

Support for Semantic Documents in Protégé

Henrik Eriksson

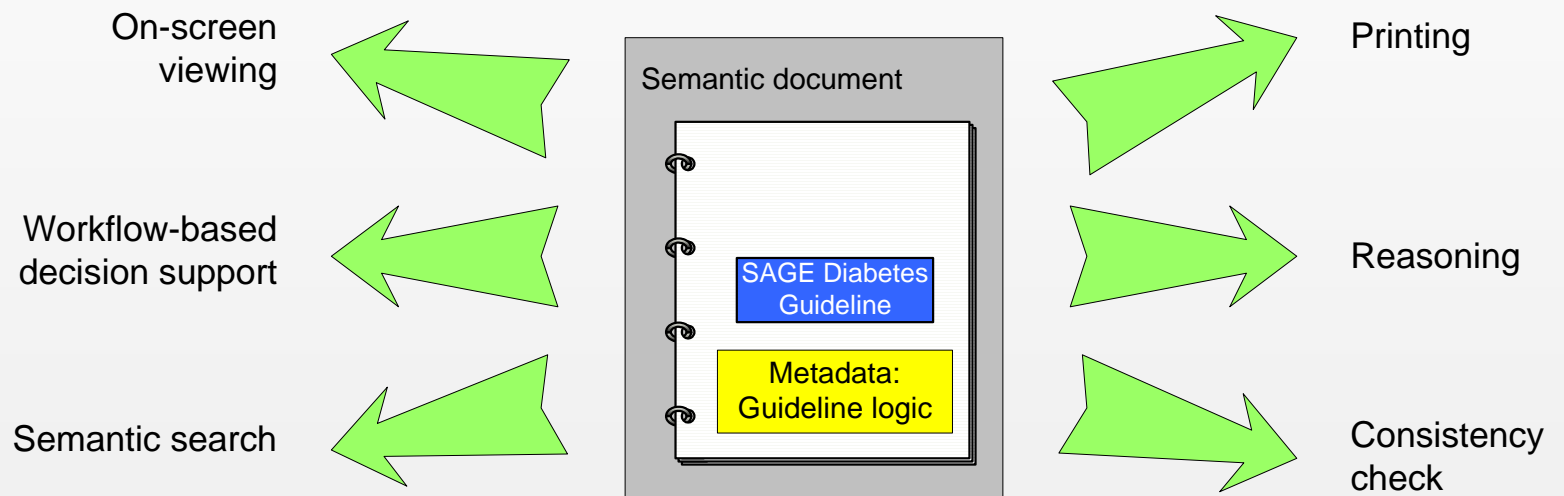
Linköping University

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

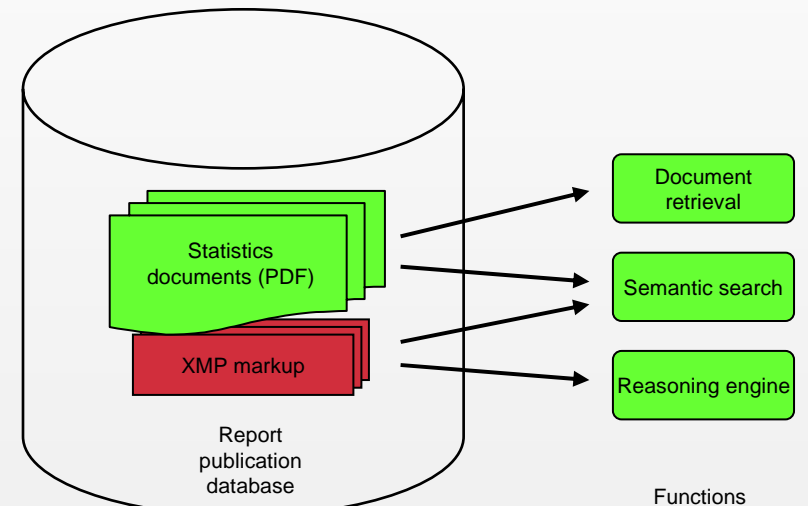


One Document—Many Applications



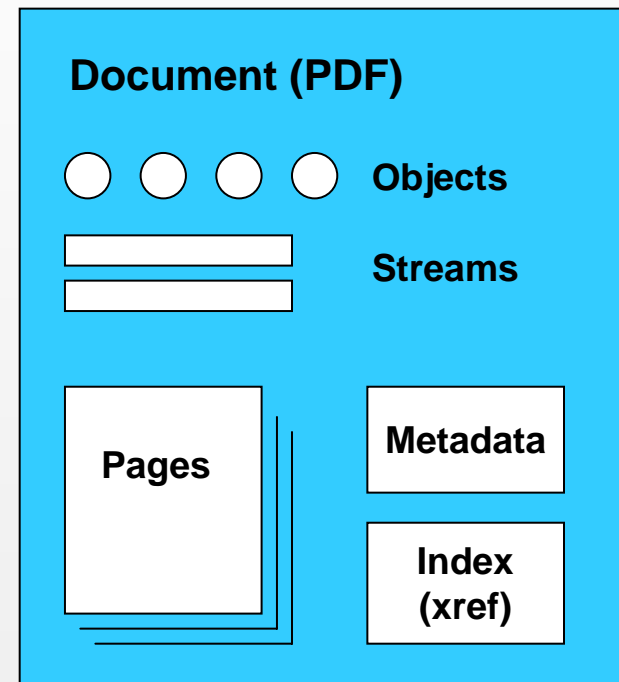
