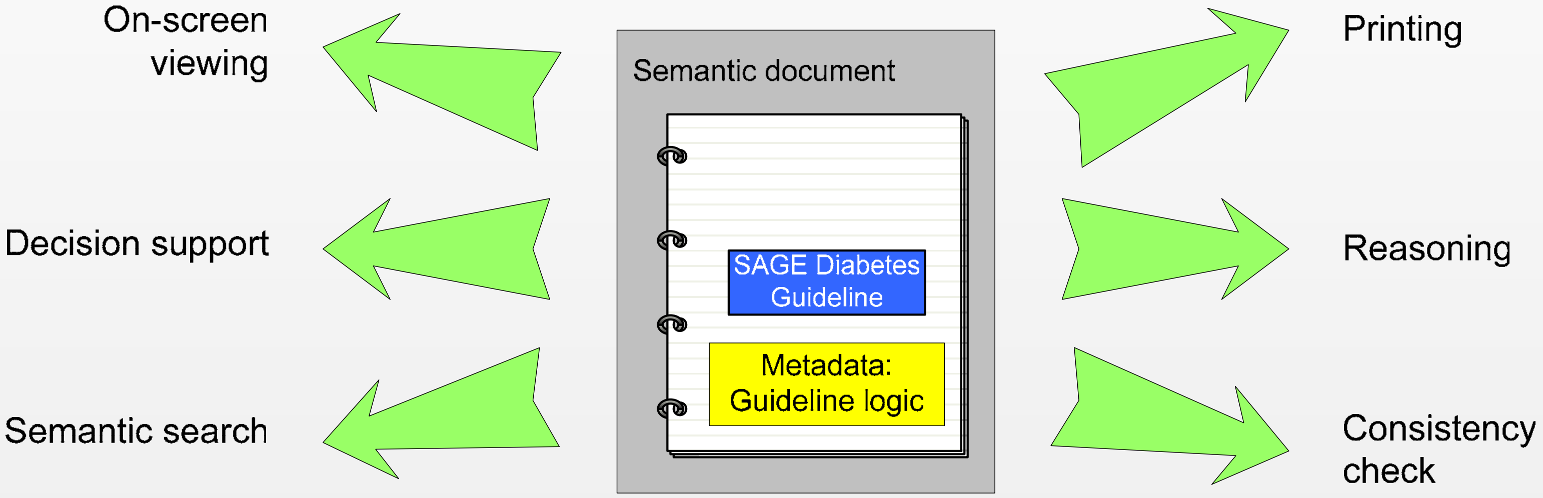
There is a storage problem here

# Background: Semantic Documents

- Combining documents with knowledge representation
  - Like semantic web, but for “real” documents
- Problem: Large amounts of information is available electronically, but it is
  - difficult to find the right information when the search query is complex, and
  - difficult to navigate content-rich information.
- Goal
  - Semantic description of document content (i.e., a meta-model for documents)
  - Support for systematic authoring of complex electronic documents
  - Adding support for PDF to Protégé – a PDF tab for Protégé

