INFORMATION TECHNOLOGY LAW

DISPENSE

(Ignazio Zangara)

… … …

SEZIONE II

APPLICAZIONI DI INFORMATICA GIURIDICA

I SISTEMI INFORMATIVI

Tra le applicazioni più rilevanti di Informatica giuridica troviamo i sistemi informativi, ossia quei sistemi informatici che hanno lo scopo di immagazzinare dati ed aiutare l'utente a reperire quelli di suo specifico interesse (la parola "informatica" deriva proprio dalla fusione dei due termini "informazione automatica"). Tali sistemi sono detti "isomorfi" perché non vi è differenza tra i dati in entrata e i dati in uscita. In tutti gli ordinamenti giuridici moderni hanno acquisito una particolare rilevanza quelle applicazioni informatiche che raccolgono, organizzano, selezionano e aiutano l'utente a reperire le norme, le decisioni giurisprudenziali e la letteratura giuridica. A questa nuova scienza il filosofo del diritto Vittorio Frosini, che è considerato il padre dell'Informatica giuridica in Italia, diede nel 1975 il nome di 'Giuritecnica', non nel senso meramente pratico e applicativo di uso delle tecnologie nella pratica del diritto, bensì in quello più alto di scienza che studia i complessi mutamenti nella mentalità, e quindi nel modo di lavorare degli operatori del diritto, indotti dalla rivoluzione tecnologica nel campo dell'informazione giuridica, che in quegli anni aveva avuto la sua esplosione.

COME PROGETTARE E REALIZZARE UN SISTEMA INFORMATIVO

La progettazione di un sistema informativo automatizzato richiede anzitutto una chiara definizione degli obiettivi da perseguire, con specifico riferimento alla materia oggetto d'informazione e ai suoi destinatari (gli utenti).

Il "documento" è l'oggetto centrale delle discipline che si occupano di documentazione. Con tale termine si individua qualsiasi oggetto portatore di informazioni. Si pensi ai libri, agli articoli di riviste, ai documenti della prassi pubblica e privata, ma anche alle fotografie, ai nastri, ai dischi, ai dipinti, ai francobolli, alle monete, insomma a tutto ciò che serve a documentare un determinato settore del sapere. Per la creazione del sistema informativo possono essere utilizzati documenti di un solo tipo (ad esempio, solo testi) oppure documenti di tipo diverso (testi, immagini, suoni) e in questo caso si avranno delle raccolte multimediali.

L'archivio di documenti è un insieme di documenti, dello stesso tipo o di tipi diversi, che, per esigenze specifiche informative, costituisce una raccolta logicamente omogenea (i dati catastali di un certo Comune). La gestione di un archivio di documenti ha come finalità quella di soddisfare le esigenze informative dell'utenza che hanno portato alla costituzione dell'archivio stesso. Nel caso del catasto, i cittadini e gli amministratori pubblici; nel caso di un catalogo di una biblioteca o di una guida di un museo, i fruitori della biblioteca o del museo ed il proprietario dei libri o delle opere d'arte. Con il progresso tecnologico, i fruitori degli archivi informatizzati sono sempre più spesso gli agenti intelligenti, cioè particolari software che analizzano in tempi rapidissimi grandi masse di dati per scopi informativi ben precisi.

Le dimensioni dell'archivio possono far sì che la sua gestione con sistemi manuali si riveli complicata o financo impossibile e l'utente non possa essere in grado di consultare la raccolta nella sua interezza in un tempo accettabile[1]. D'altra parte, vi possono essere esigenze informative più raffinate, come ad esempio quella di selezionare documenti di testo scritti in lingue diverse. Queste ragioni possono far ritenere necessario approntare uno strumento elettronico che gestisca l'archivio ed insieme informi l'utente in maniera completa ed analitica sui suoi contenuti.

[1] Si pensi, ad esempio, alla Biblioteca del Congresso di Washington, la biblioteca nazionale degli Stati Uniti, tra le più ricche al mondo, la cui consistenza è di circa 160 milioni di documenti.

Con l'espressione "base di dati" o "database" si intende una raccolta di informazioni gestite tramite elaboratore elettronico, organizzate secondo un sistema di relazioni contestuali che ne consentano il recupero. La definizione individua bene gli elementi che concorrono a formare una base di dati: un insieme di informazioni (omogenee per contenuto), la gestione automatizzata e la finalità del recupero dell'informazione.

L'espressone 'base di dati' può riferirsi a diversi tipi di archivi informativi, da quelli contenenti gli orari dei voli aerei e le relative prenotazioni, a quelli utilizzati dalle banche per gestire i depositi dei clienti o quelli che tutti noi utilizziamo nei nostri *smartphone* per gestire i contatti, ecc. Generalmente, si usa distinguere le basi di dati in due categorie: alcune costituiscono un'informazione secondaria, perché contengono una mera segnalazione (*reference data bases*); altre costituiscono fonte primaria di informazione (*source data bases*). Si tratta di una distinzione importante: nel primo caso, le informazioni contenute nell'archivio forniscono gli elementi necessari per identificare e descrivere una determinata entità, alla quale rinviare l'utente (esempi di questo tipo sono gli archivi bibliografici quali cataloghi, bibliografie, indici, ecc. o i repertori legislativi o giurisprudenziali); nel secondo caso, la base di dati contiene il dato informativo ultimo (statistiche, dati numerici, ma anche i testi pieni delle leggi, delle sentenze, degli atti amministrativi *full-text*).

Si tratta di una differenza non solo di struttura dei dati contenuti, ma anche, e soprattutto, di finalità informative; ciò comporta anche un diverso trattamento dei dati ai fini del recupero dell'informazione nonché differenti scelte sul tipo di software da adottare. Tecnicamente, le espressioni "banca dati" o "databank" indicano le raccolte informatizzate a testo pieno, in cui è possibile ottenere il documento ultimo, anche se nell'uso corrente le espressioni basi di dati e banche dati sono considerate equivalenti.

Nell'ambito delle basi di dati secondarie la disponibilità maggiore è costituita dagli archivi di dati bibliografici. Questi costituiscono la versione digitale degli strumenti a stampa (cataloghi, repertori, raccolte, ecc.) di cui si servono gli studiosi o gli operatori di un determinato settore. Le informazioni contenute nelle basi di dati bibliografiche non contengono il documento vero e proprio (libro o articolo di rivista), ma solo un riferimento ad esso.

Ciascun riferimento costituisce un'unità informativa della base di dati ed è chiamato *record*. Il *record* è normalmente diviso in campi strutturati, che hanno lo scopo di descrivere il documento.

Se trattassimo testi, i campi strutturati normalmente conterrebbero la descrizione bibliografica (titolo del documento, autore, pagine, editore, anno di pubblicazione, luogo di edizione) e, in alcuni casi, anche la descrizione del contenuto del documento, attuata con metodi diversi. La descrizione semantica si effettua in due maniere, attraverso l'"indicizzazione"[2] e l'"abstract". Si noti che le basi di dati di tipo bibliografico sono frequenti proprio per le restrizioni del diritto d'autore che non consentono di distribuire il testo pieno delle opere ad un pubblico indefinito o assai vasto, quale ad esempio gli utenti virtuali di una biblioteca o di un sito web. Tipicamente, le basi di dati delle biblioteche consentono di verificare l'esistenza di uno o più *record* di specifico interesse e di individuarne la collocazione fisica; successivamente, l'utente avendo a disposizione la collocazione (generalmente, un codice che indica il luogo in cui si trova l'opera, lo scaffale e il numero progressivo, ossia il posto in cui essa è collocata) potrà recarsi nella biblioteca fisica per prenderne visione.

Con l'ampliarsi delle capacità di memorizzare informazioni nei *server* locali e in *cloud* si stanno diffondendo sempre più le banche dati a testo pieno di dati pubblici (di norme, di sentenze, di

[2] Il termine 'indicizzazione' nel linguaggio informatico indica che ciascun termine o segno contenuto in un documento è riconoscibile ai fini della ricerca. In altri termini, il sistema di *information retrieval* indicizza ogni termine ed ogni segno contenuto nei documenti che costituiscono la base informativa sulla quale opera e, di conseguenza, quei termini e segni possono costituire chiave di interrogazione. Invece, in senso documentaristico, come lo si intende in questo contesto, il termine 'indicizzazione' è l'attività di trattamento delle informazioni al fine di ottenerne la descrizione del contenuto.

statistiche, di archivi anagrafici dei comuni, di archivi catastali, ecc.), quelle cioè che contengono i documenti nella loro interezza (*full text*, piante, schede allegate, ecc.).

