I requisiti per l'addestramento degli strumenti di Intelligenza Artificiale e il deposito legale delle risorse digitali: una riflessione sul contesto normativo italiano e il ruolo delle biblioteche[1].

«Dig*Italia*» 1-2024
DOI: 10.36181/digitalia-00092

Chiara Storti
Biblioteca nazionale centrale di Firenze (BNCF)

Nel 2022, la Biblioteca nazionale centrale di Firenze ha avviato il progetto sperimentale per l'addestramento in lingua italiana del tool per l'indicizzazione semantica automatica Annif, sviluppato e mantenuto dalla Biblioteca nazionale di Finlandia. La sperimentazione ha fatto emergere una più generale mancanza, nelle biblioteche italiane, dei prerequisiti per l'utilizzo non solo dei più tradizionali strumenti di Machine Learning, ma anche delle nuove Intelligenze Artificiali. Le cause sono da ricondursi, da una parte, alla lacunosa normativa sul deposito legale delle risorse digitali, dall'altra alla frammentazione e moltiplicazione di vocabolari e thesauri, rispondenti ad ontologie differenti e strutturati in formati non sempre adatti all'interoperabilità nel web. Scopo del contributo è quello di avviare una riflessione su questi temi, tentando di delineare il ruolo che potrebbero ricoprire le biblioteche nello sviluppo e utilizzo, etico e consapevole, dei sistemi di Intelligenza Artificiale. Con un approfondimento sul progetto, attualmente in corso, denominato V.I.CO. - Vocabolario delle Identità Controllate.

Premessa

È della fine di dicembre 2023 la notizia che il *New York Times* ha fatto causa alla Microsoft e a OpenAI, la società cui fa capo l'ormai celeberrima ChatGPT,

[1] Il presente contributo riprende in forma approfondita i contenuti dell'intervento dell'autrice dal titolo *Annif per la lingua italiana: i requisiti per l'addestramento degli strumenti di indicizzazione automatica e il progetto V.I.CO*, tenuto in occasione del 62° Congresso AIB, Firenze, 16-17 novembre 2023: <https://www.aib.it/eventi/congr62/>. Con l'occasione si rinnovano i ringraziamenti a Giovanni Bergamin, Federico Giubbolini e Lorenzo Gobbo, per il fondamentale apporto fornito allo sviluppo dei progetti descritti in questa sede, e ai colleghi dei Settori Ricerche e strumenti di indicizzazione semantica e Servizi informatici della Biblioteca nazionale centrale di Firenze per la generosità nel condividere la documentazione riguardante le prime sperimentazioni sull'indicizzazione semantica automatica condotte in biblioteca, e per gli scambi di idee sul tema sempre arricchenti.

per aver utilizzato gli articoli del quotidiano, coperti da copyright, per addestrare i propri sistemi di Intelligenza Artificiale generativa[2].

Il tema della titolarità all'utilizzo di dati prodotti da terzi per il training dei sistemi di Intelligenza Artificiale, generativi e non generativi, non è nuovo, ma è entrato anche nel dibattito comune[3] da quando, a marzo 2023, OpenAI ha rilasciato l'ultima versione del modello GPT - Generative Pretrained Transformer[4], consentendo a chiunque di testarne le potenzialità, tramite l'interfaccia di ChatGPT4[5].

Questo contributo si pone come una prima e rapida riflessione sulle criticità tecniche e legali, nel contesto italiano, dell'addestramento e dell'uso da parte delle biblioteche degli strumenti di Intelligenza Artificiale, con particolare riferimento a quelli di Natural Language Processing (NLP) e di Machine Learning (ML). Il contributo propone inoltre un approfondimento sui progetti in corso alla Biblioteca nazionale centrale di Firenze (d'ora in poi BNCF) che hanno dato luogo a tali riflessioni.

La prima sperimentazione di indicizzazione semantica automatica dei testi in BNCF

Quasi contestualmente all'avvio, tra il 2010 e il 2011, di Magazzini digitali[6], il servizio nazionale di conservazione e accesso a lungo termine alle risorse digitali, la BNCF intraprese una sperimentazione per l'indicizzazione semantica automatica delle tesi di dottorato in formato digitale[7]. Il principale strumento utilizzato fu l'indicizzatore MAUI[8], implementazione basata sull'algoritmo *tf-idf* (*term frequency–inverse document frequency*) che misura l'importanza di un termine all'interno di un testo o di una collezione di testi. I risultati della sperimentazione non furono considerati accettabili rispetto allo standard stabilito, ma iniziarono a

[2] A titolo esemplificativo, tra le tante testate che hanno riportato la notizia, si veda: *Il* New York Times *ha fatto causa a OpenAI e Microsoft per aver usato materiale protetto da copyright*, «Il Post», 27 dicembre 2023, <https://www.ilpost.it/2023/12/27/il-new-york-times-ha-accusato-openai-e-microsoft-di-aver-usato-illecitamente-materiale-protetto-da-copyright/>.

[3] Sempre a titolo esemplificativo, si veda *Il dataset è politico,* «Guerre di Rete - una newsletter di notizie cyber», a cura di C. Frediani, n. 169 (1° ottobre 2023), <https://guerredirete.substack.com/p/guerre-di-rete-il-dataset-e-politico>.

[4] https://openai.com/gpt-4.

