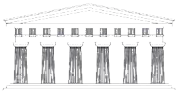
Atti Chiari

CHIAREZZA E CONCISIONE NELLA SCRITTURA FORENSE

a cura di
Riccardo Gualdo
e Laura Clemenzi



Camera Civile di Viterbo
"Carlo Alfonso Pesaresi"
aderente all'Unione Nazionale delle Camere Civili

Via Mazzini, 87 • 01100 Viterbo
Tel 0761 304967 Fax 0761 1760202
www.settecitta.eu • info@settecitta.eu

Impaginazione a cura di *Stefano Frateiacci*
Finito di stampare nel mese di settembre 2021

ISBN: 978-88-7853-928-0
ISBN ebook: 978-88-7853-929-7

Volume pubblicato con il contributo finanziario della Camera Civile di Viterbo e con contributi a valere sul fondo Prin 2017: Progetto di rilevanza nazionale (PRIN) "La chiarezza degli atti del processo (AttiChiari): una base di dati inedita per lo studioso e il cittadino" (Ministero dell'Istruzione dell'Università e della Ricerca, Prot. 2017BSECYX)

Nota iniziale

Questo volume raccoglie i contributi presentati nel seminario telematico "Atti Chiari. Chiarezza e concisione nella scrittura forense", organizzato il 20 maggio 2021 in collaborazione tra la Camera Civile di Viterbo e l'Università degli studi della Tuscia.

Il seminario non avrebbe avuto luogo se non fosse stato promosso e sostenuto con gentile energia da Rosita Ponticiello, Presidente della Camera Civile di Viterbo; grazie alla sua iniziativa è stato ospitato dal sistema documentale digitale "Plusplus 24 Diritto" del "Gruppo 24 Ore" e hanno potuto assistervi oltre 100 tra professionisti e studiosi da tutta Italia.

La pubblicazione si è avvalsa del sostegno finanziario della Camera Civile di Viterbo e dei fondi del dipartimento D.I.R.A.A.S. dell'Università di Genova (progetto P.R.I.N. "La chiarezza degli atti del processo (AttiChiari): una base di dati inedita per lo studioso e il cittadino" - Prot. 2017BSECYX, coordinato a livello nazionale dalla Prof. Jacqueline Visconti).

Agli enti finanziatori va il grazie sincero dei curatori.

Mentre allestivamo questo volume, ferveva la discussione politica e tecnica sulla riforma del processo civile e penale. Siamo convinti che la chiarezza e la sinteticità degli atti processuali sono la premessa per una giustizia più rapida – o più "liquida", come si usa anche dire tra gli esperti – e più comprensibile; e una giustizia più comprensibile è anche una giustizia più giusta.

I contributi, primo assaggio di un ampio e innovativo progetto di analisi linguistica degli atti di parte, cioè delle scritture che formano l'ossatura testuale del processo, insistono soprattutto su una fase preliminare quanto essenziale: la predisposizione dei testi in formato digitale anonimizzato, a garanzia della riservatezza delle informazioni. I dettagli tecnici, tuttavia, non trascurano questioni giuridiche e aspetti linguistici rilevanti, di natura sintattica, testuale e lessicale.

Insieme a Jacqueline Visconti, che guida il progetto Atti Chiari, ringraziamo ancora Rosita Ponticiello e chi ha partecipato al seminario del 20 maggio 2021: il Magnifico Rettore dell'Università degli studi della Tuscia, Prof. Stefa-

no Ubertini, e il Prof. Saverio Ricci, Direttore del dipartimento D.I.S.T.U., la Dottoressa Maria Rosaria Covelli, allora Presidente del Tribunale di Viterbo, gli Avvocati Salvatore Donadei e Stefano Brenciaglia, rispettivamente Coordinatore della Commissione Linguistica e Diritto dell'U.N.C.C. e Presidente dell'Ordine degli avvocati di Viterbo.

Riccardo Gualdo, Laura Clemenzi
Viterbo, 1 settembre 2021

Sommario

Digitalizzazione e marcatura XML degli atti

Daniele Fusi (Università Ca' Foscari Venezia)

Benché il mio breve intervento sia focalizzato sul processo di anonimizzazione, esso intende anzitutto evidenziare l'importanza della struttura nell'organizzazione dei dati testuali, che rende possibile farne la base di veri strumenti di analisi. Il processo di anonimizzazione qui illustrato è infatti funzionale in primo luogo al progetto Atti Chiari, e dunque a un uso linguistico, ma senza trascurare eventuali risvolti di natura più contenutistica in ambito giuridico. In quanto destinato a preparare i testi che nutriranno il *corpus* alla base dell'analisi anzitutto in vista della loro anonimizzazione, il *software* è disegnato per funzionare in un ambito totalmente disconnesso, all'interno del computer di ogni anonimizzatore, sicché è necessariamente multipiattaforma. La sua particolare natura tuttavia si deve alla sua collocazione in un quadro più ampio, dove emerge non solo la primaria esigenza di spogliare i testi di ogni dato personale, ma anche quella di dotarli di una struttura semantica adeguata alla successiva analisi.

1. Anonimizzazione

In un tradizionale processo di anonimizzazione, la procedura essenzialmente consiste nel cancellare ogni dato personale dal documento originale, ottenendo come risultato un documento costellato di lacune[1]. Questo naturalmente pregiudica la sua leggibilità, e a maggior ragione la possibilità di utilizzarlo con profitto per analisi linguistiche. Tale situazione è in effetti la conseguenza di una totale mancanza di granularità nel processo: ogni dato o si conserva, o viene obliterato, senza possibilità di altri trattamenti.

