Rivista Telematica di
Diritto Tributario

Rivista semestrale — *Fascicolo monografico* 2022

LOADING PRODIGIT
Dal diritto naturale al diritto digitale: l'intelligenza artificiale nella giustizia tributaria

a cura di
Alberto Marcheselli
e Enrico Marello

ISSN 2499-2569

Pacini Giuridica

Rivista Telematica di Diritto Tributario

Segreteria di redazione: Gloria Giacomelli
ggiacomelli@pacinieditore.it
Phone +39 050 31 30 243 - Fax +39 050 31 30 300

Amministrazione:
Pacini Editore Srl, via Gherardesca 1, 56121 Pisa
Tel. 050 313011 • Fax 050 3130300
www.pacinieditore.it • abbonamenti_giuridica@pacinieditore.it

In corso di registrazione presso il Tribunale di Pisa
Direttore responsabile: Patrizia Alma Pacini

LOADING PRODIGIT
DAL DIRITTO NATURALE AL DIRITTO DIGITALE: L'INTELLIGENZA ARTIFICIALE NELLA GIUSTIZIA TRIBUTARIA

A CURA DI ALBERTO MARCHESELLI E ENRICO MARELLO

CAPITOLO I
QUALE GIUSTIZIA PREDITTIVA?

CAPITOLO II
PRODIGIT E LA GIUSTIZIA TRIBUTARIA

CAPITOLO III
PRODIGIT E LA CONOSCIBILITÀ DEL DIRITTO

CAPITOLO IV
DIRITTO MATEMATICO E GIUSTIZIA PREDITTIVA

Prodigit come banca dati intelligente

Prodigit: an intelligent information retrieval system

Enrico Marello

Abstract
Prodigit può essere una grande opportunità per sviluppare un banca dati costruita intorno all'intelligenza artificiale. Nel contributo si analizzano alcuni dei temi principali che possono essere affrontati nella progettazione di una banca dati intelligente.
Parole chiave: Prodigit, banche dati, intelligenza artificiale

Abstract
Prodigit is a great opportunity to develop a database built around artificial intelligence. The article deals with some of the main issues that can be addressed in the design of an intelligent database.
Keywords: *Prodigit, information retrieval, artificial intelligence*

1. In questo contributo si vuole provare a tratteggiare non il tormentato tema della giustizia predittiva, ma piuttosto un tema meno trattato nella pubblicistica nel dibattito sul progetto italiano Prodigit: quello relativo all'*Information Retrieval.*

La funzione di Information Retrieval è, usando i termini burocratici, quella relativa alla «*implementazione della Banca dati nazionale di giurisprudenza di merito accessibile gratuitamente e pubblicamente dal sito Internet istituzionale del CPGT*» (linea di intervento n. 3).

Nell'ambito della digitalizzazione dell'informazione giuridica, il tema delle banche dati è uno dei più rilevanti: all'interno della Relazione sull'amministrazione della giustizia, redatta dal Ministero della Giustizia per l'inaugurazione dell'anno giudiziario 2023, la giustizia predittiva è assente, mentre l'importanza della realizzazione delle banche dati viene evidenziata in almeno in tre luoghi (relazione reperibile qui: https://www.giustizia.it/giustizia/it/mg_2_15_4.page riferimenti alle p. 579-580, 584, 597-598).

Come si vedrà nel seguito, in questo esercizio di costruzione della banca dati c'è lo spazio per un ampio uso dell'intelligenza artificiale (di seguito anche: IA), in modo molto più affidabile e costruttivo (e utile, verrebbe da dire), rispetto a quanto potrà avvenire nell'oracolare campo della predizione.

In ambito scientifico ci si può poggiare su un'ampia letteratura sul tema del *Legal Information Retrieval*, mossa anche dalle possibilità applicative (e commerciali) offerte dal tema di indagine (per una recente sintesi della letteratura sul tema: Sansone C. - Sperlí G., *Legal Information Retrieval systems: State-of-the-art and open issues*, in *Inf. Systems*, 2022, 106, 101967; per un'introduzione ai temi classici trattati da decenni Moens M.F., *Innovative techniques for legal text retrieval*, in *Artif. Intell. Law*, 2001, vol. 9, 29 ss.).

Il cuore del dibattito sull'Information Retrieval negli ultimi decenni è costituito dalla ricerca di metodi per giungere a un'informazione *semantica*: si progettano sistemi che consentano di porgere all'utente (nel nostro caso: un interprete della norma giuridica) un'informazione che offra un significato *contestuale*. Quindi, un'informazione che non sia solo apprezzabile in termini di correlazione tra documenti (la sentenza x e la sentenza y sono correlate), ma un'informazione che offra una dimensione contestuale, presentando nessi e relazioni formalizzate con altri caratteri rilevanti dell'informazione di partenza (per esempio, dalla sentenza x si può ricavare un *corpus* di sentenze correlate in diversi modi alla sentenza di partenza, quanto a esito, a contenuto, a disposizioni richiamate, a richiami alla medesima giurisprudenza ecc.). In questa direzione, al momento, le banche dati presenti sul mercato sono ancora decisamente poco performanti e Prodigit potrebbe fornire un ausilio nel miglioramento dell'intero settore.