[5] Al momento della redazione di questo contributo (gennaio 2024), l'utilizzo di ChatGPT4 è possibile solo a seguito della sottoscrizione di un abbonamento: <https://chat.openai.com/>.

[6] https://www.bncf.firenze.sbn.it/biblioteca/magazzini-digitali/.

[7] Maria Grazia Pepe – Elisabetta Viti, *Prime sperimentazioni di indicizzazione [semi]automatica alla BNCF*, presentazione al 6° incontro ISKO Italia, Firenze, 20 maggio 2013: <http://www.iskoi.org/doc/firenze13/pepe&viti.pdf> e Anna Lucarelli – Elisabetta Viti, *Thesaurus del Nuovo soggettario fra linked data, prove di indicizzazione automatica e altri sviluppi*, paper della presentazione tenuta in occasione dell'Italian Research Conference on Digital Libraries (IRCDL), Firenze, 4 febbraio 2016: <https://www.micc.unifi.it/ircdl/wp-content/uploads/2016/01/ircdl2016_paper_12.pdf>.

[8] MAUI fu sviluppato grazie a un *research grant* di Google: <https://code.google.com/archive/p/maui-indexer/>.

far emergere, già in quella sede, la mancanza nel contesto italiano dei prerequisiti per l'addestramento di questo tipo di sistemi. Mancanza di requisiti che, come si vedrà di seguito, persiste ancora oggi.

Strumenti di Intelligenza Artificiale per la valorizzazione e l'accesso alle collezioni digitali e agli archivi del web

Nel 2018, sempre nel perseguimento delle finalità di Magazzini digitali, la BNCF ha inaugurato anche un servizio di *web archiving*[9], per la conservazione di siti e altre risorse web di interesse culturale nazionale.

A causa della mancanza di uno specifico mandato normativo all'interno della legge sul deposito legale[10], le attività di raccolta periodica, conservazione e accesso sono necessariamente limitate ai soli contenuti per i quali la Biblioteca abbia ricevuto esplicito consenso da parte degli autori o dei responsabili a diverso titolo. La collezione di siti web archiviati dalla BNCF[11] è quindi da considerarsi, in termini meramente quantitativi, assai esigua[12]. Questa pur piccola collezione contiene, tuttavia, oltre 28 milioni di documenti[13]. Appare evidente come in un ecosistema in cui la quantità di risorse informative che le biblioteche si trovano a gestire sia di tale entità, e sempre in continua crescita, diventi impellente la necessità di individuare strumenti automatici che supportino le tradizionali attività di organizzazione, valorizzazione e accesso alle raccolte.

Da questo punto di vista, i progetti più noti e avviati più precocemente all'interno di biblioteche e archivi riguardano l'utilizzo di sistemi di Natural Language Processing e Machine Learning per il riconoscimento, l'analisi e la metadatazione dei testi[14]. Più recenti, invece, i tentativi di mettere a disposizione, in maniera maggiormente efficiente e usabile, le grandi quantità di dati generati dalla creazione di collezioni digitali e degli archivi del web, sia per gli utenti-macchina che

[9] https://www.bncf.firenze.sbn.it/biblioteca/web-archiving/.

[10] Per approfondimenti sul tema, Chiara Storti, *'Resource Not found': Cultural Institutions, Interinstitutional Cooperation and Collaborative Projects for Web Heritage Preservation*, «JLIS.it», 14 (2023), n. 2, p. 39-52, <https://www.jlis.it/index.php/jlis/article/view/533>, DOI: 10.36253/jlis.it-533.

[11] https://archive-it.org/home/BNCF.

[12] A gennaio 2024 i dati raccolti ammontano a circa 3 TB. Per avere un'idea delle dimensioni che possono raggiungere gli archivi del web, è sufficiente citare quello della British Library che, al termine della campagna 2023 di raccolta dello spazio web nazionale, contiene quasi 1.3 Petabyte di dati: <https://blogs.bl.uk/webarchive/2023/10/uk-web-archive-technical-update-autumn-2023.html>.

[13] In questo contesto, "documento" deve intendersi nella definizione data da Archive-it: «a document is any file on the web that has a distinct URL. Images, PDFs, videos, articles, etc., are all considered separate documents».

[14] A questo proposito si veda: Giovanni Colavizza – Tobias Blanke – Charles Jeurgens – Julia Noordegraaf, *Archives and AI: an overview of current debates and future perspectives,* «Journal on Computing and Cultural Heritage», 15 (2021), n. 1, p. 1-15, DOI: <https://doi.org/10.1145/3479010>.

per gli utenti-umani. In questa direzione si muovono, da una parte, la ricerca[15] e le sperimentazioni[16] nell'ambito del Computational Access, e dall'altra lo sviluppo di interfacce per il recupero delle informazioni che sfruttino le potenzialità degli strumenti di Intelligenza Artificiale generativa[17].

I risultati del progetto di addestramento del tool Annif per la lingua italiana

Per rispondere all'esigenza di automatizzare, almeno in parte, l'attività di metadatazione delle risorse acquisite nell'ambito del servizio di web archiving, nel 2022 la BNCF ha intrapreso un percorso di valutazione sull'utilizzo per la lingua italiana del toolkit per l'indicizzazione semantica automatica Annif[18], sviluppato e mantenuto dalla Biblioteca nazionale di Finlandia. La riapertura delle attività di ricerca della biblioteca su queste tematiche si deve, infatti, anche alla spinta fornita dall'intervento di Osma Suominen[19] al convegno "Bibliographic Control in the Digital Ecosystem" del febbraio 2021[20].

