# Affinare il contesto: estrazione di informazioni strutturate per l'arricchimento dei contesti archivistici

Lucia Giagnolini[1], Paolo Bonora[2], Francesca Tomasi[3]
[1] Università di Bologna, Italia - lucia.giagnolini2@unibo.it
[2] Università di Bologna, Italia - paolo.bonora@unibo.it
[3] Università di Bologna, Italia - francesca.tomasi@unibo.it

## ABSTRACT[1]

Gli strumenti di corredo archivistici in LOD sono spesso solo parzialmente capaci di esprimere il vero potenziale informativo dei dati, a causa della molteplicità di campi non strutturati presenti nelle descrizioni dei complessi documentari. La presenza di numerose sezioni *literal*, ovvero a testo pieno, limita da un lato la possibilità di interrogazioni a base semantica e dall'altro non consente l'apertura ai numerosi contesti latenti che tali porzioni di testo non strutturato veicolano. Si intende allora qui presentare una metodologia per acquisire nuova conoscenza dai dati, aprendoli al dialogo con nuovi contesti impliciti. A questo scopo, si è valutato quanto possano essere utili alcuni tool ad oggi disponibili per acquisire e restituire informazione strutturata rispetto alle descrizioni archivistiche. Attraverso un caso di studio del Sistema Archivistico Nazionale in LOD si sono analizzate le potenzialità di TINT, FRED e di ChatGPT nell'estrarre informazione morfosintattica, lessicale o semantica dai dati archivistici, riflettendo al contempo sulla possibilità di far dialogare il grafo di conoscenza nativo e il grafo risultante dall'analisi, e documentando gli atti interpretativi emersi.

## PAROLE CHIAVE

Linked Open Data; archivi; information retrieval; supervised annotation; contesti.

## 1. PREMESSA

Un archivio è composto da elementi che, ciascuno con le proprie caratteristiche, permettono di fornire una rappresentazione stratificata del complesso di attività ed entità coinvolte nella produzione dei documenti. Per renderlo manifesto, è necessaria una chiara esposizione della rete di relazioni che collega le singole parti tra loro e il fondo stesso con i molteplici contesti di riferimento [5]. Se, da un lato, le descrizioni basate sullo standard metodologico ISAD(G) hanno permesso una funzionale formalizzazione e strutturazione dell'atto descrittivo, dall'altro è ormai largamente appurato che l'applicazione dello standard ha determinato rappresentazioni primariamente impostate su relazioni gerarchiche – dunque strettamente verticali e scarsamente permeabili ai contesti – valorizzando solo marginalmente la restituzione in termini di legami orizzontali [5, 17]. Anche per questo motivo, da ormai più di un decennio, le istituzioni dell'ambito GLAM si sono avvicinate al paradigma dei Linked Open Data (LOD), che "ha richiesto di rivedere sistematicamente le informazioni riportate nelle descrizioni archivistiche e nelle schede catalografiche (destrutturando e ristrutturando tipologia, granularità e precisione), così da superare gli schemi documento-centrici della descrizione con approcci data-centrici, che valorizzano le relazioni con il contesto" [8]. Nella migrazione di inventari archivistici tradizionali in LOD, blocchi informativi estremamente rilevanti – come la nota biografica, la storia archivistica e i criteri di riordinamento – vengono spesso trasposti esclusivamente come corpose stringhe di caratteri *literal*, ossia in testo pieno. Si tratta di campi testuali estremamente ricchi di informazioni, che potrebbero essere strutturate in modo più organizzato e funzionale, rappresentando "il carburante indispensabile a far decollare il razzo dell'integrazione multicontestuale" [15: 5]. Infatti, il Semantic Web non ha cambiato l'approccio tradizionale delle istituzioni alla descrizione, ma ha enfatizzato la necessità di adottare una semantica esplicita, in modo tale da consentire una interoperabilità basata sull'impiego di modelli concettuali, facilitando il riuso dei dati [14]. Ogni nuova asserzione espressa in forma di tripla diventa generatrice di inferenza e di nuova informazione: più i contesti di appartenenza di queste asserzioni crescono e si intersecano, più la rete semantica si arricchisce e diventa informazione classificata [10]. In altre parole, i contenuti testuali di campi descrittivi, espressi come sequenze di stringhe e rappresentati in forma aggregata attraverso semplici nodi di tipo *literal* – pur mantenendo la loro unitarietà nel campo descrittivo – potrebbero essere esplicitati attraverso nuove triple. Ogni nuova tripla diventerebbe portatrice di una componente informativa specifica presente nel testo aggregato come, ad esempio, le attestazioni di istituzioni, persone, eventi e coordinate spazio-temporali. L'estrapolazione della specifica semantica del dato preesistente, attraverso la creazione di

