Complessità della codifica ed ergonomia strumentale nel contesto XML-TEI: dove siamo? (Bilancio a partire da un nuovo progetto di edizione digitale medievale)

Marta Materni

Complessità della codifica ed ergonomia strumentale nel contesto XML-TEI: dove siamo? (Bilancio a partire da un nuovo progetto di edizione digitale medievale)

Marta Materni
IF Marie Curie, LUHCIE, UGA-Université Grenoble Alpes
marta.materni@univ-grenoble-alpes.fr
marta.materni@gmail.com

*** This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowslka-Curie grant agreement N° 745821 (DigiFlor)**

ABSTRACT (300 parole)

Il contributo, nato dall'esperienza maturata nell'ambito di un nuovo progetto di edizione digitale medievale, intende fare il punto sulla questione, ancora problematica, della produzione di file XML-TEI. Gli strumenti, ancora oggi come un decennio fa quando Peter Robinson aveva denunciato il problema, sono ben lontani dall'essere ergonomici ed efficaci. La questione è spinosa in quanto la lacuna strumentale incide sia sulla diffusione della pratica editoriale digitale – la mancanza di strumenti non è il solo fattore che ha frenato quello che, stando all'idea di una possibile *rivoluzione*, doveva conseguentemente presentarsi come una crescita esponenziale delle edizioni digitali realizzate secondo principi scientifici; sia, ancor peggio, e assai più direttamente, sulla qualità della codifica proposta. Il rischio che si sta correndo è quello di continuare a produrre "codifiche semplici con strumenti complessi" quando ormai la quotidianità filo-digitale dovrebbe essere la produzione di "codifiche complesse con strumenti semplici". Il progetto di edizione messo in opera ha implicato la creazione di un prototipo di editor XML-TEI capace di generare il codice in modo automatico a partire da una sintassi simbolica (sul modello di Markdown) applicata a un semplice file .TXT associato con un file .CSV, con la possibilità di ottenere inoltre molteplici codifiche a partire dallo stesso file modificando le corrispondenze fra simboli e tag XML nel file .CSV. È stato possibile in questo modo realizzare una codifica complessa secondo un modello di testo concepito come un database di parole, pronto a essere utilizzato immediatamente per altre analisi al di là degli obiettivi specifici di edizione del contesto progettuale in cui il file è stato prodotto. Si precisa in questo modo, anche, il ruolo del "filologo digitale", responsabile della diffusione all'interno (e all'esterno) della comunità accademica di un testo filologicamente **E** informaticamente corretto, "pronto all'uso", secondo un modello di capitalizzazione progressiva del lavoro di analisi.

PAROLE CHIAVE

Edizione digitale, TEI, editor TEI, modellizzazione della codifica

TESTO

Nel 2005, in un articolo ancora (purtroppo) di attualità e che offre numerosi spunti di riflessione, Peter Robinson lanciava un grido di allarme: «We need some things we do not yet have: software that does not exist and established online publication systems that have yet to be created. Let us not wait too long for these» (Robinson 2005, § 31). La letteratura grigia lascia spesso trasparire del disagio al riguardo, e nemmeno tanto velato. Tara Andrews, autrice di una soluzione parzialmente analoga a quella qui proposta, TEI Markup, è piuttosto efficace nella descrizione:

> TEI XML is a wonderful thing. [...] The problem is the transcription itself. When I am transcribing a manuscript [...] I do not want to be switching back and forth between keyboard layouts in order to type <tag attr="attr">arrow-arrow-arrow-arrow-arrow</tag> every six seconds. It's prone to typo, it's astonishingly slow, and it makes my wrists hurt just to think about it. (https://metacpan.org/pod/Text::TEI::Markup)

Recentemente Roberto Rosselli del Turco ha ripreso la questione in un articolo di sintesi del 2016 che prende in considerazione i due poli della problematica: produzione e visualizzazione di un'edizione digitale. L'argomento è talmente vasto che in questo contesto vorrei soffermarmi solo sul polo produzione, che si rivela strettamente connesso alla questione della complessità del modello di codifica proposto: "complessità di codifica di una codifica complessa" potrebbe essere la definizione sintetica del problema. La conclusione di Rosselli del Turco un decennio dopo la denuncia di Robinson costituisce la base di partenza per la riflessione:

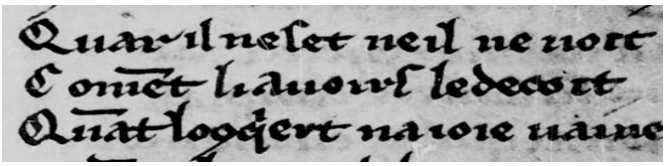
> Robinson's remark about the lack of easy-to-use production tools is unfortunately still valid: there is no software tool nor suite of tools that allows a scholar to produce a full digital edition, be it image-based with a diplomatic text or a critical edition, in a way comparable to how printed editions are prepared (Rosselli del Turco 2016, § 16)

Il problema strumentale è ovviamente scientificamente piuttosto grave in quanto rischia di condizionare in modo *inaccettabile* le nostre proposizioni teoriche di modellizzazione della codifica testuale, il che significa in ultima analisi una prospettiva in cui lo strumento tecnico ha il potere di influenzare le nostre decisioni intellettuali riguardanti la rappresentazione del pensiero. Ancora una frase di Robinson a farci da guida:

> A well-made electronic scholarly edition will be built on enoding of great complexity and richness. As well as free text searching, efficient search systems can make use of this encoding to enable sophisticated search, going considerably beyond the standard word and phrase search. (Robinson 2004)

Analisi testuali complesse sono possibili solo a partire da una codifica testuale complessa: altrimenti, a che pro ricorrere alla dimensione digitale quando quella stampata è già così rodata ed efficiente rispetto a una certa ben radicata visione dell'edizione? Gli strumenti proposti ufficialmente, come *Oxygen*, con la loro farraginosità di produzione, ci inchiodano di fatto a codifiche al limite talvolta della banalità. Di fronte alla ricchezza di informazioni estrapolabili da un

verso in antico francese tramandato da un manoscritto medievale – sistema abbreviativo, stadio evolutivo della lingua, grafie regionali, segmentazione della scrittura, forme paleografiche ecc. –



quanti dati perdiamo proponendo una codifica di questo tipo, purtroppo piuttosto diffusa?

```
<l>Quar il ne set ne il ne voit</l>
<l>Come<ex>n</ex>t li avoirs le deçoit</l>
<l>Qua<ex>n</ex>t lo <ex>con</ex>q<ex>ui</ex>ert n’a joie vaine</l>
```

Quanto possiamo invece guadagnare in potenziale di analisi con una codifica di quest’altro tipo, che trasforma il testo in un database di parole, ciascuna identificabile e quindi analizzabile e quindi manipolabile? Limitiamoci al terzo verso:

```
<l n="3" xml:id="msl3">
   <w xml:id="msl3w1">Qua<expan corresp="#abb-tild-nas"><ex>n</ex></expan>t</w>
   <w xml:id="msl3w2" ana="#aggl-s">lo</w>
   <w xml:id="msl3w3"><expan corresp="#abb-tir-9"><ex>con</ex></expan>
         <expan corresp="#lsup-q">q<ex>u</ex>i</expan>ert</w>
   <w xml:id="msl3w4" ana="#elis">n</w>
   <w xml:id="msl3w5" ana="aggl-s-unc">a</w>
   <w xml:id="msl3w6"><c ana="#lram-cons">i</c>oie</w>
   <w xml:id="msl3w7"><c ana="#lram-cons">u</c>aine</w>
</l>
```

Possiamo in questo modo ottenere una visualizzazione diplomatica in modalità “trascrizione”:

Qua*n*t lo*conqui*ert na-ioie uaine

e in modalità “interpretativa”:

Quant lo conquiert n’a joie vaine

Si tratta del modello di codifica proposto per un recente progetto di edizione diplomatica digitale dei testimoni di un romanzo francese del XII sec., e pensato come una sorta di modello di codifica proiettato sul futuro, al di là degli obiettivi contingenti legati al progetto specifico: il testo, cristallizzato nella forma documentaria che si offre ai nostri occhi, viene scomposto nei suoi

elementi minimali portatori di significato – la parola (un paleografo, interessato all'immagine del testo, probabilmente considererà come unità minima il carattere) – e ricostruito sotto forma di un file testuale codificato che vuole rappresentare un avatar digitale, cioè *formalizzato*, della realtà manoscritta. Così preparato, il file XML-TEI ha soddisfatto gli obiettivi immediati del progetto – la realizzazione di un'edizione diplomatica visualizzabile in modalità trascrizione grafematica e in modalità interpretativa modernizzata –, offrendosi al contempo come prodotto "già pronto" per essere sottoposto a ulteriori analisi: lemmatizzazione e analisi morfo-sintattica, utilizzo del *Double end-point attached method* per la creazione di un'edizione critica, analisi sintattiche, retoriche ecc. facilmente codificabili a questo punto attraverso uno stand-off markup dal momento che ciascuna parola è associata con un ID. Si tratta di applicare il principio metodologico dei "nani sulle spalle dei giganti", senza naturalmente, fuor di metafora, attribuire alcun valore di "nanismo" o "gigantismo" scientifico ai due momenti; o, in altri termini, di capitalizzare al massimo il lavoro, cioè il tempo della ricerca. La posizione pragmatica, direi quasi imprenditoriale, rispetto alla codifica, sostenuta con forza per es. da Elena Pierazzo:

> We all know how important economic considerations are in our decision-making processes; almost all of our research projects are funded for a specific time-span and budget, and so it is fundamental to ensure that the transcription (and encoding) is feasible within this lifetime. The decision whether or not to encode a specific feature will be heavily determined by the cost of encoding it, and it would be naive not to admit this. Economic considerations may then be used as the pragmatic limits of the level of transcription we are looking for, in the same way that the limits of the typography worked for print-based publications.
> [...] How far should we go in considering the needs of the Others? [...] if it is matter of considering the needs of possible future scholars in other disciplines and providing special markup for them, the temptation is to say: 'not far'. There is, in fact, a serious risk of wasting precious project time here: it is very difficult to guess other scholars' needs or principles which are, potentially, infinite – as infinite as the set of 'facts' that can be derived from the document. (Pierazzo 2011, p. 469, p. 471)

è esattamente il contrario di quanto si vuole qui proporre, e cercare di perseguire. Se tempo e denaro influenzano ovviamente le nostre attività, l'obiettivo scientifico non potrà essere tuttavia quello di assecondare le "leggi del mercato" quanto piuttosto di alleggerire il peso del loro giogo attraverso un percorso di sperimentazione e ricerca che permetta di "abbassare i costi e ottimizzare i tempi", per continuare a utilizzare questa metafora mercantile, ottenendo ciò che viene reputato scientificamente corretto e necessario e non "economico".

Inoltre, la preoccupazione riguardo «how far should we go in considering the needs of the Others» potrebbe essere in realtà una falsa preoccupazione derivante dalla tendenza a cumulare nello stesso file gli strati di codifica descrittiva e gli strati di codifica interpretativa. Piuttosto che prevedere uno «special markup» per ciascuna, possibile disciplina, in quanto filologi creatori di un'edizione digitale dovremmo in realtà proporre una sorta di "markup passepartout" che descriva ma soprattutto formalizzi il testo, al quale poi le varie discipline applicheranno la loro griglia, e codifica, interpretativa. Lo stand-off markup è dal mio punto di vista la risposta per eccellenza, risposta che presuppone, per essere praticata correntemente, la formalizzazione del testo come agglomerato di elementi minimi identificati in modo univoco e utilizzabili quindi come *anchor*.

La questione diventa dunque: come creare in tempi ragionevoli e in una modalità scevra da errori ricorrenti un prodotto complesso e utile, a me e agli altri? Non certo attraverso editor XML "prestati" al mondo TEI, bensì attraverso editor XML-TEI *specificatamente* concepiti per la codifica testuale, e, in più, attraverso l'automatizzazione massimale della produzione dei tag XML. Più il processo è automatizzato, più l'errore della digitazione è ridotto, più la complessità della codifica è incrementata.

Se, ed è evidente, il processo di produzione con strumenti come Oxygen è farraginoso e lento, parimenti non reputo che la soluzione del problema possa passare attraverso piattaforme online come T-Pen o la nuova nata TextualCommunity, che semplicemente spostano il problema dal l'ambiente di lavoro "in locale" a quello "in linea", incatenandoci fra l'altro a un utilizzo online che rappresenta una vera e propria schiavitù oltre a soffrire di una certa fragilità tecnologica. Né attraverso strumenti che nascondono il codice e che rappresentano un vero e proprio ossimoro teorico rispetto all'idea della codifica testuale: se il primo insegnamento rispetto al senso profondo della TEI insiste sulla differenza fra un sistema WYSIWYG e uno WYSIWYM, utilizzando uno strumento che cela ai nostri occhi il codice XML-TEI noi stiamo esattamente producendo un codice WYSIWYM con uno strumento WYSIWYG, stiamo cioè andando contro la stessa filosofia di cui dovremmo essere divulgatori.