I risultati della sperimentazione, condotta da Lorenzo Gobbo[21], hanno tracciato una *roadmap* abbastanza precisa dei passi da compiere per l'addestramento di Annif, facendo al contempo emergere, in maniera molto chiara, la mancanza nel contesto nazionale dei prerequisiti di carattere tecnologico, organizzativo e legale per l'adozione da parte delle biblioteche italiane di strumenti di Intelligenza Artificiale, con particolare riferimento a quelli di Machine Learning. Con un azzardato tentativo di generalizzazione, si potrebbe infatti affermare che ciò di cui necessitano sempre questi sistemi sono ampie collezioni di dati e vocabolari strutturati.

[15] Digital Preservation Coalition, *Computational Access: a beginner's guide for digital preservation practitioners*, 2022: <https://www.dpconline.org/digipres/implement-digipres/computational-access-guide>.

[16] Si veda, ad esempio, l'iniziativa *Computational access to digital collections* della International GLAM Labs Community: <https://glamlabs.io/computational-access-to-digital-collections/>.

[17] Si vedano i progetti in corso presso la Library of Congress: <https://blogs.loc.gov/thesignal/2023/08/improvements-ahead-for-the-web-archives/> e presso le Standford Libraries: <https://netpreserveblog.wordpress.com/2023/06/28/navigating-through-archived-websites-from-text-matching-to-generative-ai-enhanced-qa/>.

[18] <https://annif.org/>, inizialmente disponibile per il finlandese e l'inglese, è oggi utilizzato anche dalle Biblioteche nazionali di Germania e Svezia per l'indicizzazione dei testi in tedesco e svedese.

[19] Osma Suominen – Juho Inkinen – Mona Lehtinen, *Annif and Finto AI: developing and implementing automated subject indexing,* «JLIS.it», 13 (2022), n. 1, p. 265-282, <https://www.jlis.it/index.php/jlis/article/view/437>, DOI: 10.4403/jlis.it-12740.

[20] https://www.bc2021.unifi.it/.

[21] Lorenzo Gobbo allora tirocinante presso la BNCF, ora bibliotecario addetto ai servizi digitali presso USI - Università della Svizzera italiana (lorenzo.gobbo@usi.ch). I risultati sono stati presentati ufficialmente al convegno "Look beyond. Indicizzazione per soggetto delle risorse non librarie", Roma, 6 febbraio 2023: <https://www.aib.it/struttura/commissioni-e-gruppi/gruppo-di-studio-catalogazione-ed-indicizzazione/2022/101441-look-beyond-ita/>. I dati utilizzati per il training di Annif e il risultato della sperimentazione sono disponibili su GitHub: <https://github.com/Bncf/Annif-Bncf>.

Annif supporta due diversi formati per il vocabolario di riferimento: una versione SKOS[22], ontologia basata su RDF per la rappresentazione delle relazioni semantiche tra entità all'interno di thesauri, vocabolari e sistemi di classificazione, oppure un più semplice formato testuale TSV in cui i dati sono rappresentati come coppie di URI e etichette alternative e letterali.
Di seguito un esempio dell'entità "pubblicazioni scientifiche"[23] con le sue relazioni, come descritte all'interno del Thesaurus del *Nuovo soggettario*, in formato SKOS:

```
<?xml version="1.0" encoding="UTF-8"?>
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
xmlns:skos="http://www.w3.org/2004/02/skos/core#"
xmlns:skos-xl="http://www.w3.org/2008/05/skos-xl#"
xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
xmlns:dc="http://purl.org/dc/elements/1.1/"
xmlns:dcterms="http://purl.org/dc/terms/"
xmlns:iso-thes="http://purl.org/iso25964/skos-thes#"
xmlns:void="http://rdfs.org/ns/void#"
xmlns:nsogi="http://prefix.cc/nsogi">

<rdf:Description rdf:about="http://purl.org/bncf/tid/10042">
<rdf:type rdf:resource="http://www.w3.org/2004/02/skos/core#Concept"/>
<skos:prefLabel xml:lang="it">Pubblicazioni scientifiche</skos:prefLabel>
<dc:date>2005-06-17</dc:date>
<skos:inScheme rdf:resource="http://purl.org/bncf/tid/ThesCF15"/>
<skos:inScheme rdf:resource="http://purl.org/bncf/tid/Thes"/>
<skos:broader rdf:resource="http://purl.org/bncf/tid/584"/>
<skos:narrower rdf:resource="http://purl.org/bncf/tid/33325"/>
<skos:narrower rdf:resource="http://purl.org/bncf/tid/4473"/>
<skos:related rdf:resource="http://purl.org/bncf/tid/57934"/>
<skos:related rdf:resource="http://purl.org/bncf/tid/9070"/>
<skos:related rdf:resource="http://purl.org/bncf/tid/17374"/>
<skos:related rdf:resource="http://purl.org/bncf/tid/50001"/>
<skos:related rdf:resource="http://purl.org/bncf/tid/2762"/>
<skos:related rdf:resource="http://purl.org/bncf/tid/1977"/>
<skos:closeMatch
rdf:resource="http://www.wikidata.org/wiki/Q591041"/><skos:editorialNote>FONTE:
Soggettario; DeM: scientifico; VT; Wikipedia(IT)</skos:editorialNote>
</rdf:Description>
</rdf:RDF>
```

E un esempio di lista di entità, prive di relazioni esplicite, in formato TSV:

[22] https://www.w3.org/2004/02/skos/intro.
[23] https://thes.bncf.firenze.sbn.it/termine.php?id=10042.

