

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.

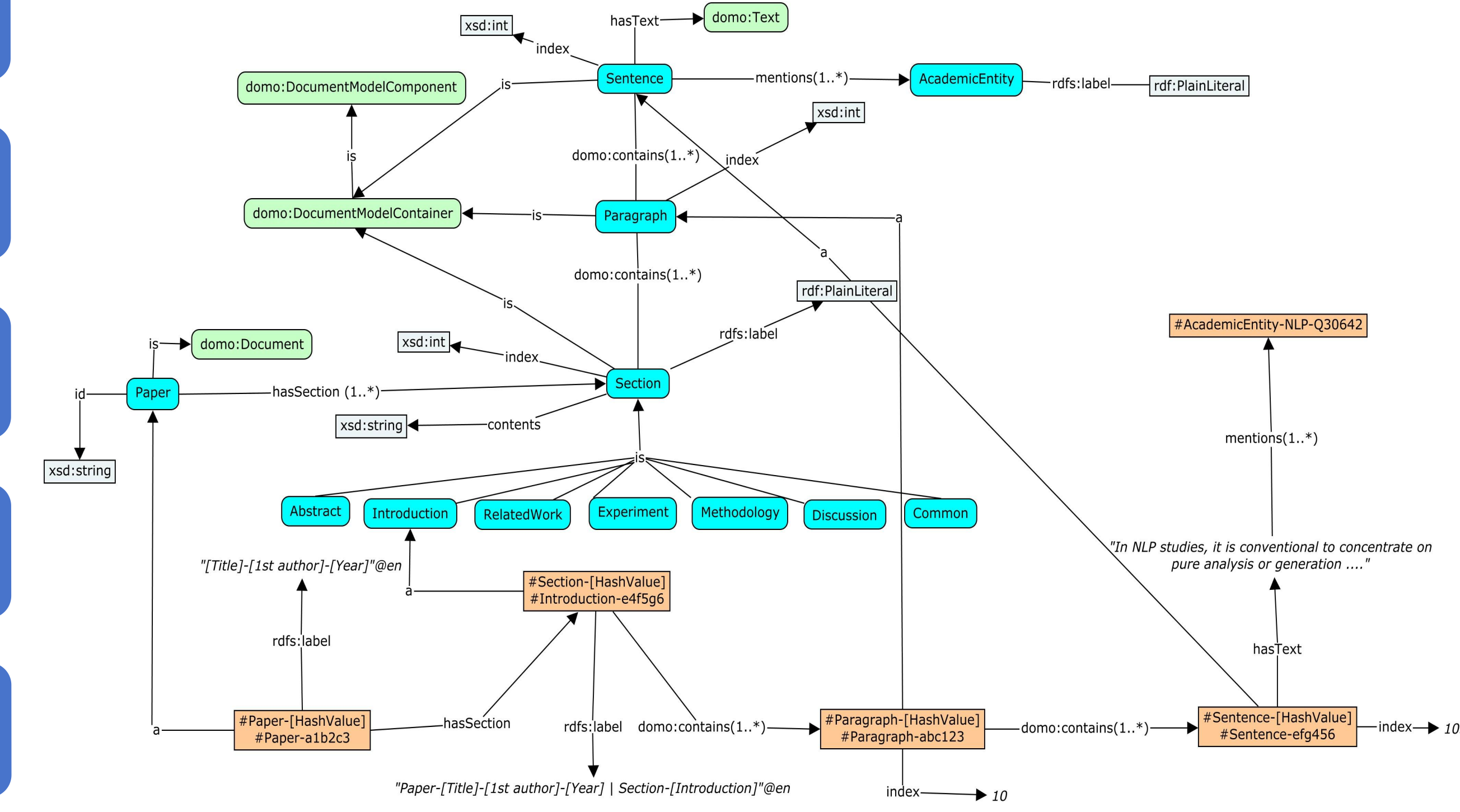
Inspiration



```
<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.

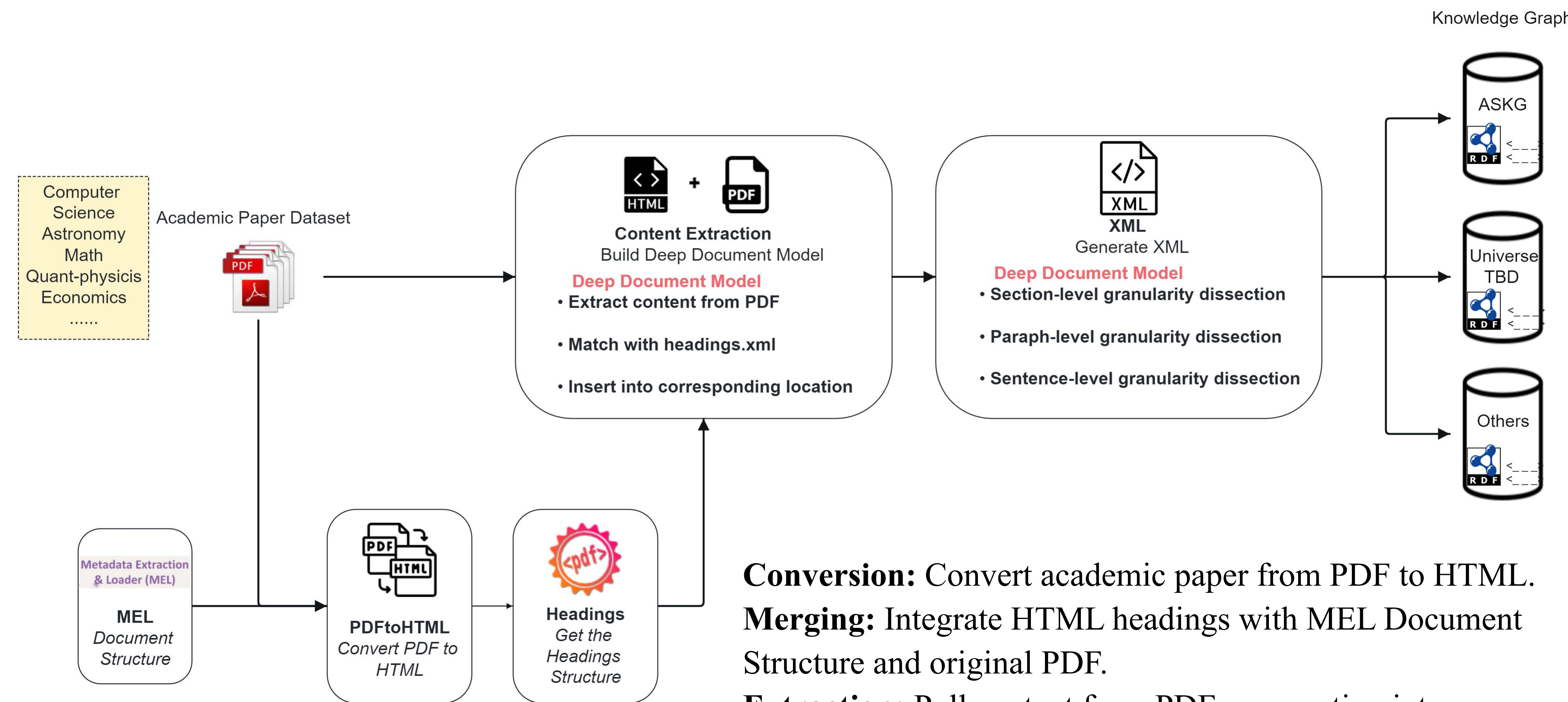
Results



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.

Contact

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

ASKG

