

PAOLO MONELLA*

FORME DEL TESTO DIGITALE

ABSTRACT. – *Forms of the digital text.* This article reflects on the modelization of texts of our cultural heritage in the current very early phase of Digital Philology by asking the question: what “form” are we forcing upon our texts? I explore three ways in which our textual models mirror our cultural and technological *bias*: first, the OHCO hierarchical model behind TEI XML; second, the sequential alphabetical model based on Western print, lying behind Unicode. A third aspect has been less explored in the scholarly discussion: citation protocols such as CITE/CTS are based on specific corpora. For example, the canonic corpora for classical Greek and Latin texts are TLG and PHI 5.3. Each of those corpora freezes author identification, authorship attribution, text boundaries, paragraph and line numbering in a specific canonic form. A dilemma arises: on the one side, interoperability requires that CTS URIs are based on a fixed textual canon, which hinders scholarly modifications of that canon; on the other side, the principles of philology suggest that URIs point to specific versions of texts, which in its turn cripples the interoperability potential of the protocol.

RIASSUNTO. – L’articolo costituisce una riflessione sulla modellizzazione dei testi appartenenti alla nostra tradizione culturale in questa fase aurorale della Filologia Digitale e parte dalla domanda: che «forma» stiamo dando ai nostri testi? Vengono esplorati tre aspetti per cui i nostri modelli testuali riflettono i nostri presupposti culturali e tecnologici: in primo luogo, il modello gerarchico OHCO che sta alla base di TEI XML; in secondo luogo, il modello sequenziale alfabetico basato sulla stampa delle lingue occidentali, che sta alla base di Unicode. Un terzo aspetto, poi, è il meno esplorato nella bibliografia: protocolli per la citazione testuale come CITE/CTS sono basati su corpora specifici. Per esempio, i corpora canonici per i testi greci e latini classici sono il TLG e il PHI 5.3. Ciascuno di questi corpora fissa in una forma canonica l’identificazione degli autori, l’attribuzione delle opere, la delimitazione dei testi, la numerazione di porzioni testuali come paragrafi e versi. Ne scaturisce un dilemma: da una parte, l’interoperabilità richiede che gli URI CTS siano basati su un canone testuale fisso, il che tende ad ostacolare le modifiche al canone da parte dei filologi; d’altra parte, i principi della filologia suggeriscono che gli URI puntino a versioni specifiche dei testi, il che però riduce il potenziale di interoperabilità del protocollo.

* Dipartimento Culture e Società – Università degli Studi di Palermo – Viale delle Scienze, Ed. 15 – 90128 PALERMO.

INTRODUZIONE

L'alba dell'era digitale porta con sé anche una «rivoluzione testuale», ovvero un cambiamento profondo nella concezione del «testo». Le due rivoluzioni che hanno preceduto quella attuale hanno avuto come luoghi emblematici la biblioteca di Alessandria (per il passaggio dal testo orale a quello scritto) e la stamperia di Gutenberg (dal testo manoscritto a quello stampato). Ognuna di esse ha contribuito a dare una nuova «forma» al testo, più stabile, chiusa, strutturata, autoriale¹.

Con questo contributo voglio proporre un esperimento di auto-posizionamento storico², ponendo una semplice domanda: nella rivoluzione testuale attuale, quale «forma» stiamo dando *noi filologi digitali* ai testi?

L'esperimento che vorrei tentare è dunque di osservare il processo di messa in forma – si sarebbe tentati di dire di «formato» o addirittura di «formattazione» – cui la nostra comunità di editori-filologi digitali sta sottoponendo i testi della nostra eredità culturale, e le specifiche (inevitabili) forzature che stiamo loro imponendo.

Vorrei insomma pormi nella prospettiva di un Raul Modenti futuro, che invece di studiare oggi la «forma» imposta dalla pressa di Gutenberg ai testi medievali, studi in futuro la forma imposta oggi dai nostri formati digitali ai testi che andiamo digitalizzando.

Mi soffermerò su tre aspetti in particolare, e più diffusamente sul terzo, che mi pare del tutto ignorato:

- A forma di albero (TEI XML)
- A forma di alfabeto (Unicode)
- A forma di corpus (CTS).

¹ Questo tema, che sta alla base di buona parte del mio lavoro in informatica umanistica, mi viene dai grandi saggi degli anni '90 e primi anni del 2000 della «scuola italiana» dell'Informatica Umanistica, tra cui Orlandi (1990), Leonardi *et al.* (1994), Nerozzi (1997), Mordenti (2001), Fiormonte (2003). Tra l'altro, proprio in questa giornata di studi Raul Mordenti ha ricordato che «la rigidità è caratteristica della stampa, mentre la fluidità è caratteristica dell'edizione digitale».

² Sulla necessità di «uno studio del nostro punto di vista» (Cozzo, 2006: 22) e di una ricerca «consapevolmente posizionata, esplicita sui suoi presupposti e sui suoi scopi» (id., 28) si vedano le lucide introduzioni di due saggi di Andrea Cozzo (Cozzo, 2002: 13-36 e soprattutto Cozzo, 2006: 9-29).

A FORMA DI ALBERO: TEI XML

Il linguaggio di marcatura TEI XML è diventato lo standard per la codifica dei testi letterari, ed un requisito *de facto* per ogni progetto di codifica digitale di testi che ambisca ad accedere ai fondi degli enti finanziatori della ricerca³.

Ma ogni testo pre-moderno codificato in TEI XML subisce già due forme di «normalizzazione». La prima è legata alla TEI in quanto vocabolario condiviso di *tag* per commentare (o, come comunemente si dice, marcare) testi, la seconda al modello di dati ad «albero» di XML.

TEI come vocabolario comune.

La TEI mira ad essere un (meta)linguaggio comune per parlare dei testi attraverso il markup, ovvero un vocabolario di riferimento per chiamare tutti (più o meno) le stesse cose con (più o meno) le stesse parole, in vista dell'«interoperabilità», cioè della possibilità di utilizzare *file* attraverso sistemi software diversi, e attraverso i decenni⁴.