<http://purl.org/bncf/tid/10041> Pubblicazioni ufficiali
<http://purl.org/bncf/tid/10042> Pubblicazioni scientifiche
<http://purl.org/bncf/tid/10043> Pubblicazioni militari

Non tutti gli algoritmi per l'indicizzazione, però, supportano o hanno le medesime performance con entrambi i formati[24], rivelando invece l'opportunità di avere a disposizione per l'addestramento un vocabolario SKOS.
Il vocabolario, inoltre, deve essere abbastanza ampio da referenziare tutti i termini usati per l'indicizzazione del *corpus* di addestramento. In Italia gli authority file tradizionalmente implementati e usati nel dominio GLAM/MAB sono gestiti in archivi separati con strutture e formati di dati spesso molto differenti.
Condizione necessaria, quindi, per l'addestramento di Annif ai fini di un suo utilizzo sistematico e non solo sperimentale, è l'individuazione di una soluzione per la gestione integrata dei dati di autorità in formati standard e adatti all'interoperabilità nel web, sul modello del tedesco GND - Gemeinsame Normdatei[25].
Allo stesso tempo, il *corpus* di addestramento deve essere costituito da porzioni di testo sufficientemente lunghe e "significative", capaci cioè di rappresentare adeguatamente il contenuto precipuo della risorsa informativa, abbinate a liste di termini riconducibili alle entità presenti nel vocabolario, che ne identifichino il contenuto semantico. Perché l'addestramento sia completo, i diversi ambiti disciplinari dovrebbero essere rappresentati in maniera omogenea. Purtroppo, anche questa seconda condizione, a causa dei limiti normativi e organizzativi di cui soffrono le biblioteche e le istituzioni culturali nel nostro Paese, non è di facile assolvimento.

Il deposito legale delle risorse digitali per la costituzione di corpora di dati affidabili e verificabili[26]

Dal punto di vista delle biblioteche, tra le principali novità introdotte nella normativa italiana sul diritto d'autore – Legge 22 aprile 1941, n. 633 –, a seguito del recepimento nel 2021 della Direttiva UE 2019/790 sul diritto d'autore e sui diritti connessi nel mercato unico digitale, si annoverano quelle presenti nel Capo V del Titolo 1 "Eccezioni e limitazioni", che possono essere sintetizzate nell'enunciato dei commi 1 e 2 dell'art. 70 ter.:

> «1. Sono consentite le riproduzioni compiute da organismi di ricerca e da istituti di tutela del patrimonio culturale, per scopi di ricerca scientifica, ai fini dell'estrazione di testo e di dati da opere o da altri materiali disponibili in reti o banche di dati cui essi hanno lecitamente accesso, nonché la comunicazione al pubblico degli esiti della ricerca ove espressi in nuove opere originali.

[24] Soltanto tre dei dieci algoritmi testati possono essere usati senza un vocabolario in formato SKOS.
[25] https://gnd.network/Webs/gnd/DE/Home/home_node.html.
[26] Su questo tema verteva anche l'intervento al 62° Congresso AIB di Elda Merenda, dal titolo *Deposito legale e intelligenza artificiale*: <https://www.aib.it/eventi/congr62/#capitolo-2>.

2. Ai fini della presente legge per estrazione di testo e di dati si intende qualsiasi tecnica automatizzata volta ad analizzare grandi quantità di testi, suoni, immagini, dati o metadati in formato digitale con lo scopo di generare informazioni, inclusi modelli, tendenze e correlazioni.».

Precedentemente, il Regolamento UE 2016/679 relativo alla protezione delle persone fisiche con riguardo al trattamento dei dati personali, nonché alla libera circolazione di tali dati (GDPR) aveva sancito la possibilità di conservare anche i dati personali «a fini di archiviazione nel pubblico interesse, di ricerca scientifica o storica o a fini statistici», come eccezione alla norma generale sul diritto all'oblio[27], stabilendo contemporaneamente le modalità del trattamento. Dal combinato disposto di queste due norme si evince chiaramente che le biblioteche, come gli altri istituti di tutela del patrimonio e gli organismi di ricerca, potrebbero utilizzare le collezioni di testi digitali e, più in generale, i *dataset* di cui sono legittimamente entrati in possesso, per lo sviluppo e il training di sistemi di Machine Learning.
Il nodo, potremmo affermare, si situa proprio nelle *legittime* modalità di acquisizione dei dati. Infatti, tradizionalmente, il canale preferenziale per l'accrescimento del patrimonio delle biblioteche è l'istituto del deposito legale[28] che prevede, in maniera similare in quasi tutto il mondo, l'obbligo di consegna presso le istituzioni individuate di tutti i documenti di interesse culturale pubblicati sul territorio nazionale. Anche in Italia, come nella maggioranza degli altri paesi europei, tale obbligo era stato esteso formalmente alle "risorse diffuse tramite rete informatica", vale a dire alle risorse nativamente digitali e alle risorse web, già con il d.P.R. 3/5/2006, n. 252. Tuttavia, esso non è mai stato regolamentato dal punto di vista tecnico, così come previsto nell'articolo 37[29] dello stesso d.P.R. 3/5/2006, n. 252, ed è quindi rimasto inattuato[30]. A causa di questa grave lacuna normativa, le risorse native digitali e web conservate nelle biblioteche, come si è già avuto modo di sottolineare in relazione al *web archiving*, sono quantitativamente ridotte, venen-

[27] Regolamento UE 2016/679 art. 17.

