Informatica e diritto, XL annata, Vol. XXIII, 2014, n. 1, pp. 181-192

Un corpus di documenti giuridici per il Nuovo Vocabolario dell'Italiano moderno e contemporaneo

Maria Vittoria dell'Anna, Elisabetta Marinai
Francesco Romano, Jacqueline Visconti*

Sommario: *1. Il progetto – 2. Il lessico giuridico – 2.1. Legislazione – 2.2. Giurisprudenza – 2.3. Dottrina – 3. Criticità*

1. Il progetto

Questo contributo si inquadra nell'ambito del Progetto di ricerca di rilevante interesse nazionale (d'ora in avanti "progetto") *Corpus di riferimento per un Nuovo Vocabolario dell'Italiano moderno e contemporaneo. Fonti documentarie, retrodatazioni, innovazione*[1] che ha come obiettivo finale la fondazione di una lessicografia italiana di nuovo impianto che, diversamente dalla tradizione, si deve basare su spogli di corpora bilanciati, con larga presenza di lingua non letteraria. In particolare si intende dare conto del lavoro dei ricercatori incaricati di selezionare un corpus di documenti giuridici che vada ad integrare la parte del corpus dell'italiano non letterario.

Il corpus finale di testi per il progetto sarà costituito da diversi corpora fra i quali ricordiamo quello dell'italiano delle arti (dall'Ottocento a oggi), quello del linguaggio del fumetto, quello dell'italiano scientifico, del linguaggio politico cattolico otto-novecentesco e del linguaggio politico e amministrativo, oltre che quelli del linguaggio cinematografico, del linguaggio della paraletteratura, della trattatistica socio-etica e divulgativa, della letteratura per l'infanzia e per la scuola e infine quello dell'italiano poetico, lirico e cantato.

A queste risorse si aggiungerà un'originale raccolta delle prime attestazioni delle parole dell'italiano moderno e contemporaneo, così come saranno riadattati altri archivi frutto di precedenti ricerche (alcuni di questi sono già

* M.V. Dell'Anna è ricercatrice di Linguistica italiana, Dipartimento di Studi Umanistici, Università del Salento; E. Marinai è tecnologo dell'Istituto di Teoria e Tecniche dell'Informazione Giuridica del CNR; F. Romano è ricercatore dell'Istituto di Teoria e Tecniche dell'Informazione Giuridica del CNR; J. Visconti è professore associato, DIRAAS, Università di Genova. I par. 1., 2. e 3. sono attribuibili a tutti gli Autori; il par. 2.1. a F. Romano; il 2.2. a M.V. Dell'Anna e J. Visconti; il 2.3. a E. Marinai.

[1] Il responsabile del progetto è il professor Claudio Marazzini. Partecipano al progetto le Università del Piemonte orientale, di Genova, Firenze, Catania, Milano, della Tuscia, di Napoli (L'Orientale) e, inoltre, per il Consiglio Nazionale delle Ricerche, l'ITTIG.

disponibili presso l'Accademia della Crusca: LIT - Lessico Italiano Televisivo, DIALIT - Lessico televisivo italiano in diacronia, LIR - Lessico Italiano Radiofonico, LIS - Lessico Italiano Scritto, LIC - Lessico Italiano della Cucina, corpus dei giornali milanesi del primo Ottocento). La costituzione di tale *repository* prevede anche la risoluzione di problemi tecnologici legati alla digitalizzazione di alcuni corpora che al momento sono in formato cartaceo e alla standardizzazione, per la consultazione unificata, di archivi attualmente in formati diversi.