In questo breve scritto cercherò di elencare alcune caratteristiche che potrebbero essere interessanti per l'interprete e alcuni caveat di metodo che potrebbero aiutare nella costruzione di questo sistema.

2. La prima scelta strutturale tocca i soggetti cui si rivolge la banca dati: da chi sarà utilizzata?

Possiamo immaginare tre destinatari diversi della banca dati:

a) giudici;
b) giudici e difensori (includendo ovviamente gli enti impositori);
c) pubblico indistinto.

I pochi documenti ufficiali relativi a Prodigit evocano la formula già riportata al paragrafo precedente: "Banca dati nazionale di giurisprudenza di merito accessibile gratuitamente e pubblicamente", il che fa supporre che si vada verso un'applicazione del modello 3.

Una banca dati di giurisprudenza tributaria di merito accessibile pubblicamente implica che vi sia: *(i)* una certa quota di utenti (che possiamo immaginare minoritaria) costituita da non addetti ai lavori, ossia contribuenti comuni che intendono avere qualche informazione su argomenti di proprio interesse e *(ii)* una quota maggioritaria di addetti ai lavori (giudici, difensori) con l'obiettivo principale di recuperare informazioni rispetto a una fattispecie che devono trattare.

I due canali implicano interfacce di consultazione differenti, come si dirà oltre al par. 8.

Quanto alla quota di utenti maggioritaria, ossia di utenti "tecnici", il primo passo procedurale dovrebbe essere quello di consultare ampie rappresentanze di giudici e difensori.

L'assunto di base è che: «*Any legal information retrieval system built without sufficient knowledge, not just of the actual legal information needs but also of the 'juristic mind', is apt to fail*» (van Opijnen M. - Santos C., *On the concept of relevance in legal information retrieval*, in *Artif. Intell. Law*, 2017, vol. 25, 66).

È anche stato rilevato che in questa fase è necessario comprendere non solo come i soggetti cercano, ma anche cosa cercano, per definire la loro definizione di rilevanza (Dadgostari F. - Guim M. - Beling P.A. - Livermore M.A. - Rockmore D.N., *Modeling law search as prediction*, in *Artif. Intell. Law*, 2021, vol. 29, 7; sulla rilevanza si veda ancora oltre, al par. 5).

Il compito dell'inchiesta iniziale dovrebbe essere la formalizzazione dei modi attraverso cui l'utente compie il proprio lavoro interpretativo: quali sono i percorsi logici ed ermeneutici che compie, come questi percorsi si integrano nella ricerca di informazioni precedenti, quali sono i dati che vengono utilizzati più di frequente ecc. È ovvio che giudici e difensori abbiano percorsi logici differenti, in momenti processuali differenti: l'informazione giuridica di cui ha bisogno un difensore nel momento in cui prepara l'atto introduttivo è differente da quella di cui necessita lo stesso difensore al momento di redigere le memorie per replicare alla controparte ed è ancora differente da quella che necessita il giudice quando deve decidere, per verificare e integrare le informazioni giuridiche affastellate dalle parti nel processo.

Quindi, la formalizzazione iniziale dovrebbe tenere conto della pluralità di percorsi logici dei soggetti processuali e delle diverse fasi in cui la ricerca giuridica diviene necessaria.

Perché questa inchiesta abbia una possibilità di successo, dovrebbe essere ampia, per riuscire a intercettare le molte variazioni sul tema, raggrupparle e sistematizzarle. Questo è un passaggio delicato,

perché le applicazioni del digitale al processo tributario non hanno sinora trovato una grande propensione all'interrogazione degli *stakeholders*.

Il processo tributario telematico (PTT) ne è un esempio rilevante: il sistema è solido e tecnicamente ben costruito, ma è in parte inadeguato a soddisfare alcuni bisogni di base dei difensori. Chiunque abbia provato altri sistemi processuali digitali (anche solo il PCT), sa che in altri sistemi si sono pensate funzionalità importanti per il difensore, che nel PTT sono o non implementate o nascoste nelle pieghe del sistema (per esempio, la possibilità di accedere in una sola interfaccia a tutte le controversie pendenti e non ancora definite (magari distinte per organo giudicante), la possibilità di distinguere nel fascicolo di parte tra i diversi tipi di atti, raggruppandoli in tab autonome (ricorsi, memorie, documenti), una visione nidificata e maggiormente leggibile della scansione temporale del fascicolo d'ufficio ecc.).

Poiché non si tratta di difficoltà tecniche, l'unica spiegazione è che il PTT sia stato realizzato senza una grande interrogazione dei soggetti coinvolti e che comunque neppure nell'implementazione si sia pensato a qualcosa di differente da un form con domande generiche.