[1] L. Giagnolini ha curato le sezioni 1 e 2, con P. Bonora la sezione 3; le conclusioni sono il risultato di una riflessione collettiva degli autori.

triple attestanti relazioni più o meno esplicite nel testo, consentirebbe un arricchimento significativo della *knowledge base* contestuale, permettendo, fra l'altro, un notevole miglioramento delle operazioni di ricerca e, potenzialmente, anche la disambiguazione delle entità citate.
Le operazioni di estrazione delle entità presenti nel testo e l'attribuzione della semantica di relazione tra di esse rappresentano a tutti gli effetti un atto interpretativo del contenuto testuale [6]. È necessario, quindi, che le nuove triple, indipendentemente dal fatto che siano il risultato di un processo di estrazione supervisionato o meno, vengano esplicitamente individuate come il risultato di una nuova attività di analisi, distinta dall'azione di descrizione archivistica, che ha prodotto il record originario. Le triple finalizzate all'arricchimento del dato, dovranno quindi essere corredate da una serie di ulteriori triple che ne dichiarino espressamente l'origine, le modalità di produzione e, in ultima istanza, l'attribuzione di responsabilità. Ovvero, ne esplicitino la cosiddetta *provenance* [14].
In sintesi, l'adozione del paradigma Linked Open Data (LOD) ha aperto la strada a una nuova prospettiva nella descrizione archivistica [12], ma per superare davvero le limitazioni gerarchiche e favorire la creazione di una knowledge base semanticamente ricca occorre sfruttare al meglio i campi di testo e strutturarne i contesti latenti, aggiungendo adeguata documentazione al processo di produzione di nuova conoscenza.

## 2. METODOLOGIA E WORKFLOW

Il tentativo di "stabilire se e in che misura le tecniche e le tecnologie di gestione del testo possano potenziare i nostri strumenti nel rispetto del contesto, arricchendoli di appigli informativi" [15: 7], si traduce nel comprendere come impiegare gli strumenti ad oggi disponibili per acquisire informazione strutturata dalle descrizioni archivistiche. A questo scopo, è necessario chiarire gli step del processo di estrazione dell'informazione, ovvero elaborare un modello di workflow che sia capace di contemplare tanto l'esigenza di definire il tipo di analisi da delegare allo strumento, quanto la necessità di valutare degli esiti della sua applicazione. L'approccio che proponiamo per l'implementazione del processo è articolato nei seguenti punti:

1. Selezionare il tipo di atto interpretativo che si delega allo strumento (ad esempio, analisi morfosintattica, lessicale o semantica) a seconda dei contenuti da analizzare.
2. Individuare le tecnologie e le rispettive implementazioni in funzione del tipo di atto interpretativo atteso (ad esempio, da tecniche NLP elementari al deep learning e LLM).
3. Definire il modello di valutazione dell'esito e della qualità dell'atto interpretativo automatico, dove per qualità si intende "la possibilità di attingere a dati ragionevolmente affidabili, perché parte di un contesto che li giustifica e li spiega" [16: 10]. Gli output dell'atto interpretativo dovranno, infatti, essere vagliati e selezionati da un esperto di dominio per essere ritenuti validi.
4. Individuare il modello di rappresentazione e sedimentazione della conoscenza estratta nell'ottica di una struttura semanticamente controllata e interoperabile dal punto di vista dell'acceso al dato (ad esempio, RDF in prospettiva LOD).
5. Modellare i criteri e le modalità di acquisizione della conoscenza estratta in funzione della capacità espressiva del relativo modello descrittivo (ad esempio, Dublin Core, RiC-O, SAN LOD) e dei criteri redazionali. A questo scopo andrà definito un modello di attestazione della *provenance* del dato che espliciti il tipo di atto interpretativo, lo strumento e il processo utilizzato per ottenerlo, le metriche di valutazione (ad esempio, *recall* e *precision*) e il riferimento al supervisore scientifico (ossia l'attribuzione di responsabilità).
6. Valutare le strategie per mettere in relazione il dato analizzato nel sistema nativo e la serie di triple esito dell'atto interpretativo.
7. Modellare l'interazione utente-sistema in termini di processo operativo e di interfacce, ovvero individuare strategie di information visualization che consentano al contenuto informativo estratto di essere adeguatamente presentato e gestito.
8. Valutare potenziali modalità di rinforzo dello strumento esterno per migliorarne le prestazioni (ad esempio, training set NLP, miglioramento del prompt per ChatGPT).