La leggibilità umana del file XML non può a mio avviso rappresentare un principio guida della codifica, a meno di rinunciare, in nome di questa leggibilità, alla possibilità di analisi radicalmente alternative al paradigma editoriale cartaceo. Quello che propongo in alternativa è un processo di produzione che passa attraverso dei file testuali "di mediazione", basati su un linguaggio simbolico ispirato alla sintassi Markdown HTML et alle *entities* XML (&…;). Si tratta della metodologia sperimentata nella realizzazione di questo progetto di edizione digitale e concretizzatasi nella creazione di un editor-prototipo: TEI-med.it, realizzato in Python. I simboli, applicati a un semplice file .TXT realizzabile con qualsiasi editor (Word *ovviamente* escluso), sono associati, in un file .CSV esterno, liberamente modificabile dall'utente, ai tag XML, con il duplice vantaggio, oltre alla rapidità di realizzazione, di poter proporre più soluzioni di codifica a partire dallo stesso .TXT, semplicemente associando un differente .CSV.

Riprendendo il verso di cui si è proposto il codice XML precedentemente, quel codice è stato ottenuto a partire da questa stringa testuale:

Qua&n;t +lo &9;&qi;ert 'n +_a *ioie *uaine

Il linguaggio TEI-Med – che è solamente, sottolineo, un linguaggio di mediazione, strumentale, non un markup alternativo a TEI – si avvale di tre categorie di simboli:

1. simboli ASCII associati a una lettera o a una parola per produrre intere stringhe di codice o per generare un attributo:

es. * associato a i, *i, produce:

```
<c ana="#lram-cons">u</c>
```

o, volendo una codifica più classica:

```
<choice><orig>u</orig><reg>v</reg></choice>
```

2. *entities* fisse, corrispondenti a un singolo tag o a una stringa intera di codice, assai utile per le abbreviazioni, la struttura è &…;

es. &n; produce:

```
<expan corresp="#abb-tild-nas"><ex>n</ex></expan>
```

o, volendo:

```
<choice><abbr><am><g ref="tild-nas"/></am></abbr>
        <expan><ex>n</ex></expan></choice>
```

3. *entities* con argomenti variabili, la struttura è &…;($)

es. &sub-ex-int;(parla,parola), corrisponde a

```
<sub><del rend="erasure"><w xml:id="msl1w1">parla</w></del>
        <add place="interline"><w xml:id="msl1w2">parola</w></add></sub>
```

Le categorie 2 e 3 sono liberamente modificabili e incrementabili dall'utente finale attraverso il .CSV. Linee <l>, parole <w>, @xml:id e @n vengono prodotti automaticamente dall'editor senza alcun intervento da parte dell'operatore sul .TXT.

Il sistema ha permesso di realizzare la trascrizione completa e la codifica di 6 manoscritti di un testo di 13.500 versi (all'inizio del progetto ne erano stati previsti 4) nello spazio di circa 18 mesi. Le edizioni sono in corso di pubblicazione sul sito *DigiFlorimont* (http://digiflorimont.huma-num.fr/).

BIBLIOGRAFIA

[1] Pierazzo, E. 2011. *A Rationale of Digital Documentary Editions*. In: Literary and Linguistic Computing, vol. 26.4: 463-477.

[2] Robinson, P. 2004. *Where We Are with Electronic Scholarly Editions, and Where We Want to Be.* http://computerphilologie.uni-muenchen.de/jg03/robinson.html

[3] Robinson, P. 2005, *Current Issues in Making Digital Editions of Medieval Texts–Or, Do Electronic Scholarly Editions Have a Future?*, In: Digital Medievalist, vol. 1.1. http://www.digitalmedievalist.org/journal/1.1/robinson DOI=http://doi.org/10.16995/dm.8

[4] Rosselli del Turco, R. 2016. *The Battle We Forgot to Fight: Should We Make a Case for Digital Editions?*. In: *Digital Scholarly Editing: Theories and Practices*, Cambridge, Open Book Publisher, http://books.openedition.org/obp/3423>. ISBN: 9782821884007.

[5] Sahle, P. 2016. *What is a Scholarly Digital Edition?*. In: *Digital Scholarly Editing: Theories and Practices*, Cambridge, Open Book Publisher, http://books.openedition.org/obp/3423>. ISBN: 9782821884007.

[6] Stolz, M. 2017. *Copying, Emergence and Digital Reproduction. Transforming Medieval Manuscript Culture into an Electronic Edition*. In: Digital Philology, vol. 6.2: 257-87.