

LA LESSICOGRAFIA ELETTRONICA E ITALIANO DELLE ORIGINI

Salvatore Arcidiacono

1. LA LESSICOGRAFIA ELETTRONICA

1.1 Interazioni disciplinari tra lessicografia e informatica 1.1.1 Il contributo della lessicografia all'informatica

A partire dalla fine degli anni Settanta dello scorso secolo, gli editori *Larousse*, *Éditions Le Robert*, *GETA* e *IBM*, così come alcuni centri di ricerca specializzati nella lessicografia scientifica, hanno attinto informazioni dai dizionari tradizionali per popolare e commercializzare banche dati a uso di sistemi digitali.

Il *Longman Dictionary of Contemporary English (LDOCE)*, pubblicato nel 1978 è l'esempio più noto di questa tendenza: i contenuti comprendevano anche informazioni non presenti nella versione a stampa. Alcuni studiosi, individuano nel passaggio del *LDOCE* a *machine-readable dictionary* la nascita del database lessicale.

L' *Oxford English Dictionary* ha contribuito alla definizione dello *Standard Generalized Markup Language (SGML)* da cui sono derivati XML e HTML.

1.1.2 Implicazioni metalessicografiche

L'editoria delle opere di consultazione ha seguito una traiettoria in virtù dell'esperienza di lettura meno sensibile alla perdita delle coordinate fisiche e al piacere della felicità della carta.

Dalla prospettiva umanistica, le interpretazioni che si rintracciano in letteratura sono diverse e degne di considerazione. C'è una generica somiglianza nel *modus operandi* delle discipline umanistiche e dell'informatica.

Quemada spiega la precoce adozione del nuovo strumento tecnico con una particolare capacità di adattamento dei lessicografi, che avrebbero saputo conciliare le esigenze della ricerca con le limitate prestazioni offerte dai primi calcolatori.

Ad aver fornito l'alchimia sarebbe stata la singolare modalità di lettura delle opere di consultazione. Queste opere rispondono alla funzione di soddisfare rapidamente bisogni specifici in opposizione ai bisogni informativi generali.

La compatibilità tra informatica e lessicografia ha rapidamente mutato gli scenari consueti, contribuendo a stimolare il crescente interesse per la riflessione teorica.

Mettere sotto accusa la tradizionale afferenza della lessicografia alle discipline linguistiche, in favore di una piena adesione alla scienza dell'informazione.

1.2 Il dizionario elettronico

1.2.1 Nuove forme di risorse lessicali

Alla diffusione delle tecnologie digitali nel mondo contemporaneo è corrisposta una proporzionale proliferazione di risorse lessicali implementate su supporti digitali: il dominio dell'informazione

lessicale si è trasformato in una ricca costellazione di prodotti che arricchiscono l'offerta di soluzioni linguistiche disponibili sul mercato.

Questi oggetti si collocano in un *continuum* indistinto tra il polo informatico e quello umanistico.

Il *Devoto-Oli* (DO) è utilizzato da tutti i dispositivi apple come principale fonte di informazione lessicografica; con la sola eccezione dell'app 'Dizionario' l'utente non deve lasciare l'applicazione che sta utilizzando e, in molti casi, non deve neppure digitare la stringa di ricerca: sono le stesse app a reperire il lemma corretto al clic di una parola.

L'utente medio non ha idea di richiedere informazioni ai dizionari.

In questo quadro, dove comunque i dizionari sono molto più diffusi che in passato, cercare di mettere ordine diventa difficoltoso.

1.2.2 Per una definizione di lessicografia elettronica

Semplificazione e definizione dizionari elettronici → questo tentativo di circoscrivere il campo di osservazione è confortato dalla tendenza in ambito internazionale a restringere l'applicazione del termine dizionario elettronico ai dizionari *human-oriented*.

Un vocabolario elettronico può e deve conciliare l'aspetto *human-oriented* con quello *software-based*.

Il crescente fenomeno dell'ibridazione, riconosciuto all'unanimità come un elemento di vanto del processo digitale, rende labile il confine tra dizionari e altri artefatti informativi, che si trovano a convergere in un unico prodotto.

I dizionari tradizionali possono accrescere il loro valore se corredati di applicazioni linguistiche o strumenti per la scrittura e viceversa.

È innegabile che, a livello di tecniche e strutture di dati, ci sia una grande distanza tra i due mondi – i lessici per il NLP prevedono notoriamente formalismi ostici per il lettore umano e una diversa complessità tecnica – ma è evidente una virtuosa tendenza ad applicare teorie e metodi nel NLP alla lessicografia.

1.3 Classificazioni e valutazioni

1.3.1 Applicazione strumentale o cambio di paradigma

La raffinatezza della rappresentazione e il grafo di elaborazione tecnica rientrano come parametri nell'esercizio definitorio e classificatorio della lessicografia elettronica. L'impiego del computer in lessicografia non può essere ritenuto sufficiente per poter parlare di un dizionario elettronico "in senso forte".

Questo atteggiamento è radicato nelle *digital humanities*; l'applicazione dell'informatica a un dominio disciplinare produce ripercussioni profonde sulle metodologie e sulle pratiche di ricerca, fino a un possibile cambio di paradigma innescato dal calcolatore elettronico e che interessa le scienze umane.

Da questi presupposti sono seguiti diversi moniti contro la "singolare cecità teorica" che applica la scienza dell'informazione come un nuovo strumento con cui affrontare i vecchi problemi senza interferire con lo spazio epistemico di riferimento.

Questo principio si è già radicato come necessità di una riflessione critica sul mutamento causato dalle tecnologie sui domini disciplinari e come stimolo a riformare gli oggetti tradizionali per pervenire a

un'innovazione significativa. → cambio di paradigma, da cui sono esclusi tutti quei dizionari elettronici che si limitano a riproporre su un nuovo supporto le forme del dizionario tradizionale.

Distinzione tra dizionari accessibili in digitale come immagine e dizionari *machine-readable* → sottoposto a una sequenza di ulteriori lavorazioni.

Sahle → introduce il concetto di "*page paradigm*" come persistenza del modello culturale del libro. La conformità alla pagina stampata permette di distinguere le "*digital editions*" dalle "*digitalized editions*".

Se non tutti i vocabolari sono *machine-readable* sono "elettronici" → catalogazione:

- a. *Copycats* → fotocopie; versioni pagina per pagina su PDF;
- b. *Faster horses* → dispositivi di lemmatizzazione delle forme flesse e di riconoscimento delle variabili ortografiche. Categoria ricondotta alla maggior parte dei dizionari elettronici in circolazione: i *born-digital*.
- c. *Ford T Model* → riescono a adattare gli articoli ai bisogni degli utenti e funzionano attraverso la rete; possibilità di rimaneggiare dinamicamente i dati.
- d. *Rolls Royce* → rirappresentare l'informazione e creare informazioni nuove a partire da quelle già esistenti.

La classificazione di Tarp potrebbe essere ulteriormente estesa oltre i confini del singolo progetto lessicografico.

Il dizionario può essere costituito come un servizio e può diventare un modulo da integrare con altri dizionari o all'interno di altri servizi come biblioteche digitali o *corpora* elettronici.

1.3.2 Dizionari *born-digital* e digitalizzati

La periodizzazione di Bernard Cerquiglini articola la storia dei dizionari elettronici in 3 fasi:

1. Lessicografia di dizionari cartacei assistita dal calcolatore;
2. Digitalizzazione di dizionari cartacei esistenti → individua le azioni di intervento retrospettive, su risorse esistenti che vengono definite in retroconversione; in opposizione ai dizionari *born-digital*. L'operazione deve comunque avvalersi di tecniche all'avanguardia in grado di determinare un reale miglioramento rispetto alla versione cartacea, in termini di estrazione di informazione implicita o di efficienza d'accesso ai dati.
3. Dizionari concepiti e realizzati espressamente per il supporto digitale → L'ultimo stadio viene talvolta visto come l'unico reale compimento della lessicografia elettronica.

