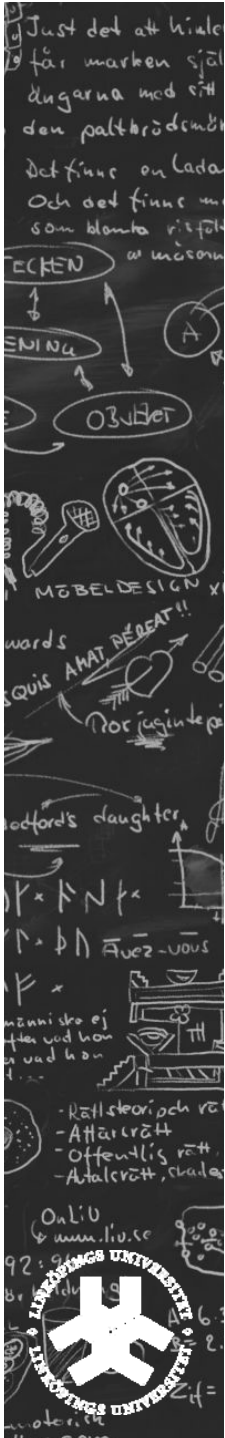


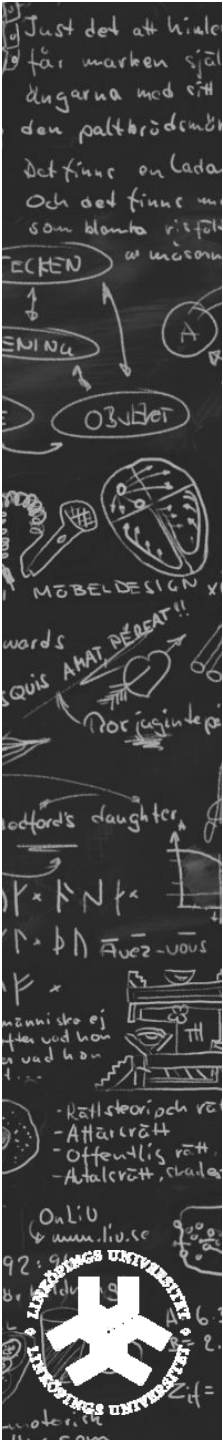
# Document Management using Protégé

Henrik Eriksson  
Linköping University



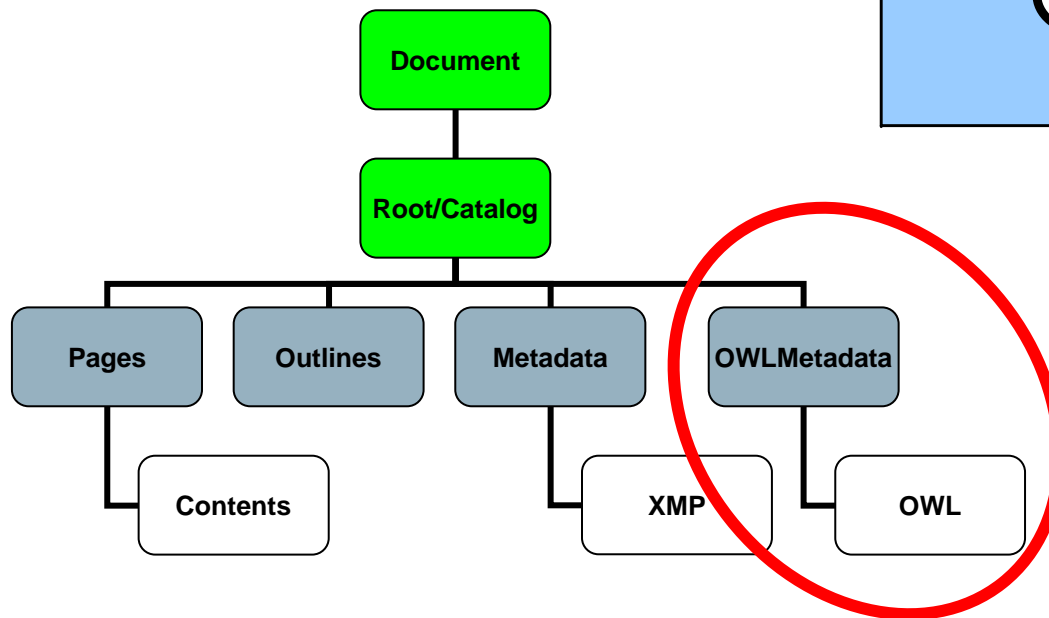
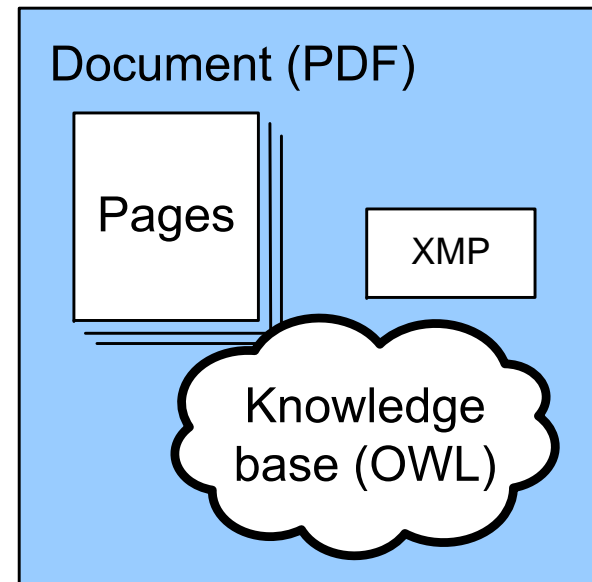
## Approach: Semantic Documents

