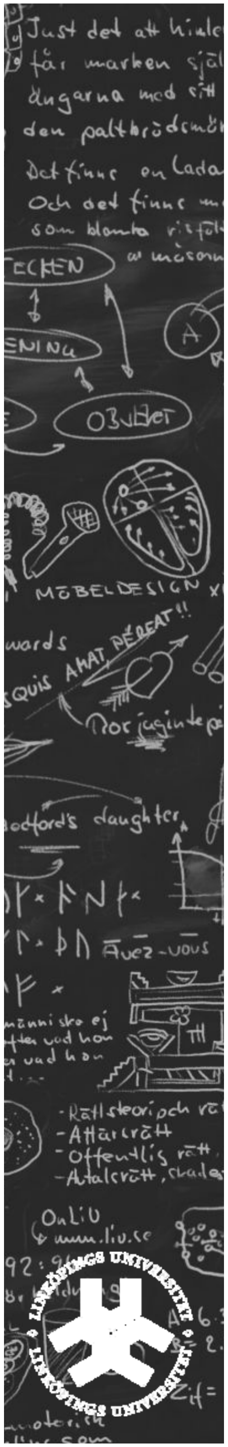
Document Management using Protégé

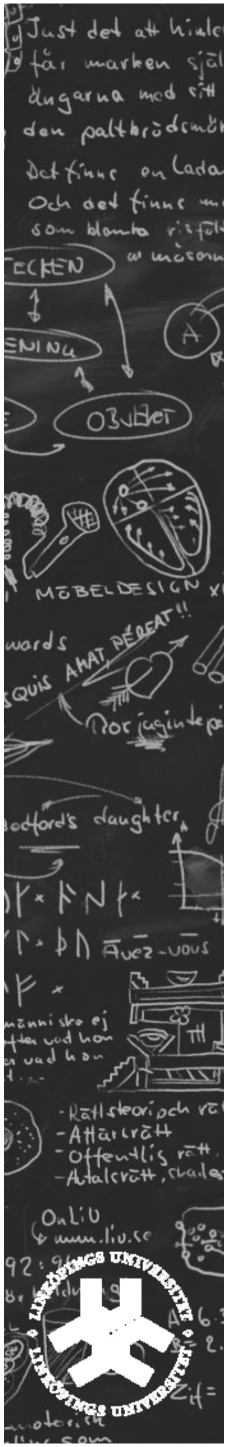
Henrik Eriksson

Linköping University



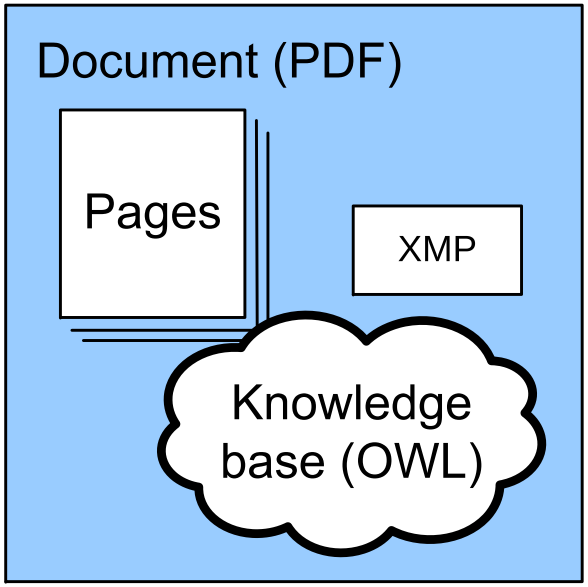
Approach: Semantic Documents

- Combine documents with knowledge representation
  - Like semantic web, but for “real” documents
- Semantic Documents
  - Printable electronic documents
  - Knowledge representation: Ontologies, workflows, and rules
  - An integrated format that keeps textual and computer-based guidelines together
  - Based on wide-spread document formats
- Currently supported format: PDF



Adding Additional Information to the PDF Structure

- Ontologies inside PDF documents
- OWL-based metadata

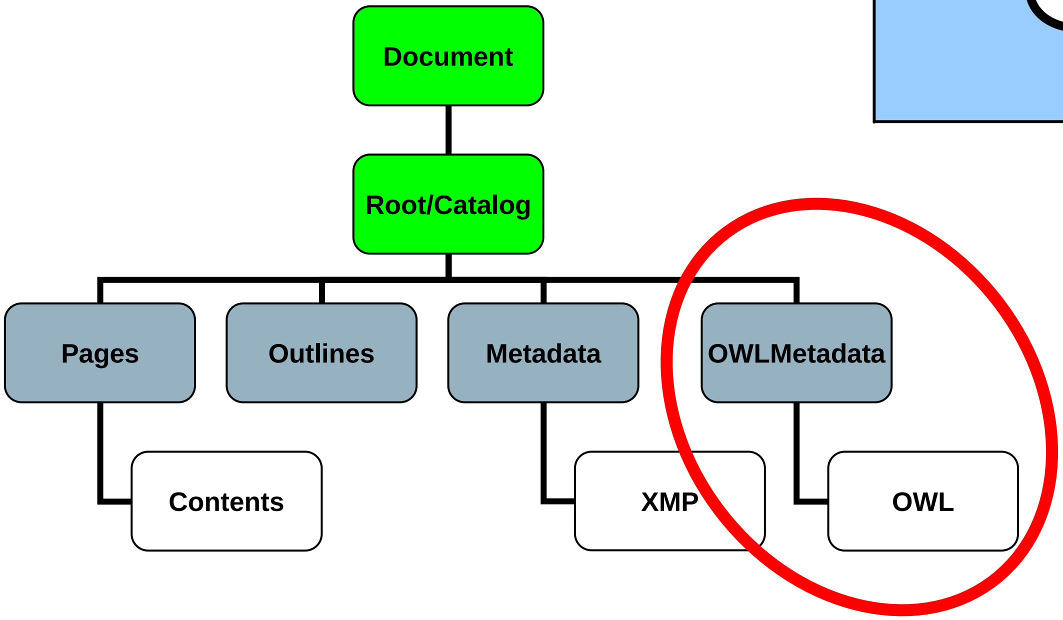


Document (PDF)

Pages

XMP

Knowledge base (OWL)



