Some Applications of WDA

different formats

- HTML, MSWord(?)
  - hierarchical structure usually reflects logical structure
  - need to label the hierarchy semantically
  - extraction of prices
  - product info
  - business intelligence
  - web-site usability
- PDF, PostScript, MSWord(?)
  - like "traditional" layout analysis
  - similar applications to "OCR" systems (except that character recognition itself is easy)

The Semantic Web

- HTML, XML, Word have the tags to mark up semantics
- Berners-Lee, others: "the semantic web is coming"
- why is it used so little?
  - formats, standards like XML and XSLT are fairly new
  - too much work for authors?
    - but most "interesting" content exists in databases–already labeled
  - no single ontology/XML standard application formats?
  - inconvenient formats have a history of being used for "DRM" in the physical world–actively trying to keep competitors from mining data?
- is this going to change?

Where is WDA going?

- questions easily answered for PDF, PS
  - lots of documents like this–come out of word processors
  - largely "traditional" document analysis techniques
  - also: a lot of Word documents fall here (no hierarchy)

- HTML: arms race or useful service?

- ontology, knowledge management, AI?
  - techniques quite different from "traditional" document analysis
  - are we getting to adv. document analysis earlier, or is this intrinsically different?