# Arricchimento semantico delle banche dati giuridiche

*Tommaso Agnoloni* [*]

ABSTRACT:

La disponibilità dei dati giuridici (documenti, metadati) di qualità, aperti, interconnessi, organizzati e arricchiti con attributi semantici espliciti è essenziale per la costruzione di applicazioni evolute di ricerca e analisi giuridica e per la valutazione indipendente delle applicazioni commerciali. La costruzione di una infrastruttura di dati pubblica basata su standard condivisi di interoperabilità e supportata da strumenti automatici di estrazione di informazione e annotazione semantica è una precondizione per la costruzione di sistemi capaci di fornire le spiegazioni ai processi che stanno alla base dei propri risultati, decisioni e previsioni.

*The availability of good quality legal data (documents, metadata), open, interconnected, organized and enriched with explicit semantic attributes is essential for the construction of advanced legal research and analysis applications and for the independent evaluation of commercial applications. The construction of a public data infrastructure based on shared interoperability standards and supported by automatic information extraction and semantic annotation tools is a precondition for the construction of systems capable of providing explanations for the processes that underlie their results, decisions and forecasts.*

Sommario: 1. Introduzione. – 2. Apprendimento statistico nel dominio giuridico. – 2.1. Modelli neurali del linguaggio. – 3. L'approccio semantico e i progetti dell'IGSG-CNR. – 3.1. Linkoln. – 3.2. Marker. – 3.3. Linked Legal Data Repository. – 3.4. DoGi – Dottrina giuridica. – 4. Applicazioni. – 4.1. Senato della Repubblica e portale Normattiva. – 4.2. Corte di cassazione, Corte costituzionale e ItalGiureWeb. – 4.3. Archivio della giurisprudenza di merito. – 4.4. Banca dati della Documentazione Economica e Finanziaria. – 5. Prospettiva: integrazione degli approcci statistico e semantico. – 6. Conclusioni.

---

[*] Primo Ricercatore all'Istituto di Informatica Giuridica e Sistemi Giudiziari del Consiglio Nazionale delle Ricerche (IGSG-CNR).

## 1. *Introduzione*

Il tema dell'applicazione degli strumenti dell'intelligenza artificiale al diritto è oggetto di studi da ormai diversi decenni seppure in una comunità scientifica ristretta di ricerca interdisciplinare. Le principali società scientifiche internazionali dedicate *IAAIL*, *JURIX*[1], raccolgono nelle proprie conferenze i contributi di ricercatori su temi e con approcci diversi che vanno dai modelli formali delle norme e del ragionamento giuridico, all'apprendimento automatico e al *data mining* per le applicazioni giuridiche e molti altri[2]. Da comunità scientifica di nicchia sta guadagnando sempre maggiori attenzioni e anche in quel contesto – viste anche le recenti accelerazioni dei progressi dell'intelligenza artificiale basata sull'apprendimento automatico da grandi quantità di dati e pur cercando di mantenere elevata la consapevolezza delle peculiarità del dominio, delle insidie, dei limiti e del potenziale dei vari livelli di automazione nel settore giuridico – si sta assistendo ad un progressivo incremento degli approcci "*data-based*" a scapito degli approcci più tradizionalmente "*knowledge-based*".

A questa tendenza già in atto si sono aggiunti la diffusione dei servizi commerciali di intelligenza artificiale sotto forma di *chatbot* accessibili al grande pubblico e un dibattito generale che ha investito tutta la società.

In questo contesto, uno dei temi che ha colpito maggiormente l'immaginario in campo giuridico è stato indubbiamente quello della "previsione automatica del giudizio" (*legal judgment prediction*) nelle sue varie declinazioni più o meno minacciose, creando un vivace dibattito e attenzione secondo alcuni ben al di là dei rischi concreti e delle potenzialità reali delle tecnologie[3]. Pur senza sottovalutare i rischi delle tecnologie digitali su funzioni delicate come quella giudiziaria, è vero che il loro potenziale in funzione di assistenza e di miglioramento dell'accesso alla giustizia è allo stato attuale di gran lunga più concreto[4].

---

[1] IAAIL – *International Association for Artificial Intelligence and Law*, http://www.iaail.org. JURIX *Foundation for legal knowledge based systems*, http://jurix.nl.

[2] Si veda, ad esempio: K. ASHLEY, *Artificial Intelligence and Legal Analytics: New Tools for Law Practice in the Digital Age*, Cambridge University Press, Cambridge, 2017, doi: 10.1017/9781316761380.

[3] A. SANTOSUOSSO, G. SARTOR, *La giustizia predittiva: una visione realistica*, in *Giur. it.*, vol. 174, f. 7, 2022, p. 1760.

[4] Conclusioni del Consiglio UE «Accesso alla giustizia – Cogliere le opportunità della

In questo senso è utile tenere presente che l'intelligenza artificiale non è una disciplina unica ma un campo disciplinare molto ampio che va dall'apprendimento automatico, ai sistemi esperti, all'elaborazione del linguaggio naturale e molte altre, ognuna con ulteriori declinazioni e approcci differenti. Non esiste l'"algoritmo" ma ne esistono molti, con presupposti diversi e che si basano e sfruttano i dati in misura e modi diversi.

L'importanza dei dati, completi di buona qualità e aggiornati è forse il fattore maggiormente determinante e trasversale a tutti gli approcci: servono dati di qualità come archivio dettagliato e accurato da interrogare, come base di conoscenza su cui costruire *reasoning* e inferenze logiche spiegabili secondo un modello, per addestrare gli strumenti basati sull'apprendimento. E servono dati per la valutazione indipendente degli strumenti e per garantire la riproducibilità dei risultati della ricerca scientifica.

La creazione di una infrastruttura pubblica di dati giuridici aperti e interoperabili e il loro arricchimento semantico è una delle linee di ricerca perseguite dall'Istituto di Informatica Giuridica e Sistemi Giudiziari del Consiglio Nazionale delle Ricerche (IGSG-CNR) con lo sviluppo e l'implementazione di standard di rappresentazione e strumenti di supporto.

Il dialogo interdisciplinare fra studiosi del diritto e dell'informatica è l'altro elemento essenziale per una comprensione più approfondita dei sistemi e del loro funzionamento da un lato, dei requisiti del dominio giuridico e degli effetti delle tecnologie sulla natura stessa del diritto dall'altro.

## 2. *Apprendimento statistico nel dominio giuridico*

Nelle applicazioni delle prime forme di intelligenza artificiale al diritto l'idea prevalente era quella di formalizzare la conoscenza di dominio in una rappresentazione eseguibile dalle macchine in c.d. "sistemi esperti", grazie alla stretta collaborazione tra giuristi e informatici.

Per diverse ragioni, legate alla crescente produzione e disponibilità di dati in formato elettronico e alla loro sempre maggiore efficienza, gli approcci all'intelligenza artificiale si sono sempre più spostati da "*knowledge based*", basati sulla formalizzazione della conoscenza del dominio, a "*data based*",

---

digitalizzazione» (2020/C 342 I/01) https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52020XG1014(01).

basati sull'apprendimento statistico di pattern e regolarità da grandi quantità di dati di addestramento.