In questa formulazione, il processo è sufficientemente astratto da poter essere applicato a contesti diversi e obiettivi di estrazione dell'informazione operanti a molteplici livelli: dall'analisi della superficie lessicale all'interpretazione della semantica del testo.

## 3. PROOF OF CONCEPT

Per presentare possibili esiti dell'approccio metodologico così illustrato, è possibile effettuare sperimentazioni con strumenti immediatamente disponibili che, anche se non ancora integrati all'interno di un concreto workflow, non presentano barriere d'accesso tali da impedirne un utilizzo dimostrativo.

Ad esempio, possiamo analizzare il campo denominato "descrizione" di una scheda "soggetto produttore" del Sistema di Archiviazione Nazionale (SAN), corrispondente alle proprietà "dc:description" e "abstract" del tracciato schema SAN. Individuiamo nella scheda SAN dedicata alla nota biografica di Andrea Costa[2] (1851-1910) il testo campione, prendendo in analisi il primo paragrafo della descrizione:

> Nasce a Imola il 29 novembre 1851 da Pietro e Rosa Tozzi in una famiglia cattolica praticante e di modeste condizioni. Il giorno successivo è battezzato nella cattedrale di S. Cassiano con i nomi di Andrea, Antonio e Baldassarre e suo padrino è Orso Orsini. Frequenta le scuole elementari gestite da un sacerdote e negli anni scolastici 1866-1867 e 1867-1868 frequenta la scuola tecnica comunale con Gaetano Darchini, Luigi Sassi e Angelo Negri. Negli anni scolastici 1868-1869 e 1869-1870 frequenta il liceo come uditore per le lezioni di letteratura italiana e latina. Il 15 dicembre 1870 si iscrive alla facoltà di filosofia e belle lettere dell'Università di Bologna come "studente libero" non avendo la possibilità di pagare le regolari tasse di ammissione e per mantenersi si impiega come scrivano in un'agenzia di assicurazioni imolese. Lì un impiegato, Paolo Renzi, lo associa, o almeno lo avvicina, all'Internazionale. A Imola e a Bologna compie il suo noviziato, nell'atmosfera che presto si accenderà degli entusiasmi per la Comune, e nel contatto con Carducci, che lo predilige fra i suoi allievi.

Stante l'obiettivo di identificare, isolare ed estrarre informazioni relative al soggetto contenute nella nota biografica, abbiamo selezionato tre strumenti progettati per operare sui tre livelli di analisi del testo (passo n. 1 e n. 2 del workflow): annotazione morfosintattica e NER (Tint[3] [1]); estrazione del significato su base lessicale e dei relativi nessi sintattici (FRED[4] [9]); interpretazione del testo su base statistica (ChatGPT[5]). I tre strumenti sono stati selezionati in considerazione del grado di maturità, dell'immediatezza d'uso, delle possibilità di ulteriore affinamento dell'addestramento. La sperimentazione, lungi dall'essere definitiva, mira alla semplice verifica della percorribilità dell'approccio basandosi su soluzioni terze, prescindendo dallo sviluppo in proprio di componenti dedicati.

L'applicazione di Tint per l'analisi del primo paragrafo della nota biografica ha permesso di individuare, tramite NER, i nomi di organizzazioni, luoghi e persone (vd. Fig. 1), così come le dipendenze sintattiche del testo, classificando in modo automatico le parti del discorso[6] (vd. Fig. 2).