[28] La vigente normativa italiana sul deposito legale è riconducibile sostanzialmente alla l. 106/2004 e al d.P.R. 3/5/2006, n. 252. Maggiori approfondimenti sono disponibili sul sito della Direzione generale biblioteche e diritto d'autore: <https://biblioteche.cultura.gov.it/it/Attivita/deposito-legale/>.

[29] L'art. 37 c. 1 del d.P.R. 3/5/2006, n. 252 così recita: «Le modalità di deposito dei documenti diffusi tramite rete informatica sono definite con successivo regolamento adottato ai sensi dell'articolo 5, comma 1, della l. 106/2004, su proposta del Ministro per i beni e le attività culturali, di concerto con il Ministro delegato per l'innovazione e le tecnologie, sentite le associazioni di categoria interessate, nonché la Commissione per il deposito legale, di cui all'articolo 42 e il Comitato consultivo permanente per il diritto d'autore».

[30] Tra i più recenti contributi sul tema si veda Paola Puglisi, *Deposito legale quattordici anni dopo: come, quando, 'quanto', e perché,* «AIB Studi», 60 (2020), n. 3, <https://doi.org/10.2426/aibstudi-12477> e Giuliano Genetasio – Elda Merenda – Chiara Storti, *In the mangrove society: a collaborative legal deposit management hypothesis for the preservation of and permanent access to the National Cultural Heritage*, «JLIS.it», 13 (2022), n. 1, <https://doi.org/10.4403/jlis.it-12732>.

do acquisite principalmente grazie a depositi di tipo volontario o a seguito di esplicita richiesta di autorizzazione ai titolari. L'assenza di un mandato normativo specifico implica anche la generale grave carenza negli istituti del patrimonio di adeguate risorse, sia professionali che economiche, che sarebbero necessarie per un'efficiente gestione del patrimonio digitale. Si può quindi affermare che le biblioteche oggi non dispongono, e difficilmente potrebbero disporre, *legittimamente* di collezioni di testi digitali, in quantità utili all'addestramento di strumenti come Annif o di altri modelli di Natural Language Processing[31].

Ed è doveroso sottolineare che non si perde così solo l'opportunità che i sistemi di Intelligenza Artificiale servano davvero a migliorare i servizi delle biblioteche, ma anche e soprattutto che tali sistemi siano addestrati in maniera etica, garantendo totale trasparenza sulla fonte di acquisizione e le modalità di trattamento dei dati, oltre che la possibilità di condivisione dei modelli così ottenuti.

Rispetto alla disponibilità di collezioni di testi adatte al training degli strumenti di NLP e ML, vi è un ulteriore limite che potremmo definire "intrinseco" all'organizzazione degli attuali flussi catalografici nelle biblioteche italiane, in parte ancora conseguenza dall'assenza di obbligo di deposito legale delle risorse native digitali. In Italia, infatti, le risorse che rientrano nei tradizionali meccanismi di controllo bibliografico[32] e che sono soggette a procedure di indicizzazione semantica "controllata" – utili quindi per il training –, non sono disponibili in versione digitale, almeno non in modalità open[33]. Viceversa, le risorse maggiormente o esclusivamente disponibili in digitale, spesso in open access e dotate di ricchi metadati, sono relative a campi del sapere che in gran parte non rientrano nel controllo bibliografico strettamente inteso. Tra queste sicuramente la quasi totalità dei prodotti della ricerca, in ambito scientifico, che restano quasi esclusivo appannaggio dei gestori dei repository accademici. Per questi motivi, per la creazione di *corpora* di testi in grado di rappresentare molteplici campi del sapere, sarebbe necessario stabilire delle forme di cooperazione con le biblioteche e i sistemi bibliotecari di ateneo, oltre che con editori e distributori specializzati e no[34].

[31] Per farsi un'idea delle potenzialità dei *dataset* conservati dalle biblioteche per il NLP: *The Spanish Web Archive as a training field for Natural Language Processing models*, «IIPC Blog», 29 settembre 2021: <https://netpreserveblog.wordpress.com/2021/09/29/the-spanish-web-archive-as-a-training-field-for-natural-language-processing-models/>.

[32] Si fa qui riferimento, in primo luogo, alla redazione della Bibliografia nazionale italiana: <https://www.bncf.firenze.sbn.it/biblioteca/bibliografia-nazionale-italiana/>.

[33] Nel contesto appena delineato relativo alla mancanza di obbligo di deposito legale delle risorse digitali, con la conseguente possibilità di effettuare legittimamente attività di *text and data mining*, le licenze aperte sono requisito indispensabile. È però doveroso precisare che la sperimentazione ha dimostrato che sono sufficienti anche porzioni testuali più ridotte, come gli abstract, e che questi si potrebbero ottenere ad esempio con un accordo con i gestori del database dei libri in commercio.

