Εισήγηση προς την ΤΕ 48 για το ISO/IEC DIS 29500

Αντώνης Χριστοφίδης, 19 Μαρτίου 2008

Εισηγούμαι στην Τεχνική Επιτροπή 48 του ΕΛΟΤ να ψηφίσει όχι στο ISO/IEC DIS 29500 (OOXML), για δύο λόγους:

1. Το κείμενο είναι προχειρογραμμένο και έχει χιλιάδες ασάφειες, που το καθιστούν εντελώς ακατάλληλο για πρότυπο.
2. Το ISO/IEC DIS 29500 είναι ανταγωνιστικό με το ISO/IEC 26300, και η ύπαρξη πολλαπλών προτύπων για το ίδιο πράγμα αναιρεί το σκοπό ύπαρξης προτύπων.

1. Προχειρογραμμένο κείμενο με χιλιάδες ασάφειες

Στη διάρκεια της εργασίας μου για τις ημερομηνίες[1] χρειάστηκε να μελετήσω προσεκτικά 30 περίπου σελίδες από τις 6 χιλιάδες που ήταν η αρχική πρόταση, και ανακάλυψα περίπου 13 τεχνικά λάθη που δεν είχαν αναφερθεί προηγουμένως. Τα λάθη συνοψίζονται στον Πίνακα 1 και παρατίθενται στο Παράρτημα Α.

Πίνακας 1: *Προβλήματα που ανακάλυψα μελετώντας τις ημερομηνίες*

Ελάσσονα προβλήματα	7
Μείζονα προβλήματα	5
Κρίσιμα προβλήματα	1
Σύνολο	13

Ελάσσονα προβλήματα είναι αυτά όπου το νόημα μπορεί να γίνει κατανοητό από το συγκείμενο. Μείζονα είναι αυτά όπου η κανονιστική (normative) περιγραφή είναι ανεπαρκής ή λανθασμένη ή αντιφάσκει, αλλά όπου το νόημα μπορεί να εξαχθεί, με κάποια βεβαιότητα, από το συγκείμενο και το πληροφοριακό (informative) κείμενο. Κρίσιμα είναι αυτά όπου κανένα συμπέρασμα δεν μπορεί να βγει για το νόημα.

Επειδή δεν υπάρχει κάποιος λόγος να θεωρήσουμε ότι στις υπόλοιπες 6 χιλιάδες σελίδες η πυκνότητα των λαθών είναι μεγαλύτερη ή μικρότερη από αυτήν στις παραπάνω σελίδες, προκύπτει, ως πρώτη προσέγγιση, ότι το κείμενο πρέπει ακόμη να περιέχει περίπου 2500 λάθη, από τα οποία περίπου τα 1000 είναι μείζονα ή κρίσιμα. Αυτά τα λάθη είναι επιπλέον των σχολίων που υποβλήθηκαν από τις χώρες το Σεπτέμβριο 2007, και επιπλέον των περίπου 1000 τροποποιήσεων της Ecma (που με τη σειρά τους έχουν λάθη).

Αυτή η εκτίμησή μου για τα λάθη είναι αισιόδοξη. Όπως φαίνεται στο Παράρτημα Α, σχετιζόμενα λάθη που βρίσκονται κοντά το ένα στο άλλο τα έχω μετρήσει ως ένα λάθος. Πολύ εύκολα μπορούν τα 13 να θεωρηθούν 20. Ο Rob Weir[2] χρησιμοποίησε γεννήτρια τυχαίων αριθμών για να επιλέξει 25 τυχαίες σελίδες, τις οποίες διάβασε προσεκτικά και βρήκε 64 τεχνικά λάθη, από τα οποία κανένα δεν είχε αναφερθεί ως τώρα. Αν η εκτίμηση του Rob Weir είναι σωστή, το κείμενο έχει ακόμα 15 χιλιάδες τεχνικά λάθη.

Η διαδικασία fast track, που ακολουθείται για το ISO DIS 29500, υπάρχει για τη γρήγορη έγκριση ώριμων προτύπων. Με πολλές χιλιάδες λάθη, το ISO DIS 29500 δεν είναι ώριμο και είναι εντελώς ακατάλληλο για τη διαδικασία fast track. Το Ballot Resolution Meeting μπόρεσε μόνο πολύ βιαστικά και επιφανειακά να εξετάσει μερικά από τα χιλιάδες ζητήματα (για περισσότερες πληροφορίες, βλ. την προσωπική μου μαρτυρία από το BRM στο Παράρτημα Β).

