CEN/TC304 N967 SC22/WG20 N821 L2/01-105

Subject/Title: Matching - Disposition of comments to the pre-draft 1.1

Source: Marc Küster

Date: 19 February 2001

Note: The Matching draft in N962 was produced after taking account of the comments in this document.

Disposition of Comments for the pre-draft 1.1

Marc Wilhelm Küster, Project manager and main editor

Comments from Ken Whistler via ISO/IEC JTC1/SC22/WG20

Misc. typos and small English usage issues

Scope, 8th paragraph.
»explicitely« → »explicitly«
»defined as« → »defined than«
Overview, page 5, 4th paragraph, first bullet
»of the technical report« — the referent is unclear here; you should spell the name out completely as in the second bullet, or pick an abbreviation for the reference that you then use consistently.
Overview, page 5, 4th paragraph, 3rd bullet.
»taxed at« → »estimated as« This usage occurs several other places below. Search for »tax(ed)« and change these to »estimate(d)«.
Also, in The Holy Grail, page 10: »is not taxable« → »cannot be estimated«. (I presume this is a mistranslation of »schätzen«; in this context, it would mean »estimate« or »assess«, but not »tax«.)

All of these points are accepted.

Scope, last paragraph

I take issue with characterizing the two dimensions as »temporal« and »spatial«. What you are talking about is the distinction between historical (diachronic) varieties of a given language and coexisting, contemporary (synchronic) varieties of languages. And in the latter case, the varieties could be dialects within a language, related languages, or unrelated languages. For the purposes of this discussion, »spatial« doesn't have anything to do with the latter, since, in the extreme case, synchronic varieties can exist at the same time, at the same place, in the *same speaker* (or written in the same multilingual document).

Also, »access to texts written in other cultures than the user's own« is a misleading characterization — since what you mean is »written in other *languages*«. It is perfectly possible to access texts written in other cultures *in the same language* as that of a particular speaker. Here, language and culture are somewhat independent variables. And a concern for sensitivity about preservation of »cultural heritage« should not be an excuse for fuzzy usage of the terms »culture« and »language« in this document.

Accepted. Will be amended

Overview. A brief look at history

1st paragraph. I wouldn't characterize Soundex as an »advanced« search strategy. Your footnote may well be accurate, but as phonological search systems go, Soundex is pretty simplistic and primitive — and comes up with lots of bizarre results because it is not dictionary based.

Accepted. Some of this is implicit in the »Phonetically aware matching« section but should be made more explicit already in the overview.

5th paragraph. 10 Gigabytes would not be considered a very »large« database by today's commercial standards. Insurance companies and consumer packaged goods companies, for example, routinely work with terabyte databases now. (Each of those scanner »beeps« at the retail point of sale (POS) stations is turned into an Insert statement in some database somewhere!)

Accepted in principle. I would still maintain that 10GB is a »large« database in many fields of application (though not a »very large« one). I'll make this more explicit.

In general

The text of this document is too heavily footnoted. This is just a style issue, I suppose, but the over-attention to documenting the details and implications of minute claims is detracting from an overall clear rhetorical structure that defines the problem(s) and articulates clear direction for accomplishing something that will have an impact on how the technological developments in this area affect Europe (positively or negatively).

Accepted. This is the German academic illness. The problem will be alleviated by creating a less footnoted »executive version« of the final report.

Overview: matching, p. 5, 4th paragraph, 2nd bullet

It is not clear to me at all why this extension of guiding principles should be such a gigantic task. Conceptually, as for all complex algorithmic text processes, the cleanest way forward is to conceive the steps in terms of the universal character encoding — abstracted away from all issues of encoding forms and particular character sets. Then the encoding form issues can all be encapsulated in relatively well-understood and widely implemented character set conversion technology. (This is like the approach of defining HTML and XML behavior in terms of Unicode as the reference character set.)

And then the other, more complex issue of representational »downshifting« when converting to incomplete character repertoires can also be treated as a modular set of issues: fallbacks within scripts; transliterations between scripts. This way a requested transliteration can be treated as just a special case of the general problem of script conversion when a target repertoire does not contain the script in question, for example.

Approached this way, there might be some hope of getting universal search engines to have adaptable plug-in interfaces (either explicitly specified, implicitly based on various cues, or heuristic) that would start to approach the level of sophistication that this scoping statement seems to be hinting at.

Accepted. I wonder, however, how to assess the time requirements for this modular approach (which is, indeed, the only sensible one). It will still be a major task to evaluate potential problems that occur because of the downshifting.

Completeness of information, page 6, 3rd paragraph

»Recurrent reports on this topic should give an incentive to industry to support European requirements.« Huh? Recurrent reports are not an incentive to do anything. This needs to be thought through further to try to find some real impetus to do something. What *would* the incentive be for any portal or search engine company to take some report recommendations into account? Why would they be driven by that, more than by the internal pressures of their marketing people responding to customers and their perception of the competition and of marketing opportunities?

Not accepted. Reports can play an important rôle in heightening customer awareness.

Overview: browsing

This entire section seems weaker to me. It is not clear what exactly is being proposed for »consistent use of cataloguing strategies«. You need a clearer statement of what the problem is with existing information organizing sites like portals, with emphasis on the multilingual and multicultural requirements of Europe. And then what would be the goal of a TC304 report on this that might have some hope of impacting what anyone is actually doing in this arena?

Accepted. This section needs some reworking. I'll take your excellent suggestions about the »strategy of exemplification« on board.

Also you might think about a strategy of exemplification in this area. Pick a monolingual portal site that poses problems for European access. And contrast that with some more exemplary site that organizes itself well for multilingual access. (Are there good examples to point to?)

Accepted.

Comments from Mats Lander, Sweden

The questions

1. P. 5, last para before section »Trends of today«, beginning »In conjunction with...« — I just don't understand the expression »the relative availability of data in various encoding schemes«.

What the text wanted to say is »the relative amounts of data which are available in various encoding schemes«. The text is going to be clarified.

2. P. 7, last para before section »Phonetically aware matching« — Does the wording »The project team sitll intends to do some scoping on this field of action« mean that a more detailed description of scope will be provided in the final report?

Yes, the final report is going to be somewhat more elaborate on this issue.

3. P. 10, European requirements — What are the lists that you are talking about here?

What is intended here is a portal site of portal sites. The phrasing will be suitably modified.

The comments

1. A list of contents would help getting an overview of the report.

Accepted. Such a list will be added to the final version.

2. P. 5, top: Is is not clear what role this (in itself very logical) dimensional classification plays in the rest of the report.

The »temporal / spatial« dichtinomy is going to be expanded upon / changed in formulation (cf. also the comments by Ken Whistler).

3. P. 5: The heading »Trends of today: Search engines« implies the existence of another section »Trends of today: Something else«, which however does not exist. Maybe the heading should be reworded to something like »The multilingual approach in today's search engines«.

Accepted. The heading has been changed accordingly.

4. P. 9: The sub-section »Summary« obviously refers only to the preceding sub-section but is presented on the same hierarchical level, which is slightly misleading. Perhaps this single paragraph could simply be included in the Background information sub-section, possibly preceded by the words In sum:«.

Accepted.

PS. I wonder about the term »scoping« which is more or less new to me and sounds sort of truncated. If the meaning is »defining the scope«, I would propose to use that instead.

This title is the title of the project as proposed by the Commission. While I agree with this comment in principle, it may be less confusing for the Commission to leave the title as it stands. However, I'll gladly follow any guidance that CEN/TC304 may be willing to give.