Dalla prospettiva delle *digital humanities*, l'uso del mezzo digitale "non come sostituto ma come antecedente della stampa è fondamentalmente errato".

Le prime esperienze di dizionari supportati dal calcolatore rientrano solo perché per il primo vocabolario consultabile su un dispositivo elettronico si dovrà attendere il 1978 con l'LK-3000 della Lexicon Corporation.

1.3.3 La digitalizzazione dei Vocabolario degli Accademici della Crusca

La sua digitalizzazione ha unito due approcci:

1. Quello francese del TLF, che ha privilegiato il testo elettronico strutturato;
2. Quello dei dizionari digitalizzati della Real Accademia Espanola che ha puntato sulla fedeltà della riproduzione facsimilare.

La versione digitale nasce con l'impostazione di un vocabolario elettronico in senso forte: oltre alla volontà di preservare e migliorare l'accesso, troviamo l'urgenza del "superamento della consultazione tradizionale del Vocabolario per una "lettura" complessa e attenta al medesimo".

È stato formulato il concetto di "rovesciamento", un rimaneggiamento sistematico dei contenuti del *Vocabolario* al fine di far affiorare tutto il lessico contenuto nell'opera e non solo quello indicizzato nel lemmario.

A spingere verso un'operazione di rovesciamento erano le stesse caratteristiche del *Vocabolario*:

- a. Sottolemmi senza rimando;
- b. Voci di definizioni, locuzioni e proverbi;
- c. Significati del lemma registrati altrove;
- d. Definizioni non evidenziate nell'articolo del lemma.

Il concetto di "rovesciamento" è un'azione del tutto assimilabile alla strutturazione di un vocabolario elettronico in senso forte. La digitalizzazione è stata effettuata in piena conformità con gli standard internazionali.

La codifica e gli strumenti di interrogazione sono stati calibrati; l'analisi dell'oggetto codificato ha così beneficiato di un adattamento granulare alle metodologie informatiche, rispettoso delle specificità testuali del *Vocabolario* fin nei dettagli.

- La tokenizzazione del testo è stata considerata secondo l'uso del tempo;
- La punteggiatura può essere ricercata al pari delle forme in tutti i tipi d'interrogazione;
- Le abbreviature sono gestite per permettere il collegamento diretto con schede dedicate a ciascun testo citato e per raccogliere esempi.

L'operazione di digitalizzazione ha consentito di conoscere meglio la ricchezza informativa del *Vocabolario* e quantificarne le dimensioni nelle diverse edizioni.

1.4 I framework di valutazione

Schall → fornisce un censimento di criteri di valutazione relativi al mezzo organizzati in 18 gruppi tematici.

Studio di Pearson / Nichols → tentativo di elaborare un *framework* per la valutazione dei dizionari in rete per la lingua inglese.

In questo lavoro vengono definiti 8 criteri generali, 3 dei quali dedicati prevalentemente agli aspetti relativi al mezzo:

- a. Qualità dell'interfaccia → in considerazione il design della piattaforma e la sua accessibilità;
- b. Funzioni di ricerca → rileva e valuta la ricerca estesa a varianti e sinonimi;
- c. Funzioni "extra" del sito → funzioni accessorie es. parola del giorno.

Lew / Szarowska → prevedono le mosse da Pearson / Nichols per proporre una versione modificata della griglia dedicata ai vocabolari bilingui *online* articolata su due assi:

1. *Medium-independent* → parametri lessicografici es. copertura degli elementi lessicali; ritenuti più importanti rispetto ai giudizi relativi all'accesso e alla presentazione dei dati.
2. Parametri di pertinenza del vocabolario elettronico.

A livello di macrostruttura si valutano gli strumenti per agevolare l'identificazione del lemma, l'inclusione tra i risultati delle polirematiche e la ricerca incrementale.

A livello della microstruttura si valutano i dispositivi per la lettura delle voci.

1.5 Lessicografia evolutiva

“Parlare di completamento per i dizionari accademici è come parlare di arcobaleni che si allontanano man mano che ci si avvicina”.

Le infrastrutture digitali forniscono un espediente concreto al problema dell'elaborazione “infinita”, mantenendo il dizionario sempre aperto a correzioni, modifiche e aggiornamenti.

Le scienze “dure” hanno sempre usato costrutti manipolabili ma solo il computer ne ha permesso una più larga diffusione e ha consentito di ridurre i tempi delle manipolazioni.

Si afferma così il modello dei dizionari *work in progress* o *under construction*: pubblicati in corso d'opera senza attendere il completamento dei lavori.

es. → *Dictionnaire du Moyen Français (DMF)*: è uno tra i primi esempi di un impianto fortemente caratterizzato da questo approccio.

Questo approccio risolve in parte l'annoso problema dei lunghi tempi di gestazione di un dizionario.

Un dizionario evolutivo può favorire l'accesso alle fonti di finanziamento, e rispettare i tempi di rendicontazione dei bandi competitivi.

La pubblicazione in rete da un lato produsse un nuovo interesse internazionale verso il progetto e dall'altro diede un'impressione favorevole ai finanziatori.

1.6 Supporti digitali, il World Wide Web e l'ipertesto

Il vocabolario elettronico dimora in un ecosistema digitale complesso che va considerato nella sua interezza: l'elemento centrale del calcolatore andrà considerato all'interno di un sistema integrato di trattamento e trasmissione dell'informazione di cui fanno parte le infrastrutture di trasmissione.

Il dizionario elettronico ha più volte cambiato supporto determinando una parallela alterazione nel concetto di vocabolario elettronico e nelle sue proprietà.

L'unico che ha determinato un reale spartiacque è quello sulla rete Internet e World Wide Web, tanto che alcuni studiosi limitano l'uso del termine ‘vocabolario elettronico’ solo ai dizionari *on-line*.

Il dizionario dimostra inoltre una buona affinità con il concetto di ipertesto perché le opere di consultazione si sono sempre sforzate di permettere una modalità di lettura multilineare, non-sequenziale e interattiva.

Qui il vocabolario, oltre ai collegamenti interni tra le sue unità di informazione, può sviluppare collegamenti con altre risorse lessicali, linguistiche e testuali in una rete di rimandi.

2. L'INFRASTRUTTURA DIGITALE IN LESSICOGRAFIA

2.1 Il sistema informativo lessicografico: Corpus Query System (CQS) e Dictionary Writing System (DWS)

All'inizio di un progetto è opportuno riconoscere l'estensione dell'operazione che, nel caso della lessicografia scientifica e storica, assume quasi sempre dimensioni notevoli per l'estensione della macrostruttura e la ricchezza della microstruttura.

Nella lessicografia tradizionale una consistente parte degli sforzi era rivolta alla costruzione del sistema informativo, soprattutto in relazione alle attività di spoglio e di schedatura.

Nella prima metà del XX secolo la lessicografia manifesta il bisogno di porre rimedio al problema meccanizzando i processi e riducendo le dimensioni dei supporti.

Solo nella seconda metà del secolo, l'infrastruttura digitale avrebbe consentito una reale dematerializzazione degli schedari e un miglioramento delle modalità di accesso agli archivi.

I *corpora* elettronici, che costituiscono la parte più consistente del sistema informativo, tendono ad assumere dimensioni monumentali e in lessicografia "ciò che sembra essere un fatto quantitativo, ha in realtà un lato qualitativo".