Linköpings universitet

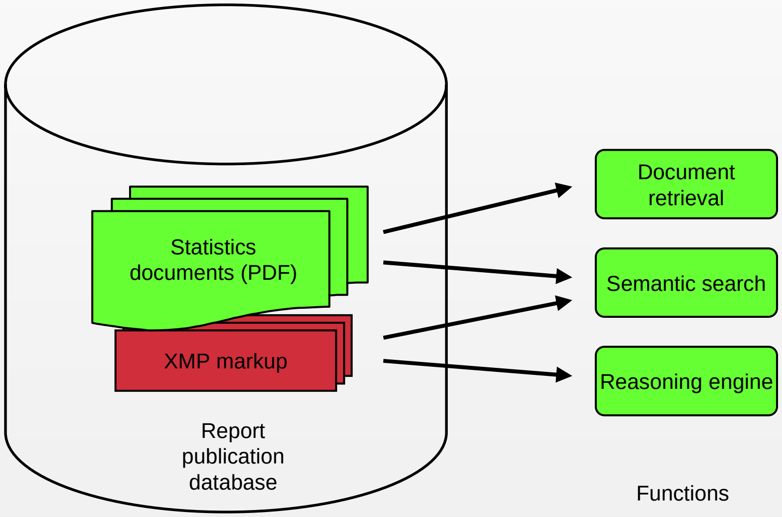
# One Document—Many Applications



One format for all applications

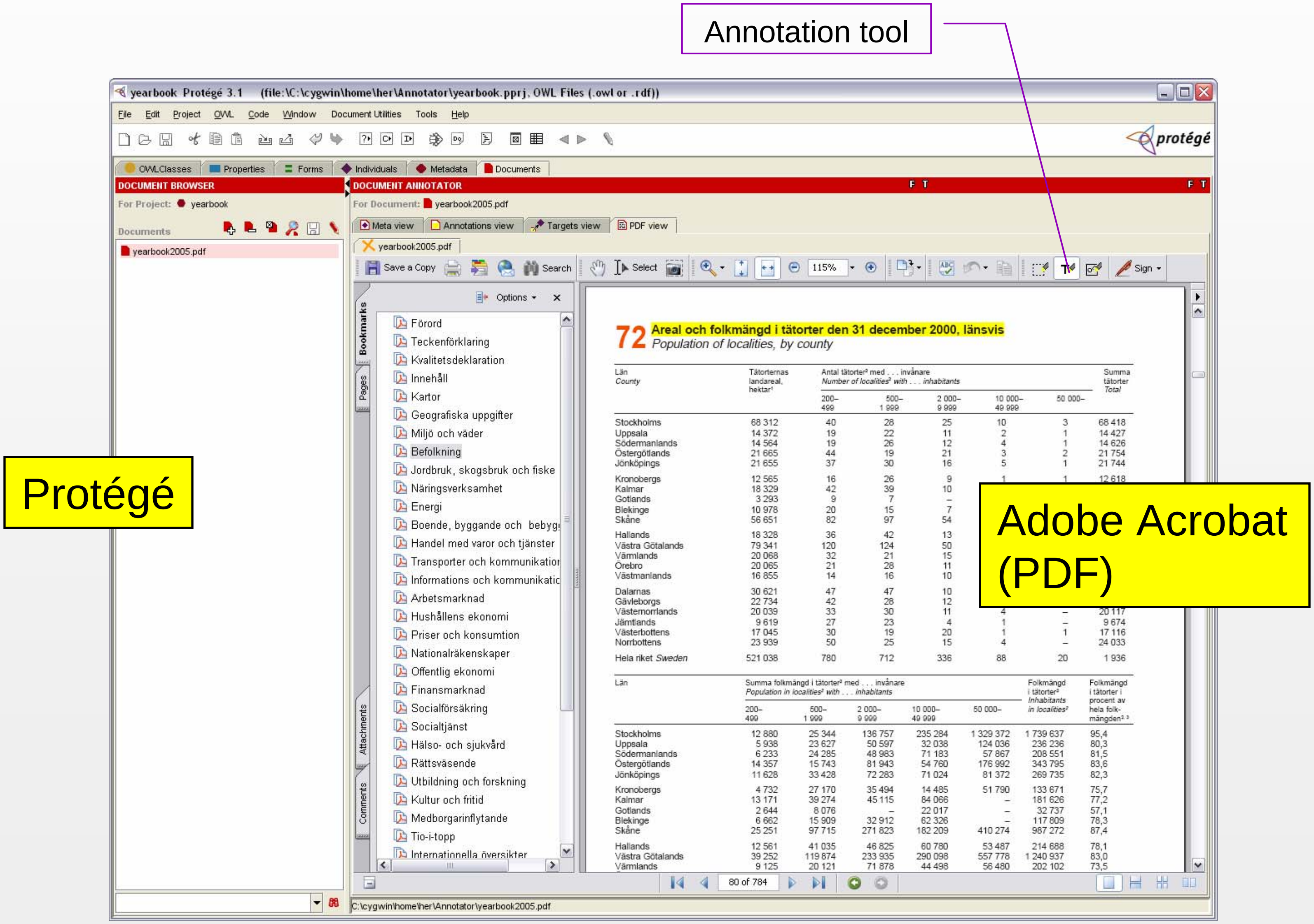
# Semantic Documents

- Knowledge representation
  - Semantic web: OWL
  - Ontologies
- Document models
  - Adobe’s Portable Document Format (PDF)
  - Extensible Metadata Platform (XMP)
  - MS Word, RTF (?)
- Functions
  - Semantic search based on metadata
  - Reasoning, inference



Linköpings universitet

# PDFTab: Annotation tool for Protégé



Annotation tool

Protégé

Adobe Acrobat (PDF)

yearbook Protégé 3.1 (file:\C:\cygwin\home\her\Annotator\yearbook.pprj, OWL Files (.owl or .rdf))
File Edit Project OWL Code Window Document Utilities Tools Help
protégé
OWLClasses Properties Forms Individuals Metadata Documents
DOCUMENT BROWSER
For Project: yearbook
Documents
yearbook2005.pdf
DOCUMENT ANNOTATOR
F T
For Document: yearbook2005.pdf
Meta view Annotations view Targets view PDF view
yearbook2005.pdf
Save a Copy Search Select 115% Sign
Options
Bookmarks Pages Attachments Comments
80 of 784
C:\cygwin\home\her\Annotator\yearbook2005.pdf