Semantic Documents

- 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

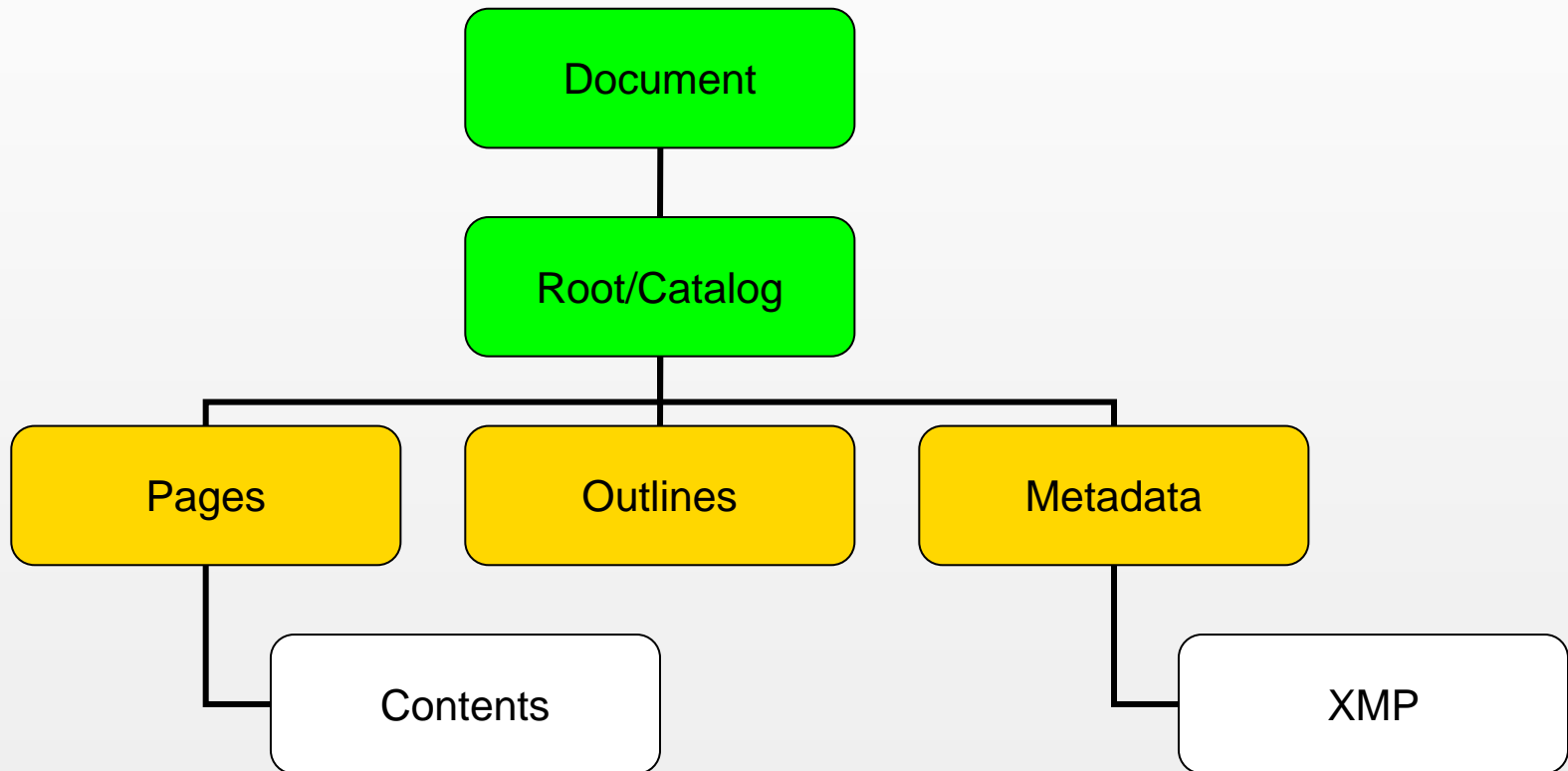


The “Secrets” of the Portable Document Format (PDF)

- 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

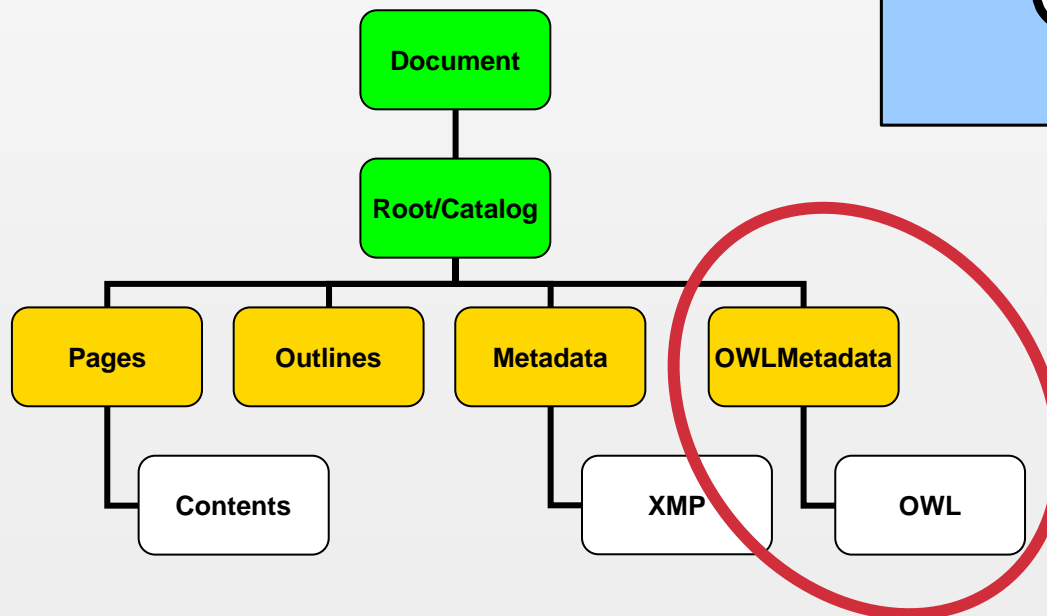
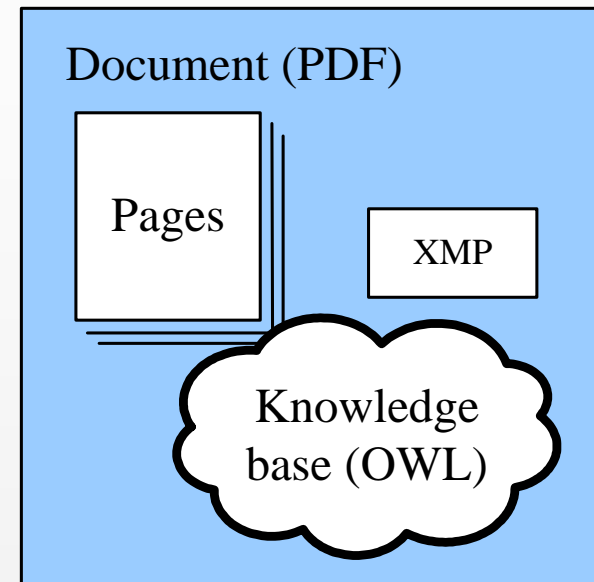