La rappresentatività di un *corpus* è proporzionale alle sue dimensioni → i fenomeni più rari es. proposizioni relative richiedono testi più lunghi per presentare distribuzioni stabili. Bisogna superare una certa soglia nel volume di dati per poter apprezzare la distinzione tra le relazioni di co-occorrenza significative.

Le tecnologie digitali sono state integrate nei processi di creazione e distribuzione dei dizionari, al punto che oggi la quasi totalità delle imprese lessicografiche fa uso di un certo numero di sudditi digitali e può dirsi seppure in minima parte informatizzata. Nelle redazioni che si affidano a sistemi specializzati sono di solito presenti due dispositivi:

1. *Corpus Query System (CQS)* → gestione per l'interrogazione delle fonti di informazione linguistica;
2. *Dictionary Writing System (DWS)* → assiste la scrittura delle voci e ne gestisce la rappresentazione digitale e l'archiviazione.

Il DWS è il sistema più vicino al vocabolario come il CQS lo è alle collezioni testuali.

2.2 Le fonti primarie e il Corpus Query System

2.2.1 La stagione delle concordanze e il progetto del Vocabolario storico

Le attività del *Vocabolario storico* e dell'opera del *Vocabolario Italiano* si inseriscono in questo filone di ricerca attraverso una fitta rete di contatti e collaborazioni.

Gli spogli di Busa convergeranno nelle attività del *CNUCE*. Sotto la direzione di Zampolli nel 1970 fu istituita una divisione per la linguistica che diventerà *l'Istituto di Linguistica Computazionale del CNR*. I contatti tra questi centri avvengono per iniziativa dell'Accademia della Crusca.

Dopo la promessa di un primo stanziamento di fondi da parte del CNR Duro viene incaricato di prendere contatti con Busa.

Duro ampliò i viaggi per stringere contatti nei maggiori cantieri lessicografici d'Europa, raccogliendo informazioni sui metodi di spoglio e schedatura, sull'organizzazione complessiva della redazione e sullo stato dell'arte dell'elaborazione elettronica dei dati linguistici.

1964 → possibilità di collaborazione con *SEIOD*: esempio di collaborazione internazionale nel campo della ricerca umanistica assistita dal calcolatore.

In questo periodo in Italia si sviluppano le tecniche di automazione per il trattamento del testo e comincia a maturare la sensibilità per i sistemi di codifica; l'ecosistema digitale non dispone di adeguati strumenti di disseminazione: i risultati dell'analisi computerizzata non possono che finire stampati come concordanze e indici in volume.

In Italia questo momento coincide con il rilascio del programma *DBT*.

2.2.2 *Corpora e lessicografia diacronica*

In lessicografia, i *corpora* elettronici diventano rapidamente uno strumento di analisi privilegiato, permettendo di osservare e analizzare agevolmente grandi quantità di attestazioni.

La semplicità di archiviazione e di accesso del CQS favorisce la realizzazione di dizionari “di prima mano” in cui l'analisi si concentra principalmente all'interno dei confini ben definiti; è una linea d'indagine empirica, assimilabile alla linguistica *corpus-based*.

Il CQS ha consentito al redattore di riappropriarsi dello spoglio integrale, mentre in precedenza il compito di spogliare e schedare i testi era commissionato a non professionisti.

La bontà di uno spoglio di scelta dipendeva:

- a. Affidabilità e sensibilità linguistica dei lettori;
- b. Accuratezza delle linee guida per lo spoglio definite inizialmente.

Il CQS ha reso lo spoglio integrale meno oneroso. La portata del suo cambiamento rende possibile la creazione di viste aggregate delle informazioni e *report* statistici per supportare le intuizioni del lessicografo con evidenze empiriche e rivelare aspetti delle parole che sfuggono all'intuizione stessa.

Accompagnare il vocabolario con l'accesso diretto alle fonti primarie ne aumenta il valore scientifico.

2.2.3 *Fonti ibride*

I *corpora* di dati testuali possono essere integrati con fonti di informazione che comprendono dati linguistici lavorativi o semi-lavorativi.

Le fonti primarie del *Vocabolario Dinamico dell'Italiano Moderno (VoDIM)* prevedono un corpus bilanciato in diacronia di dimensioni più contenute da abbinare a un corpus sincronico.

La *Stazione lessicografica del VoDIM* si configura come una fonte composita. Questa “banca linguistica”, oltre ai *corpora* principali, raccoglie archivi testuali e strumenti lessicografici digitali.

Gli aggregatori possono collocarsi a metà strada tra CQS e DWS. Nell'ambiente di redazione il DWS può fungere da aggregatore di fonti perché la raccolta dei contesti di occorrenza attraverso le risorse può essere operata all'interno dell'interfaccia di inserimento delle voci.

Grazie alla tecnologia, i vocabolari storici possono configurarsi come portali per l'accesso all'intero sistema di risorse linguistiche in rete in cui, la lettura della voce sul dizionario costituisce l'avvio di nuovi percorsi di esplorazione.

2.3 Il Dictionary Writing System

2.3.1 Il DWS come sistema di redazione

Con *Dictionary Writing System* si intende un sistema specializzato per la redazione del dizionario. I *DWS* forniscono modalità di rappresentazione e archiviazione delle voci persistenti e sostenibili ma, tendono a configurarsi come *suites* di strumenti più ricche. È un sistema molto più versatile rispetto al *CQS*.

Talvolta la piattaforma lessicografica viene adoperata anche per lavorare sui testi in un ambiente collaborativo prima dell'indicizzazione effettiva del *CQS*.

Il *DWS* opera meglio quando la sua personalizzazione è minuziosa e per tale ragione la configurazione più diffusa dei principali centri lessicografici prevede un *CQS* di terze parti e un *DWS* sviluppato appositamente “*in-house*”.

2.3.2 Lo spazio: dall'archiviazione all'interfaccia utente

L'ambiente di scrittura digitale del *DWS* offre spazio illimitato e libera il redattore dall'esercizio di compressione dell'informazione a cui prima era costretto.

È possibile ampliare indefinitamente la copertura del lemmario senza particolari controindicazioni. L'espansione interna alla macrostruttura comporta alcune cautele.

La disponibilità di spazio infinito è diventata una sfida nella lessicografia elettronica poiché gli utenti rischiano di essere soverchiati da un numero crescente di dettagli a fronte di un bisogno di informazioni molto più limitato. Andranno evitate situazioni di sovraccarico informativo.

A livello della piattaforma è bene distinguere tra infinito spazio di archiviazione e limitato spazio di presentazione.

Per i dizionari scientifici questi problemi possono essere quasi sempre risolti con un buon *layout* “responsivo”.

La letteratura specialistica da molti anni fornisce spunti e sperimentazioni interessanti, come il *simultaneous feedback* introdotto da Schryver in cui l'utente influenza la visualizzazione del dizionario in un ciclo continuo.

In risposta alle possibili perplessità sulle scelte coscienti degli utenti sono state messe a prova metodologie di analisi come l'*eye tracking*.

2.3.3 Il tempo: gestione del progetto e tracciamento della redazione

Nei requisiti di base, la presenza di strumenti per la gestione del flusso di lavoro è inseparabile dal *DWS* perché fa parte dei dispositivi collaborativi implementati nel programma informatico: l'assegnazione di attività e permessi e il tracciamento delle attività sono gestiti in maniera nativa dalla piattaforma.