Ora, un file, in TEI XML o in qualunque formato, è tanto più interoperabile quanto più standard è la sua codifica. I testi, però, e soprattutto quelli letterari e pre-moderni, sono quanto di meno standard si possa immaginare: persino un concetto semplice come quello di «verso» (tag TEI <l>, dall'inglese «line») costituisce una normalizzazione forzata, se applicato a quei manoscritti greci medievali in cui i versi dei cori di una tragedia greca classica sono scritti di seguito, senza interruzioni di riga né altra distinzione grafica.

³ Si veda Schmidt (2012: par. 2.1), che cita (a n. 23) le linee guida per gli *Scholarly Editions and Translations Grants* dell'americano *National Endowment for the Humanities* (NEH). Una versione più aggiornata del documento è ora disponibile in <http://www.neh.gov/files/grants/scholarly-editions-dec-9-2015.pdf>. Questo e tutti gli altri URL indicati nell'articolo sono stati consultati per l'ultima volta il 14 luglio 2016.

⁴ Si vedano le argomentazioni della proposta originaria al NEH per la creazione delle prime Guidelines TEI in Ide *et al.* (1988: par. 1.3): «Several projects are already underway to encode and distribute massive data bases of texts via CD-ROM, each proceeding or planning to proceed with its own encoding scheme. Together, the scanner and the CD-ROM promise to aggravate the problem of anarchic encoding practices by several orders of magnitude within a very few years unless action can be taken soon».

E cos'è la «punteggiatura» (TEI <pc>, «punctuation character») in un manoscritto medievale o in un'iscrizione antica? Cos'è un «paragrafo» (TEI <p> «paragraph») in questi sistemi scrittori? Cosa sono i «titoli» (TEI <head>) in un manoscritto medievale con rubriche? Sono lo stesso che in una raccolta poetica di Ungaretti o in un'edizione moderna di Catullo in cui, mancando i titoli latini antichi, sono sostituiti da numeri arabi⁵?

Al fondo, resta una tensione ineliminabile: il prezzo dell'interoperabilità è la normalizzazione, la «formattazione» del testo, anche pre-moderno, secondo concetti che nascono nel mondo della testualità gutenberghiana.

L'albero XML.

La seconda «forma» che, silenziosamente, imponiamo ad ogni testo codificato in XML/TEI, e che si attaglia più forzatamente ai testi di origine pre-gutenberghiana, è quella «ad albero» imposta da XML.

Il modello OHCO (il testo come «ordered hierarchy of content objects»), che sta alla base dell'applicazione di tale struttura alla codifica digitale dei testi, presuppone che ogni testo, per sua natura, sia strutturato come un albero, ovvero con «scatole» (corpus, testo, capitolo, paragrafo) che si annidano perfettamente le une dentro le altre⁶.

Le maggiori contestazioni a questo modello si sono incentrate sull'*overlapping*, per il semplice motivo che praticamente ogni testo codificato in XML/TEI si scontra con questo problema.

A chi ha elaborato il modello OHCO, appariva inizialmente naturale che un paragrafo (<p>) non possa iniziare in un capitolo (<div type=«chapter»>) e finire in un altro, ma a chi codifica testi appare assai forzato che una frase (<s>, da «sentence») non possa iniziare in un verso (<l>) e finire in un altro⁷.

⁵ Per i limiti effettivi dell'interoperabilità della codifica TEI si veda Schmidt (2012: par. 3.6).

⁶ La definizione del modello originario è in Coombs *et al.* (1987) e DeRose *et al.* (1990). Cito da Coombs *et al.* (1987: par. *Outlining and Structure-Oriented Editing*): «Documents have a natural hierarchical structure: chapters have sections, sections have subsections, and so on, until one reaches sentences, words, and letters».

⁷ Il problema è stato affrontato ben presto dagli stessi studiosi che avevano elaborato il modello OHCO (Renear *et al.*, 1993). Un'ottima e concisa discussione, con ulteriore bibliografia, è in Schmidt (2012: par. 3.3).

Il problema è aggirabile con stratagemmi, per lo più basati sull'uso di elementi vuoti⁸. Con la sua ingombrante evidenza pratica, però, l'*overlapping* rischia di coprire una questione retrostante: prima dell'età moderna e dell'elaborazione di generi testuali come il romanzo o il saggio moderno (coi loro capitoli e paragrafi), non tutti i testi erano a forma di albero.

Testi di origine orale come l'epica omerica ed i «poemi ciclici» dell'antica Grecia si realizzavano, nella *performance*, in un canto continuo, per cui il modello più adatto sarebbe piuttosto quello di una sequenza di segni, e nella competenza dell'aedo quei testi esistevano come collezioni modulari di episodi e formule adattabili e combinabili tra loro. I testi omerici sono stati «fissati», con l'espunzione dei versi considerati spurii, e divisi in sezioni (libri) assai dopo la loro composizione, probabilmente solo quando la tecnologia della scrittura lo ha richiesto⁹.

Se i testi orali sfuggono al modello gerarchico ad albero in quanto legati alla materialità monodimensionale della catena fonica, le iscrizioni, e in fondo ogni testo manoscritto, vi sfuggono in quanto legati alla materialità bidimensionale del loro supporto materiale.

Sullo specchio scrittorio di una lastra di marmo, del basamento di una statua, di un papiro o una pergamena, la mano dello scriba sfrutta tutte le possibili relazioni spaziali, su due dimensioni, per codificare il testo¹⁰.

⁸ L'impiego di elementi vuoti, ovvero privi di contenuto testuale, in TEI XML riguarda aspetti del testo anche molto importanti: così l'onnipresente `<lb/>` («line break», andata a capo) ci permette di non marcare l'inizio e la fine della riga con tag di apertura e chiusura che darebbero infiniti problemi di overlapping, e una funzione simile hanno `<cb/>` («column break») e `<pb/>` («page break») per la distinzione del testo in colonne e pagine. Le strategie di «markup esterno», poi, collocano nel file dei segnaposto vuoti (àncore) e spostano in un file separato la marcatura, che a quelle àncore fa riferimento per indicare la porzione di testo marcata. In entrambi i casi, si abdicava ad uno dei principi di fondo del markup, cioè all'idea di avere elementi che contengano (cioè «marchino») testo tramite tag di apertura e di chiusura.

