Rivista Telematica di
Diritto Tributario
Pacini Giuridica

EDITORIALE

Prodigit: alcune domande di metodo e qualche semplice proposta

I. La critica ha a che fare con l'analisi, con la distinzione. Ripescando pedantemente "critico" dal DELI, p. 416, otteniamo: "esame a cui la ragione sottopone fatti e teorie per determinare in modo rigoroso certe loro caratteristiche", "denuncia di un'imperfezione, di un difetto di un errore".
Non esistono, a ben vedere, critiche distruttive. Quando si percepisce una critica come distruttiva significa solo che nella comunicazione non è stata esplicitata chiaramente la fallacia che si vuole correggere: chiarita la fallacia, la soluzione emerge naturalmente.
Prodigit è certamente un progetto interessante ed è naturale che sia oggetto di critica. Questa è la normalità nelle società occidentali aperte e libere e la critica pubblica è uno degli inneschi principali della formazione di un pensiero democratico.
Prodigit è in uno stato di primo avvio e se da una parte è ingeneroso rivolgere a questo progetto critiche rispetto a forme che non si sa se saranno mai concretizzate, d'altra parte è pur sempre interessante ragionare in logica condizionale what-if, pensando alle forme che potrebbe assumere e cercando di pre-vederne i possibili difetti (le previsioni

su un sistema predittivo sono una divertente nemesi), indicando magari qualche correttivo.
Sia che lo si consideri come sistema di predizione, sia che lo si immagini come un grande sistema di reperimento dell'informazione giuridica, a fondamento di Prodigit vi sono i dati che comporranno il data lake, l'enorme base di dati su cui si impianteranno gli algoritmi di elaborazione.
Prima viene il dato: il sistema, per quanto intelligente, difficilmente produrrà un output qualitativamente migliore del dato da cui proviene.
Credere che il sistema di AI trasformi un dato cattivo in un output buono è resuscitare l'alchimia, che è scienza divertente e meritevole di essere studiata, ma come molti giuristi razionalisti e carenti di fantasia resto dell'idea che il piombo resta piombo e l'oro, beh, resta oro.
Su questi dati vorrei appuntare un paio di piccoli spilli.
I dati, nel caso di specie, saranno soprattutto sentenze (anche se la banca dati potrebbe contenere una buona fetta di legislazione tributaria, almeno richiamata per link).
Possiamo porci almeno tre interrogativi generali su questi dati: (1) quali sentenze verranno inserite nella banca dati consultabile; (2) come saranno rappresentate; (3) come saranno rese disponibili.
Tenterò qualche provocazione sui primi due interrogativi, lasciando il terzo a un contributo più esteso di prossima pubblicazione.

***

II. Quali sentenze nella banca dati?
L'attuale def.finanze.it contiene una selezione di sentenze di merito. Nella banca dati attuale non sono contenute tutte le sentenze: è sufficiente interrogare il database con riferimento a una specifica autorità e a uno specifico anno per rendersi conto che si tratta di una selezione. I criteri di selezione sono molteplici, in parte dipendenti dai "vecchi" massimari locali, in parte dipendenti da altre fonti di caricamento del dato.
Il contenuto delle banche dati private (senza fare pubblicità, ma in Italia le due maggiormente utilizzate sembrano Wolters Kluwer e Giuffrè Francis Lefebvre) è spesso più corposo, ma anche qui non abbiamo una copertura integrale.
Per il def.finanze i criteri di massimazione sono fortemente diseguali tra sentenza e sentenza: a parte molte sentenze non massimate, abbiamo sentenze massimate secondo i metodi tradizionali dei gloriosi massimari cartacei (o del massimario della Corte di cassazione), altre dove la massima sembra un riassunto della vicenda di fatto, altre in cui la massima sembra una piccola ricostruzione del diritto applicato dal giudice. Abbiamo poi molte massime cui non corrisponde la disponibilità della sentenza in full text.
Qui si pone il nostro primo problema logico-procedurale: la rappresentatività del campione.
Se nella banca dati sarà inserito solo un campione di sentenze, quanto queste sentenze saranno in grado di rappresentare la popolazione dell'intero delle sentenze?
Parliamo sia di rappresentatività quantitativa, intesa brutalmente rispetto all'esito del giudizio (quanto le sentenze inserite rispettano i risultati generali dell'organo