Internal PDF Structure



Adding Additional Information to the PDF Structure

- OWL-based metadata

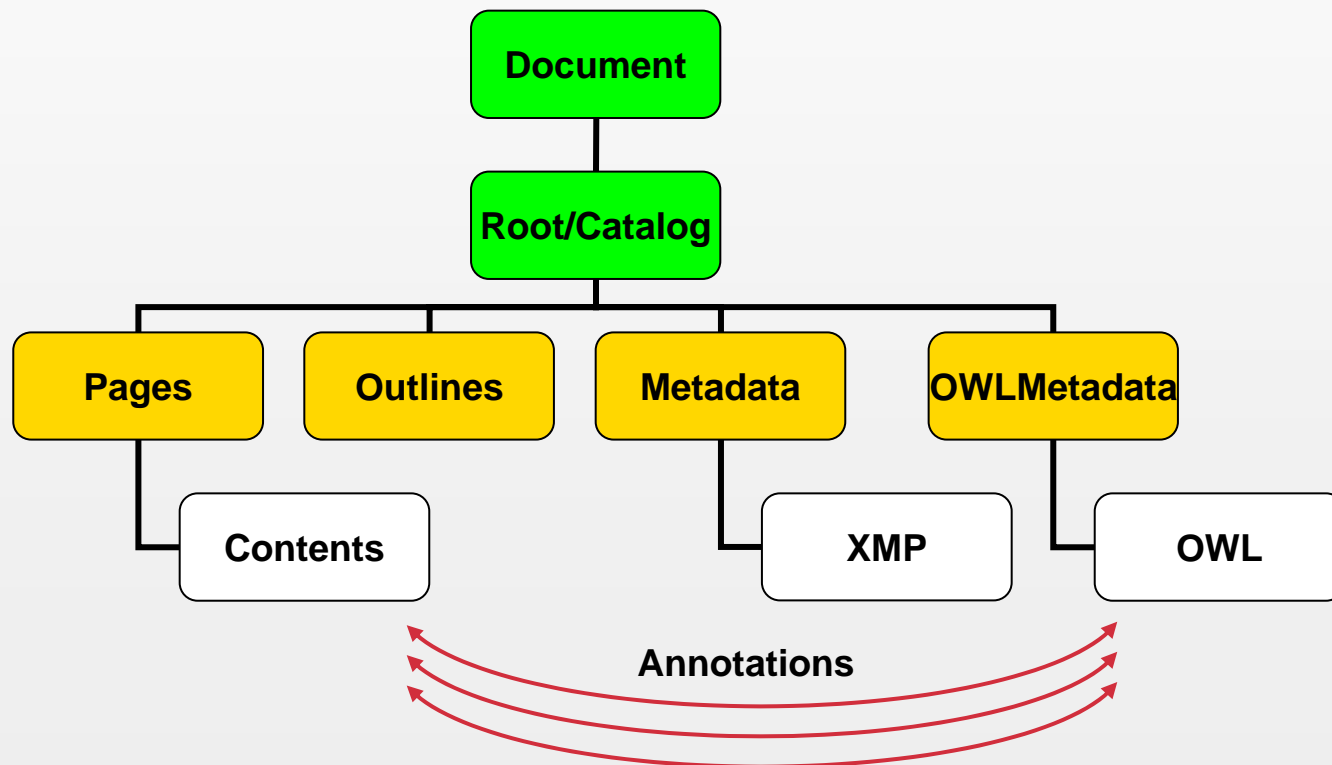


Added OWL statements



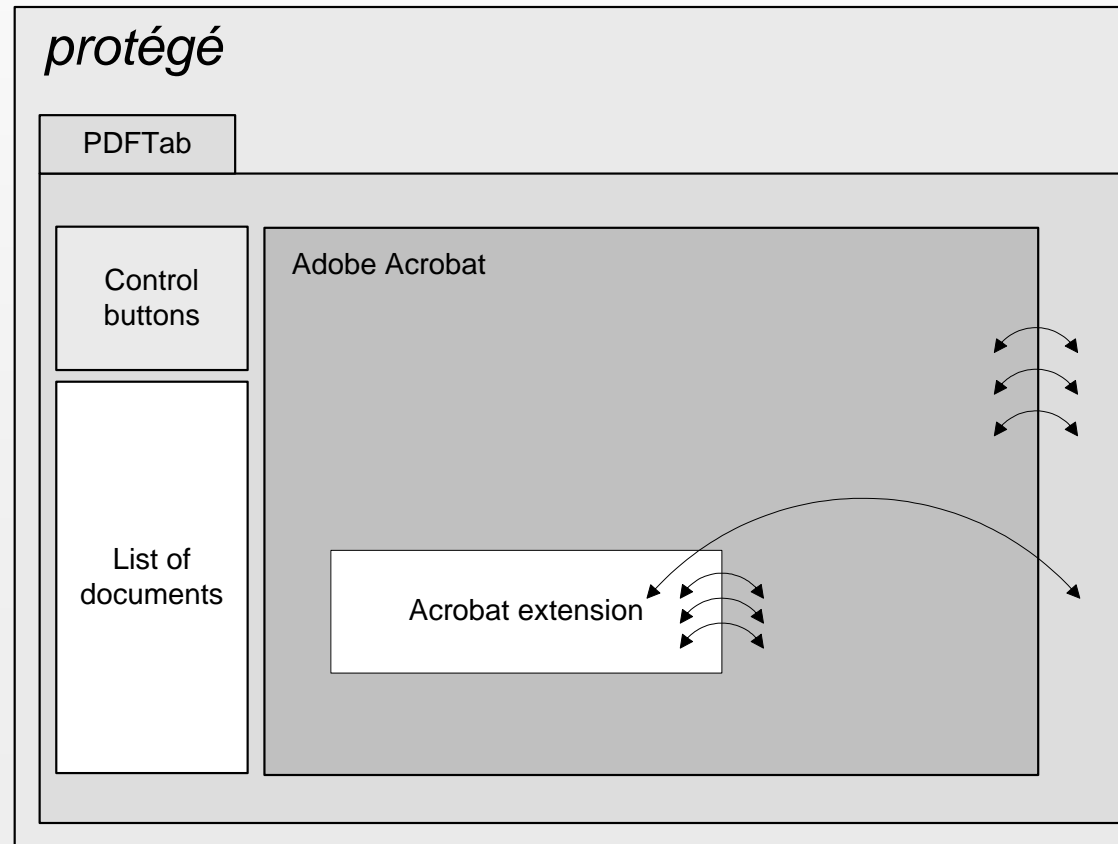
Annotations

- Relates document text to OWL individuals



A Protégé Extension for PDF

- Adobe Acrobat runs inside a Protégé tab



PDFTab: Annotation Tool for Protégé

Annotation tool

Protégé

Adobe Acrobat
(PDF)