- 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



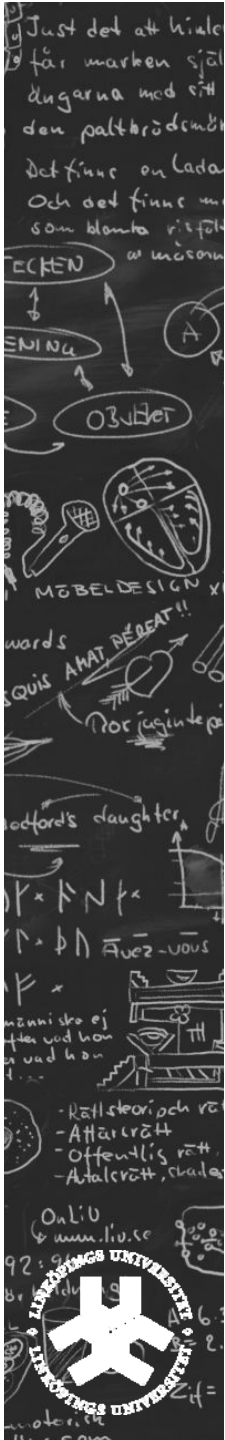
# Adding Additional Information to the PDF Structure

- Ontologies inside PDF documents
- OWL-based metadata



Added OWL statements

# PDFTab: Annotation Tool for Protégé



Annotation tool

Protégé

Adobe Acrobat (PDF)

year book Protégé 3.1 (file: C:\cygwin\home\ther\Annotator\yearbook.pprj\OWL Files (.owl or .rdf))