Data la natura di dato pubblico e la vastissima produzione, sin dagli anni Sessanta vi è stata una fortissima spinta all'informatizzazione degli archivi legislativi in Italia, così come negli altri paesi di *civil law*. Si pensi che dall'Unità d'Italia (1861) ad oggi sono state pubblicate orientativamente 200.000 norme e di esse circa la metà sono ancora in vigore.

In campo giuridico sono già diffusi da tempo gli archivi contenenti i testi di legge statali e regionali e comunitari, delle sentenze dei giudici superiori. Meno diffusa è la disponibilità dei testi pieni degli articoli di dottrina contenuti per lo più nelle riviste elettroniche, per via delle restrizioni del dritto d'autore. Uno dei sistemi documentari di carattere giuridico tra i più ricchi al mondo è ItalgiureWeb della Corte di Cassazione, ospitato presso il sito del Ministero della Giustizia, che raccoglie diversi milioni di documenti. Altri sistemi informativi di grandi dimensioni sono quelli della Camera dei Deputati e del Senato della Repubblica, quello della Gazzetta Ufficiale della Repubblica italiana, Normattiva, quello della Giustizia-amministrativa e quelli dell'Istituto di Teorie e Tecniche dell'Informazione Giuridica (ITTIG-CNR) di Firenze.

Il trattamento dell'informazione

L'inserimento dei documenti in un archivio documentario avviene tradizionalmente secondo una procedura standardizzata: ciascun documento va anzitutto individuato nella sua fisicità (descrizione formale). Ciò può avvenire in vari modi: dal più semplice, attraverso un numero progressivo d'ingresso del documento nell'archivio, a sistemi più raffinati, che individuano alcuni elementi esteriori (ad esempio, per i volumi di una biblioteca, autore, titolo, città e anno di edizione; per una legge, la data, il numero e il titolo; per una massima giurisprudenziale, il numero ufficiale attribuito dall'Ufficio del Massimario della Cassazione, il collegio giudicante, i nomi delle parti). Ciascuno di questi elementi è individuato nell'archivio informatico con un apposito campo in cui viene articolato il *record* e può costituire chiave di ricerca; vale a dire che il documento può essere ricercato attraverso ciascuno di quegli elementi formali di identificazione indicati nei campi che costituiscono il *record*.

Nei sistemi documentari informatizzati, generalmente, non ci si limita alla descrizione formale del documento, ma si va oltre, perché si aggiunge anche una indicazione relativa al contenuto, cioè all'argomento (descrizione semantica), in modo da fornire all'utente un'informazione più completa e precisa, consentendogli di utilizzare ulteriori chiavi di accesso. Si tratta di indicazioni non contenute nel documento, ma aggiunte dal tecnico (documentarista) che inserisce i documenti nell'archivio e che quindi arricchiscono il documento di un plusvalore. I dati aggiuntivi sono *metadati*, cioè sono informazioni a corredo dell'informazione principale che vanno oltre il semplice testo contenuto.

La descrizione semantica di un documento avviene, più in particolare, attraverso due diversi procedimenti: l'indicizzazione e l'*abstract*.

La prima è un'operazione mirante a rappresentare i risultati dell'analisi di un documento con gli elementi di un linguaggio controllato al fine di individuare immediatamente gli elementi significativi. Il linguaggio controllato, nella terminologia documentaria, è la terminologia convenzionale, appositamente standardizzata, che consente l'uso di chiavi di accesso univoche: è un linguaggio che può essere costruito anche a prescindere dal contenuto dei documenti, partendo cioè da schemi scientifici o concettuali già organizzati e articolati, sempre relativi ad un particolare ambito disciplinare.

Il linguaggio controllato può essere espresso in:

- linguaggio documentario: generalmente un codice numerico o alfanumerico, che esprime un determinato concetto o un argomento (codice di classificazione). Con la classificazione si possono rappresentare in modo sistematico tutti i concetti e gli oggetti di un determinato settore del sapere;

- linguaggio naturale: espresso invece mediante termini, o serie di termini, usati per definire un concetto in maniera univoca e completa (descrittori) apposti dall'autore o da un documentarista esperto della materia e generalmente estratti da una lista (soggettario o *thesaurus*).

A differenza delle parole-chiave (*keywords*), che sono parole del linguaggio naturale direttamente estratte dal documento, i descrittori sono il risultato di una scelta e di un'elaborazione del linguaggio connessa all'indicizzazione. I descrittori possono essere presentati semplicemente in una lista alfabetica, a cui si possono aggiungere alcuni sinonimi sotto i termini corrispondenti; ma possono anche essere organizzati in maniera da stabilire le relazioni fra i descrittori: si ha, in questo caso, una lista strutturata.

Uno dei più importanti esempi in area giuridica di indicizzazione in linguaggio documentario è costituito dal sistema di classificazione della banca dati Dottrina Giuridica (DOGI) curata dall'ITTIG. Tutte le discipline giuridiche sono state articolate in una classificazione alfanumerica ad albero, i cui rami possono essere percorsi dall'argomento più generale a quello più specifico e viceversa.

I 'descrittori' sono poi generalmente distinti dai 'non descrittori', parole anch'esse proprie del linguaggio naturale, ma che non possono essere usate come linguaggio controllato per l'indicizzazione dei documenti. Generalmente essi rinviano al descrittore corrispondente.

Sotto il nome di *abstract* si intende genericamente il riassunto in linguaggio libero del contenuto del documento in forma abbreviata senza interpretazione né critica, redatto dallo stesso autore o da un tecnico dell'informazione (documentarista). Il linguaggio libero è un linguaggio che contiene una terminologia non controllata, e quindi più semplice da comprendere, ma anche per questo ricca di sinonimi (es.: vendita – alienazione, affitto – locazione), omonimi (es.: la parola 'contratto', che può essere un sostantivo o un aggettivo o un verbo), polisemi (es.: 'dolo' come vizio della volontà o come elemento psicologico del reato). Si comprende perciò che sebbene il contenuto informativo di un documento corredato da *abstract* possa sembrare a prima vista più ricco di un documento che contenga soltanto alcuni descrittori verbali e uno o più codici di classificazione, in realtà spesso sia un sistema meno efficace. Infatti, essendo l'*abstract* redatto in linguaggio libero, ai fini del recupero dell'informazione, potrebbe, da un lato, restituire documenti in eccesso in risposta, fornendo risultati che niente hanno a che fare con l'esigenza informativa dell'utente e, dall'altro, non consentire il reperimento di quei documenti che, pur essendo pertinenti alla richiesta, non vengono richiamati perché non contengono, né nel testo del documento né nell'*abstract*, i termini controllati attraverso i quali si esprime un determinato concetto.

È per questa ragione che le due operazioni di descrizione semantica (indicizzazione e *abstract*) vengono spesso adoperate insieme. Così, in molti sistemi informativi l'unità documentaria comprende sia le informazioni formali del documento sia i relativi termini in linguaggio controllato sia una sintesi in linguaggio libero. Occorre aggiungere, infine, che la contrapposizione tra linguaggio libero e linguaggio controllato tende in questi ultimi anni ad essere molto meno netta che in passato, in quanto sono sempre più presenti in linea le banche dati a testo pieno sfornite di indicizzazione, ricercabili attraverso i termini del linguaggio libero. Si tratta di sistemi di recupero dell'informazione basati su modelli di rappresentazione della conoscenza fondati su analisi statistiche che assegnano pesi diversi ai termini a seconda del contesto comunicativo (sistemi basati sulla conoscenza, reti neuronali, interfacce ipertestuali, sistemi esperti che interpretano e disambiguano il linguaggio naturale).

Con lo sviluppo di sistemi di intelligenza artificiale dedicati e le pubblicazioni digitali, il trattamento delle informazioni è oggi effettuato sempre più spesso da automi che analizzano il testo e classificano le unità documentali con un'accuratezza sempre più paragonabile a quella dei documentaristi.