- Förord
- Teckenförklaring
- Kvalitetsdeklaration
- Innehåll
- Kartor
- Geografiska uppgifter
- Miljö och väder
- Befolkning
- Jordbruk, skogsbruk och fiske
- Näringsverksamhet
- Energi
- Boende, byggande och bebygg
- Handel med varor och tjänster
- Transporter och kommunikatior
- Informations och kommunikatic
- Arbetsmarknad
- Hushållens ekonomi
- Priser och konsumtion
- Nationalräkenskaper
- Offentlig ekonomi
- Finansmarknad
- Socialförsäkring
- Socialtjänst
- Hälso- och sjukvård
- Rättsväsende
- Utbildning och forskning
- Kultur och fritid
- Medborgarinflytande
- Tio-i-topp
- Internationella översikter

**72 Areal och folkmängd i tätorter den 31 december 2000, länsvis**
*Population of localities, by county*

| Län<br>*County* | Tätorternas landareal, hektar[1] | Antal tätorter[2] med . . . invånare<br>*Number of localities[2] with . . . inhabitants* | | | | | Summa tätorter<br>*Total* |
|---|---|---|---|---|---|---|---|
| | | 200–<br>499 | 500–<br>1 999 | 2 000–<br>9 999 | 10 000–<br>49 999 | 50 000– | |
| Stockholms | 68 312 | 40 | 28 | 25 | 10 | 3 | 68 418 |
| Uppsala | 14 372 | 19 | 22 | 11 | 2 | 1 | 14 427 |
| Södermanlands | 14 564 | 19 | 26 | 12 | 4 | 1 | 14 626 |
| Östergötlands | 21 665 | 44 | 19 | 21 | 3 | 2 | 21 754 |
| Jönköpings | 21 655 | 37 | 30 | 16 | 5 | 1 | 21 744 |
| Kronobergs | 12 565 | 16 | 26 | 9 | 1 | 1 | 12 618 |
| Kalmar | 18 329 | 42 | 39 | 10 | [illegible] | [illegible] | [illegible] |
| Gotlands | 3 293 | 9 | 7 | – | [illegible] | [illegible] | [illegible] |
| Blekinge | 10 978 | 20 | 15 | 7 | [illegible] | [illegible] | [illegible] |
| Skåne | 56 651 | 82 | 97 | 54 | [illegible] | [illegible] | [illegible] |
| Hallands | 18 328 | 36 | 42 | 13 | [illegible] | [illegible] | [illegible] |
| Västra Götalands | 79 341 | 120 | 124 | 50 | [illegible] | [illegible] | [illegible] |
| Värmlands | 20 068 | 32 | 21 | 15 | [illegible] | [illegible] | [illegible] |
| Örebro | 20 065 | 21 | 28 | 11 | [illegible] | [illegible] | [illegible] |
| Västmanlands | 16 855 | 14 | 16 | 10 | [illegible] | [illegible] | [illegible] |
| Dalarnas | 30 621 | 47 | 47 | 10 | [illegible] | [illegible] | [illegible] |
| Gävleborgs | 22 734 | 42 | 28 | 12 | [illegible] | [illegible] | [illegible] |
| Västernorrlands | 20 039 | 33 | 30 | 11 | 4 | – | 20 117 |
| Jämtlands | 9 619 | 27 | 23 | 4 | 1 | – | 9 674 |
| Västerbottens | 17 045 | 30 | 19 | 20 | 1 | 1 | 17 116 |
| Norrbottens | 23 939 | 50 | 25 | 15 | 4 | – | 24 033 |
| Hela riket *Sweden* | 521 038 | 780 | 712 | 336 | 88 | 20 | 1 936 |