DOCUMENT ANNOTATOR

For Project: yearbook

For Document: yearbook2005.pdf

Meta view Annotations view Targets view PDF view

Save a Copy Search Select 115% Sign

Pages

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

72 Areal och folkmängd i tätorter den 31 december 2000, länsvis  
Population of localities with ... inhabitants

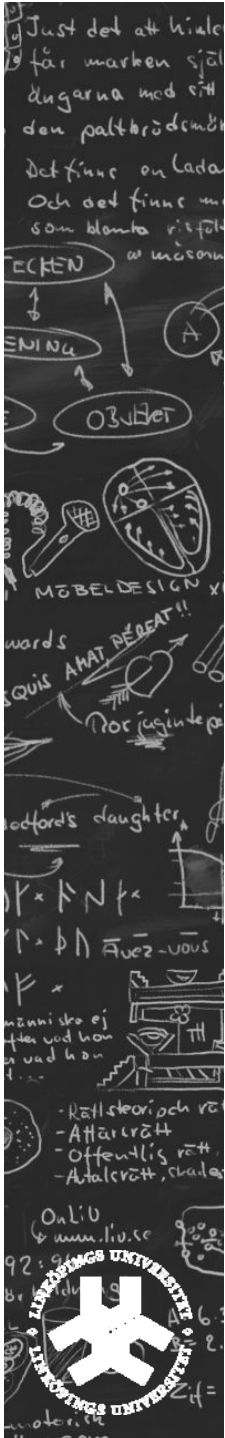
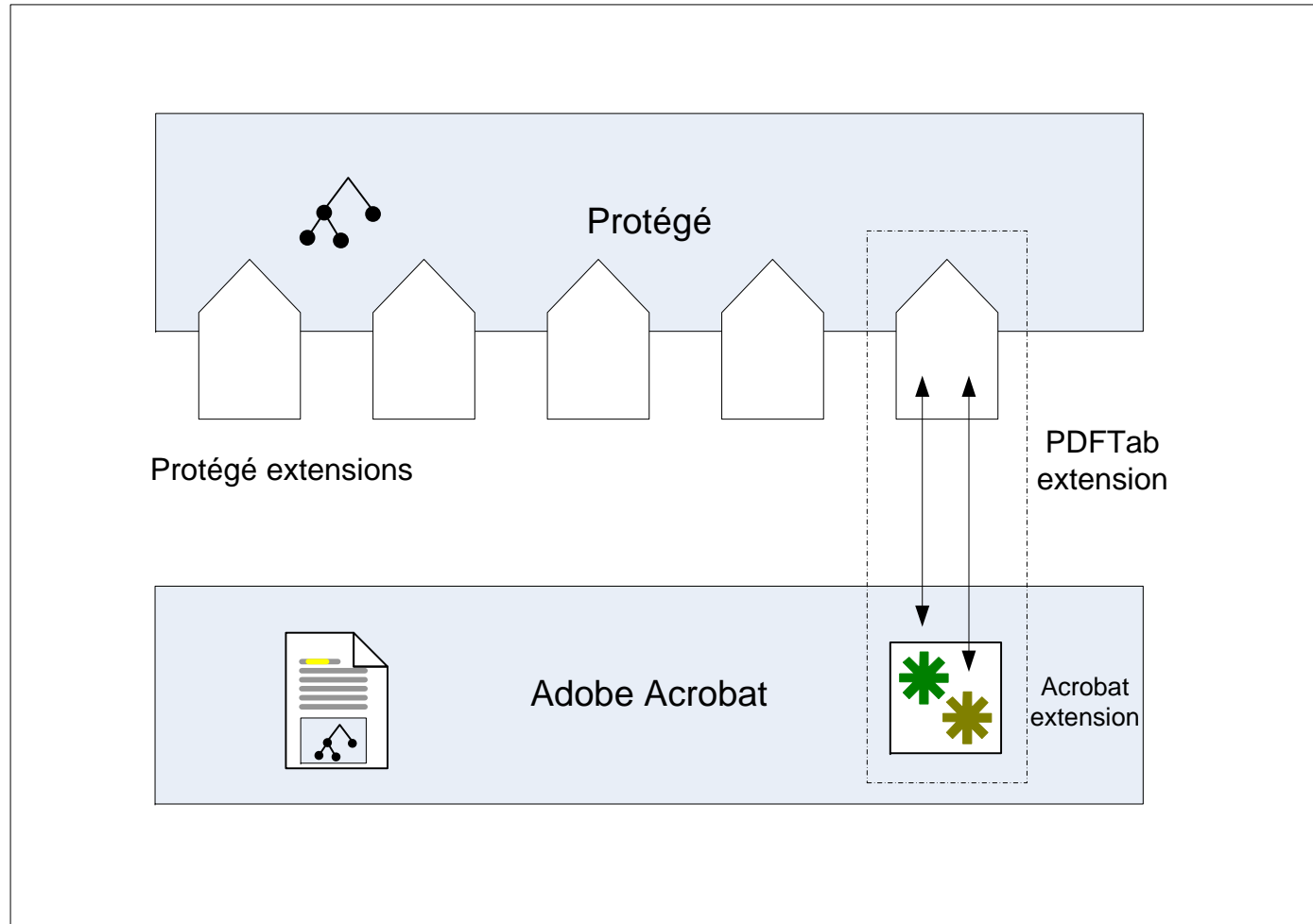
Län County	Tätorternas landareal, hektar <sup>1</sup>	Antal tätorter <sup>2</sup> med ... invånare Number of localities <sup>2</sup> with ... inhabitants				Summa tätorter Total	
		200- 499	500- 1 999	2 000- 9 999	10 000- 49 999		50 000- ...
Stockholms	68 312	40	28	25	10	3	68 418
Uppsala	14 372	19	22	11	2	1	14 427
Södermanlands	14 564	19	26	12	4	1	14 628
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	1	1	18 372
Gotlands	3 293	9	7	1	1	1	3 302
Blekinge	10 978	20	15	7	1	1	11 012
Skåne	58 651	82	97	54	1	1	58 836
Hallands	18 328	36	42	13	1	1	18 371
Västra Götalands	79 341	120	124	50	1	1	79 636
Värmlands	20 068	32	21	15	1	1	20 118
Örebro	20 065	21	28	11	1	1	20 116
Västmanlands	16 655	14	16	10	1	1	16 697
Dalarnas	30 621	47	47	10	1	1	30 727
Gävleborgs	22 734	42	28	12	1	1	22 778
Västernorrlands	20 039	33	30	11	4	1	20 117
Jämtlands	9 619	27	23	4	1	1	9 674
Västerbottens	17 045	30	19	20	1	1	17 116
Norrbottnens	23 939	50	25	15	4	1	24 033
Hela riket Sweden	521 038	780	712	336	88	20	521 936

Län Summa folkmängd i tätorter<sup>2</sup> med ... invånare  
Population in localities<sup>2</sup> with ... inhabitants