Document

Root/Catalog

Pages

Outlines

Metadata

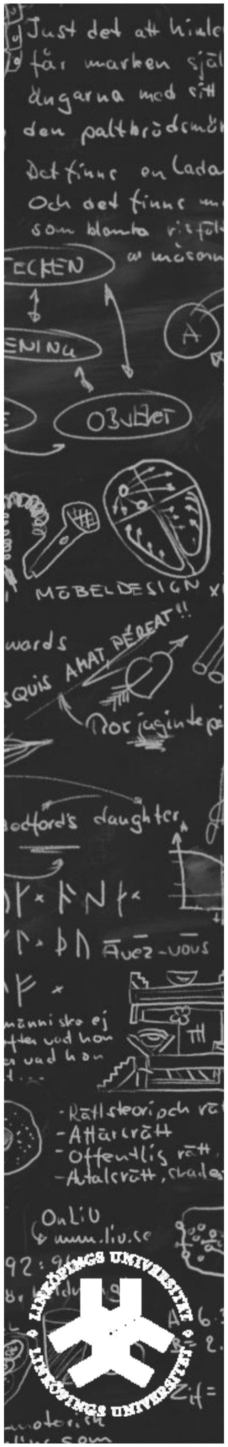
OWLMetadata

Contents

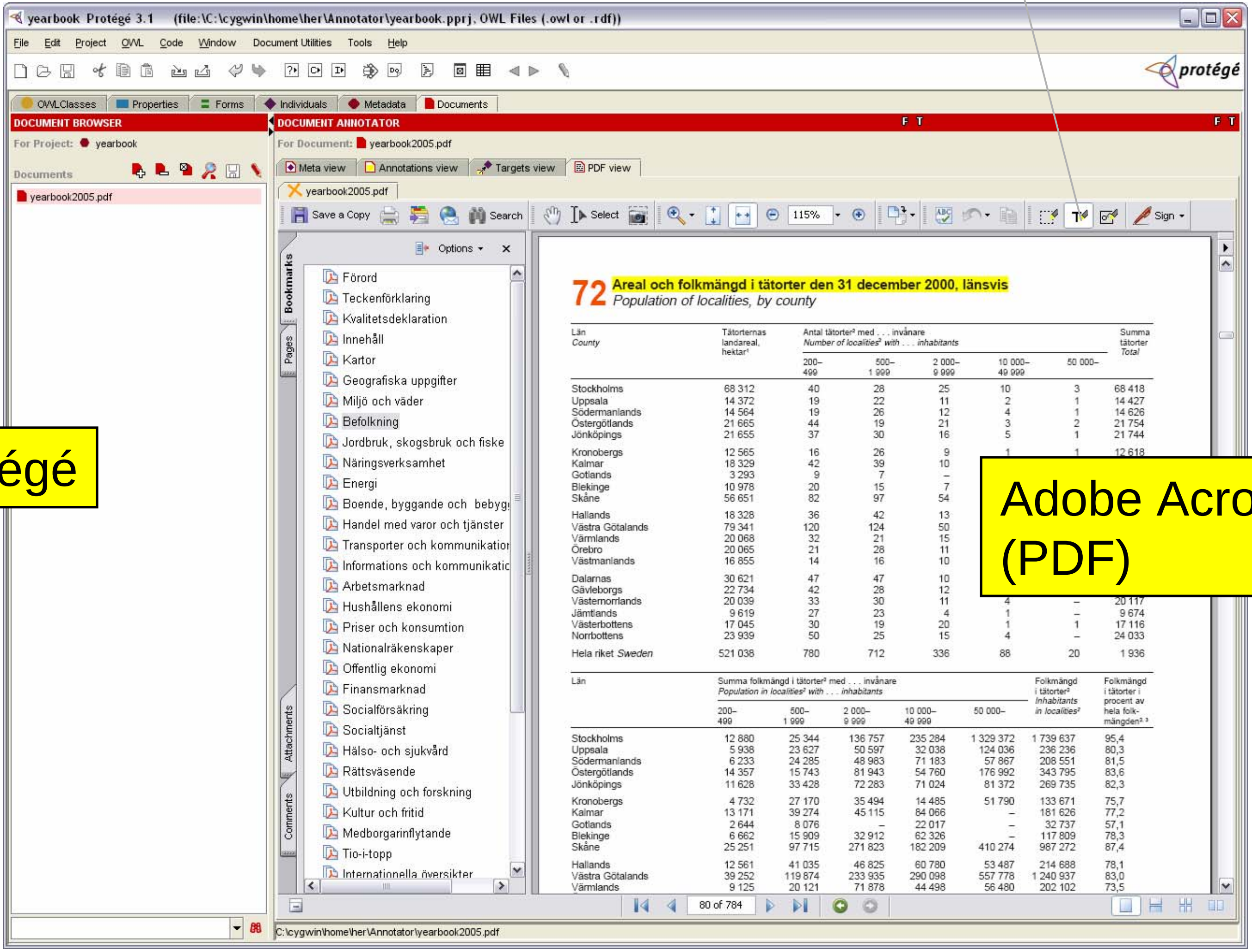
XMP

OWL

Added OWL statements



PDFTab: Annotation Tool for Protégé



Annotation tool

Protégé

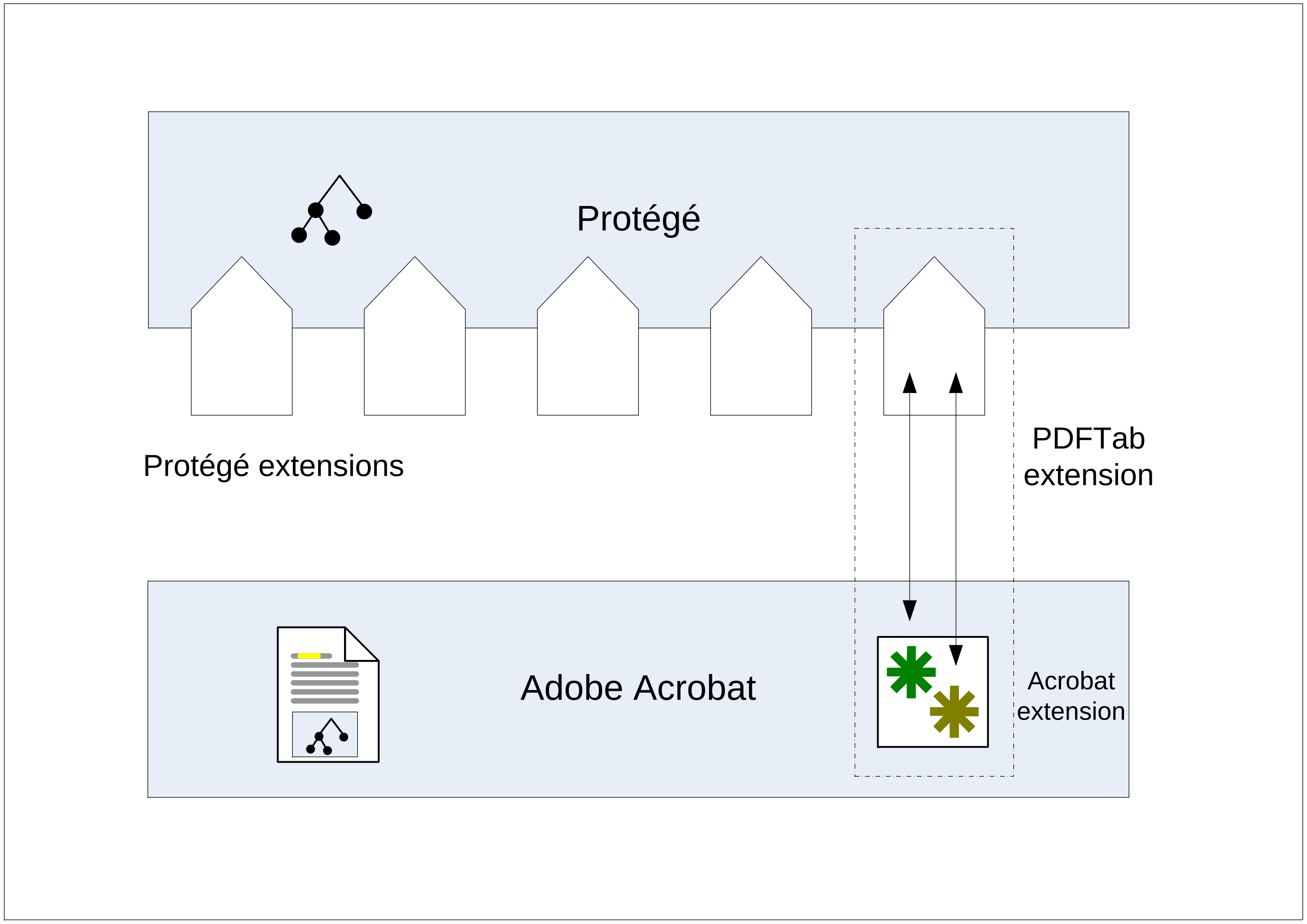
Adobe Acrobat (PDF)