*Figura 1. Entità riconosciute nel testo e relativa classificazione*

[2] http://dati.san.beniculturali.it/SAN/produttore_IT-ER-IBC_san.cat.sogP.66756
[3] Per le finalità di questo intervento, è stata utilizzata la versione "Online demo" disponibile al link: https://dh.fbk.eu/tint-demo/
[4] Per le finalità di questo intervento, è stata utilizzata la versione "Online demo" disponibile al link: http://wit.istc.cnr.it/stlab-tools/fred/demo/
[5] Per le finalità di questo intervento è stata utilizzata la versione GPT-3.5 https://chat.openai.com/
[6] Per visualizzare il risultato integrale dell'analisi effettuata tramite Tint, v. Giagnolini, Lucia, and Paolo Bonora. "Affinare Il Contesto: Estrazione Di Informazioni Strutturate Per L'arricchimento Dei Contesti Archivistici. Risultati Dell'analisi Effettuata Con Tint". figshare, January 31, 2024. https://doi.org/10.6084/m9.figshare.25119116.v4

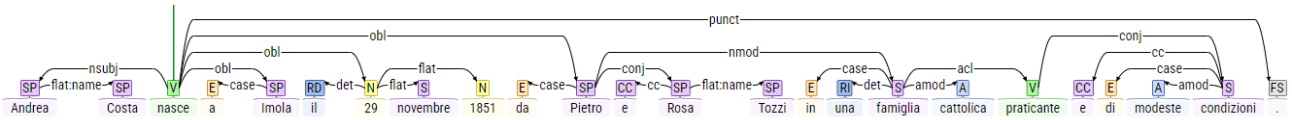


*Figura 2. Grafo delle dipendenze sintattiche relative al primo periodo*

L'applicazione di FRED ha prodotto una rappresentazione unificata e formalizzata in grafo di fatti e concetti espressi dal testo in linguaggio naturale[7] come, ad esempio, l'interpretazione delle condizioni di nascita di Andrea Costa (vd. Fig. 3).

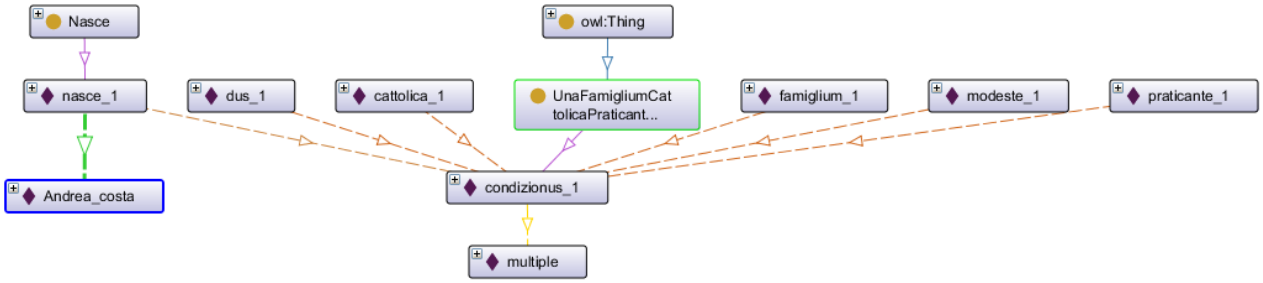


*Figura 3. Grafo dell'interpretazione delle condizioni di nascita elaborata da FRED*

ChatGPT 3.5 è stata utilizzata per sondare l'emersione di concetti più complessi, potenzialmente variabili a seconda dell'obiettivo e della strutturazione del prompt. A scopo esemplificativo, il prompt "*Elenca le entità che emergono da questo testo*" ha fornito i seguenti risultati:

- **Luogo di nascita**: Imola (29 novembre 1851)
- **Genitori**: Pietro e Rosa Tozzi
- **Ambiente familiare**: Famiglia cattolica praticante e di modeste condizioni
- **Atti religiosi**: Battezzato nella cattedrale di S. Cassiano con i nomi di Andrea, Antonio e Baldassarre; padrino Orso Orsini
- **Educazione**: Scuole elementari gestite da un sacerdote; frequenta la scuola tecnica comunale con Gaetano Darchini, Luigi Sassi e Angelo Negri; frequenta il liceo come uditore per le lezioni di letteratura italiana e latina
- **Università**: Iscritto alla facoltà di filosofia e belle lettere dell'Università di Bologna il 15 dicembre 1870 come "studente libero"
- **Situazione finanziaria**: Non ha la possibilità di pagare le regolari tasse di ammissione all'università; si mantiene lavorando come scrivano in un'agenzia di assicurazioni imolese
- **Influenza politica**: Associato o avvicinato all'Internazionale da Paolo Renzi, un impiegato nell'agenzia di assicurazioni
- **Contesto storico/politico**: Noviziato a Imola e a Bologna nell'atmosfera degli entusiasmi per la Comune; contatto con Carducci, che lo predilige fra i suoi allievi.