La raccolta dei testi giuridici acquisisce particolare importanza ai fini del *Nuovo Vocabolario* per una serie di fenomeni di evoluzione in atto nel lessico giuridico. Così, ad esempio, termini di comprovata tradizione nel diritto interno, quali mediazione, nel codice civile un contratto tipico, subiscono una radicale risemantizzazione ad opera dell'influsso degli istituti anglosassoni assonanti, in questo caso la *mediation*, per cui la mediazione diventa, secondo la definizione inserita nel decreto legislativo 4 marzo 2010, n. 28: "l'attività, comunque denominata, svolta da un terzo imparziale e finalizzata ad assistere due o più soggetti [...] nella ricerca di un accordo amichevole per la composizione di una controversia [...]"[2].

2. Il lessico giuridico

Il lessico giuridico, con tutta la sua specialità e ricchezza di termini tecnici, rientra a pieno titolo nel campo di indagine del progetto per l'italiano non letterario[3]. L'unità di ricerca ITTIG-CNR, insieme a due ricercatori di altre unità, intende individuare e organizzare un corpus di testi rappresentativi del linguaggio giuridico nel periodo di riferimento – dall'Unità di Italia ad oggi – a copertura dei tre ambiti di attività in cui suole suddividersi l'uso giuridico della lingua[4]: la legislazione, cioè la lingua della legge, la giurisprudenza, cioè la lingua dei giudici, e la dottrina, cioè la lingua dell'interpretazione e dell'osservazione giuridica.

[2] Gli esempi, raccolti nel volume F. Bambi, B. Pozzo (a cura di), *L'italiano giuridico che cambia*, Firenze, Accademia della Crusca, 2012, sono innumerevoli.

[3] Per una descrizione del lessico giuridico cfr. B. Mortara Garavelli, *Le parole e la giustizia. Divagazioni grammaticali e retoriche su testi giuridici italiani*, Torino, Einaudi, 2001, pp. 10-18; M.V. Dell'Anna, *Il lessico giuridico italiano. Proposta di descrizione*, in "Lingua nostra", Vol. LXIX, 2008, pp. 98-110; Luca Serianni, *Il linguaggio giuridico*, in Id., *Italiani scritti*, Bologna, Il Mulino, 2012, pp. 121-157.

[4] B. Mortara Garavelli, *op. cit.*, p. 22.

2.1. Legislazione

Per quanto riguarda la legislazione il progetto potrà recuperare dati da un vasto corpus di documenti legislativi presenti nella banca dati LLI - Lingua Legislativa Italiana dell'ITTIG[5] che raccoglie codici, costituzione e leggi fondamentali in lingua italiana dal 1539 al 2007.

Per il periodo di riferimento del progetto si potranno acquisire 116 documenti che coprono l'arco temporale che va dal 1865 al 2007.

La tipologia dei testi è varia, ma si tratta principalmente di fonti primarie (ad esempio codici, costituzioni, leggi costituzionali). Tuttavia sono presenti nell'archivio anche numerosi campioni di altri tipi di normativa. Vi sono infatti numerosi testi unici[6] e rilevante normativa di settore (ad esempio norme sulla cambiale e sul vaglia cambiario, sull'assegno bancario, sull'assegno circolare e su alcuni titoli speciali, la legge doganale, la legge sulla protezione del diritto d'autore e di altri diritti connessi al suo esercizio, il regio decreto regolante l'ordinamento giudiziario, le norme per la disciplina dell'elettorato attivo e per la tenuta e la revisione annuale delle liste elettorali ecc.). Vi sono inoltre gli statuti regionali (a partire da quello della Regione siciliana – approvato con decreto legislativo 15 maggio 1946, n. 455) e numerosi regolamenti sul funzionamento dei più alti organi dello Stato (ad esempio il regolamento del Senato della Repubblica e il regolamento della Corte costituzionale).