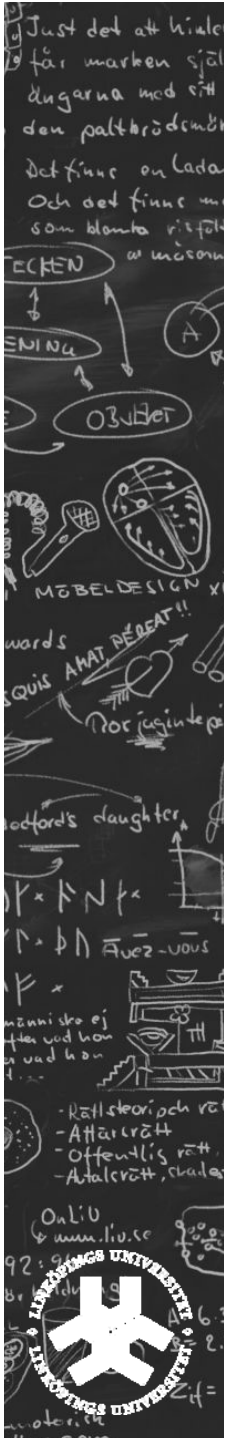
Län	Summa folkmängd i tätorter <sup>2</sup> med ... invånare Population in localities <sup>2</sup> with ... inhabitants				Folkmängd i tätorter <sup>2</sup> Inhabitants in localities <sup>2</sup>	Folkmängd i tätorter procent av hela folk- mängden <sup>1</sup>	
	200- 499	500- 1 999	2 000- 9 999	10 000- 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	84 086	106 578	181 526	77,2
Gotlands	2 644	8 078	1 017	22 017	25 119	32 737	57,1
Blekinge	6 662	15 909	32 912	62 326	117 809	137 809	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	119 674	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

80 of 784

# Tool Architecture



# Corresponding Ontology



test1 Protégé 3.0 beta (file:\C:\Program%20Files\Protége\_3.0\_beta\test1.pprj, OWL Files)

File Edit Project OWL Wizards Code Window Help

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:  
D Nummer D L...

1. Befolkningsförändringar 1994-2003
10. Utrikes födda och utländska medborgare i k
2. Immigranter och emigranter efter medborgars
3. Immigranter och emigranter efter födelselanc
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 bef
9. Folkmängden efter kön och ålder 31 decemb

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

Name: SameAs DifferentFrom

test1\_Individual\_15

RDFS:COMMENT:

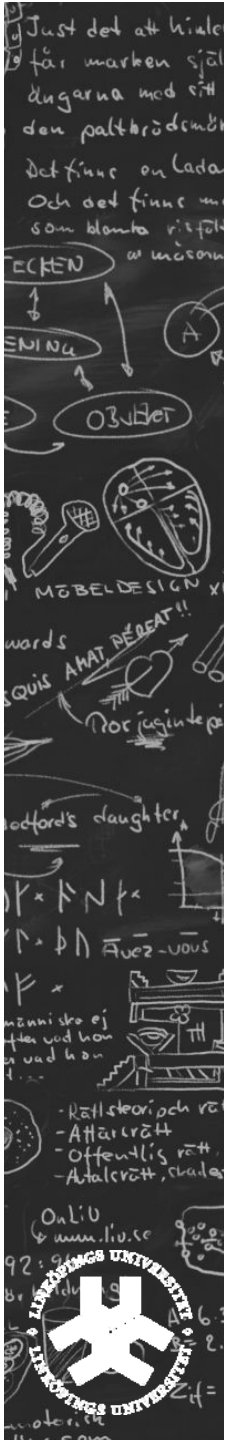
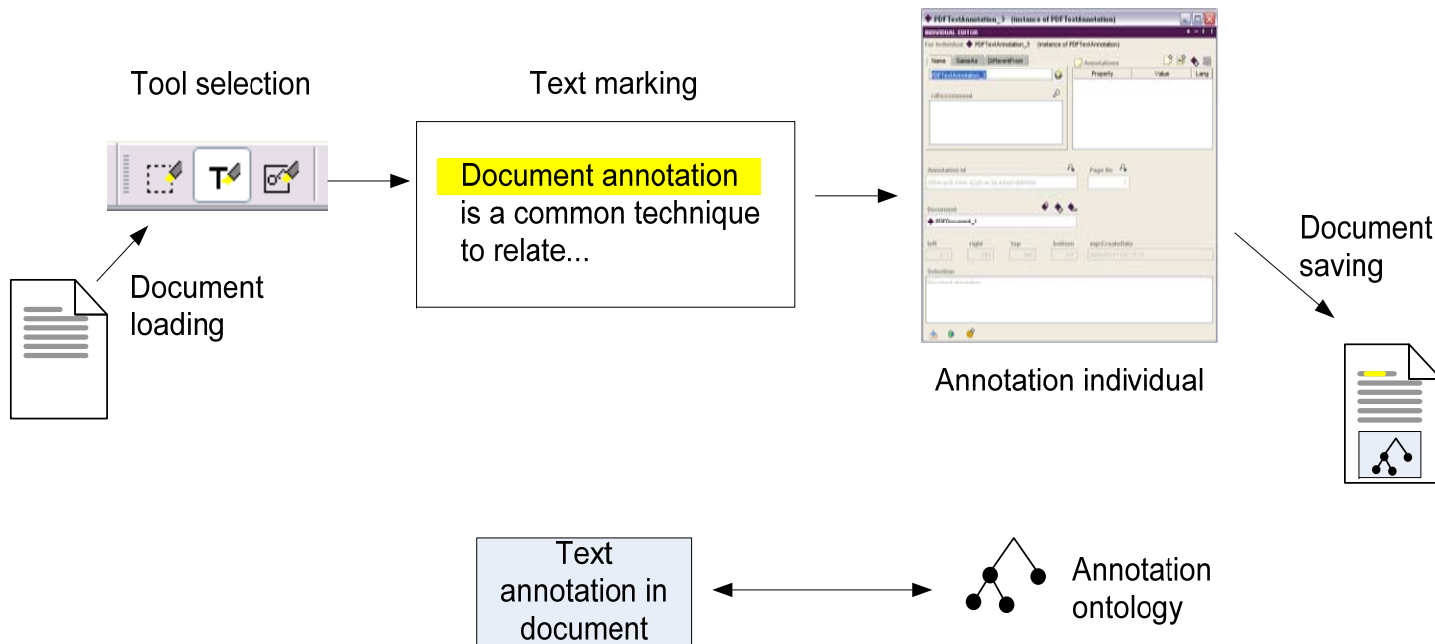
NUMMER: 5 TEXT: Länens folkmängd 31 december 2003 och befolkningsförändringar 2003

