Internet Engineering Task Force
INTERNET-DRAFT
schulzrinne-rfc-00.ps

Schulzrinne
Columbia U.

September 3, 2000
Expires: January 2001

Suggestions for Future Document Format Requirements for RFCs and Internet Drafts

Status of this Memo

Copyright Notice

Abstract

A dwindling fraction of authors of I-Ds and RFCs are still fluent in nroff, the official input format for RFCs. With the use of other tools, the consistency and layout quality of RFCs and I-Ds has suffered. Printing, hyperlinking and viewing of ASCII documents are unsatisfactory. In this document, we investigate how the I-D and RFC series can be upgraded in usability while maintaining stability, longevity and cross-platform abilities that have been hallmarks of the RFC series.

1 Introduction

We distinguish three different formats, with rather different requirements, namely the input document, the archive document and the presentation document. Input documents are submitted to the I-D or RFC editor, archive documents are reference documents available to the community, and presentation documents are alternate representations suitable for viewing or printing. Each document can be represented in multiple formats. Currently, nroff and ASCII are RFC input document formats, ASCII is the archive format and PostScript is the optional presentation format. Only about 43 out of 2894 RFCs are available in PostScript.

In general, it is desirable that all formats have the following properties. However, if multiple formats are supported for a single document type, one or more of these requirements can be relaxed.

Long-lived: The same format should be readable for at least several decades, using software available in source form on a variety of operating system platforms.

Safe: With reasonable effort, users must be able to ensure themselves that the document does not have harmful side effects, such as macro viruses and trojan horses. Note that no format other than plain text without control characters is likely to be completely safe, as rendering engines are subject to bugs. Some document formats are safely implementable, in that a safe implementation could be written,

but it is likely that existing implementations have security holes. Inherently unsafe documents have implementations that cannot be made safe. Any document with web links such as images that are automatically rendered can violate user privacy by revealing the use of the document.

Cross-platform: Creation tools SHOULD be and viewing tools MUST be available across multiple operating systems.

Searchable: Search engines and standard OS search tools (such as grep on Unix systems) should be able to search document repositories for key words.

Expressive: The format SHOULD allow content-based marking of text, such as abstracts, citations, section headings, table captions and tables, author contact information, reserved words, BNF entities and other semantically significant entities. This allows indexing and the creation of hypertext.

Extractable: Text and figures from the document should be extractable via cut-and-paste or a text editor.

Consistent: Input formats should support mechanisms for automatic cross-references to sections, tables, figures and bibliographic entries. To avoid errors and create a consistent reference style, it is desirable that authors can use a standard, publically-available database for citations of RFCs and Internet drafts. Standard references to other networking-related papers are also desirable. Such databases are currently available for BibTeX (LaTeX) and, for RFCs only, the XML RFC DTD.

Tables and figures: Architectural relationships and other explanatory material can benefit from graphical representations beyond ASCII art. Bit-level protocol definitions are reasonably well served by the standard 32-bit ASCII notation. Tables, including paragraph entries, are desirable.

Universal preparation and access: It is desirable that document submission formats and presentation formats have freely available software on all major platforms so that the creation and viewing of RFCs does not depend on the purchase of software. Open-source software is desireable, but probably cannot be an absolute requirement.

For I-Ds and possibly RFCs that update or obsolete other RFCs, change bars are helpful. For authoring, systems that allow annotation and shared editing may also be advantageous.

Generally, standards-based document formats are more likely to be cross-platform and long-lived. The following classification of document formats has been proposed [1]. Note that the terminology of "proposed" does not correspond to the IETF use of the term.

Adopted: A file format that has been studied and adopted by a competent official standards body or professional institute (e.g., ISO, BSI, IEEE, W3C). Examples include ASCII or UTF-8 plain text, HTML, GIF, JPEG, MPEG, or XML.

Proposed: A file format that has been proposed by a company, organization or group for the purpose of data interchange. The specification has been published and is freely available for other companies to build into software. Examples include Postscript and Quicktime.

Proprietary: A file format that is proposed by a company or organization for the purpose of data interchange, not published, but supported by software that is available operating systems. Examples include Adobe Acrobat file, Zip or Stuffit archive files.

Non-standard: A file format that is proposed by a company for the purpose of data interchange, not published, supported by software on only some computers and operating systems. Examples include Microsoft Word and PowerPoint, Adobe PhotoShop files.

2 Properties of Document Formats

Below, we summarize the properties of some candidate document formats. Here, "text" refers to plain text, represented currently as ASCII.

