L'interazione fa la forza
Il paradigma Linked Data applicato alle risorse linguistiche per il latino

Marco Passarotti*

The paper presents the LiLa Knowledge Base of interoperable linguistic resources for Latin. After introducing the problem of the sparsity and separation of lexical and textual resources (both for Latin and for other languages), the paper describes the principles of the Linked Data paradigm, on which the LiLa Knowledge Base is grounded. The lemma-centered architecture of LiLa is then detailed, as well as the linguistic resources for Latin currently interlinked through the Knowledge Base. Two different means to query the interoperable resources are presented and described with examples. Finally, some thoughts about the use of digital linguistic data to move from information to knowledge are provided.

1. Introduzione

Una delle principali sfide attuali nel campo della Linguistica Computazionale consiste nel dare un senso alla grande massa di raccolte digitali di evidenza empirica in forma di 'risorse linguistiche' di tipo sia testuale (come corpora) che lessicale (come lessici e dizionari) prodotta nel corso degli ultimi decenni per molte lingue (Chiarcos et al. 2017, p. 1).

Per comprendere cosa si intenda con 'dare un senso' si possono richiamare i principii che oggi guidano la produzione, gestione e pubblicazione dei dati in campo scientifico, riassunti nell'acronimo FAIR (Wilkinson et al. 2016), in base a cui i (meta)dati devono essere:

- *Findable*: i (meta)dati sono associati a un identificativo unico e persistente;
- *Accessible*: i (meta)dati sono ricercabili attraverso un protocollo di comunicazione condiviso, che consente l'autenticazione e l'autorizzazione della ricerca, dove richiesto;
- *Interoperable*: i (meta)dati sono rappresentati e tra loro posti in relazione attraverso un vocabolario di descrizione della conoscenza formale, accessibile, condiviso e di ampia applicabilità;

* Università Cattolica del Sacro Cuore, Milano.

Il progetto *LiLa: Linking Latin* è stato finanziato dallo *European Research Council* (ERC) nell'ambito del programma di ricerca e innovazione *European Union's Horizon 2020* – Grant Agreement No. 769994.

- *Reusable*: i (meta)dati sono associati a una serie di attributi che ne descrivono la licenza d'uso e la provenienza.

L'applicazione dei principii FAIR ai (meta)dati delle risorse linguistiche comporta non solo che le loro componenti (dalle singole occorrenze di parole nei testi ai metadati che descrivono una risorsa) siano identificate univocamente e persistentemente (*Findable*), ma anche che esse possano essere interrogate sul web (*Accessible*), poste in reciproca relazione adottando ontologie che consentano l'interpretazione semantica delle relazioni stesse da parte delle macchine (*Interoperable*), e infine riutilizzate il più possibile liberamente in base a esplicite licenze (*Reusable*). La sfida, dunque, consiste nello svincolare le singole risorse linguistiche dalla loro autoreferenzialità, intendendo i (meta)dati che ciascuna risorsa raccoglie non più come parte soltanto di essa, ma come una componente di un ecosistema interconnesso e, appunto, interoperabile, in cui i (meta)dati delle risorse interagiscono, al fine di valorizzare al meglio il contributo specifico che ciascuna di esse apporta.

Disporre di risorse interoperabili è un *desideratum* fondamentale della comunità scientifica, che ormai da più di un decennio ha avviato un processo di risoluzione della diaspora delle risorse linguistiche, che si è massimamente concretizzato nella realizzazione della infrastruttura CLARIN (*Common Language Resources and Technology Infrastructure*).[1] CLARIN rappresenta un luogo dove le risorse possono essere pubblicate, trovate e, nella maggior parte dei casi, scaricate, sollevando così gli utenti dalla necessità di cercare i (meta)dati che servono loro nei *repository* dei diversi sviluppatori e distributori di risorse.

Nel corso degli ultimi anni, CLARIN ha iniziato a sviluppare metodi e applicazioni a supporto non solo della raccolta, ma anche dell'interoperabilità tra le risorse in essa pubblicate. Ad oggi, tale interoperabilità può essere realizzata a livello dei metadati descrittivi delle risorse attraverso la cosiddetta *Component MetaData Infrastructure* (CMDI) (Broeder et al. 2022), che raccoglie 'componenti', ovvero gruppi di metadati semanticamente coerenti che vengono connessi a un registro di concetti condiviso (il CLARIN *Concept Registry*) (Schuurman et al. 2016). Tuttavia, tali concetti non sono (ancora) messi in relazione con quelli di altri schemi/ontologie e, soprattutto, non arrivano a consentire una rappresentazione dei dati più granulari, siano questi lessicali (come, ad esempio, le entrate lessicali dei dizionari) o testuali (le singole parole nei testi).

Una soluzione alla sfida sollevata dall'interoperabilità granulare tra le risorse linguistiche è venuta negli ultimi anni dall'applicazione dei principii del paradigma *Linked Data* a dati linguistici da parte della comunità scientifica attiva nel campo

1. www.clarin.eu.

dei cosiddetti *Linguistic Linked Open Data* (LLOD), che ha realizzato una serie di ontologie specificamente dedicate alla rappresentazione di informazione (meta)linguistica e un *cloud* di risorse linguistiche interoperabili proprio in quanto pubblicate in modalità *Linked Data* (LOD *Cloud*[2]).

Questo articolo descrive una *Knowledge Base*, nominata LiLa ('Linking Latin'), che raccoglie e rende interoperabili risorse testuali e lessicali per la lingua latina attraverso l'uso e lo sviluppo di ontologie, ovvero rappresentazioni formali dell'informazione, e la pubblicazione dei (meta)dati secondo i principii *Linked Data* e, in particolare, LLOD. Il lavoro di sviluppo di LiLa è motivato da un'esigenza pratica e diffusa nella comunità scientifica: fare il miglior uso possibile dei dati forniti dalle risorse linguistiche. L'assunto che dà ragione alla creazione di LiLa consiste nel fatto che tale uso può essere messo in pratica attraverso la reciproca interazione tra le risorse. Infatti, decenni di lavoro di ricerca nell'ambito della linguistica empirica e di quella computazionale, oltre che di digitalizzazione del patrimonio testuale delle lingue classiche, hanno reso disponibile una notevole quantità di dati e metadati (testuali e lessicali) in forma di risorse: il passo successivo consiste, ora, nel fare in modo che questi dati e metadati possano essere prodotti, interrogati e raccolti superando i confini delle singole risorse che li includono, ovvero utilizzando queste ultime in modo interoperabile.