⁹ Le sezioni testuali antiche erano spesso dette βιβλία, ovvero rotoli di papiro. La divisione in canti dei poemi omerici è peraltro assai controversa (v. Jensen, 1999), e diverse fonti antiche la attribuiscono a filologi e momenti diversi, tra cui Aristarco di Samotraccia (Ps-Plutarco, *De Homero* 2.4), Aristarco seguito da Zenodoto (Eustazio, *Commentarii ad Homeri Iliadem* 1.9.4-5) e lo stesso Pisistrato (Cicerone, *De oratore* 3.137).

¹⁰ All'interno del *TEI Consortium* è maturato di recente un nuovo interesse verso la complessità della *mise en page*, stimolato dallo *Special Interest Group* (SIG) sulle edizioni genetiche (<http://www.tei-c.org/SIG/Manuscripts/genetic.html>). Esso ha prodotto un'intera nuova sezione della TEI, il modulo *zone*, che configura un approccio completamente nuovo, «documento-centrico», alla codifica del testo, basato appunto sulle «zone» bidimensionali in cui il testo si distribuisce all'interno

Dei grafemi che non si susseguono solamente, ma si collocano sopra, dietro o sotto altri grafemi (iota sottoscritto, diacritici, abbreviazioni, *Umlaut*) dirò meglio nel prossimo paragrafo.

Intanto, restando alle relazioni strutturali del testo, penso alle glosse interlineari che si collocano logicamente «sulla soglia» del testo in quanto integrazioni, correzioni, sinonimi, parafrasi¹¹; ad altre forme di paratesto come le rubriche; alle corone di *scholia*, metatesto che scorre a margine, parallelamente al testo; ai ritmi medievali in cui la notazione musicale scorre parallela a quella testuale; ai calendari in cui le immagini costituiscono parte integrante dell'opera, spesso contengono testo e lo organizzano in una struttura grafica dalla forma imprevedibile.

Un giorno, quando nuovi modelli di testo saranno invalsi, forse anche grazie alla testualità digitale, la forma «ad albero» legata ad XML che stiamo dando a Omero e ai manoscritti medievali con le loro corone di *scholia* apparirà forse tra le più bizzarre forzature che le nostre scelte tecnologiche ci fanno imporre al multiforme mondo delle tradizioni testuali.

A FORMA DI ALFABETO: UNICODE

La scrittura gutenberghiana ci ha abituati ad una corrispondenza 1:1 tra grafemi e lettere alfabetiche (o «alfabemi», con un termine coniato da Raul Mordenti): la parola «parola» viene stampata con sei grafemi sequenziali, al-

della pagina, al di là della sequenzialità monodimensionale del testo astratto. Il capitolo 11 *Representation of Primary Sources* delle TEI P5 Guidelines ne è uscito in larga parte riscritto (<http://www.tei-c.org/release/doc/tei-p5-doc/it/html/PH.html>). Ciononostante, a causa della struttura di fondo di XML, lo stesso modulo *zone* non fa altro che creare a sua volta un nuovo albero, un'altra gerarchia ordinata di oggetti: <surface> (ad es. una pagina di manoscritto) / <zone> / <line> / testo contenuto nella riga. Si tratta di un albero sostanzialmente alternativo a quello «testo-centrico» tipicamente costituito da elementi strutturali come <div type="chapter"> / <p>, tanto che, quando entrambe le gerarchie coesistono in uno stesso documento TEI XML, esse sono nettamente distinte e convivono in parallelo, essendo la gerarchia «documento-centrica» (<surface> / <zone>) figlia di <TEI> / <facsimile>, e la gerarchia «testo-centrica» figlia di <TEI> / <text>, con strategie di collegamento tra le due strutture. Vedere il paragrafo 11.2 *Combining Transcription with Facsimile* delle *Guidelines* (<http://www.tei-c.org/release/doc/tei-p5-doc/it/html/PH.html#PH-transcr>).

¹¹ Si vedano Monella (2008) e Genette (1987).

lineati su una sola dimensione («p», «a» ecc.), ciascuno dei quali rappresenta un alfabema (la lettera «p», la lettera «a» ecc.)¹².

Questa razionalizzazione dei sistemi scrittori è stata completata nell'ambito della rivoluzione digitale, ed è culminata in Unicode¹³.

Non è stato sempre così. Come accennato nel paragrafo precedente, la scrittura a mano, nel mondo occidentale, ha sfruttato il potenziale della bidimensionalità del supporto per rappresentare più alfabemi possibile usando meno grafemi possibile. Si tratta delle cosiddette abbreviazioni.

Il caso più semplice è quello dei brevigrafi, in cui un grafema rappresenta più alfabemi: ad esempio, la «p» con un trattino orizzontale ad incrociare il tratto discendente spesso significa tre alfabemi: «p», «e» ed «r».

In altri casi, rompendo la sequenzialità monodimensionale, un grafema si colloca sopra, sotto o addirittura dietro un altro, creando un segno complesso che rappresenta una serie di alfabemi. La «e» tedesca di *Brueder*, dapprima scritta sopra la «u», è stata tracciata nel tempo in modo sempre più compendioso, fino a diventare l'*Umlaut* di *Brüder*. Nei manoscritti medievali un macron o un altro segno abbreviativo scritto sopra un grafema-base crea abbreviazioni come *fratrē* per *fratrem*, o *dns* con un lungo macron sopra le tre lettere per *dominus*¹⁴.

Gli esempi si moltiplicano se ci si allontana dal latino e dalle lingue che ne hanno ereditato alfabeto e sistemi grafici: basti pensare allo iota sottoscritto greco, ai segni diacritici arabi, fondamentali per la vocalizzazione e non solo, collocati in varie posizioni intorno al grafema consonantico di base, fino ai

¹² L'idea che il «significato» immediato di un grafema sia un alfabema era stata suggerita da Orlandi (1999). Implicitamente, essa si trova anche nel modello formale proposto in Stokes 2011. Per parte sua, Baroni (2009: 51-54) evidenzia come il «contenuto» (in termini semiotici, ovvero il significato) di un grafema non sia, semplicisticamente, un fonema, ma senza fare riferimento alle lettere alfabetiche. Una discussione ampia ed informata sui sistemi grafici alfabetici, comparati con altri (sillabici, logografici ecc.), è in Baroni (2009: 77-98). Una delle riflessioni più intelligenti sull'«alfabeto» che io abbia letto è nella prima parte di Mordenti (2011).