REFERENCE::  
Länens folkmängd 31 december 2003 och befolk

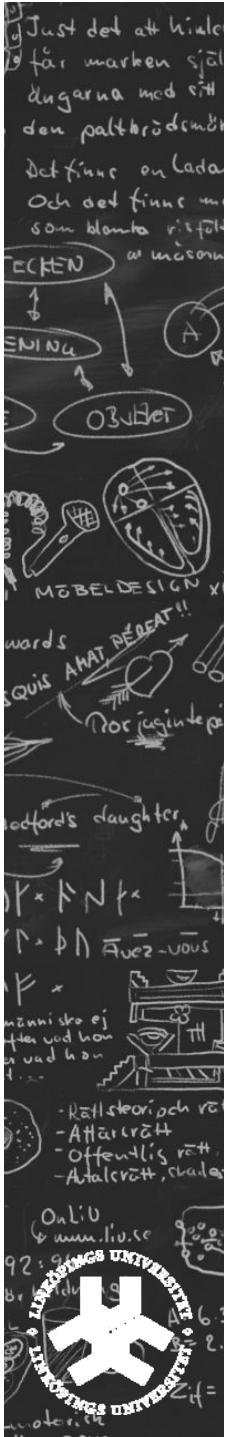
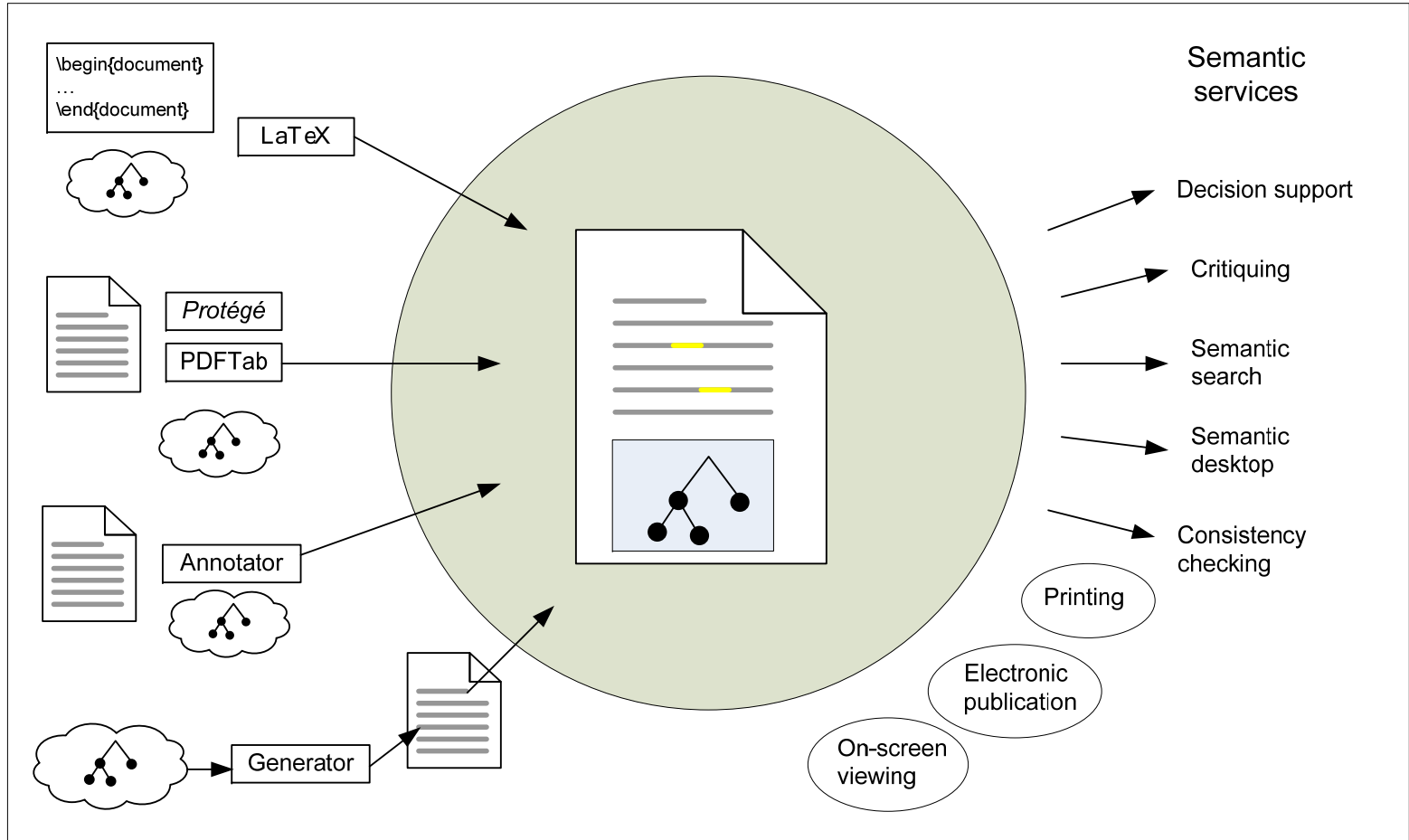
COLUMNS::  
2 Folkmängd  
3 Levande födda  
3 Folkökning  
1 Län