È da diversi anni che potenti algoritmi di Information Retrieval sono conosciuti e utilizzabili, per cui è ragionevole domandarsi come mai le banche dati giuridiche continuino a essere così poco performanti. Una delle spiegazioni è proprio in questo ambito: si trascura il fatto che che generalmente i giuristi non vogliono interrogare il sistema per trovare una risposta pre-confezionata, ma vogliono trovare un *corpus* di informazioni su cui compiere un proprio esercizio interpretativo (se vogliamo: gli informatici, spesso, non hanno compreso che le parti processuali vogliono esercitare il *potere* di interpretare).

Come detto, il rimedio è semplice: occorre selezionare un bacino ampio di utenti e interrogarli su base qualitativa e personalizzata, per poter formalizzare i modi del ragionamento indicati sopra. Questo passaggio è decisivo per la buona riuscita del sistema.

Se in Prodigit ci si limiterà a una costruzione calata dall'alto, sulla base di una formalizzazione del percorso di ricerca derivato da poche esperienze, interne esclusivamente al progetto, esiste il rischio che si costruisca una banca dati tecnicamente anche perfetta, ma sostanzialmente lontana dall'interesse concreto dei soggetti processuali. Pur in presenza di grandi risorse computazionali, di dati ottimi e di algoritmi performanti, se non si ha una nitida rappresentazione della "mente giuridica" e del suo polimorfismo, il sistema produrrà un *output* di scarsa utilità.

3. Dai soggetti possiamo passare agli oggetti. Quali dati saranno contenuti nella banca dati?

Per quel che è noto, da qualche tempo dovrebbe essere già disponibile per i soli giudici una banca dati contenente tutte le sentenze di merito. Questi dati, però, sono in larga parte di difficile accesso, perché ancora veicolati da *file* pdf-immagine. La conversione in forma elaborabile e cercabile di tutti questi documenti non so se sia in programma: richiederebbe ancora una dose di controllo umano notevole (per evitare che i programmi di conversione scambino oggettivo per soggettivo o IRPEF per IRPEG o qualsiasi altro errore che venga in mente a chi ha adoperato i programmi anche più evoluti di riconoscimento del testo oggi in circolazione), che non so se per i tempi e i costi rientri nel programma di Prodigit. Mettere in programma questa conversione sarebbe un esercizio comunque utile per la migliore e completa formazione del dato cercabile.

La banca dati sarà vincolata dal dato tecnicamente utilizzabile, come spesso avviene. Ossia, il dato (sentenza) sarà inserito in banca dati solo se rispondente ad alcune caratteristiche tecniche: nel nostro caso, sarà inseribile il dato costruito in forma nativa digitale.

Si può immaginare che da oggi o forse comunque in un immediato futuro questo porterà all'immissione in banca dati di *tutte* le sentenze tributarie di merito.

Fintanto che questo non accadrà, la banca dati Prodigit avrà, fatte le dovute proporzioni, una struttura logica simile a quella di *def.finanze.it*, ossia opererà una selezione: saranno inserite solo una parte delle decisioni pronunciate.

L'inserimento parziale pone evidentemente un problema di campionamento, ben conosciuto in statistica. Se si inserisce solo una parte delle sentenze, queste dovrebbero essere rappresentative della popolazione delle sentenze: se il campione non è rappresentativo, l'operazione è falsata.

Occorre quindi che sia trasparentemente comunicato il modo del campionamento e la rappresentatività del campione. Se il campione è casuale e non ponderato, l'affidabilità complessiva della banca dati si riduce di molto.

Su questi profili ho già scritto qualche riga ponendo delle domande che sinora sembrano senza risposta, per cui mi permetto di rinviare a quel contributo, per non ripetermi (Marello E., *Prodigit: alcune domande di metodo e qualche semplice risposta*, in questa *Rivista*, 1° febbraio 2023).

4. Sgombrato il campo dalle questioni preliminari, spostiamoci verso alcune funzioni e alcune strutture che potrebbero rendere interessante questa banca dati.

Abbiamo visto sopra che la moderna Information Retrieval punta sulla contestualità dell'informazione. Un modo forse debole, ma già utile, di intendere la contestualità è rispetto ai riferimenti formali contenuti nel documento-sentenza.

Ogni decisione contiene un'ampia gamma di citazioni: a dati esterni al procedimento, come disposizioni di legge e precedenti giurisprudenziali e a dati interni al processo, come i documenti depositati dalle parti.

Un primo obiettivo apprezzabile sarebbe quello di rendere percorribili *tutti* i riferimenti contenuti nella sentenza: ogni qual volta una decisione richiama riferimenti interni o esterni, dovrebbe essere possibile raggiungere il documento citato con un solo *clic* (e, come si dirà al par. 8, sarebbe utile avere anche la possibilità di scaricare in blocco parti selezionabili di questo albero di citazioni).

