«Tema» – Referato
https://doi.org/10.6092/issn.1721-4777/12279

«Griseldaonline» 20, 2 | 2021
ISSN 1721-4777

I progetti digitali dell'OVI

Paolo Squillacioti

Pubblicato: 15 dicembre 2021

Abstract

The paper offers an overview of the digital projects in progress at the CNR institute Opera del Vocabolario Italiano in Florence, specialized in historical lexicography and in the development of software for lexicography, and now active in the European infrastructures for humanistic research.

L'intervento propone una panoramica dei progetti digitali in corso all'Istituto del CNR Opera del Vocabolario Italiano di Firenze, specializzato in lessicografia storica e nell'elaborazione di *software* per la lessicografia, e ora attivo nell'ambito delle infrastrutture europee per la ricerca umanistica.

Parole chiave: lessicografia storica; filologia digitale; infrastrutture digitali per la ricerca.

Paolo Squillacioti: Istituto CNR Opera del Vocabolario Italiano (dizione istituzionale ufficiale: vd. https://www.cnr.it//people/paolo.squillacioti)
paolo.squillacioti@cnr.it

È dirigente di ricerca del Consiglio Nazionale delle Ricerche presso l'Istituto Opera del Vocabolario Italiano (OVI), di cui è direttore dal luglio 2020. Si è formato all'Università di Pisa, conseguendo poi il Perfezionamento in Discipline filologiche e linguistiche moderne alla Scuola Normale Superiore. I principali ambiti di ricerca sono: 1) lessicografia storica dell'italiano, con particolare riguardo alla fase medievale (risultati: cura e redazione del *Tesoro della Lingua Italiana delle Origini* e pubblicazione di articoli di argomento lessicografico); 2) filologia e linguistica romanza, con particolare riguardo all'area occitanica e antico francese e alla produzione lirica e didattica (risultati: edizione critica delle poesie del trovatore Folquet de Marselha e di varie poesie trobadoriche; l'edizione critica e traduzione del primo libro del *Tresor* di Brunetto Latini; studio della tradizione manoscritta del *Tresor* e del suo volgarizzamento toscano); 3) letteratura italiana medievale, con particolare riguardo all'influenza della letteratura cortese galloromanza (risultati: articoli su Dante Alighieri, Giovanni Boccaccio e Francesco Petrarca); 4) letteratura italiana contemporanea, in particolare l'opera Leonardo Sciascia e Giuseppe Tomasi di Lampedusa (risultati: articoli su periodici; edizione critica delle opere complete di Sciascia).

1. *Introduzione*

L'Opera del Vocabolario Italiano (OVI) è un istituto del Consiglio Nazionale delle Ricerche con la missione specifica di dotare l'Italia di un vocabolario storico della lingua nazionale dalle origini al presente, filologicamente affidabile e realizzato con metodi innovativi.

Nato in seno all'Accademia della Crusca nel 1965 con finanziamenti del CNR, l'OVI è diventato una struttura del CNR solo vent'anni dopo; l'attività che lo caratterizzava era stata delineata ancora prima, nel secondo dopoguerra, quando la punta avanzata della lessicografia nazionale era costituita dalla quinta impressione del Vocabolario della Crusca, interrotto nel 1923 alla fine della lettera *O*. L'iniziativa della casa editrice UTET di elaborare un *Grande dizionario della lingua italiana*, affidato al filologo romanzo Salvatore Battaglia, ebbe un ruolo essenziale nella scelta di concentrare il lavoro sulla prima *tranche* cronologica, fino al 1375, data di morte di Giovanni Boccaccio e limite simbolico, oltre che pragmatico.

Da quel momento l'OVI si è specializzato su ricerche sulla lingua e la testualità in volgare (letteraria e non) del medioevo italiano (e romanzo), e accanto al vocabolario, il *Tesoro della Lingua Italiana delle Origini* (TLIO), concretamente realizzato dal 1996 sotto la guida più che ventennale del filologo romanzo Pietro Beltrami, l'OVI ha definito gradualmente, e ora in modo compiuto, la sua missione.