Alcune leggi ordinarie ricomprese nel corpus assumono particolare rilievo anche per testimoniare i mutamenti sociali e di costume intervenuti nel tempo nel nostro paese. Ci si riferisce, ad esempio, alla legge 20 maggio 1970, n. 300, meglio nota come Statuto dei lavoratori, o alla legge 1° dicembre 1970, n. 898 in tema di disciplina dei casi di scioglimento del matrimonio, alla legge 19 maggio 1975, n. 151, che ha ampiamente riformato il diritto di famiglia, oppure alla legge 22 maggio 1978, n. 194 in materia di tutela sociale della maternità e sull'interruzione volontaria della gravidanza, fino ad arrivare alla legge 31 dicembre 1996, n. 675, che ha regolato la tutela delle persone e di altri soggetti rispetto al trattamento dei dati personali[7]. Tale corpus, che

[5] Cfr. www.ittig.cnr.it/BancheDatiGuide/lli/Index.html.

[6] Si fa riferimento al Testo unico delle leggi di pubblica sicurezza ma anche al Testo unico delle disposizioni concernenti lo Statuto degli impiegati civili dello Stato o al Testo unico delle leggi sulle imposte dirette, per ricordarne solo alcuni.

[7] L'elenco è lungo; si potrebbe infatti ricordare anche la legge 27 luglio 1978, n. 392: *Disciplina delle locazioni di immobili urbani* o la legge 23 dicembre 1978, n. 833: *Istituzione del servizio sanitario nazionale*, se non la legge quadro sul pubblico impiego 29 marzo 1983, n. 93.

comprende tra gli altri testi normativi anche la Costituzione della Repubblica Italiana, potrà essere integrato ad esempio con le leggi costituzionali successive al 2007.

Ma è anche pensabile integrare questo rilevante campione con una selezione (ad esempio per anno e per tipo di fonte) della legislazione ordinaria reperibile sul sito *Normattiva*, che, come noto, rende disponibile la legislazione italiana ufficiale dal 1944 ad oggi[8].

Inoltre si potrà arricchire il campione con la legislazione regionale. Esistono infatti archivi legislativi, spesso marcati in XML (*Extensible Markup Language*), delle Regioni italiane. È dunque pensabile che il corpus progettuale sia integrato con la normativa di alcune regioni, scelte ad esempio in ragione della loro collocazione geografica (nord, centro, sud Italia). Infine si potrà valutare l'ipotesi di mettere a disposizione altri archivi di settore.

Si pensi ai documenti (non solo normativi) presenti nel portale del progetto PAeSI - Pubblica Amministrazione e Stranieri Immigrati[9], o nella banca dati VIPD - Vita Indipendente delle Persone con Disabilità, che contiene 1.545 documenti nazionali e 96 regionali[10]. Tali risorse digitali sono solo una parte di quelle che l'ITTIG ha reso disponibili on-line e che possono fornire materiale di studio per il progetto. Ci si riferisce alla banca dati LGI - Lingua Giuridica Italiana (o Vocanet)[11], che contiene 24.338 lemmi appartenenti a tre diverse aree: legislazione, prassi, dottrina (e infatti come esporremo in seguito si propone l'utilizzo di questi dati anche per la selezione del corpus della dottrina). Restringendo la ricerca alla sola legislazione, per il periodo 1861-1970 si hanno circa 69.935 occorrenze di lemmi dei quali è possibile vedere le immagini digitali dei contesti da cui sono tratti.

Proprio la mole enorme dei documenti presenti in tale archivio ha indotto i ricercatori CNR a ideare un progetto di ricerca che permettesse di creare un indice semantico. Il programma per la redazione collaborativa delle voci dell'*Indice semantico* consente di associare ad ogni immagine, oltre all'accezione relativa al lemma selezionato, anche la fraseologia rilevante associata al lemma stesso[12].

[8] Cfr. www.normattiva.it.

[9] Cfr. www.immigrazione.regione.toscana.it/lenya/paesi/live/index.html.

[10] Cfr. www.ittig.cnr.it/disabilita.

[11] F. ROMANO, M.-T. SAGRI, *Tecnologie per la storia del diritto: gli archivi lessicali storici del Cnr*, in "Historia et ius", Vol. 1, 2012, n. 1, 6 p.