I ritardi nelle consegne possono essere limitati con statistiche e controlli in tempo reale delle attività effettuate sulla piattaforma; i contrattamenti possono essere limitati procedendo per tappe autonome. Le lunghe scale temporali dei processi lessicografici hanno comunque un effetto deleterio su qualunque vocabolario.

La digitalizzazione conferisce ordine e consistenza sui progetti di ampie dimensioni in cui il fattore umano può determinare ampie oscillazioni qualitative.

2.3.4 Le risorse; pubblicazione gratuita e sostenibilità a lungo termine

Nel nuovo millennio l'autore può pubblicare autonomamente il suo lavoro sulla rete a costi irrisori. I costi continuano ad esistere ma possono compromettere l'intero lavoro anche dopo molti anni dal completamento.

Le instabilità di tecnologie poco matura hanno pesantemente penalizzato i progetti del *Vocabolario storico* e del *TLIO*, che avevano scelto di adottare precocemente le nuove tecnologie digitali.

La stabilità digitale del nuovo millennio non arresta obsolescenza, ne ritarda solamente le manifestazioni più evidenti per un lasso di tempo che giunge a termine quando le risorse inizialmente disponibili sono esaurite o notevolmente ridotte.

A differenza di un libro stampato, il dizionario elettronico per essere consultato richiede una manutenzione continua e un costante supporto tecnico: un dizionario *on-line* continua a esistere finché il server che lo ospita rimane acceso.

La dipendenza dai finanziamenti esterni rende ancora più difficile la sopravvivenza a lungo termine dei dizionari.

Per i dati testuali una risposta arriva dall'uso di tecnologie di codifica *standard* come XML-TEI.

Gli standard possono costituire un buon antidoto contro l'obsolescenza dei dati: più lo standard viene usato, maggiore sarà l'interesse a tenerlo vivo.

Per i *DWS* e i *CQS*, non esistono soluzioni altrettanto promettenti: i *software* sono prodotti molto complessi, e sono sostituiti richiamando componenti e librerie di terze parti con cui deve essere sempre preservata la compatibilità.

Le piattaforme che veicolano il progetto scientifico e lo rendono accessibile nella forma in cui lo conosciamo sono purtroppo destinate a un ciclo di vita limitato.

Gli sforzi di preservazione e standardizzazione vanno concentrati sui contenuti, preferibilmente codificati in formati *standard*.

3. PLUTO: PIATTAFORMA LESSICOGRAFICA UNICA DEL TESORO DELLE ORIGINI

3.1 Un DWS per l'italiano delle origini

3.1.1 La procedura COVIREN e il vocabolario come base di dati

La prima generazione di *DWS* fa la sua comparsa nell'ultimo decennio del secolo scorso.

A quell'altezza cronologica presso l'*Opera de Vocabolario Italiano* è stato avviato lo sviluppo della procedura COVIREN → proponeva di individuare le procedure informatiche più adeguate all'allestimento del *TLIO*.

La procedura riconosceva l'esigenza "immediata del redattore relativa alla gestione dell'ingente molte di dati raccolta" e l'esigenza di una procedura di redazione interamente assistita dal calcolatore fino alla pubblicazione.

Informazioni più precise fornite sugli aspetti digitali miravano a ottenere:

- a. Un sistema di redazione modulare che si ricollegava direttamente al concetto di lessicografia evolutiva;
- b. Dispositivi di assistenza all'inserimento delle voci;
- c. Strumenti per l'analisi dei dati.

Buona parte delle prerogative COVIREN si fonda sul principio del "vocabolario considerato come base di dati".

Il *database* moltiplica le possibilità di lettura discontinua e multilineare già riscontrate con l'ipertesto.

L'organizzazione per unità discrete del *database* è inoltre profondamente compatibile con la struttura prototipica del dizionario.

Il *database* è quindi una forma di organizzazione razionale dei sistemi informativi che opera con modelli, ciascuno dei quali determina modi di costruire, archiviare e interrogare i dati.

Il modello relazionale previsto da COVIREN pone l'accento sulla corretta interpretazione dei dati al fine di:

- a. annullare la ridondanza;
- b. evidenziare le relazioni tra informazioni;
- c. garantire l'integrità referenziale dell'archivio anche in presenza di grandi quantità d'informazioni.

Il *design* efficace di una base di dati è responsabile dei successivi trattamenti possibili, perché dal modo in cui le informazioni sono distribuite nei *record*, dipenderanno le analisi che sarà possibile effettuare.

Il modello relazionale con *record* a struttura fissa viene considerato molto rigido per un vocabolario; per questo possono essere preferiti formati di dati semi-strutturati come l'*XML*.

La consolidata tecnologia dei *database* relazionali ha caratteristiche di efficienza e *performance* notevolmente superiori a quelle di sistemi alternativi.

Per la preservazione a lungo termine, i linguaggi di *markup* e gli *standard* internazionali rimangono preferibili alle basi di dati.

Per garantire la sopravvivenza del dizionario a lungo termine il *DWS* dovrà codificare le voci in formato semi-strutturato.

3.1.2 Il TLIOWeb e ReddiX

La procedura COVIED non riuscì ad assistere la stesura delle prime voci del *TLIO*. Il primo sistema digitale per la messa in rete risale al '97 e consisteva in un semplice file di testo indicizzato da uno *script* in *PERL* e operante in ambiente *SunOS*.

Nel 2002 fu sostituito dal sistema *TLIOWeb* → è un insieme di procedure che permettono di trasformare e pubblicare in rete le voci del vocabolario redatte in Word. Prevede due blocchi:

1. Le procedure di normalizzazione e indicizzazione dei *file*;
2. La piattaforma Web.

Negli anni sono stati sviluppati alcuni moduli aggiuntivi.

Nel 2011 è stato avviato lo studio e lo sviluppo di un vero e proprio *DWS* rivolto a sostituire la redazione delle voci in Word, denominato *ReddiX*.

L'architettura era divisa in due "macromoduli", uno *off-line* per la redazione delle voci e uno *on-line* per la pubblicazione in rete e a stampa del dizionario. Costituiti da *software* separati funzionavano con modalità di archiviazione differenti, interagenti con una logica *client/server*:

- a. *Client* → costituito da un'applicazione desktop per windows;
- b. *Server* → raccolta centralizzata delle voci.

Il progetto prevedeva lo sviluppo di funzioni di verifica formale delle voci, modellate sulle *Norme di Redazione* di Beltrami; doveva prevedere la possibilità di realizzare funzioni di consultazione e ricerca molto più potenti di quelle offerte.

Il successo più importante per *ReddiX* è stato il rilascio di un sistema di conversione automatico delle voci da Word a un formato XML modellato sulle specificità del foglio di stile con cui sono state redatte le voci del *TLIO*.

3.1.3 La piattaforma Pluto

Pluto ha impostato un nuovo paradigma di *DWS* per gli antichi volgari italiani. È un approccio basato su una piattaforma *on-line* unica.

La gestione centralizzata ha sollecitato il contestuale trasferimento sulla piattaforma dell'intero complesso di risorse informative sviluppate e conservate dall'Istituto sviluppando ulteriormente il concetto di piattaforma unica.

Al di là degli aspetti applicativi, la completa integrazione mira a realizzare quella complementarità "interna" tra progetti scientifici.

Valutazioni a favore dell'integrazione si ritrovano anche nello sviluppo della piattaforma digitale Pasadena dell'*OED*, riportate in Elliott/Williams.

Pasadena ha risolto i limiti del precedente sistema che distribuiva le informazioni in differenti *database* e che utilizzava una serie di moduli secondari autonomi per il montaggio del flusso di redazione.