L'articolo è organizzato come segue. La Sezione 2 introduce i principii del paradigma *Linked Data*. La Sezione 3 descrive l'architettura della *Knowledge Base* LiLa, presentando la raccolta di lemmi latini che ne è la struttura portante, le risorse linguistiche al momento allacciate a essa e due modi di interrogarne i (meta)dati. Infine, la Sezione 4 propone alcune considerazioni conclusive sull'utilizzo delle risorse linguistiche digitali per l'avanzamento della conoscenza.

2. Il paradigma Linked Data

Introdotto da Tim Berners-Lee *et alii* (2001), il concetto di *Semantic Web* si fonda sull'assunto che i documenti pubblicati nel *World Wide Web* vengano associati a informazioni e metadati strutturati in modo tale da consentirne l'interrogazione e l'interpretazione semantica da parte non solo di esseri umani, ma anche di agenti automatizzati.

Tale strutturazione è realizzata in forma di *Linked Data*, che rappresentano le colonne portanti del *Semantic Web*, inteso come un *web of data*. Diversamente da un web fatto di ipertesti, in cui i collegamenti non sono semanticamente interpretabili, il *Semantic Web* è costituto da link tra 'oggetti' associati a un identificativo unico

2. https://lod-cloud.net.

e persistente (URI: *Uniform Resource Identifier*). I collegamenti tra gli oggetti sono semanticamente interpretabili in quanto rappresentati attraverso vocabolari di descrizione della conoscenza il più possibile condivisi registrati in forma di ontologie.

Il paradigma *Linked Data* è fondato su quattro principii definiti da Berners-Lee stesso[3]:

1. usare URI come 'nomi per le cose' (*names for things*) al fine di identificarle in modo unico e persistente. Le 'cose' con cui si ha a che fare se si trattano (meta) dati linguistici in *Linked Data* sono oggetti linguistici, come ad esempio occorrenze di parole in testi, entrate lessicali in dizionari, o insiemi di parti del discorso;
2. usare HTTP URI, per consentire alle persone (e alle macchine) di 'cercare le cose' sul web (*to look up things*);
3. usare standard come RDF e SPARQL per fornire informazione utile su quanto è identificato da una URI, ai fini della rappresentazione e ricerca dei (meta)dati. RDF (*Resource Description Framework*) (Lassila & Swick 1998) è il *data model* che sta alla base del *Semantic Web*. In base a esso, l'informazione nel *Semantic Web* è organizzata e rappresentata in termini di triple, ovvero relazioni tra un Soggetto e un Oggetto attraverso una Proprietà (ovvero, un Predicato diadico). Le classi cui appartengono i Soggetti e gli Oggetti, così come la semantica delle Proprietà sono stabilite da ontologie condivise dalle diverse comunità che arricchiscono e utilizzano il *Semantic Web*. SPARQL (SPARQL *Protocol And RDF Query Language*)[4] è un linguaggio di interrogazione per (meta)dati rappresentati in RDF;
4. includere link a altre URI, in modo da consentire alle persone (e alle macchine) di 'scoprire più cose' (*to discover more things*).

Nel 2010, Berners-Lee ha sviluppato un sistema di valutazione (a stellette) dei dati pubblicati sul web basato su cinque livelli progressivi verso lo status di *Linked Data*:

- una stelletta: i dati sono disponibili sul web (in qualsiasi formato) con una licenza aperta, in modo tale da essere *Open Data*;
- due stellette: i dati sono disponibili in un formato strutturato e *machine-readable* (e.g., in una tabella Excel);
- tre stellette: come il livello precedente, ma in un formato non proprietario (e.g., in un formato a campi fissi separati da virgole invece che in Excel);
- quattro stellette: tutto quanto stabilito dai livelli precedenti più l'adozione di standard aperti (come RDF e SPARQL) per identificare le cose e collegarne altre a esse;

3. https://www.w3.org/DesignIssues/LinkedData.
4. https://www.w3.org/TR/rdf-sparql-query/.

- cinque stellette: tutto quanto stabilito dai livelli precedenti più i collegamenti con altri dati.

Applicare i principii del paradigma *Linked Data* a (meta)dati tratti da risorse linguistiche e pubblicarli sul web al più alto livello stabilito dal *rating* di Berners-Lee comporta una serie di benefici, tra cui i seguenti (Chiarcos et al. 2013):

- Rappresentazione e Modellizzazione: RDF è un *data model* molto versatile e, quindi, adatto per rappresentare metadati come, ad esempio, quelli veicolati dai vari livelli di annotazione linguistica (morfologia, sintassi, lemmatizzazione etc.);
- Interoperabilità Strutturale (o Sintattica): consiste nell'abilità di sistemi diversi di processare dati scambiati utilizzando un *data model* comune (RDF) consistente in protocolli e formati condivisi dei dati (HTTP, URI) (Ide & Pustejovsky 2010);
- Interoperabilità Concettuale (o Semantica): è l'abilità di un sistema d'interpretare automaticamente e semanticamente l'informazione scambiata, utilizzando un insieme comune di classi e categorie dei dati definite in ontologie e vocabolari (Ide & Pustejovsky 2010);
- Federazione: l'informazione può essere combinata a partire da *repository* che sono fisicamente separati (ovvero da più server web);
- Dinamicità: dal momento che chi fornisce i (meta)dati delle risorse linguistiche pubblicate in LLOD li può gestire e mantenere localmente sul proprio server, è possibile dare accesso sempre alla versione più recente della risorsa;
- Ecosistema: esiste un'ampia e vivace comunità scientifica che adotta e sviluppa strumenti e pratiche comuni di LLOD. Tra le iniziative in corso è meritevole di menzione la COST Action *Nexus Linguarum*: *European network for Web-centred linguistic data science*.[5]

3. La *Knowledge Base* LiLa

Questa sezione descrive la Knowledge Base LiLa, che consiste in una raccolta di risorse linguistiche (sia lessicali che testuali) per la lingua latina rese interoperabili in Linked Data sul web tramite la loro rappresentazione attraverso comuni ontologie e vocabolari di descrizione della conoscenza.

3.1 L'architettura di LiLa

L'architettura di LiLa si fonda sul semplice assunto che tutto ciò che fa parte di LiLa ha a che fare con le parole. La Figura 1 mostra, nella parte bassa, le fonti dei (meta) dati che LiLa rende interoperabili. Nello specifico, esse sono:

5. https://nexuslinguarum.eu.