# Annotation Process



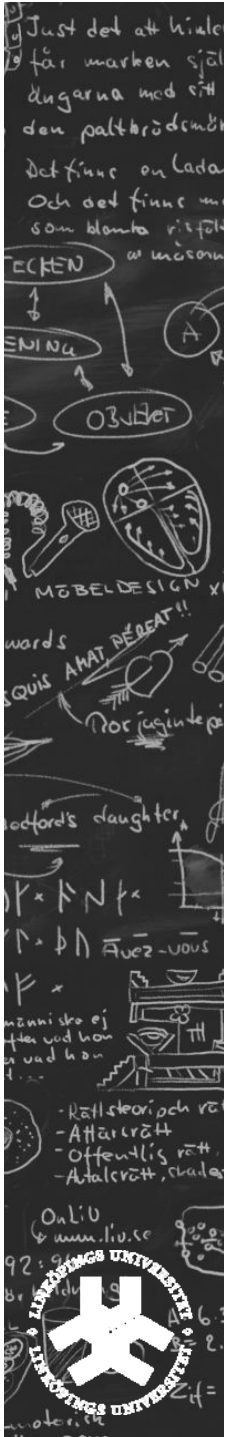
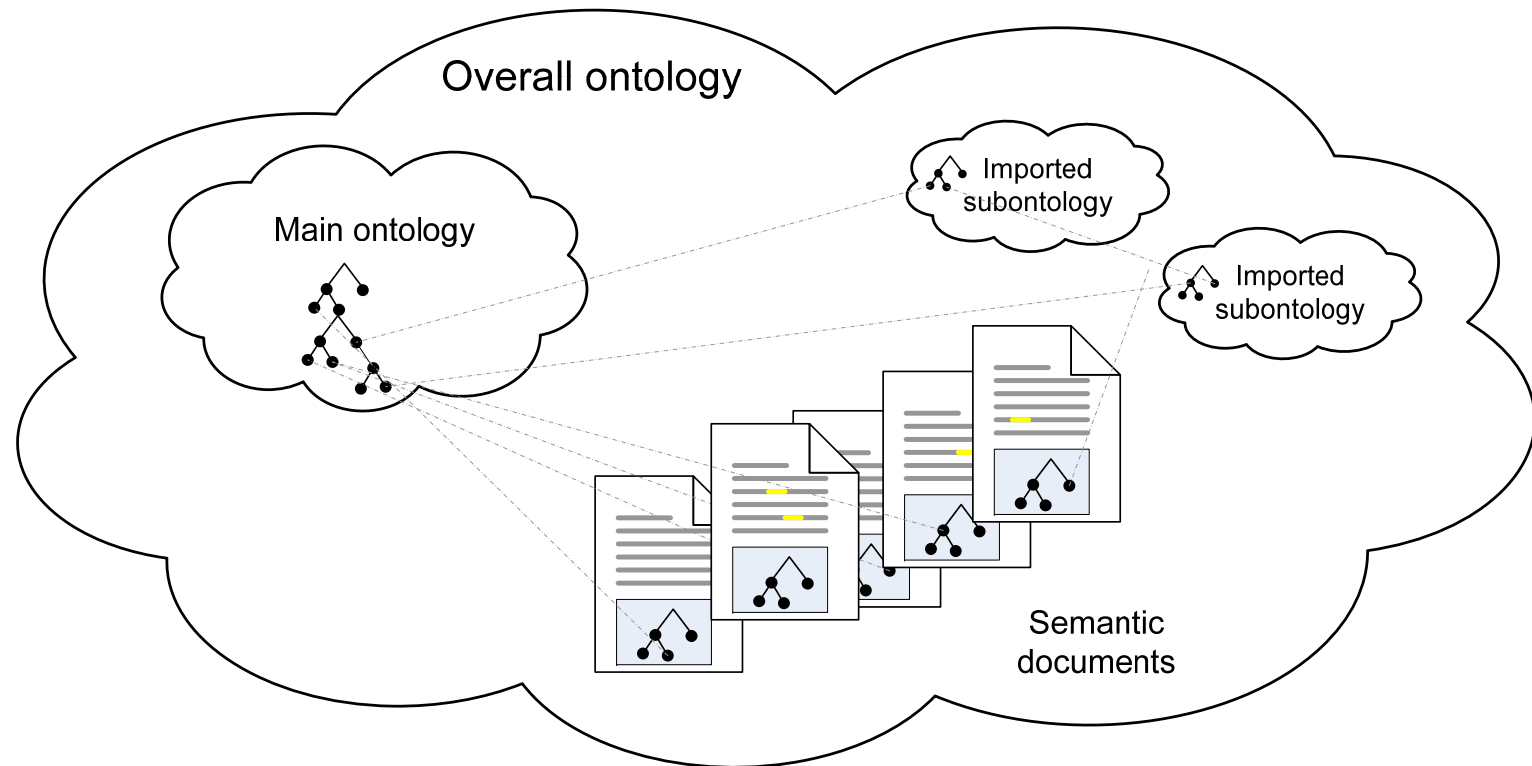
# Document-centric Annotation Framework

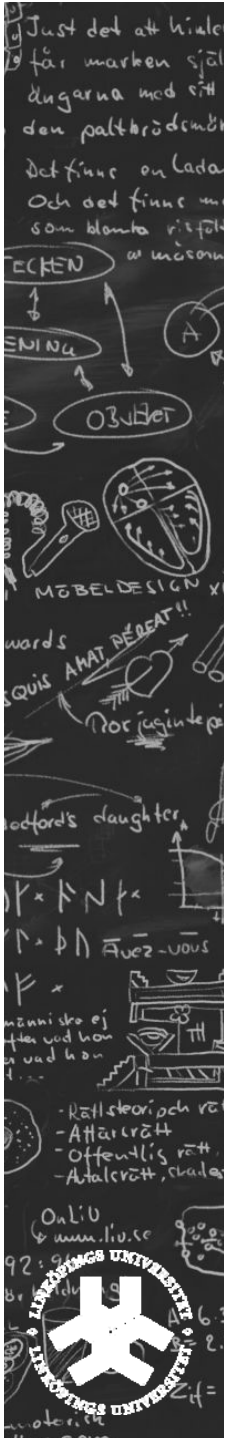




# Supporting multiple documents

- Architecture with multiple ontologies and ontology modules

