Technical Reports

L2/26-052

Draft Unicode® Technical Standard #58

UNICODE LINK DETECTION AND FORMATTING: URLS AND EMAIL ADDRESSES

Version	17.0 (draft 7)
Editors	Mark Davis, Markus Scherer
Date	2025-12-29
This Version	https://www.unicode.org/reports/tr58/tr58-1.html
Previous Version	none
Latest Version	https://www.unicode.org/reports/tr58/
Latest Proposed Update	https://www.unicode.org/reports/tr58/proposed.html
Revision	1

Summary

When URLs are stored and exchanged in structured data, the start and end of each URL is clear, and it can be parsed according to the relevant specifications. However, when URLs appear as unmarked strings in text content, detecting their boundaries can be challenging. For example, some characters that are often used as sentence-level punctuation in text, such as parentheses, commas, and periods, can also be valid characters within a URL. Implementations often do not behave intuitively and consistently.

When a URL is inserted into text, non-ASCII characters and "special" characters can be percent-encoded, which can make it easy for a later process to find the start and end of the URL. However, escaping more characters than necessary, especially normal letters, can make the URL illegible for a human reader.

Similar problems exist for email addresses.

This document specifies two consistent, standardized mechanisms that address these problems, consisting of:

1. ***link detection**: mechanisms for detecting URLs and email addresses embedded in plain text that properly handles non-ASCII characters, and*
2. ***minimally escaping**: mechanisms for minimal escaping of non-ASCII code points in the Path, Query, and Fragment portions of a URL, and in the local-part of an email address.*

The focus is on links with the Schemes `http:`, `https:`, and `mailto:` — and links where those Schemes are missing but implied. For these cases, the two mechanisms of detecting and formatting are aligned, so that: a minimally escaped URL string between two spaces in flowing text is accurately detected, and a detected URL works when pasted into address bars of major browsers.

Status

*This is a **draft** document which may be updated, replaced, or superseded by other documents at any time. Publication does not imply endorsement by the Unicode Consortium. This is not a stable*

document; it is inappropriate to cite this document as other than a work in progress.

> ***A Unicode Technical Standard (UTS)*** *is an independent specification. Conformance to the Unicode Standard does not imply conformance to any UTS.*

Please submit corrigenda and other comments with the online reporting form [Feedback]. Related information that is useful in understanding this document is found in the References. For more information see About Unicode Technical Reports and the Specifications FAQ. Unicode Technical Reports are governed by the Unicode Terms of Use.

Contents

Review Note: TBD: Sync ToC entries vs. heading numbers and titles before final.

---

1 Introduction

1.1 URLs

The standards for URLs and their implementations in browsers generally handle Unicode quite well, permitting people around the world to use their writing systems in those URLs. This is important: in

writing their native languages, the majority of humanity uses characters that are not limited to A-Z, and they expect other characters to work equally well. But there are certain ways in which their characters fail to work seamlessly. For example, consider the common practice of providing user handles such as:

- x.com/rihanna
- bsky.app/profile/jaketapper.bsky.social
- www.instagram.com/vancityreynolds/
- www.youtube.com/@핑크퐁

The first three of these work well in practice. Copying from the address bar and pasting into text provides a readable result. However, the last example contains non-ASCII characters. In many browsers this turns into an unreadable string:

- www.youtube.com/@핑크퐁 (desirable display)
- https://www.youtube.com/@%ED%95%91%ED%81%AC%ED%90%81 (in many browsers)

The names also expand in size and turn into very long strings:

- https://hi.wikipedia.org/wiki/महात्मा_गांधी
- https://hi.wikipedia.org/wiki/%E0%A4%AE%E0%A4%B9...%E0%A5%80

While many people cannot read "महात्मा_गांधी", *nobody* can read %E0%A4%AE%E0%A4%B9...%E0%A5%80. This unintentional obfuscation also happens with URLs using Latin-script characters:

- https://en.wikipedia.org/wiki/Antonín_Dvořák
- https://en.wikipedia.org/wiki/Anton%C3%ADn_Dvo%C5%99%C3%A1k

Very few languages using Latin-script characters are limited to the ASCII letters A-Z; English being a notable exception. This situation is doubly frustrating for people because the un-obfuscated URLs such as https://www.youtube.com/@핑크퐁 and https://en.wikipedia.org/wiki/Antonín_Dvořák work fine as plain text; you can copy and paste them back into your address bar — they go to the right page *and display properly in the address bar*.