Il recupero dell'informazione

Il recupero (o reperimento) dell'informazione è quella parte della scienza della documentazione che si occupa delle metodologie e delle tecniche che permettono di interrogare il sistema informativo e di ritrovare quei documenti contenuti nell'archivio che sono probabilmente pertinenti (o rilevanti) alle esigenze informative dell'utente, così come sono state espresse nella *query*.

Teoricamente, ciascuno dei dati (termini e segni) contenuti nell'archivio può costituire chiave di ricerca, cioè può diventare *keyword* di interrogazione. Da questo punto di vista, la differenza principale tra i sistemi di recupero manuali, come i repertori cartacei o i cataloghi a schede, e i sistemi automatizzati è la capacità di questi ultimi di fornire un accesso multiplo ai dati da ricercare: anziché disporre di un solo punto di accesso (la voce di repertorio, il nome dell'autore), con i sistemi automatizzati l'utente può disporre di tanti accessi all'informazione quante sono le parole significative che la compongono o che sono state aggiunte dal documentarista nel *record*. Rispetto ad un sistema informativo tradizionale (per esempio il catalogo cartaceo di una biblioteca), un sistema informativo di tipo automatico prevede un'indicizzazione (in questo caso, intesa in senso informatico) totale dei documenti memorizzati, ciò significa – si ribadisce – che, una volta memorizzato il testo del documento, tutte le parole, tutti i segni, insomma tutte le informazioni in esso contenute, possono costituire chiavi di accesso al documento stesso.

L'utente può utilizzare quegli elementi che ritiene utili per effettuare la sua ricerca, scartando gli altri meno significativi. Con i moderni sistemi di *information retrieval* è possibile effettuare ricerche per sottoinsiemi successivi di documenti, restringendo così progressivamente la base documentaria oggetto di ricerca, oppure utilizzare più chiavi di ricerca contemporaneamente anche su campi diversi.

Per effettuare ricerche sulle basi di dati è necessario adottare una strategia di ricerca, ossia aver chiari i criteri fondamentali con i quali accedere alle informazioni al fine di recuperare esattamente i dati ricercati. Poiché i linguaggi di interrogazione, i nomi dei vari comandi e la rispettiva sintassi, variano da un sistema informativo all'altro, è necessario conoscere ciascun sistema in maniera approfondita per poter effettuare una ricerca potenzialmente produttiva di risultati utili. È però vero che si vanno diffondendo sempre più linguaggi intuitivi univoci, cosicché un utente esperto riesce quasi subito ad identificare correttamente i singoli comandi per ottimizzare le ricerche, anche in sistemi documentari diversi.

Fra le strategie di ricerca va ricordato anzitutto l'uso degli operatori logici (c.d. operatori booleani, dal nome del matematico inglese George Boole che, verso la metà del secolo scorso, applicò le regole dell'algebra alla logica, creando la logica algebrica, detta appunto "algebra di Boole").

Si tratta di tre operazioni logiche che, in termini documentari, individuano:

a) la compresenza di due o più elementi;

b) l'alternativa fra più elementi;

c) l'inesistenza di uno o più elementi.

In termini informatici le tre operazioni logiche sono indicate mediante i seguenti operatori: AND, per indicare la compresenza di più chiavi di ricerca (ad esempio, più autori o più parole contenute in uno stesso titolo); OR, per indicare l'alternatività (per esempio, una oppure un'altra parola contenuta in un testo di legge); NOT, per indicare che si vuole escludere dai risultati della ricerca quei documenti che presentano l'elemento preceduto dall'operatore in parola (per esempio, intelligenza NOT artificiale).

Gli operatori logici possono essere, a loro volta, combinati tra loro in vari modi, espressi in modalità algebrica mediante parentesi: ad esempio (A or B) and C, oppure A not (B and C), ecc.

Oltre a questi operatori logici ve ne sono altri che consentono aggregazioni di più termini in un certo ordine (frase esatta) che permettono di recuperare i documenti che contengono i termini indicati solo se essi compaiono nella stessa sequenza indicata (è il caso dei c.d. sintagmi, che nel

linguaggio giuridico sono molto frequenti: capacità giuridica, locazioni brevi, domicilio digitale, reati di pericolo, ecc.).

Altri strumenti di ricerca presenti generalmente nei sistemi informativi sono i metacaratteri per attivare le funzioni di troncamento e di mascheramento. Il troncamento consente di specificare solo la parte iniziale dei vocaboli cercati, troncando il termine di ricerca e sostituendo col metacarattere (*) le lettere mancanti; ciò consente di reperire tutti i documenti che contengono una qualsiasi parola che inizi con l'espressione troncata (ad es.: la chiave di ricerca 'amministr*' consente di individuare tutti i documenti in cui sono presenti i termini amministrato/a/i/e, amministrativo/a/i/e, amministrazione/i, amministrare/ndo, ecc.). È uno strumento molto utile, perché consente di ottimizzare la *query* comprendendo le varie forme lessicali aventi la stessa 'radice', evitando di formulare più volte la domanda o di usare l'operatore OR. La funzione di mascheramento consente di sostituire uno o più caratteri della chiave di ricerca con un carattere jolly (?); ciò consente di reperire tutti i documenti che contengono oltre ai caratteri indicati qualsiasi altro carattere collocato nella posizione del jolly (ad es.: 197?, fide?ussione, b?b).

In conclusione, al di là dei metodi e delle strategie utilizzate, il recupero dell'informazione è fortemente condizionato dal trattamento delle informazioni inserite nei sistemi documentari. Più il trattamento sarà ricco di indicazioni sui singoli documenti (dati e metadati inerenti ad esso ed al dominio di conoscenza relativo) migliori saranno i risultati delle ricerche.

Se ci spostiamo sulla rete il ragionamento appena accennato va esteso. La rete, infatti, contiene centinaia miliardi di documenti statici, cioè presenti costantemente *online* nonché altri migliaia di miliardi di documenti archiviati nei *databases* – nella maggior parte dei casi, accessibili tramite *user id* e *password* (i sistemi informativi telematici protetti) –; questi ultimi, quando interrogati, generano pagine dinamiche, cioè pagine uniche ed irripetibili composte sul momento (sulla base della *query* dell'utente) che riportano i dati estratti dal *database*, non collegati ad un url specifico. Si tratta del c.d. *deep web* i cui contenuti non sono indicizzabili dai comuni motori di ricerca. Analogamente, alcuni produttori di informazioni inseriscono nel codice di compilazione delle pagine web indicazioni precise al fine di non permettere ai comuni motori di ricerca l'accesso e l'indicizzazione dei contenuti riportati nelle loro pagine. Gli scopi di queste tecniche sono i più vari: di fondo possono individuarsi profili di riservatezza, orientati ora al segreto industriale ora alla tutela del diritto d'autore ora al segreto militare ora, semplicemente, allo scambio di informazioni private solo tra gruppi controllati.

Altre reti oscure ospitano il *dark web* ossia una porzione più specifica del *deep web* non tracciabile e del tutto anonima in cui sono veicolate informazioni e sono forniti servizi che poco hanno a che vedere con i concetti di trasparenza e di legalità. Per proteggere l'utente, le reti oscure utilizzano un grande numero di *server* intermedi. I pacchetti vengono cifrati con chiavi diverse ad ogni passaggio, rendendo così molto complesso ricostruirne il percorso dato che ogni nodo conosce solamente l'indirizzo dei suoi due corrispondenti. Questa tecnica è comunemente nota con il nome di *onion routing*. Non è inutile dire che accedere a queste reti è di per sé molto compromettente poiché lì, il più delle volte, si perpetrano frodi informatiche, commerci e traffici di prodotti o di servizi illegali, informazioni per le organizzazioni criminali e sovversive e truffe di ogni genere[3].

GLI INDICI DI PRESTAZIONE DEI SISTEMI INFORMATIVI

Per calcolare la prestazione dei sistemi informativi la scienza della documentazione usa quattro indici, detti 'di prestazione': richiamo, precisione, silenzio, rumore.

