L2/20-013

Proposal to ensure maximum visibility of changes to UAXes

For consideration by Unicode Technical Committee

2020-01-06
Marcel Schneider (charupdate@orange.fr)

We should always say what we see.
Above all we should always
—which is most difficult—
see what we see.

Charles Péguy

This proposal adds to the response to Action item 161-A1. It advises to ensure that the changes to UAX #14 suggested in *Proposal to make material changes to UAX #14* or in *Proposal suggesting formal edits to UAX #14,* as well as any updates to the Unicode Standard will effectively show up in specific web searches and won't be outnumbered by a host of old versions.

Problem

Actually, changes to UAXes are not visible enough. Due to an incomplete /robots.txt, superseded versions of most UAXes stay indexed by search engines. Some old versions are disallowed individually, so for UAX #34 up to revision 18 out of 25. For UAX #14, robots.txt contains one single line, disallowing revision 15 (version 4.0.1).

In order to make sure that outdated versions of UAX #14 are effectively ruled out and won't be visited inadvertently after a web search, changes to the robots instructions file of the Unicode website are suggested in subsection *4.1 File /robots.txt* below.

Additionally, the file /sitemap.xml, where all valid pages would be listed by their (latest-version) URL in <loc> (I'd specify the HTTPS protocol and omit the www), with <lastmod> set to the latest version release date, is also missing. Hence this proposal suggests creating the sitemap of unicode.org, without going into details however since the only on-topic point is to ensure the effectiveness of the proposed changes.

1 UAX #14 in /robots.txt

Change from:

```
Disallow: /reports/tr14/tr14-15.html
```

Change to:

```
disallow:/reports/dtr14-03.html
disallow:/reports/tr14-4/
disallow:/reports/tr14-5/
disallow:/reports/tr14/ # all other old versions
allow:/reports/tr14/index.html # latest version
```

Rationale:

The path "/reports/tr14/" catches all versions that have their revision number in the URL, except the first few, that have atypical URLs needing to be disallowed specifically. With "index.html" appended, the standard path points to the latest version. Else, all old revisions would need to be disallowed individually, at the risk of the list failing to be updated, as it happened.

2 Other UAXes in /robots.txt

The same scheme will help improve the visibility and thus the practical effectiveness of updates to all other UAXes. I'm running out of time and therefore cannot suggest rewriting the file right now, but hopefully the patch will expand to the Unicode website.

3 File /sitemap.xml

The sitemap of https://unicode.org does not yet exist, per the lack of the file "sitemap.xml" in the root directory. It would be useful for the purpose of optimizing the visibility of the latest version of any part of the Unicode Standard. Hence I'd suggest creating or autogenerating this file.

Acknowledgments

Thanks to everyone who directly or indirectly helped put this paper together.

Thanks to Microsoft for Word Online, OneDrive, VS Code and MSKLC.

Thanks to Google for Google Chrome, Google Search, Google Books, Google Translate and Gmail.