Una maggiore granularità invece potrebbe consentire di adattare il risultato di questo processo a innumerevoli usi diversi, ciascuno dei quali ha le proprie esigenze. Ad esempio, un testo "mutilo" come quello prodotto da una anoni-

[1] Cfr. F. Candrilli, *Il progetto di archiviazione e anonimizzazione*, in questo volume, pp. 19-28.

mizzazione tradizionale potrebbe anche essere sufficiente per effettuare spogli statistici, o costruire una banca dati giuridica. Se tuttavia si desidera fare uno studio linguistico, occorre anzitutto riempire le lacune: ovvero non semplicemente cancellare i dati personali, ma sostituirli con qualcosa di formalmente simile, capace però nello stesso tempo di rimuovere l'informazione originale. Inoltre, sempre in un contesto linguistico potrebbe essere utile distinguere i nomi propri presenti nel testo, o indicare la presenza di lingue diverse al suo interno (ad esempio una forma latina o inglese).

In alcuni casi potrebbe anche essere utile procedere oltre su questa strada, ad esempio individuando le abbreviazioni. Questo infatti anzitutto rende più leggibile il testo per i non specialisti, consentendo di indicare le corrispondenti forme sciolte, e ha il vantaggio di fornire alla macchina un testo dove tutte le parole sono indicizzabili per esteso, evitando così di inquinare gli indici con forme surrettizie derivanti dalla loro abbreviazione. In aggiunta, questo scioglimento implica un altro notevole vantaggio sul piano più propriamente formale, ovvero la possibilità di fare affidamento sull'interpunzione per individuare con buona approssimazione i confini di frase. Dato che infatti spesso le forme abbreviate comprendono dei punti, codificati digitalmente allo stesso modo del punto che termina una frase, un'analisi puramente formale non avrebbe modo di disambiguare il ruolo di un punto (parte dell'abbreviazione, o termine della frase?). Laddove invece le abbreviazioni siano individuate, ed eventualmente sciolte, questo consentirà alla macchina di distinguere il ruolo del punto, e per conseguenza di trattarlo come terminatore di frase dove opportuno, al pari di altri segni che non presentano questa ambiguità, come il punto interrogativo o esclamativo.

In uno strumento di analisi linguistica infatti la possibilità di determinare i confini di frase non va sottovalutata, dato che consente di limitare la ricerca di più parole a un contesto sintatticamente determinabile. Ad esempio, ovunque si voglia esaminare ogni co-occorrenza di due o più parole, questo implica necessariamente la definizione di un contesto all'interno del quale le parole trovate si debbano considerare come rilevanti. Un meccanismo efficace ma più superficiale può essere qui il mero computo delle parole, definendo una distanza massima fra due di esse: in caso contrario, otterremmo fra i risultati della ricerca anche due parole poste ai due estremi di un intero documento. Se però il nostro interesse riguarda specificamente l'occorrenza in un contesto sintatticamente determinato, come la frase, è ovvio che anche due parole consecutive nel testo potrebbero ben appartenere a due frasi distinte (l'ultima parola di una frase seguita dalla prima della frase successiva). La possibilità di determinare almeno i confini sintattici maggiori utilizzando l'interpunzione definisce inve-

ce proprio un contesto sintattico, benché basato su indicatori puramente formali[2]; e per conseguenza diviene possibile limitare la ricerca di co-occorrenze non con un mero conteggio, ma all'interno della stessa frase.

Come si vede quindi, anche un aspetto apparentemente banale come un'abbreviazione può avere implicazioni in funzione del tipo di uso che si desidera fare del testo; sicché, al di là di questo esempio, appare evidente che ogni uso può imporre una serie di requisiti assai specialistici al processo di anonimizzazione. In questo ambito diviene quindi necessaria proprio una maggiore granularità nel trattamento dei dati personali presenti nel testo, che faccia posto a una serie di sfumature fra i due estremi di cancellazione o conservazione.

Ad esempio, un cognome presente in un testo potrebbe essere cancellato, ma anche sostituito con un altro fittizio; parimenti, un nome di battesimo potrebbe non solo essere cancellato o sostituito, ma anche sostituito in modo più articolato, assicurandosi non solo che allo stesso nome dell'originale corrisponda sempre lo stesso nome fittizio, ma anche conservandone il genere maschile o femminile. Questo infatti evita di turbare l'accordo grammaticale nel contesto del nome, oltre a non alterare in modo grossolano gli aspetti essenziali della vicenda trattata.

Sempre su questa via, una targa, un codice fiscale, un numero telefonico, un indirizzo *e-mail*, o qualsiasi altro dato personale rappresentato in forma alfanumerica potrebbe semplicemente essere cancellato, o magari sostituito con una serie di trattini; o piuttosto essere trasformato in modo da rimuovere la sua informazione, pur conservandone l'aspetto formale. Generalizzando, è possibile ottenere questo secondo risultato semplicemente sostituendo in un codice alfanumerico ogni lettera con una lettera casuale e diversa, e ogni cifra con una casuale e diversa. Ad esempio, una targa come AB123CD potrebbe diventare RZ308MP, dove la procedura di anonimizzazione ha creato un nuovo codice sostituendo ogni carattere con uno della stessa classe, scelto in modo casuale ma facendo attenzione che sia sempre diverso rispetto a quello originale. Il vantaggio per la leggibilità del testo qui è enorme, benché l'informazione originale sia completamente rimossa.

Gli esempi potrebbero continuare, ma dovrebbero essere sufficienti a mostrare come una maggior granularità nel processo di anonimizzazione consenta di ottenere un testo formalmente completo e leggibile, pur eliminando ogni dato personale dal suo interno. In tal modo diviene possibile adattare il processo di anonimizzazione agli scopi per cui viene costituito il *corpus* te-

[2] Analisi più complesse basate su marcature sintattiche con tecniche semiautomatiche (*tree tagging*) sono spesso uno sforzo eccessivo rispetto alle risorse del progetto e ai suoi scopi.

stuale, ottenendo dei testi formalmente indistinguibili dagli originali, anche se totalmente sicuri, piuttosto che documenti mutili e scarsamente utilizzabili per analisi formali.