1) Richiamo. È la capacità del sistema di ottenere in risposta tutti i documenti pertinenti, contenuti nell'intero archivio. Per tasso di richiamo si intende appunto rapporto fra il numero dei

[3] Il sistema dell'*onion root* in rete non necessariamente reca connotati negativi. Ha consentito, ad esempio, ai cittadini di paesi controllati dal sistema politico di dialogare con il mondo esterno e di bypassare i limiti alla libertà di espressione e di informazione.

documenti recuperati, attraverso una ricerca, e il totale dei documenti memorizzati: eseguendo una ricerca in tema di affitto di fondi rustici in un archivio che contiene 80 documenti, ove il sistema consenta di reperire tutti i documenti che trattano l'affitto di fondi rustici si dirà che esso ha una capacità di richiamo del 100%.

2) Precisione. È la capacità del sistema di fornire in risposta solo documenti pertinenti. Per tasso di precisione (o di pertinenza) si intende il rapporto tra il numero di documenti presenti in archivio e pertinenti la richiesta e il totale dei documenti che sono stati recuperati attraverso una strategia di ricerca. Ove, ad esempio, nella risposta compaiano 50 documenti, ma di questi solo 25 sono pertinenti alla richiesta, si dirà che il sistema ha il 50% di precisione.

3) Silenzio. È l'opposto del richiamo, ed indica i documenti pertinenti non richiamati dal sistema. Il tasso di silenzio è dato dal rapporto fra i documenti pertinenti che vengono perduti nella risposta e i documenti pertinenti contenuti nell'archivio. Nell'ipotesi in cui vi sia un richiamo del 100% non vi sarà alcun silenzio.

4) Rumore. È l'opposto della precisione e indica i documenti non pertinenti contenuti per errore nella risposta. Il tasso di rumore è dato dal rapporto tra il numero dei documenti non pertinenti contenuti nella risposta ed il totale dei documenti recuperati. Se, come nell'esempio di sopra, solo 25 dei 50 documenti recuperati sono pertinenti, vi sarà un 50% di rumore.

In sostanza, il richiamo tende a reperire tutto ciò che è pertinente, anche a costo di provocare del rumore con documenti non pertinenti; invece la precisione tende a fornire risposte in cui non ci siano documenti non pertinenti, ma col rischio di perdere documenti pertinenti.

I *THESAURI*

Uno strumento tipico dei sistemi documentari, che si colloca a metà strada tra i sistemi di indicizzazione ed i sistemi di recupero dell'informazione, è costituito dai *thesauri*.

Un *thesaurus* è un vocabolario di un linguaggio di indicizzazione controllato, organizzato in maniera formalizzata, in modo che le relazioni a priori tra i concetti siano rese esplicite. Il ruolo fondamentale di un *thesaurus* è il controllo applicato sia a fini dell'indicizzazione (nel senso che indica all'indicizzatore i termini controllati con i quali esprimere un determinato concetto) sia a fini di recupero dell'informazione (nel senso che indica all'utente i termini per formulare la sua domanda al sistema) sia, ancora, a prescindere da un'esigenza di ricerca documentaria, a fini repertoriali, di consultazione, quale vocabolario tecnico di un settore del sapere scientifico, simile alla funzione che può svolgere un qualsiasi vocabolario del linguaggio naturale.

La caratteristica che distingue un *thesaurus* da una semplice lista di descrittori controllata (soggettario) è il fatto che tutti i termini presenti nel *thesaurus* sono organizzati in una struttura semantica, cioè le relazioni tra loro sono compiutamente esplicitate. Tali relazioni possono essere di vario genere:

• la relazione di preferenza, indicata con l'operatore *us* (dall'inglese *use* = usa) e col suo reciproco *uf* (*used for* = usato per) o con il simbolo internazionale "=", serve per risolvere situazioni di sinonimia (case – abitazioni) o quasi-sinonimia (ville – abitazioni) o antinomia (comunismo – anticomunismo) o varianti nella forma, sigle, acronimi (XII Tabulae - Lex XII Tabularum) e per rinviare da un non-descrittore a un descrittore e viceversa (per esempio: Catasto *us* registri immobiliari; registri immobiliari *uf* catasto);

• la relazione di gerarchia, indicata con l'operatore *bt* (*broader term* = termine più ampio) e col suo reciproco *nt* (*narrower term* = termine più ristretto) o con i simboli internazionali "<" e ">", collega verticalmente tra loro i termini appartenenti alla stessa famiglia semantica e serve per mettere in evidenza i rapporti di sovraordinazione o subordinazione tra genere e specie (per esempio: usufrutto *bt* diritti reali; diritti reali *nt* usufrutto). Nella scala gerarchica viene talora indicato come *top term* (*tt*) il termine al vertice di una catena semantica;

• la relazione di associazione o affinità, indicata con l'operatore *rt* (*related term* = termine associato) o con il simbolo internazionale "-", evidenzia i rapporti di correlazione, equivalenza ed

associazione d'idee tra descrittori (esempio: educazione *rt* insegnamento; insegnamento *rt* educazione). Tale relazione è molto utile nei *thesauri* multilingue.

Non tutti i *thesauri* esplicitano le relazioni sopra indicate: alcuni si limitano alle relazioni di tipo gerarchico, altri a quelle gerarchiche e di preferenza. La relazione di gerarchia è quella più usata. La relazione gerarchica viene espressa graficamente attraverso una classificazione ad albero, con la quale dal termine più generale (*top term*), che esprime una intera classe di concetti, si scende via via ai termini più specifici.

In campo giuridico i *thesauri* più noti sono: *Teseo*, il *thesaurus* del sistema informativo del Senato della Repubblica; *Eurovoc thesaurus*, dedicato alla documentazione dell'Unione europea; *Bibliotheca iuris antiqui thesaurus*, dedicato ai diritti dell'antichità.

LE ONTOLOGIE

A differenza del *thesaurus*, l'ontologia è una rappresentazione formale di un dominio di conoscenza in cui sono esplicitati i rapporti tra i termini ed i concetti, in modo da fornire una mappa concettuale (semantica) dalla quale ricavare le classi di entità rilevanti, le classi incluse in classi più ampie, le condizioni per l'appartenenza ad una classe, gli attributi connessi ai termini, ecc.

In altri termini, le ontologie sono descrizioni di dominio condivise che, indicando in modo formale il significato ed i legami tra i termini, costituiscono ottime basi di conoscenza, soprattutto, per i sistemi di intelligenza artificiale.

La disponibilità di un'ontologia formale consente di raffinare anche la ricerca documentale. Se si dispone di un'ontologia (ed essa è collegata al lessico) è possibile individuare con precisione il concetto che interessa (ad esempio, lo scioglimento di un contratto, sia esso qualificato come annullamento, risoluzione o recesso) e reperire tutti i documenti attinenti a quel concetto o ai concetti che si trovano in particolare connessione semantica con esso (ad esempio, possiamo ricercare tutti i documenti che vertono sugli eventi che determinano lo scioglimento di un contratto). Un'ontologia formale può inoltre essere utilizzata per far ragionare le macchine e permette la comunicazione tra sistemi informatici che utilizzano diverse terminologie (*machine to machine*).

La relazione di inclusione tra classi è tra le più importanti relazioni ontologiche (gli oggetti appartenenti alla classe più ristretta appartengono tutti anche alla classe più ampia). L'ontologia permette di distinguere, ad esempio, differenti accezioni di un termine allo scopo di approfondire i concetti sottostanti disambiguandone il significato. Ad esempio, il concetto giuridico di 'contratto' può essere inteso quale evento (riferito alle azioni compiute dalle parti in un certo tempo), quale contenuto (un insieme di clausole pattuite), quale documento (il contenitore del documento) e quale regolamento (le pattuizioni che regoleranno i rapporti tra le parti). Per concludere, mentre un mero *thesaurus* si limita ad indicare che il termine 'contratto' è più specifico rispetto al termine 'accordo' e più ampio rispetto al termine 'compravendita', con l'ontologia si può distinguere ed approfondire nell'ambito del termine 'contratto, i concetti di 'parti del contratto', 'proposta', 'accettazione', 'adempimento' 'inadempimento', 'atto pubblico', 'scrittura privata', 'digitale', 'cartaceo', 'registrato', 'tipico', 'atipico', 'preliminare', 'definitivo', ecc. La mappa concettuale che si genera in tal modo è in grado di esprimere ed esplicitare una porzione di realtà collegata ad un evento/elemento.