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

File Edit Project OWL Wizards Code Window Help

OWLClasses Properties Forms Individuals Metadata PDF

PDFTab

Open... Close Remove

BE0101_2003M12_SM_BE12SM0401x1d.pdf

Meta view PDF view

BE0101_2003M12_SM_BE12SM0401x1d.pdf

SCB 22 BE 12 SM 0401

7. Kommunerna i storleksordning efter antal invånare 31 december 2003 enligt indelning 1 januari 2004

7. Municipalities by size of population on Dec. 31, 2003 according to the administrative subdivisions of Jan. 1, 2004

Folkökning				Folkökning			
Kommunerna i storleksordning	Folkmängd	Antal	Procent	Kommunerna i storleksordning	Folkmängd	Antal	Procent
1 Stockholm	761 721	3 573	0,47	56 Enköping	38 005	358	-0,39
2 Göteborg	478 055	3 134	0,66	57 Ängelholm	37 859	153	-0,17
3 Malmö	267 171	1 690	0,64	58 Upplands Väsby	37 397	47	-0,17
4 Uppsala	180 699	996	0,55	59 Lidingö	37 125	184	-0,17
5 Linköping	136 231	1 165	0,86	60 Vänersborg	37 101	76	-0,17
6 Västerås	129 987	1 085	0,84	61 Hudiksvall	37 057	9	-0,17
7 Örebro	126 288	768	0,61	62 Sandviken	36 817	52	-0,17
8 Norrköping	123 971	668	0,54	63 Västervik	36 768	-145	-0,39
9 Helsingborg	120 154	748	0,63	64 Österåker	36 183	550	1,54
10 Jönköping	119 340	759	0,64	65 Sigtuna	36 028	267	0,72
11 Umeå	107 917	1 392	1,31	66 Lerum	35 890	332	0,93
12 Lund	100 995	593	0,59	67 Alingsås	35 530	203	0,57
13 Borås	98 505	355	0,36	68 Sundbyberg	33 738	-59	-0,17
14 Sundsvall	93 307	55	0,06	69 Mark	33 218	203	0,61
15 Gävle	91 701	425	0,47	70 Partille	33 192	104	0,31
16 Eskilstuna	90 894	605	0,67	71 Värmdö	33 134	470	1,44
17 Halmstad	87 372	797	0,91	72 Katrineholm	32 418	37	0,11
18 Huddinge	87 121	664	0,77	73 Värnamo	32 350	96	0,30
19 Karlstad	81 343	409	0,51	74 Vellinge	31 300	213	0,69
20 Södertälje	80 049	436	0,55	75 Härryda	31 208	364	1,18
21 Nacka	77 470	846	1,10	76 Falköping	30 981	85	0,28
22 Växjö	76 848	812	1,08	77 Karlshamn	30 779	40	0,13
23 Botkyrka	76 432	216	0,29	78 Karlskoga	30 532	-68	-0,22

8,26 x 11,69 in

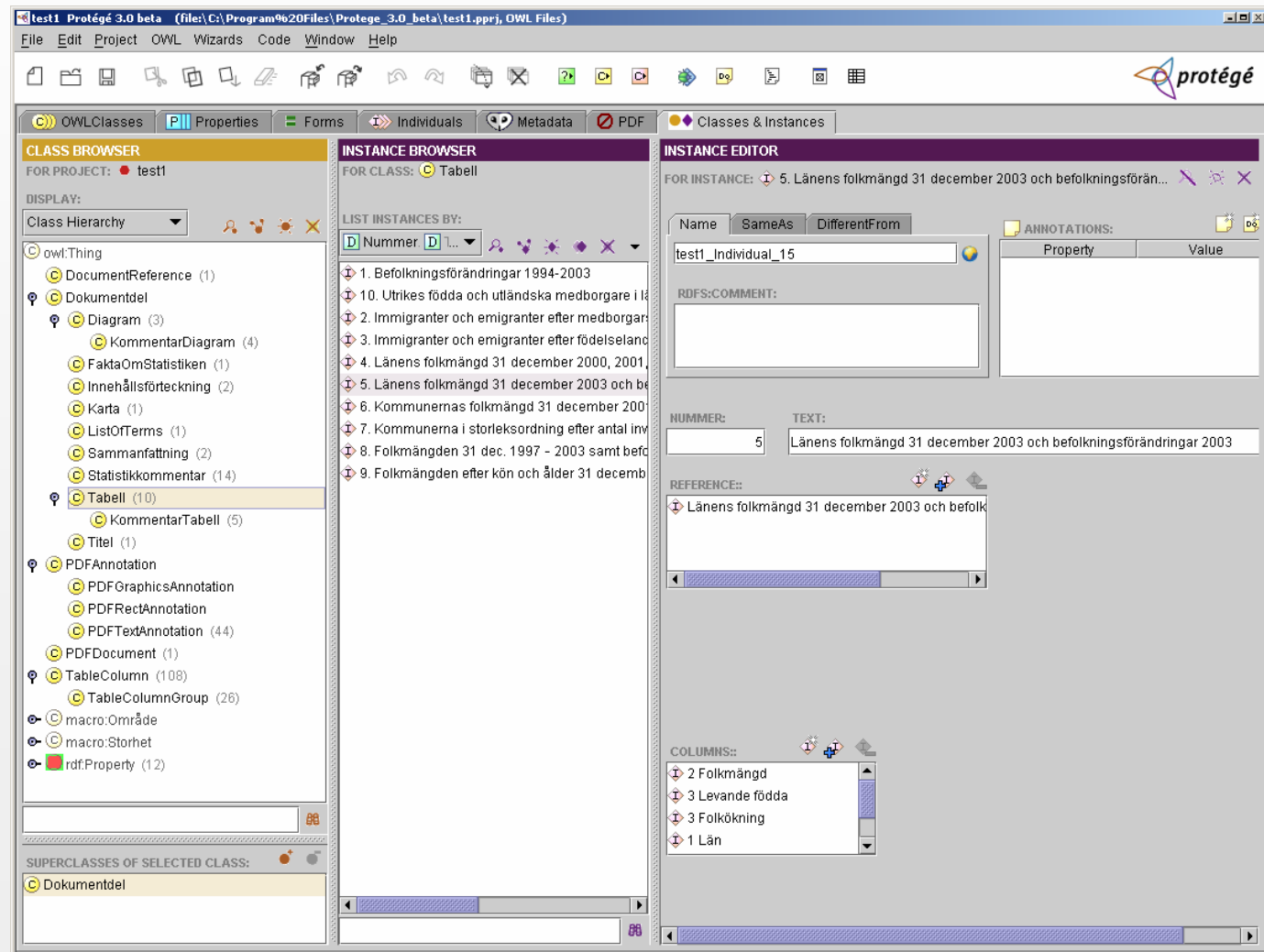
22 of 40

C:\Documents and Settings\herMy Documents\BE0101_2003M12_SM_BE12SM0401x1d.pdf



Linköpings universitet

Corresponding Ontology



Mark up of Table Headings

20Files\Protege_3.0_beta\test1.pprj, OWL Files)

