

X_YLaTeX e il formato PDF/A

Presentato da Luigi Scarso

Claudio Beccari

LaTeX

LaTeX meeting 2011

Per seri motivi personali sono all'estero.

Non ho mancato una seduta del G_UT meeting da quando partecipo al G_UT; ma questa volta non sono riuscito a rendere compatibile la data del meeting con le date dei miei frequenti viaggi all'estero.

Vi chiedo scusa.

Nello stesso tempo voglio ringraziare pubblicamente l'amico Luigi Scarso che si è prestato gentilmente a presentare il mio contributo al meeting; sono particolarmente contento che egli abbia accettato perché la sua competenza nel problema dell'archiviazione dei documenti è molto superiore alla mia; se non fosse concentrato su Lua_TE_X, questo articolo l'avrebbe certamente scritto lui.

Claudio Beccari

In precedenti contributi ad \LaTeX sia Luigi sia io abbiamo trattato del formato PDF/A e dei due sottoformati PDF/A-1a e PDF/A-1b.

In questo anno 2011 è stata presentata all'International Standards Organisation (ISO) una nuova proposta di standard PDF/A, ma che io sappia non è ancora stata approvata definitivamente.

Vale la pena di ricordare alcuni concetti.

Lo standard ISO19005:1-2005 (prossimamente ISO19005:2-2011) stabilisce le norme per l'archiviazione di lunga durata dei documenti.

Questi devono essere composti nel formato PDF archiviabile (PDF/A) che è composto con il linguaggio PDF e devono sottostare almeno a queste ulteriori specifiche:

- la versione del linguaggio PDF deve essere la **1.4** (prossimamente potrà essere anche 1.7);
- le figure devono essere in formato **RGB** e deve essere specificato con precisione il **profilo di colore**;
- i font devono essere assolutamente vettoriali (**Type 1, TrueType, OpenType**) e almeno i glifi usati devono essere inclusi nel file; i singoli glifi devono necessariamente avere una larghezza **non nulla**;
- il file deve contenere un certo insieme minimo di **metadata** ben specificati.

Questo è sostanzialmente il sottoformato **1b**; per il sottoformato **1a** valgono ulteriori restrizioni.

Se si compone con pdfL^AT_EX si dispone del pacchetto *pdfx*, specificando al quale l'opzione *a-1b*, si può ottenere un documento spesso compatibile con lo standard *1b*. Occorre solo accompagnare il documento con un file contenente i *metadata* scritto in una forma molto particolare.

L'uso di pdfL^AT_EX permette anche di conservare i collegamenti ipertestuali interni; permette di comporre con la *microgiustificazione* fornita dal pacchetto *microtype*.

Ma pdfL^AT_EX non può lavorare nativamente con font le cui polizze contengano più di 256 glifi; quindi, senza ricorrere a trucchi vari, come quello di usare i font virtuali, con pdfL^AT_EX non si possono usare i font TrueType né i font OpenType.

Ecco perché potrebbe essere necessario usare X_YL^AT_EX (oppure LuaL^AT_EX).

Il pregio maggiore di X_YL^AT_EX è quello di poter usare i font OpenType e TrueType in modo nativo, anche quelli non distribuiti con il sistema T_EX, ma presenti nella dotazione del proprio calcolatore.

Disponendo dei font giusti, non è difficile comporre con alfabeti diversi da quello “latino”: greco, cirillico, ebraico, arabo, cinese, giapponese, coreano, devanagari, IPA, eccetera. L’unico neo che rimane, per chi ha necessità di comporre testi con questi insiemi di caratteri, è eventualmente la tastiera; ma c’è modo di provvedere.

Siccome i font OpenType e TrueType sono ammessi nello standard PDF/A sotto questo aspetto X_YL^AT_EX non dovrebbe dare problemi.

In relazione allo standard PDF/A, però, X_ƎƎ_ƎX presenta alcuni difetti:

- il programma di composizione di X_ƎƎ_ƎX è **xetex**, non **pdftex**, quindi non può fare uso di un certo numero di comandi specifici del linguaggio PDF che sono stati implementati nel programma di composizione **pdftex**;
- infatti **xetex** produce la sua uscita nel formato DVI esteso (XDV) e poi invoca il programma **xdvipdfmx** che lo converte nel formato PDF. Per questo motivo non si può usare il pacchetto *pdfx*; perciò si perdono diverse funzionalità che si possono avere solo usando il programma di composizione **pdftex**, in particolare non si può produrre direttamente un file PDF **sperando che sia compatibile con lo standard PDF/A**.