Notes

- This specification uses the term **URL** broadly, as including unescaped non-ASCII characters; in other words, treating it as matching the formal definition of IRIs. Standardizing on the term “URL” and avoiding the terms “URI” and “IRI” follows the practice promoted by the WHATWG in [URL Standard: Goals].
  See also the W3C’s [An Introduction to Multilingual Web Addresses].
- In examples, links will be shown with a background color, to make the extent of the linkification clear.

1.2 Email Addresses

Email addresses should also work well for all languages. With most email programs, when someone pastes in the plain text:

- The page https://ja.wikipedia.org/wiki/アルベルト・アインシュタイン contains information about Albert Einstein.

and sends to someone else, they receive it as:

- The page https://ja.wikipedia.org/wiki/アルベルト・アインシュタイン contains information about Albert Einstein.

1.3 Displaying Unmarked URLs and Email Addresses

URLs are “linkified” in many applications, such when pasting into a word processor (triggered by typing a space afterwards, for example). However, many products (many text messaging apps, video messaging chats, etc.) completely fail to recognize any non-ASCII characters past the domain name. And even among those that do recognize such non-ASCII characters, there are gratuitous differences in where they *stop* linkifying.

Linkification is the process of adding links to URLs and email addresses in plain text input, such as in email body text, text messaging, or video meeting chats. The first step in this process is *link detection*, which is determining the boundaries of spans of text that contain URLs. That substring can then have a link applied to it in output text. The functions that perform these operations are called a *link detector* and *linkifier*, respectively.

The specifications for a URL don’t specify how to handle link detection, since they are only concerned with the structure in isolation, not when it is embedded within flowing text. The lack of a clear specification for link detection also causes many implementations to overuse percent escaping for non-ASCII characters when converting URLs into plain text.

Different implementations linkify URLs and email addresses differently even when they contain only ASCII characters. The differences are even greater when non-ASCII characters are used. Handling letters of all writing systems well is very important for usability. Consider the last example above of a sentence in an email when displayed with a percent-escaped URL:

- The page https://ja.wikipedia.org/wiki/%E3%82%A2%E3%83%AB...%E3%83%B3 contains information about Albert Einstein.

For example, take the lists of links on [List of articles every Wikipedia should have] in the available languages. When those links are tested with major products, there are significant differences: any two implementations are likely to linkify those differently, such as terminating the linkification at different places, or not linkifying at all. That makes it very difficult to exchange URLs between products within plain text, which is done surprisingly often — definitely causing problems for implementations that need predictable behavior.

This inconsistency causes problems for users and software companies. Having consistent rules for linkification also has additional benefits, leading to solutions for the following reported problems:

- If a system allows users to have their own user ids that end up in URLs, like https://www.linkedin.com/in/my.user.name, it can avoid user ids that have problematic linkification behavior, like trailing periods after path segments.
- Because linkification cannot be predicted for URLs with non-ASCII characters, common practice is to exchange them with escaped characters, which gives unreadable results such as the long line above.

If linkification behavior becomes more predictable across platforms and applications, applications will be able to do minimal escaping. For example, in the following only one character would need escaping, the %29 — representing an unmatched “)”:

- https://ja.wikipedia.org/wiki/アルベルト%29アインシュタイン

Providing a consistent, predictable solution that works well across the world’s languages requires standardized algorithms to define the behavior, and the corresponding Unicode character properties covering all Unicode characters.

2 Conformance

UTS58-C1. *For a given version of Unicode, a conformant implementation shall replicate the same link detection results as those produced by Section 3, Link Detection Algorithm.*

UTS58-C2. *For a given version of Unicode, a conformant implementation shall replicate the same minimal escaping results as those produced by Section 4, Minimal Escaping.*

UTS58-C3. *For a given version of Unicode, a conformant implementation shall replicate the same email link detection results as those produced by Section 5, Email Addresses.*

3 Link Detection

The following table shows the relevant parts of a URL. For clarity, the separator characters are included in the examples. For more information see [WhatWG URL: Example URL Components].

Table 3-1. Parts of a URL

Scheme	*Host (incl. Domain)*	*Port*	*Path*	*Query*	*Fragment*
https://	docs.foobar.com	:8000	/knowledge/area/	?name=article&topic=seo	#top

Notes:

- The Scheme, Port, Path, Query, and Fragment are each optional.
- Each of the Parts may have internal structure, such as:
  - The Host just consists of a domain, which consists of a list of one or more labels separated by "." such as `example.com`. ()The syntax of a URI actually permits a `userinfo` component, such as `username:password@example.com`, but its use is deprecated due to security concerns.
  - The Path consists of one or more segments, separated by "/".
  - The Query typically consists of one or more key-value pairs, separated by "&"; where each key is separated from its value by "=". (There are other possible structures, but this structure is seen most commonly.
  - The Fragment has various possible structures defined by web applications, and at the end can contain one or more fragment directives, starting with a separator ":~:", and separated again by the sequence ":~:".
- The goal for this specification is to handle the Query and Fragment structures that are most common, where matching brackets shouldn't typically span internal separators.