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

Län *County*	Tätorternas landareal, hektar[1]	Antal tätorter[2] med . . . invånare *Number of localities[2] with . . . inhabitants*					Summa tätorter *Total*
		200–499	500–1 999	2 000–9 999	10 000–49 999	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 *Population in localities[2] with . . . inhabitants*					Folkmängd i tätorter[2] *Inhabitants in localities[2]*	Folkmängd i tätorter i procent av hela folkmängden[2, 3]
	200–499	500–1 999	2 000–9 999	10 000–49 999	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

Tool Architecture



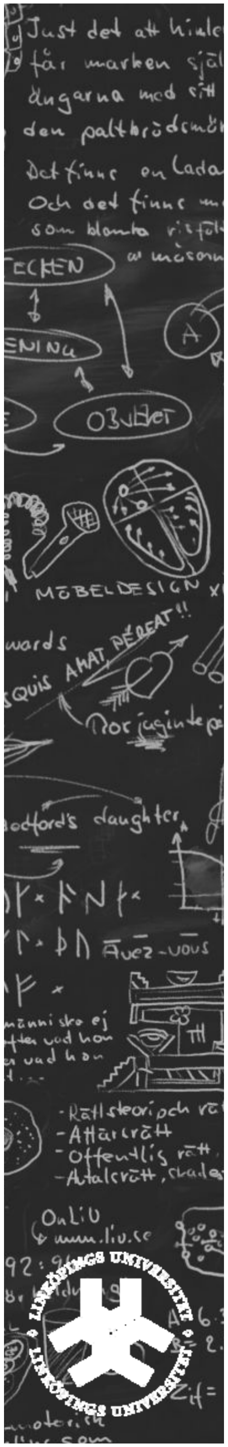
Protégé

Protégé extensions

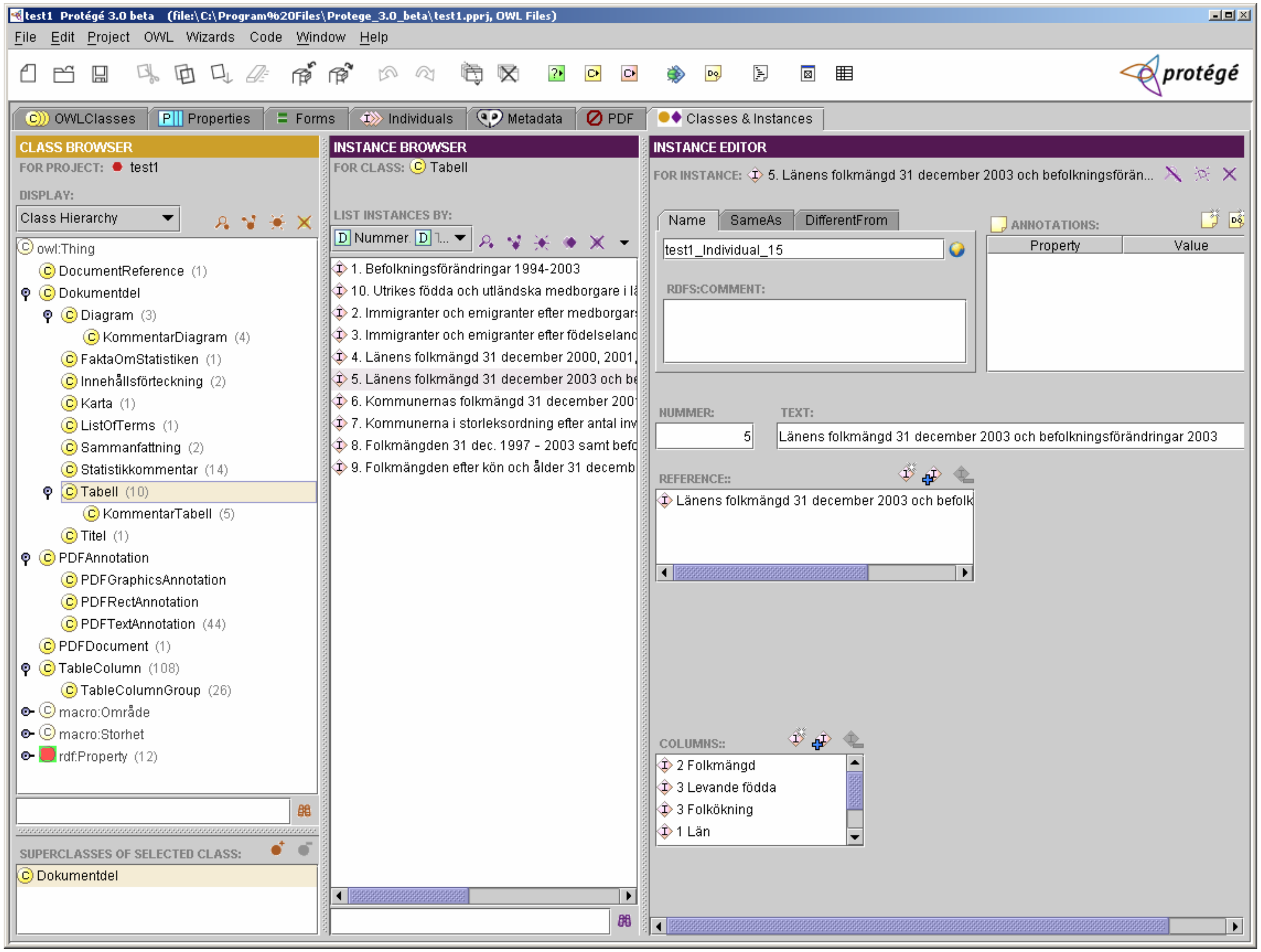
PDFTab extension

Adobe Acrobat

Acrobat extension



Corresponding Ontology



test1 Protégé 3.0 beta (file:\C:\Program%20Files\Protege_3.0_beta\test1.pprj, OWL Files)
File Edit Project OWL Wizards Code Window Help
protégé
OWLClasses Properties Forms Individuals Metadata PDF Classes & Instances

CLASS BROWSER

FOR PROJECT: ● test1

DISPLAY: Class Hierarchy

- owl:Thing
  - DocumentReference (1)
  - Dokumentdel
    - Diagram (3)
      - KommentarDiagram (4)
    - FaktaOmStatistiken (1)
    - Innehållsförteckning (2)
    - Karta (1)
    - ListOfTerms (1)
    - Sammanfattning (2)
    - Statistikkommentar (14)
    - Tabell (10)
      - KommentarTabell (5)
    - Titel (1)
  - PDFAnnotation
    - PDFGraphicsAnnotation
    - PDFRectAnnotation
    - PDFTextAnnotation (44)
  - PDFDocument (1)
  - TableColumn (108)
    - TableColumnGroup (26)
  - macro:Område
  - macro:Storhet
  - rdf:Property (12)

SUPERCLASSES OF SELECTED CLASS:

- Dokumentdel

INSTANCE BROWSER

FOR CLASS: Tabell

LIST INSTANCES BY: Nummer

- 1. Befolkningsförändringar 1994-2003
- 10. Utrikes födda och utländska medborgare i lä
- 2. Immigranter och emigranter efter medborgar
- 3. Immigranter och emigranter efter födelseland
- 4. Länens folkmängd 31 december 2000, 2001,
- 5. Länens folkmängd 31 december 2003 och be
- 6. Kommunernas folkmängd 31 december 200
- 7. Kommunerna i storleksordning efter antal inv
- 8. Folkmängden 31 dec. 1997 – 2003 samt befo
- 9. Folkmängden efter kön och ålder 31 decemb

INSTANCE EDITOR

FOR INSTANCE: 5. Länens folkmängd 31 december 2003 och befolkningsförän...