Anche nelle applicazioni dell'intelligenza artificiale al diritto si è avuto un simile spostamento verso approcci basati sui dati. D'altra parte, sono in molti a ritenere che il dominio giuridico abbia delle specificità tali da rendere inadatti o quantomeno insufficienti tali approcci [5].

In primo luogo, perché questi non consentono, almeno nella maggioranza dei casi e in una forma accessibile, di ottenere una *spiegazione*, ovvero gli argomenti a supporto del risultato prodotto dalla macchina. Nel diritto l'argomentazione a supporto di una decisione è importante almeno quanto la decisione stessa. Gli argomenti devono essere motivati e contestabili.

Inoltre, l'apprendimento statistico è *retrospettivo*, si basa sull'apprendimento di regolarità dai dati del passato. Quando lo si applica alla decisione di nuovi casi sulla base dei precedenti giurisprudenziali si favorisce il conformismo nella decisione, si dà un peso eccessivo al precedente che nei sistemi di *civil law* non dovrebbe essere vincolante, si perde uno dei grandi punti di forza del diritto, ovvero che la sua interpretazione è dinamica, capace di far fronte a nuove situazioni e di adattarsi ai cambiamenti della società.

L'apprendimento statistico ha bisogno di un *grande volume* di dati di addestramento (ad esempio casi precedenti) per ottenere un'elevata accuratezza dei risultati. Nel dominio giuridico non è detto che *dataset* di sufficiente grandezza siano sempre disponibili. Su determinate materie o su fattispecie affini al caso in esame non è detto che ci siano molti precedenti su cui basare l'apprendimento. Se per le decisioni di *routine* può essere semplice ottenere previsioni dell'esito affidabili, per i casi limite o più complessi, che sollevano questioni giuridiche di maggiore interesse, non ci sarà un sufficiente numero di casi per supportare una giustificazione statistica della previsione.

A differenza di altri campi di applicazione dell'apprendimento automatico, il diritto continua ad essere strettamente legato alla lingua e alla giurisdizione nazionali, limitando ulteriormente la disponibilità dei dati.

La qualità dei dati di apprendimento è cruciale, sotto diversi punti di vista. L'espressione *garbage-in garbage-out* richiama l'attenzione sul fatto

---

[5] T. BENCH-CAPON, *The need for good-old fashioned AI and law*, in W. HÖTZENDORFER, C. TSCHOL, F. KUMMER (eds.), *International Trends in Legal Informatics: Festschrift for Erich Schweighofer*, Editions Weblaw, Bern, 2020, pp. 23-36. http://dx.doi.org/10.38023/cefe7081-e6dd-49de-9592-9adbb6063fd6.

F. BEX, H. PRAKKEN, *On the relevance of algorithmic decision predictors for judicial decision making*, ICAIL'21, June 21-25, 2021, Sa̋o Paulo, Brazil, https://doi.org/10.1145/3462757.3466069.

che un insieme di dati di cattiva qualità, fornito come input o dataset di apprendimento ad un algoritmo, determina un risultato di cattiva qualità. Non solo, i possibili errori, la presenza di *bias*, lo sbilanciamento dei campioni nei dati di apprendimento, si propagheranno e persisteranno nei risultati prodotti.

Inoltre, i dati del passato non sono omogenei. Nel corso del tempo possono determinarsi variazioni di cui è necessario tenere conto: conferme o riforme nei successivi gradi di giudizio, variazioni delle norme applicabili, variazioni negli orientamenti giurisprudenziali. In certi casi i dati più recenti sono più rilevanti e deve essere attribuito loro un peso maggiore.

Ciò rende ancora più difficile ottenere *dataset* di apprendimento sufficientemente ampi e correttamente rappresentativi e pone l'accento sul carattere critico della qualità dei dati su cui basare gli algoritmi.

Infine, l'efficacia di algoritmi basati sull'apprendimento statistico nel raggiungimento di prestazioni di classificazione o di predizione corrette su casi nuovi, può essere valutata soltanto in relazione a *dataset* di test, su dati non usati per l'apprendimento e di cui sia stato annotato manualmente da esperti il risultato atteso corretto.

Il processo di annotazione manuale degli esempi corretti da usare per addestrare e valutare i risultati prodotti dalla macchina è oneroso e richiede una approfondita conoscenza del dominio. Tuttavia, è indispensabile, in particolare in applicazioni critiche, condividere i *benchmark* per la valutazione da terze parti e i dati annotati per consentire la riproducibilità dei risultati.

Queste cautele restano in larga misura valide e per certi versi accentuate a fronte della dirompente innovazione tecnologica introdotta dai modelli neurali del linguaggio (*Large Language Models*) e delle loro capacità generative.

## 2.1. *Modelli neurali del linguaggio*

La potenza mostrata dalle più recenti tecnologie basate su reti neurali artificiali in molti task "cognitivi" sta indubbiamente portando uno sconvolgimento in tutti i campi della conoscenza, in particolare in quelli basati sul linguaggio naturale e il testo.

I *Large Language Models* (LLM) (modelli neurali del linguaggio, modelli fondazionali) stanno rivoluzionando il campo dell'intelligenza artificiale e, in particolare, quello dell'elaborazione del linguaggio naturale. Segnano un passo avanti significativo nella capacità delle macchine di comprendere e generare il linguaggio umano.

Un modello linguistico di grandi dimensioni è un tipo di modello di intelligenza artificiale che utilizza algoritmi di apprendimento automatico per generare testo simile a quello umano. Questi modelli vengono addestrati su estesi set di dati contenenti miliardi di parole grazie agli enormi corpora testuali reperibili sul *web* (sostanzialmente ogni contenuto pubblicato in rete e raggiungibile con i classici motori di ricerca, inclusi, per dare un'idea della dimensione e della varietà, intere biblioteche digitali, tutta la conoscenza enciclopedica di Wikipedia, corpora legislativi, articoli di informazione, letteratura scientifica, ecc.) e sono progettati per individuare relazioni complesse nei dati per comprendere la semantica e la sintassi del linguaggio. Analizzando le parole e i loro contesti, gli *LLM* possono prevedere le parole successive più probabili in una frase e persino costruire interi paragrafi plausibili, sintatticamente corretti e contestualmente rilevanti, ad esempio in risposta ad una domanda espressa in linguaggio naturale (*prompt*).