¹³ Si consulti The Unicode Consortium (2016), che però al momento della stesura di questo saggio non è ancora completamente disponibile online: intanto si veda The Unicode Consortium (2015).

¹⁴ Anche limitandosi ai sistemi grafici basati sull'alfabeto latino, è straordinaria la varietà e la complessità del mondo delle abbreviazioni e delle convenzioni della scrittura a mano dischiusa da un buon manuale di paleografia come Battelli (1949) o Cencetti (1978).

sistemi grafici indiani in cui a volte una vocale, per quanto pronunciata dopo la consonante, viene scritta in legatura prima di essa¹⁵.

Tutti i sistemi esistenti di codifica digitale del testo sono fondati sul modello normalizzato post-gutenberghiano di scrittura dell'alfabeto latino, ovvero su una sequenza monodimensionale di grafemi e sulla corrispondenza 1:1 grafema/alfabema¹⁶.

Ciò è evidente per sistemi semplici come ASCII, mentre Unicode sembra abbracciare convenientemente e pluralisticamente sistemi grafici non latini. Così è per la ricchezza dei singoli caratteri. Ma il modello sottostante è comunque «latino-gutenberghiano» in quanto i rapporti «bidimensionali» tra grafemi (un grafema sopra, sotto o prima dell'altro), al livello della codifica digitale «profonda», sono comunque ricondotti ad una semplice sequenza lineare di caratteri.

Il sistema complesso rappresentato nella scrittura a mano da un grafema sormontato da un altro con funzione di diacritico (ad esempio una *fā* araba iniziale sormontata da una *fathā*) è codificato in Unicode o come un terzo grafema rappresentante la combinazione tra i due segni, o come una sequenza di due caratteri: *prima* il grafema base (*fā*), *poi* il diacritico (*fathā*). Il rapporto sopra/sotto emergerà semplicemente a livello di visualizzazione.

Di questo ho già scritto in altra occasione¹⁷: basti qui notare come, anche a livello di codifica digitale dei sistemi grafici, rendiamo poca giustizia alla specificità dei sistemi bidimensionali, imitandone l'aspetto solo a livello di visualizzazione (*rendering*), ma nella sostanza uniformandoli forzosamente al modello bidimensionale (la «forma alfabeto») consolidatosi in Occidente nell'età moderna per l'alfabeto latino a seguito dell'invenzione della stampa¹⁸.

¹⁵ Per la scrittura devanagari, usata per sanscrito e hindi, si legga Perri (2009), in particolare il par. 4 (pp. 734-737), ma nello stesso saggio si veda anche l'esempio dello hangul coreano (pp. 739-742). Sul devanagari, si consultino anche Constable e Constable (2001: par. 6.3), Baroni (2009: 91) e The Unicode Consortium (2015: par. 12.1).

¹⁶ Perri (2009: 727) parla delle codifiche digitali del testo, e di Unicode in particolare, come di «una sorta di iper-tipografia, poiché ha adottato e spinto sino alle estreme conseguenze la “logica dell'alfabeto” e la sua “tirannia”», citando a sua volta Harris (2003).

¹⁷ Fiormonte *et al.*, 2015, par. *Universalizing the typography: Unicode*.

¹⁸ Si vedano le conclusioni di (Perri, 2009). Per quanto mi riguarda, negli ultimi anni ho lavorato ad un modello di edizione critica digitale fondata sulla trascrizione integrale delle fonti primarie tale da integrare una trascrizione grafematica, che rappresenti il sistema dei grafemi e dei paragrafi (punteggiatura e simili) nel documento manoscritto, una trascrizione alfabetica, che rappre-

A FORMA DI CORPUS: CTS

Le citazioni canoniche a stampa.

La filologia classica si occupa di opere canoniche, come l’*Iliade* o l’*Eneide*, fissate e levigate da secoli di studio, cura ecdotica e «correzione» (riscrittura) in una precisa forma linguistica e testuale¹⁹.

Le opere del canone classico, come i testi canonici della Bibbia, sono così diventate facilmente citabili tramite convenzioni semi-formali come «Hom. *Il.* 1.2», ovvero «il secondo verso del primo libro dell’*Iliade* di Omero». Si presuppone, senza rifletterci troppo, che tutti siano d’accordo su chi sia Omero (o almeno su quali opere vengano comunemente raggruppate sotto questo nome), su cosa sia l’*Iliade*, sulla divisione di questa in libri e persino sulla numerazione dei versi – cioè su quanti e quali versi esattamente compongano l’opera e quali siano invece spuri.

Il corpus testuale classico appare già quanto di più rigido e formalizzato si possa immaginare. Ma anche qui può intervenire, ed in parte sta già intervenendo, una nostra scelta tecnologica ad imporgli una forma ancora più rigida.

In un certo senso, una convenzione come «Hom. *Il.* 1.2» resta ancora, come dicevo, «semi-formale»: una delle «bonarie tassonomie semiclandestine» di Avalle, spesso citate da Raul Mordenti²⁰. Una modifica della composi-

senti la sequenza lineare degli alfabemi significati da quei grafemi, ed una trascrizione linguistica, normalizzata. Nel 2012 avevo iniziato un primo esperimento di edizione, poi lasciata incompleta, del testo latino tardoantico *Iudicium coci et pistoris* di Vespa (*Anth. Lat.* 199 Riese): vedere <http://www1.unipa.it/paolo.monella/lincei/edition.html>. I principi metodologici e le soluzioni tecnologiche sperimentate in quell’occasione sono esposti in Monella (2014). Nel 2015-2016 ho ripreso il lavoro nel quadro del progetto ALIM (Archivio della Latinità Italiana del Medioevo, <http://alim.unisi.it/>), producendo un’edizione critica digitale del *De nomine*, un’operetta grammaticale di Orso di Benevento (IX secolo) trādita nel codice Casanatensis 1086. Al momento della stesura di questo saggio (luglio 2016), l’edizione è completa ma in fase di revisione finale, ed è già disponibile in <http://www1.unipa.it/paolo.monella/ursus/transcription.html>, insieme a tutti i sorgenti markup e software, e ad una documentazione che ne illustra i principi. Alla base di questo esperimento stanno le riflessioni di Sampson (1990) e soprattutto di Orlandi (2010). Va ricordato come Orlandi stesso avesse creato un’edizione-prototipo parziale in (Orlandi, 2006).