Anche le migliori banche dati private, oggi, hanno un grado di percorribilità limitato: non sono raggiungibili, di solito, tutti i riferimenti compiuti dal documento che la banca dati offre in risposta alla ricerca. Prodigit potrebbe offrire un contributo rilevante se rendesse completa questa percorribilità: un esercizio utile di IA, teso a riconoscere i riferimenti e ad arricchirli con il *link* al documento contenuto nella banca dati.

Una volta costruita una completa rete citazionale, si potrebbe pensare se offrire agli utenti anche informazioni quantitative rispetto alle citazioni, come avviene per esempio in questi due casi già in produzione da diversi anni:

- una delle banche dati più utilizzate nella ricerca giuridica negli USA (Heinoline), offre la possibilità di vedere quanto la sentenza elencata tra i risultati di ricerca sia citata da altre sentenze (e offre la possibilità di accedere, con un solo clic, alla lista di sentenze che citano il caso: è la costruzione usabile di un citation *network*, altro esercizio utile per l'intelligenza artificiale);
- altre banche dati, oltre a questo servizio, consentono anche di comprendere se il caso citato appartenga ancora a un filone giurisprudenziale prevalente rispetto alle corti superiori o se, invece, sia divenuto giurisprudenza minoritaria o superata (per esempio, Westlaw offre Keycite, che ha queste caratteristiche; per l'applicazione italiana in Italgiure v. il contributo di Mastroiacovo V., *Il progetto Prodigit e il paradigma della comodità*, in questo fascicolo telematico).

Queste ultime funzionalità, peraltro, dovrebbero essere discusse pubblicamente, perché sono in grado di mutare considerevolmente il modo nel quale la decisione del caso concreto viene assunta: offrendo una patente di "popolarità" della sentenza citata possono influenzare marcatamente il decisore. Per quello che interessa qui, merita ribadire che le reti citazionali offrono grandi opportunità di arricchimento della ricerca giuridica.

5. Al par. 1 ho messo in luce la centralità dell'informazione semantica nei recenti dibattiti sull'*Information retrieval.* Si introduce così un concetto chiave: la rilevanza dell'informazione proposta dalla banca dati. Con il crescere del dato giuridico (pensiamo al "milione" di sentenze promesso da Prodigit), il punto nodale diventa trovare il dato più rilevante rispetto alla necessità di chi interroga la banca dati.

Sul concetto di rilevanza nell'Information retrieval vi è un dibattito che è ampio, articolato e rozzo allo stesso tempo (sulla difficoltà di definire in cosa consista la rilevanza: van Opijnen M. - Santos C., *op. cit.,* 70 ss.).

Vi è una notevole tendenza al riduzionismo, forse derivata dall'uso dell'idea di rilevanza estrapolata da altri ambiti di sviluppo *dell'Information retrieval*. In particolare, il concetto di rilevanza sviluppato nell'*e-commerce* è relativamente semplice da individuare, ma sembra abbastanza lontano dal significato di rilevanza che potrebbe assumere il difensore o il giudice in un processo.

Qual è l'informazione giuridica rilevante in una banca dati di giurisprudenza come Prodigit?

Non si può dare una risposta univoca: la maggior parte degli studi adotta una soluzione unitaria o al massimo riferita a un paio di scenari. La verità è che esistono decine di accezioni di possibili attuazioni della rilevanza dell'informazione, che dipendono dalla funzione della ricerca.

Se un difensore sta cercando una sentenza riferita a un caso molto ben formalizzato (per esempio, trattamento fiscale dei dividendi nell'IRAP) e si trova nella fase di introduzione del giudizio, può darsi che l'informazione più rilevante sia quella contenuta nella più recente e più alta (intesa come emessa dall'organo più elevato presente in banca dati) sentenza che descrive correttamente la fattispecie per come interpretata dal difensore; nello stesso contesto e con una prospettiva differente, l'informazione più rilevante potrebbe essere un'informazione di contesto, ossia comprendere l'articolazione del *complesso* delle sentenze elaborate sul tema: su quali aspetti si concentrano? Con quali motivazioni? Con quali esiti? Ancora nello stesso contesto, per altri fini, potrebbe essere rilevante conoscere la giurisprudenza del giudice cui ci si deve rivolgere (per esempio, la giurisprudenza sul punto della Corte di giustizia tributaria di primo grado di Torino, anche se non recente e anche se difforme da quella nazionale).

Questa è sola una piccola variazione su un punto di partenza abbastanza univoco (introduzione del giudizio, fattispecie formalizzata chiaramente).

Elenco qui alcuni fattori che, autonomamente e composti tra loro, cambiano radicalmente, nella prospettiva dell'interprete, la percezione di cosa sia rilevante in una banca dati di giurisprudenza:

(i) formalizzazione chiara della fattispecie;
(ii) fase del giudizio in cui ci si trova: introduzione, cautelare, repliche, trattazione;
(iii) funzione svolta: giudice, avvocato, consulente tecnico;
(iv) ricerca svolta con riferimento ad argomenti o rispetto a riferimenti normativi e giurisprudenziali già individuati.