1 http://elot.ece.ntua.gr/te48/ooxml/disposition-on-dates
2 http://www.robweir.com/blog/2008/03/how-many-defects-remain-in-ooxml.html

2. Ανταγωνιστικότητα με το ISO/IEC 26300

Στο ερώτημα γιατί χρειαζόμαστε άλλο ένα πρότυπο όταν ήδη υπάρχει το ODF (ISO/IEC 26300), έχουμε πάρει δύο απαντήσεις: 1) Ότι το OOXML έχει σκοπό την προς τα πίσω συμβατότητα· 2) ότι περισσότερα πρότυπα σημαίνει περισσότερη επιλογή.

Το πρώτο από αυτά τα δύο σημεία δεν βγάζει νόημα. Ουδαμώς το OOXML βοηθάει να διαβάσουμε τα παλιά αρχεία του MS Office. Όσο για το ότι μπορεί να αναπαραστήσει επακριβώς την πληροφορία τους, το OpenOffice μπορεί να μετατρέπει αξιόπιστα σε ODF πάνω από το 90% των αρχείων MS Office που λαμβάνω με email, ενώ για όσα δεν μετατρέπει αξιόπιστα το πρόβλημα φαίνεται να είναι στο OpenOffice, όχι στο ODF. Δεν έχουμε δει συγκεκριμένα παραδείγματα πληροφορίας που δεν μπορεί να αναπαρασταθεί στο ODF, ούτε έχουμε δει κάποιο λόγο γιατί δεν θα μπορούσαν αυτά το υποτιθέμενα προβλήματα να λυθούν προσθέτοντας δυνατότητες στο ODF.

Το δεύτερο σημείο δείχνει ανυπαρξία κατανόησης του σκοπού της τυποποίησης. Τα πρότυπα υπάρχουν ακριβώς για να περιορίζεται η επιλογή. Αν έβγαινε ένα πρότυπο για συγκεκριμένο βύσμα φορτιστή σε κινητά τηλέφωνα, αλλά μια από τις εταιρείες, αντί να ακολουθήσει το πρότυπο, προσπαθούσε να τυποποιήσει το δικό της βύσμα ως δεύτερο πρότυπο, δεν θα πίστευε κανένας ότι το κάνει για να προσφέρει περισσότερη επιλογή, αλλά ότι θέλει να σαμποτάρει το πρώτο πρότυπο.

Παράρτημα Α: Πρόσφατα ανακαλυφθέντα λάθη στο Ecma 376

Τα παρακάτω λάθη ανακαλύφθηκαν στη διάρκεια της εργασίας μου για της ημερομηνίες.

Part 4, §3.17.7.104, page 2626:

The start-date parameter is frequently conveyed as date-string or as start-value.

Part 4, §3.17.7.104, page 2626:

The EDATE function description does not adequately describe what happens when the resulting month has insufficient days; for example, what is returned from EDATE(DATE(2000, 01, 31), 1)? This is apparent from the example only.

Part 4, §3.17.7, page 2530:

Several references made in the table are wrong. For example, EDATE is not §3.17.7.103, it's §3.17.7.104.

Part 4, §3.17.7.75, page 2,601:

In the DATEVALUE, there are two errors. One, it is ambiguous whether times are taken into account. Two, the first item in the example might have different meaning depending on operating system settings or other user settings.

Part 4, §3.17.7.78, page 2,605:

Two errors. One, the "and/or time" is not correct, since any time information is ignored. Two, DAY(DATE(2006, 0, 2)) does not return 31, as the example claims, but 2.

Part 4, §3.17.7.104, page 2626:

The phrases "future date" and "past date" actually mean a date after and before start-date. The same problem is in EOMONTH.

Part 4, §3.17.7.143, page 2658:

The argument to the function is listed as "number". However, in the description, it appears that it can be either a number or a string. Same problem in the MINUTE function and in the SECOND function.

Part 4, §3.17.7.143, page 2658:

The return value for this function is in the range 0-23, not 0-59.

Part 4, §3.17.7.218, page 2717:

The return value for this function is in the range 1-12, not 1900-9999.

