DH
CENTRO
INTERDIPARTIMENTALE
DI RICERCA

ASSOCIAZIONE per
l'INFORMATICA UMANISTICA
e la CULTURA DIGITALE

AIUCD 2022 | UNIVERSITÀ DEL SALENTO

CULTURE DIGITALI

INTERSEZIONI FILOSOFIA ARTI MEDIA

TESTO ARTI FILOSOFIA CONTENUTI INTELLIGENZA

PROCEEDINGS

ISBN 9788894253566

Fabio Ciracì, Giulia Miglietta, Carola Gatto (edd.), AIUCD 2022 - Culture digitali. Intersezioni: filosofia, arti, media. Preceedings della 11ª conferenza nazionale, Lecce, 2022. Fabio Ciracì, Giulia Miglietta, Carola Gatto (edd.), AIUCD 2022 - Digital cultures. Intersections: philosophy, arts, media. Preceedings of the 11th national conference, Lecce, 2022.

Salvo diversa indicazione, ogni link citato era attivo al 21 gennaio 2022. All links have been visited on 21th January 2022, unless otherwise indicated

Si prega di notificare all'editore ogni omissione o errore si riscontri, al fine di provvedere alla rettifica. Please notify the publisher of any omissions or errors found, in order to rectify them. aiucd.segreteria [at] aiucd.org

I contributi pubblicati nel presente volume hanno ottenuto il parere favorevole da parte di valutatori esperti della materia, attraverso un processo di revisione anonima mediante *double-blind peer review* sotto la responsabilità del Comitato Scientifico di AIUCD 2022.

All the paper published in this volume have received favourable reviews by experts in the field of DH, through an anonymous double-blind peer review process under the responsibility of the AIUCD 2022 Scientific Committee.

Il programma della conferenza AIUCD 2022 è disponibile online all'indirizzo/ The AIUCD 2022 conference program is available online all'apposito indirizzo

http://aiucd2022.unisalento.it http://conference.unisalento.it/ocs/index.php/aiucd2022/index/pages/view/programma

Comitato Scientifico:

General Chair: Fabio Ciracì (Università del Salento)

Local Chair: Mario Bochicchio (Università del Salento, Università di Bari)

Membri Comitato Scientifico: Marina Buzzoni (Presidentessa AIUCD, Uni. Venezia), Federico Boschetti (Ric. ILC-CNR); Federico Meschini (Uni. Tuscia); Roberto Rosselli Del Turco (Uni Torino); Rachele Sprugnoli (Ass. Ric. Univ. Cattolica); Donato Malerba (Università Bari);

Luca Bandirali, Daniela Castaldo, Francesco Ceraolo, Stefano Cristante, Domenico M. Fazio, Manolita Francesca, Marco Mancarella, Pietro Luigi Iaia, Massimiliano Rossi, Grazia Semeraro, Franco Tommasi, Luigi Patrono (Università del Salento)

Membri del Comitato di programma*:* Mario Bochicchio (Local Chair), Luca Bandirali, Daniela Castaldo, Marco Mancarella, Pietro Luigi Iaia, Federica Epifani (Responsabile Comitato di Programma), Ilenia Colonna, Patrizia Miggiano; Carola Gatto; Giulia Miglietta; Marco Giannotta; Alessia De Blasi, Isabella Hernandez.

Direttori di Area: Luca Bandirali; Mario Bochicchio; Fabio Ciracì; Roberto Rosselli Del Turco; Marco Mancarella; Grazia Semeraro.

Segreteria del Convegno: Dott.ssa Silvia Gravili
Resp. tecnico: Carlo Tafuro; web design: Dr.ssa Paola D'Amico; comunicazione: Dr.ssa Loredana De Vitis

Enti organizzatori / Organizing institutions:

AIUCD;

Università del Salento: Centro interdipartimentale in Digital Humanities in collaborazione con i corsi di laurea in Filosofia, DAMS, Beni Culturali e Digital Humanities; ISUFI, Scuola Placetelling.

Università degli Studi Aldo Moro, Dipartimento di Informatica

Sponsor

Regione Puglia; Provincia di Lecce; Città di Lecce; CINI – Consorzio Universitario Nazionale per l'Informatica; SFI-Società Filosofica Italiana; AFC - Apulia Film Commission, Teatro Pubblico Pugliese; Argo Software.

Lista dei revisori - List of the reviewers

Agnese Addone; Tommaso Agnoloni; Luca Bandirali; Nicola Barbuti; Andrea Bellandi; Armando Bisogno; Mario Alessandro Bochicchio; Andrea Bolioli; Federico Boschetti; Dominique Brunato; Paolo Buono; Dino Buzzetti; Marina Buzzoni; Luigi Catalani; Francesco Ceraolo; Daniele Chiffi; Simona Chiodo; Fabio Ciotti; Ilenia Colonna; Christian D'Agata; Elisa D'Argenio; Riccardo De Biase; Manuela De Giorgi; Daniela De Leo; Salvatore De Masi; Pierpaolo Del Coco; Angelo Mario Del Grosso; Francesca Di Donato; Giorgio Maria Di Nunzio; Federica Epifani; Daniela Fogli; Claudio Forziati; Greta Franzini; Francesca Frontini; Emiliano Giovannetti; Edmondo Grassi; Fabiana Guernaccini; Barbara Guidi; Pietro Luigi Iaia; Benedetta Iavarone; Fahad Khan; Maurizio Lana; Angelica Lo Duca; Donato Malerba; Marco Mancarella; Tiziana Mancinelli; Chiara Mannari; Valentina Marangi; Cristina Marras; Federico Meschini; Patrizia Miggiano; Giulia Miglietta; Paolo Monella; Giovanni Morrone; Serge Noiret; Deborah Paci; Antonio Pascucci; Enrico Pasini; Luigi Patrono; Igor Pizzirusso; Simone Rebora; Massimiliano Rossi; Daniela Rotelli; Enrica Salvatori; Eva Sassolini; Daria Spampinato; Rachele Sprugnoli; Enrico Terrone; Francesca Tomasi; Francesco Tommasi; Sara Tonelli; Gennaro Vessio; Marco Salvatore Zappatore.

Indice – Table of Contents

Prefazione

L'undicesima edizione del Convegno Nazionale dell'AIUCD-Associazione di Informatica Umanistica ha per titolo *Culture digitali. Intersezioni: filosofia, arti, media*. Nel titolo è presente, in maniera esplicita, la richiesta di una riflessione, metodologica e teorica, sull'interrelazione tra tecnologie digitali, scienze dell'informazione, discipline filosofiche, mondo delle arti e *cultural studies*. Per questo motivo, il Comitato Scientifico ha individuato cinque aree funzionali alla *call for paper*, in base alle quali selezionare i contributi da presentare in occasione del convegno nazionale. Tutte le aree sono connotate da un "+D" di digitale. Tale espressione non sta a indicare un addendo esornativo e accidentale, ammiccante e modaiolo, né un supplemento alle varie discipline umanistiche con funzione integrativa o sussidiaria; essa denota invece una contaminazione, profonda e trasformativa, delle discipline umanistiche con il digitale, intendendo quest'ultimo in senso ampio, come espressione di una trasformazione scientifica e tecnologica che investe e muta la cultura e la società. Alla luce dell'informatizzazione delle conoscenze e della digitalizzazione delle pratiche, che ridiscutono limiti e poteri delle discipline istituzionali, si tratta di comprendere il nuovo ruolo delle *humanities*. Si tratta di trasformazioni che pongono problematiche inedite, ma al contempo ampliano le possibilità di indagine nei campi della tradizionale ricerca umanistica. Fedeli alla massima di Terenzio – *homo sum humani nihil a me alienum puto* – siamo convinti che una tale contaminazione individui nell'umanista un interlocutore privilegiato. Siamo cioè dell'idea che i saperi si costruiscono reciprocamente, con mutua dipendenza e in maniera interrelata, travalicando i settori scientifici e le camicie di forza delle definizioni settoriali. In questo senso il *digital humanist* rappresenta una figura capace di un supplemento di conoscenza e di una visione interdisciplinare, è abilitato a una ricerca di confine spesso difficile da caratterizzare, sia in relazione agli aspetti più squisitamente teorici dell'informatizzazione, sia in riferimento agli effetti pratici e al loro portato sociale e culturale. A questa trasformazione partecipano a pieno titolo la filosofia e le arti, come discipline chiamate a riflettere sul digitale, non solo perché da sempre si interrogano sull'uomo e sul mondo, ma anche perché ambiscono a migliorare la realtà e governare il cambiamento.