	Text	nroff	PS	PDF	Word	RTF	XML	HTML
Long-lived	yes	maybe	maybe(3)	maybe	no	no	probably	maybe
Safe	yes	yes	yes(2)	yes	no	yes	yes(7)	yes(7)
Cross-platform	yes	Unix	yes	yes	no	no	yes	mostly
Searchable	yes	yes	no	no(1)	no(1)	no	yes	yes
Expressive	no	some	no	no	no	no	yes	maybe(4)
Extractable	yes	yes	no	maybe	yes	yes	yes	yes
Consistent	no	no(5)	N/A	N/A	yes(5)	yes(5)	yes(6)	no
Free viewers	yes	yes	yes	yes	yes	yes	no	yes
Free creator	yes	yes	yes	?	yes	yes	yes	yes

(1): There are special tools for searching grouups of PDF and Word files, but they are not widely used. PDF viewers and Word editors support searching within a file.

(2): PostScript is safe as long as the viewer disables commands that allow reading and writing local files.

(3): These formats are subject to version changes and vendor-specific implementations and extensions.

(4): HTML can be made expressive by adding pseudo-tags that are ignored by browsers, but this is not standardized and thus not recommended.

(5): There do not seem to be bibliographic databases for Word and refer exists for nroff.

(6): While XML would allow consistent references, it is not clear whether existing tools support automatic cross-references, for example.

(7): While HTML and presumably XML can contain scripts, they are readily removed by simple filters or disabled in browsers.

3 Use of Document Formats

3.1 Input to RFC and I-D Editor

Word and RTF are not suitable as input formats. Among other reasons, Word is not safe. (RTF does not have macro facilities and thus is safe.) Word and RTF have multiple, incompatible versions, do not have fully functioning display and print tools on many non-Windows platforms and enforcing consistent formating or mark-up is difficult. Current translation tools from Word and RTF to HTML and ASCII are less than reliable as soon as tables and other non-paragraph formatting is used.

nroff works reasonably well, as long as authors stick to standard macros. However, it is available only on Unix systems (Is this true?) It has problems with tables and tables of contents. The latter require manual intervention. nroff mark-up is only roughly content-based, i.e., it is sometimes, but not always, possible to detect semantic elements such as headers, figure captions or author addresses automatically. The

major problem with nroff is that a dwindling fraction of RFC and I-D authors are familiar with the tool. Reasonable, but not perfect, translation mechanisms from LaTeX are available. Thus, it is desirable that nroff remain an option as an input format.

ASCII as an input format has the disadvantage that it encourages inconsistent formatting. Any changes requested by the IESG that add or remove a line are likely to cause page formatting problems, with extensive hand-editing required.

It may be advantageous for the RFC editor to capture whatever format was used for document preparation, as the author cannot be relied on to be able to produce the source file after a lengthy interval. Even Word, for example, is likely to be translatable to other formats, if with difficulty. Collecting files used for document creation, with the explicit disclaimer that they are insufficient and unallowable as input format, seems to occur little cost beyond storage.

3.2 Archival

Only plain text is likely to be readable and printable across a wide variety of computing platforms for decades to come. Thus, it has to remain as a primary archival format.

XML is suitable as an archival format as long as the DTD is stable over long periods of time, publically document and all documents are validated. This validation can be performed automatically before the document even reaches the I-D or RFC editor. XML must be supplemented by a presentation format that is viewable and printable without specialized tools.

3.3 Presentation

As a presentation format, plain text is necessary but has a number of disadvantages. Courier-font (fixed-width) text is hard to read; it is space-inefficient, with only 72 characters per line while a variable-width document has about 110-120 characters per line. Since there are no font and character size distinctions, are hard to present.

Alternatives to plain text include Postscript and PDF. PostScript can be rendered on every operating system platform, using (a small set) of freely available tools. Experience has shown that older PostScript does not render properly on modern devices. Also, the fraction of PostScript-capable printers is probably shrinking. PostScript generated by standard printer drivers is often not displayable or printable on printers other than the printer the PostScript version is designed for. Some PostScript versions (notably those produced by FrameMaker) also include PostScript instructions that prevent printing on printers using a paper format other than the "native" one (A4 vs. letter-size). Care must be taken that the PostScript document does not reference fonts other than the built-in Times Roman and Symbol fonts. LaTeX (dvips) and nroff (ditroff) tools generally produce robust and device-independent PostScript. LaTeX, however, needs to be instructed to use native PostScript fonts as the Computer Modern fonts render poorly unless the dvips resolution and the screen or printer resolution agree.

PDF is probably the most widely available cross-platform display document format. It generates documents that are mostly smaller than PostScript and that print on both letter-size and A4 paper. While PDF is mostly generated by converting PostScript using programs such as distill and ps2pdf, there are also a number of tools that generate PDF directly. Such tools include printer drivers and pdfTeX (and its LaTeX version). PDF has the disadvantage that there are no printers that can accept PDF directly. Currently, version 3 of PDF is the most readily usable, with version 4 sometimes used by output filters. Some versions of distill apparently create files with missing fonts.