I tre strumenti producono output con gradi di finitura crescenti. Ai risultati prodotti sia con Tint che da FRED devono essere applicati algoritmi di estrazione in base a criteri semantici [3] per ottenere informazione strutturata da sottoporre alla valutazione degli esperti di dominio (passo n. 3 del workflow). Più immeditatamente strutturato è, invece, l'output prodotto dall'LLM (ChatGPT) che comunque dovrà essere allineato alla struttura del modello concettuale (passo n. 4).
Notiamo, inoltre, che l'estrazione delle informazioni di carattere temporale contenute nel testo, fondamentali per l'arricchimento dei contesti, prodotta dai tre strumenti oggetto della sperimentazione risulta lacunosa. Per superare questo limite, è possibile prevedere l'integrazione di strumenti dedicati come i Time Taggers [13] nel passo n.2 o un ulteriore raffinamento del prompt di ChatGPT[8].
A questo punto, occorre tenere presente che le informazioni estratte con questo approccio potrebbero non essere direttamente reintegrabili nella knowledge base d'origine, per limitazioni ontologiche della stessa (passo n. 6). Per quanto

[7] Per visualizzare il risultato integrale dell'analisi effettuata tramite FRED, v. Giagnolini, Lucia, and Paolo Bonora. "Affinare Il Contesto: Estrazione Di Informazioni Strutturate Per L'arricchimento Dei Contesti Archivistici. Risultati Dell'analisi Effettuata Con Fred". figshare, April 3, 2024. https://doi.org/10.6084/m9.figshare.25534225.v1
[8] https://platform.openai.com

riguarda i dati strutturati estratti dalla nota biografica di Andrea Costa, ad esempio, nell'ambito del modello proposto dall'ontologia SAN LOD[9], emerge l'assenza di classi e proprietà in grado di rappresentare adeguatamente le informazioni estratte. Dunque, dal momento che la rappresentatività del modello da cui sono state acquisite le informazioni potrebbe diventare un ulteriore ostacolo all'esplicitazione dei contesti latenti, è più opportuno optare per un approccio che astragga dalle infrastrutture specifiche. Gli esiti dell'analisi possono confluire in un grafo in grado di rappresentare le nuove triple nella loro massima granularità in una estensione stand-off. La nota biografica – o altri campi descrittivi dell'infrastruttura di riferimento – possono diventare l'oggetto di un asserto che stabilisce il legame tra la knowledge base d'origine e il grafo derivante dall'interpretazione del testo, sulla base, ad esempio, del modello proposto dalla Web Annotation Ontology[10] (vd. Fig. 4).

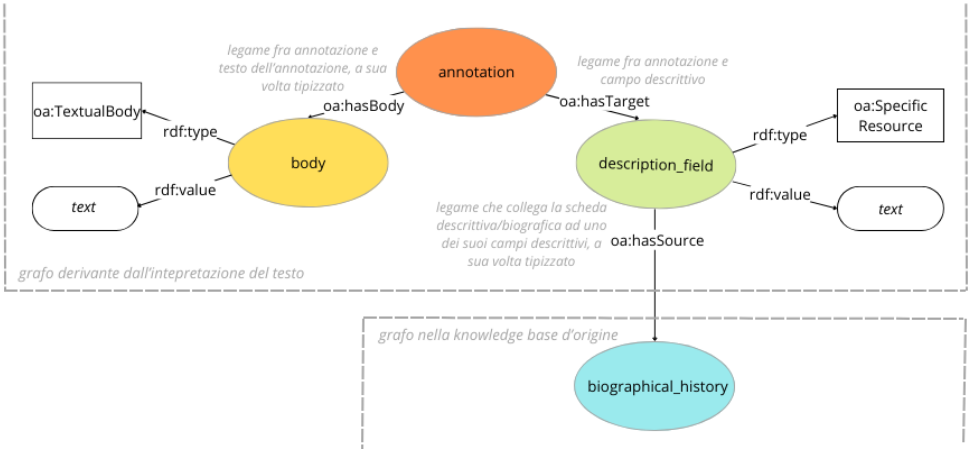


*Figura 4. Rappresentazione del legame tra la knowledge base d'origine e il grafo derivante dall'interpretazione del testo*

La Web Annotation Ontology permetterebbe di dar conto della *provenance* degli atti interpretativi e della loro potenziale molteplicità – sia in termini di strumenti che di esperti di dominio – con un conseguente ampliamento delle possibilità di esplicitazione dei contesti alla base dei nuovi grafi (passo n. 5) [6, 7]. Tuttavia, occorrerà estendere l'ontologia in modo tale da veicolare adeguatamente le informazioni estratte, a seconda delle esigenze di rappresentazione.
Sarà poi necessario individuare forme di visualizzazione dell'informazione così rappresentata, capaci di restituire all'utente nuova conoscenza attraverso un accesso sapiente alla molteplicità dei contesti che emergono dai processi di analisi (passo n. 7).

