Support for Semantic Documents in Protégé

Henrik Eriksson

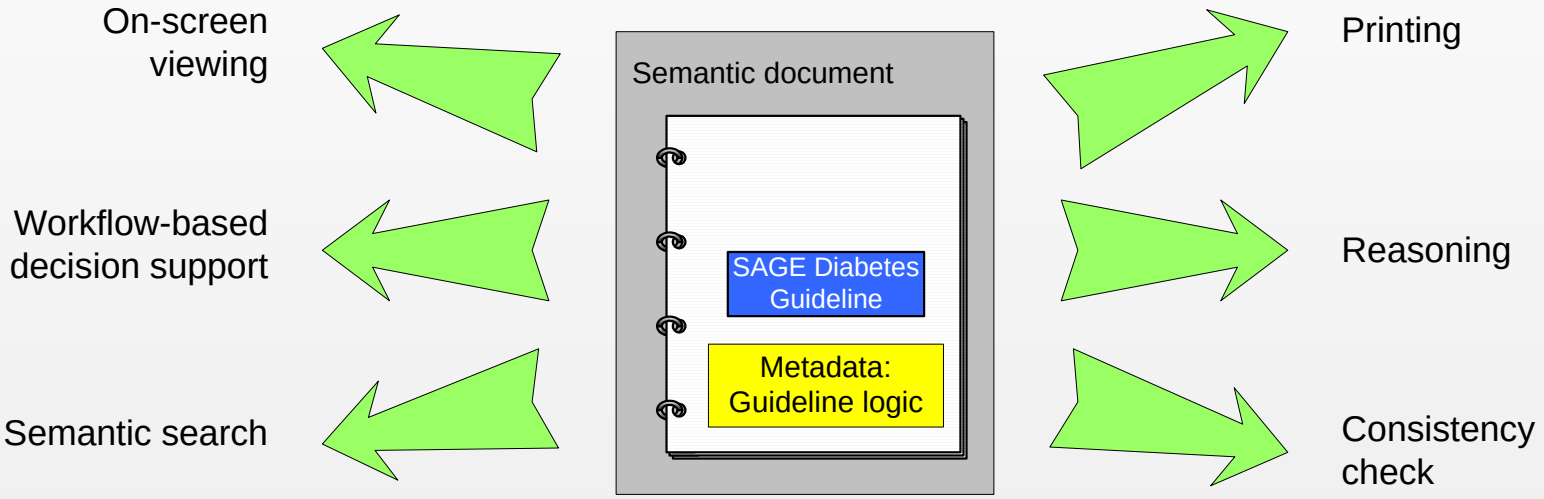
Linköping University

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

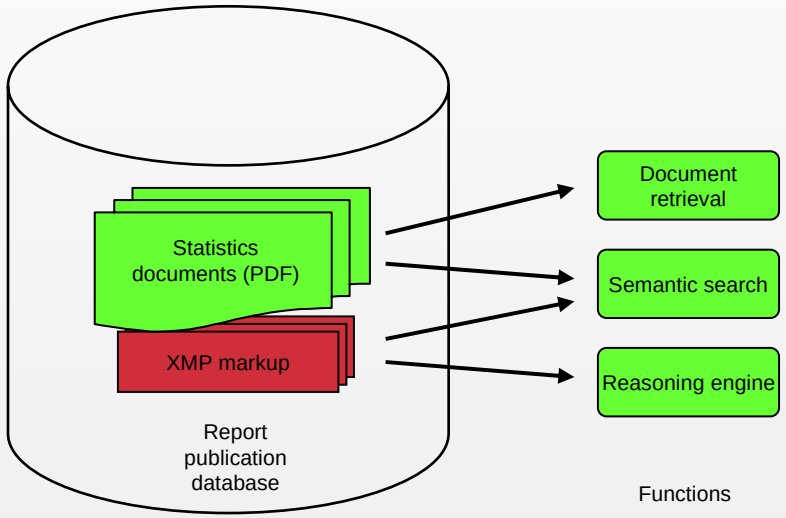
Linköpings universitet

One Document—Many Applications



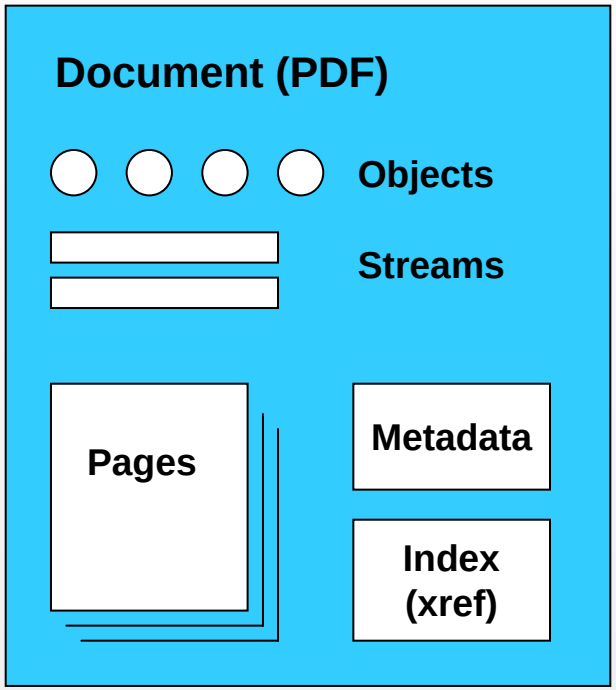
Semantic Documents

- 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

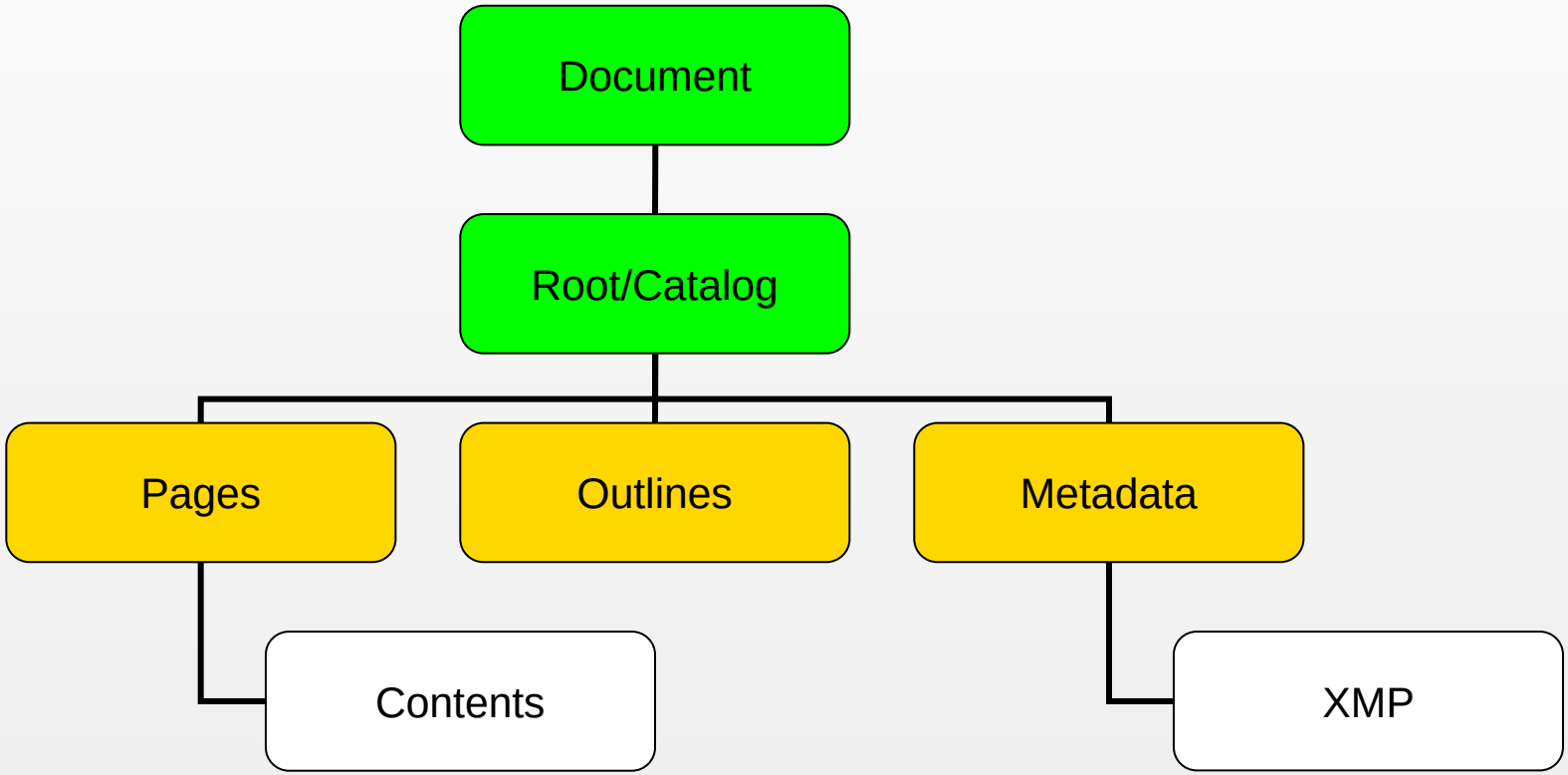


The “Secrets” of the Portable Document Format (PDF)