Processes

There are two main processes involved in Unicode link detection.

1. **Initiation.** This requires determining the point within plain text where the parsing of a URL starts. When the Scheme is present for a URL (such as “http://”), determining the start of link detection is simple. However, the Scheme for a URL is commonly omitted when URLs are represented in text. For example, the string “*adobe.com*” should be recognized as being a URL when it occurs in the body of an email message, even though it does not have a Scheme.
2. **Termination.** This requires determining the point within plain text where the parsing of a URL ends. A formal reading of the URL specs allows almost any character in certain URL parts, so it is insufficient for separating the end of the URL from the non-URL text after it.

There are two special cases. Both of these introduce some complications in the algorithm, because each of the Parts have different internal syntax and different initial characters, and can be followed by different Parts.

1. "Soft" characters are not included in the link, unless they are followed by other characters that would be included.
   - Example:
     1. With “See abc.com?def!” the ! is not included.
     2. However, with “See abc.com?def!ghi” it would be.
2. Closing brackets are not included in the link, unless they have a matching opening bracket — *that doesn’t cross syntax characters*.
   - Example: For ‘)’
     1. Not included with “(See abc.com?def=a). And…”

2. Is included with “See abc.com?def=(a). And…”

The algorithm is a single-pass algorithm with backup, that is, remembering the latest ‘safe’ point to break, and returning that where necessary. It also has a stack, so that it can determine when a closing bracket matches.

Initiation

The start of a URL is easy to determine when it has a known Scheme (eg, “https://”).

Implementations have also developed heuristics for determining the start of the URL when the Scheme is elided, taking advantage of the fact that there are relatively few top-level domains. And those techniques can be easily applied to internationalized domain names, which still have strong limitations on the valid characters. So the end of the domain name is also relatively easy to determine. For more information, see UTS #46, Unicode IDNA Compatibility Processing.

The parsing up to the path, query, or fragment is as specified in [WHATWG URL: 4.4. URL parsing].

For example, implementations must terminate link detection if a *forbidden host code point* is encountered, or if the host is a domain and a *forbidden domain code point* is encountered. Implementations must not linkify if a domain is not a *registrable domain*. The terms *forbidden host code point*, *forbidden domain code point*, and *registrable domain* are defined in [WHATWG URL: Host representation].

For example, an implementation would parse to the end of microsoft.com and google.de, foo.pф, or xn--j1ay.xn--p1ai.

Termination

Termination is much more challenging, because of the presence of characters from many different writing systems. While small, hard-coded sets of characters suffice for an ASCII implementation, there are over 150,000 Unicode characters, many with quite different behavior than ASCII. While in theory, almost any Unicode character can occur in certain URL parts, in practice many characters have very restricted usage in URLs.

Initiation stops at any Path, Query, or Fragment, so the termination process takes over with a “/”, “?”, or “#” character. Each Path, Query, or Fragment can contain most Unicode characters. The key is to be able to determine, given a URL Part (such as a Query), when a sequence of characters should cause termination of the link detection, even though that character would be valid in the URL specification.

It is impossible for a link detection algorithm to match user expectations in all circumstances, given the variation in usage of various characters both within and across languages. So the goal is to cover use cases as broadly as possible, recognizing that it will sometimes not match user expectations in certain cases. Exceptional cases (URLs that need to use characters that would terminate) can still be appropriately linkified if those few characters are represented with % escapes.

At a high level, this specification defines three features:

1. A method for identifying when to terminate link detection based on properties that define contexts for terminating the parsing of a URL.
    - This addresses the question, for example, when a trailing period should be included in a link or not.
2. A method for identifying balanced quotes and brackets that enclose a URL.
    - This addresses the distinction, for example, of enclosing the entire URL in parentheses, vs. URLs that contain a segment that is enclosed in parens, etc.
3. An algorithm for doing the above, together with an enumerated property and a mapping property.

The focus is on the high-runner cases.

- Links for http://, https://, and mailto:
- Instances where those schemes are omitted.
- Handling internal structures of queries and fragments that most often occur.

One of the goals is also predictability; it should be relatively easy for users to understand the link detection behavior at a high level.

Properties

Review Note: The names for Link_Termination and Link_Paired_Opener have changed to Link_Term and Link_Bracket. This change is not marked in the text, for ease of reading.