## 4. CONCLUSIONI E PROSPETTIVE

Questo contributo intende mettere in luce il ruolo fondamentale della destrutturazione del dato testuale per consentire un recupero semantico efficace dell'enorme e preziosissima mole di inventari archivistici pubblicati in rete. Infatti, il suo obiettivo non è valutare le performance dei singoli strumenti, bensì presentare una proposta di approccio metodologico per l'estrazione automatica di informazioni strutturate dalle descrizioni archivistiche. Ciò non toglie che, in una prospettiva di sviluppo, non ci si potrà esimere da un'analisi comparata approfondita dell'efficacia di questo approccio in termini di quantità e qualità dell'informazione estratta. In questi termini, le prime sperimentazioni, precedentemente illustrate, sembrano fornire risultati benauguranti, anche in considerazione del fatto che l'automazione della procedura non esaurisce la capacità interpretativa dell'essere umano, ma assume una funzione coadiuvante e documentale della stessa.
Le prospettive di lavoro si aprono soprattutto verso lo sviluppo e l'applicazione dei LLM [4, 11]: occorrerà determinare quali e quante operazioni di analisi siano effettivamente in grado di effettuare e con che grado di raffinatezza, anche per stabilire se i risultati forniti da tecniche NLP siano già superati o superabili, ragionando su come affinare gli strumenti utilizzati o il loro addestramento (passo n. 8). Il passo successivo si tratteggia sulla modellazione ed estrazione di concetti che vadano "oltre l'identificazione delle entità, enfatizzando sistemi di relazioni sempre più larghi e suggerendo di affiancare alla consolidata multilivellarità una multidimensionalità capace di rendere 'visibili' le idee o i fatti di cui le diverse entità sono veicoli" [16: 5]. Notiamo, infatti, che i risultati ottenuti attraverso l'utilizzo di ChatGPT offrono una gamma di informazioni che superano l'individuazione delle entità canoniche. Tuttavia, stabilizzarne e strutturarne i dati risultanti risulta notevolmente più complesso rispetto all'adozione di tecniche più tradizionali come la NER. Questo

[9] http://dati.san.beniculturali.it/lode/aggiornato.htm#d4e2193
[10] https://www.w3.org/ns/oa

sottolinea l'importanza di ulteriori ricerche e sviluppi nel campo sia della *knowledge graph generation* dal linguaggio naturale che dell'uso delle classificazioni prodotte dagli LLM, al fine di aumentare il potenziale di tali modelli per la massimizzazione del contenuto informativo delle descrizioni archivistiche [2, 11].

## 5. RINGRAZIAMENTI

Contributo parzialmente finanziato dall'Unione europea - Next Generation EU, investimento I.4.1 Borse PNRR Patrimonio Culturale, Decreto Ministeriale n. 351 del 9 aprile 2022.

## BIBLIOGRAFIA

[1] Aprosio, Alessio Palmero, e Giovanni Moretti. «Tint 2.0: An All-Inclusive Suite for NLP in Italian». In *Proceedings of the Fifth Italian Conference on Computational Linguistics CLiC-It. 10-12 dicembre 2018*, a cura di Elena Cabrio, Alessandro Mazzei, e Fabio Tamburini, 311–17. Torino: Accademia University Press, 2018. https://doi.org/10.4000/books.aaccademia.3571

[2] Babaei Giglou, Hamed, Jennifer D'Souza, e Sören Auer. «LLMs4OL: Large Language Models for Ontology Learning». In *The Semantic Web – ISWC 2023*, a cura di Terry R. Payne et al., 408–427. Cham: Springer Nature Switzerland, 2023. https://doi.org/10.1007/978-3-031-47240-4_22