- 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

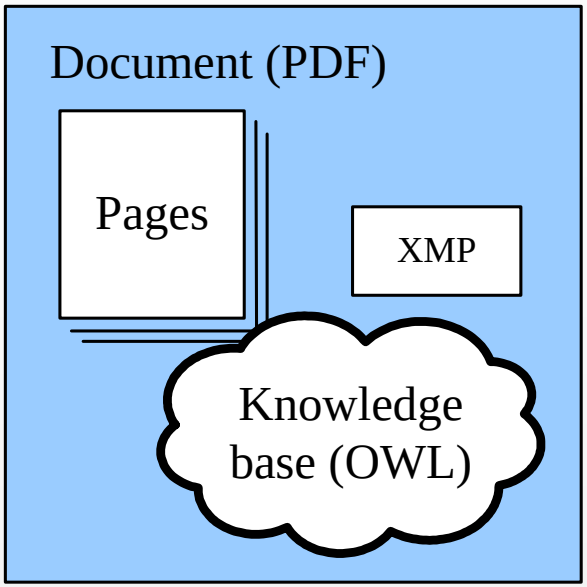


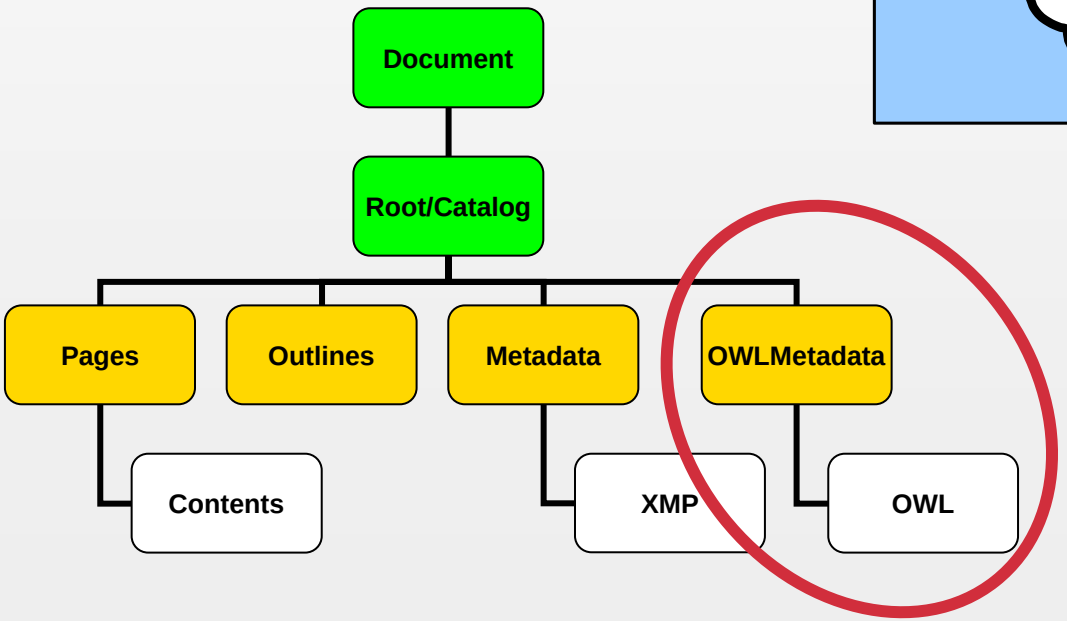
Internal PDF Structure



Adding Additional Information to the PDF Structure

- OWL-based metadata

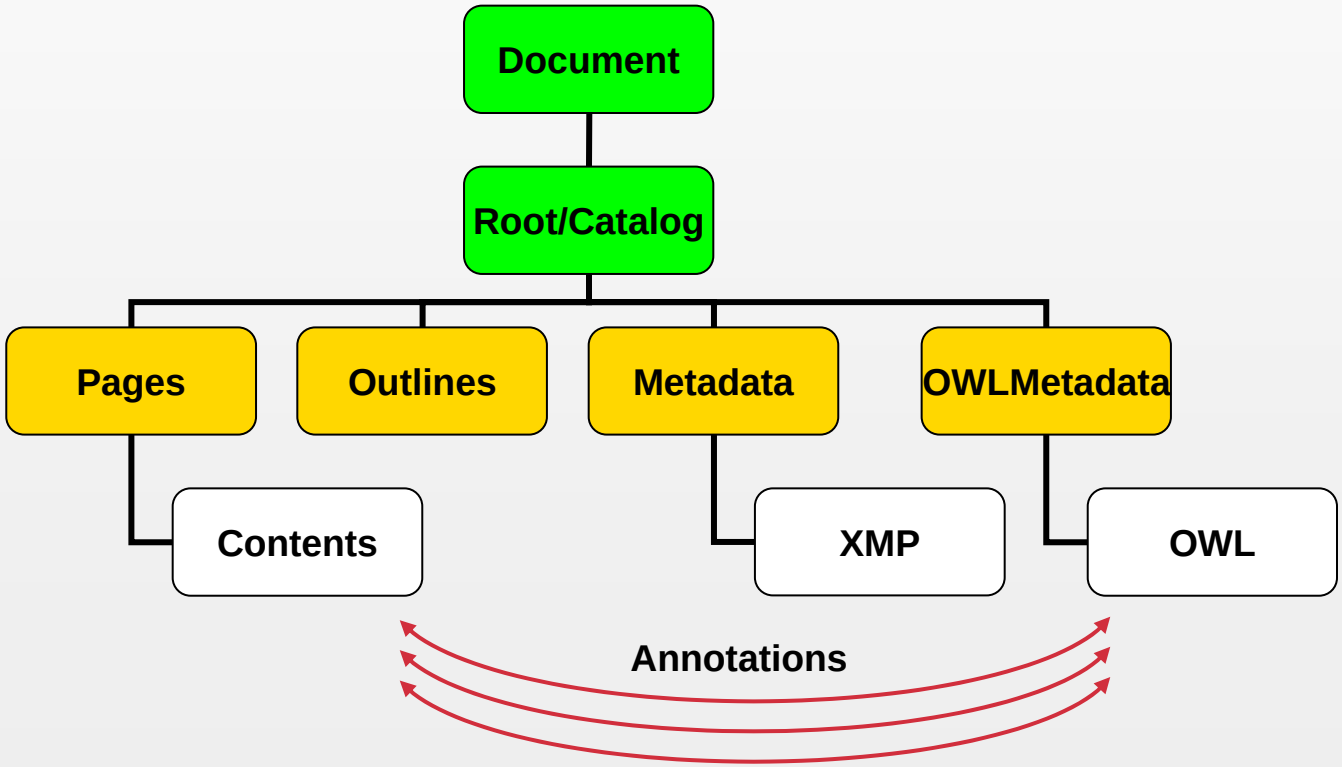




Added OWL statements

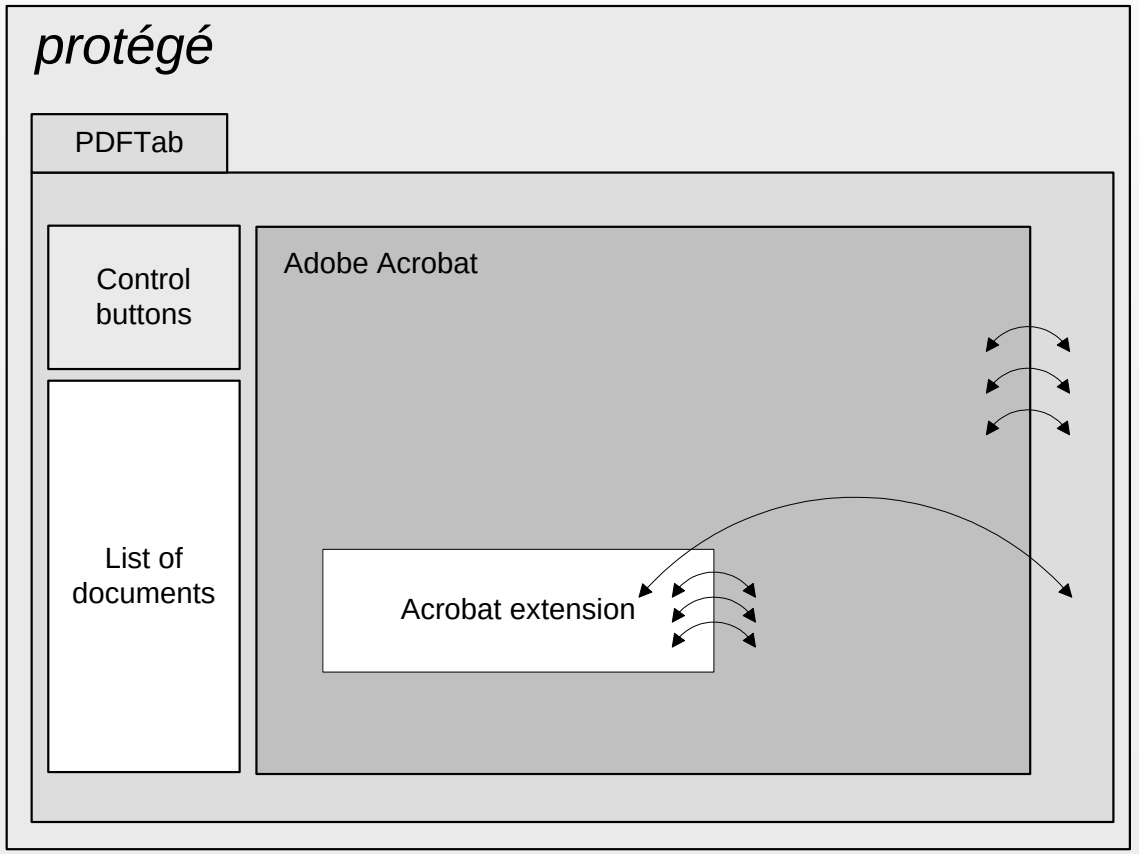
Annotations

- Relates document text to OWL individuals

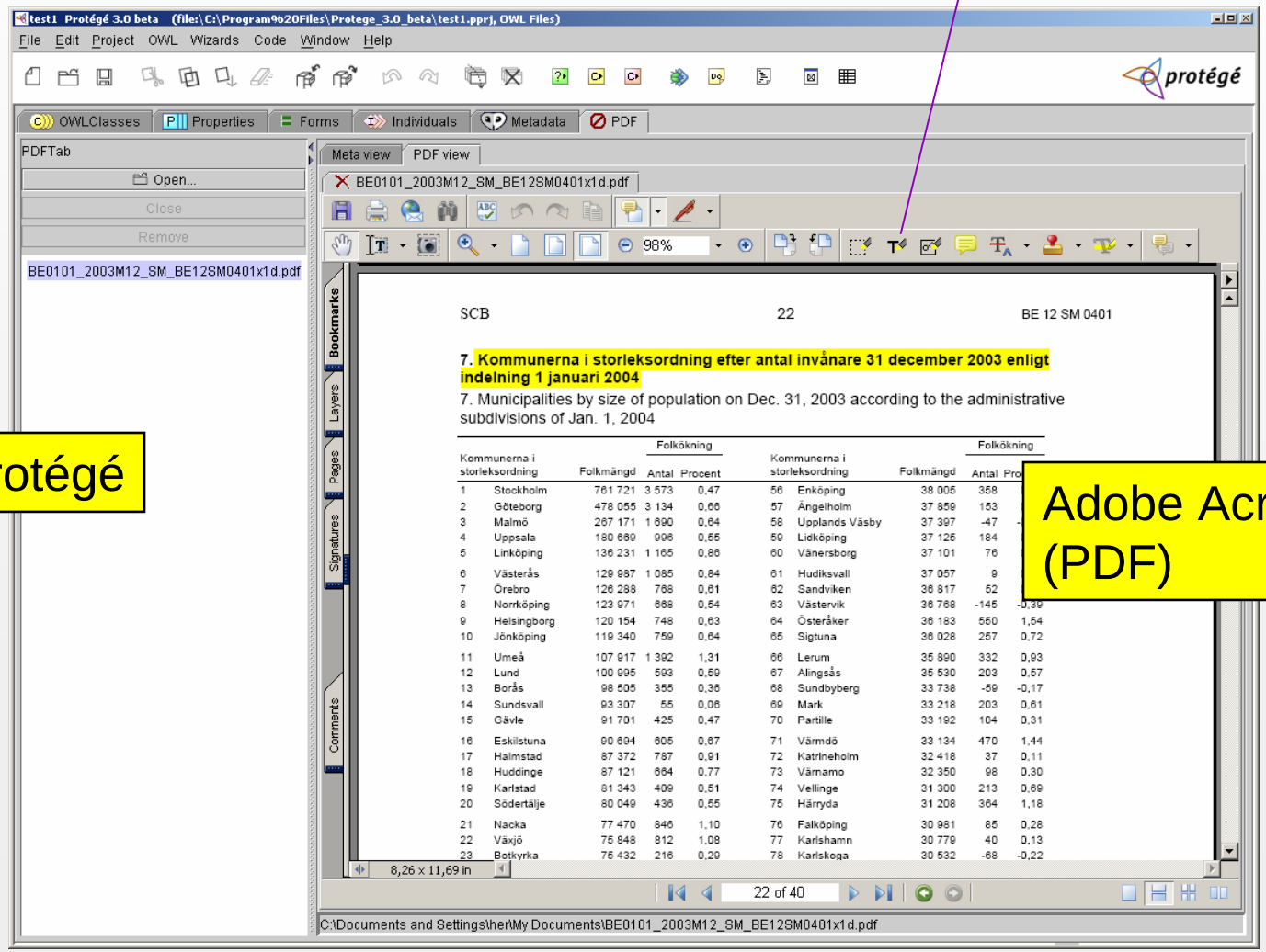