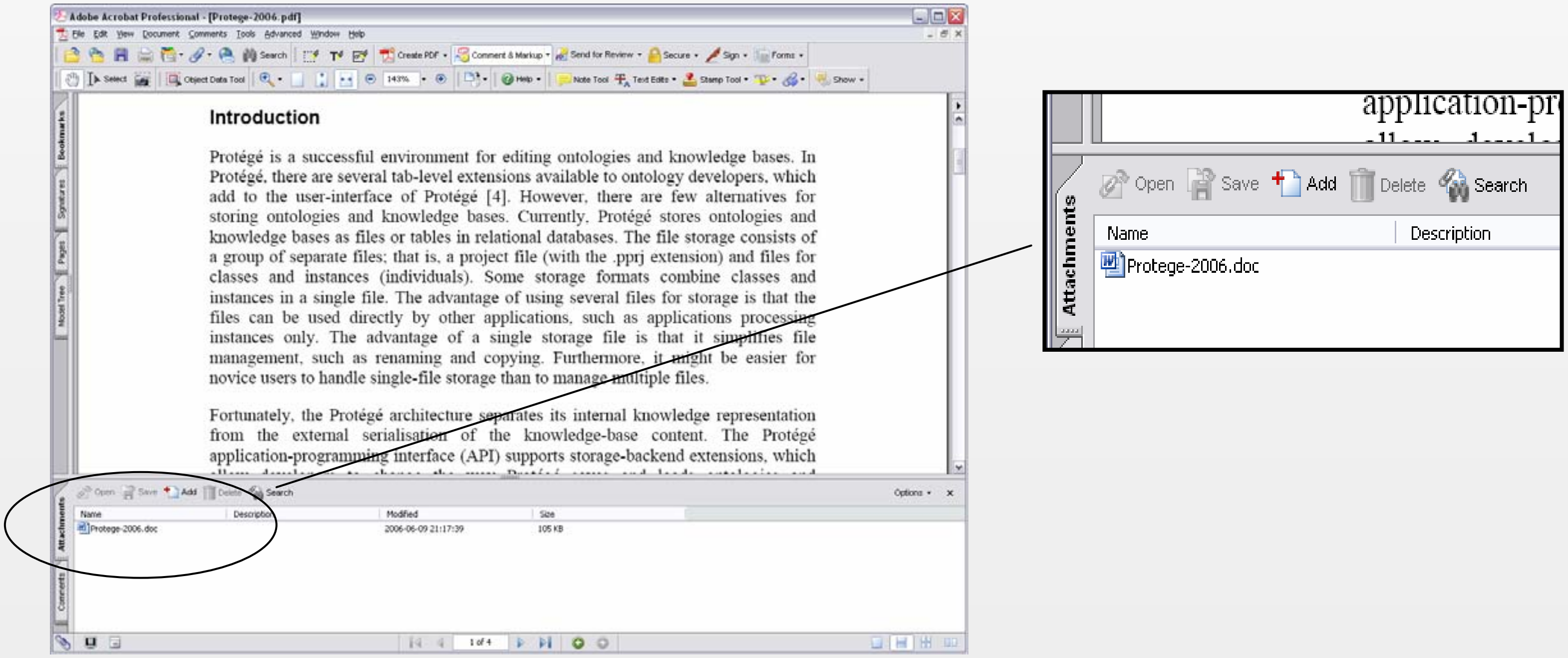
| Län | Summa folkmängd i tätorter[2] med . . . invånare<br>*Population in localities[2] with . . . inhabitants* | | | | | Folkmängd i tätorter[2]<br>*Inhabitants in localities[2]* | Folkmängd i tätorter i procent av hela folkmängden[2 3] |
|---|---|---|---|---|---|---|---|
| | 200–<br>499 | 500–<br>1 999 | 2 000–<br>9 999 | 10 000–<br>49 999 | 50 000– | | |
| Stockholms | 12 880 | 25 344 | 136 757 | 235 284 | 1 329 372 | 1 739 637 | 95,4 |
| Uppsala | 5 938 | 23 627 | 50 597 | 32 038 | 124 036 | 236 236 | 80,3 |
| Södermanlands | 6 233 | 24 285 | 48 983 | 71 183 | 57 867 | 208 551 | 81,5 |
| Östergötlands | 14 357 | 15 743 | 81 943 | 54 760 | 176 992 | 343 795 | 83,6 |
| Jönköpings | 11 628 | 33 428 | 72 283 | 71 024 | 81 372 | 269 735 | 82,3 |
| Kronobergs | 4 732 | 27 170 | 35 494 | 14 485 | 51 790 | 133 671 | 75,7 |
| Kalmar | 13 171 | 39 274 | 45 115 | 84 066 | – | 181 626 | 77,2 |
| Gotlands | 2 644 | 8 076 | – | 22 017 | – | 32 737 | 57,1 |
| Blekinge | 6 662 | 15 909 | 32 912 | 62 326 | – | 117 809 | 78,3 |
| Skåne | 25 251 | 97 715 | 271 823 | 182 209 | 410 274 | 987 272 | 87,4 |
| Hallands | 12 561 | 41 035 | 46 825 | 60 780 | 53 487 | 214 688 | 78,1 |
| Västra Götalands | 39 252 | 119 874 | 233 935 | 290 098 | 557 778 | 1 240 937 | 83,0 |
| Värmlands | 9 125 | 20 121 | 71 878 | 44 498 | 56 480 | 202 102 | 73,5 |