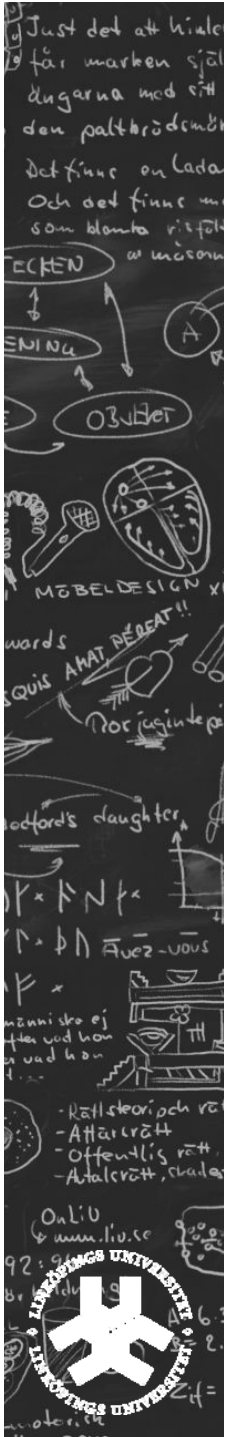




## Case Study: Document Repository in Protégé

- 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



SCB Protégé 3.3 beta (file:\C:\cygwin\home\her\Annotator\SCB.pprj, OWL / RDF Files)