Si tratta dunque di valutare quale approccio possa garantire la granularità necessaria a questo scopo. In effetti, con un'inversione solo apparentemente paradossale, il processo di anonimizzazione qui proposto aggiunge, piuttosto che togliere, informazione. In altri termini, l'operatore non fa che lavorare in un semplice applicativo di videoscrittura come Word, limitandosi a segnalare tutti gli elementi che in un testo corrispondono a dati personali, e indicandone nel contempo anche la categoria generale: una porzione di testo viene quindi marcata come indicativa di un cognome, o di un prenome femminile, o di un codice alfanumerico, ecc. Sul piano pratico, questo significa introdurre una leggerissima marcatura direttamente nel testo originale: nel nostro progetto essa consiste in un paio di parentesi graffe, caratteri certo non presenti nei documenti di origine, all'interno delle quali un breve prefisso identifica la categoria generale cui appartiene il testo racchiuso tra le parentesi. Ad esempio, si consideri questo stralcio di testo:

```
il signor {a-f-m:Mario} {a-l:Verdi}, nato a {t:Roma} il
{d:1/2/1970}, c.f. {u: VRDMRA70B01H501N}
```

Qui come si vede le graffe servono a marcare "Mario" come antroponimo maschile (*anthroponym, first name, male*: a-f-m), "Verdi" come cognome (*anthroponym, last name*: a-l), "Roma" come toponimo, e "VRDMRA70B01H501N" come un codice alfanumerico. L'opera dell'anonimizzatore termina qui: di fatto, qui lavora piuttosto come marcatore, che aggiunge informazione al documento, attribuendogli quel tanto di struttura utile agli scopi del progetto. A questo stadio, non si ha altro che il testo originale, semplicemente marcato; l'informazione personale viene eliminata in una seconda fase, e in modo diverso a seconda degli scopi. Proprio questa distinzione nelle fasi del processo consente di adattarlo di volta in volta alle diverse necessità, in una scala di granularità molto più fine.

Il sistema di anonimizzazione qui illustrato opera dunque su testi marcati in questo modo, e ha un'architettura fortemente modulare, che garantisce la sua versatilità: a ognuna delle marcature corrisponde un modulo *software* specializzato nel trattare i dati in essa contenuti, o anche più di un modulo, qualora si voglia offrire la scelta di diversi trattamenti a seconda degli scopi.

In particolare, l'anonimizzatore dispone di due grandi classi di moduli: alcuni sono deputati al trattamento dei contenuti marcati, a seconda del marcatore

utilizzato; altri invece hanno il solo compito di generare una mappatura sistematica fra ogni nome proprio e il suo corrispondente nome fittizio scelto in modo casuale, rispettando la distribuzione delle occorrenze dell'originale. Questi mappatori attingono quindi a una serie di repertori, desunti da varie risorse digitali e comprendenti decine di migliaia di nomi italiani, raggruppati per categoria: prenomi maschili, prenomi femminili, cognomi, e nomi di città. Tali repertori sono stati opportunamente pretrattati da altri programmi realizzati *ad hoc*, che ad esempio eliminano tutti gli omonimi distribuiti fra categorie diverse, in modo che un prenome che possa essere sia femminile che maschile (per esempio *Andrea*) venga eliminato, evitando così ogni ambiguità formale.

Ciascuno dei moduli segue la propria logica nel trasformare la porzione di testo marcata, sicché l'insieme dei moduli opera in modo concertato per produrre un testo formalmente completo, e calcato esattamente su quello originale, ma dove ogni dato personale è stato rimosso. Si consideri ad esempio questo testo fittizio (il grassetto è nell'originale):

```
il signor {a-f-m:Mario} {a-l:Verdi}, nato a {t:Roma} il
{d:1/2/1970}, c.f. {u: VRDMRA70B01H501N}, impiegato presso la
ditta {j-f:ACME}, con autovettura targata {u:FO392FI} , reca-
tosi {f-lat:de relato} in ritardo al lavoro
```

In esso, le marcature individuano diverse categorie di dati personali, nell'ordine:

1. antroponimo maschile (Mario).
2. cognome (Verdi).
3. toponimo (Roma).
4. data (1/2/1970). La data all'interno della marcatura può essere espressa nei modi più vari, dal semplice giorno e mese in lettere o cifre alla forma completa di anno, con vari formati. Il modulo incaricato di trattare le date ha la competenza necessaria per analizzare le diverse forme in cui si trovano tipicamente espresse le date, desumendo automaticamente il valore di giorno, mese, e anno (ad esempio, "30-5-1970", o "30/5/1970", o "30.5.1970" o "30 mag. 1970", o "30 maggio 1970", "30 MAGGIO 1970", "30 maggio", ecc.).
5. codice fiscale, appartenente alla più generale categoria dei codici alfanumerici.
6. nome di persona giuridica (ACME).
7. targa, un altro codice alfanumerico.

8. forestierismo, in tal caso un'espressione latina (*de relato*; lat è il codice standard ISO-639 per la lingua latina).

Una volta che il testo passi attraverso l'anonimizzatore, esso viene trasformato in modo casuale ottenendo un risultato formalmente identico. Si confrontino il testo originale con quello anonimizzato disposti uno a fianco all'altro:

il signor **Mario Verdi**, nato a **Roma** il 1/2/1970, c.f. VRDMRA70B01H501N, impiegato presso la ditta **ACME**, con autovettura targata FO392FI , recatosi de relato in ritardo al lavoro	il signor **Carlo Rossi**, nato a **Como** il 9/7/1984, c.f. NKEBWS85H97M356P, impiegato presso la ditta **RAZZI**, con autovettura targata GM413HY, recatosi de relato in ritardo al lavoro

Si noti qui che mentre l'essenziale stile tipografico del testo (grassetti, corsivi, paragrafi, ecc.) viene conservato, ogni dato personale è stato rimosso, pur mantenendo una chiara identità formale. Ogni nome proprio viene sostituito con un altro fittizio della stessa classe, avendo cura di mantenere sempre la stessa corrispondenza: tutte le occorrenze di "Mario" saranno sostituite sempre da (in questo esempio) "Carlo", e così via.