- le risorse lessicali, come i dizionari o i lessici, che descrivono proprietà di parole e sono costitute da entrate lessicali;
- le risorse testuali, come i corpora e le biblioteche digitali, che forniscono testi e includono occorrenze di parole in essi (*token*);
- gli strumenti di trattamento automatico del linguaggio (TAL; in inglese: *Natural Language Processing*, NLP), che processano parole e producono risultati (*NLP Output*). In particolare, l'output di uno specifico tipo di strumento di TAL (i cosidetti *tokenizer*) sono *token*, che a propria volta entrano in input ad altri strumenti di TAL, come ad esempio ai marcatori delle parti del discorso (*Part Of Speech Tagger*).

Nella Figura 1 è possibile vedere come le entrate lessicali, le occorrenze delle parole nei testi e gli output degli strumenti di TAL vengano resi interoperabili in LiLa attraverso il loro collegamento ai rispettivi lemmi, ovvero le forme convenzionali di citazione delle parole.

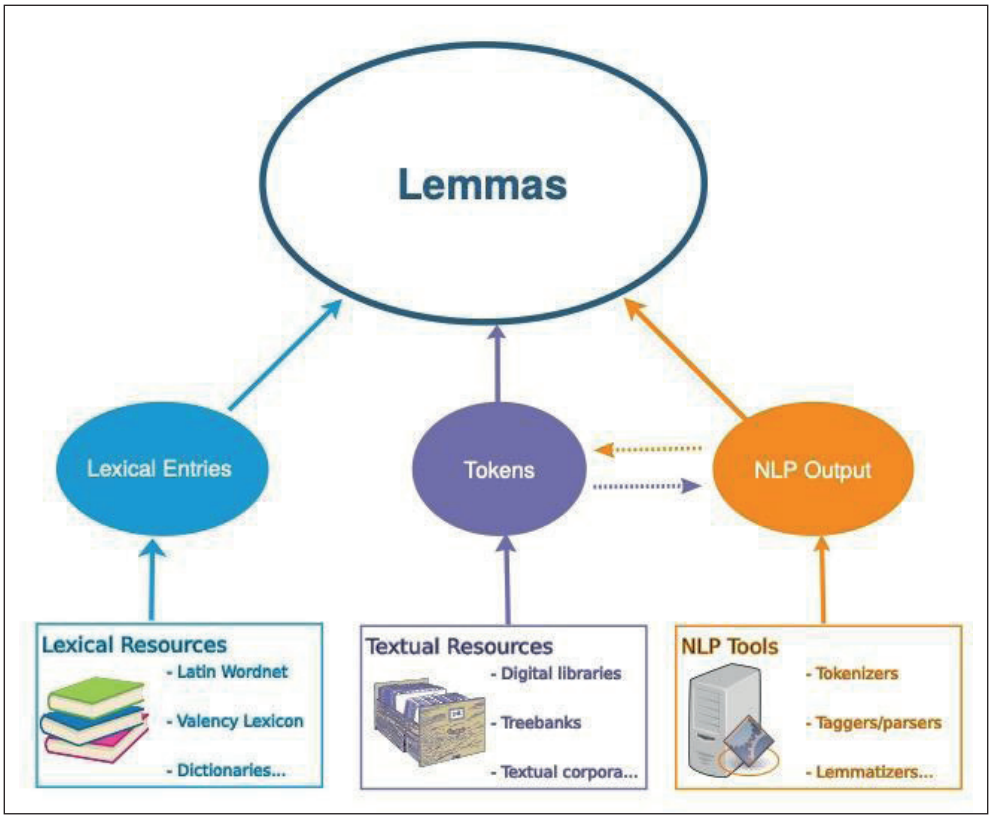

Figura 1 L'architettura di LiLa

Passando attraverso i lemmi, è dunque possibile operare ricerche distribuite sulle risorse linguistiche collegate e rese interoperabili in LiLa. Ad esempio, si possono cercare tutte le occorrenze (i *token*) del medesimo lemma in più corpora testuali; oppure si possono estrarre da più corpora tutte le occorrenze di parole che hanno certe proprietà fornite da una o più risorse lessicali.

Dato il ruolo centrale giocato dai lemmi nell'architettura di LiLa, una componente essenziale della Knowledge Base è una raccolta di forme convenzionali di citazione delle parole latine, chiamata *Lemma Bank*.

3.2 Il 'cuore' di LiLa: la Lemma Bank

La *Lemma Bank* di LiLa consiste in una raccolta di lemmi della lingua latina, ovvero forme di citazione lessicale (più o meno convenzionalmente) adottate nelle risorse linguistiche. Si tratta, cioè, dei nomi delle entrate nelle risorse lessicali e delle parole adottate per raccogliere tutte le forme di un medesimo elemento lessicale che occorrono in un testo. Come visto, la *Lemma Bank* ha un compito fondamentale nella *Knowledge Base* LiLa: attraverso essa vengono rese interoperabili le risorse linguistiche per il latino, rappresentando il punto di connessione tra le entrate delle diverse risorse lessicali e le occorrenze delle parole di quelle testuali.

Fondandosi sui principii del paradigma *Linked Data*, l'interoperabilità concettuale tra le risorse distribuite connesse in LiLa è realizzata attraverso l'applicazione di un vocabolario di descrizione della conoscenza condiviso non solo entro LiLa ma, più ampiamente, nel mondo LLOD. Nello specifico della *Lemma Bank*, ciò consiste nel ricorso all'uso del vocabolario definito da *OntoLex-Lemon* (McCrae et al. 2017), una delle ontologie più adottate nel settore per fini di rappresentazione di risorse lessicali in *Linked Data*. La Figura 2 presenta il modello di *OntoLex-Lemon*.

Nella Figura 2, le Classi di *OntoLex-Lemon* (ovvero, i 'concetti' descritti dal modello) sono graficamente rappresentate entro rettangoli. Le relazioni tra le Classi sono frecce associate al nome della Proprietà (ovvero, il predicato) che collega tra loro due Classi.

La Classe principale di *OntoLex-Lemon* è la *Lexical Entry*, intesa come l'unità di analisi del lessico che raccoglie una o più forme (*Lexical Form*) e uno o più sensi (*Lexical Sense*) e concetti (*Lexical Concept*). I *Lexical Sense* sono sensi lessicalizzati, ovvero un senso appartiene esattamente a una *Lexical Entry*. Elementi semantici che possono essere espressi da più parole sono, invece, rappresentati attraverso i *Lexical Concept*, che dunque possono avere più di una lessicalizzazione. Un tipico esempio di *Lexical Concept* sono i *synset* di una risorsa come *WordNet*, che raggruppano più parole legate tra loro da un rapporto di sinonimia concettuale (Fellbaum 2010).