# Lightweight semantic documents

- Semantic documents are nice, but
  - sometimes too heavy
  - advanced tools required (heavy)
- The PDF backend provides
  - a new save method
  - a compact storage format
  - storage using standard PDF attachments
  - file access through standard PDF tools (e.g., Acrobat)

# PDF Attachments

- Little known feature of PDF
- Just like e-mail attachments



## Introduction

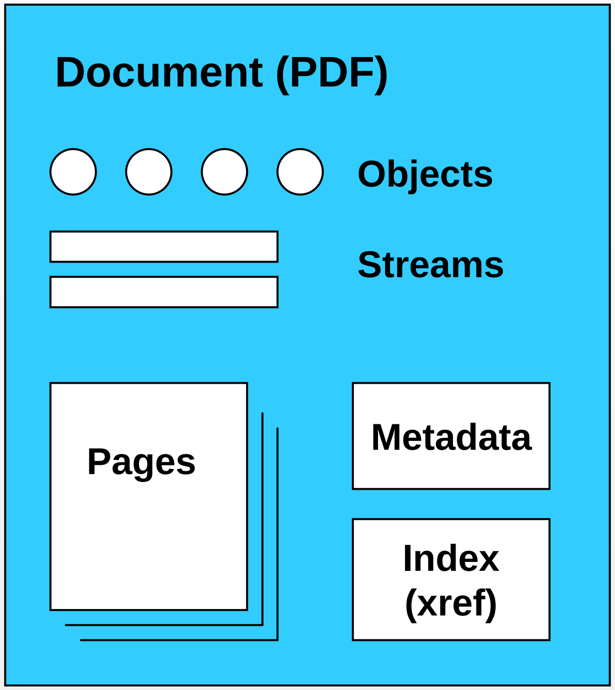
Protégé is a successful environment for editing ontologies and knowledge bases. In Protégé, there are several tab-level extensions available to ontology developers, which add to the user-interface of Protégé [4]. However, there are few alternatives for storing ontologies and knowledge bases. Currently, Protégé stores ontologies and knowledge bases as files or tables in relational databases. The file storage consists of a group of separate files; that is, a project file (with the .pprj extension) and files for classes and instances (individuals). Some storage formats combine classes and instances in a single file. The advantage of using several files for storage is that the files can be used directly by other applications, such as applications processing instances only. The advantage of a single storage file is that it simplifies file management, such as renaming and copying. Furthermore, it might be easier for novice users to handle single-file storage than to manage multiple files.

Fortunately, the Protégé architecture separates its internal knowledge representation from the external serialisation of the knowledge-base content. The Protégé application-programming interface (API) supports storage-backend extensions, which

| Name | Description | Modified | Size |
|---|---|---|---|
| Protege-2006.doc | | 2006-06-09 21:17:39 | 105 KB |