I dati alfanumerici poi sono tutti formalmente simili, anche se privati di ogni contenuto personale. Si tratta naturalmente di una corrispondenza superficiale, ma di norma sufficiente a garantire la leggibilità del testo, e per conseguenza la sua analisi linguistica. Ad esempio, è naturalmente ovvio che un codice fiscale come quello trasformato non rispetti i criteri della sua formazione: qui il codice fiscale è marcato come un generico tipo alfanumerico, sicché il modulo corrispondente non fa che sostituire ogni lettera con un'altra casuale, e ogni cifra con un'altra casuale. Di fatto, per tutti gli usi cui possono essere adibiti questi testi non ha alcuna rilevanza che il codice fiscale non sia effettivamente valido; quanto importa è solo che formalmente esso continui ad apparire come un codice fiscale, sì da lasciare la leggibilità del testo inalterata.

Peraltro, alcuni marcatori, introdotti per evidenti scopi linguistici, come quello che identifica un latinismo, neppure alterano il loro testo: nell'esempio *de relato* rimane intatto, in quanto non si tratta di un dato personale. Tuttavia, è utile che i marcatori impegnati nell'anonimizzazione introducano nel contempo anche questa marcatura, perché consente di distinguere in modo economico lingue diverse all'interno del testo. Naturalmente questo è un tipo

di marcatura essenziale all'analisi linguistica, ma che potrebbe essere tranquillamente omesso in una destinata ad altri scopi.

Il livello di granularità qui è molto fine, e potrebbe consentire con altrettanta facilità altri trattamenti. Il caso della marcatura della lingua è un esempio di specializzazione del sistema a usi linguistici, ma già il fatto che la mappatura dei nomi sia costante nello stesso contesto è la conseguenza di un'esigenza di leggibilità che va oltre il livello puramente formale, e serve non solo a favorire un minimo di coerenza contenutistica alla lettura (un testo dove il nome della stessa persona cambiasse ogni volta che viene citata risulterebbe oscuro), ma anche a consentire ricerche più orientate alla sostanza giuridica che alla forma linguistica. Grazie a questa mappatura sistematica e coerente infatti rimane comunque possibile seguire la vicenda descritta; ma su questa via di adattamento a un tipo di analisi più contenutistica si può proseguire, introducendo nuovi moduli per ottenere un risultato finale diverso.

Ad esempio, si pensi alla data: si è visto che il modulo predefinito, cui interessa solo mantenere una data formalmente uguale quale che sia la sua espressione, non fa che analizzare il contenuto per desumerne anno (quando presente), mese e giorno, per poi sostituire ogni valore con un altro casuale. Con ciò si garantisce la perfetta corrispondenza formale con l'originale, senza alcuna preoccupazione per la cronologia: le date generate sono completamente casuali. Nel caso però di un'analisi giuridica, l'anonimizzatore dovrebbe comportarsi in modo diverso, per conservare la cronologia relativa dei fatti; in caso contrario ci si troverebbe dinanzi a un fatto successivo a un altro ma datato prima, o a un genitore o pensionato dell'età di 5 anni, ecc. In tal caso, un secondo modulo può essere introdotto per generare un risultato che rispetti la cronologia originale, pur oscurandola: ad esempio sottraendo a ogni anno un numero casuale (sempre lo stesso in tutto il contesto di anonimizzazione), e lasciando inalterati giorni e mesi. Questi infatti non presentano sostanzialmente problemi per i dati personali, trattandosi di indicazioni troppo generiche: ad esempio, "30 maggio" senza specificazione di anno non ha rilevanza, mentre "30 maggio 2021" viene trasformato sottraendo un numero casuale all'anno, magari 18, per cui diventa "30 maggio 2003". In tal modo otteniamo di preservare fedelmente tutte le distanze cronologiche relative, pur rimuovendo ogni dato a rischio. Alterare anche il giorno o il mese non solo non sarebbe utile all'anonimizzazione, ma rischierebbe di produrre incongruenze sul piano cronologico: se ad esempio volessimo sottrarre 18 al giorno, ma il giorno fosse 5, si finirebbe al mese precedente, con potenziali conseguenze distruttive sulla cronologia relativa dell'intero contesto dei documenti.

Sempre in questo ambito, a proposito del contesto di mappatura, dove è necessario mantenere sempre le medesime corrispondenze per ogni nome sostituito e sempre la stessa alterazione delle date conservando la cronologia relativa, va peraltro osservato che il contesto può essere rappresentato dal singolo documento, ma anche da un insieme di documenti. Questo può ad esempio essere il caso di vari atti appartenenti tutti allo stesso procedimento, per cui il "Mario Verdi" del primo documento sostituito con "Carlo Rossi" deve continuare a essere sostituito con "Carlo Rossi" per tutta la serie dei documenti. Ancora una volta, questa è un'esigenza specifica delle analisi giuridiche, priva di rilevanza per l'analisi linguistica; sicché può essere applicata a discrezione dell'utente del programma. Il sistema infatti offre entrambe le possibilità: o trattare come un unico contesto tutti i documenti processati nella stessa sessione; o trattare ogni documento come un contesto separato.

Questi esempi mostrano quindi come lo stesso sistema possa essere utilizzato per generare testi diversi, a seconda degli scopi per cui viene costruito il loro *corpus*. Questo sistema potrebbe infatti anche effettuare anonimizzazioni tradizionali (cfr. quanto detto sopra), dove ogni dato marcato come personale viene cancellato, o sostituito con una qualche sequenza fissa di caratteri; ma può essere usato invece per una anonimizzazione più conservativa, che piuttosto che cancellare sostituisce i dati personali trattandoli però in modo diverso a seconda della loro categoria, sì da produrre un documento formalmente ineccepibile e calcato sull'originale, ma del tutto anonimizzato.