Le *Lexical Form* hanno una o più varianti grafiche (*Written Representation*) ed eventualmente fonetiche (*Phonetic Representation*). Una delle *Lexical Form* è la co-

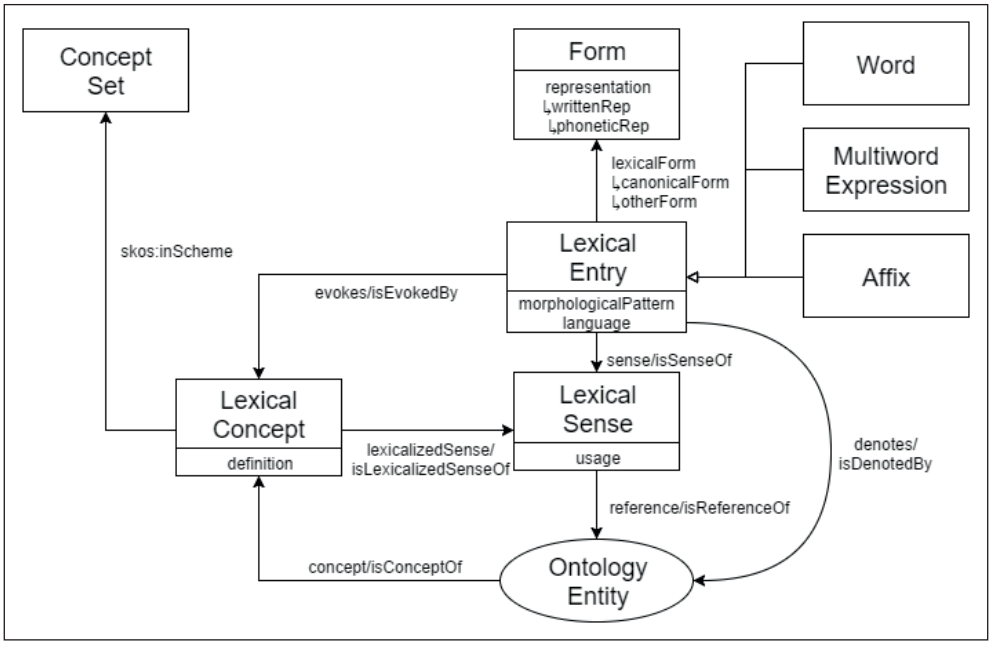

Figura 2 Il modello di *OntoLex-Lemon*

siddetta *Canonical Form*, ovvero il lemma, cioè la forma che è scelta, più o meno convenzionalmente, per rappresentare l'intero insieme delle forme flesse dell'entrata lessicale in questione. La *Lemma Bank* di LiLa è, dunque, una raccolta di *Canonical Form* di *OntoLex-Lemon*, svincolate da alcun rapporto con una *Lexical Entry*, in quanto la *Lemma Bank* non è una risorsa lessicale fatta di entrate lessicali, ma, appunto, un insieme di forme canoniche di citazione.

I lemmi che costituiscono la *Lemma Bank* sono stati tratti dalla base lessicale dell'analizzatore morfologico per il latino *LemLat* (Passarotti et al. 2017), che consiste nella collazione di tre dizionari di latino classico (Georges & Georges 1913-1918; Glare 1982; Gradenwitz 1904), nell'intero *Onomasticon* del *Lexicon Totius Latinitatis* di Forcellini (1940) (Budassi & Passarotti 2016) e nel *Glossarium Mediae et Infimae Latinitatis* di du Cange *et alii* (1883-1887), per un totale di più di 130.000 parole, corrispondenti a circa 200.000 *Canonical Form* (Cecchini et al. 2018b).

Le risorse testuali sono connesse alla *Lemma Bank* attraverso la proprietà `hasLemma` definita nell'ontologia specifica di LiLa[6], che collega un *token* in un corpus con il suo lemma nella *Lemma Bank*. Le risorse lessicali, invece, sono connesse alla Lemma Bank attraverso la proprietà di *OntoLex-Lemon* `canonicalForm`,

6. https://lila-erc.eu/ontologies/lila/.

che collega una *Lexical Entry* della risorsa in questione con il corrispondente lemma, ovvero *Canonical Form*, nella *Lemma Bank*.

3.3 Le risorse linguistiche in LiLa

Al momento, la *Knowledge Base* LiLa include e rende interoperabili le seguenti risorse linguistiche per il latino, per un totale di circa 33 milioni di triple.

Risorse testuali:

- *Index Thomisticus Treebank* (https://lila-erc.eu/data/corpora/ITTB/id/corpus) (Passarotti 2019): il più esteso corpus annotato sintatticamente disponibile per la lingua latina. Raccoglie più di 350.000 occorrenze di parola tratte da testi di Tommaso d'Aquino (tra cui l'intera *Summa contra Gentiles*). La *treebank* è disponibile in due versioni: quella annotata secondo i criteri originali della risorsa (Bamman et al. 2008) e la sua conversione nello schema *Universal Dependencies* (Cecchini et al. 2018a);
- *UDante* (https://lila-erc.eu/data/corpora/UDante/id/corpus): corpus che raccoglie le opere latine di Dante Alighieri (circa 50.000 parole) arricchite con annotazione sintattica secondo lo schema *Universal Dependencies* (Cecchini et al. 2020);
- *Querolus sive Aulularia* (https://lila-erc.eu/data/corpora/Querolus/id/citationUnit/QuerolussiveAulularia): il testo di una commedia anonima della tarda antichità latina (circa 17.000 parole) (Gamba 2020);
- *Liber Abbaci* (https://lila-erc.eu/data/corpora/CorpusFibonacci/id/corpus/Liber%20Abbaci): un trattato di aritmetica scritto nel 1202 da Leonardo Fibonacci (VIII capitolo: circa 30.000 parole) (Grotto et al. 2021);
- LASLA (https://lila-erc.eu/data/corpora/Lasla/id/corpus): un corpus che raccoglie più di 130 testi di epoca classica e tarda, per un totale di circa 1.700.000 parole (Verkerk et al. 2020).