Window Help

Forms Individuals Metadata PDF

Meta view PDF view

BE0101_2003M12_SM_BE12SM0401x1d.pdf

98%

SCB 22

7. Kommunerna i storleksordning efter antal invånare 31 december 2003 enligt indelning 1 januari 2004

7. Municipalities by size of population on Dec. 31, 2003 according to subdivisions of Jan. 1, 2004

Kommunerna i storleksordning	Folkmängd	Antal	Folkökning	Kommunerna i storleksordning
1 Stockholm	781 721	3 573		5 Vänersborg
2 Göteborg	478 055	3 134		
3 Malmö	267 171	1 890		
4 Uppsala	180 889	998	0,55	
5 Linköping	138 231	1 165	0,88	

INDIVIDUAL EDITOR

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

Name SameAs DifferentFrom

test1_Individual_17

RDFS:COMMENT:

NUMMER: 7

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

REFERENCE:

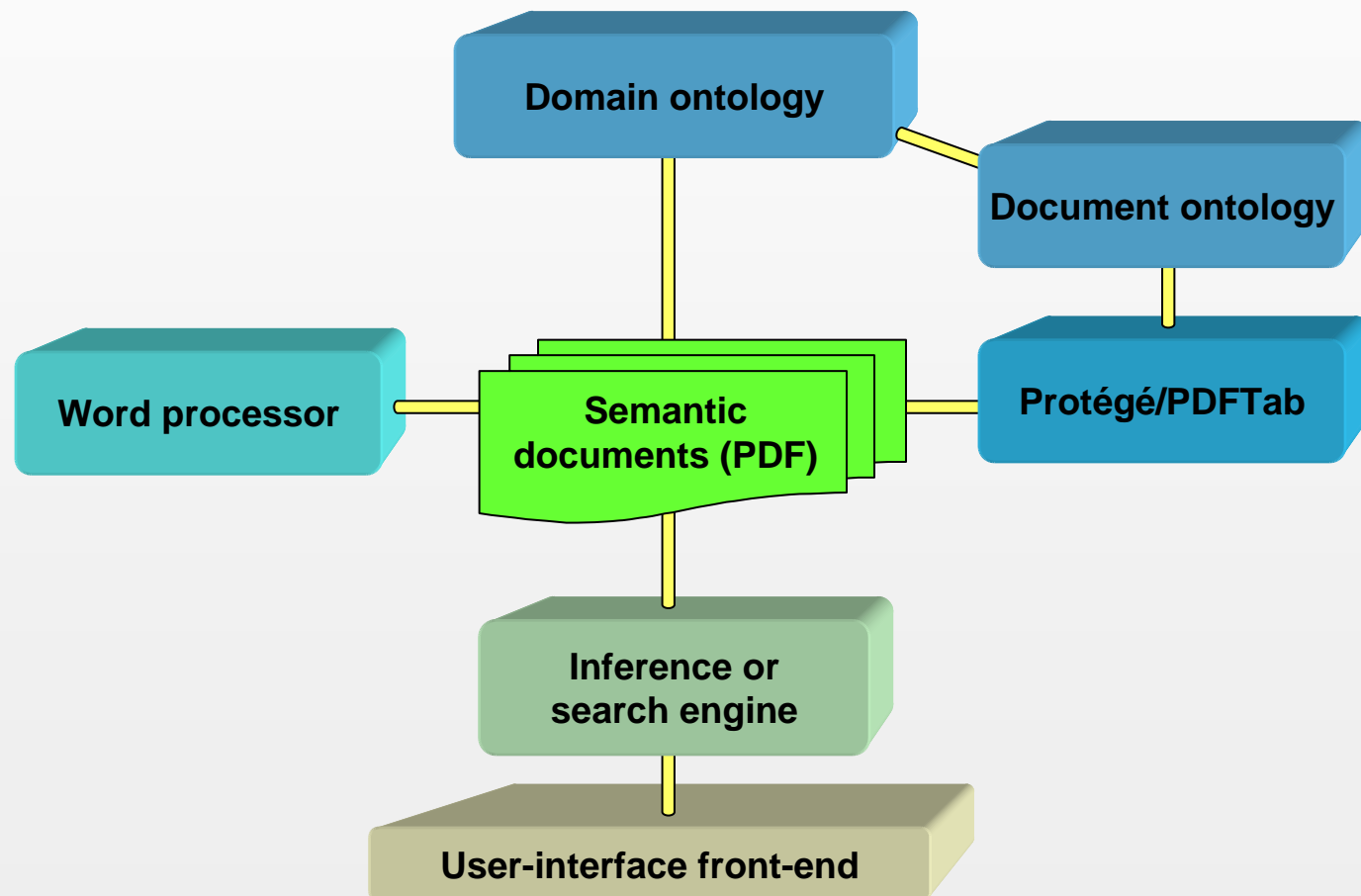
Kommunerna i storleksordning efter antal invånare 31 december 2003 enligt indelning 1 januari 2004

COLUMNS:

- 3 Folkökning
- 2 Folkmängd
- 1 Kommunerna i storleksordning

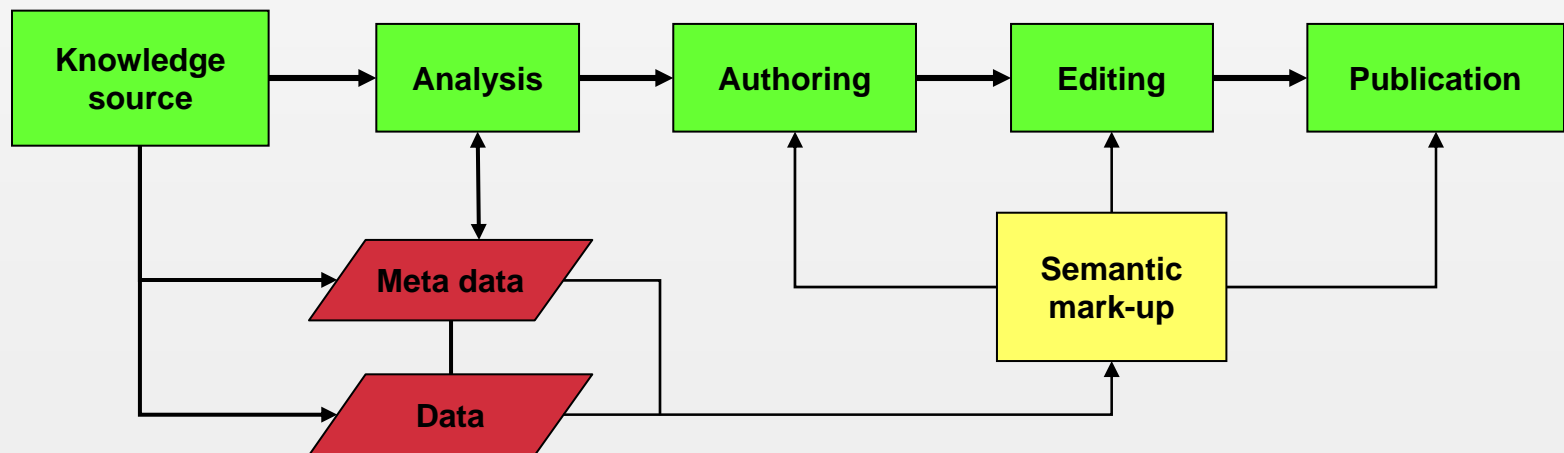


A Semantic Document Architecture for Knowledge Management



Document Production Process

- Basic idea: Tool support for the entire chain
 - Knowledge-management approach
 - Metadata is kept throughout the process
 - Support for annotation (tagging) based on data sources, including metadata



Application Areas

- **Statistics**
 - Annotation of statistics reports
 - Highly structured documents with tables and diagram
 - Report series (e.g., quarterly and annual reports)
 - Collaboration with Statistics Sweden (SCB)
- **Clinical guidelines**
 - Generation of documentation from SAGE knowledge bases
 - Highly structured documents with graphs and cross links
 - Target: Guideline documents in PDF complete with annotations
 - Collaboration with Samson Tu, Stanford University
- **Document search**
 - Searching text and metadata
 - Different levels of search
 - Test case: Statistics reports



Statistics Reports as Semantic Documents

- Statistics Reports
- Statistical Yearbook of Sweden (784 pages)
- Manual and (semi-)automated annotation
- Statistical metadata available
- Development of relevant ontologies
 - Annotation ontology
 - Document ontology
 - Macro data ontology
 - Domain ontology
 - In general, an ontology of the entire country!
- Interesting idea: Use annotation of the previous document edition as the starting point



Mark-up of Statistical Yearbook

yearbook Protégé 3.0 beta (file: C:\cygwin\home\her\Annotator\yearbook.pprj, OWL Files)

File Edit Project OWL Wizards Code Window Help

OWLClasses Properties Forms Individuals Metadata PDF

PDFTab

Open... Close Remove

yearbook2005.pdf

Options

Bookmarks

- Förord
- Teckenförklaring
- Kvalitetsdeklaration
- Innehåll
- Kartor
- Geografiska uppgifter
- Miljö och väder
- Befolkning
- Jordbruk, skogsbruk
- Näringsverksamhet
- Energi
- Boende, byggande o
- Handel med varor oc
- Transporter och kom
- Informations och kon
- Arbetsmarknad
- Hushållens ekonomi
- Priser och konsumtio
- Nationalräkenskaper
- Offentlig ekonomi
- Finansmarknad
- Socialförsäkring

Layers Pages Signatures Comments

67 Spädbarnsdödligheten per 1 000 levande födda 1751/55–2000/01
Infant deaths (under 1 year of age), rate per 1 000 live births

Döda per 1 000

1751/55 1801/05 1851/55 1901/05 1951/55 2000/01

Vid mitten av 1700-talet var spädbarnsdödligheten mycket hög och vart femte barn dog före ett års ålder. Spädbarnsdödligheten var kvar på denna höga nivå fortfarande vid början av 1800-talet men sedan började dödligheten sjunka snabbt. Omkring år 1860 hade spädbarnsdödligheten sjunkit så mycket att endast 15 procent av pojkarna och 13 procent av flickorna avled före ett års ålder. Hundra år senare var det endast 2 procent av pojkarna och 1 procent av flickorna som avled före ett års ålder. År 2002 var det endast 4 barn av 1 000 som avled före ett års ålder.

Källa: SCB, Befolkningsutvecklingen under 250 år, Historisk statistik för Sverige, Demografrapporter 1999:2, Befolkningsstatistik, Del 4 (1991–2002).