[12] Tale programma è "a titolo puramente sperimentale" a disposizione di esperti e studiosi disponibili a incrementare la compilazione delle voci. L'Istituto fornisce naturalmente i criteri

Questa fraseologia è molto ricca e varia e chiarendo molto i contesti trattati è forse il vero valore aggiunto di questa banca dati. Attraverso questo strumento si possono ad esempio apprezzare alcune sfumature tipiche del lessico legislativo (ma anche della dottrina) che variano con il trascorrere del tempo.

2.2. Giurisprudenza

L'importanza della prosa giurisprudenziale per lo studio del lessico giuridico è stata più volte sottolineata dagli studi di linguistica giuridica; insieme ai testi dottrinali, infatti, i testi della giurisprudenza:

- consentono una analisi esaustiva in sincronia del lessico giuridico nelle sue diverse componenti individuate in base a criteri di esclusività specialistica, di contatto e interferenza con la lingua comune e con altri lessici settoriali, di livello stilistico, di registro;
- esibiscono esempi e "categorie" del lessico giuridico (tecnicismi dottrinali, prassismi, forme dell'argomentazione giuridica, ecc.) tendenzialmente esclusi dai testi normativi e dalla lessicografia storica e dell'uso (che finora ha spogliato, tra i testi giuridici, quasi soltanto testi normativi);
- sono sedi di creazione concettuale e lessicale giuridica (da qui parole e accezioni nuove possono trovare accoglienza nei testi normativi);
- costituiscono possibili contesti di prime attestazioni e di retrodatazioni assolute o relative di voci e accezioni giuridiche[13].

Per il segmento relativo alla giurisprudenza il progetto allestirà un campione significativo di sentenze, mettendo a disposizione del *Nuovo Vocabolario dell'Italiano moderno e contemporaneo* un genere giuridico finora poco o per nulla frequentato dalla pratica lessicografica.

Ricordiamo che i maggiori archivi (elettronici) di testi giuridici oggi esistenti in Italia specificamente orientati alla ricerca e allo studio del lessico accolgono testi della norma, della dottrina e della prassi, ma non testi di giurisprudenza (cfr. i già ricordati archivi Vocanet-LLI e IS-LeGI dell'ITTIG).

di redazione ed il programma ha subito alcune modifiche al fine di consentire l'identificazione degli utenti redattori. A tal fine vedi P. MARIANI, *Norme di lemmatizzazione per uniformare i dati lessicali dell'archivio selettivo*, rapporto tecnico n. 9/2002, 4 p., Firenze, Ittig-Cnr, 2002, www.ittig.cnr.it/Ricerca/Testi/mariani2002.pdf.

[13] Queste e altre osservazioni sul lessico dei testi giurisprudenziali sono nel volume di M.V. DELL'ANNA, *In nome del popolo italiano. Linguaggio giuridico e lingue della sentenza in Italia*, Roma, Bonacci, 2013, pp. 141-145.

D'altro canto le numerose raccolte informatizzate di giurisprudenza – di diversa dimensione e affondo temporale – non sono accessibili al largo pubblico, sono conosciute e utilizzate tendenzialmente per soli scopi professionali dagli addetti ai lavori (magistrati, avvocati, notai, operatori del diritto e della giustizia) e non sono in ogni caso concepite per l'indagine sulla lingua (sebbene presentino tutte una pratica e snella funzione di ricerca lessicale, utilissima per il non linguista e per il linguista, che la finalizzeranno alle specificità del proprio lavoro e dei propri interessi).

Nell'amplissimo insieme di provvedimenti giurisprudenziali disponibili, il gruppo di lavoro sui testi giuridici ha deciso di orientarsi verso sentenze e pronunce emesse dalla Corte di cassazione, dalla Corte costituzionale e dal Consiglio di Stato, che più hanno inciso nel tempo sulla formazione di una fisionomia della sentenza in Italia agendo da modello linguistico per gli altri organi e gradi di giudizio.