Risorse lessicali:

- *Word Formation Latin* (https://lila-erc.eu/data/lexicalResources/WFL/Lexicon): un lessico le cui entrate (circa 30.000) sono relazionate tramite processi di formazione di parola (Litta et al. 2019);
- *Etymological Dictionary of Latin & the Other Italic Languages* (https://lila-erc.eu/data/lexicalResources/BrillEDL/Lexicon): un dizionario che include forme ricostruite protoindoeuropee e protoitaliche per spiegare la storia etimologica di circa 1.400 forme latine (De Vaan 2008; Mambrini & Passarotti 2020);
- *Latin Vallex* 2.0 e *Latin WordNet* (https://lila-erc.eu/data/lexicalResources/LatinVallex/Lexicon; http://lila-erc.eu/data/lexicalResources/LatinWordNet/Lexi-

con): una porzione corretta manualmente di Latin *Wordnet* in cui a ogni senso di una parola (corrispondente a un *synset* di *WordNet*) è associato un *frame* valenziale (Mambrini et al. 2021a);
- *Index Graecorum Vocabulorum in Linguam Latinam Translatorum* (https://lila-erc.eu/data/lexicalResources/IGVLL/Lexicon): una lista di 1.763 prestiti latini dal greco antico pubblicata nel 1874 da Günther Alexander E. A. Saalfeld (Saalfeld 1884; Franzini et al. 2020);
- *Latin Affectus* (https://lila-erc.eu/data/lexicalResources/LatinAffectus/Lexicon): un lessico che assegna un valore di polarità di *sentiment* a priori a più di 2.500 aggettivi e nomi latini (Sprugnoli et al. 2020);
- *Lewis & Short* (https://lila-erc.eu/data/lexicalResources/LewisShort/Lexicon): un dizionario bilingue latino-inglese curato da Ch. T. Lewis e Ch. Short pubblicato nel 1879 (Lewis & Short 1879; Mambrini et al. 2021b).

3.4 Ricerche su LiLa

La *Knowledge Base LiLa* può essere interrogata da due punti di partenza. Il primo modo di accedere ai contenuti di LiLa è particolarmente centrato sulla Lemma Bank. Si tratta di una semplice e intuitiva interfaccia di *query*, accessibile presso https://lila-erc.eu/query/, che consente di raccogliere gruppi omogenei di lemmi della *Lemma Bank* e consultare le triple a ciascuno di essi associate.

Tramite l'interfaccia, i lemmi della *Lemma Bank* possono essere raggruppati per parte del discorso, genere, categoria flessiva, base lessicale e affisso formativo. Le liste che si ottengono possono essere scaricate in formato CSV (*Comma Separated Values*), così come la *query* in SPARQL che l'interfaccia ha scritto a fronte della selezione delle proprietà lessicali operata dall'utente.

La Figura 3 mostra la lista dei 5 lemmi che risultano da una *query* che cerca nella *Lemma Bank* i verbi di terza coniugazione formati con il prefisso *pro-* i cui ultimi due caratteri sono *-co* (espressione regolare: `.*co$`). Nella parte destra della riga di ciascun lemma sono riportate le icone:

a. delle risorse lessicali allacciate a LiLa che includono un'entrata connessa al lemma in questione. Ad esempio, per il lemma *produco*, le tre icone corrispondono rispettivamente all'entrata di *produco* nel lessico *Word Formation Latin*, nel dizionario *Lewis & Short* e in *Latin Valle*x 2.0 + *Latin WordNet*;
b. del cosiddetto *datasheet* del lemma, ovvero una pagina in cui sono riportate le triple che hanno il lemma in questione come Soggetto, o Oggetto;
c. della visualizzazione grafica delle triple del lemma.

La Figura 4 riporta la visualizzazione grafica di alcune triple che riguardano il lemma *produco* (URI: https://lila-erc.eu/data/id/lemma/119681). Il nodo del lemma *produco*

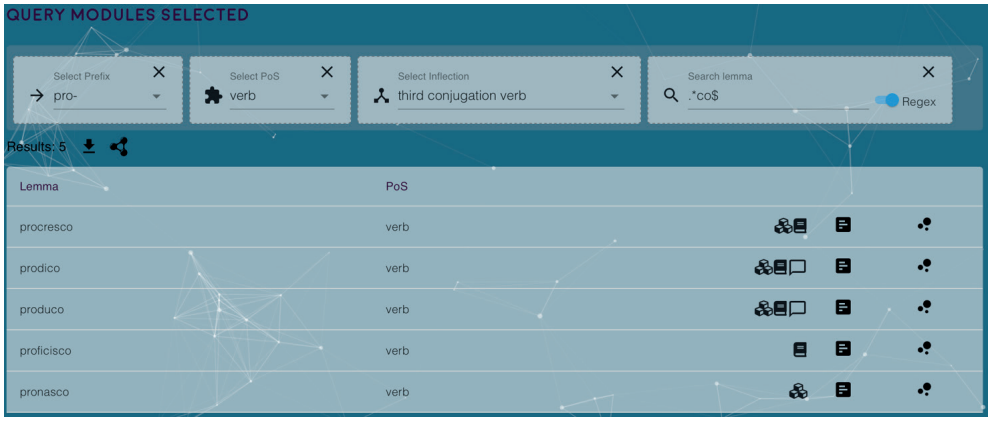

Figura 3 L'interfaccia di *query* della *Lemma Bank*

è connesso, attraverso la proprietà `canonicalForm`, al nodo dell'entrata lessicale di *produco* nel dizionario *Lewis & Short*. A propria volta, quest'ultimo è connesso, attraverso `sense`, al Lexical Sense di *produco* che è il senso lessicalizzato (`isLexicalizedSenseOf`) del *Lexical Concept* che rappresenta il contenuto della definizione di quel senso fornita dal dizionario (*to lead or bring forth, to lead forward or out*), evocata (`evokes`) dall'entrata lessicale, secondo quanto stabilito da *OntoLex-Lemon*. Attraverso la proprietà `hasLemma`, *produco* è connesso a tutti i suoi *token* nei corpora interoperabili in LiLa. Nella Figura 3 è possibile vedere il link al token *producit*, che occorre nella frase 639 del *De Monarchia* di Dante Alighieri. Infine, la Figura 3 presenta il link tra il lemma *produco* della *Lemma Bank* e il nodo che rappresenta graficamente la sua base lessicale, ovvero quella di *duco* (attraverso la proprietà `hasBase`). Tutti i lemmi della *Lemma Bank* che fanno uso della base lessicale di *duco* sono connessi a questo nodo: nella Figura 3 è possibile vedere il link tra la base di *duco* e il lemma *duco* (con variante grafica *douco*).

Il secondo modo per interrogare LiLa è il punto di accesso SPARQL alla *Knowledge Base*. Presso https://lila-erc.eu/sparql/ è possibile scrivere e lanciare *query* secondo il linguaggio SPARQL. Sono fornite alcune *query* d'esempio: tra esse, la Figura 5 riporta parte dell'output della *query* che elenca i lemmi che occorrono nei testi di Seneca (tratti dal corpus LASLA), di Dante Alighieri (da *UDante*) o di Tommaso d'Aquino (dall'*Index Thomisticus Treebank*) che hanno come base lessicale quella di *patior*. L'output consiste in una tabella formata da quattro colonne:

- la URI del lemma nella *Lemma Bank*;
- il nome del lemma;
- la URI del testo in cui il lemma occorre;
- il titolo del testo.