Ancora, si potrebbe voler soddisfare un sovrainsieme delle esigenze di un *corpus* destinato all'analisi linguistica per generarne uno destinato all'analisi giuridica. In tal caso potremmo voler conservare la cronologia relativa dei fatti, mantenere le stesse corrispondenze di nomi in contesti di più documenti quando necessario, conservare l'ordine di grandezza delle cifre di denaro, e così via, su una scala continua di granularità. Un unico sistema, a partire da un unico insieme di documenti originali, può così generare un qualsiasi numero di documenti anonimizzati diversi per tipo e scopi della procedura adottata.

Di più, anche all'interno della stessa destinazione d'uso si possono individuare diversi livelli di raffinatezza nei trattamenti. Ad esempio, pur restando nell'ambito formale dell'analisi linguistica un'opzione utile nel caso dell'anonimizzazione dei nomi è rappresentata da una mappatura più attenta ai fatti di concordanza. Non ci si limita cioè a selezionare un nome fittizio maschile per uno originale maschile, e femminile per uno originale femminile; ma, per qualsiasi nome proprio, un ulteriore raffinamento può cambiare il meccanismo di selezione del nome fittizio a partire dai repertori. Questo avviene nel caso dei nomi propri (siano essi antroponimi, toponimi, o nomi di persona giuridi-

ca) inizianti per vocale. Il modulo di base per la sostituzione dei nomi non fa che pescare un nome a caso da un repertorio, assicurandosi solo che sia diverso da quello originale. Un modulo più raffinato invece esamina il nome originale per garantire un migliore rispetto delle concordanze nel caso appunto di nomi inizianti per vocale. Potrebbe infatti darsi il caso limite in cui ad esempio un nome come Arezzo venga sostituito da uno come Roma, quando però la scelta della preposizione che lo precede fosse condizionata dalla sua forma: "ad Arezzo" con una <d> eufonica diverrebbe così "ad Roma", il che naturalmente genererebbe un palese errore formale. Per evitare questo problema, il modulo più raffinato non opera una selezione indiscriminata, ma nel caso il nome originale inizi per vocale limita il repertorio ai soli nomi inizianti per vocale, ad esempio "Ortona". Il risultato per il nostro esempio sarebbe così "ad Ortona", formalmente valido.

Su questa via si può anche decidere di procedere oltre sul piano più prettamente linguistico, restringendo ulteriormente la selezione a un nome casuale che inizi non solo per vocale, ma anche per la stessa vocale: nel caso di Arezzo quindi potrebbe essere selezionato Avellino, donde "ad Avellino". Questa modifica, non necessaria sul piano puramente formale, potrebbe infatti essere rilevante per studi molto specialistici relativi a determinati fenomeni come lo iato, dove spesso la scelta dello scrivente è condizionata da considerazioni stilistiche. Alcuni infatti possono seguire una prassi indiscriminata e considerabile più scorretta, dove in modo meccanico viene evitato qualsiasi incontro di vocali, indipendentemente dalla loro qualità. In tal caso, si finisce per usare "ad" davanti a qualsiasi vocale, sia essa uguale o diversa da <a>, o dotata di un grado di apertura elevato o ridotto. Altri invece possono seguire la prassi consigliata e meno meccanica di limitare la <d> eufonica ai casi di vocali uguali, il che come noto non implica affatto un supposto effetto cacofonico, non solo in virtù del diverso livello di apertura; in alcuni casi anzi proprio la <d> "eufonica" potrebbe condurre a effetti di solito poco piacevoli per il nostro orecchio, come l'allitterazione nel caso di "ad Udine". In questo contesto è dunque chiaro che la selezione di un nome fittizio in sostituzione di quello casuale non deve essere totalmente libera (donde errori come "ad Roma"), ma neppure semplicemente fermarsi alla categoria di vocale; piuttosto, laddove un nome da sostituire inizi per vocale, esso andrà sostituito con uno iniziante per la *stessa* vocale. Il fatto che i repertori contano comunque decine di migliaia di forme rende in ogni caso sicura anche questa soluzione: il risultato è un testo completamente anonimizzato, né peraltro il lettore finale conosce i dettagli operativi degli algoritmi di anonimizzazione, o i parametri con cui sono stati applicati.

Gli esempi potrebbero continuare, ma sono sufficienti a mostrare come dinanzi a una molteplicità di esigenze anche molto specifiche dettate dall'uso del *corpus* come strumento di analisi il processo di anonimizzazione non può limitarsi a un *aut aut*, ma piuttosto adattarsi a diversi livelli di granularità richiesti di volta in volta da ogni specifica categoria di dato.

Questa procedura di anonimizzazione peraltro si coniuga con un sistema di trasformazione dei testi originali capace di offrire un'altra essenziale caratteristica per il *corpus*, benché dalla prospettiva qui esaminata essa possa apparire una sua mera ricaduta: la conversione automatica in formato XML TEI[3], che al momento costituisce essenzialmente lo standard di fatto per la codifica digitale di testi di interesse umanistico. Non è ovviamente questo il luogo per anche solo far cenno a questa tecnologia e al suo inquadramento nel contesto delle *Digital Humanities*; basterà quindi introdurla in quanto parte della procedura di anonimizzazione che costituisce l'argomento di questo breve intervento.