A Protégé Extension for PDF

- **Adobe Acrobat runs inside a Protégé tab**



PDFTab: Annotation Tool for Protégé

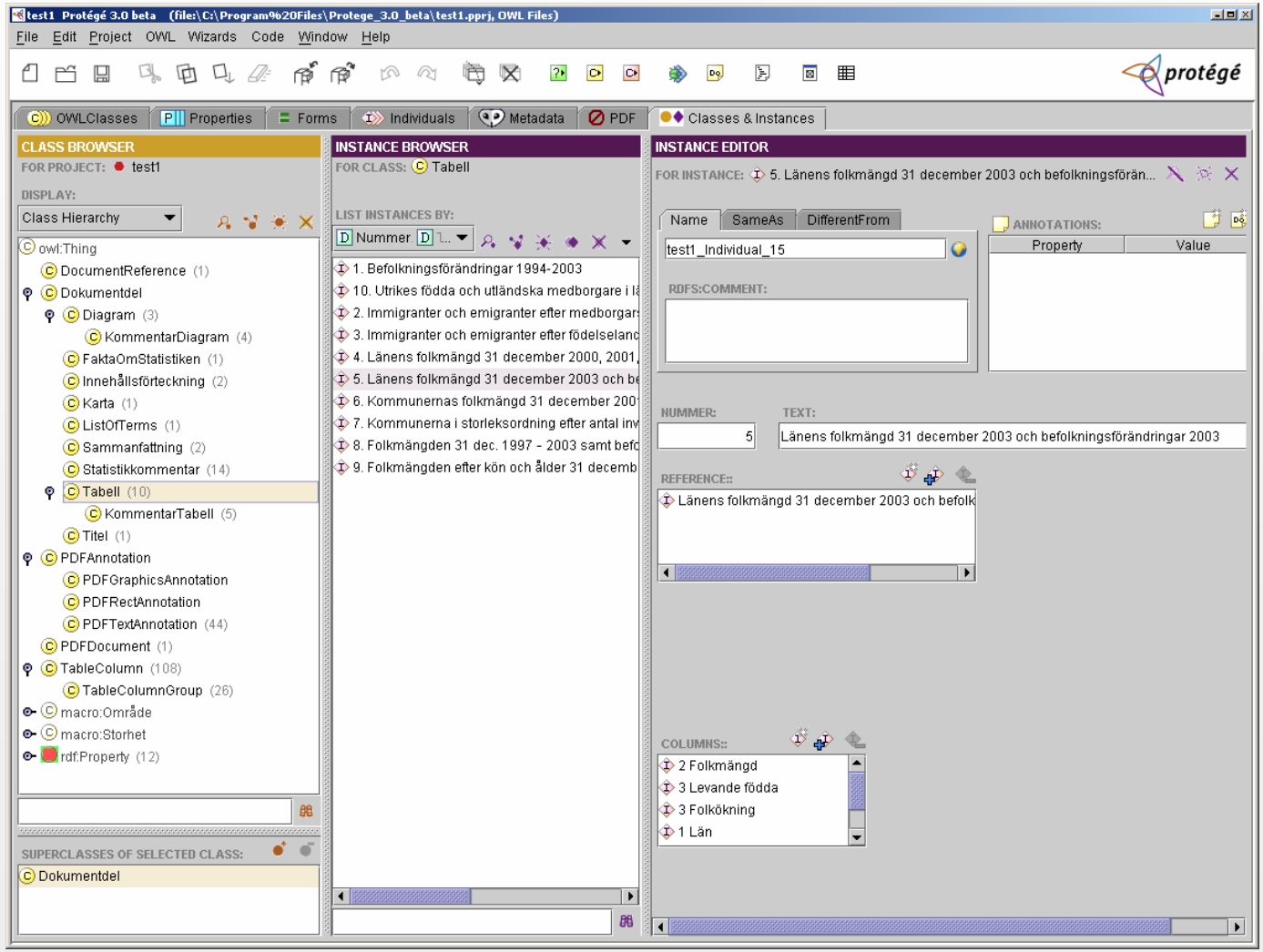


	Kommunerna i storleksordning	Folkmängd	Folkökning Antal	Procent
1	Stockholm	761 721	3 573	0,47
2	Göteborg	478 055	3 134	0,66
3	Malmö	267 171	1 690	0,64
4	Uppsala	180 669	996	0,55
5	Linköping	136 231	1 165	0,86
6	Västerås	129 987	1 085	0,84
7	Örebro	126 288	768	0,61
8	Norrköping	123 971	668	0,54
9	Helsingborg	120 154	748	0,63
10	Jönköping	119 340	759	0,64
11	Umeå	107 917	1 392	1,31
12	Lund	100 995	593	0,59
13	Borås	98 505	355	0,36
14	Sundsvall	93 307	55	0,06
15	Gävle	91 701	425	0,47
16	Eskilstuna	90 694	605	0,67
17	Halmstad	87 372	787	0,91
18	Huddinge	87 121	664	0,77
19	Karlstad	81 343	409	0,51
20	Södertälje	80 049	436	0,55
21	Nacka	77 470	846	1,10
22	Växjö	75 848	812	1,08
23	Botkyrka	75 432	216	0,29