Name | SameAs | DifferentFrom

Name: test1_Individual_15

RDFS:COMMENT:

ANNOTATIONS:

Property	Value

NUMMER: 5

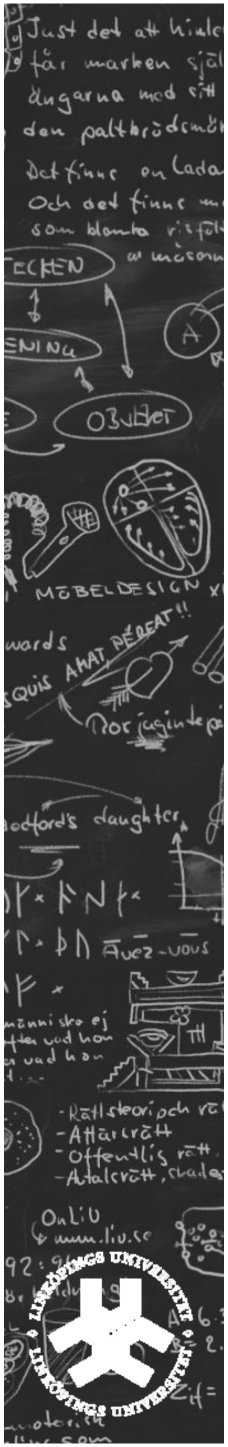
TEXT: Länens folkmängd 31 december 2003 och befolkningsförändringar 2003

REFERENCE::

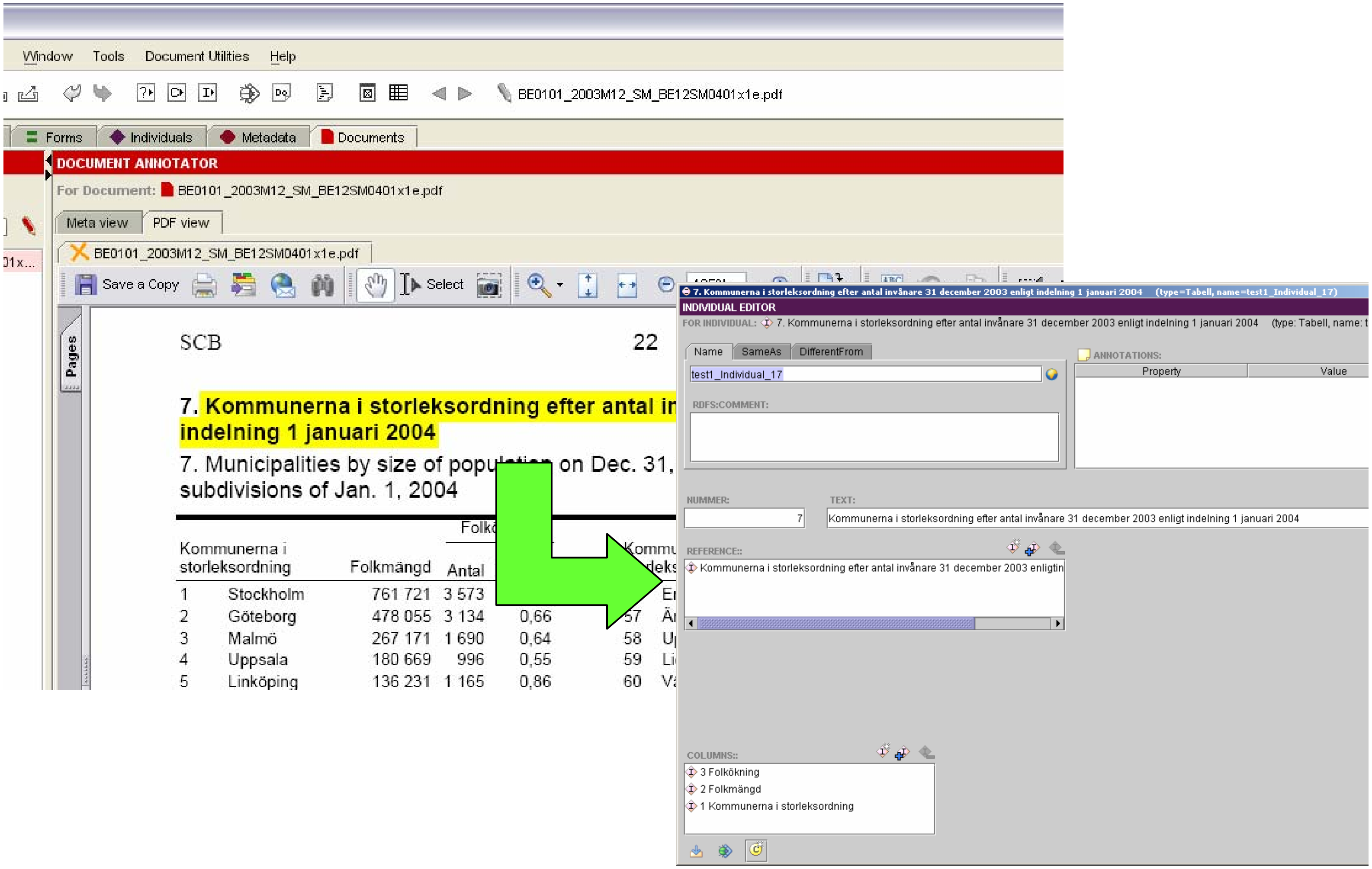
- Länens folkmängd 31 december 2003 och befolk

COLUMNS::

- 2 Folkmängd
- 3 Levande födda
- 3 Folkökning
- 1 Län



Document Mark Up



DOCUMENT ANNOTATOR

For Document: BE0101_2003M12_SM_BE12SM0401x1e.pdf

Meta view | PDF view

BE0101_2003M12_SM_BE12SM0401x1e.pdf

SCB 22

7. Kommunerna i storleksordning efter antal in
indelning 1 januari 2004

7. Municipalities by size of population on Dec. 31,
subdivisions of Jan. 1, 2004

Kommunerna i storleksordning		Folkmängd	Folkö Antal		Kommu deks	
1	Stockholm	761 721	3 573			Er
2	Göteborg	478 055	3 134	0,66	57	Är
3	Malmö	267 171	1 690	0,64	58	Up
4	Uppsala	180 669	996	0,55	59	Li
5	Linköping	136 231	1 165	0,86	60	Vä

7. Kommunerna i storleksordning efter antal invånare 31 december 2003 enligt indelning 1 januari 2004 (type=Tabell, name=test1_Individual_17)

INDIVIDUAL EDITOR

FOR INDIVIDUAL: 7. Kommunerna i storleksordning efter antal invånare 31 december 2003 enligt indelning 1 januari 2004 (type: Tabell, name: t

Name | SameAs | DifferentFrom

test1_Individual_17

RDFS:COMMENT:

ANNOTATIONS:

Property	Value

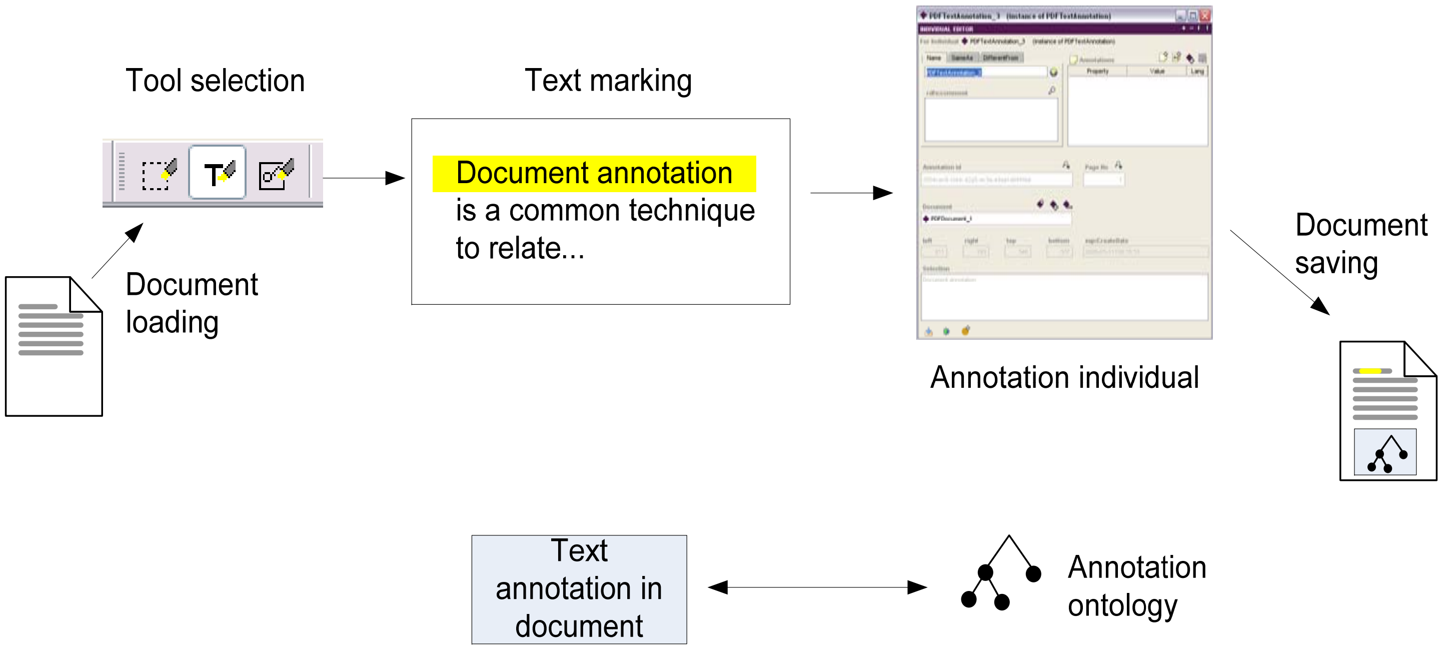
NUMMER: 7

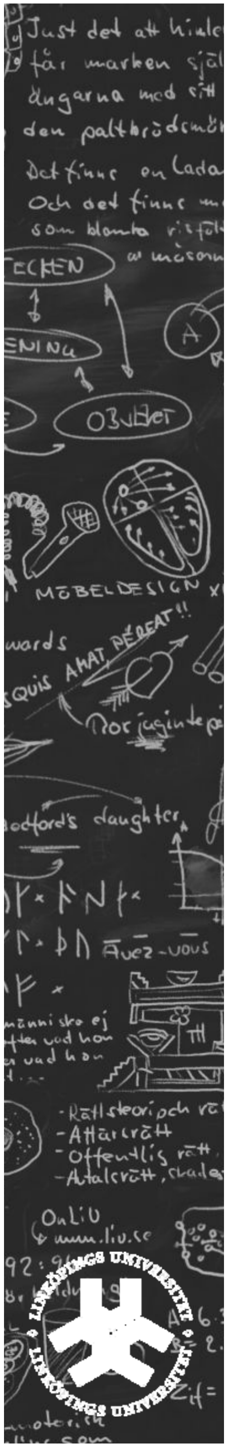
TEXT: Kommunerna i storleksordning efter antal invånare 31 december 2003 enligt indelning 1 januari 2004

REFERENCE::
Kommunerna i storleksordning efter antal invånare 31 december 2003 enligtin

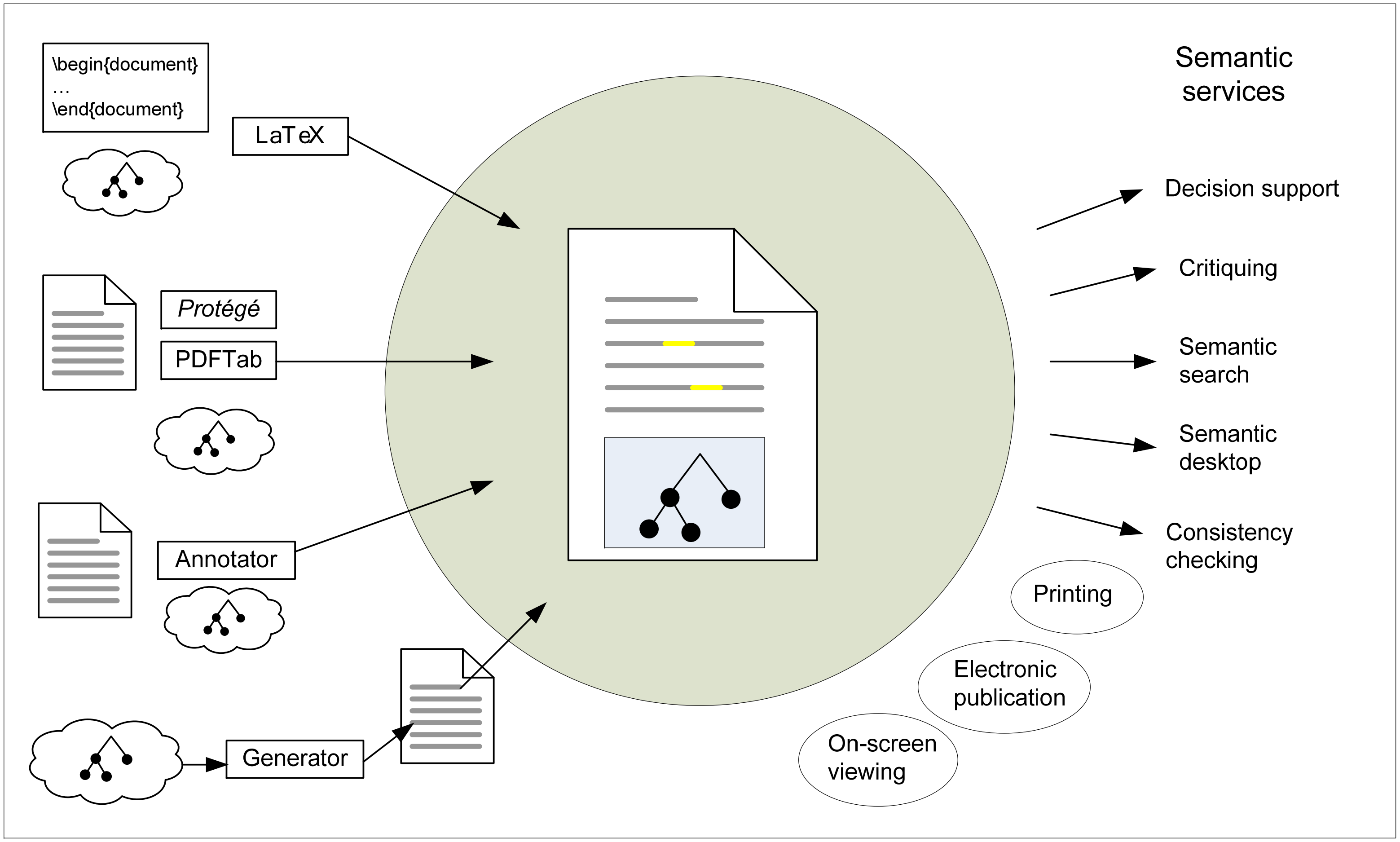
COLUMNS::
- 3 Folkökning
- 2 Folkmängd
- 1 Kommunerna i storleksordning