2. *Flusso generale*

Come si è visto, il punto di partenza per la costituzione di questo *corpus* digitale è rappresentato essenzialmente da documenti prodotti da programmi di videoscrittura, o da documenti PDF che ne derivano. Quale che sia il dettaglio del loro formato, internamente i documenti sono anzitutto livellati su un unico formato di partenza, DOCX, utilizzato dal più popolare applicativo di videoscrittura (Word)[4]. Questo consente da un lato di offrire la migliore esperienza agli anonimizzatori, che devono semplicemente aggiungere le loro marcature nel testo ricevuto utilizzando l'applicativo di videoscrittura loro più familiare; e insieme di fornire al *software* il punto di partenza, a base XML, per la trasformazione in XML TEI.

Nella Fig. 1, che mostra il flusso generale dei dati, i documenti Word rappresentano quindi il punto di partenza, siano essi derivati da altri formati o direttamente acquisiti in questo (come avviene nella maggior parte dei casi). A partire da questi documenti gli anonimizzatori applicano le marcature necessarie, per poi avviare il programma per il contesto desiderato (uno o più documenti per sessione).

[3] Cfr. https://tei-c.org.

[4] Come è noto, da tempo Microsoft ha abbandonato il formato proprietario nativo DOC, per abbracciare un formato aperto e descritto in un apposito standard a base XML (ECMA 376), DOCX.

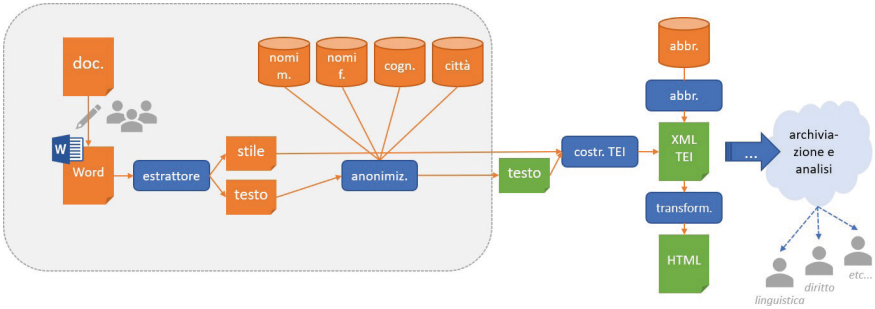


Fig. 1 - Flusso generale dei dati: il riquadro in grigio delimita l'area protetta all'interno del flusso, dalla quale nessun dato personale può uscire

Un aspetto essenziale da considerare a questo stadio è l'enorme ridondanza dell'informazione tipografica presente nei documenti di partenza, che male si adatta a un trattamento automatico finalizzato all'anonimizzazione e a una successiva analisi. Di fatto, anche un frammento di frase come quello riportato nella Fig. 2 nel formato XML che lo codifica appare densissimo di informazione tipografica che è non solo irrilevante agli scopi della presente ricerca, ma anche di ostacolo[5].

del **dott. Immacolato PAOLETTI**, già Commissario Giudiziale del Concordato Preventivo Desolina in liq., elettivamente domiciliato in Invorio, Giodoco Gottifredi, 68, presso lo studio dell'avv. Ado Liguoro (c.f.codice fiscale HMVMQY03G67Y621X; fax 389-3174293; pec: hp1462@outlook.it) che lo rappresenta ed assiste come da procura in calce alla presente comparsa

Fig. 2 - Frammento di testo Word

In effetti, anche senza conoscere i dettagli di questa codifica è evidente dalla Fig. 3, nella pagina seguente, che il breve testo riportato nella Fig. 2 risulta disperso in un mare di marcatori (il testo racchiuso fra <>), che per lo più rappresentano dettagliate informazioni tipografiche (tipi di carattere, dimensioni della pagina, margini, interlinea, ecc.) di dubbia utilità per un'analisi linguistica.

Anzitutto dunque il programma di anonimizzazione ha il compito di estrarre da questo formato il solo testo con i suoi stili essenziali, riducendo così enormemente il rumore associato al documento originale. A questo scopo, il programma incorpora funzionalità di un mio sistema sviluppato per un altro