[34] La creazione, gestione e conservazione a lungo termine di collezioni di dati, anche per finalità come quelle qui delineate, nell'ambito della cultura e della ricerca scientifica, sarebbe facilitata dalla creazione di una rete di soggetti sia pubblici che privati come quella ipotizzata in Associazione italiana

V.I.CO. - Vocabolario delle Identità Controllate

Il progetto V.I.CO. - Vocabolario delle Identità Controllate[35], avviato sempre in BNCF nel 2023, ha lo scopo di mettere a disposizione, in un ambiente nativamente linked open data, un vocabolario controllato partecipato dei termini in lingua italiana relativi alle entità (concetti, persone e organizzazioni, opere, luoghi, eventi ecc.) utili alla descrizione del contenuto dei testi, ovvero alla loro indicizzazione descrittiva e semantica. Come si è già avuto modo di rilevare, la disponibilità di un vocabolario in formato aperto è, infatti, uno dei requisiti per l'avvio dell'indicizzazione automatica dei testi in lingua italiana tramite il toolkit Annif.
In Italia i vocabolari controllati, incluse le liste di autorità, tradizionalmente implementati e usati per la descrizione e l'indicizzazione delle risorse nel dominio della cultura, sono gestiti in archivi separati, rispondenti a differenti ontologie. Un elenco non esaustivo comprende:

- il Thesaurus[36] del *Nuovo soggettario*;
- l'edizione italiana della Classificazione Decimale Dewey[37];
- gli authority file gestiti dall'ICCU relativi a Nomi, Opere e Luoghi[38];
- le liste di autorità prodotte in ambito archivistico[39];
- i vocabolari prodotti e mantenuti dall'ICCD[40].

A questi si potrebbero aggiungere i vocabolari prodotti al di fuori dell'ambito GLAM/MAB, in particolare nel dominio della ricerca, e ugualmente utili all'indicizzazione semantica delle risorse. A titolo esemplificativo:

- gli elementi Wikidata che presentano una *label* in italiano[41];
- il database dei nomi geografici GeoNames[42];
- le liste di nomi di autori presenti nei repository della ricerca[43];
- l'elenco dei settori scientifico-disciplinari (SSD) del MUR.

biblioteche, Gruppo di lavoro sulle biblioteche digitali, *Piano d'azione per l'infrastruttura nazionale della conoscenza*. Roma: Associazione Italiana Biblioteche, 2023. ISBN 978-88-7812-374-8, DOI: <https://doi.org/10.53263/9788878123748>.

35 https://vicobncf.wikibase.cloud/wiki/Main_Page.

36 https://thes.bncf.firenze.sbn.it/.

37 Attualmente l'edizione italiana della CDD ed. 23. o *Web Dewey* non è disponibile con licenza e formati aperti. A questo proposito si confrontino il contributo di Pietro Cavaleri, *WebDewey search*, «AIB Studi», 57 (2017), n. 2, <https://doi.org/10.2426/aibstudi-11644> e il progetto tedesco Web Dewey Search: <https://deweysearchde.pansoft.de/>.

38 https://www.iccu.sbn.it/it/normative-standard/authority-control/.

39 https://icar.cultura.gov.it/standard/standard-san.

40 http://www.iccd.beniculturali.it/it/strumenti-terminologici.

41 La query SPARQL effettuata sul database di Wikidata, tramite il servizio QLever, restituisce oltre 3.770.000 risultati (dati a gennaio 2024): <https://qlever.cs.uni-freiburg.de/wikidata/ewLSC3>.

42 https://www.geonames.org/.

43 A questo proposito, di particolare interesse il progetto IRIS del Gruppo Wikidata per Musei, Archivi e

V.I.CO. è stato quindi pensato come soluzione per la gestione condivisa di vocabolari controllati, alimentato da dati provenienti da diversi sistemi di dominio, sia centrali che locali. È ospitato su una istanza Wikibase Cloud[44]. La disponibilità di Wikibase come SaaS[45] consente di concentrare le limitate risorse a disposizione sulla analisi e gestione dei dati più che sulle necessità sistemistiche. Inoltre, da qualche anno, Wikibase è utilizzato a livello globale da numerose istituzioni MAB/GLAM per la gestione dei dati di dominio[46]. In particolare, è in rapida crescita il numero di biblioteche nazionali europee che lo adottano per l'allestimento di vocabolari controllati, in maniera esclusiva o, più frequentemente, come "repository LOD" dei dati. Tra queste, le biblioteche nazionali francese[47], tedesca[48], ceca[49], greca e portoghese.

Per il popolamento di V.I.CO. sono stati, innanzitutto, individuati i vocabolari di interesse prioritario rispetto all'utilizzo dei termini in essi contenuti per l'indicizzazione dei testi delle collezioni a disposizione per il training, e il loro ordine di importazione nell'istanza Wikibase[50].

In questa fase si è scelto di non riconciliare esplicitamente eventuali entità descritte in più di un vocabolario, come ad esempio le entità persona presenti sia in Wikidata che nell'Authority file Nomi dell'ICCU, ma per ciascuna entità sono stati importati, laddove presenti, gli identificativi dei vocabolari esterni, per garantire comunque la possibilità di riconciliazione lato macchina[51]. Per ciascuna tipologia di entità (nomi, luoghi ecc.), e per ciascun vocabolario, è stato poi definito un *subset* di dati di interesse ai fini dell'indicizzazione con Annif. Limitando in questo modo

Biblioteche:
<https://www.wikidata.org/wiki/Wikidata:Gruppo_Wikidata_per_Musei,_Archivi_e_Biblioteche/IRIS>, presentato anche in occasione della Giornata di studi dal titolo "La valutazione della ricerca alla prova dei fatti: modelli, politiche e progetti in corso", organizzata dall'Osservatorio della ricerca dell'Università di Firenze e dal Sistema bibliotecario di Ateneo, in collaborazione con Wikimedia Italia, Firenze, 17 gennaio 2024.