Lo scopo di identificare categorie astratte nelle quali cogliere gli aspetti fondamentali della realtà rende il concetto di ontologia in senso informatico assimilabile e continua rispetto a ciò che con questo termine intendevano gli antichi filosofi greci (Platone, Aristotele, Porfirio), al di là delle tecniche utilizzate e dei fini perseguiti[4].

[4] La nozione di ontologia in senso informatico-giuridico qui rappresentata è del prof. Giovanni Sartor.

Le ontologie, quindi, aprono scenari interessanti, tanto di ordine filosofico quanto di ordine informatico-applicativo, con riferimento ai sistemi di intelligenza artificiale, oggetto del prosieguo della presente sezione.

I SISTEMI COGNITIVI

I sistemi cognitivi sono sistemi informatici che producono nuova conoscenza. Essi costituiscono l'evoluzione dell'ambito c.d. 'logico-decisionale' della Giurimetria. Data una base di conoscenza giuridica formalizzata, mediante una serie di operazioni logiche, si ottiene una risposta in termini di decisione/soluzione su un quesito proposto. Questi sistemi sono detti eteromorfi, nel senso che l'*output* è differente dall'*input*.

I sistemi eteromorfi, in cui è predominante l'algoritmo di intelligenza artificiale (IA), studiano la possibilità di riprodurre mediante l'elaboratore le attività intellettuali dell'uomo. Le prime sperimentazioni di intelligenza artificiale appartengono agli anni '70. Alla base di quegli studi stava il sorgere di una nuova disciplina, nell'ambito delle scienze cognitive, il cui oggetto è lo studio dei sistemi intelligenti naturali ed artificiali e delle loro interazioni.

Rispetto all'informatica tradizionale le differenze sono essenzialmente tre:

a) i dati che vengono memorizzati non sono i documenti nella loro forma originale, bensì descrizioni o rappresentazioni di essi (testi, immagini, suoni non quali entità statiche, bensì quali portatori di significati). Ogni elemento utile per la decisione deve essere rappresentato;

b) i processi sui dati non sono semplici raffronti tra stringhe di caratteri, come avviene nei programmi di reperimento dei documenti, bensì elaborazioni complesse che in qualche modo riproducono i ragionamenti dell'uomo;

c) i risultati del processo non sono costituiti dal reperimento dei dati che sono stati immagazzinati precedentemente nell'elaboratore, ma piuttosto costituiscono una nuova conoscenza prodotta dall'elaborazione.

Un sistema cognitivo (o un sistema basato su tecniche di intelligenza artificiale) è quindi un sistema che, sulla base di un'elaborazione complessa sui dati forniti all'elaboratore, tenta di riprodurre il ragionamento della mente umana per risolvere un problema, riconoscere la realtà, comprendere un messaggio, proporre una risposta/soluzione ad un quesito/problema.

Sulla base di questo obiettivo, tra gli studiosi di scienze cognitive si sviluppò, durante tutti gli anni '70, un dibattito teorico tra due tesi contrapposte: quella dell'intelligenza artificiale forte, e quella dell'intelligenza artificiale debole. Secondo i sostenitori della prima tesi il computer potrebbe essere davvero dotato di una intelligenza non distinguibile in alcun modo da quella della mente umana, anzi l'intelligenza artificiale sarebbe in grado di valutare la correttezza dei processi mentali umani, producendo da sé modelli di ragionamento. La seconda tesi sostiene invece che un computer non potrà mai eguagliare la mente umana, ma al massimo potrà simulare alcuni processi mentali umani senza mai riprodurli nella loro complessità.

È evidente che la prima tesi, anche se certamente più affascinante (si pensi alle problematiche che essa introduce, quali ad esempio, se la mente è autonoma rispetto al cervello e all'individuo; se una mente artificiale, una volta realizzata, può avere un'esistenza autonoma e quindi se un programma di IA è un soggetto giuridico e se, di conseguenza, può essere responsabile o dotato di poteri, come per esempio il voto!), presenta tuttavia molti motivi di perplessità, sia di ordine teorico che di ordine pratico. Ma, specialmente, essa non ha retto alla prova del tempo: dopo decine di anni di ricerche i risultati non sono stati certo così confortanti come le aspettative. È per questo che oggi quel dibattito risulta superato e certamente la tesi dell'intelligenza artificiale debole sembra la più realistica. Per il prossimo futuro ci si aspetta una interazione crescente tra uomo e macchina volta al completamento reciproco. L'uomo presterà alle macchine l'intelligenza sensibile, le macchine presteranno all'uomo elaborazioni straordinarie.

In una logica di apprendimento nozionistico ed esperienziale, le macchine – che a differenza dell'essere umano non si riposano e non hanno bisogno di svaghi, ma lavorano senza sosta –

raggiungono risultati sempre più sorprendenti in svariati campi del sapere e per tutte le attività di *routine* hanno già sostituito l'uomo, perché eseguono molto meglio e molto più rapidamente i lavori ad esse affidati.

I sistemi cognitivi strutturalmente sono basati su due moduli fondamentali:

• le informazioni sul settore della realtà di cui si occupano;

• le istruzioni per elaborare tali informazioni ('ragionare'), ricavandone informazioni originali.

I due moduli sono chiamati rispettivamente 'base di conoscenza' e 'motore di inferenza'. Essi stanno ai sistemi cognitivi come la base di dati e il motore di ricerca stanno ai sistemi informativi. La differenza rispetto a questi è che nei sistemi cognitivi la base di conoscenza, come già accennato, non è costituita dai dati, ma dalla loro rappresentazione formalizzata in funzione dell'ambito di realtà che si vuole rappresentare e che il motore di inferenza non va a ricercare i dati, bensì li utilizza e li elabora per ottenere una conoscenza nuova, che può essere, di volta in volta, il riconoscimento di immagini, la comprensione del linguaggio naturale e della voce umana, la soluzione di problemi, la robotica, la pianificazione, l'osservazione, la predizione, ecc.

Per costruire un sistema cognitivo la prima operazione da fare consiste nell'individuazione della base di conoscenza, ossia la delimitazione del dominio di conoscenza che si intende rappresentare. Nel caso del diritto non si tratta solo delle norme, ed eventualmente delle sentenze o della dottrina, ma di tutti gli elementi, anche fattuali e di esperienza, che possono essere utili alla decisione.

La fase più delicata è, dunque, quella della rappresentazione della realtà. La rappresentazione è un processo di astrazione mediante il quale si descrive un dato di realtà attraverso l'uso di simboli, detti pure formalismi. Ciascun tipo di realtà può essere descritto mediante formalismi diversi. In genere nella teoria dell'IA si è soliti distinguere due diversi tipi di formalismi: a) il c.d. formalismo procedurale, che privilegia appunto gli aspetti procedurali della conoscenza, cioè descrive le concatenazioni di cause ed effetti tra gli eventi (ad esempio: per vendere un bene occorre averne la proprietà); b) il c.d. formalismo dichiarativo, che descrive invece gli elementi che concorrono a formare uno stato di cose (ad esempio: un contratto richiede un soggetto, una forma, una causa).

Il motore di inferenza utilizza la conoscenza memorizzata mettendo in atto processi di ragionamento differenti a seconda del formalismo di rappresentazione adottato: deduttivi per i formalismi procedurali; analogici per i formalismi dichiarativi.

Per restare nell'ambito del diritto, un esempio di formalismo procedurale è il seguente. Alla domanda: "può sposarsi Mario?", occorre individuare anzitutto la norma da applicare (in questo caso l'articolo del codice civile che disciplina la capacità matrimoniale), indi confrontare il caso concreto con la norma in oggetto (se si è maggiorenni, la risposta è sì; in caso contrario, no, a meno che non si abbia l'autorizzazione del Tribunale). È questo il tipico sillogismo giuridico: una premessa maggiore (la norma) ed una premessa minore (i fatti del caso di specie), da cui deriva la soluzione (ci si può sposare o no). Questo tipo di formalismo si presta molto bene a riprodurre il modo di ragionare dei giuristi dell'area continentale (*civil law*).