¹⁹ Vedere Cozzo (2006: 223-227) e Monella (in stampa).

²⁰ Vedere Avalle (1985: 380), Mordenti (1999: 183) e Mordenti (2007: 152).

zione o dell'organizzazione del canone può essere infatti facilmente recepita e riflessa da questa convenzione.

Se uno studioso volesse attribuire i *Cypria* (uno dei «poemi ciclici») ad Omero, gli basterebbe scrivere e far stampare «Hom. *Cypria* fr. 1» e al massimo aggiungere una nota per motivare la propria posizione al riguardo²¹.

Chi vuole mettere in discussione l'autenticità della *Epistola Sapphus* attribuita a Ovidio e collocata in quindicesima posizione tra le sue *Heroides*, non ha che da citarla, come si fa per i testi adespoti, semplicemente come *Epistola Sapphus*, o come Ps-Ov. *Her.*, piuttosto che come Ov. *Her.*²².

E così, chi vuole vedere nell'elegia properziana tradizionalmente numerata come 2.22 due componenti distinti li distingue in Prop. 2.22a e 2.22b, e ricomincia la numerazione dei versi all'inizio di quest'ultimo.

Sotto gli occhi di tutti è poi la sacrosanta libertà degli editori di ri-numerare versi, paragrafi o frammenti in caso di aggiunte, espunzioni o trasposizioni, non infrequenti anche nel corpus canonico dei testi classici. Chi non ha mai sbattuto la testa con un passaggio introvabile per questioni di numerazione? Giungono in soccorso, nei casi migliori, note esplicative e tabelle di concordanza con le numerazioni dei principali editori precedenti.

Un formalismo bonario, dunque, mediato dall'intuizione del lettore umano, che vede l'identità tra espressioni come Hom. *Il.* 1.2; Hom. *Il.* 1, 2; Hom. *Il.* I 2; Hom. *Iliad* 1.2; *Iliad* 1.2 ed interpreta assai elasticamente ogni sorta di abbreviazione, omissione, convenzione tipografica.

Il protocollo CTS: sfide pratiche.

Il sistema tradizionale delle citazioni canoniche, per i testi greco-latini classici, ma anche per quelli biblici o per i testi più reverendi delle tradizioni letterarie occidentali, appare comunque assai formale, già pronto per l'ulteriore formalizzazione digitale. Il progetto più coerente in tal senso, nel mondo classicistico, è quello del protocollo CITE/CTS, cui lavora da anni un gruppo di umanisti informatici, e di cui è attualmente in cantiere una profonda revisione²³.

²¹ L'attribuzione era circolata nell'antichità, ma già Erodoto (*Hist.* 2.117) la contestava.

²² Si confrontino Tarrant (1981), contro la paternità ovidiana, e Rosati (1996), a favore.

²³ CITE sta per «Collections, Indices, Texts, and Extensions», CTS per «Canonical Text Services». Molti dei siti che avevo annotato nel 2012 come contenenti informazioni sul protocollo CTS e sulla sua espansione CITE sono oggi, a luglio 2016, irraggiungibili, a prova del fatto che il pro-

Secondo il protocollo CTS, una citazione tradizionale come Hom. *Il.* 1.2 diventa urn:cts:greekLit:tlg0012.tlg001:1.2, in cui tlg0012 rappresenta Omero nel canone TLG (per cui vedi sotto), tlg001 l'Iliade, 1 il primo libro e 2 il secondo verso²⁴.

Anche qui, come per gli esempi visti in precedenza, nascono questioni pratiche – sempre in qualche modo aggirabili – e questioni di fondo. Queste ultime sono, anche in questo caso, quelle che più mi interessano in questa sede.

I problemi pratici sono facilmente intuibili:

- Ad ogni autore, ad ogni opera e ad ogni sotto-divisione del testo va attribuito un codice identificativo. Chi crea il canone con questi codici? Per i testi latini classici esiste il canone PHI²⁵, e per quelli greci classici il canone TLG²⁶, ma molto lavoro resta da fare per i testi greco-latini meno «canonici»: papiri, frammenti, tardo-antico, padri della Chiesa, medioevo, età moderna. Considerato però che il protocollo, nato in ambito classicistico, mira ad un'applicazione molto più ampia anche alle altre letterature, i campi da arare sono ancora più ampi.
- Creato il canone, chi lo manterrà stabile nel tempo, curato, aggiornato e facilmente accessibile tramite API (sostanzialmente su un portale accessibile a qualunque programma faccia delle interrogazioni specifiche)?

Per entrambe le sfide, un ovvio candidato è il *Perseus Project*, che ha già lanciato un *Perseus Catalog* (<http://catalog.perseus.org/>) basato sui canoni PHI e TLG ma in ulteriore espansione²⁷.

Il protocollo ha stentato a trovare un'ufficializzazione istituzionale. Esso sopravvive però nelle sue implementazioni in vari progetti, variamente collegati o comunque in contatto col *Perseus Project*. Le introduzioni più complete disponibili al momento sono la sezione *The CITE architecture* all'interno della documentazione del progetto *Homer Multitext*, in <http://www.homermultitext.org/hmt-doc/cite/index.html> e Tjepmar *et al.* (2014), che illustra il protocollo, enumera e discute le sue attuali implementazioni e rimanda alla bibliografia relativa. Un gruppo di lavoro informale organizzato da Bridget Almas con cui io stesso collaboro, seppur marginalmente, ha messo in cantiere una riformulazione ed espansione del protocollo.

²⁴ «CTS URNs are citations, and they express the semantics of a text according to the OHCO2 model» (da <http://www.homermultitext.org/hmt-docs/cite/cite-overview.html>).

²⁵ <http://latin.packhum.org/canon>

²⁶ <http://stephanus.tlg.uci.edu/canon.php>

²⁷ <http://catalog.perseus.org>. Per il rapporto tra i codici identificativi del *Perseus Catalog* e quelli dei canoni PHI e TLG, vedere https://github.com/PerseusDL/catalog_data/wiki. Per le strate-

Testi in movimento.