This specification defines three properties. The first two are used in URL link detection and formatting, and the last is used in email link detection and formatting.

- Link_Term
- Link_Bracket
- Link_Email

The short property names are identical to the long property names.

Link_Term Property

Link_Term is an enumerated property of characters with five enumerated values: {**Include**, **Hard**, **Soft**, **Close**, **Open**}
The short property value aliases are the same as the long ones.

Table 3-2. Link_Term Property Values

Value	Description / Examples
Include	There is no stop before the character; it is included in the link.
	Example: *letters* • https://ja.wikipedia.org/wiki/アルベルト・アインシュタイン
Hard	The URL terminates before this character.
	Example: *a space* • Go to https://ja.wikipedia.org/wiki/アルベルト・アインシュタイン to find the material.
Soft	The URL terminates before this character, **if** it is followed by `/\p{Link_Term=Soft}*(\p{Link_Term=Hard}|$)/`
	Example: *a question mark* • https://ja.wikipedia.org/wiki/アルベルト・アインシュタイン?abc • https://ja.wikipedia.org/wiki/アルベルト・アインシュタイン? abc • https://ja.wikipedia.org/wiki/アルベルト・アインシュタイン?
Close	If the character is paired with a previous character *in the same URL Part* (path, query, fragment) and within the same sequence of characters delimited by separators as described in the Termination Algorithm below, it is treated as **Include**. Otherwise it is treated as **Hard**.

Value	Description / Examples
	Example: *an end parenthesis* • https://ja.wikipedia.org/wiki/(アルベルト)アインシュタインアインシュタイン) • (https://ja.wikipedia.org/wiki/アルベルト)アインシュタイン • (https://ja.wikipedia.org/wiki/アルベルトアインシュタイン
Open	Used to match **Close** characters.
	Example: *same as under **Close***

Link_Bracket Property

Link_Bracket is a string property of characters, which for each character in \p{Link_Term=Close}, returns a character with \p{Link_Term=Open}.

Example

1. Link_Bracket('**}**') == '**{**'

Link_Email Property

Link_Email is a binary property of characters, indicating the characters that can normally occur in the `local-part` of an email address, such as `σωκράτης@example.om`

Example

1. Link_Email('**σ**') == '**Yes**'

The specification of the characters with each of these property values is given in Property Assignments.

Termination Algorithm

The termination algorithm assumes that a domain (or other host) has been successfully parsed to the start of a Path, Query, or Fragment, as per the algorithm in [WHATWG URL:3. Hosts (domains and IP addresses)].

This algorithm then processes each final URL Part [path, query, fragment] of the URL in turn. It stops when it encounters a code point that meets one of the terminating conditions and reports the last location in the current URL Part that is still safely considered inside the link. The common terminating conditions are based on the Link_Term and Link_Bracket properties:

- A `Link_Term=Hard` character, such as a *space*.
  - In addition, while processing a certain URL Part, its corresponding terminator characters and sequences also terminate that URL Part.
- A `Link_Term=Soft` character, such as a *?* that is followed by a sequence of zero or more `Soft` characters, then either a `Hard` character or the end of the text.
- A `Link_Term=Close` character, such as a *]* that does **not** have a matching `Open` character *in the same Part* of the URL. The matching process uses the Link_Bracket property to determine the correct Open character, and matches against the top element of a stack of Open characters.

More formally:

The termination algorithm begins after the Host (and optionally Port) have been parsed, so there is potentially a Path, Query, or Fragment. In the algorithm below, each Part has three sets of Action strings that affect transitions within and between Parts:

Sequence Sets	Actions
Initiator	Starts the Part
Terminator Set	Terminates the Part
ClearStackOpen Set	Clears the stack of open brackets within the Part

Here are the sets of zero or more strings in each Sequence Set for each Part. (UnicodeSet notation is used here and elsewhere in this document, in which a backslash is used to escape characters like &.)

Table 3-3. Link Termination by URL Part

Part	Initiator	Terminator set	ClearStackOpen set
path	'/'	[?#]	[/]
query	'?'	[#]	[=\&]
fragment	'#'	[{:~:}]	[]
fragment directive	:~:	[]	[\&,{:~:}]

Fragment directives:

- The initiator is only activated if already in a fragment or in a fragment directive.
  There may be multiple fragment directives in a single URL.
- Currently the only fragment directive that has been defined is the `text` directive, as in `https://example.com#:~:text=foo&text=bar`.
- Additional fragment directives may be defined in the future, and their internal structure may differ from that of the text directive. At that time, this algorithm will need to be adjusted, including new rows in the table above and adjusting the initiators, terminators, and clearStackOpen.
- For more information, see [URL Fragment Text Directives].