Oltre all'enorme aumento dei dati di addestramento consentito dalla crescita esponenziale dei contenuti pubblicati sul *web*, diverse innovazioni scientifiche e tecnologiche (nel campo dell'informatica, della linguistica computazionale, dell'intelligenza artificiale, della visione artificiale e in molti altri) hanno abilitato questo progresso: la possibilità di addestrare reti neurali di dimensioni sempre più grandi – con miliardi di parametri – grazie ai progressi nello sviluppo di processori con sempre maggiore potenza di calcolo; lo sviluppo di innovative architetture di reti neurali (c.d. *Transformer*) per l'analisi di sequenze (inclusi i testi, riconducibili a sequenze di parole). Queste architetture[6] si sono rivelate eccellenti (ad esempio, inizialmente nella traduzione automatica) nel catturare regolarità (*pattern*) e contesto in lunghe sequenze di dati e nel gestire enormi *dataset*. Poiché sono addestrate su dati non etichettati (apprendimento non supervisionato) hanno consentito di abbattere i costi di addestramento e ridotto drasticamente la necessità di dati etichettati da annotatori umani.

Inoltre, questo tipo di modelli consentono di sfruttare la conoscenza linguistica e semantica appresa nell'addestramento sui grandi dataset generali-

[6] Due dei Transformer più conosciuti di oggi: Generative Pretrained Transformer (GPT) A. RADFORD *et al*., *Improving Language Understanding by Generative Pre-Training* (2018) *preprint* OpenAI e Bidirectional Encoder Representations from Transformers (BERT), J. DEVLIN *et al*., *BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding* (2019) In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pp. 4171-4186, Minneapolis, Minnesota. Association for Computational Linguistics.

sti (*pre-training*) e utilizzarla per inizializzare modelli specializzati su specifici domini e per compiti specifici (c.d. *Transfer Learning*) abbattendo le barriere di accesso per lo sviluppo di modelli con elevate prestazioni. Rispetto all'apprendimento supervisionato tradizionale, questo approccio produce in genere modelli di alta qualità che possono essere addestrati in modo molto più efficiente per una varietà di attività e con molti meno dati etichettati.

La rivoluzione introdotta dai modelli neurali di apprendimento basati sui dati solleva numerosi interrogativi e accesi dibattiti nella comunità scientifica e nella società. Sicuramente uno dei maggiori limiti risiede nell'opacità intrinseca di questi sistemi (c.d. *black box*) essendo impossibile o comunque molto difficile spiegare o correggere, avendo accesso soltanto a miliardi di pesi numerici che determinano l'attivazione dei nodi della rete neurale, le risposte fornite dalla macchina. L'unico modo con cui si può sapere se, e non come, i modelli funzionano in determinati compiti, è tramite la valutazione dei risultati su dataset di controllo (*benchmark*, *gold standard*).

Almeno allo stato attuale dello sviluppo, i problemi e i rischi sollevati dall'introduzione di questi modelli sono numerosi e complessi. Per menzionarne soltanto alcuni, le c.d. "allucinazioni" (risposte sintatticamente corrette ma completamente infondate o prive di senso), la difficoltà di correggerli ed aggiornarli, la predominanza dell'inglese come lingua dei contenuti, numerose questioni etiche fra cui la perpetrazione di *bias* e stereotipi negativi appresi dai dataset di addestramento, i problemi di *copyright* sui testi usati per l'addestramento, gli elevati costi legati alla loro costruzione e al loro utilizzo, la cessione di controllo e la forte dipendenza dalle aziende private c.d. *Big Tech* che per concentrazione di risorse (computazionali, di dati, economiche, di ricercatori) detengono un monopolio di fatto nel loro sviluppo, oltre alla ulteriore opacità dei modelli commerciali chiusi e protetti da segreti industriali. In aggiunta ai problemi evidenziati, c'è il fatto che alla straordinaria potenza dimostrata nelle abilità linguistiche e nella capacità generative di testo, non corrisponde una altrettanto sviluppata capacità di ragionamento, né è prevedibile che possa essere raggiunta con un approccio basato esclusivamente sull'apprendimento dai dati. L'integrazione con basi di conoscenza fattuale può essere una delle strade per superare almeno alcuni di questi limiti.

## 3. *L'approccio semantico e i progetti dell'IGSG-CNR*

I grafi della conoscenza (*knowledge graph*) sono grandi reti di entità che rappresentano oggetti del mondo reale e concetti astratti, e le loro relazioni e attributi semantici. Le relazioni semantiche tra le entità sono fondamentali per fornire agli esseri umani e alle macchine il contesto e i mezzi per il ragionamento automatizzato.

Ciò che rende un grafo della conoscenza una soluzione dati unica e potente è il modello semantico dei dati, che ne fa parte. Il modello semantico dei dati, o ontologia, contiene definizioni formali ed esplicite dei concetti e delle relazioni all'interno di un dominio. Un'ontologia arricchisce i dati all'interno di un grafo della conoscenza con contesto e significato che gli esseri umani e i computer possono interpretare. Nell'approccio di modellazione semantica i documenti testuali sono segmentati e arricchiti da metadati espliciti che ne descrivono, a diversi livelli, proprietà e relazioni.

L'approccio semantico nel dominio giuridico [7] ha una tradizione fondata nelle banche dati giuridiche e nei sistemi di rappresentazione della conoscenza e si è evoluto parallelamente alle tecnologie, ai principi di interoperabilità e agli standard del *web,* del *semantic web* e dei *linked open data* e in modo complementare alle tecnologie dell'intelligenza artificiale.

Risponde all'esigenza, particolarmente urgente nel dominio giuridico, di accedere ad informazioni integrate, organizzate e contestualizzate. L'informazione giuridica è infatti estremamente frammentata, sparsa in una varietà di fonti e in cui la conoscenza emerge dalla connessione delle diverse tipologie di informazione. Le principali tipologie di informazione giuridica – legislazione, giurisprudenza e dottrina – hanno caratteristiche, finalità, valenza giuridica e contenuto informativo molto diverso tra loro, ma è solo dall'accesso integrato a tutte le diverse tipologie che è possibile ricostruire in modo approfondito la struttura reticolare del discorso giuridico.

Ciascun *corpus* documentale giuridico (legislativo, giurisprudenziale, dottrinale) ha una vastità e una ricchezza di connessioni e relazioni sia interne

---

[7] M. VAN OPIJNEN, *The European Legal Semantic Web: Completed Building Blocks and Future Work* (November 22, 2012). European Legal Access Conference, November 2012, Available at SSRN: https://ssrn.com/abstract=2181901.

T. AGNOLONI, *Dall'informazione giuridica agli open data giuridici*, in G. PERUGINELLI, M. RAGONA (a cura di), *L'informatica giuridica in Italia. Cinquant'anni di studi, ricerche ed esperienze*, Collana ITTIG-CNR, Serie "Studi e documenti", n. 12, ESI, Napoli, 2014, pp. 581-602.

che verso gli altri corpora da costituire di per sé una rete fortemente interconnessa di documenti, frammenti testuali e metadati. Ad esempio, le citazioni testuali rendono il corpus legislativo una rete ipertestuale da ben prima dell'avvento della rete. Ognuno di questi riferimenti e di queste connessioni ha poi un significato ben preciso: sono relazioni semantiche, qualificate per descrivere la natura e il valore delle interazioni.