Si sarà notato che ho detto sia con pdf \LaTeX sia con Xe \LaTeX che il file prodotto **potrebbe essere compatibile con lo standard PDF/A**; non è detto che il file prodotto sia certamente compatibile con quello standard.

Questa certezza non c'è perché esistono diversi problemi:

- ci sono i problemi con i colori e i profili di colore;
- altri problemi nascono con i font, sia per la loro natura, sia per la larghezza nulla di alcuni glifi, sia per gli accenti.

Il fatto che non si possano conservare i link verso obbiettivi esterni al file prodotto è voluto dallo stesso standard. Ma i due problemi citati sopra sono i più comuni e sono problemi seri.

pdf \LaTeX e X \LaTeX sotto questo aspetto non presentano differenze sostanziali. I problemi sono gli stessi, e sono stati affrontati nell'articolo di Luigi presentato al *Meeting 2010*.

Rispetto all'anno scorso, forse, si aggiunge un problema; il fatto cioè che sia pdf \LaTeX sia X \LaTeX possono ora includere i file grafici in formato EPS (convertendoli al volo in formato PDF). In generale non ci sono problemi se i file EPS sono generati “in casa”. Ma se essi sono acquisiti dalla rete, o da persone diverse, potrebbero essere stati predisposti per la stampa, con un modello di colore CMYK che stride con le specifiche PDF/A.

Inoltre i file EPS importati dall'esterno potrebbero non avere inclusi i 35 font Adobe che si suppone abbiano già a bordo tutte le macchine fisiche o virtuali in grado di stampare il formato PostScript.

I font usati o da usare nel documento potrebbero presentare diversi problemi:

- essi potrebbero essere **bitmapped**, cioè **non vettoriali**; questo è difficile da realizzare se si usano i font del sistema $\text{T}_{\text{E}}\text{X}$ quando si compone con $\text{pdfL}_{\text{A}}\text{T}_{\text{E}}\text{X}$ o con $\text{X}_{\text{Y}}\text{L}_{\text{A}}\text{T}_{\text{E}}\text{X}$, ma può succedere, e deve essere evitato a tutti i costi. Basta usare, per esempio, Adobe Reader chiedendogli di mostrare la lista dei file inclusi nel documento: se ci sono dei **font Type 3**, quelli sono bitmapped e il documento deve venire ricomposto senza fare uso di quei font.
- Essi potrebbero non essere inclusi, per esempio i 35 font Adobe che si presume non sia necessario includere. Vanno inclusi anche quelli, naturalmente solo quelli effettivamente usati.

- Alcuni glifi potrebbero avere larghezza nulla; questo è difficile da scoprire, ma il programma **Preflight** della Adobe incluso nella versione Adobe Acrobat Pro 8 o successive permette di scoprirli. Se sono stati usati questi glifi bisogna trovare il modo di sostituirli con qualcosa che svolga la stessa funzione ma senza usare glifi di larghezza nulla.

Questo problema si manifesta con i font matematici della famiglia cmsy; per pdf \LaTeX ho già scritto come risolverlo (vedi bibliografia negli atti del *Meeting*).

Con X \LaTeX il problema non si manifesta se si usano i font matematici UNICODE (**XITS Math**, **Asana Math**, **Latin Modern Math**); ma se ci si affida ai font matematici di default il problema potrebbe presentarsi e va risolto nella stessa maniera che ho indicato per i font da usare con pdf \LaTeX .

Gli accenti nei file prodotti con font in codifica *T1* o in codifica UNICODE, vanno generalmente bene.

Ma potrebbe succedere che si debba importare un file composto con uno dei programmi del sistema \TeX , ma composto con font aventi la codifica di default *OT1*, cioè i font Computer Modern o qualsiasi altro font che usi quella codifica.

