News Article Format
draft-kohn-news-article-01.txt

Status of this Memo

This document is an Internet-Draft and is in full conformance with all provisions of Section 10 of RFC2026.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF), its areas, and its working groups. Note that other groups may also distribute working documents as Internet-Drafts.

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress."

The list of current Internet-Drafts can be accessed at http://www.ietf.org/ietf/1id-abstracts.txt.

The list of Internet-Draft Shadow Directories can be accessed at http://www.ietf.org/shadow.html.

This Internet-Draft will expire on August 19, 2003.

Copyright Notice

Copyright (C) The Internet Society (2003). All Rights Reserved.

Abstract

This document defines the format and procedures for interchange of network news articles. It updates and obsoletes RFC 1036, in particular adding support for internationalization of headers and message bodies and multimedia support in message bodies. It does this in a manner designed to maximize backward compatibility with news and mail servers, gateways, and user agents.

Network news articles resemble mail messages but are broadcast to potentially-large audiences, using a flooding algorithm that propagates one copy to each interested host (or group thereof), typically stores only one copy per host, and does not require any central administration or systematic registration of interested
users. Network news originated as the medium of communication for Usenet, circa 1980. Since then Usenet has grown explosively, and many Internet sites participate in it. In addition, the news technology is now in widespread use for other purposes, on the Internet and elsewhere.

Table of Contents

1. Introduction ............................................ 3
  1.1 Scope .............................................. 3
  1.2 Requirements Notation ................................. 3
  1.3 Syntax Notation .................................... 4
  1.4 Structure of This Document .......................... 4
2. Format .................................................. 5
  2.1 Base ............................................... 5
  2.2 MIME Conformance ................................... 5
  2.3 Other MIME Support ................................. 5
3. Headers .................................................. 6
  3.1 New Internet Message Format Headers ............... 6
  3.2 Mandatory Headers .................................. 6
  3.3 News headers ....................................... 6
    3.3.1 Newsgroups .................................... 6
    3.3.2 Path ........................................... 8
    3.3.3 Followup-To .................................... 8
    3.3.4 Expires ......................................... 9
    3.3.5 Control ......................................... 9
    3.3.6 Distribution .................................... 9
    3.3.7 Summary ......................................... 9
    3.3.8 Approved ........................................ 9
    3.3.9 Organization .................................... 9
    3.3.10 Xref ........................................... 10
    3.3.11 Supersedes ..................................... 10
  3.4 Other Mail Headers ................................. 10
4. Control Messages ...................................... 11
5. Security Considerations .............................. 12
  5.1 Normative References .............................. 13
  5.2 Informative References ............................ 14
  5.3 Author’s Address .................................. 15
A. Architectural Decisions ............................. 16
  A.1 Encoded Words vs. Raw UTF-8 ...................... 16
  A.2 Why Use Punycode Instead of UTF-7 .............. 16
B. Acknowledgements .................................... 17

Intellectual Property and Copyright Statements .......... 18
1. Introduction

1.1 Scope

"Netnews" is a set of protocols for generating, storing and retrieving news "articles" (which use the Internet Message Format) and for exchanging them among a readership which is potentially widely distributed. It is organized around "newsgroups", with the expectation that each reader will be able to see all articles posted to each newsgroup in which she participates. These protocols most commonly use a flooding algorithm which propagates copies throughout a network of participating servers. Typically, only one copy is stored per server, and each server makes it available on demand to readers able to access that server.

The predecessor to this document [RFC1036] said that: "In any situation where this standard conflicts with the Internet [email standard, the latter] should be considered correct and this standard in error." The basic philosophy of this document follows that previous convention, so as to standardize news article syntax firmly in the context of Internet Message Format syntax. In the context of the Internet messaging architecture, different protocols (such as IMAP [RFC2060], POP3 [RFC1939], NNTP [RFC0977] and SMTP [RFC2821]) are seen as alternative ways of moving around the same content. That content is the Internet Message Format as specified by [RFC2822], including optional enhancements such as MIME headers or bodies. A user should be able to ingest an article via NNTP, read it via IMAP, forward it off to someone else via SMTP and have them read it via POP3 all without having to alter the content.

This document uses a cite by reference methodology, rather than trying to repeat the contents of other standards, which could otherwise result in subtle differences and interoperability challenges. Although this document is as a result rather short, it requires complete understanding and implementation of the normative references to be compliant.