L'importanza dello spoglio di tali testi è inoltre accresciuta da alcuni recenti sviluppi della riflessione giuridica, e dalle conseguenze di tali sviluppi sull'evoluzione del lessico giuridico. Il "dialogo delle Corti" è ormai considerato uno dei grandi temi del diritto contemporaneo[14]. Alla tradizionale declinazione nazionale della stratificazione della giurisprudenza si aggiunge ora una pluralità di fonti sovranazionali difficilmente gerarchizzabili: nel dialogo delle nostre Corti supreme con la Corte di giustizia del Lussemburgo e con la Corte europea dei diritti dell'uomo si definiscono e ri-definiscono regole e categorie, cioè, in ultima analisi, termini.

Per la Corte costituzionale, in particolare, tale ruolo di definizione – e ridefinizione – di regole e categorie deriva dalla sua funzione di cerniera tra il diritto interno e le norme contenute nella Convenzione europea dei diritti dell'uomo. Dall'approvazione della Convenzione, infatti, molte parole hanno subito una risemantizzazione; tra gli esempi più significativi il termine "pena", ricordato da Grazia Mannozzi, che nell'ottica della Corte europea dei diritti dell'uomo azzera le distinzioni elaborate dal diritto interno tra pena e misura di sicurezza, d'un canto, e pena e misura di prevenzione, dall'altro[15].

La selezione delle sentenze avverrà secondo una campionatura quantitativa, bilanciata in rapporto alla dimensione del corpus giuridico e dell'intero

[14] R. CAPONI, *Dialogo tra Corti: alcune ragioni di un successo*, in Barsotti V., Varano V. (a cura di), "Il nuovo ruolo delle Corti supreme nell'ordine politico e istituzionale", Napoli, ESI, p. 121.

[15] G. MANNOZZI, *Riflessioni sulla lingua del diritto penale*, in Bambi F., Pozzo B. (a cura di), "op. cit.", p. 117.

corpus di testi del progetto. Non si terrà conto invece di possibili criteri di contenuto, ininfluenti per le finalità della nostra ricerca. Nei testi della giurisprudenza gli esiti di indagini sul lessico sono infatti rappresentativi indipendentemente dal ramo del diritto e dalla materia trattata nelle singole cause, a cui è più tipicamente collegata soltanto la quota dei cosiddetti tecnicismi specifici.

L'arco temporale indagato dal progetto (dagli anni dell'Unità d'Italia a oggi) sarà diversamente rappresentato, anche in ragione della possibilità di reperimento dei materiali e soprattutto dei dati storici sull'ordinamento giudiziario in Italia e sull'organizzazione delle due Corti. Le pronunce della Corte costituzionale copriranno a campione il periodo che va dalla sua nascita nel 1956 ad oggi; le sentenze della Corte di cassazione copriranno il periodo 1923-2014 (dopo la riunione delle competenze penali attuata con la legge n. 5825 del 1888, nel 1923 con il regio decreto n. 601 si completa la riunione nell'unica Corte di cassazione di Roma anche delle competenze in materia civile prima spettanti alla stessa corte e alle altre quattro cassazioni regionali di Firenze, Torino, Napoli, Palermo, di cui vengono soppresse di conseguenza le rispettive sezioni); le sentenze del Consiglio di Stato copriranno il periodo 1998-2014.

La giurisprudenza della Corte costituzionale sarà ricavata dall'archivio digitale completo delle pronunce emesse dal 1956 accessibile liberamente sia al sito Internet della Corte[16] sia da *Consulta online*[17]. Per le sentenze della Corte di cassazione e del Consiglio di Stato si può pensare di contattare editori privati che da anni sono specializzati nella raccolta digitale di banche dati di giurisprudenza, affinché mettano a disposizione una selezione dei loro materiali. Molte fra esse sono in testo integrale, complete di estremi, intestazione, fatto, diritto e dispositivo. Ad esempio l'editore Giuffrè offre un'ampia selezione di sentenze – in testo integrale – pronunciate dalla Corte Suprema in materia penale dal 1995 a oggi, nonché una raccolta di sentenze per esteso pronunciate dal Consiglio di Stato dal 1998 a oggi. Dal 2006 ad oggi sono poi presenti tutte le pronunce della Suprema Corte, ad eccezione di quelle della sez. VII di inammissibilità.