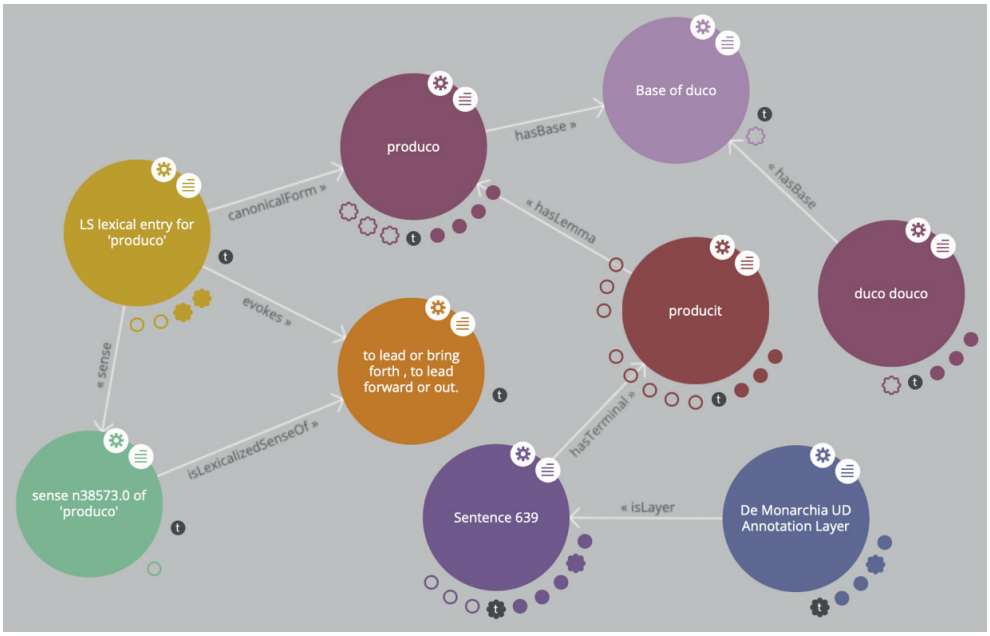

Figura 4 Visualizzazione grafica delle triple

Un esempio di scrittura di *query* in SPARQL è il seguente, che fornisce l'elenco delle risorse lessicali connesse a LiLa. Nello specifico, la *query* produce in output una tabella formata da due colonne:

- la URI della risorsa lessicale;
- il titolo della risorsa lessicale.

```
SELECT ?URI_lex ?title
WHERE {
 ?URI_lex?pred<http://www.w3.org/ns/lemon/lime#Lexicon>;
       <http://purl.org/dc/terms/title> ?title
}
order by ?title
```

La *query* si apre con l'operatore `SELECT`, che seleziona (e nomina) le colonne della tabella da produrre in output, ciascuna preceduta da un punto di domanda. Quindi, la *query* impone una serie di condizioni di selezione tra le triple di LiLa (`WHERE`). In questo caso, sono indicate due triple:

- `?URI _ lex ?pred <http://www.w3.org/ns/lemon/lime#Lexicon>`: il Soggetto, riferito con la variabile nominata `?URI _ lex` (URI della risorsa les-

Table | Response | 76 results in 0.163 seconds

sameBaseLemma	sameBaseLemmaLabel	doc	docTitle
http://lila-erc.eu/data/id/...	impatientia	http://lila-erc.eu/data/c...	Ad Lucilium Epistulae Morales
http://lila-erc.eu/data/id/...	patientia	http://lila-erc.eu/data/c...	Ad Marciam De Consolatione
http://lila-erc.eu/data/id/...	patientia	http://lila-erc.eu/data/c...	Ad Lucilium Epistulae Morales
http://lila-erc.eu/data/id/...	perpessicius	http://lila-erc.eu/data/c...	Ad Lucilium Epistulae Morales
http://lila-erc.eu/data/id/...	perpessio	http://lila-erc.eu/data/c...	Ad Lucilium Epistulae Morales
http://lila-erc.eu/data/id/...	impatiens	http://lila-erc.eu/data/c...	Agamemnon
http://lila-erc.eu/data/id/...	impatiens	http://lila-erc.eu/data/c...	De Beneficiis
http://lila-erc.eu/data/id/...	perpetior	http://lila-erc.eu/data/c...	De Constantia
http://lila-erc.eu/data/id/...	passio	http://lila-erc.eu/data/c...	De Monarchia
http://lila-erc.eu/data/id/...	patientia	http://lila-erc.eu/data/c...	Epistole
http://lila-erc.eu/data/id/...	passio	http://lila-erc.eu/data/c...	Epistole
http://lila-erc.eu/data/id/...	perpetior	http://lila-erc.eu/data/c...	Oedipus
http://lila-erc.eu/data/id/...	impassibilis	http://lila-erc.eu/data/c...	Summa contra Gentiles

Figura 5 Output di una *query* SPARQL su LiLa

sicale), è connesso attraverso una proprietà non specificata e identificata dalla variabile `?pred` a un Oggetto che è un lessico, ovvero un individuo della classe *Lexicon* del vocabolario per esprimere metadati linguistici LIME (*LInguistic MEtadata*) (Fiorelli et al. 2015). La classe *Lexicon* ha URI http://www.w3.org/ns/lemon/lime#Lexicon;

- `<http://purl.org/dc/terms/title> ?title`: il Soggetto `?URI _ lex` (non riportato nuovamente per economia di scrittura della *query*) è altresì connesso attraverso la proprietà `title` (Dublin Core: https://www.dublincore.org/specifications/dublin-core/dcmi-terms/#title) a un Oggetto identificato dalla variabile nominata `?title` (titolo della risorsa lessicale).

Infine, l'output è elencato alfabeticamente per titolo della risorsa lessicale (`order by ?title`).

4. Conclusioni

Affrontando la sfida di rendere interoperabili le risorse linguistiche per il latino tramite l'applicazione dei principii di un paradigma adottato nel *Semantic Web* per rappresentare (meta)dati di ogni tipo, la *Knowledge Base* LiLa mira a contribuire alla scalata della cosiddetta *wisdom hierarchy* lungo le pareti della piramide DIKW (Rowley 2007). Questa si riferisce a una classe di modelli per la rappresentazione delle relazioni strutturali (sintattiche) e/o concettuali (semantiche) tra dati (*Data*), informazione (*Information*), conoscenza (*Knowledge*) e saggezza (*Wisdom*). I quattro liv-

elli della gerarchia sono strettamente interrelati tra loro; infatti, "typically information is defined in terms of data, knowledge in terms of information, and wisdom in terms of knowledge" (Rowley 2007, p. 177).

