Proposed Method of Information Retrieval and Display from the US Federal Register

Matthew J Wiecek
Texas A&M University
College Station, TX
matthewwiecek@tamu.edu

Chun-Chan (Bill) Cheng
Texas A&M University
College Station, TX
aznchat@tamu.edu

Divyesh M Tekale
Texas A&M University
College Station, TX
tekale2@tamu.edu

1. AUTHORS NOTE

The following is to be considered a continuation of Report 1.

2. LOW LEVEL OVERVIEW OF SUBSYSTEMS

2.1 Indexer

The indexer will be written in *Java* and will interface with Apache Solr via the *Solrj* library.

2.1.1 Raw Data & Crawler Intergration

It is clear that before data may be indexed, it must first have been crawled. The web crawler dumps each crawled document into its own text file. The indexer will then read the text files from the directory in which the crawler has put the text files; these are stored as strings in a *BlockingQueue*. The indexer threads will then be able to parse the XML and extract the relevant fields. The reason for such separation is that it allows indexing and crawling to be done independently. This was especially helpful during development of the minimizes web traffic to the Federal Register. If the documents have to be re-indexed due to a change in the schema, we do not need to burden the Federal Register with a request for already crawled documents. This helps to maximize web crawler politeness.

2.1.2 Indexed Fields & SolrJ intergration.

Solr takes in various fields for indexing. There are four well defined types of documents published in the Federal Register: Notices, Proposed Rules, Final Rules, and Presidential Documents. The indexing threads pull the raw strings from the *BlockingQueue* and parse the string as an XML document. First, the type of document is identified. Once the document has been identified, the appropriate XML fields are extracted as defined by the schema in section 3. The fields are then added to a *SolrInputDocument* and are then passed to the Solr server by the *HTTPSolrServer* class provided by the *Solr* Java API.

2.2 XML Parser

All the pages we crawled from the federal register was in xml format, so we had to read through the xml data with org.w3c.dom.NodeList library. We parsed the key words `TYPE`, `AGENCY`, `AGENCY TYPE`, `SUBAGY`, `SUBJECT`, and for `PRESDOC` we parsed `HD`, `FP`.

3. INDEXING SCHEMA

As there are four distinct types of documents published in the Federal Register, it is necessary that a different schema is in place for indexing each type of document.

3.1 Notice

The notice is perhaps the most common document published in the Federal Register. When indexed, documents of this type will contain the following fields:

Type of Document shall always be "Notice"

Cabinet Department the federal cabinet department which issued the notice

Agency the agency within the department which issues the notice

Action description of the type of publication

Summary a brief summary of the notice

Date the date of publication in the Federal Register, if available

3.2 Proposed Rule

A proposed rule is issued as a guidance document to the public. It describes the current draft of a rule an executive agency is considering imposing. The publication of the draft rule gives the public a chance to provide feedback about the rule as well as giving the public time to prepare for the effects of the potential rule.

Type of Document shall always be "Proposed Rule"

Cabinet Department the federal cabinet department which issued the proposed rule

Agency the agency within the department which is issuing the proposed rule

Action description of the type of publication

ACM ISBN 978-1-4503-2138-9.
DOI: 10.1145/1235

Summary a brief summary of the notice

Date the date of publication in the Federal Register, if available

3.3 Rule

A rule is issued as notice that a proposed rule has been finalized and is now in effect. It serves as both a means to communicate to the public about the new rule, as well as being the legal source of the rule, which may be referenced in court.

Type of Document shall always be “Rule”

Cabinet Department the federal cabinet department which issued the rule

Agency the agency within the department which is issuing the rule

Action description of the type of publication

Summary a brief summary of the notice

Date the date of publication in the Federal Register, if available

3.4 Presidential Document

A presidential document is one which has been issued by the President of the United States. They may be executive orders, notices, or other noteworthy documents the President has decided to publish. These do not conform to the same structure as other documents published in the Federal Register, and thus, are harder to index.

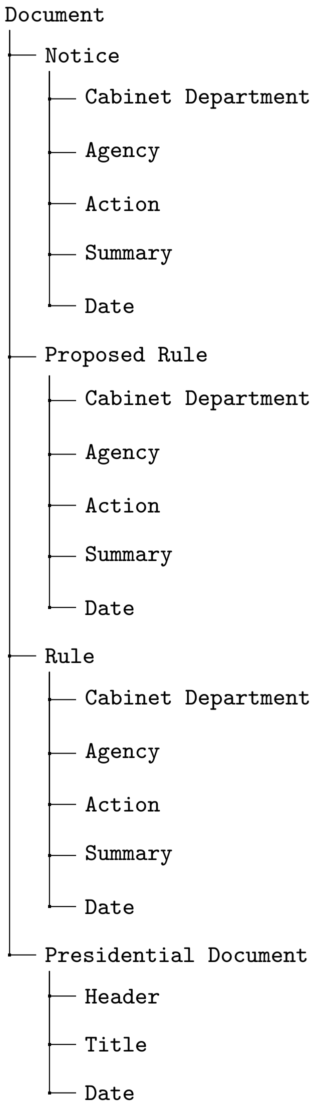
Type of Document shall always be “Presidential Document”

Header the header of the document

Title from which Title of the US Code does this document derive it’s authority (if applicable)

Date the date of publication in the Federal Register, if available

3.5 Tree Stucture Overview


Document
Notice
Cabinet Department
Agency
Action
Summary
Date
Proposed Rule
Cabinet Department
Agency
Action
Summary
Date
Rule
Cabinet Department
Agency
Action
Summary
Date
Presidential Document
Header
Title
Date