[5] Naturalmente tutti i documenti qui riportati sono in realtà già anonimizzati. Data la loro equivalenza formale agli originali la trattazione non ne risente.

```
<w:document xmlns:wpc="http://schemas.microsoft.com/office/word/2010/wordprocessingCanvas"
    xmlns:cx="http://schemas.microsoft.com/office/drawing/2014/chartex" xmlns:cx1="http
    ://schemas.microsoft.com/office/drawing/2015/9/8/chartex" xmlns:cx2="http://schemas
    .microsoft.com/office/drawing/2015/10/21/chartex" xmlns:cx3="http://schemas.microsoft.com
    /office/drawing/2016/5/9/chartex" xmlns:cx4="http://schemas.microsoft.com/office/drawing
    /2016/5/10/chartex" xmlns:cx5="http://schemas.microsoft.com/office/drawing/2016/5/11
    /chartex" xmlns:cx6="http://schemas.microsoft.com/office/drawing/2016/5/12/chartex" xmlns
    :cx7="http://schemas.microsoft.com/office/drawing/2016/5/13/chartex" xmlns:cx8="http
    ://schemas.microsoft.com/office/drawing/2016/5/14/chartex" xmlns:mc="http://schemas
    .openxmlformats.org/markup-compatibility/2006" xmlns:aink="http://schemas.microsoft.com
    /office/drawing/2016/ink" xmlns:am3d="http://schemas.microsoft.com/office/drawing/2017
    /model3d" xmlns:o="urn:schemas-microsoft-com:office:office" xmlns:r="http://schemas
    .openxmlformats.org/officeDocument/2006/relationships" xmlns:m="http://schemas
    .openxmlformats.org/officeDocument/2006/math" xmlns:v="urn:schemas-microsoft-com:vml"
    xmlns:wp14="http://schemas.microsoft.com/office/word/2010/wordprocessingDrawing" xmlns:wp
    ="http://schemas.openxmlformats.org/drawingml/2006/wordprocessingDrawing" xmlns:w10="urn
    :schemas-microsoft-com:office:word" xmlns:w="http://schemas.openxmlformats.org
    /wordprocessingml/2006/main" xmlns:w14="http://schemas.microsoft.com/office/word/2010
    /wordml" xmlns:w15="http://schemas.microsoft.com/office/word/2012/wordml" xmlns:w16cex
    ="http://schemas.microsoft.com/office/word/2018/wordml/cex" xmlns:w16cid="http://schemas
    .microsoft.com/office/word/2016/wordml/cid" xmlns:w16="http://schemas.microsoft.com/office
    /word/2018/wordml" xmlns:w16sdtdh="http://schemas.microsoft.com/office/word/2020/wordml
    /sdtdatahash" xmlns:w16se="http://schemas.microsoft.com/office/word/2015/wordml/symex"
    xmlns:wpg="http://schemas.microsoft.com/office/word/2010/wordprocessingGroup" xmlns:wpi
    ="http://schemas.microsoft.com/office/word/2010/wordprocessingInk" xmlns:wne="http
    ://schemas.microsoft.com/office/word/2006/wordml" xmlns:wps="http://schemas.microsoft.com
    /office/word/2010/wordprocessingShape" mc:Ignorable="w14 w15 w16se w16cid w16 w16cex
    w16sdtdh wp14"><w:body><w:p w14:paraId="2E7F2207" w14:textId="771FD352" w:rsidR="0018381E"
    w:rsidRPr="0018381E" w:rsidRDefault="0018381E" w:rsidP="0018381E"><w:pPr><w:ind w:right
    ="1655"/><w:jc w:val="both"/></w:pPr><w:r w:rsidRPr="0018381E"><w:t xml:space="preserve"
    >del </w:t></w:r><w:r w:rsidRPr="0018381E"><w:rPr><w:b/><w:bCs/></w:rPr><w:t>dott.
    Immacolato PAOLETTI</w:t></w:r><w:r w:rsidRPr="0018381E"><w:t>, già Commissario Giudiziale
    del Concordato Preventivo Desolina in liq., elettivamente domiciliato in Invorio, Giodoco
    Gottifredi, 68, presso lo studio dell'avv. Ado Liguoro (c.f.codice fiscale
    HMVMQY03G67Y621X; fax 389-3174293; pec: hp1462@outlook.it) che lo rappresenta ed assiste
    come da procura in calce alla presente comparsa</w:t></w:r></w:p><w:sectPr w:rsidR
    ="0018381E" w:rsidRPr="0018381E"><w:pgSz w:w="11906" w:h="16838"/><w:pgMar w:top="1440" w
    :right="1440" w:bottom="1440" w:left="1440" w:header="708" w:footer="708" w:gutter="0"/><w
    :cols w:space="708"/><w:docGrid w:linePitch="360"/></w:sectPr></w:body></w:document>
```


Fig. 3 - Codice XML (ridotto) del documento Word della Fig. 2

progetto[6], producendo dei nuovi *file* a partire dai documenti Word. Il formato di questi *file* è un altro dialetto XML, che rappresenta semplicemente da un lato il testo nei suoi blocchi (i paragrafi, e al loro interno le aree diversamente formattate; Fig. 4), e dall'altro un sottoinsieme dei suoi aspetti tipografici (Fig. 5).

```
<pick>
  <par nr="1" fmtId="1">
    <run nr="1" fmtId="1">del </run>
    <run nr="2" fmtId="2">dott. Immacolato PAOLETTI</run>
    <run nr="3" fmtId="1">, già Commissario Giudiziale del
        Concordato Preventivo Desolina in liq.,
        elettivamente domiciliato in Invorio, Giodoco
        Gottifredi, 68, presso lo studio dell'avv. Ado
        Liguoro (c.f.codice fiscale HMVMQY03G67Y621X; fax
        389-3174293; pec: hp1462@outlook.it) che lo
        rappresenta ed assiste come da procura in calce
        alla presente comparsa</run>
  </par>
</pick>
```

Fig. 4 - Codice XML con estrazione di blocchi di testo dal documento della Fig. 2

[6] Per una sommaria illustrazione di questo strumento, relativo al recupero digitale di documenti di testo, rimando a D. Fusi, *Recovering Legacy in the Digital World: Tales and Tools*, in "Ratione Rerum", 12 (2018), pp. 225-230.

```
<formattings>
  <fmt id="1">
    <pr name="rFonts" value="Calibri" />
    <pr name="sz" value="22" />
    <pr name="pIndentRight" value="1655" />
    <pr name="pJustify" value="both" />
  </fmt>
  <fmt id="2">
    <pr name="rFonts" value="Calibri" />
    <pr name="sz" value="22" />
    <pr name="pIndentRight" value="1655" />
    <pr name="pJustify" value="both" />
    <pr name="b" value="1" />
  </fmt>
</formattings>
```

Fig. 5 - Codice XML con la legenda di alcuni aspetti tipografici applicati ai blocchi della Fig. 4

Una volta estratto il testo e la sua essenziale formattazione, l'anonimizzatore può procedere a trattarlo per rimuovere tutti i dati personali, secondo le modalità illustrate sopra. Una serie di moduli *software* vengono utilizzati assieme a quattro repertori di nomi italiani di persona maschili e femminili, cognomi, e città. In questo caso[7], il risultato del processo di anonimizzazione è una copia di questi stessi *file*, dove tutti i dati personali sono stati eliminati.

```
<p rend="j">del <hi rend="b"><choice><abbr>dott.</abbr><expan xml:lang="ita">dottor
    </expan></choice></hi> <hi rend="b"><persName type="mn">Immacolato</persName>
    <persName type="s">PAOLETTI</persName></hi>, già Commissario Giudiziale del
    Concordato Preventivo <orgName type="f">Desolina</orgName> in liq., elettivamente
    domiciliato in <placeName>Invorio</placeName>, <address><addrLine>Giodoco
    Gottifredi, 68</addrLine></address>, presso lo studio dell'<choice><abbr>avv
    .</abbr><expan xml:lang="ita">avvocato</expan></choice> <persName type="mn">Ado
    </persName> <persName type="s">Liguoro</persName> (<choice><abbr>c.f.</abbr
    ><expan xml:lang="ita">codice fiscale</expan></choice> <num>HMVMQY03G67Y621X</num
    >; fax <num>389-3174293</num>; pec: <email>hp1462@outlook.it</email>) che lo
    rappresenta ed assiste come da procura in calce alla presente comparsa</p>
```

