7.7.1 Knowledge Acquisition

At the bottom of figure 7.1 we find tools that use surface analysis techniques to obtain content from documents. These can be either unstructured natural language documents or structured and semistructured documents (such as databases, HTML tables, and spreadsheets).

In the case of unstructured documents, the tools typically use a combination of statistical techniques and shallow natural language technology to extract key concepts from documents.

In the case of more structured documents, one can use database conversion tools as described above. Induction and pattern recognition techniques can be used to extract the content from more weakly structured documents.

7.7.2 Knowledge Storage

The output of the analysis tools is sets of concepts, organized in a shallow concept hierarchy with at best very few cross-taxonomical relationships, which along with RDF and RDF Schema are sufficiently expressive to represent the extracted information. This information also includes instance data.

Besides simply storing the knowledge produced by the extraction tools, the repository must of course provide the ability to retrieve this knowledge, preferably using a structured query language such as SPARQL. Any reasonable RDF Schema repository will also support the RDF model theory, which includes the deduction of class membership based on domain and range definitions and the derivation of the transitive closure of the subClassOf relationship.

Note that the repository will store both the ontology (class hierarchy, property definitions) and the instances of the ontology (specific individuals that belong to classes, pairs of individuals between which a specific property holds).