Il formalismo dichiarativo, che invece bene si adatta ai sistemi anglo-americani (*common law*), procede secondo ragionamenti di tipo analogico-induttivo. Di fronte ad un quesito simile, ma un po' più complesso ("possono sposarsi Giovanni e Sandra, dei quali uno ha compiuto 25 anni e l'altra ha compiuto 16 anni?") il giurista di formazione anglosassone va a ricercare tra le sentenze dei giudici se ve ne sono alcune su una fattispecie identica, o almeno con elementi simili (1. "Franco e Sara si sono sposati; Franco è un uomo di 25 anni; Sara è una donna di 23 anni"; 2. "Cesare ed Eleonora si sono sposati; Cesare è un uomo di 22 anni; Eleonora è una donna di 16 anni; Eleonora ha ottenuto l'autorizzazione del Tribunale"), e quindi, sulla base di un ragionamento analogico, elabora la decisione ("Giovanni e Sandra possono sposarsi, laddove Sandra ottenga l'autorizzazione del Tribunale").

In conclusione, i sistemi cognitivi si distinguono in base al tipo di conoscenza immagazzinata, al metodo di rappresentazione di tale conoscenza e al modo in cui tale conoscenza viene utilizzata. Il

processo decisionale tuttavia si svolge, in entrambi i casi, con una o più verifiche di condizioni che nel linguaggio computazionale sono rappresentate, semplificando, con la sequenza di *if-then*.

Tra le applicazioni delle tecniche di intelligenza artificiale, quella che più ha interessato i giuristi è relativa alla soluzione dei problemi giuridici (*problem solving*). Si tratta in sostanza di analizzare il processo di ragionamento attraverso cui il giurista (giudice, avvocato, pubblico amministratore, o legislatore) mette a confronto una situazione fattuale che gli viene presentata con tutti i dati che ha a disposizione (leggi, sentenze, concetti dottrinali e prassi) e giunge ad una decisione sul caso concreto (una nuova legge, una sentenza, un atto amministrativo).

Un sistema esperto giuridico è appunto un sistema cognitivo composto da una base di conoscenza giuridica, rappresentata in maniera formalizzata, e da un motore inferenziale che, sulla base di una logica (deduttiva o analogica, secondo il formalismo utilizzato), produce una soluzione al caso proposto. In altre parole, un sistema esperto non fornisce all'utente gli strumenti (norme, precedenti giurisprudenziali, riferimenti dottrinali) per costruire un ragionamento giuridico, come avviene nei sistemi informativi, ma sulla base di questi stessi strumenti enuncia una soluzione del caso (ad es., la facoltà, l'obbligo o il divieto di compiere un atto). Per fare questo potrebbe porre delle domande all'utente per accertare se nella fattispecie concreta si siano verificate certe condizioni previste in astratto nella legislazione o in casi analoghi.

L'esempio più vicino a quello della decisione giuridica è quello della diagnostica medica, per la quale i sistemi esperti già vantano applicazioni avanzate e largamente diffuse. Anche qui, infatti, siamo in presenza di processi sia deduttivi che induttivi, attraverso cui data una certa sintomatologia si giunge alla formulazione di una diagnosi.

Così come nella medicina, tuttavia, rimangono alcune perplessità di fondo circa l'applicazione sistematica di queste tecnologie al campo della decisione giuridica. È proprio vero che la decisione che un giurista deve prendere (sia essa un provvedimento o una sentenza o un atto amministrativo) sia solo conseguenza di un ragionamento, deduttivo o induttivo che sia, tale da potere escludere che essa possa essere influenzata da motivazioni estranee alla logica stessa? Ed è sempre vero che tali motivazioni, quando vi siano, proprio perché estranee ad un procedimento logico astratto, costituiscano un elemento negativo e riprovevole della decisione umana, così da far preferire la 'decisione' fornita dal computer? Sono interrogativi la cui soluzione non può che essere lasciata al buon senso dell'utente che usa questi sistemi. Quello che è certo è che per quanti elementi si potranno fornire al computer, sia con riferimento al caso concreto sia destinati a costituire la base di conoscenza, difficilmente si riuscirà a rappresentare tutti gli elementi di senso comune, di esperienza, di originalità e di intuito che caratterizzano i processi della mente umana. È per questo che si preferisce parlare di sistemi di aiuto alla decisione, a voler sottolineare che si tratta di fornire a colui che deve prendere una decisione elementi certamente utili, ma non certo definitivi, giacché sarà la mente e la volontà dell'uomo a prendere, ancora per qualche tempo, la decisione ultima.

Negli ultimi anni è stata sviluppata un'altra categoria di sistemi intelligenti per il diritto, quella dei c.d. sistemi di reperimento concettuale, che aiutano l'utente interessato ad una ricerca nella formulazione della domanda di informazione, interpretando il linguaggio naturale e disambiguando i significati degli omonimi (*smart query*), specie quando si è in presenza di basi di dati vastissime.

Sono applicazioni di intelligenza artificiale per i sistemi informativi. La richiesta che si pone al sistema è sempre quella di reperire documenti o informazioni, ma muta il modo con il quale si raggiunge il risultato che è molto più complesso di quanto accade nei programmi di *information retrieval*, per così dire, di prima generazione. Il sistema, da un lato, possiede una base di conoscenza basata sulla descrizione semantica avanzata dei documenti, vale a dire un gruppo di concetti significativi aggregati a ciascun documento (rappresentati e marcati in sgml, xml, skos/rdf, owl, tei, uri, ecc.)[5] in cui sono descritte anche le relazioni tra i concetti presenti (ontologie), dall'altro, esegue il lavoro non limitandosi a comparare la stringa di ricerca con le occorrenze

[5] Si tratta, in estrema sintesi, di linguaggi di rappresentazione e di identificazione dei dati per il web.

presenti nei documenti, ma navigando nella rete di concetti tramite inferenze logiche e guidando via via l'utente nella costruzione di domande appropriate riesce a reperire dalla massa di dati solo le informazioni realmente pertinenti, lavorando appunto sui significati. Appartengono a questa categoria i cc.dd. agenti intelligenti, cioè *robot* iniettati nel sistema con lo scopo di ottenere risultati e svolgere operazioni in cooperazione con i sistemi di "document understanding" e di "pattern recognition", ossia tecniche di riconoscimento di documenti largamente utilizzate per la creazione in modo automatico o semiautomatico di archivi ragionati, banche dati, ed, in generale, biblioteche digitali avanzate. A far da corredo ai sistemi esperti orientati alla ricerca di informazioni possono citarsi gli strumenti di "data mining", deputati all'estrazione di conoscenza su ambiti particolari da grandi masse di dati e gli strumenti di "parsing" orientati all'analisi sintattica del lessico ai fini della classificazione e descrizione dei documenti. In senso ampio, si sta avviando la strada verso il "web semantico", che ha lo scopo, tra gli altri, di arricchire le informazioni in rete di descrizioni analitiche di sé stesse (metadati), utili per la rappresentazione della conoscenza (reti relazionali tra i concetti e descrittori di realtà) e quindi per il reperimento, l'aggregazione ed il ragionamento sulle informazioni. Le prospettive di questi sistemi sono di grande interesse nell'ottica di una crescente integrazione tra sistemi di *information retrieval* tradizionali e sistemi di intelligenza artificiale orientati al "machine learning" o al "deep learning".

Le crescenti capacità di calcolo applicate all'IA stanno fornendo risultati vieppiù interessanti per i modelli di ragionamento e per le procedure di interpretazione della realtà sia sotto il profilo del *problem solving* sia sotto quello del *problem setting*, grazie all'incrocio di una enorme quantità di dati processati in unità di tempo ridottissime. Risultati inaspettati stanno progressivamente arrivando in tutte le scienze e gli studiosi stanno traendo nuovi modelli di costituzione del pensiero proprio dai collegamenti e dalle deduzioni raggiunti con i sistemi esperti[6].