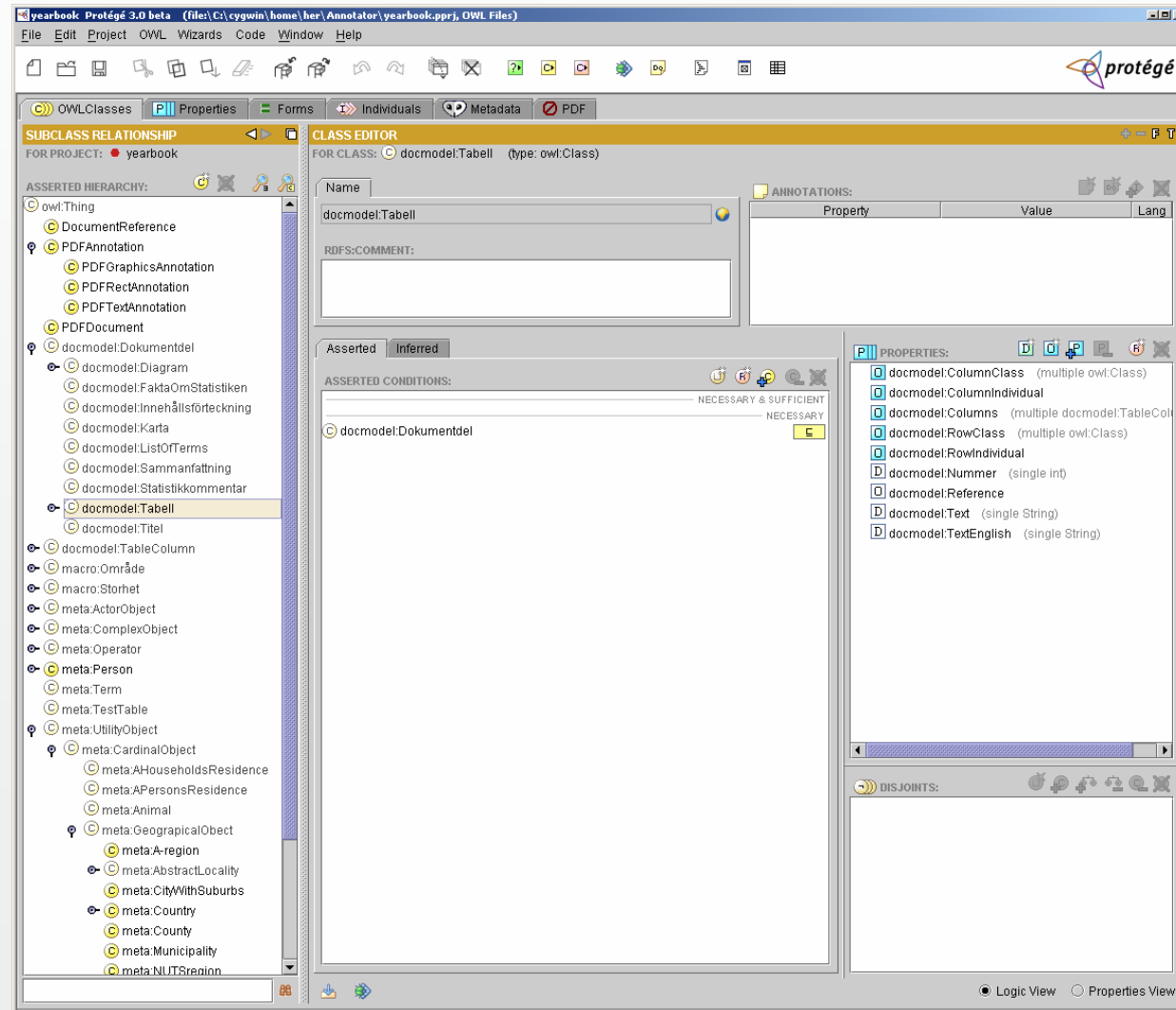
6,5 x 9,53 in

77 of 784

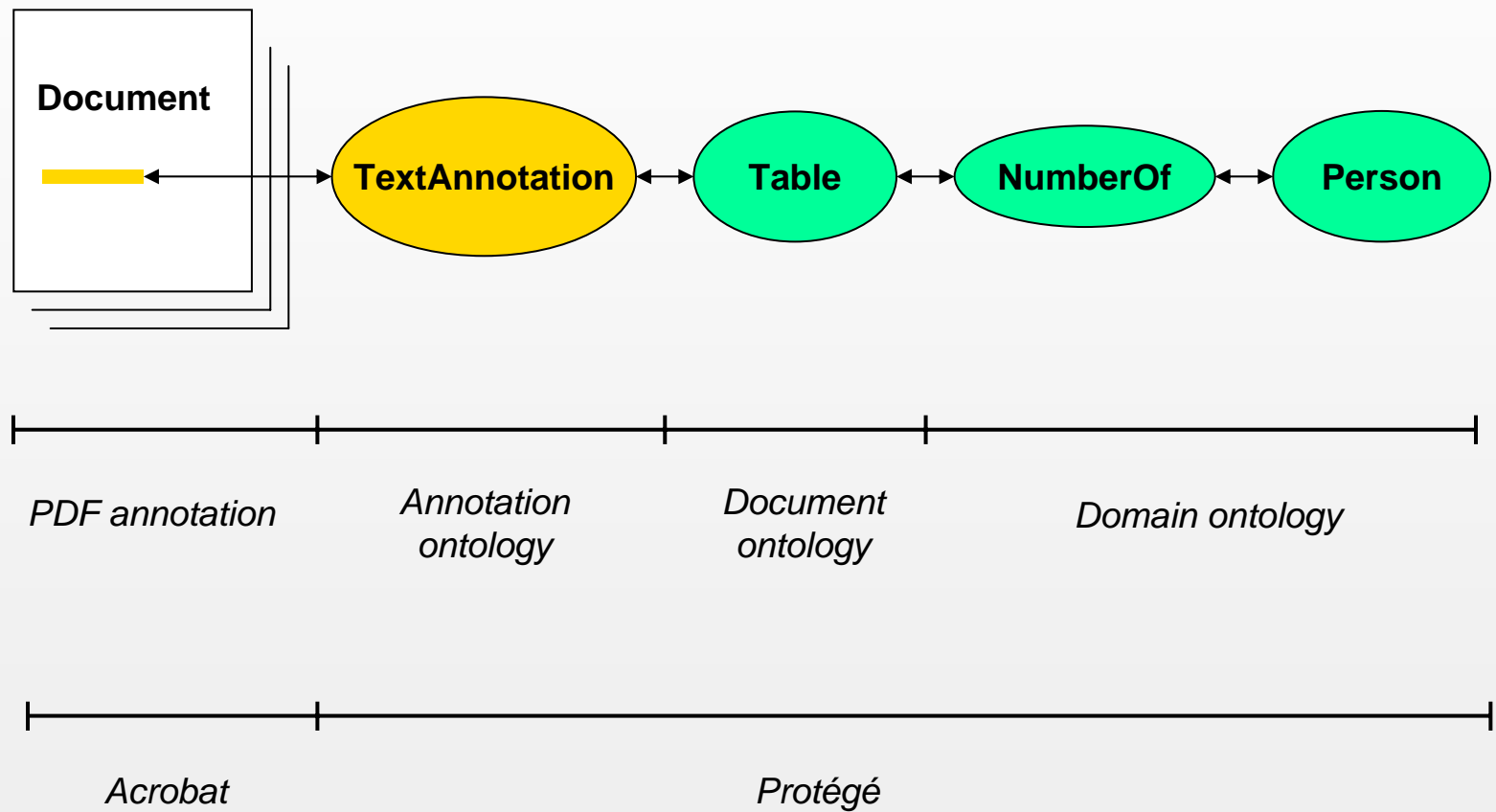
C:\cygwin\home\her\Annotator\yearbook2005.pdf



Statistics Ontologies in Protégé



Document and Domain Modeling



Questions to the OWL Experts...

1. How would you model thinks like:

- “Asylum applicants, rejections at border and persons granted residence permits as refugees or similar, by basis of residence permit,” or
- “Number of divorces in each marriage cohort by number of years since marriage”?