Review Note: In a fragment directive, the comma and ampersand are separators, and thus cause the stack of open brackets to be cleared. The dash '-' is an affix to the comma, rather than a separator, as the following syntax shows:
`#:~:text=[prefix-,]start[,end][,-suffix]`

Link-Detection Algorithm

In the following:

- `cp[i]` refers to the i^{th} code point in the string being parsed, `cp[start]` is the first code point being considered, and `n` is the length of the string.
- A stack (`openStack`) is used for matching brackets. A limit is required for security; the value 125 is chosen deliberately to far exceed any reasonable number of paired brackets.

---

1. Set `lastSafe = start` — *this marks the offset after the last code point that is included in the link detection (so far).*
2. Set `part = none`.
3. Clear the `openStack`.
4. Loop from `i = start` to `n - 1`
    1. If `part ≠ none` and one of the `part.terminators` matches at `i`
        1. Set `previousPart = part`.
        2. Set `part = none`.
    2. If `part == none` then try to match one of the URL Part `initiator`s at `i`.
        1. If none of the `initiator`s match, then stop and return `lastSafe`.

2. Set `part` according to which URL Part's `initiator` matches.
3. If `part` is a Fragment Directive and `previousPart` is neither a Fragment nor a Fragment Directive, then stop and return `lastSafe`.
4. Set `i` to just after the matched `part.initiator`.
5. Set `lastSafe = i`.
6. Clear the `openStack`.
7. Continue loop

3. If one of the `part.clearStackOpen` elements matches at `i`
    1. Set `i` to just after the matched `part.clearStackOpen` element.
    2. Set `lastSafe = i`.
    3. Clear the `openStack`.
    4. Continue loop
4. Set `LT = Link_Term(cp[i])`.
5. If `LT == Include`
    1. Set `lastSafe = i + 1`.
    2. Continue loop
6. If `LT` == `Soft`
    1. Continue loop
7. If `LT` == `Hard`
    1. Stop and return `lastSafe`
8. If `LT` == `Open`
    1. If `openStack.length() == 125`, then stop and return `lastSafe`.
    2. Push `cp[i]` onto `openStack`
    3. Set `lastSafe = i + 1`.
    4. Continue loop.
9. If `LT` == `Close`
    1. If `openStack.isEmpty()`, then stop and return `lastSafe`.
    2. Set `lastOpen = openStack.pop()`.
    3. If `Link_Bracket(cp[i]) == lastOpen`
        1. Set `lastSafe = i + 1`.
        2. Continue loop.
    4. Else stop and return `lastSafe`.
5. After the loop terminates, return `lastSafe`.

As usual, any algorithm that produces the same results is conformant. Such algorithms can be optimized in various ways, and adapted to be single-pass.

For ease of understanding, this algorithm does not include all features of URL parsing. In implementations, the algorithm can be optimized in various ways, of course, as long as the results are the same.

4 Minimal Escaping

The goal is to be able to generate a serialized form of a URL that:

1. is correctly parsed by modern browsers and other devices
2. minimizes the use of percent-escapes
3. is completely link-detected when isolated.
    1. For example, “abc.com/path1./path2.” would serialize as "abc.com/path1./path2%2E" so that linkification will identify all of the serialized form within plain text such as “See abc.com/path1./path2%2E for more information”.

2. If not surrounded by Hard characters, the linkification may extend beyond the bounds of the serialized form. For example, “See Xabc.com/path1./path2%2EX for more information”.

The minimal escaping algorithm is parallel to the linkification algorithm. Basically, when serializing a URL, a character in a Path, Query, or Fragment is only percent-escaped if it is: Hard, Close when unmatched, or Soft when it is (in) a URL Part terminator in the enclosing URL Part.

Minimal Escaping Algorithm

This algorithm only handles the formatting of the Path, Query, and Fragment URL Parts. Formatting of the Scheme, Host, and Port should be done as is customary for those URL Parts. For the Host (domain name), see also UTS #46: Unicode IDNA Compatibility Processing and its ToUnicode operation.