emittente, o quanto le sentenze inserite rispettano i risultati nazionali sulla materia trattata), sia di rappresentatività qualitativa (supposto che esista un "buon diritto" in termini di logica, comprensibilità, fondatezza, quanto le sentenze inserite rappresentano esercizio di buon diritto e quanto invece rappresentano cattivo diritto; e, al minimo, quanto le sentenze inserite sono state confermate nei gradi successivi).
Selezionare significa condizionare.
Se si rende disponibile solo un certo tipo di dato (sentenza), possono esistere in circolazione dati migliori, più rappresentativi, più utili per lo sviluppo della scienza (giuridica), ma questi dati migliori -non disponibili- non saranno utilizzati, perché irraggiungibili.
Ecco un primo suggerimento costruttivo: la banca dati dovrebbe riportare una chiara nota metodologica in cui si indica quanto i campioni siano rappresentativi e come sia stato scelto il campione reso disponibile.
Se il database contenesse tutte le sentenze tributarie non avremmo più un problema di rappresentatività e questo sarebbe certamente il golden standard.
Oggi, però, non è possibile inserire tutte le sentenze delle Commissioni (Corti) tributarie: considerata la disponibilità in formato digitale nativo, potrebbero essere inserite poche annualità e da anni molto recenti. Anche in questo caso avremmo un problema di campionamento (temporale) che qualche questione potrebbe generare (perché avremmo, poniamo, il 2023 in intero e il 2022 in campione).
Secondo suggerimento costruttivo: chiarire quali annualità saranno inserite in intero, quali in campione e quindi rendere disponibile un'ulteriore nota metodologica che indichi la copertura temporale e quanto questa incida sulle diverse materie trattate.

***

III. Quale struttura del dato?
Il dato, per essere elaborato dalla macchina, va strutturato.
Per passare dal dato al significato, occorre un'operazione di processamento del dato.
Nella data science adottano una bellissima espressione metaforica: il dato va mappato.
La mappatura del dato significa che il dato va elaborato secondo le caratteristiche che interessano ai costruttori del sistema.
Occorre una scelta esplicita che individua le caratteristiche del dato e i pesi relativi di queste caratteristiche; questa mappatura è quella che va comunicata alla macchina, perché possa elaborare una qualsiasi informazione.
Nella strutturazione del dato sentenza possiamo fissare alcuni parametri formali: l'autorità giudicante, i nomi dei giudici, l'esito della sentenza, il valore controverso, la materia trattata, le disposizioni richiamate etc. In aggiunta o in alternativa, possiamo pensare a parametri logico-motivazionali: la ricostruzione del fatto compiuta dal giudice, il tipo di ragionamento giuridico adoperato, il tipo di catena argomentativa adoperata, la riconducibilità della motivazione a un orientamento maggioritario o minoritario etc.
Una struttura nell'analisi dei dati è sempre presente.
Quando la struttura non è comunicata significa in alternativa: che la struttura è implicita, ossia che il costruttore del sistema non è consapevole del procedimento (e

quindi agisce in maniera random, passando una struttura non ponderata alla macchina), oppure che per i motivi più vari (anche degni, per carità) il costruttore del sistema non intende condividere con il pubblico la struttura.
Ecco il terzo suggerimento costruttivo: sarebbe opportuno che in Prodigit, quando fosse chiara la strutturazione del dato, si condividesse questa struttura con la comunità applicativa, per una comprensione e una discussione. Quali caratteri della sentenza saranno presi in considerazione? In che modo avverrà la mappatura? Quali pesi saranno attribuiti? Per ottenere quali output?

***

IV. Da una parte sarebbe interessante avere dei report periodici di avanzamento di questo interessante progetto pubblico (altro suggerimento costruttivo: report dettagliati di obiettivi e stato di avanzamento), dall'altra sarebbe necessario anche prevedere un meccanismo trasparente di accesso ai dati.
Sinora, la comunità scientifica ha prestato scarsa attenzione ai dati e alle statistiche rappresentative (mea culpa).
Questi dati, che pure sono pubblici e nevralgici, sono oggi difficilmente accessibili, come sa chiunque abbia provato a domandare copia massiva delle sentenze di un organo giurisdizionale per finalità di studio e ricerca: si tratta di un percorso a ostacoli in cui difficilmente si taglia il traguardo.
A oggi, il def.finanze.it, a differenza di altri database di interesse pubblico, non è progettato per un accesso massivo ai dati per finalità di studio: è pensato per consultazioni parcellizzate dei singoli documenti.
Anche le statistiche sul contenzioso tributario meriterebbero forse una revisione scientifica attenta (revisione che non è possibile se non si ha accesso al dato-sentenza: come si può comprendere se le statistiche riferite a un certo organo giurisdizionale sono attendibili se non si ha accesso, per una verifica a campione, a tutte le sentenze emesse da quell'organo?).
Quarto suggerimento costruttivo: delineare modi ampi, facili e trasparenti di accesso massivo ai dati, per finalità di studio e ricerca, come avviene per altre banche dati pubbliche.

***

V. Le critiche mosse sinora a Prodigit possono essere state più o meno aspre, ma sottintendono tutte una domanda e una proposta.
A queste domande e a queste proposte sarebbe opportuno che il progetto Prodigit desse una risposta pubblica, per favorire un dialogo e una condivisione di metodi e risultati.
E' così che di solito nasce un dibattito, con domande e risposte.

ENRICO MARELLO