Merita notare che tanto più ampia la banca dati, tanto più potente è l'algoritmo che controlla la rilevanza. Ho usato "potente" nel senso proprio, di aggettivo che indica la disposizione del potere. Tanto più cresce la banca dati, tanto meno questa è facilmente percorribile, tanto più ci si deve affidare alle ricerche per rilevanza e similarità. In questo ambito l'algoritmo che seleziona e individua decide (ecco il potere) la vita e la fortuna delle decisioni pronunciate: se l'algoritmo, a ragione o a torto, ritiene poco rilevante una decisione, questa finirà tra i risultati in coda. Come sappiamo bene dal quotidiano uso dei motori di ricerca generalisti, l'utente si accontenta generalmente delle prime pagine, soprattutto quando i risultati sono molti: il *ranking* attribuito dall'algoritmo governa la diffusione dell'informazione.

Il problema è forse limitabile consentendo lo scaricamento a pacchetti indicato sopra, magari accompagnato da più indicizzazioni del materiale, orientate a diversi criteri (e non da una sola indicizzazione di rilevanza). Poiché le classificazioni sono un esercizio tipico da intelligenza artificiale, ecco un altro uso utile che si potrebbe dare alle risorse del progetto: proporre la possibilità di definire in maniera personalizzabile i criteri di rilevanza (e quindi l'*output* dell'indicizzazione).

Per favorire la flessibilità delle interpretazioni si potrebbe poi ancora pensare a un sistema in cui le ricerche per concetti portassero anche a diversi *output* a seconda dell'accezione più frequente in cui questo viene assunto (come descritto in Šavelka J. - Ashley K.D., *Legal information retrieval for understanding statutory terms*, in *Artif. Intell. and Law*, 2022, vol. 30, 248).

Anche in questa prospettiva, sembra quindi che plurale sia meglio che singolare e che la mappatura della pluralità di esigenze sia una buona pratica da perseguire.

6. Veniamo a un tema complesso, su cui in questa sede mi limiterei a qualche suggestione, per tornarci successivamente in uno scritto dedicato: i rapporti tra la grande banca dati, le massime e sistemi di indicizzazione repertoriale (fondati tradizionalmente su lemmari e attribuzione del dato a una o più

voci del lemmario). Per una ricca analisi strutturale del sistema della massimazione richiamo ancora il contributo di Mastroiacovo V., *Il progetto Prodigit e il paradigma della comodità*, in questo numero monografico).

Massimazione e indicizzazione repertoriale nascono in un mondo analogico dove i parametri di riferimento erano questi: *(i)* la pubblicazione in esteso in un unico contenitore di tutto il dato giuridico era particolarmente difficile; *(ii)* l'indicizzazione cercava di ottenere una mappatura sintetica dell'intero dato giuridico; *(iii)* l'indicizzazione offriva un *output* per la ricerca che per l'interprete era articolata in almeno tre fasi: scrutinio dei repertori, valutazione della rilevanza dell'*output* sulla base della lettura del solo repertorio, eventuale ricerca del materiale ritenuto rilevante in esteso sulle fonti primarie (riviste, volumi); *(iv)* la massimazione era riportabile alla logica descritta ora al punto (iii) di questo elenco: la massima era un buon *proxy* di rilevanza, in un mondo in cui la ricerca *full text* non era possibile; la massimazione, però, non era onnicomprensiva: vi erano dati giurisprudenziali cui non corrispondeva una massima.

Questo sistema concettuale ha già subito un primo rilevante mutamento con la digitalizzazione del dato giuridico e con l'avvento, qualche decina di anni fa, delle banche dati ad accesso *full text*. Le principali mutazioni sembrano essere state: *(a)* l'indicizzazione appare meno curata, in quanto l'indicizzazione è costosa e le banche dati investono più sui contenuti che sull'indicizzazione (basti pensare al sempre peggiore servizio di attribuzione della sentenza alla voce repertoriale); *(b)* una sempre maggiore propensione a maschere di ricerca generaliste che privilegiano il linguaggio naturale rispetto a maschere analitiche (lo si percepisce anche negli OPAC); *(c)* le massime sono un *proxy* sempre meno adoperato, sia per una copertura percentualmente in discesa rispetto al materiale pubblicato, sia perché le possibilità delle ricerche booleane mettono le massime spesso in una posizione di *second best* rispetto a molte delle necessità della ricerca; *(d)* i gestori delle banche dati private tendono a offrire in risposta anche prodotti di secondo livello, ottenuti con un piccolo lavoro autoriale e compattando materiale presente nella banche dati (per esempio, "Argomenti" di One Fiscale).

Prodigit si colloca in un momento di transizione, in cui (almeno per il mercato nazionale), i gestori delle banche dati stanno iniziando a utilizzare un poco di IA, essendo però molto lontani dall'usarla anche solo alla metà delle potenzialità già consolidate. In questo senso, Prodigit potrebbe assumere addirittura un ruolo di anticipatore nel disegno di banche dati intelligenti, sia per il livello dei consulenti tecnici di progetto che per le dotazioni economiche.