Oggi offre varie risorse per l'italianistica, in particolare per la medievistica, grazie allo sviluppo del *software* proprietario GATTO (scaricabile gratuitamente dal sito dell'OVI), che consente di gestire banche dati testuali su PC o in rete, grazie all'interfaccia GattoWeb, e ad altri strumenti informatici elaborati negli ultimi anni.

Prima di passarli in rassegna vorrei fare due brevi premesse. Invece di una rassegna ampia ma inevitabilmente celere e superficiale avrei potuto fare una più approfondita illustrazione di una sola delle risorse; l'articolazione variegata che l'istituto ha assunto negli ultimi anni, grazie all'attività del precedente direttore, oggi docente alla Scuola Normale Superiore, Lino Leonardi (anche lui filologo romanzo), consiglia piuttosto una panoramica quasi esaustiva.

In secondo luogo, ho già usato alcuni acronimi, probabilmente già sentiti, che però talora vengono usati a sproposito. È vero che una delle regole dello *show business*, ormai estesa a vaste aree della società, recita «bene o male purché se ne parli», tuttavia la vocazione filologica dell'istituto mi impone una precisazione terminologica:

- non si può dire 'fare ricerche nell'OVI' (a meno che non si venga a cercare qualcosa nella Villa di Castello dove l'OVI ha la sua sede), ma semmai 'fare ricerche nel corpus OVI' (se si interroga la banca dati testuale su cui è fondato il vocabolario) o 'fare ricerche nel TLIO' (se si consulta il vocabolario stesso);
- il TLIO acronimo di *Tesoro della Lingua Italiana delle Origini* è appunto il vocabolario che l'OVI realizza interpretando gli esempi del *Corpus TLIO per il vocabolario*;
- i corpora testuali dell'OVI legati al TLIO sono due: quello appena nominato, il *Corpus TLIO per il vocabolario*, utilizzato dai redattori del vocabolario perché le 23 milioni di occorrenze sono lemmatizzate, ossia hanno avuto un'etichettatura linguistica funzionale alla stesura delle voci del TLIO; e il *Corpus OVI dell'italiano antico*, che ingloba il *Corpus TLIO* e

ha vari testi in più (circa sette milioni di occorrenze, che porta a quasi 30 milioni il totale) ed è in crescita, grazie al sostegno delle risorse di due progetti PRIN, finanziati nel 2015 (CoVo. *Corpus per il Vocabolario*) e nel 2017 (RENOVO. *Rigenerare il corpus OVI: rinnovo e ottimizzazione di metodi, contenuti, strumenti*), entrambi coordinati da Leonardi.

È al *Corpus OVI dell'italiano antico* che conviene indirizzarsi per le ricerche non strettamente lessicografiche, per esempio per studi di impianto letterario, linguistico o filologico.

Il progetto di punta dell'OVI è il TLIO, vocabolario che interpreterà il patrimonio lessicale dalle Origini alla fine del Trecento, ovvero circa 57.300 voci, di cui oltre 46.000 già redatte e circa 40.500 diffuse ad oggi online.

Queste sono risorse bene note e utilizzate, su cui non sarebbe inutile soffermarsi per illustrare funzioni di ricerca forse non pienamente divulgate, ma è bene avviarsi alla rassegna delle risorse che organizzerei in quattro categorie, distinguendo i progetti sviluppati direttamente all'OVI e quelli a cui l'OVI ha partecipato con vari gradi di coinvolgimento; e fra le risorse quelle già disponibili online e quelle ancora a livello di prototipo. Le categorie rispecchiano i settori di attività di un istituto piccolo rispetto alla media degli istituti del CNR, ma che vuol essere dinamico e produttivo.