Ad esempio, una decisione giudiziaria che ne cita un'altra può fare riferimento alla decisione precedente nello stesso caso o alla citazione giurisprudenziale di una sentenza importante. Un giudice può pronunciare una sentenza come conferma o riforma di un altro provvedimento e sulla base di certe disposizioni legislative. Una fonte normativa può determinare la soppressione o la modifica di un'altra e così via. Queste tipologie di *link* ovviamente non hanno lo stesso valore e non forniscono le stesse informazioni. L'informazione giuridica costituisce poi un corpus dinamico molto variabile nel tempo.

Il livello di modellazione semantica codificato nelle ontologie fornisce le basi per il ragionamento tramite assiomatizzazioni e regole logiche esplicite che consentono di inferire nuova conoscenza.

A differenza della conoscenza "immersa" in forma numerica nelle reti neurali, la conoscenza nei grafi semantici può essere letta e interpretata, è spiegabile e può essere facilmente aggiornata, verificata e corretta.

Il paradigma dei dati collegati consente di rispondere a domande di ricerca anche complesse tramite singole *query* [8] poste (da utenti o da agenti software) su un *knowledge graph*, in alternativa all'aggregazione manuale delle risposte a più interrogazioni a database separati.

L'approccio semantico alla descrizione dei dati non è necessariamente in alternativa o in contrapposizione con l'approccio basato su apprendimento statistico o neurale dai dati. Mira piuttosto a mitigarne alcuni limiti, primo fra tutti quello della opacità e della spiegabilità. Molte possibili interazioni e sinergie possono essere sfruttate per coglierne le rispettive potenzialità [9].

Una delle linee di ricerca presso l'Istituto di Informatica Giuridica e Sistemi Giudiziari del Consiglio Nazionale delle Ricerche (IGSG-CNR) riguarda lo sviluppo di standard, strumenti software, dataset curati e arricchiti

---

[8] K. MOODLEY, P. HERNANDEZ-SERRANO, A. ZAVERI, M. SCHAPER, M. DUMONTIER, G. VAN DIJCK, *The Case for a Linked Data Research Engine for Legal Scholars*, in *European Journal of Risk Regulation*, 11(1), 2020, pp. 70-93, doi:10.1017/err.2019.51.

[9] Si veda, in proposito, il punto 5 di questo capitolo su "Prospettiva: integrazione degli approcci statistico e semantico" per una breve descrizione di alcune linee di ricerca e sviluppo.

per la realizzazione di un grafo semantico delle risorse giuridiche (documenti, metadati), anche in collaborazione con le principali istituzioni. Una infrastruttura di dati giuridici aperti, ricchi e strutturati nel pubblico dominio come base di conoscenza per lo sviluppo di applicazioni e servizi.

Nelle prossime sezioni si farà cenno ad alcuni di questi strumenti e risorse e alle loro applicazioni.

### 3.1. *Linkoln*

Nell'infrastruttura dei dati giuridici, l'identificazione stabile e persistente dei documenti o di loro specifiche partizioni, espressa in un formato standardizzato interpretabile in modo non ambiguo da una macchina, è un elemento costitutivo essenziale per la costruzione di applicazioni per migliorarne l'accesso (motori di ricerca, strumenti di analisi).

L'identificazione stabile dei documenti consente poi la risoluzione delle citazioni (legislative, giurisprudenziali) ampiamente utilizzate nei testi giuridici come tecnica di rinvio a norme o precedenti. I collegamenti ipertestuali alle norme citate – preferibilmente risolti al livello di granularità delle specifiche disposizioni – e ai precedenti giurisprudenziali, sono essenziali per migliorare la leggibilità di un testo. Oltre a migliorare la navigazione per gli utenti, l'estrazione e l'annotazione di riferimenti giuridici leggibili dalla macchina come metadati associati ai documenti è un'informazione relazionale essenziale per consentire interrogazioni complesse.

Le tecnologie per l'analisi dei testi e per l'estrazione automatica di informazione consentono di gestire la grande variabilità di espressioni testuali con cui gli elementi costitutivi di un riferimento possono essere espressi in una citazione testuale, in modo da individuare all'interno dei documenti le parti del testo che esprimono una citazione e normalizzarne gli attributi (tipi di documento, autorità, date, numeri …) per ricondurla ad un identificatore univoco per il documento citato.

Il software Linkoln [10] è il risultato di anni di esperienza e di lavoro svolto presso l'IGSG-CNR sul tema dell'estrazione di *link* giuridici da testi scritti in italiano:

– è un software a codice sorgente aperto, sviluppato presso l'Istituto

[10] https://linkoln.gitlab.io/. Si veda anche L. BACCI *et al.*, *Improving Public Access to Legislation Through Legal Citations Detection: The Linkoln Project at the Italian Senate*. *Knowledge of the Law in the Big Data Age*, IOS Press, Amsterdam, 2019, pp. 149-158.

nell'ambito di diversi progetti e collaborazioni istituzionali in ambito nazionale ed europeo;

– consente l'estrazione dei riferimenti ad atti normativi e provvedimenti giurisprudenziali nazionali ed europei;

– aderisce agli standard di identificazione dei provvedimenti giuridici urn:nir ed ELI (European Legislation Identifier) [11] per la legislazione, ECLI (European Case Law Identifier) [12] per la giurisprudenza;

– l'approccio all'analisi testuale è basato su regole in modo da poter avere il controllo completo in ogni fase dell'analisi soprattutto in termini di accuratezza complessiva del software e facilità di estensione a nuovi casi;

– l'analisi viene effettuata in più iterazioni di scansione testuale, individuando di volta in volta entità sempre più complesse, sulla base dei risultati delle iterazioni precedenti.

L'applicazione del software su un corpus di documenti consente di estrarre rapidamente e con elevata affidabilità una quantità di meta informazioni di contesto che possono essere sfruttate in numerose applicazioni, dalla ricerca documentale all'analisi.

### 3.2. *Marker*

La struttura formale della legislazione, e in misura minore della giurisprudenza, sottende una parte non trascurabile della semantica dei documenti. Renderla esplicita tramite l'annotazione con marcatori delle diverse parti del testo e delle loro relazioni (ad esempio gerarchiche) è essenziale non soltanto per migliorare la fruibilità dei testi da parte degli utenti ma soprattutto per mettere a disposizione dell'elaborazione della macchina tale informazione di contesto.

Nel caso della legislazione le diverse parti che compongono il testo sono ad esempio l'intestazione, il titolo, il preambolo, e in particolare l'articolato, con la sua struttura gerarchica, la numerazione delle partizioni, le rubriche, ecc. Nel caso della giurisprudenza ad esempio le parti, l'oggetto, il fatto, i motivi della decisione e il dispositivo.

---

[11] Conclusioni del Consiglio, del 6 novembre 2017, sull'identificatore della legislazione europea https://eur-lex.europa.eu/legal-content/IT/TXT/?uri=CELEX:52017XG1222(02).