Le possibilità di automatizzazione erano limitate, la bibliografia era rimasta scarsamente informatizzata e, le annotazioni amministrative ed editoriali avevano finito per confluire sui testi stessi, con la conseguente esigenza di allestire meccanismi di "pulizia" per la visualizzazione e ricerca.

La pubblicazione di Pluto ha tentato di potenziare il più possibile la sinergia tra redazione e pubblicazione.

Con Pluto è stata messa alla prova l'ipotesi che sia possibile separare un livello lessicografico più astratto, trasferibile integralmente su altre piattaforme, dalle personalizzazioni richieste per il *TLIO*.

Oltre a contenere le classi astratte dedicate alle principali entità lessicografiche, Lexicad comprende anche le librerie e gli strumenti che descrivono i dati e i comportamenti di un ambiente collaborativo per la redazione e pubblicazione di dizionari *web-based*.

3.2 La maschera di redazione in Pluto

3.2.1 La compilazione della voce

Muller-Spritzer osserva che il dizionario digitale si manifesta solo per frammenti virtuali perché non può essere pienamente colto nella sua interezza ma solamente come una combinazione dinamica tra base di dati e *output* finale.

Il dato rimane in movimento a opera del sistema di interrogazione e presentazione, che riadatta continuamente il contenuto del vocabolario in risposta alle specifiche richieste dell'utente.

Il *LDOCE* prima degli altri ha dimostrato la versatilità di un'architettura che prevede un nucleo informativo strutturato.

Dalla stessa base di dati sarà così possibile ricavare una serie di pagine web, una versione stampabile in PDF, un eBook, un'app per mobile, diversi report statistici o nuove fonti di dati ad uso di ulteriori *software*.

Se il dizionario si manifesta solo nell'interazione con un'interfaccia, anche il redattore deve essere considerato un utente e anche per lui si dovrà "costruire la maschera di una scheda che consenta di introdurre i dati in maniera rapida e facile".

3.2.2 Mappatura della microstruttura e punti della voce semplici

Il cuore del *back-end* è la maschera di redazione della voce, in cui la microstruttura del vocabolario viene mappata su specifici componenti dell'interfaccia.

I campi separati e gli elementi a inserimento guidato hanno l'effetto di ridurre l'onere della formattazione del testo e dell'impaginazione della voce.

I riferimenti interni sono istituiti a partire da caselle di ricerca e selezione. Tutti gli elementi a inserimento guidato o vincolato concorrono ad accelerare ulteriormente l'inserimento della voce.

Disponendo dell'accesso alla maschera, non tutti i campi potrebbero essere visualizzati o modificabili.

Tre caratteristiche del dizionario elettronico:

1. La lessicografia evolutiva → tramite la struttura modulare le voci possono essere redatte per campagna di redazione successive che possono essere assegnate a utenti diversi.
2. La microstruttura della voce su Lexicad/Pluto viene abitualmente ampliata con punti "di servizio".
3. Alcuni punti saranno direttamente connessi alla gestione del flusso di lavoro e rimarranno di esclusiva pertinenza degli utenti con ruoli gerarchicamente più alti.

3.2.3 Punti della voce complessi

In qualsiasi punto della maschera è possibile incapsulare *script* o vere e proprie mini applicazioni che rispondono alle esigenze particolari della microstruttura.

La maschera di redazione, associando i contesti di occorrenza al lemma, può compilare autonomamente, riordinare e formattare l'elenco delle forme attestate nel *corpus* principale. Il formario definitivo è sempre il risultato di ripetute operazioni di confronto tra diverse fonti di informazione.

Nella mini applicazione sono presenti un comando per rilevare eventuali duplicati del formario e uno per convertire tutte le forme in minuscolo. L'applicazione, inoltre, su occupa di uniformare punteggiatura e spazi e riordinare alfabeticamente le forme al momento del salvataggio.

In nessun caso le procedure automatiche mirano a sostituirsi al lavoro del lessicografo.

Il sistema segnalerà ai responsabili la presenza di *record* orfani per l'eventuale rimozione manuale dei singoli elementi non più necessari.

3.2.4 Punti della voce trasformati in elementi autonomi

Il *back-end* mira a fornire un ambiente intuitivo e a mantenere un'impostazione simile per tutte le risorse gestite dalla piattaforma.

Nella *homepage* del pannello di redazione viene visualizzato l'elenco degli elementi per i quali l'utente autenticato dispone dei privilegi di lettura o di modifica; selezionando un elemento si accede alla relativa "pagina di gestione".

Ogni interfaccia di modifica non si limita a mostrare i campi di stretta pertinenza dell'elemento corrente, ma può coinvolgere altre risorse con cui questo entra in relazione: per limitare esempi alla voce, la maschera di redazione della voce mostra l'elenco di tutti i testi citabili; analogamente, lavorando a una voce, è possibile scegliere un etimo tra quelli già registrati o inserirne uno nuovo e associarlo contestualmente.

Un punto della voce può essere trasformato in un elemento autonomo pur rimanendo integrato nella microstruttura della voce.

Una parte della microstruttura può essere trasformata in un elemento autonomo in diversi casi:

1. Quando la medesima unità è suscettibile di essere riferita a due diversi elementi contemporaneamente o su diverse implementazioni;
2. Quando l'entità può essere associata a più punti della voce;
3. Quando può essere utile creare una tabella separata sul *database* per la singola entità o disporre di un ambiente redazionale separato per la loro gestione;
4. Tutte le volte in cui ci siano ragioni tecniche o logiche.

Alcuni di questi elementi autonomi possono essere creati e modificati dalla maschera di inserimento della voce.

Ogni volta che nella maschera di redazione della voce viene introdotto un *record* non censito nel sistema, questo verrà registrato anche nel suo archivio dedicato.

3.3 Le fonti primarie nel TLIO e il gestore dei contesti di Pluto

3.3.1 Le fonti primarie del TLIO

Nel 1964 fu confermato il doppio criterio di schedatura delle fonti. Negli anni il *TLIO* si è definito ma è rimasto basato sullo spoglio integrale del *Corpus TLIO*.

Il corpus mira alla completezza per l'intera produzione volgare scritta fino al 1375.

Come affermano Beltrami/Forana il *TLIO* è sospeso tra un dizionario di un *corpus* e uno di lingua.

Un dizionario storico come il *TLIO* è orientato alla lingua, di cui mira a interpretare il sistema lessicale nel suo complesso.

Il *corpus* rappresenta il nucleo di informazione primario ma la ricerca del redattore può essere allargata a tutte le fonti pertinenti. Il *corpus*, in questo contesto, non implica una limitazione preventiva delle attestazioni.

Il *TLIO* rivela la propria natura di vocabolario di lingua per dettagli:

- Predilige i significati di carattere generale e non le spiegazioni dei singoli esempi;
- Accorda la preferenza alle definizioni perifrastiche e non sinonimiche;
- Non ha l'obbligo di spiegare tutte le attestazioni;
- Non segnala le forme ricostruite accolte a lemma.

Rientra tra questi aspetti anche la scelta di organizzare la descrizione dei significati seguendo lo sviluppo logico semantico.

3.3.2 GATTO e GATTOWeb

Per una serie di motivi, gli sforzi dell'OVI si sono inizialmente concentrati sull'allestimento di un adeguato complesso di *corpora* elettronici.

Nell'idea originaria del progetto per il *Vocabolario storico* avviato dalla Crusca, l'archivio delle fonti primarie doveva essere aperto al pubblico → 1993 pubblicazione in rete ad accesso libero del *Corpus TLIO*.

Il sistema di *Gestione degli Archivi Testuali del Tesoro delle Origini (GATTO)*, è stato progettato per la costruzione, gestione e interrogazione del corpus.