(Indeed, many broken PostScript files fail the translation, so that PDF translation is useful even to ensure that the PostScript file is likely to be well-formed.)

3.4 XML

Tools that convert this DTD into plain text or HTML are available [2].

The author believes that the current DTD [2] lacks a number of structuring features, including tables. For graphics, SVG may be a suitable format, as there are a number of tools that can produce this format already. As an W3C recommendation, it is also likely to be stable. The main problem is the lack of an automated generation of ASCII. Given the different size requirements and constraints, it is unlikely that this conversion can be automated. Thus, it may be appropriate to allow an element within the figure element for the SVG rendition of the figure, in addition to the current 'artwork' for ASCII art. It may also be useful to provide for marking of text according to purpose. For example, a number of documents have started to include motivational material that indicates why design choices were taken. Having these discussions publically available may avoid repeating discussions when documents are progressed along the standards track. This material may not be needed by every implementor and it would thus be desirable to mark it for selective printing or viewing. Similarly, it may be helpful to mark examples, syntax descriptions or code. Among other benefits, it allows to automatically extract them for verification.

3.5 HTML

HTML is not suitable as an input or archive format and barely suitable as a presentation format. Unless crafted by hand, almost all HTML editors and output filters insert proprietary or extraneous formatting instructions that try to mimic the page layout rather than reflect the logical document structure. HTML printing is generally poor, particularly the generation of headers and footers. Printed versions have no page-based tables of content. However, until XML-displaying browsers are available, it may be desirable to produce HTML versions of RFCs. If XML or nroff is given, this can be readily done by third parties.

4 Internationalization

There are at least three possible ways the current ASCII-only, English-only RFC and I-D series could be internationalized, namely by allowing non-English content, supporting UTF-8 for protocol examples and for author names. However, not all of these are desirable.

4.1 Documents in Languages Other than English

It has been proposed to allow non-English-language documents, regardless of character set. This would cause a fragmentation of the Internet technical community and undue burden on the RFC and I-D editors. Authors that prefer to present their ideas in languages other than English are advised to circulate them within their language community and then translate them (or have them translated) when they are ready for wider discussion. Some countries and language communities also have local RFC-like series of documents, primarily in the local language and character set.

4.2 Character Sets beyond US-ASCII

An increasing number of documents need to refer to character sets other than US-ASCII. This includes examples in protocol documents, mostly in application-layer descriptions, but also for, say, DNS. Currently, they have to be rendered in ASCII, making presentation awkward and examples harder to understand. It has been suggested to create an alternate version of documents, labeled as rfcXXXX.iso, that contains UTF-8.

Similarly, many authors have their names rendered imperfectly in ASCII. It would be desirable to provide both ASCII and local representations of author names. Whether it is necessary to create a separate version of the document just for non-ASCII author names remains to be determined. At worst, printers and viewers tend to produce random characters for UTF-8 characters with the MSB set. Once host names allow non-ASCII characters, a similar problem may arise for email addresses.

There are existing tools, such as Omega (http://www.gutenberg.eu.org/omega/), Microsoft Wordpad, Microsoft Word and many other Windows word processors, that generate UTF-8.

5 Recommendations

- Make all RFCs available as PDF, using the original nroff sources where available or the PostScript sources. This has the slight disadvantage of creating possible ambiguities in page number references, but given the long use of PostScript RFCs, this seems to not have caused major problems.
- Create a tool that converts the XML RFC DTD or its successor to nroff.
- Initiate an effort to move an XML DTD for Internet drafts and RFCs onto the standards track, e.g., as a BCP. This may be based on RFC 2629 and the efforts within the W3C. Consider enhancing the DTD with tables, alternate graphics, and content marking (e.g., examples, BNF, motivation).
- Publish additional hints for RFC authors for producing documents from common sources.
- Investigate whether limited use of UTF-8 for examples and author names is feasible.

6 Acknowledgements

This document attempts to summarize the discussion on the IAB, wg-chairs and IESG mailing lists that took place during August 2000. Any misrepresentations are the fault of the author.

References

[1] G. Coulouris, “A note on the use of standards: Adopted, proposed and proprietary,” Mar. 1998.

[2] M. Rose, “Writing i-ds and RFCs using XML,” Request for Comments 2629, Internet Engineering Task Force, June 1999.

7 Authors’ Addresses

Henning Schulzrinne
Dept. of Computer Science

Columbia University
1214 Amsterdam Avenue
New York, NY 10027
USA
electronic mail: schulzrinne@cs.columbia.edu