Part 4, §3.17.7.224, page 2722:

The holidays argument, according to the description, can be an array constant of the serial values of the dates. However, in the example, it is an array of strings.

Part 4, §3.17.7.321, page 2,803:

According to the argument description and the example, the date information is ignored. According to the function description and the return type and value, it is not ignored.

Part 4, §3.18.13, page 2,842:

Several errors in timePeriod. First, 7/14/2006 must be written as 2006-07-14. Second, 38913 is 2006-07-15, not 2006-07-14. Third, it is not specified that this example applies only to the 1900 date base.

Part 4, §3.17.7.75, page 2,601:

The description for DATEDIF does not provide enough information. For example, what is the result of the following?

```
DATEDIF(DATE(2001,6,15),DATE(2002,9,10),"MD")
DATEDIF(DATE(2004,2,15),DATE(2005,3,10),"MD")
```

Παράρτημα Β: Πώς ήταν το BRM

I have been one of the two Greek delegates to the OOXML Ballot Resolution Meeting (BRM). Lots of things have already been written about the meeting. I will not repeat them here, but will only make a few clarifications on things that I think are not well understood. If you want a general account, I think that Tim Bray's[3] is the best so far.

The opinion of the BRM

First, it is important to clarify that the BRM did not say either that the specification is OK, or that it is not OK, because it is not within its competence to say such a thing. There was good co-operation, and, to a large extent, good will, from all sides, because the BRM has a single purpose: make the specification better. Let me repeat that: the BRM has the single purpose of making the specification better, so that if it is accepted, it is the best possible. Or, the same thing viewed from another viewpoint, the BRM attempts to make it better, to maximize its chances of going through. Well, this is in theory, of course; in practice some want to maximize its chances of not going through, and I'm certain that it's not the first standard in which this happens.

The BRM, therefore, has made no statement that the fast-track process was appropriate, nor that it was inappropriate. No claim that there was enough time to address the issues, nor that there was not. No recommendation to approve the outcome, or to not approve the outcome. The BRM only provides an outcome, implying nothing more than the fact that these are the improvements that we were able to make within the given time frame.

You don't need to take my word for that: you can read ISO's description of the objective of the BRM[4] and Alex Brown's clarification on this issue[5] (Alex Brown was the chairman of the meeting).

The paper ballot

Lots of things have been written about the paper ballot. Andy's blog entry[6], which has been the centre of attention, was one of the posts that discusses the paper ballot very much. I told Andy that, in my opinion, he overinterprets the paper ballot, which is not such a significant issue. Let me explain from the start, however, because it is quite complicated.

After the national bodies submitted their comments in September 2007, Ecma studied and proposed changes to the original six-thousand-page proposal in order to address these comments. Ecma's proposed changes, called Disposition of Comments, was submitted to the national bodies on 14 January 2008. It is a 2300-page document containing about one thousand proposed changes. However, Ecma does not have the power to make changes to ISO DIS 29500; only the BRM has such power. Therefore, Ecma's changes would not make it in unless approved by the BRM.

It became clear soon enough that there was no time to discuss all one thousand of Ecma's proposed changes, but at the same time it was thought that, if they constitute improvements to the original document, it would be a pity to disapprove them for lack of time. Therefore, the BRM attempted to find a solution to the problem. Several options were discussed, and the option that gained most support was for each national body to fill in a form like the following:

3 http://www.tbray.org/ongoing/When/200x/2008/02/29/BRM-narrative
4 http://www.iso.org/iso/pressrelease.htm?refid=Ref1114
5 http://www.consortiuminfo.org/standardsblog/comment.php?mode=view&cid=18803
6 http://www.consortiuminfo.org/standardsblog/article.php?story=20080229055319727

```
For all ECMA dispositions which have not been explicitly accepted or rejected
in the meeting, we want to
    Approve [ ]    Disapprove [ ]    Abstain [ ]
Except for the following:
    Response ...    Approve [ ]    Disapprove [ ]    Abstain [ ]
    Response ...    Approve [ ]    Disapprove [ ]    Abstain [ ]
    Response ...    Approve [ ]    Disapprove [ ]    Abstain [ ]
    Response ...    Approve [ ]    Disapprove [ ]    Abstain [ ]
    Response ...    Approve [ ]    Disapprove [ ]    Abstain [ ]
```

Initially it was thought that, e.g. 1 vote in favour for a specific response, and 31 abstentions, should probably mean that it is somewhat extreme to consider as in favour. But later it was pointed out that if, for example, a specific response affects one country only, that country's opinion should weigh heavily. Examples that were discussed include right-to-left writing, which none in the room understood besides Israel; and some measurement units, which few people besides UK and Australia had a grasp on.