Condividono un solido *framework* di procedure e parametri es. tokenizzazione.

L'applicazione conserva i dati in un *database* relazionale che conferisce solidità, affidabilità e assicura buone *performance* complessive.

L'interfaccia web dispone di quasi tutte le funzioni di interrogazione della piattaforma *desktop* con l'eccezione di quelle dedicate alla creazione e lemmatizzazione del corpus.

Il funzionamento poco intuitivo respinge l'utente inesperto alle prime consultazioni ma la curva di apprendimento di questi programmi è abbastanza rapida.

GATTO e *GATTOWeb* sono stati elementi centrali di un sistema in cui la sinergia tra voci e *corpus* è probabilmente tra le più sviluppate nell'ambito della lessicografia storica.

3.3.3 Trasferimento dei contesti localizzati GATTO a Pluto

Da *GATTO* si estrae la documentazione da prendere in considerazione dal *Corpus TLIO* e dal *Corpus OVI* secondo le *Norme*. Dopo aver ricostruito un elenco completo delle forme pertinenti si ottiene una sequenza di contesti.

Una piattaforma lessicografica digitale evoluta deve puntare a migliorare la comunicazione con la banca dati testuale e favorire i collegamenti tra dizionario e corpus.

In occasione della realizzazione della piattaforma redazionale del *Vocabolario Dantesco* sono stati creati due *parser* per interpretare i contesti in *RTF (Rich Text Format)* accantonati per un limite nel modo in cui erano rappresentate le informazioni. Al fine di non perdere informazione in *output* → codifica XML.

Il gestore dei contesti si apre a partire dalla maschera di redazione e permette di trattare i risultati estratti da *GATTO* rimanendo nello stesso ambiente.

La formulazione delle definizioni su *Lexicad* comincia proprio con l'apertura del gestore dei contesti che farà apparire in contesti disposti in ordine cronologico.

Il redattore deve interpretare il primo contesto visualizzato nel gestore e collegarlo a un nuovo significato: Pluto provvederà a creare il significato su cui può essere formulata la definizione.

Il redattore passa a interpretare il contesto successivo: se questo richiede una nuova definizione si procederà come detto precedentemente, altrimenti si dovrà indicare il significato pertinente per procedere all'associazione.

Per la sistemazione del contesto da citare il lacerto originale estratto da *GATTO* rimane inalterato, mentre il redattore può intervenire su una sua replica in HTML.

Il dato originale può essere consultato nel gestore e ripristinato in qualsiasi momento.

Il gestore dei contesti permette di accedere allo schedario con i contesti archiviati o di visualizzare solo i contesti marcati come "da citare"; per ogni visualizzazione possibile filtrare i contesti.

Per i testi non presenti nel corpus è prevista un'interfaccia di inserimento manuale della citazione.

Il gestore dei contesti rappresenta il collegamento principale tra *DWS* e *CQS*.

Analizzando l'XML delle voci estratte da ReddiX nei corrispondenti punti dedicati alle citazioni, è rilevata la presenza di un solo elemento XML <abbrtit_form> in chi sono presenti:

- a. Abbreviazione del titolo;
- b. Riferimento organico;
- c. Riferimento tipografico.

Nel *VD* la corretta separazione di queste informazioni permette a chi consulta la voce di ristrutturare l'albero dei significati in ordine di occorrenza sulla *Commedia*.

Partendo dalla segnalazione della sigla univoca, la piattaforma integrata può separare e verificare l'abbreviazione del titolo riportata nella citazione attraverso la bibliografia.

3.4 La Bibliografia dei Testi Volgari (BTV)

3.4.1 Repertori e bibliografie in *Lexicad/Pluto* e la BTV

Il *TLIO* ha quattro bibliografie principali:

1. *Bibliografia dei Testi Volgari (BTV)* → raccoglie i dati biografici dei testi indicizzati nei corpora; testi fuori dal corpus citati nel *TLIO* e testi databili entro il 14° secolo. Prima della migrazione in Pluto la bibliografia era archiviata in un vecchio formato di Microsoft Access;
2. *Bibliografia citata nelle voci*;
3. *Bibliografia dei citati* → riconversione in MySQL;
4. *Bibliografia dei volgarizzamenti*.

Nel codice *Lexicad* è presente una forma semplice di bibliografia a "chiave" → usata per la *Bibliografia citata nelle voci* del *TLIO*.

Bibliografie più complesse possono essere implementate come applicazioni autonome e disporre di flessibilità di *framework*.

La *BTV* è stata il primo modulo a entrare in funzione su Pluto, trasformata in applicazione *on-line*, la nuova *BTV* può essere gestita in modo collaborativo.

È stato disegnato un *front-end* con un *template* responsivo dedicato con le funzioni di ricerca e consultazione più comuni.

La *BTV* dialoga attivamente con le voci del *TLIO* ed è capace di esportare i dati in un formato compatibile con *GATTO* per la creazione del *database* bibliografico dei *corpora*.

3.4.2 Il recupero delle schede filologiche

L'indagine filologica costituisce la necessaria premessa a qualsiasi tipo di lessicografia *corpus-based* o *corpus-driven*.

La dinamica tra lessicografia e filologia pare oggi essersi assestata sui ruoli complementari. Questa prassi non era scontata nel momento dell'avvio del *Vocabolario storico* italiano. Pasquali aveva presagito il rischio di generare un circolo vizioso tra lessicografia storica e filologia; il circolo venne spezzato contestandosi delle edizioni esistenti, senza tornare ai manoscritti.

Lo stato di fatto poteva essere accolto nel corpus solo dopo un'attenta valutazione e una puntuale revisione. Nel 1965 → istituito un ufficio filologico con il compito di allestire la tavola dei citati, reperire i testi e prepararli per lo spoglio lessicale.

L'ufficio si è occupato della scelta dell'edizione più attendibile, accertamento dei criteri editoriali e dell'integrazione del testo con una selezione di contributi successivi; attivo nel promuovere la realizzazione di nuove edizioni.

Il lavoro è registrato negli "schedoni" di accompagnamento ai testi.

Grazie al progetto *LIVS* (*Lingua Italiana e Vocabolario Storico: metodi antichi e moderni*), le schede filologiche sono state digitalizzate e riordinate da Zeno Verlato ha organizzato l'archivio con le digitalizzazioni in una struttura di *directory* nominate con una parte fissa concatenata con la sigla *GATTO*. Il rigore nella classificazione ha permesso alla *BTV* in Pluto di poter 'navigare' autonomamente tra le cartelle dell'archivio.

3.4.3 Il modello repertoriale nell'ecosistema dell'OVI e le estensioni dell'applicazione BTV

Prima di essere trasformata in una nuova applicazione per Pluto, la *BTV* in Access è stata sottoposta a un processo di normalizzazione e pulizia che ha rivelato come il *database* bibliografico dell'OVI sia stato assemblato progressivamente a partire da continue aggiunte e appendici dettate dalle esigenze di ricerca 'sul campo'.

Con l'insieme dei suoi archivi testuali, l'OVI ha definito una struttura bibliografica condivisa da tutti coloro che utilizzano i dati dell'Istituto o il programma Gatto.

Il progetto ARTESIA ha dovuto predisporre strumenti di conversione per rendere compatibili i dati del proprio *database* repertoriale degli anni Duemila con i campi della bibliografia accettati da *GATTO*.

L'applicazione *BTV* fornisce adesso un dispositivo di ricerca interoperabile e subito trasferibile su altri progetti. Per la nuova versione del *database* repertoriale di ARTESIA, il riadattamento dell'applicazione *BTV* di Lexicad ha permesso di poter avviare i lavori a partire da un sistema già collaudato, compatibile con il modello dell'OVI e già perfettamente integrato con la piattaforma lessicografica del *VSM*.