Fig. 6 - Stralcio del contenuto del documento XML prodotto dalla trasformazione dei documenti delle Figg. 4 e 5

[7] Il sistema essendo modulare non è dipendente da uno specifico formato, sicché lo stesso processo di anonimizzazione potrebbe essere applicato anche ad altri formati digitali, incluso il semplice testo. Nel contesto del progetto Atti Chiari tuttavia si opera con i *file* estratti da Word in quanto è necessario conservare un piccolo gruppo di aspetti tipografici essenziali sia per visualizzare il documento in un formato gradevole, sia per conservare informazione che potrebbe essere a sua volta oggetto di studio per aspetti paragrafematici.

Naturalmente, questo dialetto XML è proprietario, e il suo unico scopo è di costituire una tappa intermedia nel corso del processo di trasformazione dei documenti. Esso viene quindi sottoposto dal programma a una ulteriore conversione, che produce un vero e proprio documento XML TEI standard (Fig. 6). Anche senza entrare nei dettagli di questa codifica, dovrebbe emergere con evidenza la sostanziale differenza di un testo strutturato con TEI rispetto all'originale: in questo esempio infatti i marcatori (il testo racchiuso fra <>) non hanno mai un valore tipografico (tranne quando questo si voglia registrare allo scopo di conservare un dato del documento originale), ma piuttosto semantico: essi indicano infatti blocchi di testo (p), nomi di persona (persName) col loro tipo (nome maschile o femminile, cognome, ecc.); nomi di persone giuridiche (orgName), indirizzi (address), codici numerici o alfanumerici (num), indirizzi *e-mail* (email), ecc. Non si tratta quindi del minuto dettaglio tipografico del testo, ma del ruolo semantico di alcune sue parti. Ciò rappresenta un radicale salto qualitativo nella codifica del testo, che passa così da una marcatura tipografica a una semantica.

In questo ambito, persino le annotazioni relative al formato tipografico originale, come (hi) per il testo in maiuscolo, sono semantiche nella misura in cui non intendono indicare come vada visualizzato un testo, ma solo come si trovava rappresentato graficamente nel documento di origine. Si tratta di un'informazione eminentemente storica e puntuale, che viene qui registrata al pari di altri aspetti del documento potenzialmente utili per l'analisi.

Infine, come si può osservare da questo esempio il programma qui applica un ulteriore trattamento automatico al testo, rappresentato dalla sistematica marcatura e relative indicazioni di scioglimento delle abbreviazioni presenti nel testo (*abbr*, che sta per *abbreviazione*, con vari *tag* TEI *expan*). Come spiegato sopra, è questo un trattamento necessario per gli scopi dell'analisi linguistica nell'ambito del presente progetto, utile non solo a indicare che una forma rappresenta una abbreviazione piuttosto che una parola intera, ma anzitutto a disambiguare il ruolo dei punti rispetto alla individuazione della fine di ogni frase. Tale marcatura, a differenza di quelle applicate manualmente dagli operatori, è del tutto automatica, in base a un repertorio di abbreviazioni a disposizione del programma.

Una volta dunque che il documento sia stato riversato in un formato standard come TEI, questo diviene il fondamento su cui poggiare qualsiasi altra analisi o processo, dalla semplice archiviazione in un *corpus* dotato di funzioni di ricerca, alla sua visualizzazione nei formati più vari (da pagine *web* a documenti PDF), fino a costituire il punto di partenza per analisi linguistiche. In effetti, all'interno del flusso di anonimizzazione e conversione il programma

genera anche una serie di visualizzazioni diagnostiche in formato HTML, prodotte da semplice trasformazione del documento TEI (tramite una semplice tecnologia standard chiamata XSLT), che consentono anzitutto all'operatore di verificare la bontà della marcatura inserita e la resa grafica complessiva del documento. Accanto a queste risorse inoltre il sistema produce costantemente una serie di *file* di dettaglio comprendenti anche le puntuali indicazioni di eventuali errori commessi dall'operatore. Al di là di questi supporti tuttavia, il prodotto finale di questa procedura, che è insieme una anonimizzazione e una conversione, è rappresentato dai documenti TEI, che possono a questo punto prendere la via dell'archiviazione e indicizzazione all'interno di un altro sistema che costituirà lo strumento operativo per linguisti e giuristi.

A partire da un semplice documento di videoscrittura, prodotto direttamente dal suo autore, si giunge così a un vero documento digitale, anonimizzato eppure leggibile, dotato di struttura e marcatura semantica, e adatto a usi molto specialistici; piuttosto che al mero clone digitalizzato della carta, con una marcatura puramente tipografica, dove al più si potrebbe cercare una parola. Questa trasformazione dunque non è solo un cambiamento di codifica digitale, ma una vera *rimodellazione* dei contenuti in senso semantico, e rappresenta il frutto da un lato delle marcature introdotte a scopo di anonimizzazione granulare, e dall'altro di quelle automaticamente desunte dal formato di videoscrittura originale. Questo frutto è l'esito finale del processo di anonimizzazione qui illustrato; ma come ogni prodotto digitale, lungi dall'essere un'opera statica o passiva, si cala piuttosto nell'eracliteo flusso delle risorse digitali, facendosi fondamento e punto di partenza per ogni tipo di analisi si voglia innestare sul *corpus* così costituito.