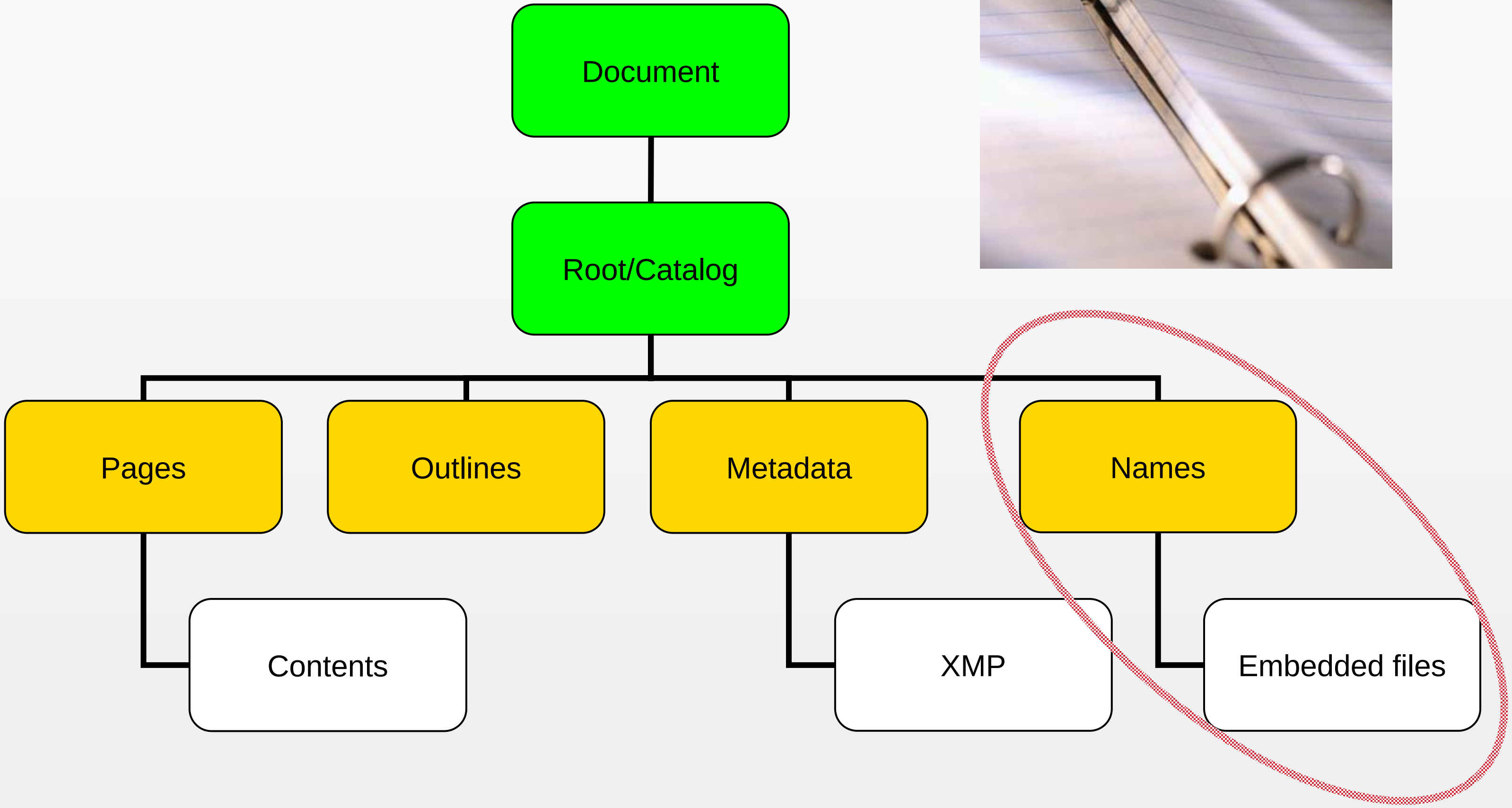
# The “Secrets” of the Portable Document Format (PDF)

- Open and documented format
- PDF files contain something like a file system
  - Indexing for fast random access
  - Like the .doc format of MS Word
- Extendible file layout
  - Custom additions
- Different object and streams with support for text, binary data, compression, and encryption