This document specifies only the syntax of compliant news articles. A companion document will be necessary to specify the policy requirements and recommendations of news agents, servers, and gateways.

1.2 Requirements Notation

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in [RFC2119].
1.3 Syntax Notation

Headers defined in this specification use the Augmented Backus-Naur Form (ABNF) notation specified in [RFC2234] and many constructs (including <date-time>, <mailbox-list>, <msg-id>, <unstructured>, and <utext>) defined in [RFC2822].

1.4 Structure of This Document

Section 2 defines the format of news articles. Section 3 defines some additional headers necessary for the netnews environment.
2. Format

2.1 Base

News articles MUST conform to the "legal to generate syntax" specified in Section 3 of [RFC2822]. News agents SHOULD also support the obsolete syntax specified in Section 4 of [RFC2822], particularly to support old news messages and gatewayed obsolete mail messages, but they MUST NOT generate such syntax.

2.2 MIME Conformance

News agents MUST meet the definition of MIME-conformance in [RFC2049]. In addition, news agents MUST support the i18n extensions for parameters, continuations, and language tagging specified in [RFC2231].

Section 2.10 of [RFC2049] describes the display of encoded-words. This document adds an additional requirement that for encoded words using the UTF-8 charset, the news agent MUST at least be able to display the characters which are also in the US-ASCII charset.

The one change from [RFC2047] is that while Section 3 of that document recommends that "members of the ISO-8859-* series be used in preference to other character sets", this document specifies that news agents SHOULD use UTF-8 as the charset for encoded words. Among other things, this is conformant with the IETF recommendations of [RFC2277].

2.3 Other MIME Support

News agents conformant with this document SHOULD support receipt (and automatic reassembly) of message/partial MIME messages, as specified in Section 5.2.2 of [RFC2046] and SHOULD support generation of message/partial articles for excessively large articles.

News agents SHOULD send regular paragraph text as "text/plain; format=flowed" as specified in [RFC2646] and SHOULD preserve flowed text (including quoting) when replying or forwarding, as described in that specification.

News agents MAY support Content-Disposition [RFC2183] and Content-Language [RFC3282].
3. Headers

3.1 New Internet Message Format Headers

Section 3.6 of [RFC2822] defines a series of permitted headers. This document extends that list as follows:

\[
\text{fields} \equiv / * ( \text{newsgroups} / \\
\text{path} / \\
\text{followup-to} / \\
\text{expires} / \\
\text{control} / \\
\text{distribution} / \\
\text{summary} / \\
\text{approved} / \\
\text{organization} / \\
\text{xref} / \\
\text{supersedes} )
\]

3.2 Mandatory Headers

Each news article conformant with this specification MUST have exactly one of each of the following headers: Date, From, Message-ID, Subject, Newsgroups, and Path. The first 4 are specified in [RFC2822].

3.3 News headers

3.3.1 Newsgroups

The Newsgroups header specifies to which newsgroup(s) the article is posted.
A newsgroup name consists of one or more components separated by periods, with no more than 71 characters total. Each component consists of less than 30 or less lowercase letters and digits, or is an encoded component. The order of newsgroup names in the Newsgroups header is not significant.

### 3.3.1.1 Encoded Components

Encoding of i18n newsgroup names follows the general approach laid out in [I-D.ietf-idn-idna]. Encoded components are strings of Unicode characters that have been normalized using [I-D.ietf-idn-nameprep] and encoded using [I-D.ietf-idn-punycode]. The main difference from [I-D.ietf-idn-idna] is that this specification limits encoded components to 30 characters, not 63. With the 4 character ACE prefix, that means that the output of punycode is limited to 26 characters.

This example encodes the newsgroup name that would be displayed as "test.3<nen>b<gumi><kinpachi><sensei>.misc" (the text in brackets is Japanese). The middle component consists of the Unicode string U+0033 U+5E74 U+0062 U+7D44 U+91D1 U+516B U+5148 U+751F. Punycode encodes that string as "3b-ww4c5e180e575a651sy2b". So, the resulting newsgroup name, which has been encoded so as to comply with this document, is "test.xn--3b-ww4c5e180e575a651sy2b.misc".
3.3.1.2 Interaction with Wildmat

The main value of using punycode newsgroup names is that the infrastructure of servers and gateways does not need to be upgraded before users can start taking advantage of i18n newsgroup names. The one exception of this is use of wildmat pattern matching within components, as specified by Section 3.3 of [RFC2980]. As specified, wildmat continues to work normally when doing matches between components, such as "test.*" matching the example newsgroup from the previous section. However, wildmat matching within encoded components does not work correctly, due to the presence of the ACE prefix. In the above example, "test.3*" should match the newsgroup name from the previous section, but will not.

