Case mapping on Unicode is hard

devblogs.microsoft.com/oldnewthing/20030905-00

September 5, 2003

Raymond Chen

Occasionally, I'm asked, "I have to identify strings that are identical, case-insensitively. How do I do it?"

The answer is, "Well, it depends. Whose case-mapping rules do you want to use?"

Sometimes the reply is, "I want this to be language-independent."

Now you have a real problem.

Every locale has its own case-mapping rules. Many of them are in conflict with the rules for other locales. For example, which of the the following pairs of words compare case-insensitive equal?

1.	gif	GIF
2.	Maße	MASSE
3.	Maße	Masse
4.	même	MEME

Answers:

1. no in Turkey, yes in US
2. no in US, yes in Germany
3. no in US, no in Germany, yes in Switzerland! (Though you would likely never see it written as "Maße" in Switzerland.)
4. yes in France, no in Quebec!

(And I've heard that the capitalization rules for German are context-sensitive. Maybe that changed with the most recent spelling reform.) Unicode Technical Report #21 has more examples.

Just because you’re using Unicode doesn’t mean that all your language problems are solved. Indeed, the ability to represent characters in nearly all of the world’s languages means that you have more things to worry about, not less.

Raymond Chen

Follow