Con l'intenzione, quindi, di coinvolgere la nostra comunità a riflettere sull'intersezione fra i saperi, nell'ottica di una pluralità di culture, il Comitato Scientifico ha individuato le seguenti aree di interesse: "Testo +D", che tesaurizza ed estende la tradizione di ricerca dell'AIUCD, rivolgendosi agli studi di linguistica computazionale, edizioni digitali, progetti ipertestuali, filologia ed ecdotica digitali; "Arti +D", relativa alle tecnologie digitali per il mondo dell'arte, *digital* e *cultural heritage*; "Filosofia +D", riguardante la filosofia dell'informazione, etica ed epistemologia del digitale; "Contenuti +D", con un focus su realtà virtuale e aumentata, contenuti multimediali e transmediali, ecosistemi narrativi e spazio dei media; "Intelligenza +D", orientata alla comunicazione mediata dal computer, apprendimento digitale e sistemi di traduzione automatizzata.

Per sviluppare al meglio le aree tematiche proposte per il convegno, nei mesi di ottobre e novembre 2021 il *Centro di ricerca in Digital Humanities* dell'Università del Salento, in collaborazione con l'AIUCD, ha organizzato il ciclo di seminari "Loading AIUCD2022", a cura di Fabio Ciracì e di Patrizia Miggiano, con sette incontri in modalità telematica, in cui numerosi accademici ed esperti del settore si sono confrontati sui seguenti temi: 20 ottobre 2021, *AI: quali rischi per l'autonomia dell'umano*, (Intelligenza + D), con relatori Angelo Alù, Mariagiovanna Gianfreda, Guglielmo Tamburrini, discussant Mario Bochicchio e moderatrice Ilenia Colonna; 27 ottobre 2021, *Immagini del passato, immagini del futuro* (Media + D), con relatori Malvina Giordana, Alma Mileto e Francesco Zucconi, discussant Luca Bandirali e moderatrice Isabella Hernandez; 9 novembre 2021, *Cultural Heritage & Digital Humanities: sfide di accessibilità* (Arte + D), con relatori Eva Degl'Innocenti, Lucio Tommaso De Paolis, Anna Maria Marras, Paola Moscati, discussant Grazia Semeraro e moderatrice Carola Gatto; 10 novembre 2021, *Textual scholarship: forme, strumenti, metodi* (Testo + D), con relatori Marina Buzzoni, Tiziana Mancinelli, Federico Meschini, Andreas Speer, discussant Fabio Ciracì e moderatrice Giulia Miglietta; 12 novembre 2021, *Politiche pubbliche per la costruzione di un ecosistema digitale* (Diritto + D), con relatori Bianca Bronzino, Mino Elefante, Claudia Morini, discussant Marco Mancarella e moderatore Marco Giannotta; 17 novembre 2021, *Tecnologia e umano: quale futuro per la conoscenza* (Filosofia + D), con relatori Simona Chiodo, Riccardo Fedriga, Cristina Marras e Viola Schiaffonati, discussant Fabio Ciracì e moderatrice Patrizia Miggiano; 24 novembre 2021, *Costruire mondi possibili: i videogiochi e le realtà sociali* (Media + D), con relatori Donata Bologna, Marco-Benoît Carbone, Riccardo Fassone e Pietro Luigi Iaia, discussant Luca Bandirali e moderatrice Alessia De Blasi.

La risposta alla *call for papers* è stata, ci pare, all'altezza delle aspettative: sono giunte 86 proposte, con una media di paper accettati del 77%, esattamente 18 paper e 5 poster accettati nell'area Testo+D, 3 paper accettati nell'area Arti+D, 11 paper e 2 poster in area Filosofia+D, 5 paper e 1 poster per Contenuti+D, infine 7 paper per Intelligenza+D, per un totale di 44 paper e 8 poster. Già da una rapida lettura dei titoli si evince non solo la molteplicità dei temi ma anche la varietà degli approcci metodologici, che attestano declinazioni interne anche alle medesime aree tematiche. Infine, per garantire una selezione dei contributi conforme alle aree di ricerca selezionate, abbiamo identificato la figura dei direttori di area, ai quali è stato assegnato il compito di individuare i revisori più adeguati ai temi dei contributi da revisionare, per un'analisi competente e puntuale: per l'area testo, Roberto Rosselli Del Turco; per l'area arti, Grazie Semeraro; per l'area filosofia, Fabio Ciracì; per l'area contenuti, Luca Bandirali e Marco Mancarella; per l'area intelligenza, Mario Bochicchio. A tutti loro va il ringraziamento del Comitato Scientifico e di AIUCD. Una tale suddivisione del lavoro e il supporto della piattaforma digitale *conference* hanno permesso di seguire con efficacia tutto il processo di selezione dei contributi: individuare i revisori idonei, confrontare le valutazioni e richiedere pareri ulteriori in caso di dubbio, controllare che le modifiche richieste agli autori in fase di revisione fossero correttamente apportate alla versione finale del *paper*, selezionare i contributi da presentare alla conferenza. Ciascun contributo è stato valutato da almeno due *referee* in caso di giudizio positivo, almeno tre in caso di giudizio incerto o di giudizi discordanti, o parere negativo. I 75 revisori hanno svolto un lavoro fondamentale di revisione che ha garantito una selezione seria e competente, assicurando al convegno dell'AIUCD la qualità delle proposte e il riconoscimento internazionale duramente conquistato dall'Associazione in questi undici anni di attività.

AIUCD2022 è patrocinato dalla Regione Puglia, dalla Provincia di Lecce e dalla Città di Lecce, la qual cosa è certo indice di una certa sensibilità territoriale ai temi della cultura e della innovazione. Inoltre, il convegno è stato sponsorizzato da: Dipartimento di Informatica dell'Università degli Studi di Bari, AFP – Apulia Film Commission, Il Teatro Pubblico Pugliese, CINI – Consorzio Interuniversitario nazionale per l'Informatica, SFI-Società Filosofica Italiana, il Teatro Pubblico Pugliese e Argo Software, che hanno generosamente sovvenzionato l'iniziativa.

Purtroppo, come nella scorsa edizione, nonostante il ricorso ai vaccini, anche quest'anno il *covid* ha ripreso a correre, improvvisamente, a poco più di una settimana dal Convegno, previsto per il 19-21 gennaio 2022, e ci ha costretti a rinviare il Convegno alla prossima estate. La scelta è stata sofferta e sicuramente ha determinato disagi, ma abbiamo inteso dare priorità alla sicurezza e alla salute pubblica, pur in assenza di decreti restrittivi o limitazioni governative all'attività convegnistica. Non abbiamo inteso proporre invece il convegno in modalità *online*, perché non abbiamo voluto rinunciare al nostro amato convegno in presenza. La virtuosa trasposizione in modalità digitale di AIUCD2021 offerta, in emergenza, per il Convegno di Pisa è stata sicuramente un esperimento riuscito. Tuttavia, dopo due anni di pandemia, il Comitato Scientifico, di concerto con il Direttivo AIUCD, ha reputato opportuno scegliere comunque di rinviare, per privilegiare il convegno in presenza, senza ovviamente rinunciare ai vantaggi offerti dalla modalità ibrida. Un ulteriore convegno solo in remoto avrebbe altrimenti gravato immancabilmente sugli aspetti sociali e relazionali, per nulla secondari, che costituiscono la vera sostanza del convegno nazionale, rendendolo un luogo di confronto vivo, un'insostituibile occasione di relazione e di partecipazione attiva. Siamo dell'opinione che il digitale debba rappresentare un'opportunità, non già una dimensione sostitutiva ed esclusiva, ma complementare e inclusiva.