2. *Risorse connesse al settore corpus in GattoWeb*

È il settore più produttivo, il primo che ha consentito l'allargamento dell'attività oltre la redazione del TLIO (che rimane, sia detto una volta per tutte, l'attività principale e quella intorno alla quale ruota gran parte delle altre attività), e il settore per cui l'OVI viene più spesso coinvolto in progetti e iniziative di ricerca. Tale attività si giova del *software* proprietario GATTO, realizzato all'OVI da Domenico Iorio-Fili e portato avanti da Andrea Boccellari.

3. *Progetti OVI*

Il primo corpus di testi medievali italoromanzi in GattoWeb che si è affiancato al corpus allestito per la redazione del TLIO, è stato il *Corpus Datini*, che a partire dal 2006 ha raccolto la gran parte delle lettere edite conservate all'Archivio di Stato di Prato nell'ampio fondo contenente i documenti relativi all'attività commerciale e finanziaria di Francesco di Marco Datini, celebre mercante pratese vissuto a cavallo fra il XIV e XV secolo. Il progetto, diretto all'OVI da Pär Larson, è stato commissionato dall'Archivio, e dato avvio all'attività 'conto terzi' dell'OVI.

È stato invece progettato e realizzato all'OVI il corpus DiVo (*Dizionario dei volgarizzamenti*), corpus di volgarizzamenti italoromanzi di testi latini classici e tardoantichi (il limite stabilito è l'opera di Severino Boezio), completato dal corpus reciproco CLaVo (*Corpus dei classici latini volgarizzati*) dei testi latini fonti dei volgarizzamenti; una funzionalità di GattoWeb consente il collegamento fra i due corpora. Si è avvalso di un finanziamento FIRB-Futuro in ricerca 2010, grazie a un progetto coordinato da Elisa Guadagnini all'OVI e da Giulio Vaccaro alla Scuola Normale Superiore, cui hanno partecipato attivamente gli altri due direttori della coppia di corpora, Cosimo Burgassi e Diego Dotto, e vari altri collaboratori.

https://doi.org/10.6092/issn.1721-4777/12279

Altro progetto promosso in istituto da Elena Artale e da Ilaria Zamuner (docente all'Università Gabriele d'Annunzio di Chieti e Pescara, ma associata da anni all'OVI) è ReMediA (*Repertorio di Medicina Antica*), un corpus plurilingue che include testi medici e farmaceutici medievali.

4. *Progetti in partnership*

Nato dall'attività delle unità di ricerca dell'OVI e dell'Università degli Studi di Siena nell'ambito del progetto PRIN 2008 L'affettività romanza: lemmi e temi coordinato da Roberto Antonelli, il corpus LirIO (*Lirica Italiana delle Origini*), annovera un'ampia raccolta (esaustiva per il XIII secolo) di testi in versi (non solo appartenenti al genere lirico) nei volgari italoromanzi databili entro il 1400.

Legato alla produzione in versi del XIII secolo e all'attività di Leonardi è la versione in GattoWeb delle *Concordanze della Lingua Poetica Italiana delle Origini* (CLPIO) allestite da d'Arco Silvio Avalle, al momento interrogabile solo per forme, ma di cui è stato avviato, con l'apporto determinante di Andrea Boccellari, il recupero dell'articolata lemmatizzazione.

Si segnala infine la versione in GattoWeb di una parte significativa dei risultati del corpus ChVA (*Chartae vulgares antiquiores*), frutto del lavoro dell'*équipe* diretta da Vittorio Formentin dell'Università di Udine, che annovera anche Nello Bertoletti (Università di Trento) e Antonio Ciaralli (Università di Perugia), che sta producendo importanti ricerche nell'ambito dei testi italoromanzi più antichi.

Un settore particolarmente produttivo dell'attività cui collabora l'OVI è quello delle ricerche sul lessico di aree linguistiche specifiche, alle quali l'OVI ha offerto e continua a offrire supporto informatico, realizzando corpora in GattoWeb, sviluppando software lessicografico, allestendo portali e siti web, ospitando le risorse sui suoi *server*.