Rectifying this will require a change in the standards-track successor to [RFC2980]. Specifically, such an upgraded wildmat format would probably need to specify that matching occurs in decoded form as Unicode characters.

3.3.2 Path

The Path-header shows the route taken by a message since its entry into the Netnews system.

\[
\text{path} = "\text{Path:}\" [\text{FWS}] \\
* ( \text{path-host} [\text{FWS}] \text{path-delimiter} [\text{FWS}] ) \\
\text{path-host} [\text{FWS}] \text{CRLF}
\]

\[
\text{path-host} = ( \text{ALPHA} / \text{DIGIT} ) \\
* ( \text{ALPHA} / \text{DIGIT} / "-" / "." / ":" / "_" )
\]

\[
\text{path-delimiter} = "/" / ":" / ":" / "," / ":!
\]

3.3.3 Followup-To

The Followup-To header specifies to which newsgroup(s) followups should be posted.

\[
\text{followup-to} = "\text{Followup-To:}\" ( \text{newsgroup-list} / \text{poster-text} ) \\
\text{CRLF}
\]

\[
\text{poster-text} = [\text{FWS}] \%70.6F.73.74.65.72 [\text{FWS}] \\
; "\text{poster}\" \text{in lower-case}
\]

The syntax is the same as that of the Newsgroups content, with the exception that the magic word "poster" (which is always lowercase) means that followups should be mailed to the article’s reply address rather than posted. In the absence of Followup-To, the default
newsgroup(s) for a followup are those in the Newsgroups header.

### 3.3.4 Expires

The Expires header content specifies a date and time when the article is deemed to be no longer useful and should be removed ("expired").

```
expires       = "Expires:" date-time CRLF
```

### 3.3.5 Control

The Control-header marks the article as a control message, and specifies the desired actions (additional to the usual ones of storing and/or relaying the article). Control messages are further specified in Section 4.

```
control       = "Control:" verb *( FWS argument ) CRLF
```

### 3.3.6 Distribution

The Distribution header content specifies geographic or organizational limits on an article’s propagation.

```
distribution = "Distribution:" dist-name *( "," dist-name ) CRLF
dist-name    = [FWS] ALPHA *( ALPHA / \+ / \- / \_ ) [FWS]
```

### 3.3.7 Summary

The Summary header content is a short phrase summarizing the article’s content.

```
summary      = "Summary:" unstructured CRLF
```

### 3.3.8 Approved

The Approved header content indicates the mailing addresses (and possibly the full names) of the persons or entities approving the article for posting.

```
approved     = "Approved:" mailbox-list CRLF
```

### 3.3.9 Organization

The Organization header content is a short phrase identifying the poster’s organization.

```
organization = "Organization:" unstructured CRLF
```
3.3.10 Xref

The Xref header content indicates where an article was filed by the last relayer to process it.

\[
xref = "Xref:" [CFWS] path-host
1*( CFWS location ) [CFWS]
\]

location = newsgroup-name ":" utext

3.3.11 Supersedes

The Supersedes header content specifies articles to be cancelled.

\[
supersedes = "Supersedes:" 1*msg-id CRLF
\]

3.4 Other Mail Headers

The headers Reply-To, Sender, In-Reply-To, References, Comments, and Keywords are often used in news articles and have the identical syntax to that specified in [RFC2822].
4. Control Messages

Describe control messages here, including definition for <verb> and <argument>.
5. Security Considerations

The news article format specified in this document does not provide any security services, such as confidentiality, authentication of sender, or non-forgery. Instead, such services need to be layered above, using such protocols as S/MIME [RFC2633] or PGP/MIME [RFC3156], or below, using secure versions of news transport protocols. Additionally, several currently non-standardized protocols [PGPVERIFY] will hopefully be standardized in the near future.
Normative References

[I-D.ietf-idn-nameprep]