[12] Conclusioni del Consiglio sull'identificatore europeo della giurisprudenza (ECLI) e una serie minima di metadati uniformi per la giurisprudenza 2019/C 360/01, https://eur-lex.europa.eu/legal-content/IT/TXT/?uri=CELEX:52019XG1024(01).

La segmentazione dei testi tramite l'annotazione esplicita della loro struttura e della semantica delle parti del documento consente di fornire all'elaborazione della macchina dati più granulari e meno "rumorosi", arricchiti di meta informazioni per tutte le elaborazioni successive basate sui documenti (ad es. indicizzazione, classificazione, trattamento automatico del contenuto testuale).

La rappresentazione dei documenti giuridici con formati standard aperti capaci di coglierne caratteristiche strutturali e di contesto (metadati) è un tema classico nell'informatica giuridica, culminato con la definizione delle specifiche del formato *AkomaNtoso*/*LegalDocML* [13].

Su questa linea presso l'IGSG-CNR è stato sviluppato il software *Marker*, finalizzato all'analisi strutturale dei testi legislativi e alla loro rappresentazione in un formato strutturato *XML* compatibile con *AkomaNtoso*. Tramite l'applicazione di regole di riconoscimento di pattern testuali specializzate sulla struttura dei testi legislativi italiani ed europei, il *Marker* è in grado di identificare automaticamente le partizioni dei testi, la loro relazione gerarchica nell'articolato e ulteriori segmenti significativi e metadati estraibili dai testi annotati esplicitamente con *tag* semantici. Con questo tipo di tecnologia, rilasciata come software *open source* e in combinazione con *Linkoln*, è possibile spingersi in livelli di annotazione semantica esplicita più profonda, ad esempio per l'individuazione delle disposizioni di modifica per la ricostruzione e l'annotazione esplicita delle versioni vigenti del corpus legislativo.

Applicato massivamente sul corpus legislativo italiano consente con elevata precisione e completezza di costruire dataset "puliti" e semanticamente ricchi e fornire le basi per ridurre il rumore nelle applicazioni basate sui testi.

### 3.3. *Linked Legal Data Repository*

L'Unione Europea sta lavorando per armonizzare la descrizione delle risorse giuridiche in tutta Europa, al fine di migliorarne lo scambio transfrontaliero. Questo obiettivo è supportato, ad esempio, da standard come lo *European Legislation Identifier* (ELI) e lo *European Case Law Identifier* (ECLI), che forniscono specifiche tecniche per gli identificatori *web* e raccomandazioni per i vocabolari da utilizzare per descrivere i metadati relativi ai documenti giuridici in un formato leggibile dalla macchina. In particolare, questi standard di metadati ECLI ed ELI aderiscono al formato dati RDF (*Re-*

[13] https://www.akomantoso.org/.

*source Description Framework*) che costituisce la base dei *Linked Data* e hanno quindi il potenziale per costituire una base per un grafo della conoscenza giuridica paneuropeo [14].

Per l'Italia l'IPZS ha implementato l'identificazione secondo lo standard ELI di tutti i provvedimenti pubblicati in Gazzetta Ufficiale e l'attribuzione di un insieme minimo di metadati dal vocabolario ELI.

Uno dei progetti in corso presso l'IGSG-CNR prevede la raccolta e la ripubblicazione dei metadati della legislazione secondo la filosofia dei *Linked Open Data* arricchiti con ulteriori metadati raccolti o estratti da fonti aperte. Sono previsti ad esempio la raccolta e la rappresentazione in conformità con lo stesso modello dei dati ELI, delle relazioni di consolidamento fra atti che consentono di ricostruire e interrogare l'evoluzione temporale degli atti, la stratificazione normativa e l'accesso alle versioni consolidate più recenti o vigenti ad una certa data. Le citazioni fra norme estratte dai testi tramite strumenti automatici come il software *Linkoln* sono a loro volta rappresentate come relazioni fra atti e integrate nel grafo arricchendolo di ulteriori connessioni.

La strutturazione dei testi nei formati di XML legislativo ottenuta ad esempio tramite l'applicazione di strumenti automatici come il software *Marker* consente di integrare nel grafo ulteriori metadati e relazioni più granulari al livello delle partizioni legislative.

Grazie alle tecnologie del *web semantico*, in particolare gli standard per il *web dei dati* URI e RDF, ogni ulteriore aggiornamento, affinamento o aggiunta di metadati e relazioni da fonti diverse è integrato in modo trasparente e immediatamente disponibile per l'interrogazione.

Le relazioni e i metadati, espressi in modo comprensibile alle macchine e agli umani e aventi una semantica esplicita descritta nel modello dei dati, consentono di sviluppare applicazioni basate sull'interrogazione del grafo [15] i cui risultati sono spiegabili, differentemente dall'opacità delle relazioni emergenti dall'apprendimento statistico sui corpora testuali.

Un processo analogo è attualmente più difficile nella giurisprudenza ita-

---

[14] E. FILTZ, S. KIRRANE, A. POLLERES, *The linked legal data landscape: linking legal data across different countries*, in *Artif Intell Law*, 29, 2021, pp. 485-539, https://doi.org/10.1007/s10506-021-09282-8.

[15] V. W. ANELLI *et al*., *Navigating the Legal Landscape: Developing Italy's Official Legal Knowledge Graph for Enhanced Legislative and Public Services in Proceedings of Workshop on AI for the Public Administration Ital-IA 2023*: 3rd National Conference on Artificial Intelligence, organized by CINI, May 29-31, 2023, Pisa, Italy https://ceur-ws.org/Vol-3486/.

liana per la limitata disponibilità di metadati e documenti da fonti aperte, ma alcuni degli elementi costitutivi per l'integrazione delle risorse giurisprudenziali nel grafo della conoscenza giuridica sono già disponibili. Lo standard ECLI prevede identificatori ed uno schema di metadati per la giurisprudenza ed è stato implementato dalle corti apicali italiane nell'ambito di progetti e collaborazioni: Corte costituzionale, Corte di cassazione, Consiglio di Stato e Corte dei conti utilizzano lo standard di identificazione dei provvedimenti e almeno in parte descrivono le proprie risorse coi metadati ECLI. Il caso della Corte costituzionale che pubblica e mantiene aggiornato il proprio portale *Linked Open Data* segna una *best practice* di condivisione dei dati che possono essere direttamente integrati nella più ampia base di conoscenza costituita dal grafo delle risorse giuridiche italiane in corso di costruzione.

### 3.4. *DoGi – Dottrina giuridica*

DoGi – Dottrina Giuridica [16] è una banca dati di riferimenti bibliografici e abstract di articoli pubblicati nelle riviste giuridiche italiane. Creata nel 1970, la banca dati è un prodotto delle attività di ricerca condotte dall'IGSG-CNR in tema di accesso e diffusione dell'informazione giuridica.