File Edit Project OWL Code Tools Window Document Utilities Help

OWLClasses Properties Forms Individuals Metadata (unnamed.owl) Documents Document Search

**DOCUMENT BROWSER** For Project: SCB

Documents

- AM0102\_2006M04\_SM\_AM17SM0606\_annot.pdf
- PRO301\_2006M02\_SM\_PR10SM0603\_annot.pdf
- NV0109\_2005A01A\_SM\_NV19SM0602\_annot.pdf
- HA0201\_2006M07S\_SM\_HA17SM0607\_annot.pdf
- HA0201\_2006M03D\_SM\_HA22SM0601\_annot.pdf
- JO0601\_2006A01\_SM\_JO19SM0601\_annot.pdf
- OV0904\_175004\_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

**DOCUMENT ANNOTATOR** For Document: ME0201\_2006M05\_SM\_ME60SM0601\_annot.pdf

Meta view Annotations view Targets view PDF view

AM0301\_2006M06\_SM\_AM39SM0608\_annot.pdf ME0201\_2006M05\_SM\_ME60SM0601\_annot.pdf

83%

Sign

**1. Arbetskostnadsindex för arbetare inom privat sektor**  
**1. Labour cost index for wage-earners in the private sector**

Näringsgren SNI	Juni 2006	Förändring från juni 2005, % <sup>2</sup>	
		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

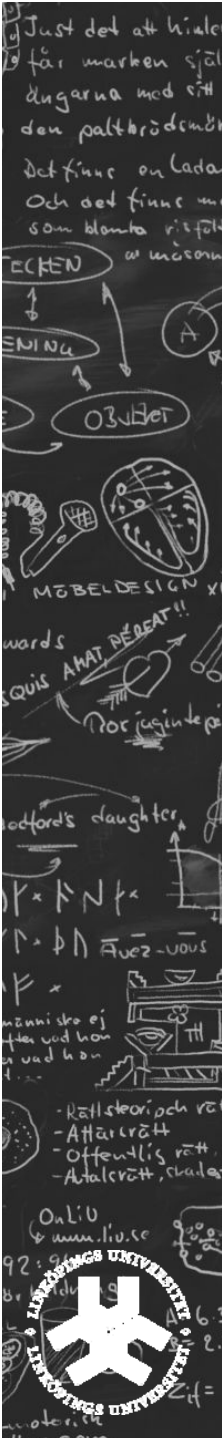
1) 1:a kvartalet 1994=100  
2) Förändringstalen beräknas på indextal med fler än en decimal.  
1) 1<sup>st</sup> kvartal 1994=100  
2) The calculation of percentage changes includes several decimals

**2. Arbetskostnadsindex för tjänstemän inom privat sektor**  
**2. Labour cost index for salaried employees in the private sector**

Näringsgren SNI	Juni 2006	Förändring från juni 2005, % <sup>2</sup>	
		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

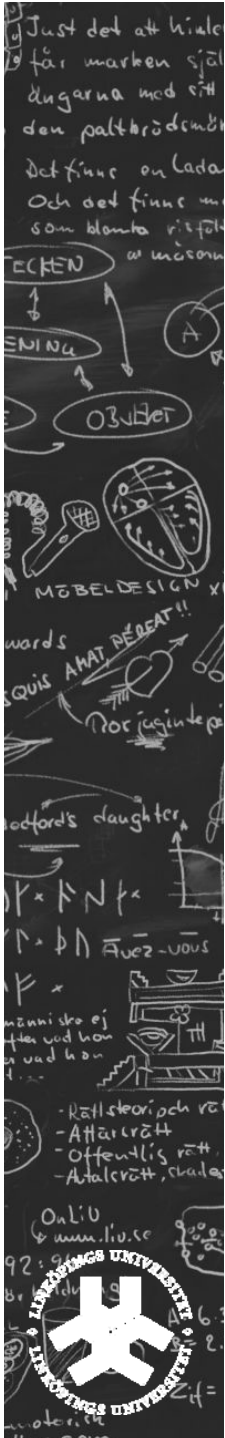
5 of 19

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