Linköpings universitet

# Internal PDF Structure

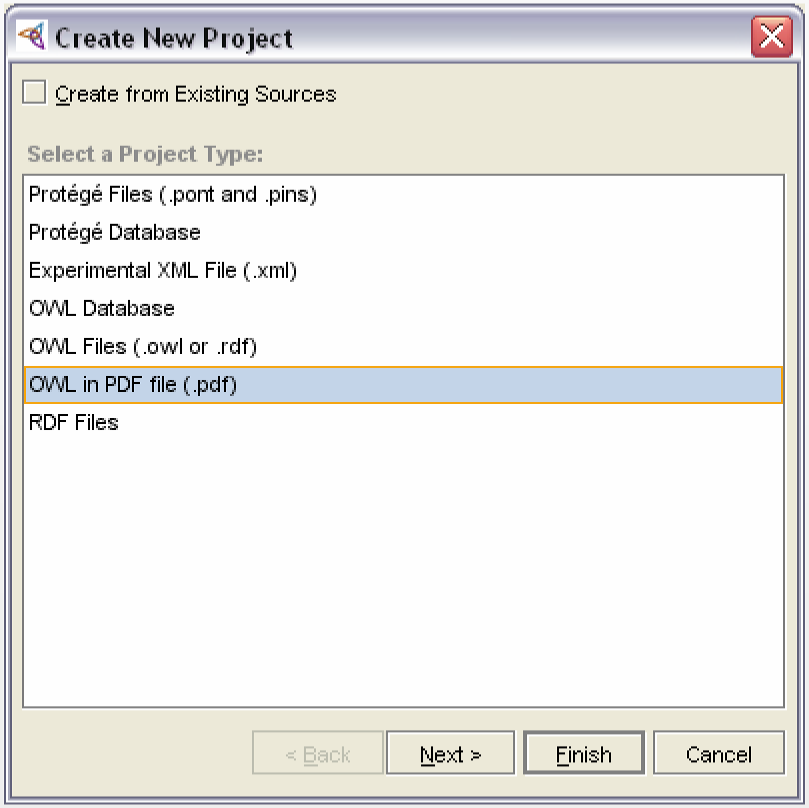


# Inserting ontologies in documents

Storage backend

# Experimental implementation

- New knowledge base format/project type



Create New Project

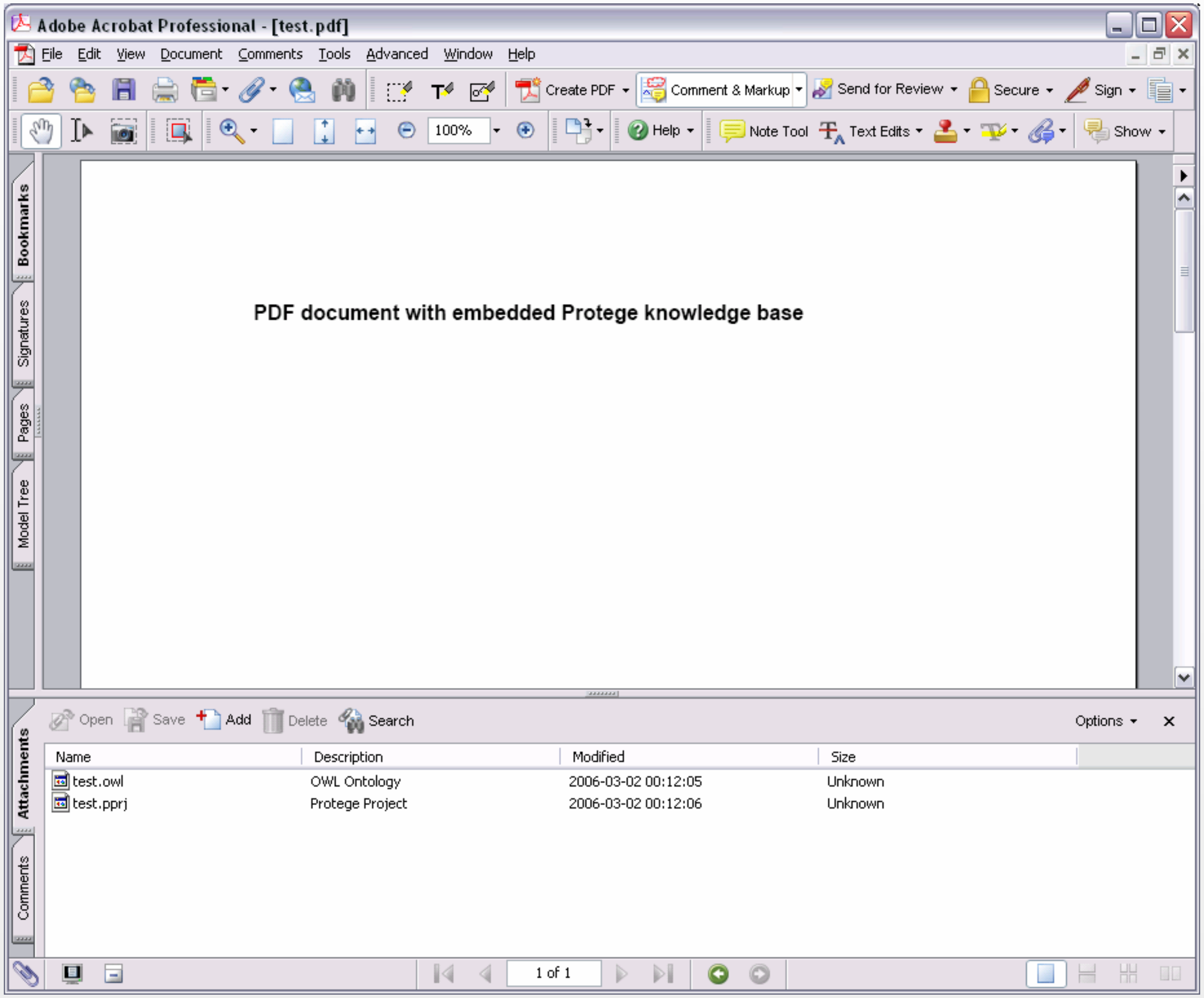
□ Create from Existing Sources

Select a Project Type:

Protégé Files (.pont and .pins)
Protégé Database
Experimental XML File (.xml)
OWL Database
OWL Files (.owl or .rdf)
OWL in PDF file (.pdf)
RDF Files