Il Convegno previsto per il 19-21 gennaio 2022 indicava la partecipazione di prestigiosi studiosi che arricchivano la proposta tematica di AIUCD2022, che intendiamo confermare anche per il rinvio di giugno. Innanzitutto, i nostri *keynote*: Luciano Floridi – Professore Ordinario di filosofia ed etica dell'informazione presso l'Oxford Internet Institute e direttore del Digital Ethics Lab dell'Università di Oxford, nonché Professore di Sociologia della comunicazione presso l'Università di Bologna – inaugurerà il convegno con una lezione su *Semantic capital: its nature, value, and preservation*; Maurizio Ferraris – Professore Ordinario di filosofia teoretica presso la Facoltà di Lettere e Filosofia dell'Università degli Studi di Torino e noto studioso della *documedialità* – concluderà i lavori con una lezione intitolata *Webfare*. Si aggiungeranno gli *invited speaker* che, per ogni giorno della conferenza, sviluppano un tema specifico del convegno: Maria Grazia Mattei – umanista, critica d'arte e direttrice di *Meet the Media Guru* – si soffermerà sull'*Arte digitale: storia e panoramica attuale*; Gino Roncaglia – Professore Associato dell'Università Roma Tre, esperto di digitale e cultura del libro, consulente RAI – discuterà di *Simulismi;* Anna Bisogno – Professore Associato di Cinema Radio e Televisione dell'Università Telematica Mercatorum – analizzerà *La rete-visione. Televisione e schermi nell'era digitale;*

infine, Riccardo Fedriga – Professore Associato dell'Università di Bologna, esperto di editoria digitale, storico delle idee – esaminerà le *Fruttuose debolezze. Fragilità e indeterminismi digitali.*

Lavoreremo affinché il programma, così faticosamente costruito per gennaio, non subisca variazioni strutturali. Inoltre, al posto del consueto *Book of Abstracts,* per l'edizione del 2022 l'AIUCD ha scelto di pubblicare i *Proceedings*, come segno tangibile di un processo di aggiornamento continuo del Convegno Nazionale e di crescita intellettuale dell'Associazione. Essi vedono la luce nonostante il rinvio del convegno in presenza, per fornire una base alla discussione che si svolgerà questa estate, con la consapevolezza che gli studi pubblicati fotografano lo stato dell'arte, ma che la ricerca è in continua evoluzione. Quindi, in sede di convegno, faremo i conti con i progressi avvenuti nei mesi trascorsi dalla pubblicazione dei *Proceedings*, di cui terremo conto per l'eventuale pubblicazione dei *selected papers*.

Vorremmo chiudere la prefazione rivolgendo un particolare ringraziamento ai membri del Comitato Scientifico e, *last but not least*, esprimendo profonda gratitudine ai componenti del Comitato di programma, coordinati da Federica Epifani: tutte giovani e promettenti energie intellettuali a cui è dedicato il presente volume di *Proceedings*, non a caso edito a cura di Giulia Miglietta e Carola Gatto.

Fabio Ciracì
Mario Bochicchio

Analisi linguistica e pseudonimizzazione: strumenti e paradigmi

Laura Clemenzi[1], Francesca Fusco[2], Daniele Fusi[3], Giulia Lombardi[4]
[1] Università degli Studi della Tuscia, Italia - laura.clemenzi@unitus.it
[2] Università del Salento, Italia - francesca.fusco@unisalento.it
[3] Bamberg University, Germania - daniele.fusi@unive.it
[4] Università di Genova, Italia - giulia.lombardi@edu.unige.it

ABSTRACT
In questo contributo si presenta la procedura innovativa messa a punto nell'ambito del progetto PRIN "La chiarezza degli atti del processo (AttiChiari): una base di dati inedita per lo studioso e il cittadino" per il trattamento dei testi giuridici, funzionale sia alla pseudonimizzazione dei dati sensibili, sia all'analisi linguistica. Si introduce inoltre il motore di ricerca che consentirà di esplorare il *corpus* in fase di costruzione.

PAROLE CHIAVE
Analisi linguistica, *corpora*, marcatura, motore di ricerca, pseudonimizzazione.

INTERVENTO

1. IL PROGETTO, GLI OBIETTIVI, LE QUESTIONI[1]
Il PRIN 2017 "La chiarezza degli atti del processo (AttiChiari): una base di dati inedita per lo studioso e il cittadino" – progetto a cui collaborano linguisti e giuristi degli atenei di Genova, Firenze, Lecce e Viterbo – si prefigge di creare una nuova risorsa per una scrittura efficace degli atti processuali[2]. In particolare, in una prima fase l'obiettivo è allestire, per fini di studio linguistico, un *corpus* sincronico di atti di parte di circa tre milioni di parole rappresentativo, per tipologie testuali e provenienza geografica, delle diverse prassi di scrittura degli avvocati. Successivamente, con i testi raccolti, si intende realizzare una base dati interrogabile che in una specifica sezione includa esempi di scrittura chiara ed efficace, utili per il giurista e anche per il cittadino[3].
La peculiarità dei testi che compongono il *corpus* è la presenza al loro interno di dati sensibili, la cui diffusione violerebbe il diritto alla riservatezza delle parti, di eventuali terzi coinvolti e dei procuratori costituiti. È dunque necessaria, propedeuticamente a qualsiasi tipo di studio, e come requisito stesso per ottenere l'accesso agli atti, un'attività di anonimizzazione dei documenti che renda irriconoscibili le vicende e i soggetti.
Le prassi di anonimizzazione usate tradizionalmente in Italia per riprodurre e diffondere testi giuridici che contengono dati sensibili, come ad esempio i provvedimenti giudiziari, consistono nella mera eliminazione di tali dati tramite l'omissione o la cancellatura con tratti neri, oppure nella loro sostituzione con asterischi, *omissis*, lettere, o altri segni grafici[4]: tutte prassi non compatibili con le esigenze del linguista, che necessita di testi massimamente leggibili e quanto più possibile completi per poter analizzare appieno le strategie usate dagli avvocati nel riferirsi alla parte assistita, alla controparte e agli altri soggetti del processo, sia all'interno di uno stesso atto, sia, in un'ottica di studio di tipo "verticale" e intertestuale, negli altri atti relativi allo stesso giudizio. Oscurando nomi, toponimi, date e ogni altro dato sensibile, difatti verrebbe meno la possibilità di individuare e distinguere le parti processuali e di ricostruire le vicende narrate: sarebbe quindi impossibile dipanare l'intreccio delle voci scriventi (cfr. [12]: 30).
Con questo contributo intendiamo proporre un modello possibile di trattamento semiautomatico degli atti di parte italiani come operazione propedeutica e funzionale sia alla tutela dei dati sensibili, sia all'analisi linguistica e contenutistica dei

[1] Il testo è stato concordato e rivisto da tutti gli autori; tuttavia, ai fini dell'attribuzione della paternità delle singole parti di cui si compone, vanno attribuiti a Laura Clemenzi il paragrafo 1, a Francesca Fusco i paragrafi 2-3, a Daniele Fusi i paragrafi 6-7, a Giulia Lombardi i paragrafi 4-5.
[2] Per alcuni primi studi sulla lingua degli atti di parte, cfr. ([22];[18];[19];[3];[5];[16];[8];[2]).
[3] Per maggiori dettagli sugli obiettivi del progetto e sulle procedure adottate, si rinvia agli interventi raccolti nel volume curato da Gualdo e Clemenzi ([15]); per alcuni esempi di fenomeni linguistici ricercabili nella base dati Atti Chiari, cfr. in particolare ([4]).
[4] Per alcuni esempi di prassi di anonimizzazione tradizionali, cfr. ([1]); in questo testo, si veda più avanti la Figura 4. Segnaliamo di passaggio, come eccezione, il caso del Consiglio di Giustizia Amministrativa della Regione Sicilia, che con la sentenza n. 1134/2020 ha deciso di sostituire gli *omissis* con nomi di fantasia.

testi. Il modello sviluppato deve ancora essere testato nella sua interezza, ma sono stati già condotti, con esito positivo, alcuni test pilota sugli atti a disposizione del gruppo di ricerca del PRIN Atti Chiari.

2. I REQUISITI DEL PROGRAMMA

Sulla base degli obiettivi e delle esigenze del gruppo di ricerca descritte poco sopra, si rende necessario sostituire i dati sensibili contenuti negli atti con dati fittizi della stessa categoria, secondo una procedura di "pseudonimizzazione", definita dal *Regolamento generale sulla protezione dei dati* (Reg. U.E. n. 2016/679), art. 4, c. 5, come «il trattamento dei dati personali in modo tale che i dati personali non possano più essere attribuiti a un interessato specifico senza l'utilizzo di informazioni aggiuntive, a condizione che tali informazioni aggiuntive siano conservate separatamente e soggette a misure tecniche e organizzative intese a garantire che tali dati personali non siano attribuiti a una persona fisica identificata o identificabile»[5].

In particolare, ai fini dell'analisi linguistica è fondamentale mantenere la coerenza concettuale-semantica tra i dati originali e quelli fittizi e la coerenza morfosintattica dei dati fittizi con il contesto (è importante, dunque, che il dato nuovo corrisponda in maniera univoca all'originale in tutte le occorrenze del testo e che ne conservi il genere, per non alterare la morfosintassi della frase in cui è inserito) (cfr. [12]: 30-34). A tale fine, una sostituzione automatica dei dati sensibili, tramite un programma che attinga a liste predefinite per i nomi e che modifichi sequenze numeriche alfanumeriche (quali date, targhe, fax, numeri di telefono, ecc.) si rivela la soluzione più efficace, in grado di ridurre il rischio di errore e di garantire un risultato uniforme.