In the following:

- `cp[i]` refers to the i^{th} code point in the URL `part` being serialized, `cp[0]` is the first code point in the `part`, and `n` is the number of code points.
- The algorithm assumes that the Path, Query, Fragment, and Fragment directives already have the normal interior escaping for syntactic characters, including the the `part.terminators` and `part.clearStack`, to prevent them from being interpreted as literals:
  - For Path, that means that literal [?#/] must be escaped.
  - For Query, that means that literal [+#=\&] must be escaped. The + is in addition, because of its use as a replacement for space.
  - For Fragment, that means that the first character of a literal ":~:" must be escaped.
  - For Fragment directive, that means that [\&,] must be escaped, as well as the first character of a literal ":~:".
- A URL's internal model may contain bytes that arise from a page being in a legacy (non-UTF-8) character encoding. It is important, especially in the Query, to maintain those bytes even when they are invalid in UTF-8, such as %FF or %C2%C2. If the URL is known to originate in a page with a legacy character encoding (such as in an href value in that page), or is otherwise detected to have any invalid UTF-8 sequences, then an alternate formatting strategy should be used, such as percent-escaping each non-ASCII byte.

---

1. Set `output = ""`
2. For each URL `part` in any non-empty Path, Query, Fragment, successively:
    1. Append to `output`: `part.initiator`
    2. Set `copiedAlready = 0`
    3. Clear the `openStack`
    4. Loop from `i = 0` to `n - 1`
        1. If one of the `part.terminators` matches at `i`
            1. Set `LT = Hard`
        2. Else set `LT = Link_Term(cp[i])`
        3. If one of the `part.clearStackOpen` elements matches at `i`, clear the `openStack`.
        4. If `LT == Include`
            1. Append to `output`: any code points between `copiedAlready` (inclusive) and `i` (exclusive)
            2. Append to `output`: `cp[i]`
            3. Set `copiedAlready = i + 1`
            4. Continue loop
        5. If `LT == Hard`
            1. Append to `output`: any code points between `copiedAlready` (inclusive) and `i` (exclusive)
            2. Append to `output`: `percentEscape(cp[i])`

3. Set `copiedAlready = i + 1`
4. Continue loop

6. If `LT == Soft`
    1. Continue loop
7. If `LT == Open`
    1. If `openStack.length() == 125`, then do the same as `LT == Hard`.
    2. Else push `cp[i]` onto `openStack` and do the same as `LT == Include`
8. If `LT == Close`
    1. Set `lastOpen = openStack.pop()`, or 0 if the `openStack` is empty
    2. If `Link_Bracket(cp[i]) == lastOpen`
        1. Do the same as `LT == Include`
    3. Else do the same as `LT == Hard`

5. If `part` is not last
    1. Append to `output`: all code points between `copiedAlready` (inclusive) and `n` (exclusive)
6. Else if `copiedAlready < n`
    1. Append to `output`: all code points between `copiedAlready` (inclusive) and `n - 1` (exclusive)
    2. Append to `output`: `percentEscape(cp[n - 1])`
3. Return output.

As usual, any algorithm that produces the same results is conformant. Such algorithms can be optimized in various ways, and adapted to be single-pass.

Additional characters can be escaped to reduce confusability, especially when they are confusable with URL syntax characters, such as a ? character in a path. See Security Considerations below.

5 Email Addresses

Email address link detection applies similar principles. An email address is of the form `local-part@domain-name`. The local-part can include unusual characters by quoting: enclosing it in "…", and using backslash to escape those characters. For example, `"john\ doe"@example.com` contains an escaped space. However, the quoted local-part format is very rarely implemented, so this algorithm does not support it. Implementations are free to extend this algorithm to support such quoted email local-part formats. The algorithm is invoked whenever an '@' character is encountered at index `n`, and another process has determined that the '@' sign is followed by a valid domain name. The algorithm scans *backward* from the '@' sign to find the *start* of the local-part, terminating at index `end` (exclusive). if there is a "mailto:" before the local-part, then that is also included.

The only complications are introduced by the requirement in the specifications that the local-part cannot start or end with a ".", nor contain "..". For details of the format, see [RFC6530].

The algorithm uses the property `Link_Email` to scan backwards, as follows.

---

1. If `n = 0`, fail to match.
2. If `n > 0` and `cp[i] == '.'`, fail to match.
3. Scan backward through the text from `i = n - 1` down to `0`.
    1. If `cp[i] == '.'`
        1. If `cp[i + 1] == '.'`, fail to match.
        2. Else continue scanning backward.
    2. Else if `cp[i]` is not in `Link_Email`, set `start = i + 1` and terminate scanning.
    3. Else continue scanning backwards.

4. If `cp[start] == '.'`, fail to match.
5. If `start = n`, fail to match.
6. If "mailto:" is immediately before *start*, then set `start = start-7`.
7. Return a match for the pair `start, end`.

As usual, any algorithm that produces the same results is conformant. Such algorithms can be optimized in various ways, and adapted to be single-pass.

A quoted local-part may include a broad range of Unicode characters. See [RFC6530]. For linkification, the values in a quoted local-part — while broader than in an unquoted locale-part — are more restrictive to prevent accidentally including linkifying more text than intended, especially since those code points are unlikely to be handled by mail servers in any event.