Il valore della banca dati rispetto a risorse informative analoghe fornite da editori giuridici commerciali, ma anche prodotte sulla base di iniziative istituzionali volontarie, risiede nel fatto che tale risorsa permette l'accesso alla letteratura non solo attraverso i classici riferimenti bibliografici, ma anche tramite riferimenti normativi e giurisprudenziali citati nel testo dottrinale. Sono proprio la ricchezza e la specificità delle informazioni disponibili le qualità capaci di offrire nuove opportunità per creare relazioni tra diverse entità e risorse. Dando accesso alle fonti normative e giurisprudenziali citate nell'articolo, nonché mettendo a disposizione dell'utente una classificazione altamente specifica per il diritto, l'utente giurista, come anche il cittadino comune hanno l'opportunità di ottenere un quadro generale delle questioni giuridiche ed avere accesso ad una completa documentazione per risolvere il caso in esame [17].

---

[16] http://dati.igsg.cnr.it/dogi.

[17] E. MARINAI, G. PERUGINELLI, *La banca dati DoGi e la condivisione dei dati giuridici: nuovi orizzonti*, in O. BONORA, D. COLTELLACCI, L. GARBOLINO, M.C. PIAZZA, B. PARADISO, A. PERIN, E. SECINARO (a cura di), *Ecosistemi per la ricerca Atti Convegno ACNP/NILDE. Trieste, 22-23 maggio 2014*, EUT Edizioni Università di Trieste, Trieste, 2015, pp. 57-74.

I dati alla base dell'archivio DoGi, prodotti con spogli manuali periodici delle principali riviste giuridiche italiane, prevedono dettagliate schede di metadati associati alle risorse. I metadati non sono soltanto quelli bibliografici dei contributi dottrinali oggetto di spoglio ma anche quelli relativi alle fonti legislative e giurisprudenziali che siano oggetto o citate nei contributi, identificate e descritte secondo gli standard per il *web* menzionati nelle sezioni precedenti e quindi collegati nel più ampio spazio del *web dei dati*.

Ad esempio nello spoglio di una nota a sentenza pubblicata su una rivista giuridica, una delle tipologie di contributo oggetto di spoglio nella banca dati DoGi, saranno riportati oltre agli estremi bibliografici del contributo e della rivista che lo ospita, i riferimenti al provvedimento oggetto della nota e i riferimenti ad altre fonti di legislazione e giurisprudenza italiane ed europea citati nel contributo, resi immediatamente accessibili sia come metadati che come documenti grazie all'utilizzo degli standard di identificazione e descrizione delle risorse giuridiche. Viceversa, la ricerca su un provvedimento di giurisprudenza consentirà di reperire tutti i contributi dottrinali che hanno ad oggetto quel provvedimento.

Il risultato è un ricco *dataset*, curato manualmente con attività di spoglio bibliografico da esperti giuristi e integrato da connessioni da e verso le altre fonti del diritto che forniscono contesto, canali di accesso e interrogazione espressi da relazioni semantiche esplicite che possono essere sfruttati da applicazioni e algoritmi come base di conoscenza curata e validata.

## 4. *Applicazioni*

Gli strumenti e le risorse resi disponibili nel tempo sono stati sviluppati anche in collaborazione e col supporto di diverse iniziative e progetti istituzionali e integrati in banche dati pubbliche come strumenti di supporto all'accesso e al miglioramento della ricerca legislativa e giurisprudenziale.

### 4.1. *Senato della Repubblica e portale Normattiva*

Il Parlamento italiano ha contribuito fin dai tempi del progetto Normeinrete ai temi dell'informatica giuridica, in particolare alla definizione di standard di identificazione e rappresentazione dei testi legislativi e alla loro implementazione sui siti istituzionali. In particolare, la collaborazione dell'isti-

tuto con il Senato della Repubblica è stata particolarmente durevole e fruttuosa.

Il Senato ha promosso lo sviluppo collaborativo e aperto delle prime versioni del software per l'estrazione dei riferimenti *Linkoln*, coinvolgendo un'ampia rete di soggetti istituzionali interessati – autorità emittenti e autorità di pubblicazione – nella condivisione degli obiettivi e nella raccolta dei requisiti, con l'obiettivo di favorire l'adozione, il riuso e il miglioramento iterativo dello strumento. Il primo rilascio pubblico del software nel 2017 è stato l'integrazione nell'applicazione per la visualizzazione degli atti ufficiali (inclusi disegni di legge, emendamenti, dossier) sul sito del Senato.

Più di recente sul portale Normattiva della legislazione vigente, a cura dell'Istituto Poligrafico IPZS, è stata attivata la funzionalità di navigazione ipertestuale delle citazioni legislative grazie all'integrazione delle più recenti versioni del software *Linkoln*.

### 4.2. *Corte di cassazione, Corte costituzionale e ItalGiureWeb*

Anche il Centro Elettronico di Documentazione (CED) della Corte di cassazione ha contribuito storicamente ai temi dell'informatica giuridica in Italia, in particolare con la visione pionieristica di accesso integrato al "dato giuridico globale" realizzata nel sistema ItalGiureWeb.

Più di recente il CED, in collaborazione con l'IGSG, ha implementato fra le prime corti supreme europee lo standard di identificazione ECLI (European Case Law Identifier) per l'identificazione univoca dei provvedimenti della cassazione, l'attribuzione dei metadati prescritti dalle raccomandazioni ECLI e la loro indicizzazione nel motore di ricerca della giurisprudenza nazionale degli Stati membri nel portale e-Justice gestito dalla Commissione UE.

Lo stesso percorso di implementazione dello standard ECLI è stato realizzato dall'IGSG in collaborazione con la Corte costituzionale oltre che con la Corte dei conti il Consiglio di stato per la giurisprudenza contabile e amministrativa.

Anche grazie all'implementazione degli identificatori univoci dei provvedimenti è stato possibile integrare nel sistema ItalGiureWeb gli strumenti di analisi forniti dal software *Linkoln* per l'estrazione e la formalizzazione delle citazioni ai provvedimenti legislativi e ai precedenti giurisprudenziali identificate nei testi dei provvedimenti della Cassazione, consentendo una ulteriore interconnessione degli archivi.

## 4.3. *Archivio della giurisprudenza di merito*

In un progetto volto alla ricostituzione di una banca dati della giurisprudenza di merito integrata nel sistema ItalgiureWeb, il Consiglio Superiore della Magistratura, supportato da un gruppo di lavoro di magistrati ha definito alcune priorità e linee guida per la selezione dei provvedimenti.

Il progetto [18] mira a ricostituire uno strumento efficace, finalizzato alla diffusione della giurisprudenza di merito, onde assicurare lo scambio e dunque la circolarità delle informazioni su materie rilevanti fra i giudici dei diversi distretti sul territorio nazionale, garantire un costante dialogo bidirezionale tra la giurisprudenza di legittimità e quella di merito, nonché dare modo agli operatori del diritto di avere un quadro completo della giurisprudenza non solo di legittimità, ma anche di merito su questioni d'interesse.