Le informazioni pubbliche su quale sia il percorso di Prodigit in relazione a questo profilo sono poche, come al solito. Sembra che sia in corso un lavoro, in fase di *training*, per l'estrazione di parti rilevanti della sentenza, mentre non ho notizie quanto a una eventuale repertoriazione del materiale.

Quanto al rapporto con la massimazione, occorre chiarire qualche ambiguità.

La *summarization* è una pratica ben consolidata nel mondo dell'intelligenza artificiale ed è un banco di prova particolarmente complesso: come estrarre da un testo di partenza un breve riassunto che esprima le informazioni interessanti senza perdere troppe informazioni contestuali? Nelle pratiche di intelligenza artificiale, si distingue la *summarization* astrattiva, che elabora una parafrasi sintetica, dalla *summarization* estrattiva, che si limita a selezionare le parti del documento (nel nostro caso: la sentenza) ritenute più rilevanti dal sistema.

Non so se oggi, nel poco di tempo di realizzazione di Prodigit, sia possibile realizzare una *summarization* astrattiva (mentre lo è certamente quella estrattiva) in un ambito così deformalizzato come quello delle sentenze di merito tributarie. In ogni caso, supponendo che una qualche forma di *summarization* sia realizzabile, si può probabilmente credere che finisca per essere un esercizio più interessante per gli informatici che utile per i giuristi. Come detto sopra, la massima era un *proxy*, un mediatore imperfetto (ma realizzabile) per un mondo dove la fonte primaria (la sentenza in esteso) non era immediatamente e facilmente raggiungibile. In un sistema in cui la sentenza in esteso diviene centrale (e in coerenza con i mutamenti già in corso descritti sopra), la massima perde di centralità nel flusso della ricerca e quindi un grande sforzo di massimazione automatizzata potrebbe portare a un risultato che poi sarebbe trascurato dagli interpreti. Con a disposizione un sistema di potente indicizzazione e classificazione per argomenti,

come quello dell'intelligenza artificiale, più che leggere 100 massime di cui si dubita dell'attendibilità (e di cui bisognerebbe consultare la fonte in esteso), meriterebbe avere invece un risultato con diversi indici che selezionano, su direttrici diverse, le 100 sentenze più rilevanti (nelle diverse sfaccettature di rilevanza indicate sopra), oppure il suggerimento su altri 10 concetti correlati a quelli intorno a cui si sta cercando.

Lo scenario cambia se l'opera di "massimazione" in corso non è indirizzata a produrre una massima come quelle tradizionali, ma un prodotto riassuntivo differente (che per esempio, costituisca una sorta di scheda riassuntiva della decisione, con una sintesi del fatto, la regola espressa, le motivazioni addotte, la decisione adottata ecc.). In questa prospettiva, l'operazione potrebbe avere un significato che dipende dall'utilizzabilità e dall'ampiezza della scheda riassuntiva che si vuole produrre. Una scheda sintetica rischia di essere di nuovo un *proxy* imperfetto e quindi di essere ritenuta scarsamente rappresentativa del contesto, mentre una scheda eccessivamente ampia non avrebbe significato nella direzione di riduzione dell'informazione per una più rapida lettura da parte dell'interprete.

Un ultimo pensiero dedicato ai repertori e ai nostri tradizionali lemmari. Una prima tentazione sarebbe di considerarli completamente superati e probabilmente questo è uno degli scenari che si concretizzeranno. Merita, però, mettere in luce come i repertori prodotti sulla base dei lemmari consentissero una ricerca "aperta" ad ampio raggio, pur contenuta in termini di percorribilità: molte volte l'interprete era alla ricerca di un'ispirazione che non trovava una soluzione nei risultati immediatamente correlati e percorreva l'intera voce generale del repertorio per trovare un qualche spunto analogico, o una qualche indicazione trasversale per arricchire l'argomentazione. Il percorso della voce repertoriale era simile al percorso di un lettore in una biblioteca a scaffale aperto: una volta individuata la sezione di interesse è sempre fruttuoso scorrere i testi alla ricerca di un'ispirazione. Forse, meriterebbe allora provare a dare nuova vita ai lemmari proprio in questa direzione: anche in questa direzione l'intelligenza artificiale può essere di aiuto, con le sue capacità di indicizzazione di attribuzione di argomenti.

7. La sentenza è un documento complesso che si presta a innumerevoli scomposizioni.

In una delle scomposizioni più seguite, possiamo dire che una sentenza tributaria presenta almeno quattro livelli, *layers* per l'informatico: la delimitazione di un accadimento (fatto), la vicenda processuale (rito), l'applicazione di una regola giuridica (diritto), l'attribuzione di una prevalenza (decisum). Ogni livello ha al suo interno livelli inferiori, a volte interconnessi: per esempio, il fatto contiene al proprio interno il sotto-livello "prova" che ha connessioni con il livello "diritto".