Risalire la *wisdom hierarchy* rappresenta uno degli obiettivi di maggiore impatto dell'utilizzo delle risorse linguistiche digitali nella ricerca umanistica. Infatti, la trasposizione su supporto digitale di lessici e dizionari, così come di raccolte di testi più o meno annotati, consente e impone uno scarto metodologico nelle scienze umanistiche che non è solo quantitativo e operativo, ma anche e soprattutto qualitativo. Le parole in proposito di uno dei pionieri del trattamento computazionale dei dati linguistici, padre Roberto Busa, sono cristalline (Busa 1980, p. 89):

> In this field one should not use the computer primarily for speeding up the operation, nor for minimizing the work of the researchers. It would not be reasonable to use the computer just to obtain the same results as before, having the same qualities as before, but more rapidly and with less human effort. [...] To repeat: the use of computers in the humanities has as its principal aim the enhancement of the quality, depth and extension of research and not merely the lessening of human effort and time.

Qualità, profondità ed estensione della ricerca sono oggi consentite non solo dall'ampia disponibilità di (meta)dati linguistici in forma di risorse, ma anche e soprattutto dalla possibilità di analizzarli computazionalmente nella loro interezza e di farli interagire utilizzando standard di rappresentazione della conoscenza ampiamente condivisi anche oltre i confini della comunità scientifica.

Risalire la piramide della conoscenza è l'obiettivo di ogni ricerca: nel campo umanistico, e specificamente in quello linguistico, assistiamo a una svolta empirica (ed empirista) che si impone nella propria ineludibilità. La Knowledge Base LiLa vuole porsi al servizio di questa svolta, mirando a diventare il luogo di pubblicazione di ogni risorsa linguistica digitale per il latino e, più ambiziosamente, un modello a supporto dell'interoperabilità tra i (meta)dati forniti dalle risorse di ogni lingua.

Bibliografia

Bamman D., Passarotti M., Busa R., Crane G. 2008, The annotation guidelines of the Latin Dependency Treebank and Index Thomisticus Treebank: the treatment of some specific syntactic constructions in Latin, in *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC 2008)*, Marrakech, Morocco, European Language Resources Association (ELRA), pp. 71-76.

Berners-Lee T., Hendler J., Lassila O. 2011, The Semantic Web, *Scientific American* 284 (n. 5), pp. 34-43.

Broeder D., Windhouwer M., Van Uytvanck D., Goosen T., Trippel T. 2022, CMDI: a component metadata infrastructure, in *Proceedings of the workshop "Describing language resources with metadata: towards flexibility and interoperability in the documentation of language resources". LREC 2012*, Istanbul, Turkey, European Language Resources Association (ELRA), pp. 1-4.

Budassi M. & Passarotti M. 2016, Nomen Omen. Enhancing the Latin Morphological Analyser Lemlat with an Onomasticon, in N. Reiter, B. Alex, K.A. Zervanou (ed.), *Proceedings of the 10th SIGHUM Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH 2016)*, Berlin, Germany, The Association for Computational Linguistics, pp. 90-94.

Busa R. 1980, The annals of humanities computing: The Index Thomisticus, *Computers and the Humanities* 14 (n. 2), pp. 83-90.

Cecchini F.M., Passarotti M., Marongiu P., Zeman D. 2018a, Challenges in Converting the Index Thomisticus Treebank into Universal Dependencies, in M.C. De Marneffe, T. Lynn, S. Schuster (ed.), *Proceedings of the Second Workshop on Universal Dependencies (UDW 2018)*, Bruxelles, Belgium, The Association for Computational Linguistics, pp. 27-36.

Cecchini F.M., Passarotti M., Testori M., Ruffolo P., Draetta L., Fieromonte M., Liano A., Marini C., Piantanida G. 2018b, Enhancing the Latin Morphological Analyser LEMLAT with a Medieval Latin Glossary, in E. Cabrio, A. Mazzei, F. Tamburini (ed.), *Proceedings of the Fifth Italian Conference on Computational Linguistics*, Torino, Italy, Accademia university press, pp. 87-92.

Cecchini F.M., Sprugnoli R., Moretti G., Passarotti M. 2020, UDante: First Steps Towards the Universal Dependencies Treebank of Dante's Latin Works, in J. Monti, F. Dell'Orletta, F. Tamburini (ed.), *Proceedings of the Seventh Italian Conference on Computational Linguistics*, Bologna, Italy, CEUR Workshop Proceedings, pp. 1-7.

Chiarcos C., Hellmann S., Nordhoff S. 2012, Introduction and overview, in C. Chiarcos, S. Hellmann, S. Nordhoff (ed.), *Linked Data in Linguistics. Representing and Connecting Language Data and Language Metadata*, Berlin Heidelberg, Springer-Verlag, pp. 1-12.

Chiarcos C., Moran S., Mendes P.N., Nordhoff S., Littauer R. 2013, Building a Linked Open Data Cloud of Linguistic Resources: Motivations and Developments, in I. Gurevych, J. Kim (ed.), *The People's Web Meets NLP*, Berlin Heidelberg, Springer-Verlag, pp. 315-348.

De Vaan M. 2008, *Etymological dictionary of Latin and the other Italic languages*, Leiden-Boston, Brill.

du Cange C., Bénédictins de Saint-Maur, Carpentier P., Henschel L., Favre L. 1883-1887, *Glossarium Mediae et Infimae Latinitatis*, Niort, L. Favre.

Fellbaum C. 2010, WordNet, in R. Poli, M. Healy, A. Kameas (ed.), *Theory and Applications of Ontology: Computer Applications*, Dordrecht, Springer, pp. 231-243.

Fiorelli M., Stellato A., McCrae J.P., Cimiano P., Pazienza M.T. 2015, LIME: the metadata module for OntoLex, in F. Gandon, M. Sabou, H. Sack, C. d'Amato, P. Cudré-Mauroux, A. Zimmermann (ed.), *The Semantic Web. Latest Advances and New Domains*, Cham, Springer, pp. 321-336.

Forcellini E. 1940, *Lexicon Totius Latinitatis / ad Aeg. Forcellini lucubratum, dein a Jos. Furlanetto emendatum et auctum; nunc demum Fr. Corradini et Jos. Perin curantibus emendatius et auctius meloremque in formam redactum adjecto altera quasi parte Onomastico totius latinitatis opera et studio ejusdem Jos. Perin*, Padova, Typis Seminarii.