Il ricostituendo Archivio Merito potrà assicurare il tempestivo accesso alle prime letture ed applicazioni concrete delle novità normative – nazionali e comunitarie – e giurisprudenziali, sia delle Corti nazionali (in particolare, le Sezioni Unite della cassazione e la Corte costituzionale), sia delle Corti sovranazionali (in particolare, la Corte di Giustizia dell'Unione Europea e la Corte Europea per i diritti dell'uomo).

In linea generale, nell'archivio potranno essere inseriti tutti i provvedimenti di merito aventi ad oggetto: decisioni che facciano applicazione di disposizioni di nuova introduzione; decisioni che costituiscano prima applicazione di orientamenti innovativi della Corte di cassazione, a maggior ragione se espressi a Sezioni Unite; decisioni che costituiscano prima applicazione di pronunce della Corte costituzionale; decisioni che facciano applicazione di disposizioni comunitarie di nuova introduzione e dei principi espressi dalle Corti sovranazionali; decisioni riguardanti materie normalmente non oggetto di pronunce di Cassazione, a condizione che nell'archivio non siano già presenti provvedimenti dello stesso distretto sulla medesima materia e di analogo tenore; decisioni che costituiscano espressione di soluzioni concrete adottate dai giudici della cognizione su temi decisori particolarmente rilevanti; provenienza geografica omogenea da tutto il territorio nazionale (raccolta a livello di distretti di Corte d'appello); rappresentatività rispetto ad una classificazione per materie sia di area civile che penale.

---

[18] Con delibera del 31 ottobre 2017 (e con successivo provvedimento in data 9 maggio 2018 modificato in data 12 settembre 2018) il Plenum del CSM ha approvato le linee guida volte alla individuazione delle modalità di ricostituzione di una banca dati della giurisprudenza di merito di ItalgiureWeb.

Infine, rilevato che la giurisprudenza di merito è tanto più rilevante quanto più è nuova ed ancora che generalmente conserva la sua attualità fintanto che non intervenga una pronuncia della Corte di cassazione che la recepisca, appare necessario anche introdurre un meccanismo di cancellazione automatica dei provvedimenti di merito dopo l'intervento della Corte e comunque dopo che l'orientamento espresso risulti ormai superato e, non più utile, a fini di approfondimento giurisprudenziale e scientifico.

Tralasciando per il momento il dibattito fra i giuristi relativo agli effetti e alle implicazioni prodotti dalla visibilità dei provvedimenti selezionati (e lo speculare oblio di quelli non selezionati), da un punto di vista informativo i criteri di selezione individuati coinvolgono attributi dei provvedimenti in larga misura estraibili automaticamente e resi disponibili come metadati secondo i modelli semantici di descrizione delle risorse giuridiche menzionati. I criteri di selezione coinvolgono proprietà e relazioni tra fonti eterogenee (giurisprudenza e legislazione nazionale ed europea, attributi temporali, decisioni in successivi gradi di giudizio) la cui interrogazione integrata è essenziale per individuare i provvedimenti rilevanti secondo i parametri prescritti.

Nell'attività di supporto alla selezione del nucleo iniziale dei provvedimenti di merito fornita al gruppo di lavoro costituito presso il CSM dall'IGSG, a fianco delle attività di spoglio e revisione manuale della selezione, si è cercato di sfruttare il più possibile l'automazione consentita dalle informazioni di contesto fornite dai metadati dei provvedimenti e da risorse come il dataset DoGi utilizzando fra gli altri, come indicatore di rilevanza e novità dei provvedimenti di merito, l'attenzione loro rivolta dagli studiosi del diritto nei propri contributi dottrinali.

### 4.4. *Banca dati della Documentazione Economica e Finanziaria*

Uno degli impulsi più recenti allo sviluppo e alla reingegnerizzazione dei software *Linkoln* e *Marker* sono stati l'estensione e l'adattamento nell'ambito di una collaborazione per l'evoluzione della piattaforma di redazione e pubblicazione del servizio di documentazione a cura del Centro di Ricerche e Documentazione Economica e Finanziaria del Ministero dell'Economia e delle Finanze (CeRDEF). Il servizio cura la pubblicazione gratuita e aggiornata sulle specifiche tematiche riguardanti la normativa nazionale, regionale e comunitaria, in materia economico-finanziaria e fiscale. Ogni provvedimento è riportato in tutte le eventuali diverse versioni, così come modificate nel

tempo, con l'indicazione dei relativi periodi di vigenza. La banca dati include la prassi amministrativa costituita dalle circolari e risoluzioni ufficiali emesse dalle agenzie fiscali e dal Dipartimento delle finanze, dalle circolari emesse da altri enti o amministrazioni pubbliche a contenuto economico-fiscale pubblicate nella Gazzetta ufficiale, dai comunicati stampa emessi dall'Agenzia delle entrate oltre alla giurisprudenza italiana e dell'Unione Europea in materia economico-finanziaria e fiscale.

I software forniti da IGSG sono stati integrati nel flusso editoriale di redazione per consentire l'arricchimento della documentazione pubblicata. La marcatura strutturale di tutte le tipologie di atto di interesse agevola i redattori nell'annotazione delle diverse versioni così come modificate nel tempo, con indicazione delle relative vigenze. Il riconoscimento delle citazioni e il collegamento ipertestuale fra i diversi atti oltre a migliorare la navigazione per gli utenti consente di offrire canali di ricerca più evoluti sul corpus documentale tematico basati su meta informazioni corrette e verificate.

## 5. *Prospettiva: integrazione degli approcci statistico e semantico*

Come già accennato, i sistemi basati su modelli neurali del linguaggio (*LLM*) funzionano molto bene nella comprensione e nella generazione del linguaggio naturale, sono integrati in interfacce conversazionali che migliorano significativamente l'esperienza degli utenti ma sono sistemi opachi, non sono in grado di spiegare su quali fatti basino le proprie risposte, sono deboli nella logica e nel ragionamento, sono difficili da mantenere e aggiornare.

I dati semantici, aperti e di buona qualità, organizzati in strutture concettuali esplicite all'interno di un grafo della conoscenza, sono fra i candidati ideali per superare o mitigare molti dei limiti dei sistemi basati su apprendimento statistico e modelli neurali del linguaggio. Un grafo modellato in *RDF* può rappresentare il significato con ricchezza e flessibilità illimitate. Come insieme di fatti curato, può fornire vincoli a un *LLM* che può migliorare la qualità del suo *output*. È affidabile perché è possibile sapere quali informazioni contiene, i fatti che ha incorporato in esso, come può rispondere alle domande e cosa può dedurre o ragionare.