Inoltre, dal momento che lo studio che si intende condurre sui testi è non solo di tipo linguistico, ma anche giuridico, è opportuno prevedere insiemi di metadati diversi a seconda degli scopi: se, ad esempio, l'analisi linguistica presuppone l'inserimento di metadati relativi al paratesto, quella giuridica richiede che la sostituzione delle date non pregiudichi la ricostruzione cronologica dei fatti.

3. LA MARCATURA E LA PSEUDONIMIZZAZIONE

Ai fini del progetto è stato ideato un nuovo metodo per il mascheramento dei dati sensibili funzionale alle analisi da condurre. Si tratta di un metodo di annotazione a due fasi, ispirato ai modelli di Douglass et al. ([9]), Noumeir ([20]), Elger ([10]) e Dalianis ([7]) per la pseudonimizzazione delle cartelle cliniche, e di Oksanen ([21]) per la pseudonimizzazione degli atti giudiziari finlandesi.

Inizialmente si interviene sul testo con una leggera marcatura manuale, che, invece di togliere, aggiunge informazioni: l'operatore annota il testo di partenza direttamente in un applicativo di videoscrittura, secondo una sintassi concordata, che segnala sia la categoria del dato sensibile, sia il genere[6]. Posto che, come si è detto, il trattamento dei testi non è funzionale solo alla pseudonimizzazione in senso stretto, ma anche a uno studio di tipo linguistico, vi sono poi altri marcatori che non comportano la sostituzione della porzione di testo marcata, bensì sono propedeutici solo alla successiva analisi linguistica: ad esempio si è deciso di marcare i forestierismi, per i quali si usano i codici ISO 639, preceduti da f- (*foreign*).

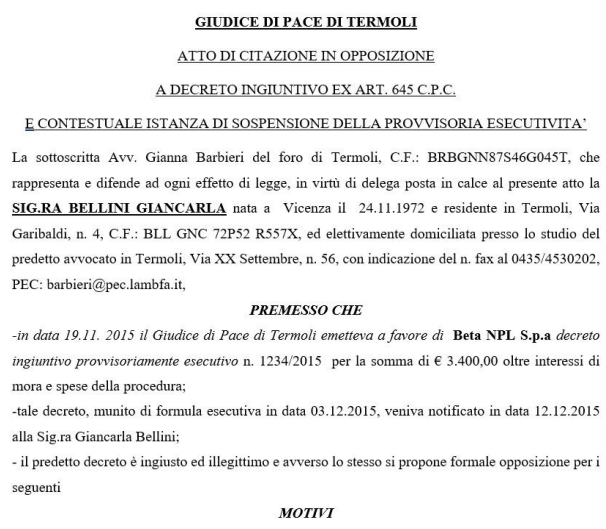
Alla marcatura manuale segue la pseudonimizzazione automatica: per sostituire i dati identificativi il programma attinge a repertori di prenomi maschili e femminili, cognomi e toponimi[7] (nel caso di nomi inizianti per vocale e preceduti da un *d* eufonica, il programma attinge a repertori di soli nomi inizianti per vocale). La coerenza concettuale-semantica all'interno del documento – o dei documenti, nel caso di più atti afferenti allo stesso giudizio – è garantita dal fatto che uno stesso dato sensibile, quando preceduto dallo stesso marcatore, è sostituito dal programma con il medesimo dato fittizio in tutte le sue occorrenze (cfr. [12]: 33-34). Cifre e sequenze alfanumeriche (come targhe, fax, numeri di telefono, ecc.) sono invece sostituite dal programma con stringhe di numeri e lettere casuali di pari estensione; un'attenzione particolare meritano le date, visto che per uno studio di tipo giuridico degli atti è necessario mantenere la coerenza dei riferimenti cronologici delle vicende fattuali e processuali in essi narrate: a tale scopo, il programma, per impostazione predefinita, lascia intatti mese e

[5] La tecnica della pseudonimizzazione è richiamata in più parti del Reg. U.E. n. 2016/679 proprio come misura di «garanzia adeguata» della riservatezza dei dati: cfr. gli artt. 6, c. 4, 25, c. 1, 32, c. 1, 40, c. 2, 89, c. 1 (oltre ai *considerando* 26, 28, 29, 75, 78, 85, 156). Cfr. anche ([12]: 30-31).

[6] I marcatori finora usati per identificare i dati sensibili sono: a-f-f (*anthroponym*, *female*, *first*) per gli antroponimi femminili; a-m-f (*anthroponym*, *male*, *first*) per gli antroponimi maschili; a-l (*anthroponym*, *last*) per i cognomi; j-f (*juridic person*, *female*) per i nomi propri di persone giuridiche di genere grammaticale femminile; j-m (*juridic person*, *male*) per i nomi propri di persone giuridiche di genere grammaticale maschile; t (*toponym*) per i toponimi; ad (*address*) per gli indirizzi; m (*e-mail*) per gli indirizzi di posta elettronica; d (*date*) per le date; n (*number*) per le cifre (es. numeri di telefono, importi in denaro, particelle catastali, ecc.); u per le stringhe alfanumeriche (es. codici fiscali, sigle delle province, targhe, ecc.); x per i dati da oscurare che non rientrano in nessuna delle precedenti categorie (sostituiti con ###).

[7] Per gli scopi previsti dal progetto, non si è reso necessario distinguere ulteriori sottocategorie di toponimi, quali ad esempio città, paesi, Stati.

giorno, sottraendo all'anno un valore compreso fra un minimo (ad es. 5) e un massimo (ad es. 15), uguale in tutta la sessione di analisi (anche se resta comunque possibile optare per una sostituzione randomica delle date).
I documenti che si ottengono non contengono dati sensibili, ma restano perfettamente leggibili (e quindi ben si prestano ad analisi sia linguistiche, sia giuridiche): per vedere concretamente il funzionamento del procedimento di marcatura e pseudonimizzazione adottato nell'ambito del PRIN Atti Chiari e i suoi vantaggi in termini di leggibilità rispetto alle prassi anonimizzatorie tradizionali, si riporta di seguito l'*incipit* di un facsimile di atto di citazione in opposizione a decreto ingiuntivo nelle versioni originale, marcata e pseudonimizzata, cui viene affiancata, per confronto, la versione del medesimo documento anonimizzata mediante mero oscuramento dei dati[8].



GIUDICE DI PACE DI TERMOLI

ATTO DI CITAZIONE IN OPPOSIZIONE

A DECRETO INGIUNTIVO EX ART. 645 C.P.C.

E CONTESTUALE ISTANZA DI SOSPENSIONE DELLA PROVVISORIA ESECUTIVITA'

La sottoscritta Avv. Gianna Barbieri del foro di Termoli, C.F.: BRBGNN87S46G045T, che rappresenta e difende ad ogni effetto di legge, in virtù di delega posta in calce al presente atto la **SIG.RA BELLINI GIANCARLA** nata a Vicenza il 24.11.1972 e residente in Termoli, Via Garibaldi, n. 4, C.F.: BLL GNC 72P52 R557X, ed elettivamente domiciliata presso lo studio del predetto avvocato in Termoli, Via XX Settembre, n. 56, con indicazione del n. fax al 0435/4530202, PEC: barbieri@pec.lambfa.it,

PREMESSO CHE

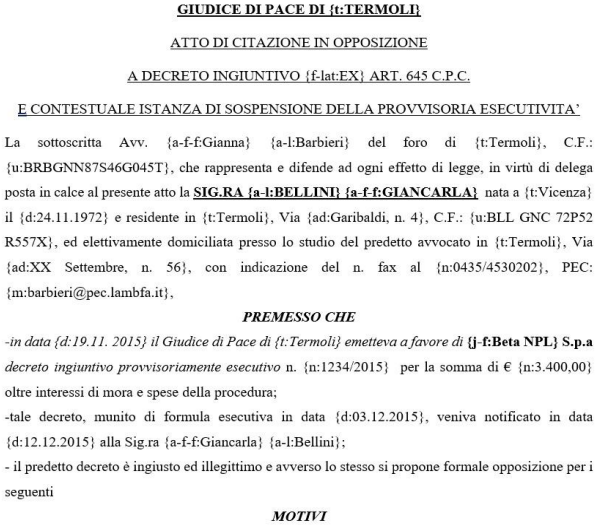
-in data 19.11. 2015 il Giudice di Pace di Termoli emetteva a favore di **Beta NPL S.p.a** *decreto ingiuntivo provvisoriamente esecutivo* n. 1234/2015 per la somma di € 3.400,00 oltre interessi di mora e spese della procedura;

-tale decreto, munito di formula esecutiva in data 03.12.2015, veniva notificato in data 12.12.2015 alla Sig.ra Giancarla Bellini;

- il predetto decreto è ingiusto ed illegittimo e avverso lo stesso si propone formale opposizione per i seguenti

MOTIVI

Figura 1 - Facsimile atto di citazione (versione originale)



GIUDICE DI PACE DI {t:TERMOLI}

ATTO DI CITAZIONE IN OPPOSIZIONE

A DECRETO INGIUNTIVO {f-lat:EX} ART. 645 C.P.C.