2. How you then search for this information?



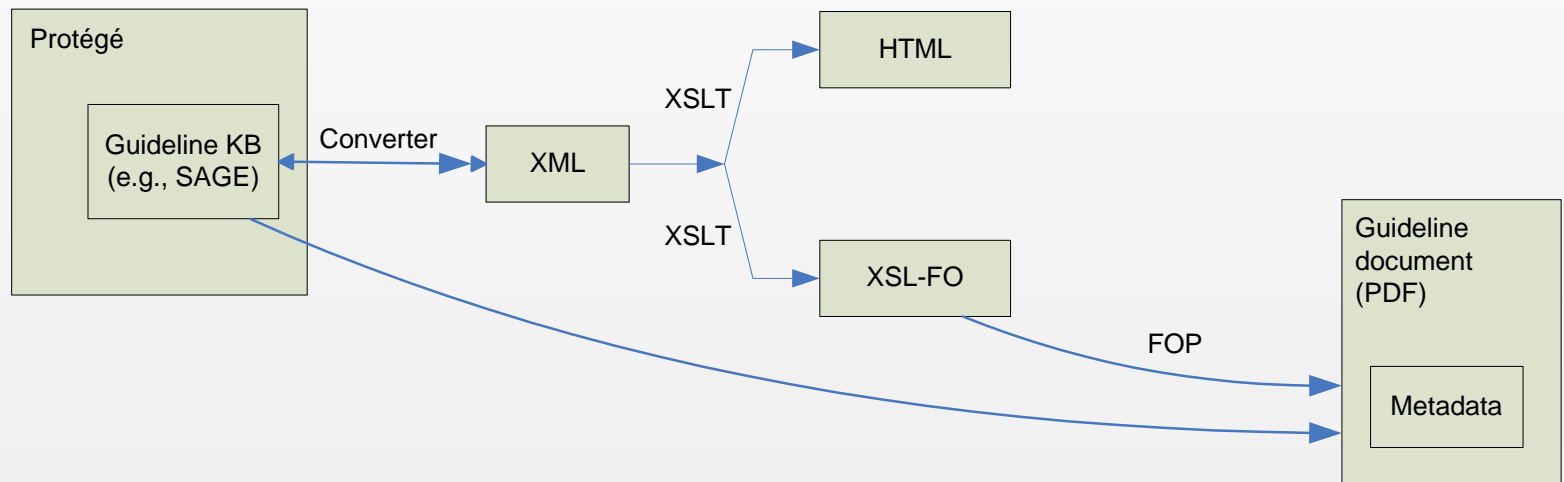
Clinical Guidelines as Semantic Documents

- Experiments with SAGE clinical guideline knowledge bases in collaboration with Samson Tu
- SAGE uses knowledge bases to store authoritative guidelines
- Uses of the knowledge bases
 - Inference
 - Workflow engines
 - Generation of guideline documentation (XML, HTML, and PDF)
- Goal: Semantic document with the knowledge base
 - PDF file with annotations and embedded SAGE knowledge base

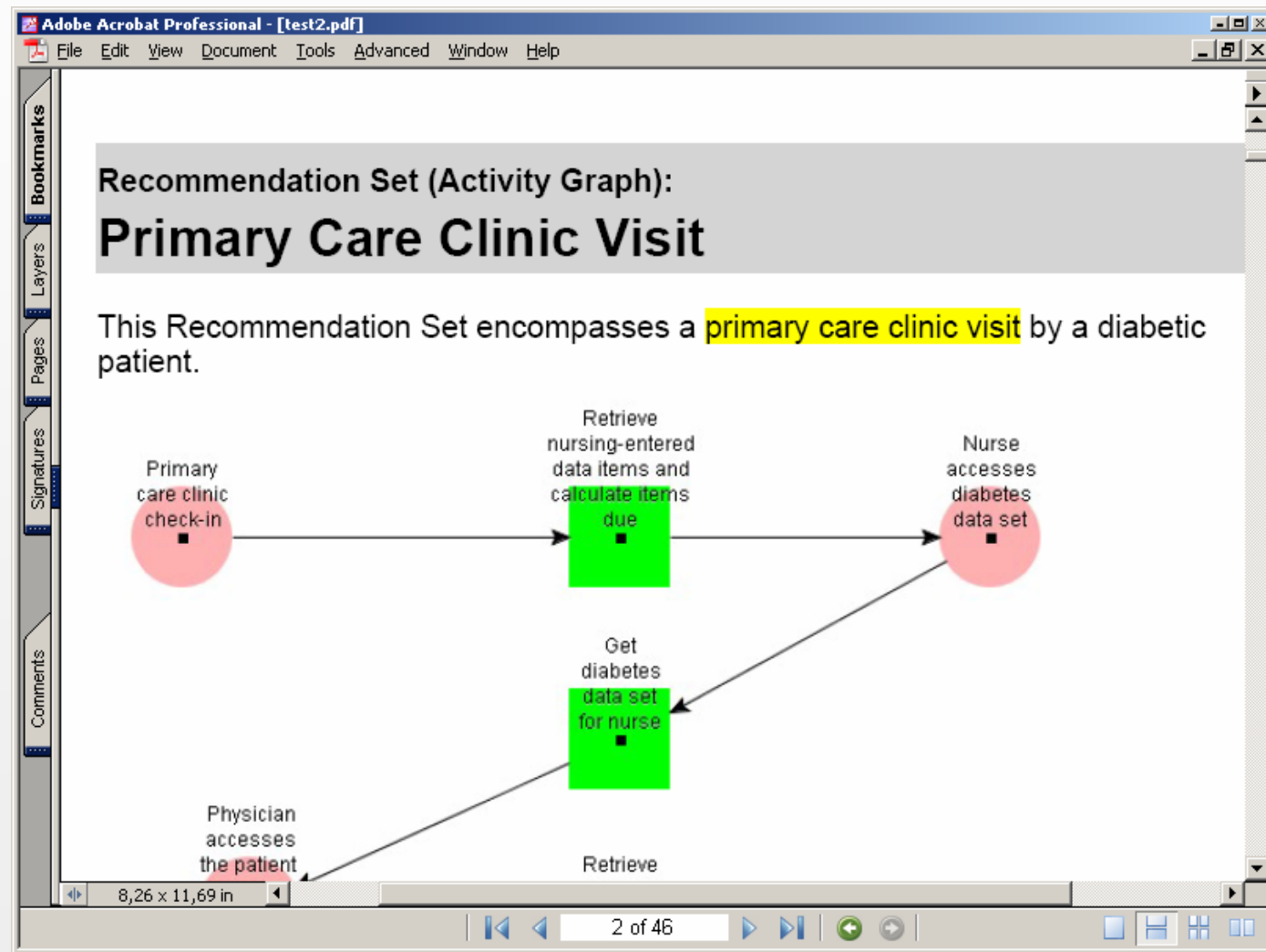


Document Generation from XML

- Generation of guideline documentation in PDF



The Resulting Guideline Document



Summary

- **Semantic documents**
 - An approach to combining printable documents with ontologies and knowledge bases
 - Combined documentation (human-readable) and reasoning (machine-readable)
 - One document with several applications
- **Tool support: PDFTab**
 - Creation of semantic documents
 - Support for document annotation
 - Editing of ontologies and knowledge bases stored in PDF files