Per come il comando “primitivo” `\accent` sovrappone l’accento alla lettera che lo deve ricevere, **il file di uscita diventa incompatibile con lo standard.**

Prima di includerlo in un nostro documento dobbiamo usare il Reader per vedere quali font siano stati usati e, certamente, se si tratta dei Computer Modern e il testo contiene anche una sola parola con un accento, il file da includere è incompatibile e va ricomposto usando almeno la codifica *T1*.

Siccome non si può produrre il file PDF/A direttamente con X_YLaTeX, anche se abbiamo seguito tutti gli accorgimenti indicati prima, come si può risolvere il problema?

Ricorrendo al programma **ghostscript** con versione 8.60 o successiva e lavorando dalla linea di comando.

Bisogna seguire scrupolosamente le linee guida che ora verranno illustrate.

- 1 preparare in formato PDF con $\text{X}_{\text{E}}\text{L}_{\text{A}}\text{T}_{\text{E}}\text{X}$ il documento da convertire in formato PDF/A; le operazioni successive possono essere eseguite quando il documento è sostanzialmente nella sua versione finale e presumibilmente non si devono più eseguire correzioni o modifiche. Il nome del file composto sia `miofile.pdf`
- 2 copiare dalla cartella del ramo di `ghostscript` `.../ghostscript/versione/lib/` il file `PDFA_def.ps` e incollarlo nella cartella dove si trova anche il file da convertire curando di cambiargli il nome in: `miofile-def.ps`.
- 3 Modificare questo file con un editor di puro testo per modificare le righe commentate con `%% Customize` per inserire i *metadata*: modello di colore, profilo di colore, autore, titolo, parole chiave, eccetera.
- 4 eseguire la procedura bash (UNIX) o batch (Windows) `pdf2pdfa` descritta nel testo presente negli atti di questo *Meeting*.

Trasformazione eseguita!

Chiaramente la procedura `pdf2pdfa` va preparata in anticipo una volta per tutte; le uniche cose che bisogna fare per ciascun file da convertire sono quelle di generare il file `miofile-def.ps` con **il nome e i metadata specifici del file da convertire**.

Per il resto la procedura è molto semplice e abbastanza rapida; ma...

Purtroppo non si è ancora sicuri che il file convertito sia davvero conforme allo standard PDF/A-1b (questo sottoformato è quello che si riesce a fare con **ghostscript**).

Ad oggi non conosco ancora nessun programma gratuito o freeware che consenta di fare la verifica necessaria.

Bisogna innanzitutto diffidare di quanto il programma Adobe Reader X dice quando apre un file potenzialmente conforme allo standard PDF/A. Esso infatti informa l'utente che il file aperto è **conforme allo standard** e che quindi lo ha aperto in modalità di sola lettura. È una mezza verità: dovrebbe dire “probabilmente conforme” perché sia tutta la la verità.

Il programma quasi assolutamente affidabile per la verifica è la voce di menù **Preflight** di Adobe Acrobat Pro, versione 8 e successive. In teoria questa voce di menu dovrebbe essere anche capace di convertire un file non conforme in un file conforme ma non è chiaro come si faccia ad inserire i *metadata*; quindi probabilmente esso è capace *talvolta* di convertire in un file conforme un file che vorrebbe esserlo ma non lo è e gli manca poco per esserlo.

Nel mio articolo segnalo anche un sito internet dove si può inviare un file PDF per la verifica gratuita. Ho provato: ma un file dichiarato valido da Acrobat Pro è risultato non conforme con questa verifica on line.

Di chi ci si può fidare allora?

Non so rispondere. Io preferirei potermi fidare dei programmi della Adobe, visto che è la Adobe che ha inventato il formato PDF e ha proposto all'ISO (insieme ad altri enti) il testo dello standard PDF/A.

Ma anche nell'articolo di Luigi dell'anno scorso egli metteva in luce alcune manchevolezze di questo programma.

Probabilmente le cose non sono così brutte come le descrivo, e la verifica positiva si ha praticamente sempre se si seguono le linee guida che ho illustrato. Ma resta un senso di incertezza che preferirei non avere.

BUON X_ƎƎ^AT_EX

Questa presentazione *beamer* è stata composta con X_ƎƎ^AT_EX.
Font: Linux Libertine O, Iwona, UM Typewriter