Il rapporto più antico è con il gruppo diretto da Mario Pagano all'Università di Catania, che sta allestendo un *Vocabolario del Siciliano Medievale* (VSM)[1] fondato sul corpus in GattoWeb ARTESIA[2] alla cui direzione sono affiancati Salvatore Arcidiacono e Ferdinando Raffaele, la più ampia raccolta di testi medievali limitati a un'area linguistica 'regionale' (mi si passi l'anacronismo) disponibile per il settore italoromanzo. L'OVI collabora fattivamente anche alla redazione del VSM, grazie all'attività di Rossella Mosti.

L'ingresso più recente fra le risorse allestite all'OVI è il corpus VEV (*Vocabolario Storico-Etimologico del Veneziano*), diretto da Lorenzo Tomasin all'Università di Losanna e Luca D'Onghia alla Scuola Normale Superiore, che si propone di integrare in un vocabolario digitale (comunque collegato con una diffusione a stampa)[3] i principali vocabolari del veneziano con un corpus di testi, in cui nucleo medievale è costituito proprio dal corpus VEV.

[1] M. Pagano, *Il Vocabolario del Siciliano Medievale (VSM) e il TLIO*, in L. Leonardi, P. Squillacioti (a cura di), *Italiano antico, italiano plurale. Testi e lessico del Medioevo nel mondo digitale*, Atti del convegno internazionale in occasione delle 40.000 voci del TLIO, Alessandria, Edizioni dell'Orso, 2020, pp. 191-205.

[2] M. Pagano, M. Spampinato (a cura di), ARTESIA *(Archivio Testuale del Siciliano Antico)*, «Zeitschrift für romanische Philologie», CXXX, 2014, 4, pp. 1222–1231.

[3] L. Tomasin, L. D'Onghia (a cura di), *Parole veneziane. 1. Una centuria di voci del Vocabolario Storico-Etimologico del Veneziano (VEV)*, coordinamento redazionale di F. Panontin e G. Verzi, Venezia, Lineadacqua, 2020.

https://doi.org/10.6092/issn.1721-4777/12279

All'esterno dell'area italoromanza, ma connesso per ragioni culturali, storiche e geografiche, è il corpus ATLiSOr (*Archivio Testuale della Lingua Sarda delle Origini*), il primo corpus digitale di testi sardi medievali, ideato e diretto da Giovanni Lupinu dell'Università di Sassari.[4]

Sono inoltre in allestimento altri corpora in GattoWeb, come risultati di progetti in corso o appena conclusi: il corpus CAO (*Corpus dell'antico occitano*), sdoppiato in due corpora, delle edizioni diplomatiche e delle edizioni interpretative di vari canzonieri trobadorici e di prose occitaniche inedite, risultato di un progetto PRIN 2015 coordinato da Maria Careri dell'Università Gabriele d'Annunzio di Chieti e Pescara; il corpus ArsNova, con le edizioni di testi per musica dell'Ars Nova in volgari italo- e galloromanzi realizzate nell'ambito del progetto ERC *European ars nova* coordinato da Maria Sofia Lannutti dell'Università di Firenze; il corpus *Leys d'amor*, con l'edizione dell'importante trattato metrico-stilistico edito recentemente da Beatrice Fedi, connesso agli altri risultati del progetto ERC *MiMus* (*Ioculator seu mimus. Performing Music and Poetry in medieval Iberia*), coordinato da Anna Alberni.[5]