Come nel caso dell'*overlapping* per l'XML, la difficoltà di queste sfide operative rischia di mettere in ombra una questione di fondo: un canone digitale è fisso, la filologia è sempre in movimento. Una volta fissato il canone, come lo si potrà modificare?

Abbiamo attribuito (sulla base del canone PHI) l'identificativo phi0959 a Ovidio, phi0959.phi002 alle sue *Heroides* e dunque il nome uniforme (URN) urn:cts:latinLit:phi0959.phi002:15 alla lettera di Saffo a Faone (la quindicesima).

Immaginiamo di aver costruito tutto un ecosistema di strumenti digitali fondato su questo protocollo e questi identificativi, come una Pauly-Wissowa digitale²⁸, dizionari online, motori di ricerca lessicale e semantica, saggi in formato *ebook* con link diretti al testo e quant'altro.

Tutti questi strumenti funzioneranno sulla base della stabilità del protocollo, cioè del fatto che la lettera dell'infelice Saffo sarà sempre identificata come urn:cts:latinLit:phi0959.phi002:15 (la quindicesima delle *Heroides* di Ovidio).

Cosa succederà se, facendo il suo lavoro di filologo, uno studioso come Richard Tarrant dovesse refutare in modo definitivo – magari confortato da un nuovo papiro – la paternità ovidiana della lettera?

La filologia è storia della tradizione e, ahimè, anche critica del testo. E l'attività critica, con la sua carica eversiva di ogni fissità, non risparmia nessun livello di quelli previsti attualmente dal protocollo CTS («text group», saggia sostituzione per il concetto di autore, «work», «passage»):

1. A livello di «text group», una scoperta papirologica potrebbe farci cambiare idea sull'identificazione o la distinzione di autori (ad esempio, il canone TLG distingue attualmente tra due diversi «Aeschylus Trag.»: tlg0031 e il più noto tlg0085).
2. Per il rapporto tra i livelli «text group» e «work», ho già parlato delle opere di paternità incerta: un altro esempio potrebbe essere il *Culex*, catalogato oggi, in mancanza di un autore certo, come

gie di ampliamento, vedere https://github.com/PerseusDL/catalog_pending/wiki.

²⁸ La monumentale enciclopedia cartacea di riferimento per gli studi classici: Wissowa *et al.* (1893-1980).

urn:cts:latinLit:phi0692.phi003 sotto il «text group» 0692 (*Appendix Vergiliana*), ma che in futuro potrebbe teoricamente essere attribuito a Virgilio.

3. Ai livelli «text group», «work» e «passage», nella tradizione manoscritta può essere anche problematica l'attribuzione di sezioni testuali ad un'opera e ad un autore, sezioni tradite insieme al resto dell'opera nella tradizione manoscritta: la paternità cesariana di intere sezioni del *Bellum civile* è oggetto di discussione tra gli studiosi.
4. Per quanto riguarda i soli livelli «work» e «passage» (all'interno di uno stesso «work group»), alcuni editori dividono quella che un editore di riferimento come Keil 1961 pubblica come un'opera unica, l'*Ars grammatica* di Donato, in una serie di operette autonome. Tra l'altro, il gruppo di lavoro del *Perseus Catalog*, seguendo quest'ultimo orientamento, si avvia ad attribuire a ciascuna un identificativo specifico²⁹.
5. Al livello di «passage», ho già accennato alla divisione di un'opera in sezioni (ad es. l'elegia properziana 2.22, che alcuni dividono in Prop. 2.22a e Prop. 2.22b), ma le aggiunte, espunzioni o trasposizioni di intere porzioni testuali numerate (ad esempio di versi) costituiscono un campo comunissimo di lavoro per la filologia testuale.

Convenzionalità degli URN.

Un argomento forte a favore del CTS è che gli URN (i nomi uniformi, ovvero gli identificativi univoci) sono puramente convenzionali.

Nei canoni di riferimento si trovano già oggi opere considerate spurie catalogate «sotto» i loro (non) presunti autori. La *Nux* pseudo-ovidiana ha come URN urn:cts:latinLit:phi0959.phi013 (laddove phi0959 indica Ovidio), ma il collegamento con Ovidio sarà visibile solo al software che userà questo URN, laddove i metadati (le informazioni collegate all'opera nei cataloghi canonici)

²⁹ Queste le operette elencate nel catalogo (ancora provvisorio, per quanto riguarda questo autore), con l'indicazione dell'URN (l'identificativo assegnato) e della corrispondenza con le pagine dell'edizione di Keil 1961: *Ars Minor* (*De partibus orationis*), URN stoa0110.stoa001 (Keil, pp. 367-392); *De Barbarismo*, stoa0110.stoa003 (pp. 392-393); *De ceteriis vitiiis*, stoa0110.stoa004 (pp. 394-395); *De Metaplasmo*, stoa0110.stoa005 (pp. 395-397); *De Schematibus*, stoa0110.stoa006 (pp. 397-399); *De Solecismo*, stoa0110.stoa007 (pp. 393-394); *De Tropis*, stoa0110.stoa008 (pp. 399-402).

possono benissimo indicare la natura spuria dell'operetta. In effetti, sia il corpus PHI 5.3 su Cd-ROM sia il *Perseus Catalog* riportano «Nux [sp.]» come titolo dell'opera.

Che il *Culex* sia catalogato «al di fuori» di phi0690 (Virgilio) non nega che possa essere considerato virgiliano in un saggio o addirittura nei metadati del catalogo, e viceversa il fatto che la *Nux* sia catalogata «sotto» phi0959 (Ovidio) non implica che essa sia ovidiana.

Versioni del testo.

Quando entra in gioco il livello «passage» (punti 3-5 della lista), viene in soccorso un meccanismo ulteriore di CTS, a mio parere assai apprezzabile: la possibilità di indicare una *versione* (cioè un'edizione) dell'opera e di riferire la numerazione del passaggio al testo specifico di quella edizione. Dei due URN

- (a) urn:cts:latinLit:phi0959.phi013:2
- (b) urn:cts:latinLit:phi0959.phi013.**opp-lat1:2**

il primo indica il secondo verso della *Nux* pseudo-ovidiana, intesa come testo astratto, mentre il secondo indica il secondo verso *dell'edizione del 1879 di Emil Baehrens* di quell'opera. Il blocco aggiuntivo *opp-lat1* identifica infatti tale edizione³⁰.

