Towards Building an RDF-based Deep Document Model —— Use Cases ASKG & UniverseTBD

Runsong Jia u7530405

Supervisor: Sergio Rodriguez Mendez, Pouya Ghiasnezhad Omran

Abstract

In the digital era, the volume of textual data, especially in the form of documents, has been growing exponentially. These documents, ranging from academic papers to business reports, encapsulate a wealth of knowledge and insights.

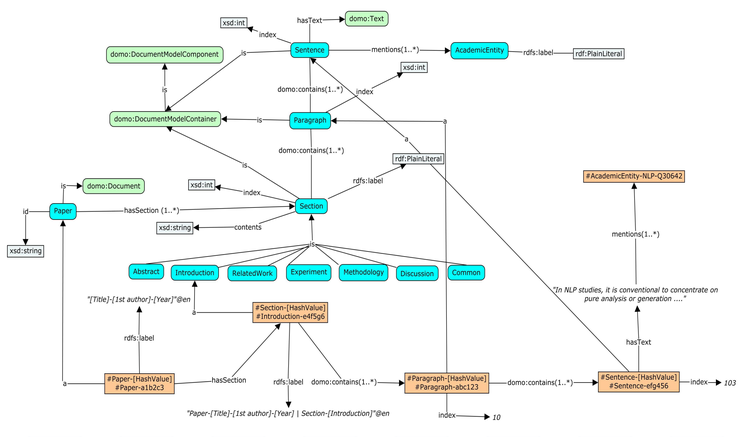
The Deep Document Model aims to recognize and outline the basic structural elements that make up a document. For instance, in a research paper, this could include the title, abstract, introduction, sections, subsections, references and so on. So we can see the document in a clean and easy-to-understand way.

Inspiration

```
<root>
  <child>
    <subchild>.....</subchild>
  </child>
</root>
```

This idea was inspired by the XML DOM, as the example above, it identifies elements, attributes, and textual nodes in a tree-structured XML document. Our result is as the similar structure as it.

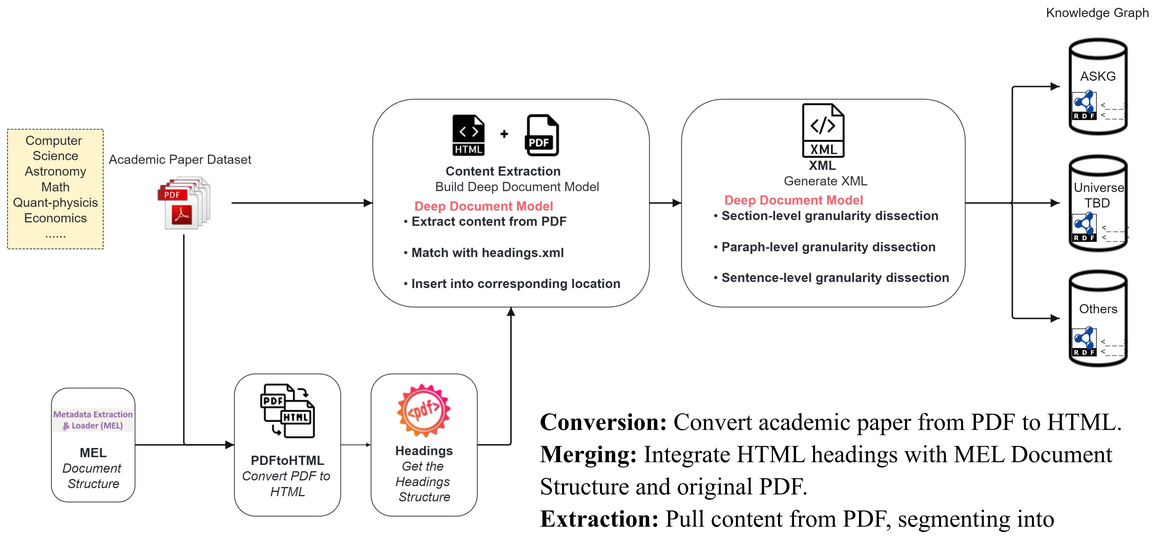
Results



Document Object Model Ontology (DOMO) is an ontology for dissecting and analyzing academic papers, inspired and derived with the HTML Document Object Model (DOM) main concepts.

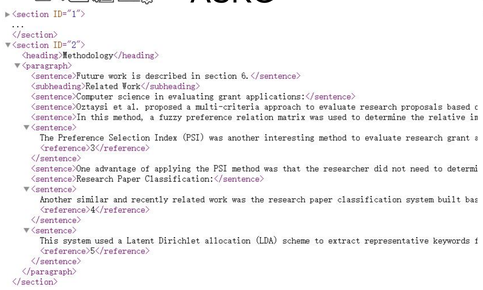
In our project, we update the DOMO ontology and integrate it into our own ontology. The various sub-components from decomposing academic papers will be mapped onto the updated ontology.

Methodology

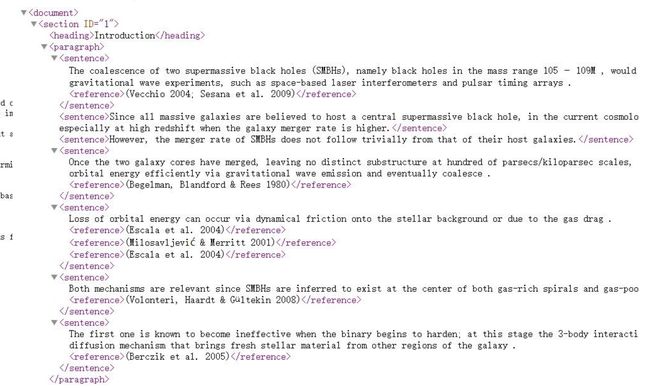


Conversion: Convert academic paper from PDF to HTML.
Merging: Integrate HTML headings with MEL Document Structure and original PDF.
Extraction: Pull content from PDF, segmenting into sentences.
Alignment: Match sentences to XML headings and organize by section.
Granularity: Model functions at levels from sections to sentences.
Transformation: Convert XML to RDF format.
Enrichment: Enhance Scholarly Knowledge Graph for interfacing with language models.

ASKG



UniverseTBD



Future Work

- **Integration:** Combine with MEL & TNNT for Entity Recognition.
- **Enrichment:** Incorporate graphs and charts into DDM.
- **DDM Refinement:** Automate the DDM construction pipeline.
- **Utilization:** Employ RDFLib to build RDF.

Contact

Runsong Jia
u7530405@anu.edu.au
+61 0493556509