Annotation Process





Document-centric Annotation Framework



\begin{document}
…
\end{document}

LaTeX

Protégé

PDFTab

Annotator

Generator

Semantic services

Decision support

Critiquing

Semantic search

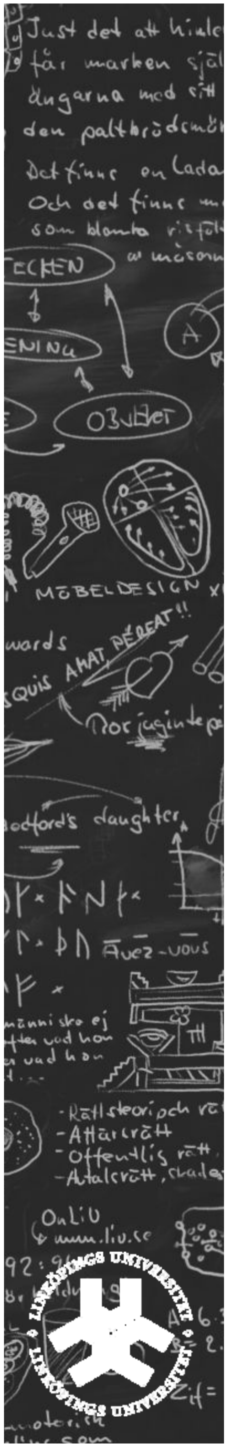
Semantic desktop

Consistency checking

Printing

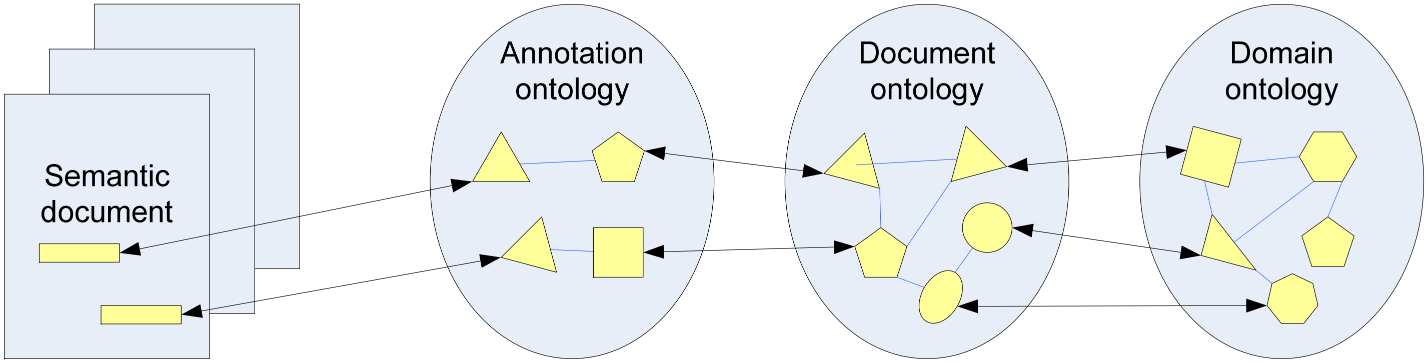
Electronic publication

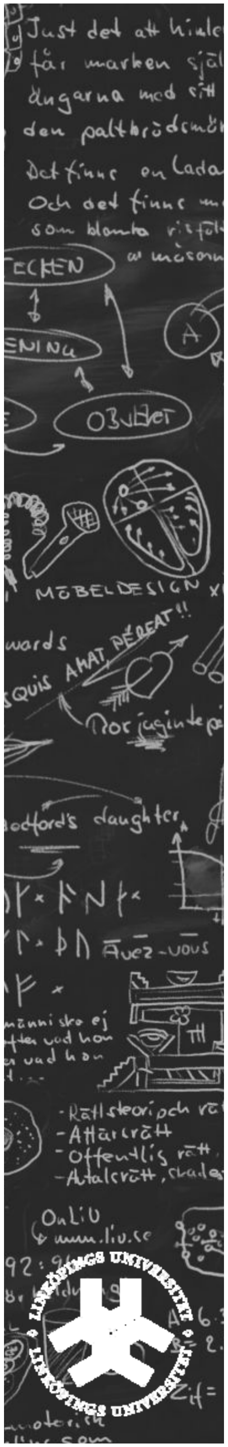
On-screen viewing



Ontology Structure

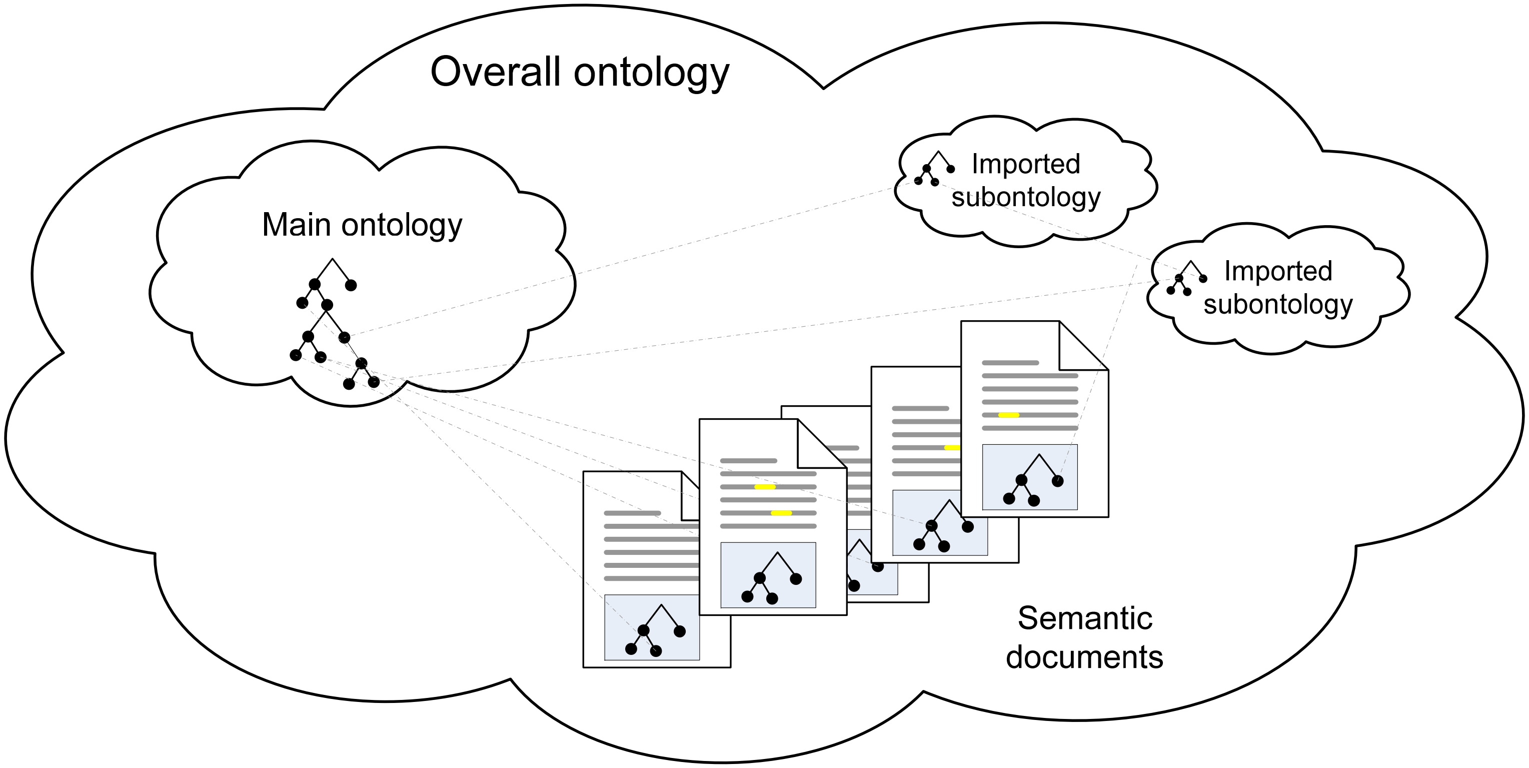
- Linking documents and ontologies
- Standard ontology structure
  - Annotation ontology
    - The annotation types
  - Document ontology
    - The key document parts
  - Domain ontology
    - The “regular” ontology

