Intent to remove support for some invalid characters in hostnames

brettw@chromium.org

Tracking bug: 652808
Status entry: https://www.chromestatus.com/feature/6059320944885760

This is on hold because other browsers (at least Firefox and Edge) can make URL objects with garbage names like:
new URL("http://@#$@#$@#$/");
They do not seem to do any (or at least much) validity checking at this level. Chrome canonicalizes these which leads to exceptions in some cases, and different strings in others. Restricting the characters in the way proposed will make this happen a lot more for some common cases like "http://*" (which Chrome's settings itself uses).

Background

Chrome's URL library supports several classes of characters in the host name:

- Valid characters which are always unescaped even if we see a %XX escape code.
- Invalid characters per the spec that we allow as if they were valid.
- Invalid characters which we reject. You will not be able to load any URLs with these characters in the host.
- Characters which are allowed but which are percent escaped in the hostname: like%20so.com. These will be called "scary" for the purposes of this document.

The scary character handling violates the spec but was what IE6 on Windows XP did when we did the URL rules, and there were examples of real sites that used this.

Scary characters

Erikv did an analysis of ~5B links with host names on the web. 0.0011% had one of these scary characters in it. This shows counts for various characters.

- . : 54,959,752 (most popular valid host character)
- 9 : 4,997,607 (least popular valid host character)
- _ : 144,702 (technically invalid but allowed)
- + : 400 (technically invalid but allowed)
- [: 134 (technically invalid but allowed)
-] : 110 (technically invalid but allowed)

- <space> : 19,244 (most popular scary character)
- , : 4,388 (usually a typo for dot)
- } : 4,176
- { : 4,169
- * : 2,073
- ‘ : 499
- (: 425
- & : 396
-) : 361
- > : 234
- < : 229
- | : 179
- “ : 171
- = : 167
- $: 165
- ! : 159
- @ : 86
- ` : 56
- # : 1 (hard to make since it affects parsing, page would have had to escape it!)

Microsoft tightened up the DNS rules in Windows 7 so that these invalid characters are no longer allowed.

In looking at the list of host names, it’s obvious that almost all of them are typos. I checked some more plausible ones and almost all of them did not resolve. My impression is that most of these are links people mistyped on the blogs and the authoring software tried to make them look plausible.

One valid host name I found that will be affected is
http://information%20systems-unkris.eksport.web.id
This page is an Indonesian University page and I suspect most users have Windows XP which allows `%20` in host names. Docs will not allow clicking on this link even if you’re on a system where it will load. To test paste this in the omnibox and click the link:

```
data:text/html,<a href="http://information%20systems-unkris.eksport.web.id">Click</a>
```

Another valid link is:
http://reggio-nell'emilia.paginegialle.it
Which appears to be a wildcard DNS entry that interprets the part to the left of the ‘.’ as a query. The source of this link has the same link in other places without the apostrophe.

Proposal

Remove the special-case handling for scary characters and mark them as invalid. Mark '+' as invalid also (it's currently allowed in violation of the spec, and is used much less than even other completely disallowed characters).

Keep accepting _, [, and] as valid characters in violation of the spec. Underscore is very popular and works across browsers. Square brackets are used for IPv6 addresses and we don't differentiate between those and other host names at the current level of the code.

Expected impact

Places where this already fails:

- All browsers on Windows 7+
- Firefox
- Safari
- Android
- Chrome typing in the omnibox (you must follow links instead)

Places these URLs do currently work:

- IE6 on Windows XP
- Chrome links on Linux, ChromeOS, and Mac (these users will see a difference in Chrome if they encounter these URLs).
- Google search will show these links [example] which will work on the above configs