Table 5-1. Email Address Link Detection Examples

See abcd@example.com	Stop backing up when a space is hit
See x.abcd@example.com	Include the medial dot.
See アルベルト.アルベルト@example.com	Handle non-ASCII
See @example.😎	No valid domain name
See @example.com	No local-part
See john.@example.com	No valid local-part
See john..doe@example.com	No valid local-part
See .john.doe@example.com	No valid local-part

Review Note: The algorithm causes linkification to fail in where the dots are illegal, such as: the last 3 examples. For the last two cases, instead of failing, the linkification could stop just before the problematic dots, such as: "john..doe@example.com" and ".john.doe@example.com". That approach is more error-prone, but could be supported with a customized algorithm.

Minimal Quoting Algorithm

The Minimal quoting algorithm for email addresses is trivial, given that the quoted forms are not supported.

6 Property Data

The assignments of Link_Term and Link_Bracket property values are defined by the following files:.

- LinkTerm.txt
- LinkBracket.txt
- LinkEmail.txt

Property Assignments

The initial property assignments are based on the following descriptions. However, their values may deviate from these descriptions in future versions. See Stability. Note that most characters that cause link termination are still valid, but require % encoding.

Link_Term=Hard

Whitespace, non-characters, deprecated characters, controls, private-use, surrogates, unassigned,...

- `[\p{whitespace}\p{NChar}[\p{C}-\p{Cf}]\p{deprecated}]`

Link_Term=Soft

Termination characters and ambiguous quotation marks:

- `\p{Term}`
- `\p{lb=qu}`

Link_Term=Open, Link_Term=Close

if Bidi_Paired_Bracket_Type(cp) == Open then Link_Term(cp) = Open

else if Bidi_Paired_Bracket_Type(cp) == Close then Link_Term(cp) = Close

else if cp == "<" then Link_Term(cp) = Open

else if cp == ">" then Link_Term(cp) = Close

Link_Term=Include

All other code points

Link_Bracket

if Bidi_Paired_Bracket_Type(cp) == Close then Link_Bracket(cp) = Bidi_Paired_Bracket(cp)

else if cp == ">" then Link_Bracket(cp) = "<"

else Link_Bracket(cp) = `<none>`

Only characters with Link_Term=Close have a Link_Bracket mapping.

See Bidi_Paired_Bracket_Type.

Review Note: For comparison to the related General_Category values, see the characters in:

- (Close_Punctuation + Final_Punctuation - BidiPairedBracketType=Close)
- (Initial_Punctuation + Open_Punctuation - BidiPairedBracketType=Open)

Link_Email

In the ASCII range, the characters are as specified for ASCII, as per RFC 5322, Section 3.2.3. That is:

- [[a-zA-Z][0-9][_ \- ! ? ' \{ \} * / \& # % ` \^ + = | ~ \$]]

Outside of the ASCII range, the characters follow UAX31 identifiers. That is:

- \p{XID_Continue}

The reasons for this are that non-ASCII in the `local-part` are less commonly supported at this point, and the `local-part`s supported on most mail servers that go beyond ASCII are likely to have restrictions similar to programming identifiers. Implementations could also customize the set, and it can be broadened in the future.

Review Note: We could have other exclusions to start with, such as only NFC characters; or only Identifier_Status=Allowed from UTS #39 Unicode Security Mechanisms?

7 Test Data

The following test files supply data for testing conformance to this specification. The format of each test is explained in the header of the test.

- LinkDetectionTest.txt
- LinkFormattingTest.txt

8 Security Considerations

The security considerations for Path, Query, and Fragment are far less important than for Domain names. See UTS #39: Unicode Security for more information about domain names.

Review Note: We could add something like the following:
Implementers may consider some additional restrictions on characters based on security considerations, such as using UTS #39 Unicode Security Mechanisms. For example, an implementation could test whether any characters in a detected link span had Identifier_Status=Restricted, and if so, not apply a link to that span. Note that simply forcing those characters to be percent-escaped in link formatting doesn't generally solve any problems; if anything, percent-escaping obfuscates characters even more than showing their regular appearance to users.

~~NOTE: the following seems misplaced; if anything, a longer discussion of this should be in UTS #39 Unicode Security Mechanisms. There are documented cases of how Format characters can be used to sneak malicious instructions into LLMs; see Invisible text that AI chatbots understand and humans can't?. URLs are just a small aspect of the larger problem of feeding *clean text* to LLMs, both in building them and in querying them: making sure the text does not have malformed encodings, is in a consistent Unicode Normalization Form (NFC), and so on.~~

For security implications of URLs in general, see UTS #39: Unicode Security Mechanisms. For related issues, see UTS #55 Unicode Source Code Handling. For display of BIDI URLs, see also HL4 in UAX #9, Unicode Bidirectional Algorithm.