Questi quattro livelli, nella loro connessione, formano il complesso unico di una certa sentenza. L'interprete compie normalmente un'operazione analitica per comprendere quale dei livelli di una certa sentenza possa presentare un significato per il proprio caso.

Non so dire se il futuro ci riserverà una separazione netta dei livelli già nella fase della costruzione telematica del processo in fieri. Sto facendo riferimento a uno scenario, in certi processi già in parte attuato, secondo cui gli atti processuali vengono inseriti secondo modelli rigidi che scompongono gli atti a seconda dei livelli di interesse. Si può immaginare che, nella fase introduttiva del processo tributario, invece di caricare un documento sostanzialmente a forma libera come è l'attuale (fatti salvi i vincoli contenutistici imposti dall'art. 18 D.Lgs. n. 546/1992), il sistema imponga di distinguere in maschere rigide i fatti (e le relative prove), le regole giuridiche da applicare (magari estratte da un elenco), la domanda rivolta al giudice ecc. Questo è un modello che avrebbe indubbiamente una facilità di processamento da parte del sistema, ma che pone qualche domanda non di poco conto quanto alla sclerotizzazione dell'interpretazione indotta dalla rigidità delle forme: induce senza dubbio conformismo e distorsioni tanto nei difensori quanto nel giudice.

In ogni caso, un simile modello è ancora da venire e poi ci dobbiamo confrontare con un grande numero di sentenze, completamente destrutturate formalmente, che comporranno la banca dati di Prodigit.

Sarebbe allora interessante un esercizio di IA teso a trarre informazioni dai diversi livelli, anche per comprendere quanto i sistemi oggi siano pronti a elaborare informazioni realmente semantiche da una sentenza. Sarebbe utile comprendere, in relazione a ogni sentenza, quanto il sistema addestrato è in gra-

do di individuare i fatti poi rilevanti per la decisione giudiziale (almeno nella percezione del giudice, che poi la rappresentazione del fatto compiuta dalle parti apre un nuovo capitolo che andrà trattato in altre pubblicazioni): il timore è, per contro, che il sistema si limiti a indicare le sentenze rilevanti senza chiarire quali siano, secondo il sistema, i fatti rilevanti nella decisione concreta (con il rischio di assimilazioni casuali e inferenze negative). L'indicazione dei fatti rilevanti avrebbe, inoltre, un'utilità per l'interprete che, in certi momenti della ricerca, potrebbe essere interessato a indagare il corpus delle sentenze che presentano elementi fattuali comuni al proprio caso, indipendentemente dalle regole applicate.

Un'ultima notazione su due rilevanti *layer* della sentenza: i nomi delle parti e dei giudici.

Sui nomi delle parti vi è in primo luogo un equivoco giuridico e concettuale. In una *pruderie* denominatoria, sembra che al Ministero e a Sogei siano intenzionati a cancellare tutti i nomi dalle sentenze. Forse per memoria di qualche altra infelice vicenda che ha visto coinvolti i dati fiscali in prospettiva di tutela dei dati personali, il decisore pubblico è intenzionato a trattare il dato-sentenza con la massima cautela possibile. Questo profilo meriterà un ulteriore successivo approfondimento (anche per la presenza di un già vivace dibattito in corso in altri ambiti giuridici), ma non pare che le sentenze possano e debbano essere trattate alla stregua di altri documenti contenenti dati, anche solo per il fondamentale principio di cui all'art. 101 Cost.

In ogni caso, è evidente che la pseudonimizzazione, intesa come attribuzione di un alias alle parti consente una migliore lettura rispetto all'anonimizzazione per rimozione; l'attribuzione di un alias è ormai acquisita in informatica e dovrebbe essere la via principale da seguire.

Questo vale al più per il nome delle parti: i nomi dei giudici, invece, non dovrebbero mai essere rimossi, mentre desta una certa preoccupazione il fatto che, all'interno delle sentenze caricate da ultimo sul *def.finanze.it* i nomi dei giudici siano stati completamente rimossi, con esiti anche bizzarri come quello per cui la stessa sentenza di Cassazione se pubblicata sul Italgiureweb (versione pubblica) riporta i nomi dei giudici, mentre sul *def.finanze.it* i nomi dei giudici scompaiono. Non esiste alcuna disposizione che tuteli il diritto alla riservatezza dei giudici rispetto a una sentenza pronunciata e anzi i principi generali derivati dall'art. 101 Cost. imporrebbero esattamente il contrario. Quindi, non solo andrebbero ripristinati i nomi dei giudici nelle sentenze già caricate, ma anche andrebbe evitato di ripetere l'errore per le prossime sentenze che saranno inserite nella banca dati.

8. Veniamo infine alla consultazione e alla reportistica.

L'intelligenza artificiale è in grado di analizzare grandi quantità di dati fornendo un *output* articolabile variamente: da *report* sontuosi per ricchezza di dati e per adattabilità alla creazione di nuovi oggetti utilizzabili nell'applicazione (da *dataframe* creati sulla base delle ricerche a pacchetti di documenti indicizzati e scaricabili).