E CONTESTUALE ISTANZA DI SOSPENSIONE DELLA PROVVISORIA ESECUTIVITA'

La sottoscritta Avv. {a-f-f:Gianna} {a-l:Barbieri} del foro di {t:Termoli}, C.F.: {u:BRBGNN87S46G045T}, che rappresenta e difende ad ogni effetto di legge, in virtù di delega posta in calce al presente atto la **SIG.RA {a-l:BELLINI} {a-f-f:GIANCARLA}** nata a {t:Vicenza} il {d:24.11.1972} e residente in {t:Termoli}, Via {ad:Garibaldi, n. 4}, C.F.: {u:BLL GNC 72P52 R557X}, ed elettivamente domiciliata presso lo studio del predetto avvocato in {t:Termoli}, Via {ad:XX Settembre, n. 56}, con indicazione del n. fax al {n:0435/4530202}, PEC: {m:barbieri@pec.lambfa.it},

PREMESSO CHE

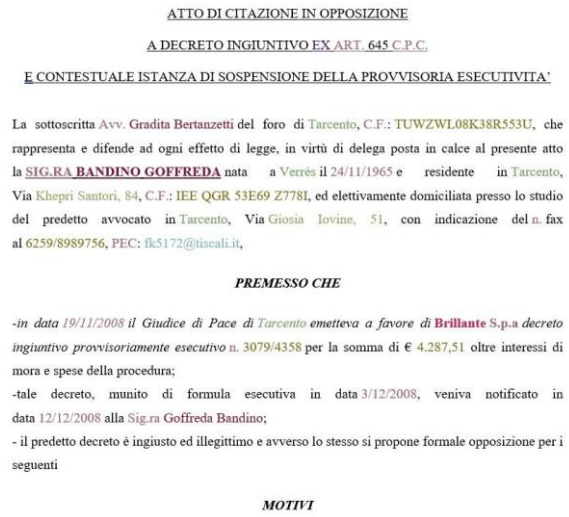
-in data {d:19.11. 2015} il Giudice di Pace di {t:Termoli} emetteva a favore di **{j-f:Beta NPL} S.p.a** *decreto ingiuntivo provvisoriamente esecutivo* n. {n:1234/2015} per la somma di € {n:3.400,00} oltre interessi di mora e spese della procedura;

-tale decreto, munito di formula esecutiva in data {d:03.12.2015}, veniva notificato in data {d:12.12.2015} alla Sig.ra {a-f-f:Giancarla} {a-l:Bellini};

- il predetto decreto è ingiusto ed illegittimo e avverso lo stesso si propone formale opposizione per i seguenti

MOTIVI

Figura 2 - Facsimile atto di citazione (versione marcata)



ATTO DI CITAZIONE IN OPPOSIZIONE

A DECRETO INGIUNTIVO EX ART. 645 C.P.C.

E CONTESTUALE ISTANZA DI SOSPENSIONE DELLA PROVVISORIA ESECUTIVITA'

La sottoscritta Avv. Gradita Bertanzetti del foro di Tarcento, C.F.: TUWZWL08K38R553U, che rappresenta e difende ad ogni effetto di legge, in virtù di delega posta in calce al presente atto la SIG.RA **BANDINO GOFFREDA** nata a Verrès il 24/11/1965 e residente in Tarcento, Via Khepri Santori, 84, C.F.: IEE QGR 53E69 Z778I, ed elettivamente domiciliata presso lo studio del predetto avvocato in Tarcento, Via Giosia Iovine, 51, con indicazione del n. fax al 6259/8989756, PEC: fk5172@tiscali.it,

PREMESSO CHE

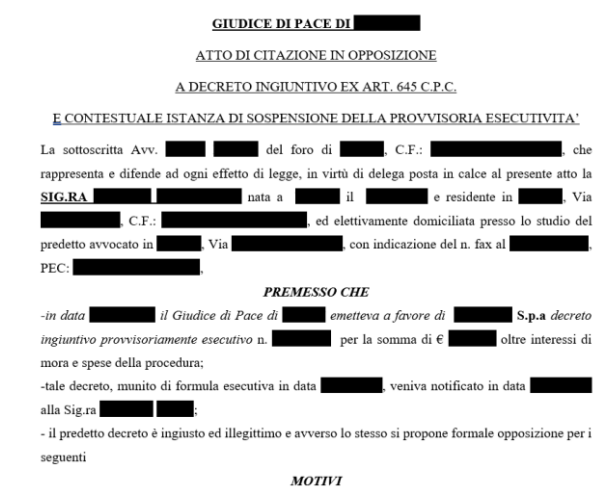
-in data 19/11/2008 il Giudice di Pace di Tarcento emetteva a favore di **Brillante S.p.a** *decreto ingiuntivo provvisoriamente esecutivo* n. 3079/4358 per la somma di € 4.287,51 oltre interessi di mora e spese della procedura;

-tale decreto, munito di formula esecutiva in data 3/12/2008, veniva notificato in data 12/12/2008 alla Sig.ra Goffreda Bandino;

- il predetto decreto è ingiusto ed illegittimo e avverso lo stesso si propone formale opposizione per i seguenti

MOTIVI

Figura 3 - Facsimile atto di citazione (versione pseudonimizzata)



GIUDICE DI PACE DI [redacted]

ATTO DI CITAZIONE IN OPPOSIZIONE

A DECRETO INGIUNTIVO EX ART. 645 C.P.C.

E CONTESTUALE ISTANZA DI SOSPENSIONE DELLA PROVVISORIA ESECUTIVITA'

La sottoscritta Avv. [redacted] del foro di [redacted], C.F.: [redacted], che rappresenta e difende ad ogni effetto di legge, in virtù di delega posta in calce al presente atto la **SIG.RA** [redacted] nata a [redacted] il [redacted] e residente in [redacted], Via [redacted], C.F.: [redacted], ed elettivamente domiciliata presso lo studio del predetto avvocato in [redacted], Via [redacted], con indicazione del n. fax al [redacted], PEC: [redacted],

PREMESSO CHE

-in data [redacted] *il Giudice di Pace di* [redacted] *emetteva a favore di* [redacted] **S.p.a** *decreto ingiuntivo provvisoriamente esecutivo* n. [redacted] per la somma di € [redacted] oltre interessi di mora e spese della procedura;

-tale decreto, munito di formula esecutiva in data [redacted], veniva notificato in data [redacted] alla Sig.ra [redacted];

- il predetto decreto è ingiusto ed illegittimo e avverso lo stesso si propone formale opposizione per i seguenti

MOTIVI

Figura 4 - Facsimile atto di citazione (versione anonimizzata)

Sempre per gli scopi del progetto e per la natura dei testi, non si è reso necessario prevedere un sistema di de-pseudonimizzazione (cfr. [10];[20]), ovvero il processo inverso alla pseudonimizzazione che permette di recuperare in maniera inequivocabile i dati personali univocamente associati ai dati fittizi.