9 Stability

As with other Unicode Properties, the algorithms and property derivations may be changed in successive versions to adapt to new information and feedback from developers and end users.

- Unassigned code points may change property values as they are assigned. All new characters will be extremely low-frequency.
- Characters that are assigned may, in rare cases, change values as more information about the character is determined.

The practical impact is very limited, such as when character is not escaped on a formatting system, but terminates the link on the detecting system.

10 Migration

An implementation may wish to just make minimal modifications to its use of existing URL link detection and formatting code. For example, it may use imported libraries for these services. The following provides some examples as to how that can be done.

Migration: Link Detection

The implementation may call its existing code library for link detection, but then post-process. Using such post-processing can retain the existing performance and feature characteristics of the code library, including the recognition of the Scheme and Host, and then refine the results for the Path, Query, and Fragment. The typical problem is that the code library terminates too early. For code libraries that 'mostly' handle non-ASCII characters this will be a fraction of the detected links.

1. Call the existing code library.

2. Let S be the start of the link in plain text as detected by the existing code library, and E be the offset at the end of that link.
3. If E is at the end of the string, or if the code point at E, that is, the character immediately after the offset at the end of the detected link, has the value Link_Term=Hard, then return S and E.
4. Scan backwards to find the last `initiator` of a Path, Query, or Fragment URL Part.
5. Follow the Termination Algorithm from that point on.

Migration: Link Formatting

The implementation calls its existing code library for the Scheme and Host. It then invokes code implementing the Minimal Escaping algorithm for the Path, Query, and Fragment.

References

Review Note: TBD, put the references into the standard format.

[RFC6530]
[URL Fragment Text Directives]
[WHATWG URL: 3. Hosts (domains and IP addresses)]
[WHATWG URL: 4.4. URL parsing]
[WhatWG URL: Example URL Components]
[WHATWG URL: Host representation]

Acknowledgments

Thanks to the following people for their contributions and/or feedback on this document: Arnt Gulbrandsen, Dennis Tan, Elika Etemad, Hayato Ito, Jules Bertholet, Markus Scherer, Mathias Bynens, Peter Constable, Robin Leroy, [TBD flesh out further]

Modifications

The following summarizes modifications from the previous revision of this document.

Draft 7

- Fixed Link_Email set in the ASCII range, now a positive set.
- Noted that the email linkification fails with illegal dot placement.

Draft 6

- Removed older change highlighting.
- Corrected TOC.
- The names for Link_Termination and Link_Paired_Opener have changed to Link_Term and Link_Bracket. This change is not marked in the text, for ease of reading.
- Added the property Link_Email and the property file LinkEmail.txt
- Moved Property Assignments to under Property Data, and made it clearer that it is the initial derivation, and can change later.
- Dropped quoted local-part formats from email detection and minimized format.
- Moved section 6 (Security) down to just above 9 (Stability).
- Added some clarifying text.
- Narrowed the characters in email local-parts (Link_Email) to align better with XID_Contimue.

~~**Draft 5**~~

- ~~In the title, changed Serialization to Formatting; added a subtitle.~~
- ~~Changed the data folder from `links` to `linkification`.~~
- ~~Expanded the Summary at the top, and clarified the Introduction.~~

- ~~Clarified the derivation of property values.~~
- ~~Fixed and simplified handling of fragment directives.~~
- ~~Fixed link detection algorithm bugs and changed to substring matching of separators.~~
- ~~Made openStack bounded.~~
- ~~Fixed link escaping issues: String matching, bounded openStack.~~
- ~~Fixed email address termination issues: In a quoted local-part, require an initial double quote and forbid escaping the final double quote.~~
- ~~More consistent use of “part”/“URL Part”.~~
- ~~Settled inconsistent “Protocol” vs. “Scheme” in favor of “Scheme”.~~
- ~~Various minor editorial changes, bug fixes, and typo fixes.~~

~~Draft 4~~

- ~~Fleshed out Table of Contents (not highlighted).~~
- ~~Rationalized the handling of fragment directives.~~
- ~~Removed old review notes and Review Issues section.~~
- ~~Fleshed out Email section, and added a corresponding conformance clause.~~
- ~~Added Stability and Migration sections.~~
- ~~Various copy-edits, only highlighted where material.~~

Modifications for previous versions are listed in those respective versions.