Interoperabilità o filologia?

Forse portato per indole ad una visione dilemmatica della realtà, vedo qui un paradosso:

- (a) Gli URN del primo tipo offrono il massimo della interoperabilità ed aprono prospettive applicative entusiasmanti: costituiscono infatti la traduzione digitale di notazioni onnipresenti negli studi letterari come «Ps-Ov. *Nux*, 2», e non è difficile prevederne una implementazione

³⁰ Baehrens (1879-1883).

molto diffusa. Il problema è, naturalmente, a *quale* testo della *Nux* ci riferiamo, a quale autore, a quale opera: la questione filologica. Essa si nascondeva anche in «Ps-Ov. *Nux*, 2», ma nel mondo della carta veniva gestita implicitamente dalla competenza e dall'elasticità del lettore/studioso. Come spesso succede, il problema metodologico latente «esplosivo» nel passaggio al digitale: un software costruito per gestire un URN come urn:cts:latinLit:phi0959.phi013:2 (senza indicazione dell'edizione) non funzionerebbe più correttamente se il verso 2 in un'edizione diversa non fosse più numerato come 2, ad esempio a causa di una trasposizione di versi.

- (b) Gli URN del secondo tipo, viceversa, costituiscono il trionfo della filologia: l'identificazione di un passaggio testuale è ancorata ad una specifica versione del testo. Ma evidentemente ciò avviene a scapito dell'interoperabilità: complesse tabelle di concordanze sarebbero necessarie per insegnare al software che il verso 8 di un'edizione della *Nux* corrisponde al verso 2 della «forma canonica» del testo nel corpus PHI 5.3, e magari al verso 5 di un'altra edizione ancora.

CTS: due questioni aperte.

Penso che la nostra generazione di informatici umanisti, mentre costruisce le prossime infrastrutture di citazione digitale – ovvero le strutture portanti dei nuovi archetipi digitali, debba tenere aperte e presenti queste due questioni:

- Ammesso l'argomento forte della convenzionalità degli URN, ci si può comunque chiedere se la *definizione* di autori, opere e attribuzioni autoriali all'interno dell'organizzazione gerarchica del CTS davvero non avrà alcuna influenza culturale, cioè se ogni programmatore, ogni curatore di progetto ed ogni utente avrà consapevolezza piena di quella convenzionalità.
- Quanto alla possibilità di collegare la numerazione delle sezioni testuali (versi, paragrafi ecc.) ad una specifica versione/edizione del testo, vale il dilemma esposto sopra: le implementazioni di tipo (b), più «filologiche», mancano quasi del tutto di quell'interoperabilità che è la ragion d'essere principale di CTS; viceversa, le implementazioni del tipo (a) aprono prospettive entusiasmanti di interoperabilità – e per ciò

sono prevedibilmente destinate a grandissima diffusione – ma rischiano di consegnare alle prossime generazioni il corpus dei testi classici (e forse non solo quello) in una forma fissa, archetipale, basata sugli attuali corpora digitali di riferimento, PHI e TLG.

COPIA PER CONSULTAZIONE

BIBLIOGRAFIA

- AVALLE D.S., 1985. *I canzonieri: definizione di genere e problemi di edizione*. In: *La critica del testo: problemi di metodo ed esperienze di lavoro. Atti del Convegno di Lecce, 22-26 ottobre 1984*. pp. 363-382.
- BAEHRENS E. (ed.), 1879-1883. *Poetae Latini minores 2.2: Ovidi Nux, Consolatio ad Liviam, Priapea*, Teubner, Leipzig.
- BARONI A., 2009. *La grafematica: teorie, problemi e applicazioni*, Tesi di laurea specialistica, Università di Padova. http://unipd.academia.edu/AntonioBaroni/Papers/455456/La_grafematica_teorie_problemi_e_applicazioni
- BATTELLI G., 1949. *Lezioni di paleografia*, Scuola Vaticana di Paleografia, Diplomatica e Archivistica, Libreria Editrice Vaticana.
- CENCETTI G., 1978. *Paleografia latina*. Guide allo studio della civiltà romana, Jouvence.
- CONSTABLE P., CONSTABLE P.G., 2001. *Understanding Unicode: a general introduction to the Unicode standard*. NRSI: Computers & Writing. http://scripts.sil.org/cms/scripts/page.php?itemsc_id=IWS-Chapter04b
- COOMBS J.H., RENEAR A.H., DE ROSE S.J., 1987. *Markup systems and the future of scholarly text processing*. Communications of the Association for Computing Machinery, 30: 933-947. <http://xml.coverpages.org/coombs.html>
- COZZO A., 2002. *Sapere e potere presso i moderni e presso i Greci antichi: una ricerca per lo studio come se servisse a qualcosa*. Mathesis (Università di Palermo. Centro interdipartimentale «Forme di produzione e trasmissione del sapere nelle società antiche e moderne»), Carocci, Roma.
- COZZO A., 2006. *La tribù degli antichisti: un'etnografia ad opera di un suo membro*, Studi e ricerche: Sezione greca, Carocci, Roma.
- DE ROSE S.J., DURAND D.G., MYLONAS E., RENEAR A.H., 1990. *What is text, really?* Journal of Computing in Higher Education, 1.2: 3-26.
- FIORMONTE D., 2003. *Scrittura e filologia nell'era digitale*. Nuova didattica: Arte e letteratura, Bollati Boringhieri, Torino.
- FIORMONTE D., SCHMIDT D., MONELLA P., SORDI P., 2015. *The politics of code. How digital representations and languages shape culture*. In: *Proceedings of the International Society for Information Studies (ISIS) Summit Vienna 2015 - The Information Society at the Crossroads*. International Society for Information Studies (ISIS). <http://infolet.it/files/2015/06/politics-of-code-fiormonte-et-al-def.pdf>
- GENETTE G., 1987. *Seuils*. Éditions du Seuil, Paris.
- HARRIS R., 2003. *La tirannia dell'alfabeto*. Stampa Alternativa & Graffiti, Viterbo.
- IDE N., SPERBERG-MCQUEEN C.M., AMSLER R., WALKER D., HOCKEY S., ZAMPOLLI A., 1988. *Proposal for funding for an initiative to formulate guidelines for the encoding and interchange of machine-readable texts*. NEH Proposal. <http://projects.oucs.ox.ac.uk/teiweb/Vault/SC/scg02.html#N453>

