

Language Matters: Cyrillic search

Let's suppose you are interested in ¹Ukrainian politics and for example want to find out more about the visit of President Victor Yuschenko to Poland. You would go to your favourite Web search engine and enter the words:

Official visit Victor Yuschenko Poland

When this query was tried on ²Google with preferences set to 'any language' it gave 5060 results - all in English.

Ideally of course you would want to get a wider perspective than from an Anglo-Saxon viewpoint - the same search query expressed in Ukrainian is:

Офіційний візит Віктора Ющенка Польщі.

This query gave 471 results (8 pages of 30 results actually viewable)

Perhaps if you were not able to get access to a Ukrainian keyboard, you might have tried a Russian keyboard and entered:

Официальный визит Виктора Ющенко Польши

This query gave 7 results, none of them the same as the previous search!

Why can that be you wonder? The explanation is that the webmasters of those sites also probably used a Russian keyboard; unfortunately a Russian keyboard does not contain a key for the letter 'i', so instead the webmaster has substituted an 'English i'. The difference is that the Ukrainian 'i' has a Unicode encoding of U+0456, whereas the English 'i' is encoded as U+0069.

Unfortunately at the time of writing Google and many other search engines do not take this into account, as a result you or your organisation could miss vital information.

dtSearch version 7.00 includes a mapping from the Cyrillic 'i' to the Latin 'i' and thus if you had searched on web pages spidered by dtSearch you would have found all the web pages, irrespective of whether you had used a Ukrainian keyboard or had used a Russian keyboard and made the error of substituting the Latin 'i'. It is this depth of experience that distinguishes dtSearch from many of the newer entrants to the world of search technology.

Consider this:

Not all search tools support Unicode search queries, those that do may not always return the results you expect. What does your current search tool offer?

¹ Although this article highlights a particular Ukrainian search query problem, the same problem exists with several other languages that use a Cyrillic based alphabet such as Belarusian and Kazakh. There are over 50 languages that use the Cyrillic alphabet.

² All the Internet search results were obtained on www.google.com on 12 May 2005, results at other times and on other web search engines may differ.