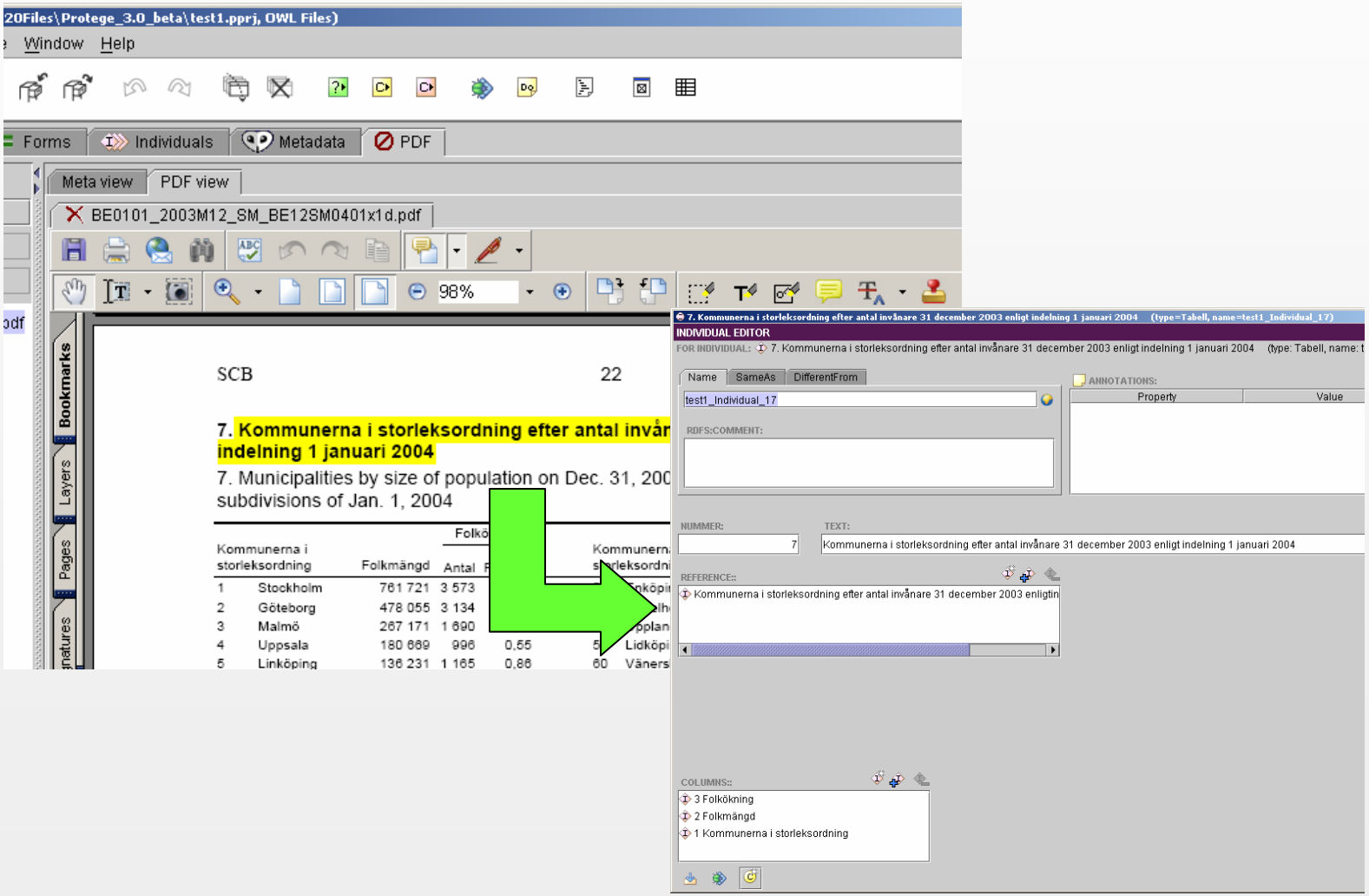
	Kommunerna i storleksordning	Folkmängd	Folkökning Antal	Pro
56	Enköping	38 005	358	
57	Ängelholm	37 859	153	
58	Upplands Väsby	37 397	-47	
59	Lidköping	37 125	184	
60	Vänersborg	37 101	76	
61	Hudiksvall	37 057	9	
62	Sandviken	36 817	52	
63	Västervik	36 768	-145	-0,39
64	Österåker	36 183	560	1,54
65	Sigtuna	36 028	257	0,72
66	Lerum	35 890	332	0,93
67	Alingsås	35 530	203	0,57
68	Sundbyberg	33 738	-59	-0,17
69	Mark	33 218	203	0,61
70	Partille	33 192	104	0,31
71	Värmdö	33 134	470	1,44
72	Katrineholm	32 418	37	0,11
73	Värnamo	32 350	98	0,30
74	Vellinge	31 300	213	0,69
75	Härryda	31 208	364	1,18
76	Falköping	30 981	85	0,28
77	Karlshamn	30 779	40	0,13
78	Karlskoga	30 532	-68	-0,22