[3] Bonora, Paolo, e Angelo Pompilio. «Automatic Extraction of Opera Character Characteristics through Lexical-Syntactic Patterns». *Umanistica Digitale* 5, fasc. 10 (gennaio 2021): 193–210. https://doi.org/10.6092/issn.2532-8816/12426

[4] Colavizza, Giovanni, Tobias Blanke, Charles Jeurgens, e Julia Noordegraaf. «Archives and AI: An Overview of Current Debates and Future Perspectives». *Journal on Computing and Cultural Heritage* 15, fasc. 1 (14 dicembre 2021): 4:1-4:15. https://doi.org/10.1145/3479010

[5] Damiani, Concetta. «Archival Description and Conceptual Transversality». *JLIS.It* 13, fasc. 3 (15 settembre 2022): 154–161. https://doi.org/10.36253/jlis.it-485

[6] Daquino, Marilena, e Francesca Tomasi. «Historical Context Ontology (HiCO): A Conceptual Model for Describing Context Information of Cultural Heritage Objects». In *Metadata and Semantics Research. MTSR 2015. Communications in Computer and Information Science*, a cura di Emmanouel Garaoufallou, Richard J. Hartley, e Panorea Gaitanou, 544:424–436. Springer International Publishing, 2015. https://doi.org/10.1007/978-3-319-24129-6_37

[7] Daquino, Marilena, Valentina Pasqual, e Francesca Tomasi. «Knowledge Representation of digital Hermeneutics of archival and literary Sources». *JLIS: Italian Journal of Library, Archives and Information Science = Rivista italiana di biblioteconomia, archivistica e scienza dell'informazione: 11, 3, 2020*, fasc. 3 (2020): 59–76. https://doi.org/10.4403/jlis.it-12642

[8] Daquino, Marilena. «Linked Open Data native cataloguing and archival description». *JLIS* 12, fasc. 3 (2021): 91–104. https://doi.org/10.4403/jlis.it-12703

[9] Gangemi, Aldo, Valentina Presutti, Diego Reforgiato, Andrea Giovanni Nuzzolese, Francesco Draicchio, e Misael Mongiovì. «Semantic Web Machine Reading with FRED». *Semantic Web* 8, fasc. 6 (2017): 873–893. https://doi.org/10.3233/SW-160240

[10] Guerrini, Mauro, e Tiziana Possemato. «Linked data: un nuovo alfabeto del web semantico». *Biblioteche oggi* 30, fasc. 3 (2012): 7–15.

[11] Mihindukulasooriya, Nandana, Sanju Tiwari, Carlos F. Enguix, e Kusum Lata. «Text2KGBench: A Benchmark for Ontology-Driven Knowledge Graph Generation from Text». In *The Semantic Web – ISWC 2023*, a cura di Terry R. Payne et al., 247–265. Cham: Springer Nature Switzerland, 2023. https://doi.org/10.1007/978-3-031-47243-5_14

[12] Polley, Katherine Louise, Vivian Teresa Tompkins, Brendan John Honick, e Jian Qin. «Named Entity Disambiguation for Archival Collections: Metadata, Wikidata, and Linked Data». In *Proceedings of the Association for Information Science and Technology 58*, 1:520–524, 2021. https://doi.org/10.1002/pra2.490

[13] Strötgen, Jannik, e Michael Gertz. «A Baseline Temporal Tagger for All Languages». In *In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 541–547. Lisbon, Portugal: Association for Computational Linguistics, 2015. https://doi.org/10.18653/v1/D15-1063

[14] Tomasi, Francesca. «Archival Finding Aids in Linked Open Data between Description and Interpretation». *JLIS.It* 14, fasc. 3 (2023): 134–146. https://doi.org/10.36253/jlis.it-557

[15] Valacchi, Federico. «Not the Institutions but the Subjects Matter. Beyond the Necessary Approximation of Finding Aids?» *JLIS.It* 14, fasc. 3 (2023): 1–14. https://doi.org/10.36253/jlis.it-539.

[16] Valacchi, Federico. «The Parts and the Whole. Integrate Knowledge». *JLIS.It* 13, fasc. 3 (2022): 1–11. https://doi.org/10.36253/jlis.it-477

[17] Vitali, Stefano. «La descrizione degli archivi nell'epoca degli standard e dei sistemi informatici». In *Archivistica. Teorie, metodi, pratiche*, a cura di Linda Giuva e Maria Guercio, 179–210. Roma: Carocci, 2024.