Poiché *De Jure* Giuffrè non ospita giurisprudenza più remota del 1986, per il periodo precedente (1923-1986) si dovrà pensare ad altre fonti. Ad esempio

[16] Cfr. www.cortecostituzionale.it.

[17] Periodico ideato dal costituzionalista Pasquale Costanzo, che permette ricerche di tipo sia cronologico sia testuale, on-line su www.giurcost.org.

i testi potrebbero essere ricavati da riviste specializzate in versione cartacea e convertiti in formato digitale tramite i consueti programmi OCR.

2.3. Dottrina

Bice Mortara Garavelli indica come produzione dottrinale i testi meramente interpretativi e di osservazione del diritto[18], con ciò intendendo articoli di riviste, interventi a convegni, note a sentenza e commenti a legislazione, saggi in opere collettanee, monografie, enciclopedie, trattati, manuali, lezioni, tesi di laurea e di dottorato, ma anche necrologi e recensioni.

L'enorme mole di materiale impone l'individuazione di criteri di selezione per il rispetto del vincolo di bilanciamento all'interno del corpus giuridico, quindi nel più ampio insieme di testi dell'italiano non letterario, e di copertura temporale.

2.3.1. La dottrina giuridica dell'ITTIG

L'ITTIG è nato alla fine degli anni Sessanta con il nome di Istituto per la Documentazione Giuridica (IDG). Fra i principali progetti curati e realizzati dall'IDG, e poi proseguiti dall'ITTIG, ci sono la banca dati bibliografica DoGi - Dottrina Giuridica e gli studi per il VGI - Vocabolario Giuridico Italiano.

DoGi[19] è l'unica banca dati bibliografica nazionale per la diffusione della conoscenza della dottrina giuridica per il mondo istituzionale, accademico, professionale e per il cittadino. Si tratta di una banca dati di riferimenti bibliografici di articoli pubblicati su riviste giuridiche italiane.

Per ciascun articolo spogliato, il documento DoGi offre le informazioni bibliografiche (autore, titolo, rivista, fascicolo, anno e pagine) arricchite da:

1. riassunto e/o sommario dell'articolo;
2. una o più voci classificatorie tratte da uno schema di classificazione delle materie giuridiche;
3. riferimenti delle fonti normative e giurisprudenziali principali citate nell'articolo;
4. abstract in inglese, se presente.

Ed ancora, un insieme di altri metadati per descrivere tratti significativi dell'articolo, fra cui l'indicazione del tipo di bibliografia, la tipologia (contri-

[18] B. Mortara Garavelli, *op. cit.*, pp. 26-29.

[19] Cfr. www.ittig.cnr.it/dogi.

buto indipendente, nota a sentenza o a legislazione, comunicazioni in convegni, recensioni, necrologi, rassegne).

DoGi conta attualmente circa 382.000 documenti derivati dallo spoglio di oltre 400 riviste italiane nei suoi 40 anni di esistenza. Oggi le riviste in spoglio sono circa 240.

DoGi non è una banca dati di testi pieni, ma sono circa 43.000 i documenti con riassunto a cura dell'autore o della redazione della rivista (oltre 33.000 spogliati a partire dal 2000, di cui 12.783 note a sentenza, 1.369 commenti a legislazione).