4. IL FLUSSO DEI DATI

Il processo descritto, inoltre, è guidato da un insieme variabile di regole configurabili a seconda degli obiettivi: infatti,

[8] La Figura 1 riproduce il testo e la formattazione dell'originale ma contiene dati già fittizi. I metadati relativi alle informazioni paratestuali vengono conservati dal programma e dunque lo stile dell'originale (grassetti, corsivi, ecc.) è riprodotto anche nella versione pseudonimizzata. Per altri esempi di atti pseudonimizzati, cfr. ([12]: 31-39).

come abbiamo già sottolineato, nel nostro caso il trattamento dei testi non è funzionale solo alla pseudonimizzazione in senso stretto, ma anche all'analisi linguistica; alcuni accorgimenti come la marcatura dei forestierismi e l'attenzione posta al rispetto dei fenomeni fonosintattici (ad esempio la *d* eufonica) anche in fase di pseudonimizzazione, contribuiscono alla raccolta dei metadati. Come anticipato sopra e come si dirà meglio più avanti, altre fonti di metadati sono lo stesso formato digitale in *rich text*, che consente di recuperare aspetti tipografici, e altri strumenti esterni come i *POS taggers*[9].
Un ulteriore beneficio offerto da questo apparentemente paradossale approccio, che aggiunge informazione solo per poterla togliere, è inoltre costituito dal fatto che il sistema di pseudonimizzazione diviene in grado di rimodellare il documento di partenza, dalla struttura puramente tipografica, in un documento semanticamente strutturato. In effetti, avvalendosi delle diverse fonti di metadati incluse nell'*input* il sistema ha la capacità di aggiungere al processo di pseudonimizzazione anche quello di conversione del documento, che dal formato di videoscrittura viene convertito in un vero e proprio documento TEI. In tal modo, nel processo completo vengono accodate le fasi di decodifica del formato originale, di pseudonimizzazione secondo un insieme variabile di regole, e di generazione di un documento TEI, corredato da eventuali rese tipografiche in HTML (v. Figura 3), sì da fornire agli operatori un immediato riscontro del loro operato (v. Figura 5)[10].
In effetti, le fonti dei metadati di un documento sono molteplici. Anzitutto, la leggera marcatura applicata, destinata sia all'offuscamento delle informazioni sensibili, sia all'annotazione di aspetti utili solo in funzione dell'analisi linguistica. In secondo luogo, un'ulteriore fonte è costituita dal formato di videoscrittura (DOCX nello specifico) nel quale vengono raccolti la pressoché totalità degli atti. Da esso naturalmente interessa estrarre solo un minimo sottoinsieme di informazioni tipografiche ritenute utili in fase di analisi degli aspetti paratestuali. Fra questi, un sottoinsieme della formattazione del testo viene direttamente estratto dal formato Office Open XML (ISO/IEC 29500). Infine, l'utilizzo di sistemi di *POS tagging* consente di ottenere con una buona approssimazione ulteriori metadati relativi alla lemmatizzazione e alla classificazione morfologica di ogni parola. Tutti questi metadati devono poi trovare posto nell'indice che nutrirà la base del motore di ricerca.

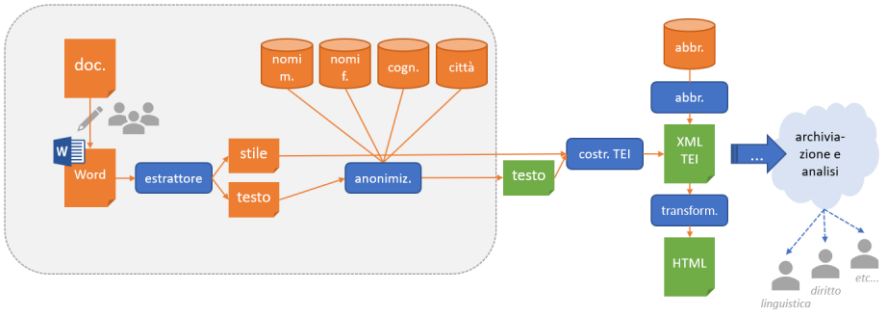


Figura 5 - Prima parte del flusso generale dei dati: il riquadro in grigio delimita l'area protetta, dalla quale nessun dato personale può uscire

Si riproducono di seguito alcuni estratti dei file intermedi del processo di trasformazione illustrato nella Figura 5, a partire dallo stesso atto usato per le esemplificazioni riportate nelle Figure 1-4. Per il riconoscimento delle abbreviazioni cui si fa riferimento nelle didascalie delle Figure 8 e 9, si veda più avanti il par. 6.

```
<pick>
  <par nr="5" fmtId="4">
    <run nr="1" fmtId="4">La sottoscritta Avv. {a-f-f:Gianna} {a-l:Barbieri} del foro di {t:Termoli}, C.F.: {u:BRBGNN87S46G045T}, che rappresenta
    e difende ad ogni effetto di legge, in virtù di delega posta in calce al presente atto la </run>
    <run nr="13" fmtId="5">SIG.RA {a-l:BELLINI} {a-f-f:GIANCARLA} </run>
    <run nr="19" fmtId="4"> nata a {t:Vicenza} il {d:24.11.1972} e residente in {t:Termoli}, Via {ad:Garibaldi n. 4}, C.F.: {u:BLL GNC 72P52 R557X},
    ed elettivamente domiciliata presso lo studio del predetto avvocato in {t:Termoli}, Via {ad:XX Settembre n. 56}, con indicazione del n. fax al
    {n:0435/4530202}, PEC: {m:barbieri@pec.lambfa.it},</run>
  </par>
</pick>
```

Figura 6 - Estratto del documento DOCX in un formato intermedio XML con i dati originali (gli attributi fmtId rimandano a insiemi di caratteristiche tipografiche ricavate dall'originale, e sono sciolti in un'apposita sezione)

[9] Sui diversi livelli di annotazione dei *corpora*, tra cui il *POS* (*part of speech*) *tagging*, cioè l'attribuzione delle categorie grammaticali, si vedano almeno ([11]: 18-25) e ([6]: 84-94).
[10] La Figura 5 è tratta da ([14]: 69); sul funzionamento e sui vantaggi del programma, cfr. ancora ([14];[15]).

```
<pick>
  <par nr="5" fmtId="4">
    <run nr="1" fmtId="4">La sottoscritta Avv. {a-f-f:Gradita} {a-l:Bertanzetti} del foro di {t:Tarcento}, C.F.: {u:TUWZWL08K38R553U}, che rappresenta
    e difende ad ogni effetto di legge, in virtù di delega posta in calce al presente atto la </run>
    <run nr="13" fmtId="5">SIG.RA {a-l:BANDINO} {a-f-f:GOFFREDA} </run>
    <run nr="19" fmtId="4"> nata a {t:Verrès} il {d:24/11/1965} e residente in {t:Tarcento}, Via {ad:Khepri Santori, 84}, C.F.: {u:IEE QGR 53E69 Z778I},
    ed elettivamente domiciliata presso lo studio del predetto avvocato in {t:Tarcento}, Via {ad:Giosia Iovine, 51}, con indicazione del n. fax al
    {n:6259/8989756}, PEC: {m:fk5172@tiscali.it},</run>
  </par>
</pick>
```

Figura 7 - Estratto del documento DOCX in un formato intermedio XML con i dati pseudonimizzati

```
<p rend="j">La sottoscritta Avv. <persName type="fn">Gradita</persName> <persName type="s">Bertanzetti</persName> del foro di
<placeName>Tarcento</placeName>, C.F.: <num>TUWZWL08K38R553U</num>, che rappresenta e difende ad ogni effetto di legge, in virtù di delega
posta in calce al presente atto la <hi rend="b">SIG.RA <persName type="s">BANDINO</persName> <persName type="fn">GOFFREDA</persName> </hi>
nata a <placeName>Verrès</placeName> il <date>24/11/1965</date> e residente in <placeName>Tarcento</placeName>, Via
<address><addrLine>Khepri Santori, 84</addrLine></address>, C.F.: <num>IEE QGR 53E69 Z778I</num>, ed elettivamente domiciliata presso
lo studio del predetto avvocato in <placeName>Tarcento</placeName>, Via <address><addrLine>Giosia Iovine, 51</addrLine></address>, con
indicazione del n. fax al <num>6259/8989756</num>, PEC: <email>fk5172@tiscali.it</email>,</p>
```

Figura 8 - Estratto del documento TEI senza il riconoscimento delle abbreviazioni

```
<p rend="j">La sottoscritta <choice><abbr>Avv.</abbr><expan xml:lang="ita">avvocato</expan></choice> <persName type="fn">Gradita</persName>
<persName type="s">Bertanzetti</persName> del foro di <placeName>Tarcento</placeName>, <choice><abbr>C.F.</abbr>
<expan xml:lang="ita">codice fiscale</expan></choice>: <num>TUWZWL08K38R553U</num>, che rappresenta e difende ad ogni effetto di legge,
in virtù di delega posta in calce al presente atto la <hi rend="b"><choice><abbr>SIG.RA</abbr><expan xml:lang="ita">signora</expan></choice>
<persName type="s">BANDINO</persName> <persName type="fn">GOFFREDA</persName> </hi>nata a <placeName>Verrès</placeName> il <date>24/11/1965</date>
e residente in <placeName>Tarcento</placeName>, Via <address><addrLine>Khepri Santori, 84</addrLine></address>,
<choice><abbr>C.F.</abbr><expan xml:lang="ita">codice fiscale</expan></choice>: <num>IEE QGR 53E69 Z778I</num>, ed elettivamente domiciliata
presso lo studio del predetto avvocato in <placeName>Tarcento</placeName>, Via <address><addrLine>Giosia Iovine, 51</addrLine></address>, con
indicazione del <choice><abbr>n.</abbr><expan xml:lang="ita">numero</expan></choice> fax al <num>6259/8989756</num>,
<choice><abbr>PEC</abbr><expan xml:lang="ita">posta elettronica certificata</expan></choice>: <email>fk5172@tiscali.it</email>,</p>
```