Franzini G., Zampedri F., Passarotti M., Mambrini F., Moretti G. 2020, Græcissare: Ancient Greek Loanwords in the LiLa Knowledge Base of Linguistic Resources for Latin, in J. Monti, F. Dell'Orletta, F. Tamburini (ed.), *Proceedings of the Seventh Italian Conference on Computational Linguistics*, Bologna, Italy, CEUR Workshop Proceedings, pp. 1-6.

Gamba F. 2020, *Including a New Textual Resource into the LiLa Knowledge Base. Lemmatization, PoS Tagging and Linking of* Querolus, Pavia, Università di Pavia [tesi di laurea magistrale non pubblicata].

Georges K.E. & Georges H. 1913-1918, *Ausführliches Lateinisch-Deutsches Handwörterbuch*, Hannover, Hahn.

Glare P.G.W. 1982, *Oxford Latin Dictionary*, Oxford, Oxford University Press.

Gradenwitz O. 1904, *Laterculi Vocum Latinarum*, Leipzig, Hirzel.

Grotto F., Sprugnoli R., Fantoli M., Simi M., Cecchini F.M., Passarotti M. 2021, The Annotation of Liber Abbaci, a Domain-Specific Latin Resource, in E. Fersini, M. Passarotti, V. Patti (ed.), *Proceedings of the Eighth Italian Conference on Computational Linguistics*, Milan, Italy, CEUR Workshop Proceedings, pp. 1-8.

Ide N. & Pustejovsky J. 2010, What does interoperability mean, anyway? toward an operational definition of interoperability for language technology, in *Proceedings of the Second International Conference on Global Interoperability for Language Resources*, Hong Kong, China.

Lassila O. & Swick R.R. 1998, World Wide and Web Consortium, *Resource description framework (RDF) model and syntax specification*. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.44.6030

Lewis C.T. & Short C. 1879, *A Latin Dictionary. Founded on Andrews' edition of Freund's Latin dictionary*, Oxford, Clarendon Press.

Litta E., Passarotti M., Mambrini F. 2019, The treatment of word formation in the LiLa knowledge base of linguistic resources for Latin, in Z. Žabokrtský, M. Ševčíková, E. Litta, M. Passarotti (ed.), *Proceedings of the Second International Workshop on Resources and Tools for Derivational Morphology (DeriMo 2019)*, Prague, Czech Republic, Charles University, Faculty of Mathematics and Physics, Institute of Formal and Applied Linguistics, pp. 35-43.

Mambrini F. & Passarotti M. 2020, Representing etymology in the LiLa knowledge base of linguistic resources for Latin, in I. Kernerman, S. Krek, J.P. McCrae, J. Gracia, S. Ahmadi, B. Kabashi (ed.), *Proceedings of the Globalex Workshop on Linked Lexicography. LREC 2020 Workshop*, Paris, France, European Language Resources Association (ELRA), pp. 20-28.

Mambrini F., Passarotti M., Litta E., Moretti G. 2021a, Interlinking Valency Frames and WordNet Synsets in the LiLa Knowledge Base of Linguistic Resources for Latin, in M.

Alam, P. Groth, V. de Boer, T. Pellegrini, H.J. Pandit, E. Montiel, V. Rodríguez Doncel, B. McGillivray, A. Meroño-Peñuela (ed.), *Further with Knowledge Graphs. Proceedings of the 17th International Conference on Semantic Systems*, Series: Studies on the Semantic Web – Volume 53, Amsterdam, The Netherlands, IOS Press, pp. 16-28.

Mambrini F., Litta E., Passarotti M., Ruffolo P. 2021b, Linking the Lewis & Short Dictionary to the LiLa Knowledge Base of Interoperable Linguistic Resources for Latin, in E. Fersini, M. Passarotti, V. Patti (ed.), *Proceedings of the Eighth Italian Conference on Computational Linguistics (CLiC-it 2021)*, Milan, Italy, CEUR Workshop Proceedings, pp. 1-7.

McCrae J.P., Bosque-Gil J., Gracia J., Buitelaar P., Cimiano P. 2017, The Ontolex-Lemon model: development and applications, in I. Kosem, C. Tiberius, M. Jakubíček, J. Kallas, S. Krek, V. Baisa (ed.), *Proceedings of eLex 2017 conference*, Brno, Czech Republic, Lexical Computing, pp. 19-21.

Passarotti M. 2019, The Project of the Index Thomisticus Treebank, in M. Berti (ed.), *Digital Classical Philology. Ancient Greek and Latin in the Digital Revolution*, Berlin-Boston, De Gruyter, pp. 299-319.

Passarotti M., Budassi M., Litta E., Ruffolo P. 2017, The Lemlat 3.0 Package for Morphological Analysis of Latin, in G. Bouma, Y. Adesam (ed.), *Proceedings of the NoDaLiDa 2017 Workshop on Processing Historical Language*, Gothenburg, Linköping University Electronic Press, pp. 24-31.

Rowley J. 2007, The wisdom hierarchy: representations of the DIKW hierarchy, *Journal of Information and Communication Science* 33 (n. 2), pp. 163-180.

Saalfeld G.A.E.A. 1884, *Tensaurus Italograecus: Ausführliches historisch-kritisches Wörterbuch der griechischen Lehn- und Fremdwörter im Lateinischen*, Wien, Austria, Carl Gerold's Sohn.

Schuurman I., Windhouwer M., Ohren O., Zeman D. 2016, CLARIN concept registry: the new semantic registry, in *Selected Papers from the CLARIN Annual Conference 2015, October 14–16, 2015*, Wroclaw, Poland, Linköping University Electronic Press, pp. 62-70.

Sprugnoli R., Passarotti M., Corbetta D., Peverelli A. 2020, Odi et Amo. Creating, Evaluating and Extending Sentiment Lexicons for Latin, in N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, S. Piperidis (ed.), *Proceedings of the Twelfth International Conference on Language Resources and Evaluation (LREC 2020)*, Paris, France, European Language Resources Association (ELRA), pp. 3078-3086.

Verkerk P., Ouvrard Y., Fantoli M., Longrée D. 2020, LASLA and Collatinus: a convergence in lexica, *Studi e Saggi Linguistici* 58 (n. 1), pp. 95-120.

Wilkinson M.D. et al. 2016, The FAIR Guiding Principles for scientific data management and stewardship, *Scientific data* 3 (n. 1), pp. 1-9.