[44] <https://www.wikibase.cloud/>. Wikibase è un software open source per la creazione di basi di dati collaborative rispondenti allo standard RDF, nativamente multi-ontologie. Il software è mantenuto dall'associazione no profit Wikimedia Deutschland e da un'ampia comunità globale: <https://wikiba.se/>. La sua implementazione più nota è Wikidata, database libero e collaborativo contenente oltre 100 milioni di dati: <https://www.wikidata.org/>.

[45] SaaS – *Software as a Service* indica la modalità di distribuzione di un software o programma: il software è ospitato su una infrastruttura cloud ed è reso disponibile agli utenti attraverso un semplice client, come un web browser. Si veda la definizione fornita dal National Institute of Standards and Technology (NIST): <https://csrc.nist.gov/glossary/term/software_as_a_service>.

[46] https://learningwikibase.com/usecases/.

[47] https://www.transition-bibliographique.fr/fne/fichier-national-entites/.

[48] <https://wiki.dnb.de/pages/viewpage.action?pageId=147754828> e <https://docs.google.com/presentation/d/1_xdkofNIadhGJsNnINgovgqsz2UHJSVwpPkc3dpXeUs>.

[49] https://autority.wikimedia.cz/auto/Hlavn%C3%AD_strana.

[50] Per la scelta dei vocabolari si veda l'homepage di V.I.CO.: <https://vicobncf.wikibase.cloud/wiki/Main_Page>.

[51] A titolo esemplificativo è possibile confrontare l'entità riferita all'autore Carlo Collodi descritta nativamente nell'authority file dell'ICCU: <https://vicobncf.wikibase.cloud/wiki/Item:Q17>, e quella rappresentante la medesima entità in Wikidata: <https://vicobncf.wikibase.cloud/wiki/Item:Q19>.

la quantità di dati da importare in V.I.CO. e gli eventuali conseguenti problemi sia di carattere sistemistico che applicativo[52]; ferma restando l'attribuzione ad ogni entità di una proprietà relativa alla fonte originale dei dati, così da facilitare un possibile necessario ampliamento del tracciato dati e, più in generale, il loro aggiornamento. Successivamente sono state mappate le ontologie dei dataset originali ed è stato implementato il meta-modello dati di V.I.CO[53]. È bene specificare che il modello dati di V.I.CO. non ha ambizioni di tipo "ontologico", ma è puramente funzionale alla gestione in unico ambiente di vocabolari strutturati in base a logiche di dominio differenti. Infine, sono state sperimentate le modalità di import di ciascun dataset in V.I.CO., dipendenti dai formati e dalle possibilità di export dei vocabolari di origine.

A questo proposito una delle principali criticità da risolvere per la messa in produzione del servizio è la definizione dei criteri e delle modalità di aggiornamento dei dati nell'istanza Wikibase, sia in relazione all'aggiornamento dei dataset-fonte, sia rispetto alla possibilità di introdurre direttamente in V.I.CO. entità utili all'indicizzazione semantica e non già descritte altrove.

Appare evidente, in conclusione, come un progetto di tale portata richieda l'apporto di altre istituzioni e organizzazioni nazionali, non solo in qualità di fornitrici di dati ma, soprattutto, come portatrici di competenze sui temi in oggetto. La finalizzazione di accordi di collaborazione con i principali attori nazionali in questo ambito[54], tutti purtroppo affetti da gravi carenze di personale, si configura come una ulteriore sfida, questa volta di carattere amministrativo e organizzativo, per l'utilizzo dei sistemi di Intelligenza Artificiale nelle biblioteche.

Ampliando la prospettiva: Annif, Semantic Search e IA generativa

Il 2023 è stato, senza dubbio, l'anno dell'Intelligenza Artificiale generativa. Sistemi come ChatGPT4 rappresentano, per ora, uno dei punti di arrivo più raffinati e allo stesso tempo più "pop" della generazione degli strumenti che utilizzano modelli linguistici (LLM) creati grazie all'impiego delle reti neurali profonde con un'architettura basata sui *transformer*[55]. Tali reti neurali profonde prendono in considera-

[52] In generale non è ignorabile la necessità di disporre di risorse tecnologiche e computazionali per la gestione di grandi dataset.

[53] Classi e proprietà generali ontologia VICO: <https://tinyurl.com/36z3ukht>.

[54] Tra questi, in prima istanza, gli istituti del MiC deputati alla definizione degli standard nazionali di descrizione e metadatazione e alla gestione degli authority file ovvero ICCU, ICAR e ICCD, ma anche l'ICDP-DL che, nell'ambito delle attività previste dall'investimento "M1C3 1.1 Strategie e piattaforme digitali per il patrimonio culturale" del Piano nazionale di ripresa e resilienza (PNRR), sta lavorando alla costituzione di una infrastruttura nazionale dei dati della cultura. Inoltre, colloqui informali sono stati avviati con il gruppo GWMAB che ha ormai un'esperienza pluriennale nella gestione di dati in ambiente Wikibase/Wikidata: <https://www.wikidata.org/wiki/Wikidata:Gruppo_Wikidata_per_Musei,_Archivi_e_Biblioteche>.