Linköpings universitet

# Resulting PDF document

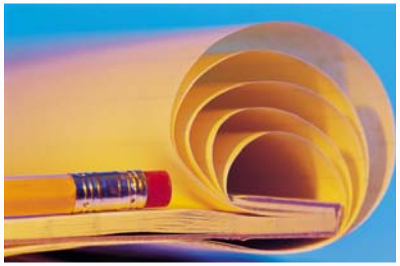


PDF document with embedded Protege knowledge base

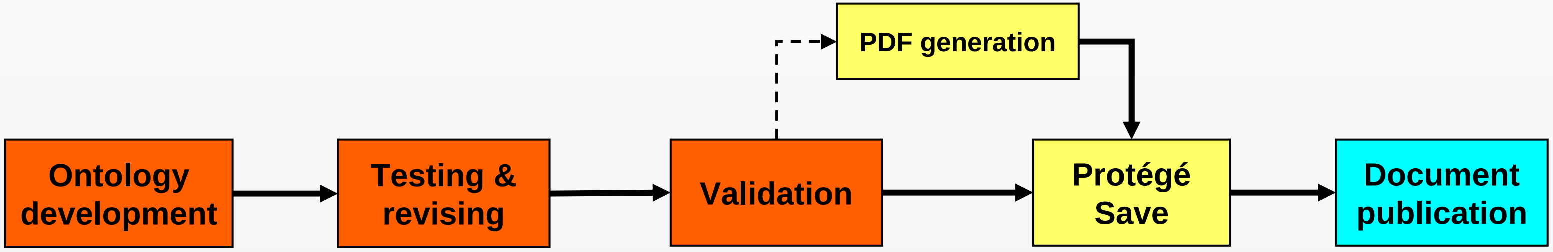
| Name | Description | Modified | Size |
|---|---|---|---|
| test.owl | OWL Ontology | 2006-03-02 00:12:05 | Unknown |
| test.pprj | Protege Project | 2006-03-02 00:12:06 | Unknown |

Linköpings universitet

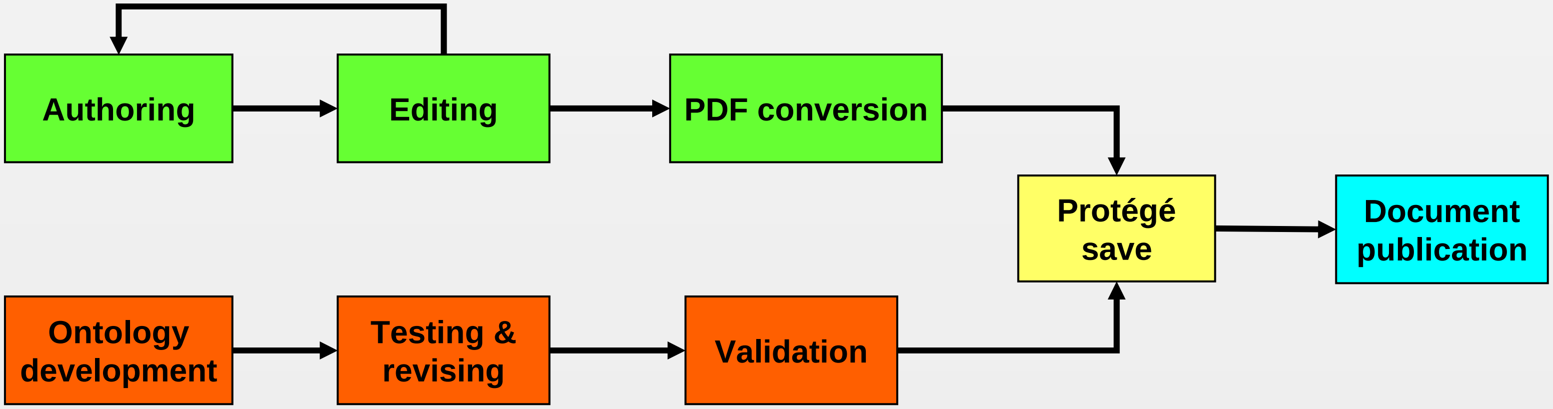
# Scenarios



- Generated documents



- Authored documents



# Discussion

- Architecture for storage (packaging) formats
  - Other formats possible
  - Examples: zip, tar, tgz, …
- Implementation issues
  - Currently “research prototype”
  - API changes/additions/debugging required
    - pdfbox, OWL plug-in, Protégé core
  - One PDF kb format required for each major storage type
    - Example: PDF-Protégé-Frames, PDF-Protégé- OWL, PDF-Protégé-RDFS
    - Should really be separated in a general PDF filter (more API changes required)

Linköpings universitet

# Summary

- Semantic documents
  - Combine printable documents with ontologies and knowledge bases
  - Combined documentation (human-readable) and reasoning (machine-readable)
  - One document with several applications
- PDF storage backend
  - Lightweight semantic documents
  - Attaching ontology files to PDF documents
  - Straightforward access from Acrobat