I grafi della conoscenza sono quindi complementari agli *LLM* e una delle direzioni più interessanti di sviluppo nelle tecnologie di intelligenza artifi-

ciale generativa riguarda lo studio di come farli funzionare insieme [19]. Come combinare approcci simbolici basati sulla modellazione esplicita della conoscenza, e approcci sub-simbolici, basati su rappresentazioni numeriche dei significati emergenti in modelli neurali. È possibile sfruttare l'interazione fra i due approcci in modi diversi. Ad esempio:

– utilizzare un *LLM* per creare o arricchire di entità e relazioni un grafo della conoscenza. In questo caso si utilizzano le funzionalità di elaborazione del linguaggio naturale dei modelli del linguaggio per elaborare un grande corpus di dati di testo. Si chiede quindi al *LLM* (che è opaco) di produrre un grafo della conoscenza (che è trasparente). Il grafo della conoscenza può essere ispezionato, sottoposto a controllo di qualità e curato.

– utilizzare un grafo della conoscenza per addestrare un *LLM*. Questo è l'approccio inverso: invece di addestrare i *LLM* su un ampio corpus generale, si specializza l'addestramento esclusivamente sul grafo della conoscenza di dominio. In questo modo è possibile costruire *chatbot* specializzati molto abili nel dominio di interesse e servizi che rispondono senza "allucinazioni".

– utilizzare un grafo della conoscenza nell'interazione con un *LLM* per arricchire query e risposte. In questo caso si intercettano i messaggi in entrata e in uscita dal *LLM* e si arricchiscono con i dati del grafo della conoscenza. I dati memorizzati nel grafo funzionano come contesto aggiuntivo nelle domande ai modelli neurali e come vincolo fattuale per le risposte.

La ricerca in queste direzioni appare molto interessante per lo sviluppo di strumenti evoluti in funzione ausiliaria in ambito giuridico che combinino la naturalezza dell'interazione in linguaggio naturale con l'affidabilità e la verificabilità della base di conoscenza fattuale su cui il sistema basa le proprie risposte.

## 6. *Conclusioni*

Dalla discussione svolta fino a qui, emerge come la disponibilità dei dati, la loro qualità e la loro completezza siano essenziali per la costruzione di sistemi automatici qualunque sia l'approccio tecnologico utilizzato.

---

[19] Unifying Large Language Models and Knowledge Graphs: A Roadmap Shirui Pan, Linhao Luo, Yufei Wang, Chen Chen, Jiapu Wang, Xindong Wu https://arxiv.org/abs/2306.08302.

La creazione di una infrastruttura di dati pubblici aperti, interconnessi, contestualizzati e di buona qualità costituirebbe di per sé una base affidabile e verificabile per la costruzione di applicazioni capaci di automatizzare determinate attività della pratica del diritto o quantomeno assistere nel reperimento dell'informazione giuridica corretta e aggiornata.

Analogamente i *dataset* utilizzati per l'addestramento dei modelli statistici e neurali e i *benchmark* di valutazione dell'accuratezza dei loro risultati, dovrebbero essere resi disponibili come dati aperti in modo da consentire valutazioni indipendenti e riproducibilità dei risultati.

Le accuratezze dei risultati promesse dai sistemi basati sull'apprendimento dai dati sono spesso da prendere con cautela. Revisioni sistematiche [20] hanno evidenziato una "crisi della riproducibilità" dei risultati scientifici basati sul *machine learning* già prima dell'avvento dei grandi modelli linguistici. In molti casi il problema era dovuto al c.d. *data leakage*: quando i dati di valutazione non sono indipendenti dai dati di addestramento, si ottengono risultati eccessivamente ottimistici. Un fenomeno osservato anche nel campo della "predizione automatica delle decisioni giudiziarie" [21].

In questo senso sono degne di nota iniziative come lo strumento *Typology of Legal Technologies* [22] che propone una metodologia sistematica di valutazione multidisciplinare di una selezione di strumenti di *legal tech,* della fondatezza delle affermazioni fatte dai loro sviluppatori e in particolare del tipo di effetto giuridico che la loro implementazione potrebbe avere. Oppure *LegalBench* [23], un benchmark per la valutazione di quali tipi di ragionamento giuridico siano in grado di svolgere i modelli *LLM*, costruito in modo collaborativo attraverso un processo interdisciplinare, che raccoglie compiti progettati e realizzati manualmente da giuristi.

Le pratiche di apertura dei dati, del software, delle pubblicazioni (*Open Data*, *Open Source, Open Access*) racchiusi in campo scientifico dal paradigma della scienza aperta (*Open Science*) [24], risultano essenziali anche nei

---

[20] REFORMS: Reporting Standards for ML-based Science https://reforms.cs.princeton.edu/.

[21] M. MEDVEDEVA, M. WIELING, M. VOLS, *Rethinking the field of automatic prediction of court decisions*, in *Artif Intell Law*, 31, 2023, pp. 195-212, https://doi.org/10.1007/s10506-021-09306-3.

[22] L. DIVER, P. MCBRIDE, M. MEDVEDEVA, A. BANERJEE, E. D'HONDT, T. DUARTE, D. DUSHI, G. GORI, E. VAN DEN HOVEN, P. MEESSEN, M. HILDEBRANDT, *'Typology of Legal Technologies'* (COHUBICOL, 2022), available at https://publications.cohubicol.com/typology.

[23] LegalBench: A Collaboratively Built Benchmark for Measuring Legal Reasoning in Large Language Models https://doi.org/10.48550/arXiv.2308.11462.

[24] La "scienza aperta" è un approccio al processo scientifico basato su collaborazione,

settori dell'informatica giuridica e della "scienza giuridica computazionale" che si voglia confrontare con i principi di funzionamento dei "sistemi intelligenti" applicati al diritto.

Analogamente l'effettiva adozione dei principi dell'*Open Data* nelle istituzioni e nelle amministrazioni pubbliche, prescritta dalla normativa relativa all'apertura dei dati e al riutilizzo dell'informazione del settore pubblico – con tutti i limiti dovuti alla tutela di dati personali e alle altre garanzie – è essenziale oltre che per liberare un potenziale di sfruttamento pubblico del patrimonio informativo e per il miglioramento dei servizi, per il progresso della ricerca scientifica nella disciplina. L'accesso ai dati per finalità di ricerca è in questo settore un indispensabile strumento di controllo dell'integrità della ricerca oltre che un requisito essenziale per il controllo dei sistemi e servizi commerciali adottati per finalità pubbliche.

Un terreno comune di dialogo e una comprensione reciproca fra studiosi – dei presupposti, del funzionamento e delle implicazioni delle tecnologie digitali fra gli studiosi del diritto; dei principi, delle pratiche e delle tutele del diritto fra gli studiosi di informatica – sono sempre più necessari per mantenere il controllo e direzionare l'impatto delle tecnologie digitali sul mondo giuridico, ai diversi livelli e nei diversi ruoli, in particolare nell'ambito dell'amministrazione della giustizia e della funzione decisionale.

---

condivisione aperta e tempestiva dei risultati, modalità di diffusione della conoscenza basate su tecnologie digitali in rete e metodi trasparenti di validazione e valutazione dei prodotti della ricerca. Piano Nazionale per la Scienza Aperta: https://www.mur.gov.it/sites/default/files/2022-06/Piano_Nazionale_per_la_Scienza_Aperta.pdf.