[I-D.ietf-idn-punycode]


Informative References

[I-D.ietf-idn-idna]
Hoffman, P., Faltstrom, P. and A. Costello,
"Internationalizing Domain Names In Applications (IDNA)",

[PGPVERIFY]


Author’s Address

Dan Kohn
Skymoon Ventures
3045 Park Boulevard
Palo Alto, California 94306
USA

Phone: +1-650-327-2600
EMail: dan@dankohn.com
URI: http://www.dankohn.com/
Appendix A. Architectural Decisions

A.1 Encoded Words vs. Raw UTF-8

A significant amount of work was done proposing that Usenet use raw UTF-8 [RFC2279] in headers to accomplish i18n rather than 2047/2231 encoded words and punycode for newsgroup names. The main problem with raw UTF-8 is that every user agent, server, and gateway in an article’s path needs to be upgraded in order to ensure successful transmission. This is especially problematic in news-to-mail gateways used by most moderators. By contrast, no upgrades are necessary for successful transmission of articles with i18n headers encoded with 2047/2231/punycode. Of course, support for these encodings is necessary in transmitting and receiving user agents to properly display i18n text, but even un-upgraded user agents can still interact with i18n articles like they do with existing ones (such as by selecting an i18n newsgroup by entering its punycode encoded name), with the exception that i18n headers may look garbled. This, of course, provides an incentive for the user to upgrade. Upgrades of the infrastructure remain unnecessary, with the exception of wildmat as specified in Section 3.3.1.2.

A.2 Why Use Punycode Instead of UTF-7

Since punycode support will already need to be implemented in user agents that support IDNA [I-D.ietf-idn-idna], support for nameprep and punycode is expected not to require much additional development. Punycode compresses much better than UTF-7 [RFC2152], and for much text, better than UTF-8. Punycode doesn’t apply special meaning to the "+" character which is currently used by newsgroup names. Finally, the "xn--" delimiter uniquely identifies encoded components.
Appendix B. Acknowledgements

Portions of this text were taken from "son-of-1036" by Henry Spencer and other portions from a draft by Charles Lindsay. Comments on ietf-822@imc.org inspired this approach. The idea of punycode-encoded newsgroups was suggested in a draft by Claus Faerber. Useful comments were provided by Mark Crispin and Ken Murchinson.
Intellectual Property Statement

The IETF takes no position regarding the validity or scope of any intellectual property or other rights that might be claimed to pertain to the implementation or use of the technology described in this document or the extent to which any license under such rights might or might not be available; neither does it represent that it has made any effort to identify any such rights. Information on the IETF’s procedures with respect to rights in standards-track and standards-related documentation can be found in BCP-11. Copies of claims of rights made available for publication and any assurances of licenses to be made available, or the result of an attempt made to obtain a general license or permission for the use of such proprietary rights by implementors or users of this specification can be obtained from the IETF Secretariat.

The IETF invites any interested party to bring to its attention any copyrights, patents or patent applications, or other proprietary rights which may cover technology that may be required to practice this standard. Please address the information to the IETF Executive Director.

Full Copyright Statement

Copyright (C) The Internet Society (2003). All Rights Reserved.

This document and translations of it may be copied and furnished to others, and derivative works that comment on or otherwise explain it or assist in its implementation may be prepared, copied, published and distributed, in whole or in part, without restriction of any kind, provided that the above copyright notice and this paragraph are included on all such copies and derivative works. However, this document itself may not be modified in any way, such as by removing the copyright notice or references to the Internet Society or other Internet organizations, except as needed for the purpose of developing Internet standards in which case the procedures for copyrights defined in the Internet Standards process must be followed, or as required to translate it into languages other than English.

The limited permissions granted above are perpetual and will not be revoked by the Internet Society or its successors or assignees.

This document and the information contained herein is provided on an "AS IS" basis and THE INTERNET SOCIETY AND THE INTERNET ENGINEERING TASK FORCE DISCLAIMS ALL WARRANTIES, EXPRESS OR IMPLIED, INCLUDING BUT NOT LIMITED TO ANY WARRANTY THAT THE USE OF THE INFORMATION
HEREIN WILL NOT INFRINGE ANY RIGHTS OR ANY IMPLIED WARRANTIES OF MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE.

Acknowledgement

Funding for the RFC Editor function is currently provided by the Internet Society.