Figura 9 - Estratto del documento TEI con il riconoscimento delle abbreviazioni

5. I REQUISITI DEL MOTORE DI RICERCA

Per soddisfare una serie di requisiti del progetto si è introdotto un particolare motore di ricerca (*Pythia*) nel flusso di lavoro che conduce dai documenti Word ai loro *output* pseudonimizzati e trasformati in TEI. Anche se l'obiettivo del contributo non è quello di illustrare in dettaglio Pythia, trattandosi di un prototipo ancora in via di sperimentazione, ci limitiamo qui ad accennare alla sua impostazione generale in funzione del progetto di ricerca qui trattato e rimandiamo alla bibliografia ([13]) e alla documentazione che accompagna il suo codice *open source* (github.com/vedph/pythia) per ulteriori approfondimenti.

I requisiti del motore destinato ad accogliere indici e metadati sono infatti piuttosto complessi: il primo è che si tratti di un motore capace di fornire concordanze, un attributo non scontato se confrontato al panorama tecnologico dei motori di ricerca testuale più diffusi in ambito informatico e nati con scopi diversi da quelli del progetto Atti Chiari (per esempio, individuare un documento in un *corpus*, oppure localizzare con precisione le occorrenze di ogni parola nel loro contesto). Il secondo requisito è quello di incorporare i metadati descritti nel contesto di un livello di astrazione più elevato, che consenta di trattare in modo omogeneo non solo le parole e i loro metadati, ma anche le strutture testuali più estese (come frasi, versi e strofe) con i loro eventuali metadati.

Simili strutture, naturalmente, molto spesso non sono affatto sovrapponibili, in quanto giacciono su livelli di analisi linguistica completamente distinti: per esempio, non sempre l'albero sintattico di un testo trova corrispondenze puntuali con la sua organizzazione metrica in versi o strofe, o con la sua disposizione colometrica a livello grafico.

6. IL MOTORE DI RICERCA

La necessità di delimitare alcune essenziali strutture (come la frase) determina ulteriormente l'evoluzione del sistema di pseudonimizzazione qui illustrato. La possibilità di incorporare i confini di frase in un indice, per quanto approssimativamente determinabili in base all'interpunzione, consente infatti ricerche contestuali più precise all'interno di un contesto sintatticamente definito, piuttosto che affidato al solo computo della distanza relativa.

L'individuazione dei confini di frase viene effettuata da uno dei numerosi filtri inseriti nella *pipeline* del sistema di indicizzazione, che opera per moduli. Nel caso specifico, trattandosi di input in formato TEI, su un generico algoritmo di *sentence splitting* viene innestato un approccio configurabile che considera anche la natura di determinati marcatori: ad esempio, un marcatore come *head*, associato all'intestazione, viene considerato come corrispondente a una frase, anche se il suo testo manca dei consueti indicatori come la punteggiatura. Il sistema può così disporre di un modulo di *sentence splitting* che si avvale di ulteriori informazioni fornite dalla marcatura XML (TEI o meno, dato che è parametrizzabile), accanto a uno che prende in considerazione solo il testo, adatto ad esempio a input *plain text*; la *pipeline* di indicizzazione viene poi configurata, come per ogni altro suo aspetto, inserendo l'uno o l'altro modulo a seconda dei documenti trattati. Questo approccio modulare è proprio dell'intero sistema di indicizzazione, utilizzando una serie di componenti destinati a

estrarre i testi da una fonte (che non necessariamente è un *file system*), filtrarli in vario modo per prepararli all'analisi, estrarne metadati, calcolarne, secondo vari, algoritmi data e chiave di ordinamento, *tokenizzare* e filtrare i *token*, e individuare una serie di strutture testuali (frasi, versi, strofe, ecc.), in qualsiasi numero e di qualsiasi genere, anche quando esse si sovrappongano.
Ulteriori componenti configurabili riguardano poi la mappatura dell'articolazione interna del testo (ad esempio divisioni in sezioni, paragrafi, ecc.), in modo tale da fornire una mappa di navigazione del testo interattiva nel *frontend* del sistema e l'estrazione di porzioni di testo da presentare come contesto semanticamente congruo (basato su questa medesima mappa), e la trasformazione del formato originale del testo in un formato destinato alla sua presentazione, tipicamente HTML e CSS. Nel caso dei testi TEI qui trattati, il modulo di trasformazione utilizza semplicemente uno script XSLT fornitogli tra i suoi parametri operativi.
In questo ambito, la peculiare natura dei testi trattati ha determinato un'ulteriore evoluzione del sistema di pseudonimizzazione destinata a individuare in modo automatico (sulla base di un elenco e su una rosa di variazioni formali trattate in modo algoritmico) le numerosissime abbreviazioni, che non sarebbe economico affidare alla marcatura manuale. Infatti, poiché l'individuazione delle strutture relative alle frasi si basa essenzialmente sull'interpunzione (anche se non esclusivamente, come nel caso dei documenti dotati di una marcatura in grado di implicare i confini sintattici), la massiccia presenza di abbreviazioni contenenti punti costituirebbe una rilevante fonte di errore. In considerazione di ciò, oltre che a vantaggio della chiarezza del testo per un pubblico non necessariamente specialista in ambito giuridico, si è allora scelto di affidare al sistema di pseudonimizzazione anche il compito di marcare automaticamente le abbreviazioni in una fase distinta e successiva del suo operato. Si tratta quindi di un ulteriore esempio di come la natura stessa di questo sistema sia modellata sulle esigenze del suo uso in sede di analisi, anzitutto linguistica ma anche di altra specie.
Si è infatti visto che un ruolo essenziale in questa analisi linguistica rivestono i metadati e spesso anche le strutture testuali: metadati relativi a informazioni linguistiche (ad es. un latinismo, una classificazione morfologica, un'abbreviazione, un antroponimo, un toponimo, un numero, ecc.), paratestuali (ad es. una parola in grassetto o in corsivo), e sintattiche (qui essenzialmente le strutture rappresentate dalle frasi). In questo ambito, il motore deve poter ricercare allo stesso modo qualsiasi entità estratta dal testo con i suoi metadati, sicché l'approccio adottato consiste nell'elevare il livello di astrazione: un testo non viene più trattato come una sequenza di caratteri all'interno dei quali individuare delle sequenze (*token*) variamente filtrate e indicizzate per essere ricercabili; piuttosto, tale sequenza viene in certo modo smaterializzata per produrre un semplice insieme di oggetti. Ogni oggetto dell'insieme può essere dotato di un qualsiasi numero di metadati appartenenti a un elenco aperto, fra cui anche la posizione nel documento di origine. A questo punto, la ricerca consiste solo nell'individuare gli oggetti di proprio interesse attraverso questi metadati, per poi presentarli nel loro contesto originale. Tali oggetti non sono quindi più solo 'parole', ma anche un qualsiasi tipo di struttura testuale estratta dal testo, la cui posizione viene definita con due punti (primo e ultimo *token*) anziché uno solo (come nel caso di una singola 'parola'). Inoltre, una serie di operatori consente non solo di operare un confronto molto articolato fra il valore ricercato e quello indicizzato, ma anche di rappresentare indicazioni posizionali. Assimilando un oggetto con una singola posizione a un punto (per esempio una 'parola'), e uno con due posizioni a un segmento, questi operatori consentono di trovare un elemento dentro l'altro, o parzialmente sovrapposto a un altro, o alla testa o alla coda di un altro, e così via: è il caso della ricerca di una parola a inizio di frase o a fine di verso, o a fine di frase e di verso, o di una frase parzialmente sovrapposta a un verso o strofe, ecc. Qualsiasi tipo di elemento, derivi esso da una parola o da una struttura, non è che un oggetto con dei metadati: il motore interroga i metadati per giungere agli oggetti e ai loro rapporti, poi li localizza nel testo di origine, e li presenta opportunamente trasformati all'utente finale.
In questo ambito, il sistema offre l'ulteriore vantaggio di fornire un ambiente di lettura dei testi completo, sia in funzione del testo trovato sia in base alle esigenze dell'utente, che dispone anche di una mappa navigabile automaticamente generata per ogni documento. Tutto questo inoltre opera all'interno di un insieme di tecnologie standard e di uso universale: l'indice non è che un database relazionale, facilmente integrabile in qualsiasi progetto e consultabile in vario modo anche al di là del motore di ricerca; inoltre, tutto il processo che conduce dal documento nel suo formato di *input*, quale esso sia (TEI nel nostro esempio), e ovunque sia contenuto (*file system*, *cloud storage*, *web*, *database*, ecc.), è configurato in una *pipeline* componibile, dove ogni stadio viene configurato da una serie di parametri, all'interno di un semplice file JSON di configurazione. Il sistema può dunque arricchirsi di nuove funzionalità semplicemente introducendo nuovi moduli in questa *pipeline*: ad esempio, per introdurre il dettagliato esito di analisi fonologiche o metriche automatiche in seno all'indice, o quello dell'analisi prodotta da sistemi esterni di *POS tagging*, ecc. Nel caso qui esemplificato dunque, l'uso di questo motore può risultare particolarmente vantaggioso proprio in ragione delle peculiarità dei testi trattati e delle soluzioni adottate, a cominciare dal sistema di pseudonimizzazione da cui questo intervento ha avuto principio, modellandosi in funzione dei suoi obiettivi.

