

start

FullTextXML

BeautifulSoup

FullText

**spaCy**

tokenize  
tagger  
parser  
NER

Doc

SourceClass

Doc  
PMC ID  
FullText  
TextDict

Section Names

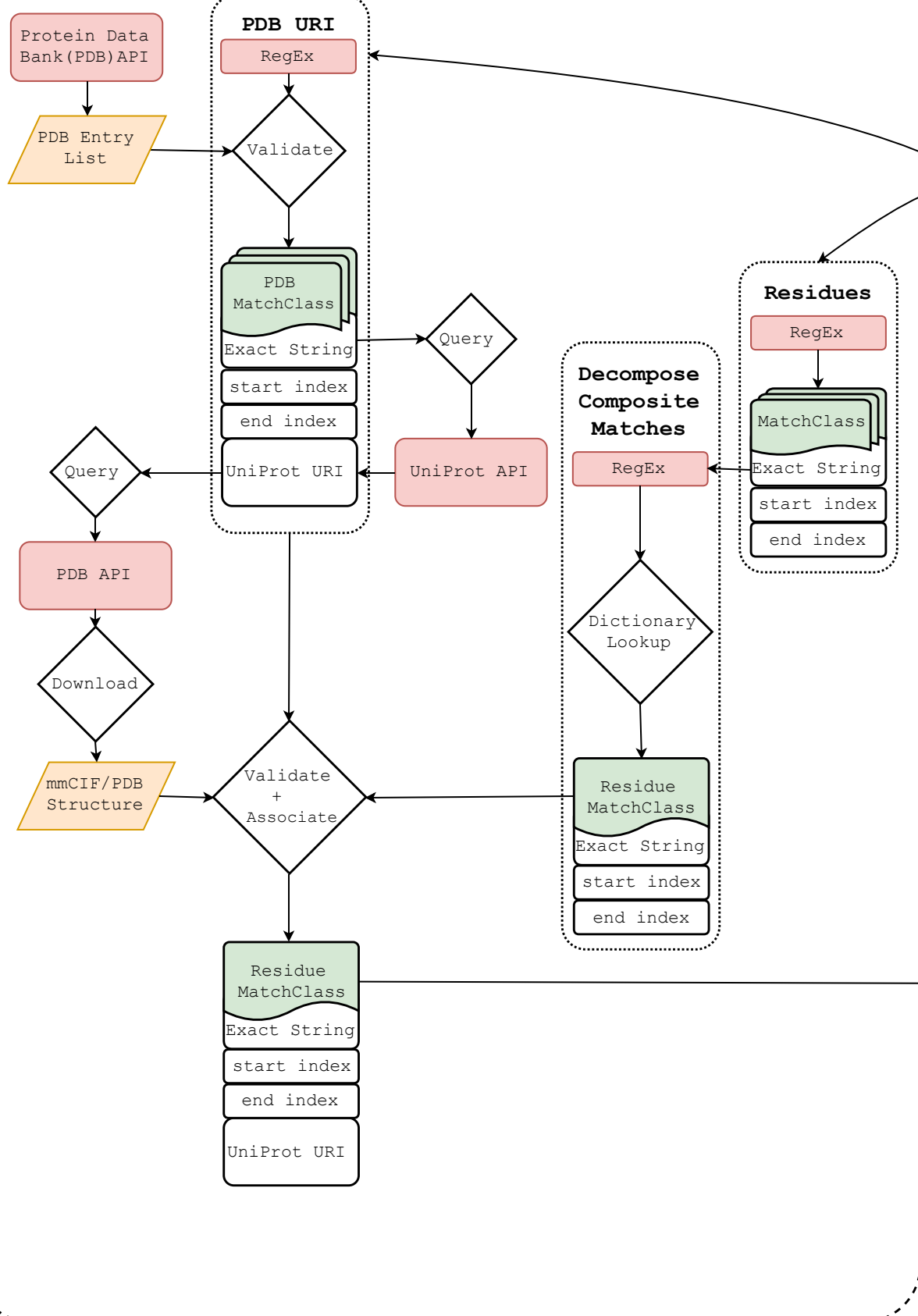
Section Whitelist

FuzzyWuzzy

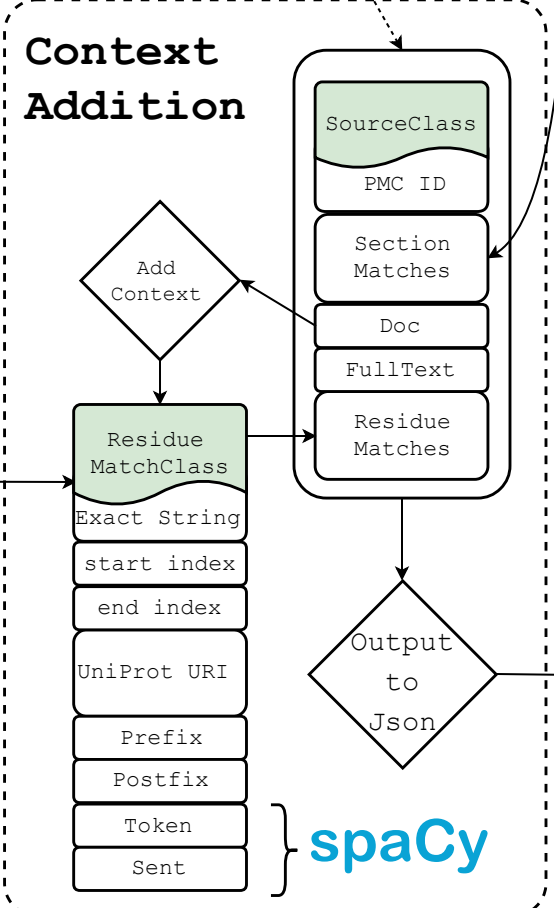
MatchClass

Exact String  
start index  
start Token  
end index  
end Token

## Identification and Location



## Context Addition



## Data Product

Annotations

- +PMID/PMCID
- +Annotations
  - prefix
  - exact (residue)
  - postfix
  - section
  - URI (protein)

Europe PubMed Central