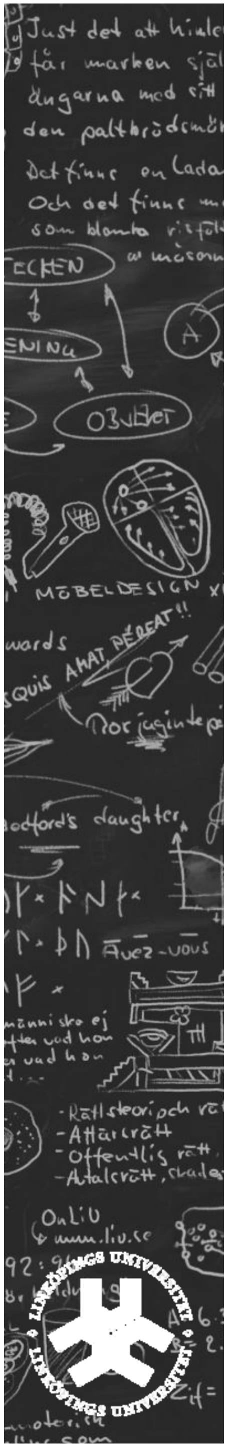




Supporting multiple documents

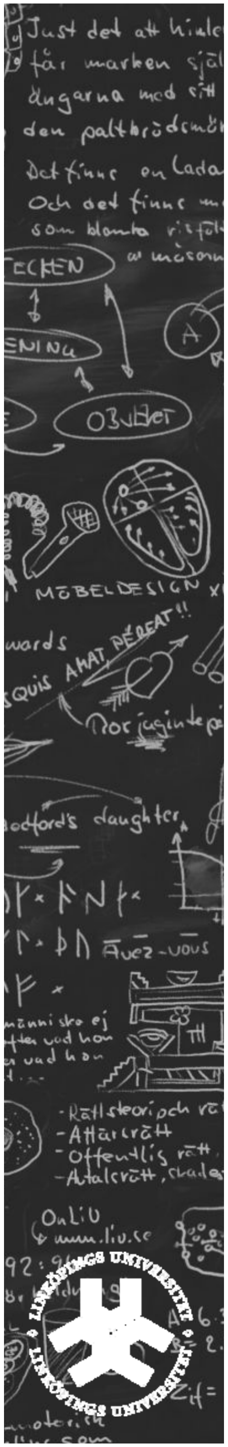
- Architecture with multiple ontologies and ontology modules



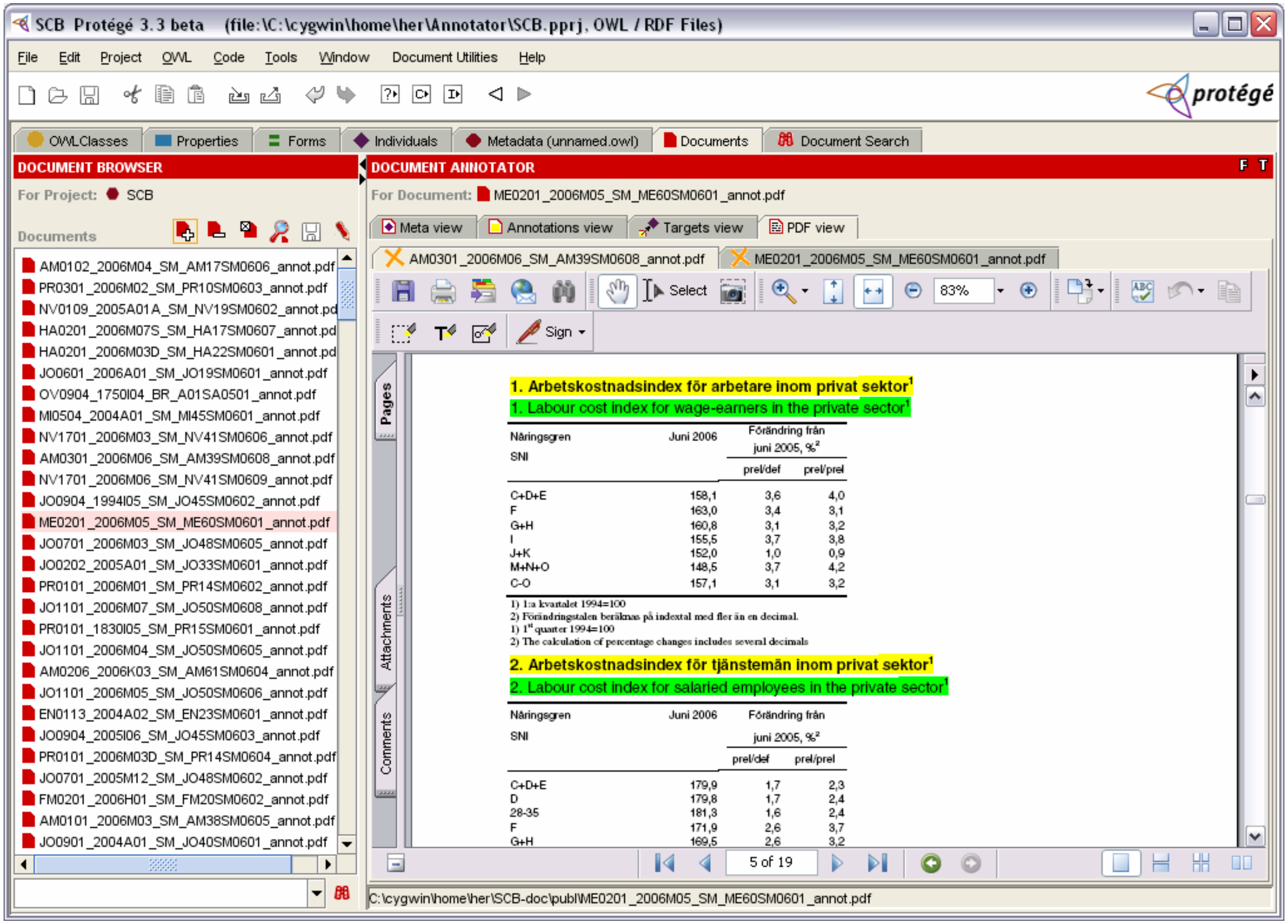


Case Study: Document Repository in Protégé

- Document data set
  - All statistics reports (PDF) published by Statistics Sweden in 2006
  - Five volumes of Statistical Yearbook (2002–2006)
- Method
  - Document acquisition
  - Ontology development
  - Automated annotation (through annotator program)
- Number of automatically-annotated documents: 302
- Total number of annotations for these documents: 17,470