After considering several options, such as distinguishing between "abstain" and "no position", which all proved to be infeasible, it was decided, on Wednesday afternoon, with an overwhelming majority of 29 votes in favour, that the paper ballot would be conducted, and that a simple majority of Approvals, without counting Abstentions, would mean approval. The paper ballot affected only Ecma Responses that had not explicitly been addressed in other ways during the meeting.

So, from the 80 or so responses that Greece had studied, we specified Approve on something like 70 of them, Abstain on the others, and Abstain on the 900 or so others that we had not studied. I believe that most countries voted similarly to Greece.

(The "default" option, besides Approve, Disapprove and Abstain also had "we do not wish to record any position", which for practical purposes is the same as Abstain.)

The way the paper ballot was conducted, especially that 1 approval and 31 abstentions counted as approval, was an issue of course, but national bodies were aware of it. There was no solution that did not have problems, and the national bodies chose to follow this one. Canada attempted to work around the problem by posting a short list of responses which they considered worse than the original text, encouraging countries to vote "no" to those instead of "abstain". Greece, however, did not manage to study Canada's list.

The fact that 98% of Ecma responses were adopted means that the BRM believes that they improve the original text. It does not mean that the BRM believes that the resulting text is OK. My opinion, for example, and many delegates agree with me, is that the Ecma responses make the text slightly better, but though slightly better it is still abysmal. But since they were an improvement, we adopted them.

The success or failure of the BRM

Brian Jones[7] and Jason Matusow[8] of Microsoft have said that the BRM was a success because it fulfilled its purpose, which was to make changes to the text. Although this is technically correct, if the original text got 1 out of 10 and the BRM managed to improve it to 1.1, it is somewhat misleading to call it a success. Brian Jones says that there was consensus in the changes. This is also true, but the reason there was consensus was that we quickly became disillusioned, lowered our standards, and only discussed modifications which we knew could pass within the given constraints. Let me give an example.

7 http://blogs.msdn.com/brian_jones/archive/2008/02/29/brm-is-done-time-to-sleep.aspx

8 http://blogs.msdn.com/jasonmatusow/archive/2008/02/29/the-open-xml-ballot-resolution-meeting-brm-was-an-unqualified-success.aspx

One of the issues brought up by the Greek delegation was that of Office Open Math ML (OOMML). Given that the purpose of OOXML is to faithfully represent the existing corpus of legacy documents, and OOMML is an entirely new language that has nothing to do with such backwards compatibility, we were at a loss as to why it was designed from scratch rather than being based on the existing standard of MathML. Ecma's reply was that OOMML has features that MathML does not.

It took almost an entire week of discussions in the corridors and in emails to sufficiently clarify the issue. During our investigation we learned that Office 2003 documents converted to OOMML by Office 2007 retain their equation fields as equation fields, and equation editor objects as embedded binary objects; you can use OOMML only in new equations. After some discussion with Rick Jelliffe, Australian delegate, on the extensibility of MathML, we figured out that you can extend MathML, although the resulting extended language will not be MathML anymore. There were reservations by Canada because extending MathML might break existing MathML accessibility tools, and there was the argument, on my side, that it is easier to fix the accessibility tools for an extended MathML than to fix them for an entirely new language.
The purpose of the BRM was not just to discuss things in theory, but also to propose specific resolutions. What resolution could be proposed on this issue? "Redesign OOMML on top of MathML" was not realistic, not only during the BRM, but also for the time left for Ecma to make the final changes in March. Therefore, the only realistic resolution could be "drop OOMML" entirely, since the old ways of writing equations (equation fields and equation editor) are unaffected, and MathML had also been added (although without extensions). Would that, however, gain enough support?

While the chairman was reviewing remaining issues on late Friday, he asked Greece about it. Greece replied that we are still uncomfortable about OOMML, but that there is no time to resolve the issue. Therefore, we removed it from the agenda. The time was 15:45.