- JENSEN M.S., 1999. *Dividing Homer: When and how were the Iliad and the Odyssey divided into songs?* *Symbolae Osloenses*, 74: 5-91.
- KEIL H., 1961. *Grammatici Latini*. Olms, Hildesheim.
- LEONARDI C., MORELLI M., SANTI F. (eds.), 1994. *Macchine per leggere: tradizioni e nuove tecnologie per comprendere i testi. Atti del Convegno di studio della Fondazione Ezio Franceschini e della Fondazione IBM Italia (Certosa del Galluzzo, 19 novembre 1993)*. Quaderni di cultura mediolatina, Centro italiano di studi sull'Alto medioevo.
- MONELLA P., 2008. *Towards a digital model to edit the different paratextuality levels within a textual tradition*. *Digital Medievalist*, 4. <http://www.digitalmedievalist.org/journal/4/monella/>
- MONELLA P., 2014. *Many witnesses, many layers: the digital scholarly edition of the Iudicium coci et pistoris (Anth. Lat. 199 Riese)*. In: F. CIOTTI (ed.), *Digital Humanities: progetti italiani ed esperienze di convergenza multidisciplinare, Atti del convegno annuale dell'Associazione per l'Informatica Umanistica e la Cultura Digitale (AIUCD) Firenze, 13-14 dicembre 2012*. Sapienza Università Editrice, Rome:173-206. http://digilab2.let.uniroma1.it/ojs/index.php/Quaderni_DigiLab/article/view/190
- MONELLA P. (in stampa), *Why are there no comprehensively digital scholarly editions of classical texts?* In: *IV Incontro di Filologia Digitale, Verona, Italy*. <http://www.unipa.it/paolo.monella/lincei/why.html>
- MORDENTI R., 1999. *Per un'edizione ipertestuale dello Zibaldone Laurenziano di Boccaccio (PLUT. XXIX, 8)*. In: *I nuovi orizzonti della filologia: ecdotica, critica testuale, editoria scientifica e mezzi informatici elettronici. Convegno internazionale (Roma, 27-29 maggio 1998)*. Atti dei Convegni Lincei, 151. Accademia Nazionale dei Lincei.
- MORDENTI R., 2001. *Informatica e critica dei testi*. Informatica e discipline umanistiche, Bulzoni, Roma.
- MORDENTI R., 2007. *L'altra critica: la nuova critica della letteratura tra studi culturali, didattica e informatica*, Meltemi, Roma. http://books.google.it/books?id=IGGrhR6m_mAC
- MORDENTI R., 2011. *Paradosis. A proposito del testo informatico*, Memorie lincee. Scienze mor.,stor.,fil.IX, Accademia Nazionale dei Lincei, Roma.
- NEROZZI P. (ed.), 1997. *Internet e le muse: la rivoluzione digitale nella cultura umanistica*. Mimesis.
- ORLANDI T., 1990. *Informatica umanistica*, Studi superiori, Nuova Italia scientifica, Roma.
- ORLANDI T., 1999. *Linguistica, sistemi, e modelli*. In: *Il ruolo del modello nella scienza e nel sapere (Roma, 27-28 ottobre 1998)*. Roma. <http://www.cmcl.it/orlandi/pubblimodello.html>
- ORLANDI T., 2006. *Edizione digitale sperimentale di Niccolò Machiavelli, De principibus*. CISADU. <http://www.cmcl.it/orlandi/principe/>
- ORLANDI T., 2010. *Informatica testuale. Teoria e prassi*. Laterza, Roma.

- PERRI A., 2009. *Al di là della tecnologia, la scrittura. Il caso Unicode*. Annali dell'Università degli Studi Suor Orsola Benincasa, 2: 725-748.
- RENEAR A., MYLONAS E., DURAND D., 1993. *Refining our notion of what text really is: The problem of overlapping hierarchies*. <http://www.stg.brown.edu/resources/stg/monographs/ohco.html>
- ROSATI G., 1996. *Sabinus, the heroides and the poet-nightingale: some observations on the authenticity of the Epistula Sapphus*. The Classical Quarterly, 46: 207-216.
- SAMPSON G., 1990. *Writing Systems: A Linguistic Introduction*. Stanford University Press.
- SCHMIDT D., 2012. *The role of markup in the digital humanities*. In: M. THALLER (ed.), *The Cologne Dialogue on Digital Humanities 2012*. Universität zu Köln, Cologne Center for eHumanities. http://www.cceh.uni-koeln.de/files/Schmidt_final.pdf
- STOKES P.A., 2011. *Describing handwriting, part IV: Recapitulation and formal model*. DigiPal Blog. <http://www.digipal.eu/blogs/blog/describing-handwriting-part-iv/>
- TARRANT R.J., 1981. *The authenticity of the letter of Sappho to Phaon (Heroides XV)*. Harvard Studies in Classical Philology, 85: 133-153.
- THE UNICODE CONSORTIUM, 2015. *Unicode 8.0*. The Unicode Consortium. <http://www.unicode.org/versions/Unicode8.0.0/>
- THE UNICODE CONSORTIUM, 2016. *Unicode 9.0*. The Unicode Consortium. <http://www.unicode.org/versions/Unicode9.0.0/>
- TIEPMAR J., TEICHMANN C., HEYER G., BERTI M., CRANE G., 2014. *A new implementation for canonical text services*. In: *Proceedings of the 8th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH)*. Association for Computational Linguistics, pp. 1-8.
- WISSOWA G., KROLL W., MITTELHAUS K., ZIEGLER K., GÄRTNER H. (eds.), 1893-1980. *Paulys Realencyclopädie der classischen Altertumswissenschaft. Neue Bearbeitung*, Metzler.