```
<?xml version="1.0" encoding="UTF-8" ?>
<document>
  <section ID="1">
    <heading>Introduction</heading>
  </section>
  <section ID="2">
    <heading>Methodology</heading>
    <paragraph>
      <sentence>Future work is described in section 6.</sentence>
      <sentence>Computer science in evaluating grant applications:</sentence>
      <sentence>Oztayisi et al. proposed a multi-criteria approach to evaluate research proposals based on</sentence>
      <sentence>In this method, a fuzzy preference relation matrix was used to determine the relative in</sentence>
      <sentence>The Preference Selection Index (PSI) was another interesting method to evaluate research grant a</sentence>
      <reference>3</reference>
    </paragraph>
    <paragraph>
      <sentence>One advantage of applying the PSI method was that the researcher did not need to determi</sentence>
      <sentence>Research Paper Classification:</sentence>
    </paragraph>
    <paragraph>
      <sentence>Another similar and recently related work was the research paper classification system built bas</sentence>
      <reference>4</reference>
    </paragraph>
    <paragraph>
      <sentence>This system used a Latent Dirichlet allocation (LDA) scheme to extract representative keywords f</sentence>
      <reference>5</reference>
    </paragraph>
  </section>
</document>
```

UniverseTBD

```
<?xml version="1.0" encoding="UTF-8" ?>
<document>
  <section ID="1">
    <heading>Introduction</heading>
    <paragraph>
      <sentence>
        The coalescence of two supermassive black holes (SMBHs), namely black holes in the mass range 105 – 109 M⊙, would
        gravitational wave experiments, such as space-based laser interferometers and pulsar timing arrays .
        <reference>(Vecchio 2004; Sesana et al. 2009)</reference>
      </sentence>
      <sentence>Since all massive galaxies are believed to host a central supermassive black hole, in the current cosmolo
        especially at high redshift when the galaxy merger rate is higher.</sentence>
      <sentence>However, the merger rate of SMBHs does not follow trivially from that of their host galaxies.</sentence>
    </paragraph>
    <paragraph>
      <sentence>
        Once the two galaxy cores have merged, leaving no distinct substructure at hundred of parsecs/kiloparsec scales,
        orbital energy efficiently via gravitational wave emission and eventually coalesce .
        <reference>(Begelman, Blandford & Rees 1980)</reference>
      </sentence>
      <sentence>
        Loss of orbital energy can occur via dynamical friction onto the stellar background or due to the gas drag .
        <reference>(Escala et al. 2004)</reference>
        <reference>(Milosavljević & Merritt 2001)</reference>
        <reference>(Escala et al. 2004)</reference>
      </sentence>
      <sentence>
        Both mechanisms are relevant since SMBHs are inferred to exist at the center of both gas-rich spirals and gas-poo
        <reference>(Volonteri, Haardt & Gultekin 2008)</reference>
      </sentence>
      <sentence>
        The first one is known to become ineffective when the binary begins to harden; at this stage the 3-body interacti
        diffusion mechanism that brings fresh stellar material from other regions of the galaxy .
        <reference>(Berczik et al. 2005)</reference>
      </sentence>
    </paragraph>
  </section>
</document>
```

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.