3.4.4 Dalla BTV alle edizioni digitalizzate di ItalArt

Il progetto *ItalArt* nella sua prima fase si è concentrato sui *Documenti per la storia dell'arte senese*.

I testi della nuova edizione sono stati indicizzati in un *corpus* interrogabile di GATTO. L'attivazione nel portale di una copia dell'applicazione della *BTV* ha permesso di compilare i dati *on-line* con gli strumenti di controllo ed esportazione già attivi.

Al ruolo di supporto per la raccolta dei metadati è stata aggiunta la funzione di permettere la pubblicazione dell'edizione digitalizzata nella nuova edizione.

Sono stati aggiunti tre ulteriori elementi:

1. Editori WYSIWYG integrato nell'ambiente di redazione;
2. Un *repository* collegato all'archivio con le edizioni semi-diplomatiche dei testi in formato PDF;
3. Un *plugin* di terze parti per la visualizzazione di gallerie di immagini.

Il sistema è scalabile, in previsione dell'allargamento del *corpus*.

ItalArt rientra tra le risorse dell'OVI connesse alla filologia digitale.

3.4.4 Alternative per la gestione di edizioni digitali su Lexicad/Pluto: le collezioni Europeana e RdP

Grazie alla maggiore flessibilità e alla pubblicazione in rete, la nuova *BTV* è già stata impiegata come strumento di supporto per l'esportazione dei metadati in formati *standard*.

Attraverso un insieme di *query*, sono state allestite sulla *BTV* alcune pagine di *report* per selezionare i testi e intraprendere un processo di conversione automatica.

Le raccolte di testi sono state disaggregate dal sistema e gestite come singoli testi e, alla fine, è stata persa la corrispondenza tra *BTV* e i testi delle due collezioni.

Questa procedura e la *BTV* non costituiscono approcci realmente alternativi.

Il convertitore è stato sviluppato con la denominazione interna di *BTO* → *Biblioteca del Tesoro delle Origini*.

Lexicad è inoltre stato utilizzato per il progetto *RdP* (*le rime disperse di Petrarca: l'altra faccia del Canzoniere*) per la pubblicazione *on-line* dell'edizione critica e commentata delle rime disperse.

Questo sistema sviluppa il gestore bibliografico di *default* in Lexicad, estendendone la classe di base.

Pur non disponendo degli strumenti di integrazione con GATTO presenti nella *BTV*, la piattaforma di *RpP* ha raggiunto una perfetta integrazione con il *Corpus RdP* e il *Corpus OVI dell'italiano antico* per mezzo di uno *script*.

3.5 Oltre Pluto

3.5.1 La DTD di ReddiX come "grammatica del dizionario"

La DTD di Reddix raccoglie l'inventario degli elementi ammessi nella microstruttura del *TLIO* in formato *machine-readable*.

Secondo il concetto di *dictionary grammar*, la DTD del *TLIO* sembrerebbe costituire un componente di controllo logico sufficiente a descrivere la microstruttura del *Tesoro* ma la DTD di ReddiX non è in grado di catturare alcuni dettagli essenziali nella microstruttura.

Nella progettazione del sistema PLUTO, ulteriori motivazioni concrete suggeriscono di riportare l'analisi su *file XML*.

Inoltre, sela *dictionary grammar* serve per imporre dei vincoli, nel caso del *TLIO* il processo è stato invertito: la DTD è stata desunta dalla configurazione dei documenti consegnati dai redattori, col risultato che alcune strutture presenti nei *file* ma non esplicitamente codificate potrebbero essere sfuggite all'analisi.

3.5.2 Filiera del trattamento e strumenti di analisi

Dai *file* in formato .doc → *file* XML.

Il formato ReddixML è considerato come lo stato di fatto da cui partire: i *file* codificati in ReddixML vengono acquisiti da Pluto come *file* XML ben formati.

L'applicazione, che ripercorre tutte le gerarchie XML delle voci per importarle su Pluto, è chiamata a effettuare la mappatura tra gli elementi del ReddixML e i campi del *database* di Pluto.

Il processo prevede l'estrazione del contenuto grezzo dei *file* che viene registrato su un *database* di lavoro. Al momento dell'estrazione vengono registrati l'identificativo univoco della voce così come riportato dal *file* di Reddix.

3.5.3 Importazione e revisione sistematica

La prima versione *beta* di Pluto rendeva accessibile in rete un blocco di circa mille voci. L'implementazione nel caso del *TLIO* si è configurata come un'impresa collettiva e complessa.

Quando sono state avviate le ricerche sul nuovo DWS, il *TLIO* era stato già completato per due terzi.

L'elaborazione digitale finisce per far emergere incongruenze e imperfezioni nascoste tra le voci. I *file* XML prodotti da Reddix sono validati per il loro aspetto strutturale, in conformità con lo schema impostato nel progetto ma l'importazione in Pluto ha aggiunto ulteriori controlli:

1. Un primo gruppo di verifiche comprende controlli tecnici che rilevano piccole incompatibilità tra il formato XML e il sistema di importazione del DWS;
2. Un secondo tipo di controlli fa uso degli stessi *script* che la piattaforma integra nei campi della maschera di redazione.
3. L'ultimo gruppo di procedure è derivato dallo studio di nuovi algoritmi volti a estrarre quanta più informazione implicita dalle voci del *TLIO*.

Quando si applicano queste verifiche sui *file* contenenti molte voci, la conversione innesca un ciclo di revisione.

Terminata l'importazione su Pluto, gli errori rilevati dai controlli vanno verificati e corretti.

I tempi non trascurabili delle conversioni spesso si allungano per le frequenti anomalie che si generano. Le

correzioni, per motivi di sincronizzazione delle modifiche, sono sempre riportate sui *file* Word originali.

La piattaforma Pluto è tutt'oggi impegnata in questa lunga fase di importazione e correzione delle vecchie voci. Sebbene Pluto abbia concentrato finora la maggior parte delle attività sul modulo di importazione, lo sviluppo del sistema è proseguito su tutte le funzioni dell'applicazione.

Queste nuove realizzazioni sono già entrate in esercizio su tutte le implementazioni parallele del sistema.

4. IMPLEMENTAZIONI SECONDARIE: VSM, AGLIO, VD, VDL, VEV

4.1 L'architettura modulare

La doppia articolazione del DWS tra un livello linguistico e lessicografico generale e un livello di personalizzazione sovraordinato è diventata sempre più rigorosa col susseguirsi delle versioni in quanto funzionale.

2018 → PlutoVD: derivato di Pluto anch'esso basato su Lexicad, progettato per il *Vocabolario Dantesco*. Tutta la piattaforma è stata organizzata come un insieme di 'applicazioni' semi-autonome che possono essere trasportate da una piattaforma all'altra.

La moltiplicazione dei contesti applicativi produce una naturale frammentazione delle attività e delle energie su diversi fronti di ricerca ma, rende necessarie determinate ottimizzazioni del codice. Quanto più si allarga il dominio applicativo che il modello deve riuscire a comprendere, tanto meno è possibile ricorrere a un approccio empirico.

La comune adozione di un modello metodologico unitario garantirà lo scambio e il trasferimento virtuoso di metodi e strumenti.

4.2 Il Vocabolario del Siciliano Medievale (VSM)

4.2.1 Il progetto ARTESIA e il VSM

Nel ventennio che va dal *desideratum* di Ruffino, alla concreta definizione del progetto del *Vocabolario del Siciliano Medievale*, il progetto ARTESIA (*Archivio Testuale del Siciliano Antico*), ha trasformato lo stato dell'arte degli studi sul volgare siciliano.