OPEN DATA E BIG DATA

Con l'espressione *open data* si indica il formato 'aperto' con cui i dati digitali possono essere distribuiti nel web per essere accessibili, riusabili ed integrabili. In senso lato, il fenomeno degli *open data* costituisce una delle novità più rilevanti nella realtà di Internet, in quanto, consentendo di trattare e rielaborare l'immensa conoscenza disponibile contenuta nella rete, apre un panorama illimitato di applicazioni. Con i dati pubblicati in formato aperto è possibile costruire servizi, creare nuovi serbatoi di conoscenza attraverso la loro connessione collaborativa (*linked data*), rafforzare il percorso verso l'accesso libero alla totalità della conoscenza digitale. Tra le caratteristiche di base vi è la leggibilità anche da parte dalle macchine, la facilità di accesso ed utilizzo, la libertà e, tendenzialmente, la gratuità. Il fenomeno degli *open data* si incontra con un altro tema di grande rilievo che riguarda la riusabilità dei dati pubblici. Sotto la spinta delle norme comunitarie sta rapidamente crescendo la tendenza delle amministrazioni pubbliche, sia locali che nazionali, a pubblicare i propri dati in formato aperto nell'ottica di raggiungere un duplice obiettivo: da un lato, facilitare lo sfruttamento del potenziale scientifico, operativo ed economico dei dati pubblici; dall'altro, realizzare in concreto il paradigma dell'*Open Government*, vale a dire dell'apertura e della trasparenza delle pubbliche amministrazioni, che acquista in questo contesto un'accezione totalmente nuova[7]. Intorno al tema dell'apertura dei dati pubblici (*Openess of Public Sector Information*) si stanno sviluppando iniziative, esperienze e buone pratiche e, in parallelo, stanno emergendo dibattiti ed analisi circa gli impatti positivi, gli ostacoli e i rischi che il fenomeno dell'*Open Society* può produrre.

[6] La comprensione del testo da parte dei sistemi esperti è già in grado, ad esempio, di esplicitare quali obbligazioni derivanti da un contratto devono osservare le parti, scadenzandole nel tempo e fornendo gli ausili per semplificare i relativi adempimenti.

[7] Il catalogo generale dei dati aperti delle pubbliche amministrazioni italiane è disponibile sul portale www.dati.gov.it, gestito dall'Agenzia per l'Italia Digitale.

Il Codice dell'amministrazione digitale enuncia, all'art. 50, il principio di "disponibilità dei dati pubblici" che consiste nella possibilità, per soggetti pubblici e privati, "di accedere ai dati senza restrizioni non riconducibili a esplicite norme di legge" e, all'art. 52, precisa che i dati e i documenti pubblicati dalle pubbliche amministrazioni si intendono rilasciati in formato aperto, fatti salvi i riferimenti ai dati personali.

Tim Berners Lee, il padre del web, nel 2009 ha classificato gli *open data* in cinque livelli di completezza, contrassegnandoli dai meno strutturati con una stella ai più strutturati con cinque stelle. Per semplificare, si procede dai dati in formato proprietario, ma accessibile, ai dati strutturati in formati con licenza aperta, ai livelli più alti di *open data* rappresentati dai dati strutturati e codificati in un formato non proprietario che sono dotati di uri (*uniform resource identificator*) che li rende permanenti sulla rete e quindi utilizzabili direttamente *online*. L'ultimo livello, con cinque stelle, comprende quelli che vengono definiti *linked open data* (lod). Sono dati aperti che, dal punto di vista del formato, oltre a rispondere alle caratteristiche indicate ai livelli precedenti, sono strutturati in modo da potersi agganciare ad altri dati, formando *dataset*[8] in collegamento tra di loro. In altri termini, grazie al ricorso al modello di descrizione dei dati rdf[9], è possibile collegare dinamicamente tra loro più *dataset*, incrociando così informazioni provenienti da fonti diverse. Si pensi, ad esempio, al caso del *dataset* contenente i provvedimenti amministrativi di un determinato ente pubblico. Tale *dataset* potrebbe essere collegato al *dataset* del Dipartimento della Funzione Pubblica che raccoglie in archivio tutti i provvedimenti amministrativi presenti in ciascuna sede italiana di quell'ente e delle Camere di Commercio. In questo caso, un sistema software potrebbe, mappare i diversi esiti di richieste di autorizzazioni connesse per l'esercizio di una determinata attività imprenditoriale, al fine di orientare l'azione proprio laddove ci sia maggiore possibilità di ottenere un parere favorevole nel più breve tempo o dove ci sia maggiore richiesta/esigenza di un determinato servizio. I *linked open data*, quindi, consentono di combinare i contenuti di *dataset* diversi grazie a costrutti formali formulati secondo il modello rdf in uno dei diversi formati esistenti (xml/rdf, *notation* 3, ecc.). Ciò aumenta esponenzialmente il valore dei *dataset* reciprocamente correlati, consentendo il passaggio dal livello dei dati a quello dell'informazione e quindi a quello della conoscenza e fornendo così un quadro di contesto strutturato a partire dalla correlazione di informazioni provenienti da fonti diverse.

L'espressione *big data* identifica ampi volumi di informazioni digitali raggruppati e immagazzinati al fine di effettuare analisi automatiche. Le caratteristiche fondamentali sono: il volume, cioè che il sistema interessa vasti bacini di raccolta di dati; la velocità, vale a dire che i dati hanno una rapidità di crescita esponenziale; la varietà, cioè che le informazioni sono raccolte in qualsiasi tipo di formato – da dati strutturati e numerici in database tradizionali a non strutturati come documenti di testo, video, audio, dati provenienti da sensori, dati di rielaborazioni automatiche, transazioni finanziarie, risorse geografiche –.

Le tecnologie digitali, che pervadono, pressoché completamente, il nostro agire quotidiano hanno la capacità di numerizzare ogni aspetto della realtà che ci circonda e hanno la capacità di registrare, di trasmettere, di analizzare, di confrontare e di aggregare i dati raccolti in maniera massiva in unità di tempo ridottissime, anche allo scopo di fornire nuove tendenze, previsioni e prospettive.

Il progresso tecnologico conduce verso un mondo parallelo a quello analogico in favore di sistemi di produzione e di raccolta informatizzati (si pensi alla guida autonoma, alla realtà virtuale ed a quella aumentata, ai rapporti interpersonali mediati dai *social*, al medico o allo psicologo virtuale). L'interazione con sistemi informatici connessi, performanti e integrati che regolano e ci aiutano nelle attività quotidiane è aumentata a dismisura negli ultimi anni. Con il sistema pubblico di identità digitale (Spid - nel Codice dell'amministrazione digitale previsto all'art. 64 e regolato nel

[8] I dataset sono insiemi di dati strutturati in relazione tra loro.
[9] V. *infra*.

DPCM del 24/10/2014) i nostri rapporti (cittadini ed imprese) con l'Autorità pubblica, che offre sempre maggiori servizi digitali e telematici, è destinato a divenire tendenzialmente virtuale, come avviene già con i *provider* Internet e telefonici. I nostri dati sono oggetto di trattamento (spesso concessi in pasto) in maniera distribuita ad una moltitudine di sistemi di gestione, di archiviazione e di rielaborazione.

I dati e le informazioni costituiscono la materia prima per la generazione di prodotti e di servizi. Secondo il report *Big Data Vendor Revenue and Market Forecast* 2012-2017, il valore del mercato relativo ai *big data* si è aggirato intorno ai 18 miliardi di dollari nel 2013 e si è attestato intorno ai 47 miliardi di dollari nel 2017. Da qui si può constatare come la crescita non è proporzionale al trascorrere del tempo, bensì esponenziale.

Ogni istante della vita quotidiana riempie di nuovi dati sistemi informatizzati. Produciamo dati in continuazione, appena ci alziamo al mattino per tutto il corso della nostra vita attiva, perfino mentre dormiamo (pensiamo ai *backup* ed alle elaborazioni dei dati contenuti nei *device* personali, ai sistemi di monitoraggio delle condizioni di salute, alle registrazioni notturne di videocamere di sorveglianza).

I sistemi informatici raccolgono e lavorano in maniera massiva ogni nostro movimento in rete e nel territorio (visite nei negozi, alberghi, uffici, parchi, web, calcolo della frequenza di visite, tempi di permanenza, celle dei ponti radio – spostamenti – indirizzi ip di collegamenti – casa, ufficio, mobile, assistenti vocali <spia>: Siri Apple, Echo Amazon, Home Google, *webcam*, microfoni e sensori che si attivano e registrano, anche a nostra insaputa).