Statistics Reports Loaded in Protégé



For Project: SCB

- AM0102_2006M04_SM_AM17SM0606_annot.pdf
- PR0301_2006M02_SM_PR10SM0603_annot.pdf
- NV0109_2005A01A_SM_NV19SM0602_annot.pd
- HA0201_2006M07S_SM_HA17SM0607_annot.pd
- HA0201_2006M03D_SM_HA22SM0601_annot.pd
- JO0601_2006A01_SM_JO19SM0601_annot.pdf
- OV0904_1750I04_BR_A01SA0501_annot.pdf
- MI0504_2004A01_SM_MI45SM0601_annot.pdf
- NV1701_2006M03_SM_NV41SM0606_annot.pdf
- AM0301_2006M06_SM_AM39SM0608_annot.pdf
- NV1701_2006M06_SM_NV41SM0609_annot.pdf
- JO0904_1994I05_SM_JO45SM0602_annot.pdf
- ME0201_2006M05_SM_ME60SM0601_annot.pdf
- JO0701_2006M03_SM_JO48SM0605_annot.pdf
- JO0202_2005A01_SM_JO33SM0601_annot.pdf
- PR0101_2006M01_SM_PR14SM0602_annot.pdf
- JO1101_2006M07_SM_JO50SM0608_annot.pdf
- PR0101_1830I05_SM_PR15SM0601_annot.pdf
- JO1101_2006M04_SM_JO50SM0605_annot.pdf
- AM0206_2006K03_SM_AM61SM0604_annot.pdf
- JO1101_2006M05_SM_JO50SM0606_annot.pdf
- EN0113_2004A02_SM_EN23SM0601_annot.pdf
- JO0904_2005I06_SM_JO45SM0603_annot.pdf
- PR0101_2006M03D_SM_PR14SM0604_annot.pdf
- JO0701_2005M12_SM_JO48SM0602_annot.pdf
- FM0201_2006H01_SM_FM20SM0602_annot.pdf
- AM0101_2006M03_SM_AM38SM0605_annot.pdf
- JO0901_2004A01_SM_JO40SM0601_annot.pdf

For Document: ME0201_2006M05_SM_ME60SM0601_annot.pdf

AM0301_2006M06_SM_AM39SM0608_annot.pdf | ME0201_2006M05_SM_ME60SM0601_annot.pdf

1. Arbetskostnadsindex för arbetare inom privat sektor[1]

1. Labour cost index for wage-earners in the private sector[1]

Näringsgren SNI	Juni 2006	Förändring från juni 2005, %[2]	
		prel/def	prel/prel
C+D+E	158,1	3,6	4,0
F	163,0	3,4	3,1
G+H	160,8	3,1	3,2
I	155,5	3,7	3,8
J+K	152,0	1,0	0,9
M+N+O	148,5	3,7	4,2
C-O	157,1	3,1	3,2

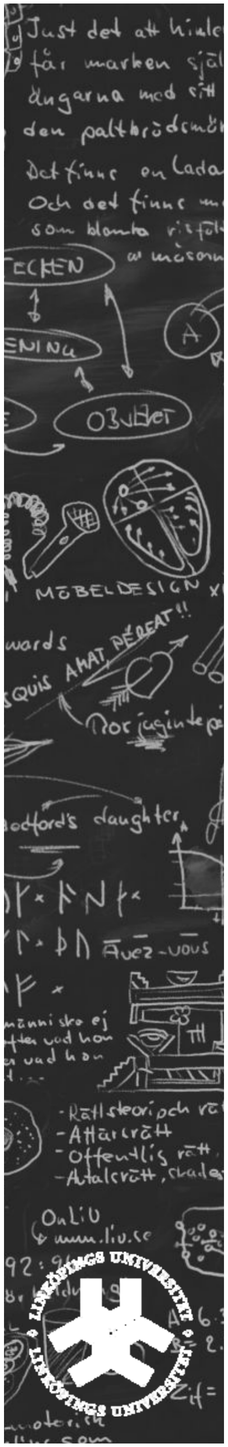
1) 1:a kvartalet 1994=100
2) Förändringstalen beräknas på indextal med fler än en decimal.
1) 1st quarter 1994=100
2) The calculation of percentage changes includes several decimals

2. Arbetskostnadsindex för tjänstemän inom privat sektor[1]

2. Labour cost index for salaried employees in the private sector[1]

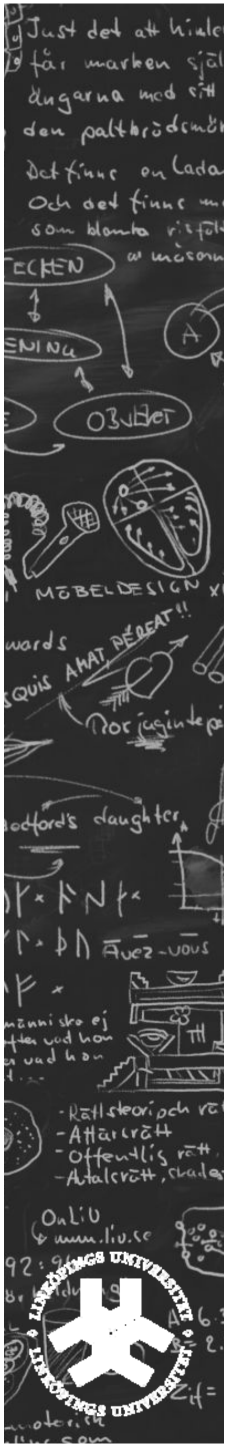
Näringsgren SNI	Juni 2006	Förändring från juni 2005, %[2]	
		prel/def	prel/prel
C+D+E	179,9	1,7	2,3
D	179,8	1,7	2,4
28-35	181,3	1,6	2,4
F	171,9	2,6	3,7
G+H	169,5	2,6	3,2

C:\cygwin\home\her\SCB-doc\publ\ME0201_2006M05_SM_ME60SM0601_annot.pdf



Discussion

- Scalability issues
  - Beyond hundreds of documents
  - Too many ontologies for the current Protégé implementation
  - How can we scale to thousands or millions of documents
- Vision: Repository storage backend
  - Possibly backend based on a document-repository database (e.g., Dspace)
  - Normal document services and semantic services



Summary

- Semantic Documents
- Protégé — a platform for document management
- Ontologies as model document repositories
- Furthermore, ontologies can act as document repositories
- However, large document sets will require a custom-tailored database backend