L'insieme degli abstract della banca dati DoGi è un esempio interessante di lingua giuridica dottrinale contemporanea: non può essere considerato un campione pienamente rappresentativo di dottrina giuridica per il progetto in quanto ha una copertura temporale limitata (dal 1970 ad oggi), il tipo di pubblicazione è unico (pubblicazione seriale) e non è informazione di livello primario (testo pieno).

Il già citato progetto che prevedeva la realizzazione del Vocabolario Giuridico Italiano è stato interrotto verso la fine degli anni Settanta[20].

Gli archivi di spogli lessicali integrali di legislazione e di spogli lessicali selettivi xerografici (legislazione, dottrina e prassi) raccolti fra il 1965 e il 1977 hanno dato vita a due banche dati (con interrogazione unificata on-line dal sito ITTIG[21]).

L'archivio che qui interessa è quello degli *spogli selettivi xerografici* (il già ricordato LGI) perché riguarda anche la dottrina. LGI copre il periodo che va dal X al XX secolo. Come scrive Fiorelli, "il lavoro fu avviato sulla base di uno spoglio selettivo, compiuto da intelligenze umane sopra un corpo di testi molto esteso (alla fine avrebbe sfiorato il milione di pagine) e rappresentativo dell'uso giuridico di secoli e territori diversi, di diversi rami del diritto e livelli di stile"[22]. La selezione di testi ha condotto ad un insieme di circa 2.000 testi di legislazione, dottrina e prassi. Il termine "selettivi" accanto a "spogli" sta a significare che il materiale era, ed è tutt'oggi, costituito non già da testi pieni ma da contesti (parte del testo) rilevanti per un termine giuridico. I contesti, si parla degli anni Sessanta e Settanta, venivano copiati o riprodotti su schede cartacee (da qui il termine "xerografici" che accompagna "spogli

[20] P. FIORELLI, *Premessa* a P. MARIANI (a cura di), *Indice della Lingua Legislativa Italiana*, Vol. I, 1993, pp. V-XII.

[21] Gli spogli selettivi xerografici sono messi on-line in formato digitalizzato. Questo ha permesso di renderli disponibili e di preservare il loro contenuto.

[22] P. FIORELLI, *Premessa*, cit.

selettivi"), su cui veniva riportato il lemma del termine giuridico rilevante e le informazioni bibliografiche del testo da cui era stato preso il contesto.

Si ritiene che il corpus di testi di dottrina (227 testi dal 1865 al 1971) che sono stati individuati per il VGI possa essere considerato un'ottima base di partenza per la realizzazione della parte di dottrina del sotto-corpus giuridico per il progetto, vista la competenza di chi tra il novembre del 1964 e il gennaio del 1965 partecipò alle riunioni della commissione istituita appositamente dal comitato per le scienze politiche e sociali del CNR per gli studi preliminari per il vocabolario giuridico italiano[23]. Il corpus abbraccia manuali, monografie, commentari di giurisprudenza e legislazione, che coprono tutti i sottoregistri giuridici e possono ritenersi rappresentativi sotto il profilo insieme storico, linguistico e giuridico.

La banca dati Vocanet-LGI può essere utilizzata non solo per l'ampio patrimonio di bibliografia di dottrina giuridica: si sta infatti analizzando la fattibilità di creare un collegamento tra un termine del corpus finale di riferimento e il medesimo termine presente in Vocanet.

Analoghe valutazioni di fattibilità si stanno conducendo per l'integrazione nel corpus dei contesti "tagliati" delle schede selettive della banca Vocanet-LGI[24], procedendo ad una loro immissione manuale oppure attraverso il collegamento all'immagine. In questo secondo caso si avrebbe una situazione simile a quanto si vede nell'interrogazione congiunta del Vocanet con LLI per la parte legislativa: interrogando le banche dati congiunte la risposta è data dai contesti di LLI insieme all'informazione proveniente da Vocanet, con la possibilità di scorrere le immagini dei vari contesti.