Grazie all'attività di Salvatore Arcidiacono, che ha sviluppato il *framework* LexiCad, pensato per il *Vocabolario del Siciliano Medievale* ma adattabile a varie applicazioni web, l'OVI ha realizzato alcune piattaforme lessicografiche che condividono l'architettura profonda pur avendo oggetti, obiettivi e aspetti piuttosto diversi: oltre alla piattaforma lessicografica del VEV, il *Vocabolario Dantesco*, progetto promosso dall'Accademia della Crusca e coordinato da Paola Manni (Università di Firenze per l'Accademia della Crusca) e Lino Leonardi, a cui l'OVI collabora a vari livelli, non solo per quanto riguarda la parte informatica;[6] il *Vocabolario Dantesco Latino*, coordinato da Gabriella Albanese dell'Università di Pisa, insieme con Paolo Chiesa (Università di Milano) e Mirko Tavoni (Università di Pisa), che annovera un'ampia partnership e si propone di realizzare il primo vocabolario dedicato alle opere latine dell'Alighieri;[7] AGLIO (*Atlante Grammaticale della Lingua Italiana delle Origini*), progetto ideato da Marcello Barbato dell'Università di Napoli L'Orientale, che si propone di fornire alla comunità scientifica uno strumento specifico e innovativo per l'analisi fono-morfologica delle varietà italoromanze.[8]

Anche il TLIO si gioverà del sistema LexiCad, e contiamo di offrire presto una nuova interfaccia del TLIO con un incremento significativo delle funzioni e delle possibilità di ricerca; l'obiettivo finale è lo sviluppo di Pluto, ossia la *Piattaforma lessicografica unica del tesoro delle origini*, di cui è stato realizzato il modulo bibliografico (Pluto-BTV), essenziale per la gestione dell'esemplificazione inclusa nelle voci, che in vocabolario storico non è meno importante della

[4] G. Lupinu *et al.*, *ATLiSOr: un nuovo strumento per la ricerca linguistica e filologica sul sardo medievale*, in S. Retali-Medori (a cura di), *Actes du colloque de lexicographie dialectale et étymologique en l'honneur de Francesco Domenico Falcucci* (Corte-Rogliano, 28-30 ottobre 2015), Alessandria, Edizioni dell'Orso, 2018, pp. 467-485.

[5] A. Alberni, *Ioculator seu Mimus. Performing Music and Poetry in Medieval Iberia*, «Journal of Transcultural Medieval Studies», V, 2018, 2, pp. 435-441.

[6] R. Mosti, Z. Verlato, *Le corrispondenze del VD: TLIO, lessicografia storica, corpora dell'Ovi*, in P. Manni (a cura di), *S'i' ho ben la parola tua intesa*, Atti della giornata di presentazione del Vocabolario Dantesco (Firenze, 1 ottobre 2018), Firenze, Quaderni degli Studi di Lessicografia Italiana, 2020, pp. 93-121.

[7] Il progetto *Vocabolario Dantesco Italiano*, annunciato da Albanese nella Tornata della Crusca summenzionata (v. *Per il Vocabolario Latino di Dante*, cit., pp. 169-185), è ormai entrato nella fase operativa.

[8] M. Barbato, *Per un atlante grammaticale della lingua italiana delle origini*, «Zeitschrift für romanische Philologie», CXXXIII, 2017, pp. 820-843; Id., *L'Atlante grammaticale della lingua italiana delle origini (AGLIO)*, «Bollettino del Centro di studi filologici e linguistici siciliani», XXX, 2019, pp. 109-123.

https://doi.org/10.6092/issn.1721-4777/12279

rete semantica. Stiamo lavorando alla piena realizzazione del sistema Pluto, ma almeno il secondo step (Pluto interfaccia) è prossimo alla realizzazione.[9]

5. *Risorse connesse al settore filologia digitale*

La flessibilità del sistema LexiCad, ha consentito all'OVI di mettere a punto un sistema di filologia digitale pensato per il progetto RDP (*Le 'rime disperse' di Francesco Petrarca: l'altra faccia del Canzoniere*), coordinato da Roberto Leporatti all'Università di Ginevra.[10]

Questo aspetto dell'attività sarà sviluppato nella URT (Unità di Ricerca) che l'OVI ha attivato nella primavera 2020 a Pisa presso la Scuola Normale Superiore; la struttura, presso cui lavorano Elena Artale e Sara Ravani, è pienamente operativa dal 15 settembre di quest'anno per le limitazioni conseguenti dall'emergenza sanitaria. Rinvio per ulteriori indicazioni alla relazione di Lino Leonardi, responsabile scientifico della URT.