Another issue that came in late was that ISO 29500 should be backwards compatible with Ecma 376. I first heard this by Brian Jones on Thursday, during lunch, and on Friday morning it was made clear to the BRM that this was an issue. This, of course, was an entirely new argument, which affected everything. The BRM simply did not have time to discuss whether we want this compatibility or not. It was around 16:30 on Friday when the BRM decided to add the clarification "Microsoft Office 97 to 2008" to the Scope, where it says about the existing corpus of MS Office documents. Now although the "97" was undisputable, it was not clear that the BRM wanted "2008" rather than "2003". It was clear to everyone, however, that there was absolutely no time to discuss about that, so when Greece proposed the "97 to 2007" it went to vote after minimal discussion and only the technical correction of replacing 2007 with 2008 (the Mac OS X version of MS Office 2007).

Incidentally, while researching the equations, an issue of contention was the phrase "MathML renderers are allowed, but not required, to accept non-standard elements and attributes," in the MathML specification. This did not make much sense, and it took considerable effort and several email exchanges, before David Carlisle, MathML editor, clarified in an email to Rick Jelliffe that this phrase appears to be an error (although the rest of the section (2.4.2 in MathML 3.0,[9] 7.3.2 in MathML 2.0[10]) appears to be ok). The reason I'm mentioning this is that it shows how a slight carelessness in a minor detail of a standard can create trouble, and all of us who have used standards know that all too well, because no standard is perfect.

The contrast with OOXML is sharp, and this brings us to another issue of contention. The Greek workgroup on OOXML had been handed only the Ecma Responses for Greece. It was at the BRM when we found out that we should have studied all responses, not only those for Greece. It is not clear if this is an error by Ecma or by the Greek NB, but, in both cases, we did not have

9 http://www.w3.org/TR/MathML3/chapter2.html#id.2.4.2
10 http://www.w3.org/TR/MathML2/chapter7.html#id.7.3.2

the time to study one thousand responses, so there would have been no difference. In fact, even the 80 responses that Greece studied, we did not study at the level of scrutiny that is required when you inspect a standard. There was no time for that. What we did was glance through, and make fast decisions based on what seems right at a quick glance.

Dates

The date issue highlights several other problems. The original text proposes to store and manipulate dates by using two different representations, i.e. number of days from either 1900 or 1904, which contain a deliberate error for backwards compatibility: 1900 is considered to be a leap year. Other problems that were pointed out were that specifying dates before 1900 is not possible, that there is no reason to have two representations when one is enough, and that new ways of storing dates should not be invented since we already have ISO 8601. In the Disposition of Comments, Ecma proposed to add two additional ways of representation, which solved some of the problems but complicated the problem too much.

Before the BRM, I had prepared an alternative proposal on dates,[11] which is much cleaner. There are three issues that demonstrate how the whole process was affected by the time constraints.

The first issue is that my proposal for dates, despite the fact that I had written it with care in the course of a few weeks, and that it had been reviewed by several people, did have shortcomings. By comparing my work to the work by Ecma and Germany I found out that I had failed to notice that data types were described in two places, 3.17.2.6 and 3.18.12, and I had only noticed 3.17.2.6. And while thinking about the whole issue I saw that the way I had proposed for handling durations probably did not work (although now I'm having second thoughts - but I'll need very careful studying of ISO 8601 to arrive at a good conclusion). Both problems were easy to fix, and I submitted an updated proposal on Friday morning. However, this shows that it was very hard to write error-free proposals when the magnitude of changes were so large.

The second issue is related, but more important. When I discussed my proposal with Brian Jones on Thursday morning, he pointed out that it would be difficult for Ecma to accept it, because they did not have the time to verify that it actually works in all cases. Now this was a very valid concern. My proposal was more than 30 pages. Even if it were well thought and error free, Ecma had no way of knowing that. Therefore, the BRM was essentially confined to making changes that only scratched the surface of the problems.

The third issue is that, while writing my proposal, I and my reviewers found 13 additional errors in the original specification. However, national bodies were not allowed to submit new comments (and rightly so, otherwise there would have been total chaos). Therefore, there was no way to submit and correct them.

The changes that actually passed in Friday morning did add the possibility to store dates in ISO 8601 format, but they also keep the old ways, and in addition they add all ways proposed in the Disposition of Comments. Therefore we now have five different ways of representing the same thing.

11 https://elot.ece.ntua.gr/te48/ooxml/disposition-on-dates