2.3.2. Altra dottrina giuridica

Alla dottrina che si trova "naturalmente" all'ITTIG, allo scopo di avere un campione esaustivo per copertura temporale e per tipo di pubblicazione, si può pensare di aggiungere fascicoli di riviste di dottrina giuridica fondate prima del 1970[25]: solo per citarne alcune, "Giurisprudenza italiana" (1849),

[23] P. FIORELLI, *L'Accademia della Crusca per il Vocabolario Giuridico italiano*, in "Atti della giornata di studio sul 'Vocabolario Giuridico Italiano'", Firenze, IDG, 1981, pp. 15-22.

[24] Questa idea nasce dalla lettura di P. MARIANI, *Introduzione* a P. MARIANI (a cura di), *op. cit.*, pp. XIII-XVIII, che cita un esempio paradossale relativo ad un testo famoso e fondamentale, anche se non rientra nel periodo di interesse per il progetto, come *Dei delitti e delle pene* (1764) di Beccaria dal quale furono "estratte" solo 45 schede. Ovviamente è un'idea che richiede una riflessione molto approfondita e ragionata.

[25] Nel corpus delle riviste DoGi le riviste fondate prima del 1970 sono oltre un centinaio.

"Il Foro Italiano" (1876), "Il diritto marittimo" (1899). Non solo: si possono aggiungere manuali, commentari, monografie e opere collettanee pubblicate dopo il 1970, provando a seguire i criteri di selezione che furono adottati per i testi di dottrina del VGI.

3. Criticità

Come detto il progetto dovrà affrontare preliminarmente il problema della digitalizzazione di alcuni documenti che non sono disponibili in formato elettronico. Questo problema, per il dominio giuridico, riguarda soprattutto i documenti della dottrina e della giurisprudenza.

Sul versante puramente informatico, per quanto riguarda lo standard documentario, è stata identificata una particolare declinazione del linguaggio XML e cioè il TEI - *Text Encoding Initiative*, che per completezza dei marcatori disponibili sembra assicurare una codifica in grado di prevedere ogni elemento del testo.

Al momento sembra che gli elementi da considerarsi ai fini della marcatura del corpus di riferimento siano:

- le frasi in lingua straniera,
- le tavole e le figure,
- la paginazione (elemento indicato ma non obbligatorio),
- le partizioni del libro per capitoli, parti ecc.,
- le note a piè di pagina o a fondo testo.

A questa prima indicazione di marcatura generica si potranno aggiungere altri elementi anche più specifici e peculiari di determinati lessici (si pensi ai fumetti).

Per le sentenze la conversione di ogni testo prevederà:

- l'inserimento di ulteriori metadati (per esempio: genere, organo di giudizio [equivalente al metadato "autore"], numero e data [equivalenti al metadato "titolo"]);
- il caricamento del file delle singole sentenze;
- la marcatura e la lemmatizzazione del corpo del testo, pronto dunque per essere indicizzato.

Analoghe previsioni riguardano la legislazione e la dottrina, per le quali i metadati saranno costituiti dagli estremi di ciascun provvedimento (legislazione) o dai riferimenti bibliografici (dottrina).

I dati contenuti nell'archivio LLI dovranno essere convertiti in XML, ma ciò non sembra comportare rilevanti difficoltà. Analoghe considerazioni

possono farsi per i dati digitali che saranno importati dalla banca dati DoGi[26]. Per tutto ciò che è in formato cartaceo alla fase di marcatura si dovrà far precedere la digitalizzazione con l'ausilio di software OCR.

Ancora prima di questa fase, ciò che verrà studiato dal gruppo di ricerca sarà il bilanciamento del materiale all'interno del corpus giuridico e del complessivo corpus obiettivo del progetto.

[26] La struttura dei documenti della banca dati DoGi è estremamente granulare e il *repository* dei documenti è un database MySQL. Queste due caratteristiche rendono i documenti DoGi, e quindi gli abstract, facilmente esportabili in qualsiasi formato, compreso il formato XML-TEI scelto dal progetto.