Linköpings universitet

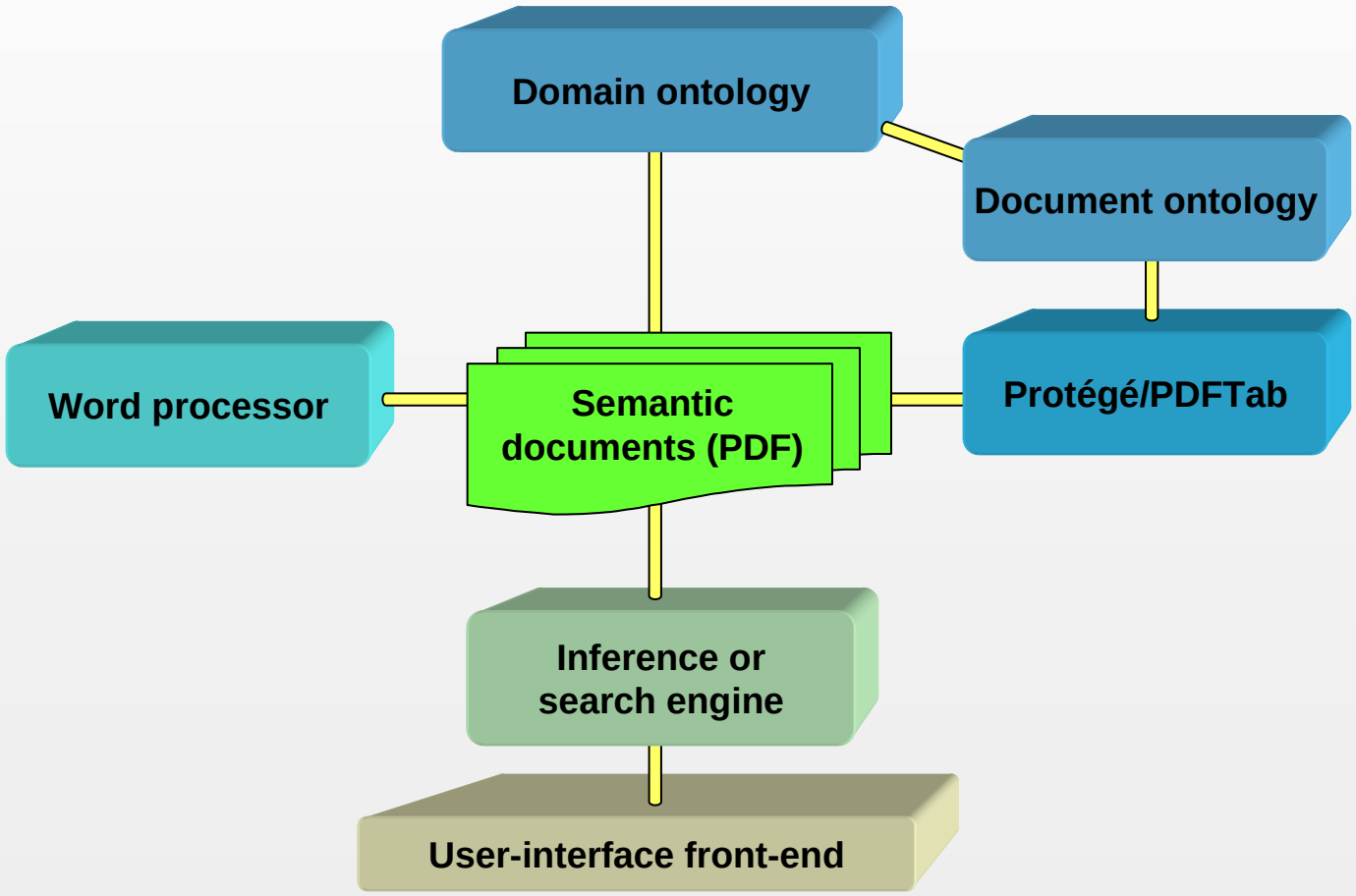
Corresponding Ontology



Mark up of Table Headings

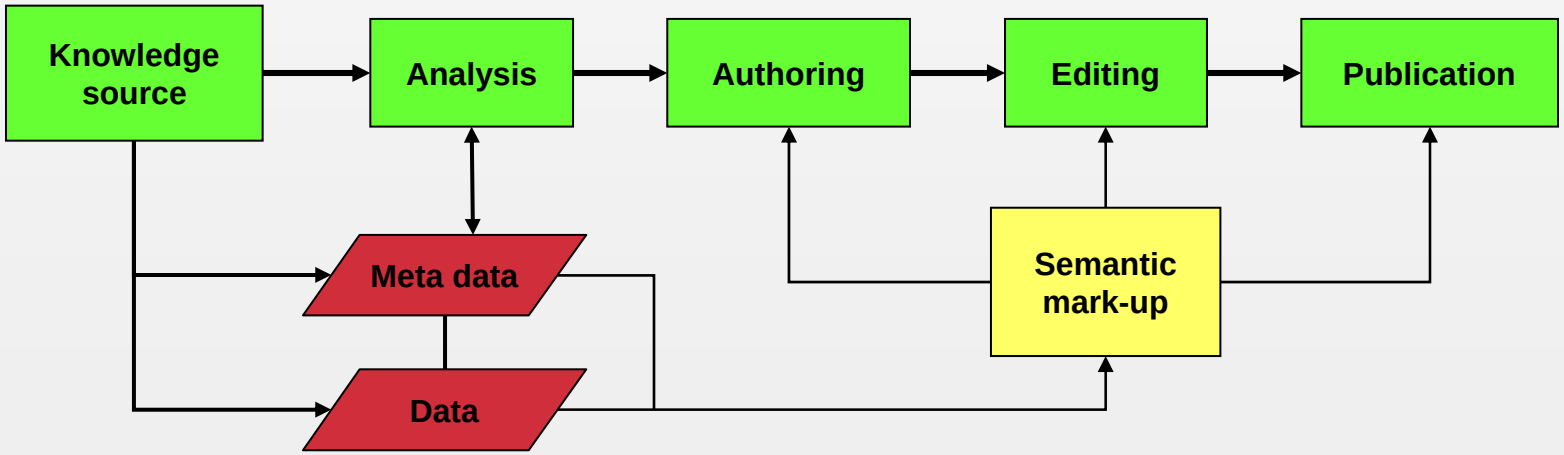


A Semantic Document Architecture for Knowledge Management



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



Linköpings universitet

14

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

Linköpings universitet

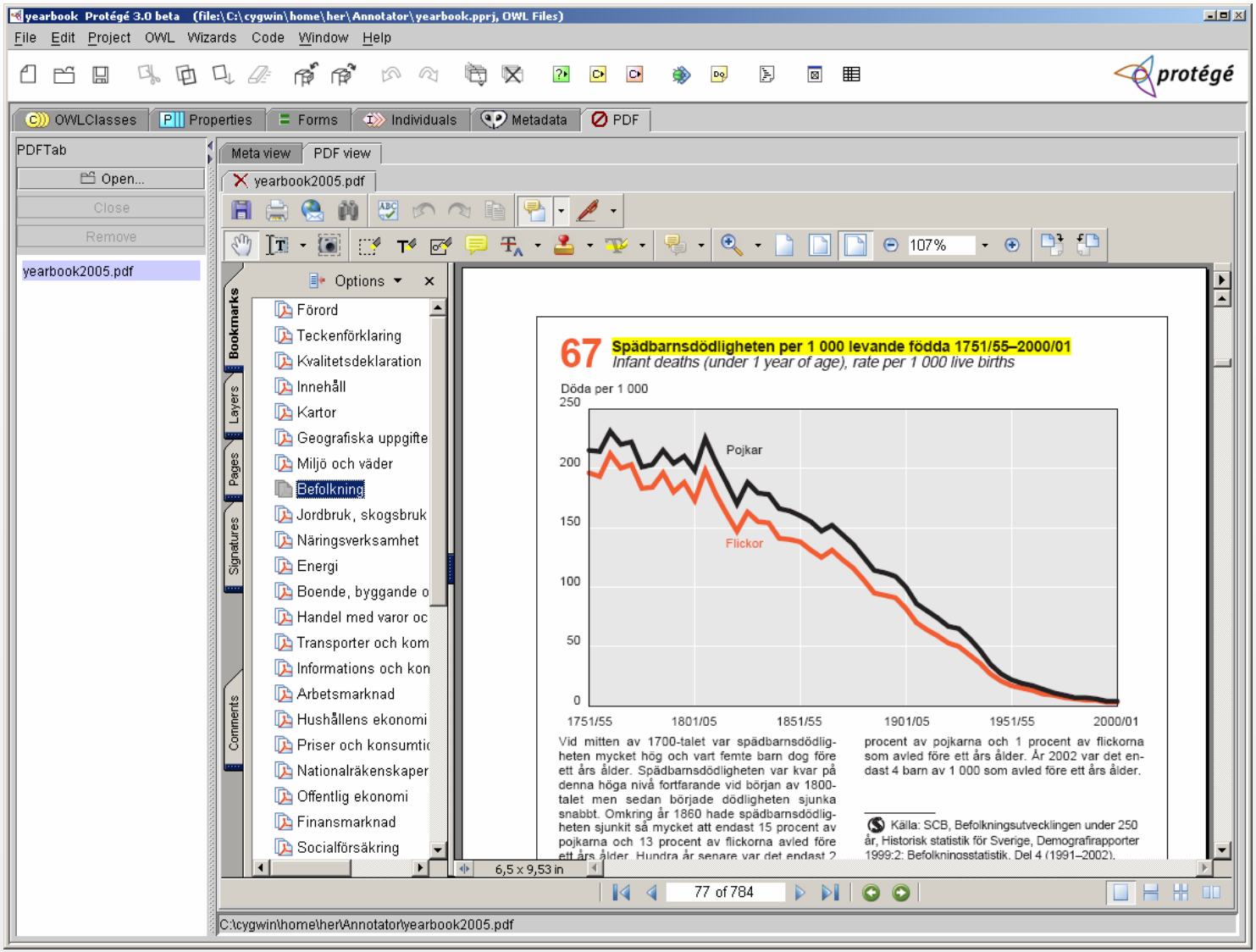
15

Statistics Reports as Semantic Documents

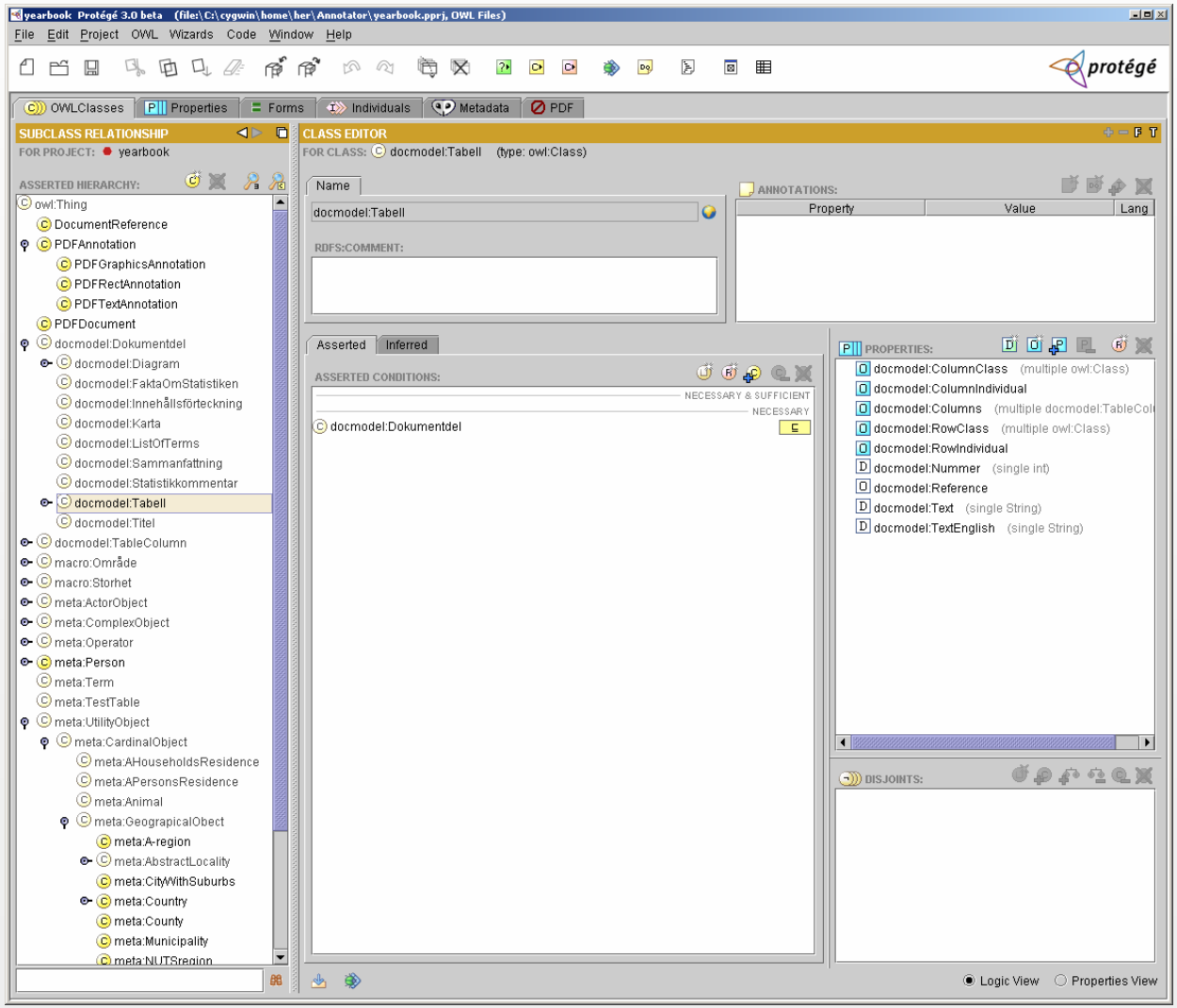