7. CONCLUSIONI

Il processo di pseudonimizzazione adottato nel progetto PRIN Atti Chiari e illustrato in questo contributo coniuga l'esigenza di tutelare la riservatezza e la necessità di disporre di testi formalmente completi che possano consentire l'analisi linguistica e l'individuazione di esempi di scrittura forense chiara ed efficace.

L'approccio qui adottato è funzionale, in primo luogo, ad assicurare una completa e non reversibile anonimizzazione dei dati, che non è solo un ovvio requisito legale, ma rappresenta un aspetto fondamentale per ottenere la fiducia di chi contribuisce alla costituzione del *corpus* di atti. In secondo luogo, esso serve a operare sul testo uno o più tipi di trasformazioni, in rapporto agli scopi della procedura, che pur garantendo questo primo requisito preservino la leggibilità del testo e la sua usabilità per analisi di ampio spettro. In queste convergono non solo le annotazioni e i metadati di ogni documento inseriti dagli anonimizzatori, ma anche ulteriori informazioni provenienti dalla conversione del formato del testo da DOCX a TEI (come, ad esempio, gli stili tipografici), o aggiunte da processi supplementari (come lo scioglimento delle abbreviazioni o il *tagger* di terza parte). A sua volta, questo richiede un sistema di ricerca capace di ingerire un insieme aperto di annotazioni, estese sulle parole come su altre strutture linguistiche (ad es. la frase), e di fornire una ricerca per concordanze che integri sul medesimo livello tutte queste fonti di dato, finendo così per mettere in campo un insieme di strumenti il cui valore pratico e metodologico può superare i confini del singolo progetto di ricerca.

BIBLIOGRAFIA

[1] Candrilli, Fernanda. 2021. «Il progetto di archiviazione e anonimizzazione». In *Atti Chiari. Chiarezza e concisione nella scrittura forense*, a cura di Riccardo Gualdo e Laura Clemenzi, 19–29. Viterbo: Sette Città.

[2] Caponi, Remo. 2014. «Il processo civile telematico tra scrittura e oralità.» In *Lingua e processo. Le parole del diritto di fronte al giudice, Atti del Convegno*, 176–86. Firenze: Firenze: Accademia della Crusca.

[3] Cavallone, Bruno. 2010. «Un idioma coriaceo: l'italiano del processo civile». In *L'italiano giuridico che cambia, Atti del Convegno*, 85–95. Firenze: Firenze: Accademia della Crusca.

[4] Clemenzi, Laura. 2021. «L'interrogazione della base dati Atti Chiari». In *Atti Chiari. Chiarezza e concisione nella scrittura forense*, a cura di Riccardo Gualdo e Laura Clemenzi, 41–52. Viterbo: Sette Città.

[5] Conte, Giuseppe. 2013. «Il linguaggio della difesa civile». In *Lingua e diritto. Scritto e parlato nelle professioni legali*, Alarico Mariani Marini e Federigo Bambi, 35–67. Pisa: Pisa University Press.

[6] Cresti, Emanuela, e Alessandro Panunzi. 2013. *Introduzione ai corpora dell'italiano*. Bologna: Il Mulino.

[7] Dalianis, Hercules. 2019. «Pseudonymisation of Swedish Electronic Patient Records Using a Rule-Based Approach». In *Proceedings of the Workshop on NLP and Pseudonymisation, a cura di Lars Ahrenberg e Beáta Megyesi*, 16–23. Turku: Linköping Electronic Press.

[8] Dell'Anna, Maria Vittoria. 2014. «Fra attori e convenuti. Lingua dell'avvocato e lingua del giudice nel processo civile.» In *Lingua e processo. Le parole del diritto di fronte al giudice, Atti del Convegno a cura di Federigo Bambi*, 83–101. Firenze: Accademia della Crusca.

[9] Douglass, Margaret, et al. 2004. «Computer-Assisted De-Identification of Free Text in the MIMIC II Database». *Computers in Cardiology* 31: 341–44.

[10] Elger, Bernice S., e et al. 2010. «Strategies for health data exchange for secondary, cross-institutional clinical research». *Computer Methods and Programs in Biomedicine* 99 (3): 230–51.

[11] Freddi, Maria. 2019. *Linguistica dei corpora*. Roma: Carocci.

[12] Fusco, Francesca. 2021. «Marcatura linguistica e tutela della riservatezza nello studio di un corpus di scritture forensi». In *Atti Chiari. Chiarezza e concisione nella scrittura forense*, a cura di Riccardo Gualdo e Laura Clemenzi, 29–40. Viterbo: Sette Città.

[13] Fusi, Daniele. 2020. «Text Searching Beyond the Text: a Case Study». *Rationes Rerum* 15: 199–230.

[14] ———. 2021. «Digitalizzazione e marcatura XML degli atti». In *Atti Chiari. Chiarezza e concisione nella scrittura forense*, a cura di Riccardo Gualdo e Laura Clemenzi, 59–73. Viterbo: Sette Città.

[15] Gualdo, Riccardo e Laura Clemenzi (a cura di). Atti Chiari. Chiarezza e concisione nella scrittura forense. Viterbo: Sette Città, 2021.

[16] Gualdo, Riccardo, e Maria VIttoria Dell'Anna. 2014. «Per prove e per indizi (testuali). La prosa forense dell'avvocato e il linguaggio giuridico.» In *La lingua variabile nei testi letterari, artistici e funzionali contemporanei. Analisi, interpretazione, traduzione, Atti del XIII Congresso SILFI. A cura di Giovanni Ruffino e Marina Castiglione*, 623–35. Firenze: Cesati.

[17] Lombardi, Giulia. 2021. «I vantaggi del programma an-tool.» In *Atti Chiari. Chiarezza e concisione nella scrittura forense, a cura di Riccardo Gualdo e Laura Clemenzi*, 29–40. Viterbo: Sette Città.

[18] Mortara Garavelli, Bice. 2003a. «L'oratoria forense: tradizione e regole». In *L'avvocato e il processo. Le tecniche della difesa, a cura di Alarico Mariani Marini e Maurizio Paganelli*, 66–91. Milano: Giuffrè.

[19] ———. 2003b. «Strutture testuali e stereotipi nel linguaggio forense». In *La lingua, la legge, la professione forense, a cura di Alarico Mariani Marini*, 3–19. Milano: Giuffrè.

[20] Noumeir, Rita. 2007. «Pseudonymization of Radiology Data for Research Purposes». *Journal of Digital Imaging* 20 (3): 284–95.
[21] Oksanen, Arttu, et al. 2019. «A Pseudonymization Service for Finnish Court Documents». In *Legal Knowledge and Information Systems. JURIX 2019: The Thirty-second Annual Conference, a cura di Michał Araszkiewicz e Víctor Rodríguez-Doncel*, 251–54. Amsterdam: IOS Press.
[22] Sabatini, Francesco. 2015. «Dalla lingua comune al linguaggio del legislatore e dell'avvocato». In *L'avvocato e il processo. Le tecniche della difesa, a cura di Alarico Mariani Marini e Maurizio Paganelli*, 3–14. Milano: Giuffrè.