Provo a indicare alcune delle opportunità realizzative di Prodigit in questa direzione:

a) le maschere per interrogare la banca dati dovrebbero rispondere a un criterio di flessibilità e pluralità. Come indicato sopra, non esiste una sola modalità di interrogazione di una banca dati giurisprudenziale: vi sono esigenze differenti che dipendono dal soggetto, dall'attività richiesta, dal *momentum* processuale. Sarebbe bene che, allora, le interfacce di accesso al sistema fossero plurali e più flessibili possibili: qualcosa che superi, proprio grazie all'intelligenza artificiale, le rigidità delle attuali maschere di consultazioni presenti sulle banche dati tradizionali;

b) la stessa flessibilità dovrebbe accompagnare il tipo di dati restituito alla ricerca. In alcuni contesti può essere interessante un elenco (per esempio, le sentenze di un certo organo pronunciate in una certa materia, o un elenco di paragrafi salienti di sentenze già pronunciate in un certo ambito), in altri un database ordinato per le caratteristiche osservate (per esempio, una certa configurazione del fatto unita a un certo esito processuale), in altri ancora un insieme di documenti consultabili (per esempio, l'insieme di sentenze pronunciate da un certo organo in una certa materia, ordinate secondo un qualche indice logico indicato dall'utente, unito alla legislazione citata nei documenti);

c) la risposta dovrebbe essere segmentabile e incrementabile facilmente. Segmentazione: si dovrebbe consentire all'interprete, una volta individuato un documento di interesse, di canalizzare la propria

ricerca su una parte sola dei dati. Se per esempio, si è trovata una decisione che risponde ai parametri attesi, si dovrebbe poter interrogare il sistema (e scaricare l'*output*) non solo secondo la similarità rispetto all'intero documento, ma rispetto a porzioni dello stesso (fatti, disposizioni richiamate, giurisprudenza citata). Si dovrebbe poi consentire all'utente di riaprire una precedente ricerca e svilupparla secondo una nuova direzione;

d) potrebbe essere interessante, in relazione ad ogni segmento di ricerca, avere una quantificazione dei topic più ricercati in quel micro-ambito, o una indicazione immediata delle decisioni più citate riferite a un certo l*ayer* di interesse: fatto, diritto, regola ecc. (usando la *network analysis* indicata sopra al par. 4);

e) i dati di base dovrebbero essere scaricabili massivamente per consentire ricerche secondo parametri personalizzabili (anche non ricollegati all'intelligenza artificiale). Le sentenze sono pronunciate nel nome del popolo italiano, nell'ambito di una delle più rilevanti funzioni pubblicistiche dell'ordinamento, e lo stesso Prodigit è finanziato con denari della collettività tutta. Il dato sentenza non può essere considerato di "proprietà" dell'organo giurisdizionale, o del progetto, o di Sogei. Le sentenze sono pubbliche e il dato sentenza, elaborato con finanziamenti pubblici non può essere ristretto. Il dato dovrebbe essere scaricabile massivamente, in modalità facilmente utilizzabili da parte di chi voglia ri-elaborare il dato, per dimostrare l'esistenza di differenti soluzioni giuridiche. Le principali banche dati pubbliche si stanno muovendo in questa direzione, mentre lo stato attuale di *def.finanze.it* non risponde per nulla a questa logica. Occorre, allora, che con semplicissimi accorgimenti tecnici, sia consentito di scaricare tutti i provvedimenti che rispondono a filtri stabiliti dall'utente (per esempio, tutte le sentenze pronunciate nel 2022, o tutte le sentenze pronunciate da un certo organo giurisdizionale).

Anche in questo ambito, Prodigit potrà dimostrarsi una grande risorsa, di aiuto per l'interprete, se nella costruzione del sistema si rispetteranno i principi di pluralità e trasparenza.

BIBLIOGRAFIA ESSENZIALE

Dadgostari F. - Guim M. - Beling P.A. - Livermore M.A. - Rockmore D.N., *Modeling law search as prediction*, in *Artif. Intell. Law*, 2021, vol. 29, 3 ss.

Marello E., *Prodigit: alcune domande di metodo e qualche semplice risposta*, in *Riv. tel. dir. trib.,* 1° febbraio 2023

Moens M.F., *Innovative techniques for legal text retrieval*, in *Artif. Intell. Law*, 2001, vol. 9, 29 ss.

Sansone C. - Sperlí G., *Legal Information Retrieval systems: State-of-the-art and open issues*, in *Inf. Systems*, 2022, 106, 101967

Šavelka J. - Ashley K.D., *Legal information retrieval for understanding statutory terms*, in *Artif. Intell. and Law*, 2022, vol. 30, 245 ss.

van Opijnen M. - Santos C., *On the concept of relevance in legal information retrieval*, in *Artif. Intell. Law*, 2017, vol. 25, 65 ss.