- Statistics Reports
- Statistical Yearbook of Sweden (784 pages)
- Manual and (semi-)automated annotation
- Statistical metadata available
- Development of relevant ontologies
  - Annotation ontology
  - Document ontology
  - Macro data ontology
  - Domain ontology
    - In general, an ontology of the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

Linköpings universitet

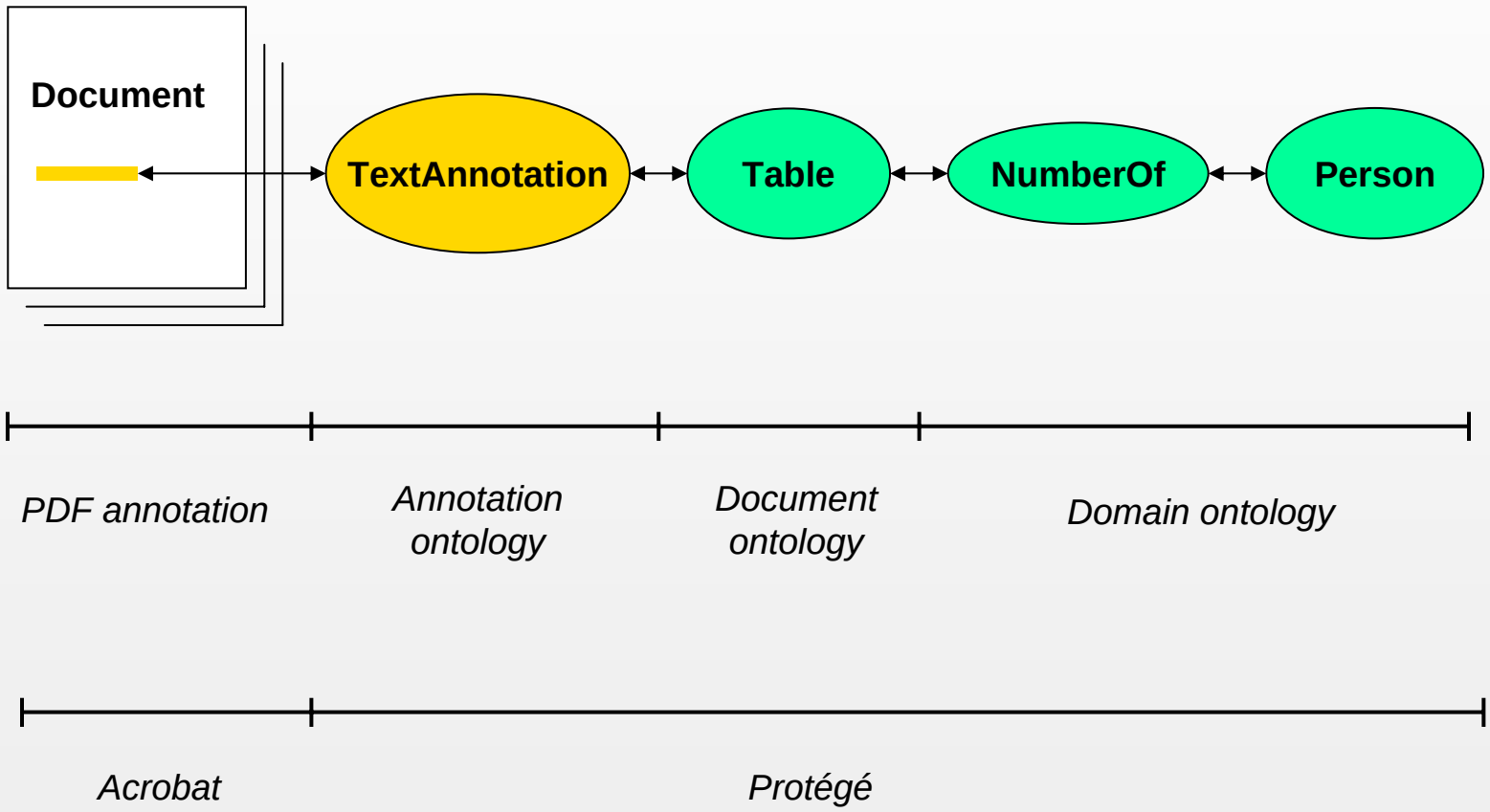
Mark-up of Statistical Yearbook



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?

