Overview

The RFC Editor intends to accept submission in plain text and XML, but will be working primarily in xml using the xml2rfc v3 grammar. The most common submission format supplied now is Internet-Draft formatted text. Tools already exist to convert other formats to text.

The goal of this project is to simplify the creation of an initial xml2rfc v3 version of a document submitted in another non-xml format. A perfect automated conversion is not expected. Rather, the application should produce a document that is well-formed, and sufficiently close to correct that an editor can complete the conversion with minimal effort.

Deliverables/Tasks

This project will create an application to convert an Internet-Draft formatted text file to an xml2rfc v3 document. The development effort will include

- Designing the command line interface
- Demonstrating the conversion of a specified set of text documents
- Providing an extensible test suite for the application
- Documentation, and training for the RFC Production Center staff

Detailed Description and Requirements

The application must run as a command-line program under Linux and os/x shells. Running at a Windows command prompt would be nice to have but is not required. The application must be easily adaptable to become part of a web-service offering the translation.

The developer will work with the Program Manager to agree on an initial command-line interface before beginning development work.

During development, the following documents will be used to refine the application. These documents were chosen to highlight difficult cases in determining the best conversion from text to xml (such as recognizing figures and tables, and correctly identifying references).

- http://trac.tools.ietf.org/tools/xml2rfc/trac/browser/trunk/cli/tests/valid/draft-miek-test.txt
- http://tools.ietf.org/html/draft-ietf-mip4-multiple-tunnel-support-07 (particularly section 4.2)
- http://datatracker.ietf.org/doc/rfc7118/ (particularly section 8.2)

The conversion must preserve the content of the input text file. In certain cases, some input text may be removed including

- Headers, footers, and other content that would be generated algorithmically when otherwise expressed semantically in the xml format
- Boilerplate that matches what is currently defined in RFC5741 (and would be accepted by id-nits). The appropriate attributes to the <rfc> element will replace such boilerplate. A section that appears where boilerplate is expected, but does not match currently acceptable boilerplate, must be preserved and converted the same as an ordinary section would be.

In general, the conversion is not expected to preserve whitespace. Elements identified as artwork, however, must preserve whitespace. When it is ambiguous whether an input block is artwork, a table, or a list, the application should favor converting the block to artwork. Artwork should be captured in a CDATA construct rather than using xml-escaping to simplify the work of correcting a mis-classification or adjusting the boundaries of the artwork.

To the extent possible, content that is explicitly represented in the xml2rfc v3 grammar should be converted to use that representation. For example, the application is expected to

- Identify references and produce populated <reference> elements and appropriate <xref> elements
- Parse any author information section and produce populated <author> elements
- Identify special sections (such as the abstract) and represent them with the appropriate element