[55] Un'efficace rappresentazione del funzionamento dei *transformer* è quella fornita nell'articolo *Generative AI exists because of the transformer*, pubblicato il 12 settembre 2023 sul «Financial

zione un grandissimo numero di parametri ma, soprattutto, sfruttano una modalità di analisi dei *token* linguistici che potremmo definire altamente "semantica", poiché li esamina non soltanto in base alla loro sequenza ma ne soppesa il contesto d'uso. Queste tecniche di apprendimento sono generalmente non supervisionate, cioè non necessitano di intervento umano per l'etichettatura dei dati, almeno non in una prima fase[56]. Il risultato sono interfacce di *Semantic Search*[57] e di IA generativa le cui capacità di recupero dell'informazione, anche in presenza di enormi quantità di dati, sono evidentemente sorprendenti, e sempre meno affette da "allucinazioni"[58]. In questo contesto, ci si potrebbe chiedere quale sia il senso di continuare a investire sullo sviluppo di strumenti di indicizzazione semantica automatica[59], come Annif, basati su forme tradizionali di Machine Learning. Utilizzando dati pre-addestrati secondo un modello tradizionale di organizzazione della conoscenza, Annif sembra rispondere alla sola esigenza di velocizzare e rendere maggiormente efficienti processi consolidati che vedono il bibliotecario (o un qualsiasi esperto in un dato ambito del sapere) tentare di prevedere non solo *cosa* gli utenti cercheranno, ma anche *come* lo faranno, utilizzando una sintassi e una semantica che sono appannaggio quasi esclusivo del mondo MAB/GLAM.

La *Semantic Search* e l'AI generativa ribaltano, invece, la prospettiva cercando di restituire risultati quanto più *probabilmente* attinenti alla ricerca dell'utente, indipendentemente dalla sua formulazione. Questi sistemi, infatti, forniscono risposte proprio sulla base di modelli probabilistici complessi creati, come appena accennato, grazie all'utilizzo di reti neurali profonde che analizzano grandi quantità di dati. Ci troviamo cioè davanti a quell'apparente dualismo nei sistemi di organizzazione della conoscenza che Gino Roncaglia ha di recente egregiamente sintetizzato nella metafora dell'Architetto e dell'Oracolo[60]. Come affermato anche dallo stesso Roncaglia, non siamo ora in grado di dire se e quale approccio prevarrà sull'altro, e con che tempi. Restano però valide, qualsiasi sia lo scenario futuro, tutte le considerazioni circa la necessità di avere finalmente una normativa nazionale che preveda l'obbligo di deposito legale delle risorse native digitali, che consenta alle biblioteche di essere parte attiva nella costruzione di tali strumenti. Allo stesso tem-

Times», a firma del Visual Storytelling Team e di Madhumita Murgia: <https://ig.ft.com/generative-ai/>.

[56] I modelli ottenuti partendo da un processo di apprendimento autonomo possono essere perfezionati tramite processi di apprendimento supervisionato.

[57] https://www.sbert.net/examples/applications/semantic-search/README.html.

[58] Andrea Fontana, *Anche l'intelligenza artificiale può avere le allucinazioni*, «Wired.it», 18 giugno 2023: <https://www.wired.it/article/intelligenza-artificiale-allucinazioni-cause-conseguenze/>.

[59] Interessante la serie di progetti e soluzioni presentate in «Cataloging & Classification Quarterly», 59 (2021), n. 8, fascicolo interamente dedicato ai sistemi di indicizzazione semantica automatica: <https://www.tandfonline.com/toc/wccq20/59/8>.

[60] Gino Roncaglia, *L'architetto e l'oracolo: forme digitali del sapere da Wikipedia a ChatGPT*, Bari; Roma: Laterza, 2023.

po, all'interno delle istituzioni culturali si dovrebbe avviare una riflessione più profonda sul concetto di *qualità* dei dati prodotti, che non può più essere misurata solo in base all'aderenza a standard stabiliti all'interno delle stesse comunità di dominio, ma che dovrebbe piuttosto considerare la possibilità di un loro riuso anche per l'addestramento dei sistemi di Intelligenza Artificiale. Se non si invertono le attuali tendenze, le biblioteche in Italia non saranno in grado di utilizzare le reali potenzialità delle Intelligenze Artificiali, e soprattutto non potranno esercitare il ruolo di produttori e gestori di conoscenza secondo i principi, che sono loro propri, di legalità, eticità e trasparenza.

In 2022, the National Central Library of Florence launched an experimental project to train the Annif tool for automatic semantic indexing in the Italian language, developed and maintained by the National Library of Finland. The project revealed a more general lack of the prerequisites in Italian libraries for using not only the more traditional Machine Learning tools but also new Artificial Intelligence systems. The causes can be traced, on the one hand, to deficient legislation on the legal deposit of digital resources and, on the other hand, to the fragmentation and multiplication of vocabularies and thesauri, which respond to different ontologies and are structured in formats not always suitable for web interoperability. The purpose of this contribution is to initiate a reflection on these issues, attempting to outline the role that libraries could play in the ethical and conscious development and use of Artificial Intelligence systems, with an in-depth look at the ongoing project called V.I.CO. - Vocabulary of Controlled Identities.

L'ultima consultazione dei siti web è avvenuta nel mese di giugno 2024