6. *Risorse connesse al settore infrastrutture digitali (europee)*

Per concludere non può mancare un accenno alle linee di ricerca più innovative che si stanno portando avanti all'OVI, e che ancora devono sviluppare il loro potenziale.

I ruoli di responsabilità di Lino Leonardi ed Emiliano Degl'Innocenti nel nodo italiano di DARIAH-ERIC (*Digital Research Infrastructure for the Arts and Humanities*), hanno consentito un graduale e sempre più ampio coinvolgimento dell'OVI nelle attività delle infrastrutture della ricerca umanistica a livello europeo. Lo scopo è inizialmente quello di descrivere l'articolazione delle attività dell'OVI nell'ambito delle Digital Humanities e delle infrastrutture di ricerca a partire dal progetto RESTORE (*smaRt accESs TO digital heRitage and mEmory*), contestualizzandone le attività all'interno di Dariah.it — mediante la partecipazione al cluster SSHOC (*Social Sciences and Humanities Open Cloud*) — arrivare alla European Open Science Cloud.

L'obiettivo di RESTORE, progetto coordinato da Emiliano Degl'Innocenti e finanziato dalla Regione Toscana, e con il cofinanziamento di una Pmi come Space spa, è migliorare l'accessibilità delle risorse digitali delle istituzioni culturali pratesi (l'Archivio di stato, il Museo di Palazzo Pretorio), per valorizzare risorse documentali di grande valore anche per la lessicografia e la storia della lingua, tra cui spicca il Fondo Datini, di cui (come si ricordava all'inizio) è disponibile un corpus in GattoWeb.

Con questa linea di attività, l'OVI si sta impegnando a sviluppare soluzioni digitali per rendere FAIR (*findable*, *accessible*, *interoperable*, *reusable*, ovvero ricercabili, accessibili, interoperabili, riusabili) le risorse e gli strumenti digitali di ambito sia umanistico (aspetti immateriali del patrimonio culturale), sia scientifico (analisi chimico-fisiche, immagini multi-spettrali, modelli 3D e simulazioni), mediante l'elaborazione di una piattaforma per la loro fruizione integrata.

[9] S. Arcidiacono, *Pluto – Piattaforma lessicografica unica del tesoro delle origini*, in *Italiano antico, italiano plurale*, cit., pp. 209-17; Id., *L'informazione lessicografica nel Dictionary Writing System del TLIO*, «Bollettino dell'Opera del Vocabolario Italiano», XXIV, 2019, pp. 381-389.

[10] R. Leporatti, T. Salvatore (a cura di), *Le rime disperse di Francesco Petrarca. Problemi di definizione del corpus, edizione e commento*, Atti dell'Atelier (Vandœuvres, 23 novembre 2018), Roma, Carocci, 2020.

L'obiettivo è ridurre la frammentazione dell'ecosistema digitale, superare le barriere disciplinari e facilitare la creazione della sezione umanistica della European Open Science Cloud, portando al suo interno contenuti e strumenti di alta qualità scientifica e tecnologica.

Fra le risorse già disponibili si segnala la biblioteca digitale di testi tratti dal *Corpus OVI dell'italiano antico* (comunque fuori dal diritto d'autore) disponibile nel portale di Europeana[11], un'iniziativa promossa e finanziata dall'UE per la condivisione del patrimonio culturale europeo, al quale abbiamo avuto accesso grazie al coordinamento di DARIAH-IT.

Ma siamo solo all'inizio.

[11] Il portale è molto ricco e variegato: la biblioteca digitale OVI si può visualizzare digitando 'edizioni critiche' nella casella di ricerca indicata con l'icona usuale della lente d'ingrandimento.

https://doi.org/10.6092/issn.1721-4777/12279