Linköpings universitet

Clinical Guidelines as Semantic Documents

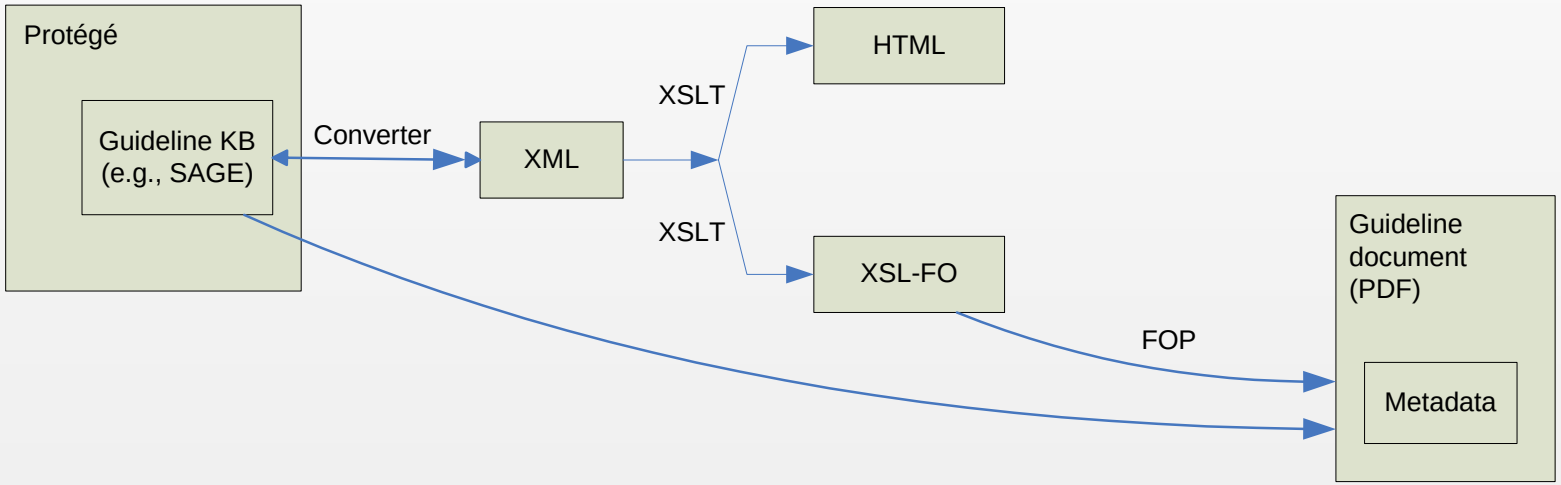
- Experiments with SAGE clinical guideline knowledge bases in collaboration with Samson Tu
- SAGE uses knowledge bases to store authoritative guidelines
- Uses of the knowledge bases
  - Inference
  - Workflow engines
  - Generation of guideline documentation (XML, HTML, and PDF)
- Goal: Semantic document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

Linköpings universitet

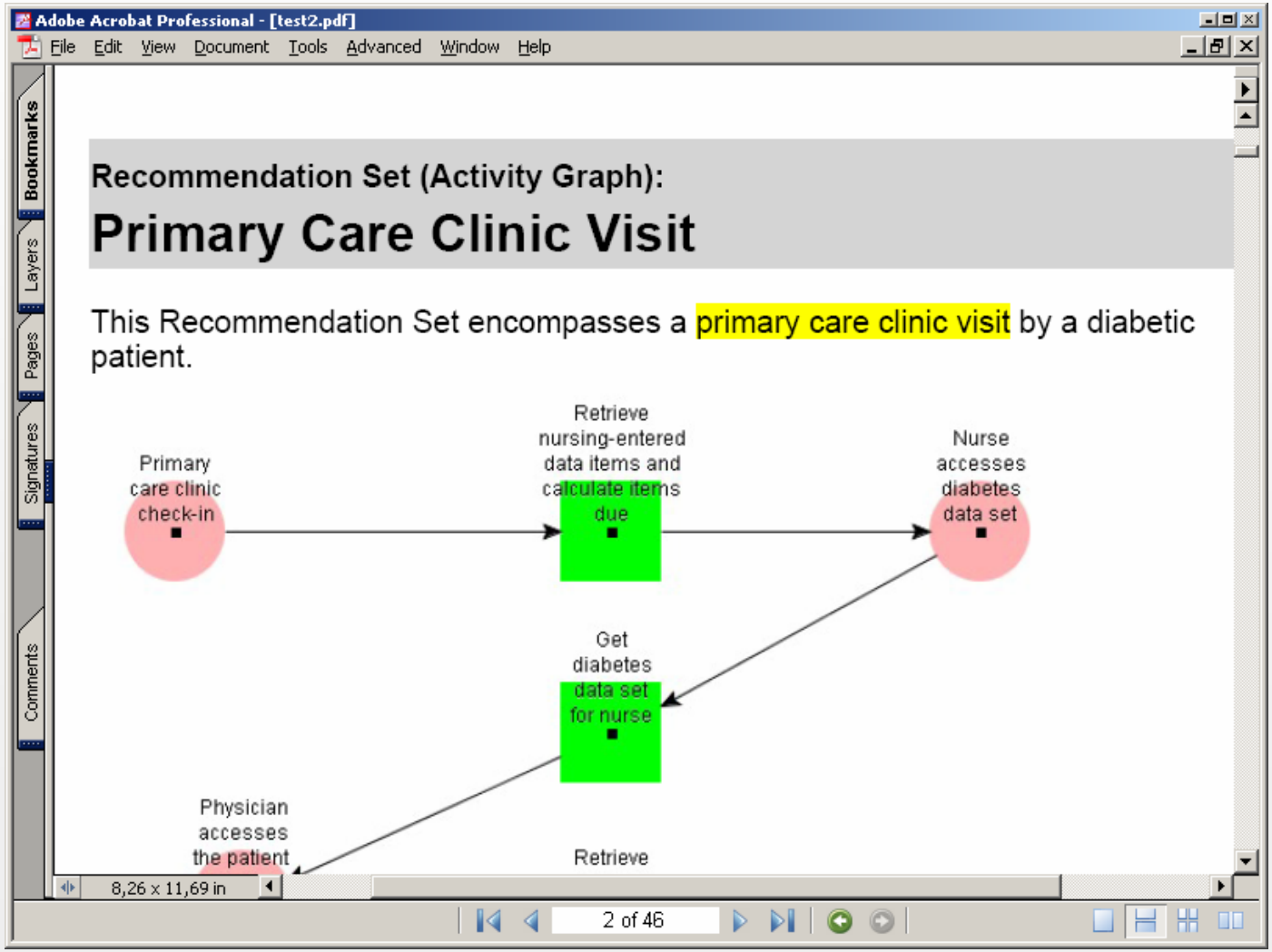
21

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

Linköpings universitet