Occorre ragionare sul crescente valore che per tutti noi assume la vita privata e il sistema di rapporti con l'esterno (il tema sarà approfondito nel capitolo dedicato alla tutela della riservatezza).

Dire di saper manovrare bene un computer è cosa ben diversa dal comprendere i funzionamenti interni e gli effetti di ciò che si "fa".

Di *open data* e di *big data* si nutre l'intelligenza artificiale, la *business intelligence*, la scienza, la medicina, la metereologia, i robot iniettati nella rete per qualunque scopo, le *smartcities*, ecc.).

Dei dati disponibili, attualmente, ne utilizziamo poco più dell'1%. Ogni giorno vengono prodotti, si stima (fonte IBM), 2,5 quintilioni di dati, pari a un miliardo di triliardi (10 elevato a 30) che possono corrispondere più o meno al *download* di mezzo miliardo di film in alta definizione, divisi tra testi, report, foto, chat, video, mail (di queste ogni giorno ne vengono spedite circa 205 miliardi).

Il paradigma con cui dobbiamo confrontarci oggi è quello di una contemporaneità che pietrifica il substrato della modernità liquida baumaniana (teorizzata dal filosofo polacco Zygmunt Bauman, vissuto intorno alla metà del '900). L'immagine che rende l'idea è quella di un oceano dove la parte viva di acqua è la liquidità delle relazioni quotidiane fatte di foto, pensieri, condivisioni, rapporti in genere; una parte sempre crescente di queste interazioni (le conversazioni, i dati che ci riguardano, le azioni che compiamo, gli spostamenti, le preferenze, i pensieri che postiamo sui *social*), che forse riteniamo che si perdano nel nulla, invece precipitano e sedimentano su un fondale che le fossilizza per sempre. Questo substrato è l'ambiente digitale di archiviazione dati attraverso cui tutte le informazioni si cristallizzano e poi vengono elaborate dai sistemi di intelligenza artificiale, *data analythics* (*descriptive*, *predictive*, *prescriptive* e *automated*), *pattern recognition*, *data matching* e *mining*, *face detection*, *geotagging*, *reputation*, ecc.

Un concetto utile da tenere in mente quando si 'ragiona' con i dati reperiti in Internet è quello espresso da Umberto Eco, che si può così sintetizzare: Internet è una grande memoria che ricorda tutto. Occorre pesare i dati disponibili per produrre informazioni corrette. Questa attività è ancora una prerogativa dell'uomo.

Alcune biblioteche, archivi documentari, musei che contenevano e custodivano parte del sapere, in passato hanno preso fuoco o si sono allagati e più di qualcosa è andata perduta davvero per sempre. L'uomo è sopravvissuto reagendo con orgoglio e con spirito di rinnovamento, talvolta ripartendo anche dalle fondamenta di una disciplina. D'ora innanzi molto difficilmente si perderà

qualcosa (documenti, testi, fotografie, riprese, messaggi, rilevazioni, ecc.) e bisogna chiedersi se l'umanità saprà gestire utilmente la massa enorme di informazioni disponibili oppure se la sovrabbondanza di dati sarà solo un fardello inutile da sostenere e in alcuni casi responsabile di disinformazione[10]. Queste domande, per nulla banali, unite a quelle che riguardano chi potrà avere accesso ai dati e chi dovrà gestire l'eventuale scarto sono oggetto di discussioni accese tra gli studiosi.

IL WEB SEMANTICO

Il web sta evolvendo e si sta concretizzando la possibilità di relazionare automaticamente, sotto il profilo logico-concettuale, le informazioni in esso contenute. In altri termini, la struttura dei dati pubblicati e le tecnologie di intelligenza artificiale sono sempre più in grado di collegare in rapporti di significato le diverse parole e i vari documenti presenti in rete al di là dell'intervento dell'uomo. Il limite che ha condizionato e che ancora condiziona l'accesso e l'elaborazione delle risorse pubblicate sul web è costituito dal modo in cui le stesse sono state inserite. Il web è colmo di significanti e non di significati. In mancanza di un adeguato trattamento, le risorse rimangono disponibili, ma isolate. Il web semantico muove nella prospettiva di rendere il patrimonio informativo presente sul web comprensibile, connesso e condiviso. Le informazioni strutturate in maniera significativa, tramite elementi descrittivi e di contesto, costituiscono il punto di partenza del web semantico.

Gli obiettivi del web semantico sono:

- realizzare la catalogazione dei contenuti delle singole pagine web e le relazioni tra di essi;
- riunire dinamicamente in un unico documento logico-concettuale collezioni di pagine web semanticamente correlate anche se distribuite in più siti;
- migliorare la precisione e l'efficienza dei motori di ricerca;
- aumentare il livello di fiducia degli utenti sulla qualità dei servizi pubblici e privati offerti nel web;
- semplificare ed aumentare la sicurezza per l'automazione di transazioni di tipo commerciale;
- favorire la condivisione, lo scambio e l'interpretazione di informazioni tra operatori umani e sistemi intelligenti.

I tre pilastri sui quali poggia il web semantico sono dettati dal W3C[11].

1) L'xml (*extensible markup language*) è un metalinguaggio di marcatura del testo – con struttura nidificata come l'html – utilizzato per caratterizzare le informazioni contenute in un documento; è estensibile, implementabile cioè in ragione di esigenze specifiche di descrizione del contenuto con la possibilità di creare nuovi *tags*; consente di attribuire ai dati elementi di significato; favorisce lo scambio documentale tra applicativi informatici che cooperano; offre, ancora, la possibilità di controllare la correttezza sintattica e strutturale definita per un determinato tipo di documento (*document type definition* – dtd).

Un documento marcato in xml costituisce il primo tassello necessario per rivelare alla macchina quali elementi di significato sono in esso rappresentati.

2) L'rdf (*resource description framework*) è una sorta di vocabolario o *thesaurus* per il riconoscimento dei *tags* xml. Consente alla macchina di comprendere la sintassi delle informazioni contenute nelle pagine web per come sono descritte in xml. Dà l'opportunità di esprimere affermazioni che siano 'machine-processable', cosicché, anche se i computer non sono in grado di

[10] Si pensi alla strumentalizzazione dei dati che se decontestualizzati possono fornire informazioni errate, fuorvianti, ingannevoli, ecc.

[11] Il *World Wide Web Consortium* (W3C), fondato a Ginevra nel 1994, ha l'obiettivo di sviluppare tecnologie e strategie per il web attraverso raccomandazioni specifiche sui protocolli comuni, le linee guida, gli strumenti di programmazione e i software dedicati.

comprenderne effettivamente il significato, possono comunque elaborare le affermazioni come se le comprendessero, producendo risultati significativi per gli utenti.

3) Le ontologie, quali enciclopedie volte a coprire i livelli di conoscenza, che esplicitano le relazioni tra i termini in xml e i concetti in rdf. Esse permettono la individuazione di un concetto complesso ('entità') che viene automaticamente e progressivamente interpretato e compiuto dalla macchina attraverso passaggi logici al fine di svolgere un intero processo interpretativo in maniera automatica.

Nel momento in cui le informazioni contenute nel web saranno complete di descrizioni in xml, da interpretare tramite l'rdf con l'utilizzo delle ontologie, molte saranno le applicazioni anche pratiche del web semantico. In ambito giuridico, ad esempio, volte al reperimento intelligente delle informazioni, alla gestione dei flussi documentali, agli strumenti di democrazia diretta e indiretta, al miglioramento dell'efficienza della PA e dei rapporti tra PA, cittadini e imprese.

Attualmente, alcune applicazioni sperimentali del web semantico sono in corso con risultati apprezzabili che fanno ben sperare per il prossimo futuro.

Oltre ai sistemi informativi ed ai sistemi cognitivi, vi sono ulteriori strumenti di Informatica giuridica come, ad es., i sistemi didattici, i sistemi redazionali, i sistemi gestionali o manageriali che, tuttavia, non sono oggetto di approfondimento in questa sede.