Il *corpus* si configura come un insieme sufficientemente rappresentativo delle varie tipologie testuali in volgare siciliano, dalla prima attestazione della fine del XIII secolo sino alla prima metà del XVI secolo.

Un *corpus* di oltre 1m di occorrenze. La prima idea di Ruffino, che prevedeva di basare il lessico sui glossari della *Collezione di testi siciliani dei secoli XIV e XV* edita dal Centro di studi filologici e linguistici siciliani, ha lasciato posto a quella di un dizionario di prima mano. → costruito sul modello del *TLIO*, il VSM è concepito come un vocabolario elettronico *born-digital* e fa uso di un *corpus* elettronico interrogabile in GATTO.

4.2.2 La microstruttura del VSM

Il VSM è stato tra i primi progetti a definire la sua microstruttura sulla falsariga di quella del *TLIO*, limitando il più possibile le personalizzazioni, affinché i due modelli di voce fossero in larga parte sovrapponibili. Il VSM dimostra come un'estesa compatibilità tra le implementazioni possa agevolare considerevolmente la sostenibilità dei progetti.

Alcune API orientate alle *performance* possono essere trasferite da un sistema all'altro con un copia/incolla del codice.

Una delle particolarità del VSM è il consistente numero di campi supplementari nella sezione 'Corrispondenze' che documentano la presenza del lessema nella lessicografia siciliana per mettere in evidenza elementi di continuità/discontinuità dal volgare siciliano al dialetto.

4.2.3 Il prototipo di redazione

Per i lavori per il *Vocabolario del Siciliano Medievale* la prima bozza di microstruttura è stata codificata direttamente sugli *standard* della TEI.

Tenuto conto della prevedibile difficoltà dei redattori non esperti nella codifica con un linguaggio di *markup*, la definizione di un modello formalizzato è stata accompagnata dall'allestimento di una maschera di inserimento guidato delle voci. La maschera serviva a nascondere la complessità della banca dati dietro un'interfaccia intuitiva.

Il pannello di redazione era costituito da un modulo, realizzato con un *design* responsivo e interattivo, per mezzo del quale in redattore poteva scrivere la voce e visualizzarne un'anteprima impaginata nello stile 'Tutto/Stampa' del *TLIO*.

La maschera di redazione è stata da subito concepita come un sistema *web-based* basato su tecnologie *open source* dello *stack LAMP*, così da poter funzionare sulla maggior parte dei servizi di *hosting* senza troppe pretese tecniche.

4.2.4 Lexicad per le banche dati repertoriali

Il progetto ARTESIA ha realizzato negli anni 2000 una banca dati repertoriale sul modello di *TLIon* (*Tradizione della Letteratura Italiana online*) sviluppata anche grazie alla collaborazione diretta con due progetti della rete *TLIon*, *CASVI*.

Così come per molti preziosi *database* di quel periodo, ARTESIA ha condiviso con *TLIon* il modello e anche la sorte, diventando rapidamente obsoleto.

Tutte le funzioni generali di Lexicad/Pluto sono utilizzabili anche su una banca dati repertoriale, con gli stessi vantaggi in termini di integrazione osservati sul *TLIO*.

Il repertorio ARTESIA è stato reso conforme allo schema di base della BTV.

Gli autori sono stati sostati su schede dedicate; i *record* delle opere sono stati messi in relazione con i manoscritti e con gli altri elementi previsti nel primo portale.

Con l'integrazione tra repertorio e *DWS*, i percorsi di esplorazione e i nessi logici tra gli elementi potrebbero essere ulteriormente sviluppati. Per rendere realizzabile questa possibilità sarà necessario superare un limite tecnico del CQS.

4.3 L'Atlante Grammaticale della Lingua Italiana delle Origini (AGLIO)

4.3.1 Atlanti linguistici e Lexicad

Il primo riadattamento del codice del prototipo del *VSM* non ha riguardato un vocabolario in senso stretto ma l'*Atlante Grammaticale della Lingua Italiana delle Origini* (AGLIO) → idea di costruire una sorta di prospetto grammaticale del *Corpus OVI* o un atlante grammaticale per l'italiano antico.

Le fonti sono i testi linguistici rappresentativi per una determinata varietà e filologicamente sicuri del *Corpus TLIO*.

Per l'analisi → idealmente due fasi:

1. Marcatura dei tratti morfologici
2. Relazione tra la forma e entità della banca dati: lemma e tratti fonologici.

Il sistema informativo dell'*AGLIO* comprende un insieme di lemmi a cui sono collegate le forme attestate; l'entità che accoglie i risultati dell'analisi è la forma.

La struttura dei dati dell'*AGLIO* è costituita da pochissime entità elementari collocate in una fitta serie di rimandi; organizzare i dati in numerosi percorsi di consultazione e di interrogazione: i primi aggregano automaticamente le informazioni in indici e tabelle; i secondi permettono all'utente di impostare e lanciare ricerche complesse sulle forme e aree.

Le forme individuate dall'*Atlante* sono di *default* ordinate alfabeticamente, ma possono anche essere aggregate per informazioni:

- Geografiche;
- Aree e testi;
- Paradigmi;
- Aree e paradigmi.

La corrispondenza tra le interfacce del *back-end* e quelle del *front-end* è limitata al caso della scheda forma.

4.3.2 Rappresentazioni cartografiche

Il dato geografico si crea nella relazione di occorrenza tra forma e testo.

Le aree in formato stringa di *GATTO* e nella *BTV* non prevedono ulteriori specificazioni per il trattamento cartografico e informatico.

In un primo momento, i valori possibili per ciascuna area sono stati mappati su un vettore di chiavi numeriche ordinate. Per ciascuna delle tre tipologie di area è stata predisposta una tabella per la conversione della stringa con l'abbreviazione di un codice numerico.

La prospettiva geolinguistica è stata sviluppata in direzione di una compiuta rappresentazione cartografica.

Il progetto *MIRA* prevede due operazioni di mappatura:

1. l'implementazione di alcune interfacce cartografiche per l'accesso ai dati dell'*AGLIO*;
2. una serie di carte linguistiche 'secondarie', che interpretano gli esiti di un centinaio di dati linguistici.

Per il punto 1. di pertinenza del *DWS*, è ragionevole ipotizzare l'implementazione di una possibile carta per ogni *output* del livello prestazionale che riporta dati areali. Per raggiungere lo scopo: ulteriore tabella di geolocalizzazione.

Parte di questa ricerca di metodi e strumenti geolinguistici è stata ripresa in un progetto presso il Notre Dame Center for Italian Studies.

Generalizzare l'applicazione di questi metodi, estendendoli dall'*AGLIO* a tutti gli studi sull'italiano delle origini.

Il centro ideale per elaborare e archiviare le indicazioni di geolocalizzazione è la *BTV*.

L'*OVI*, come promotore di una standardizzazione sui metadati per l'italiano antico, può garantire la completezza dell'informazione sulla base del continuo aggiornamento dei *corpora*.

Il sistema standardizzato di metadati è indipendente dagli strumenti di visualizzazione e i progetti che decideranno di aderire a questo sistema potranno utilizzare diverse combinazioni di tecnologie.

4.4 Il Vocabolario Dantesco (VD) e il Vocabolario Dantesco Latino (VDL)

4.4.1 Lessicografia dantesca e metodi computazionali

Le prime significative applicazioni tecnologiche alla lessicografia dantesca risalgono almeno al 1965. In quell'anno venne pubblicato l'indice di frequenza della *Commedia*, nell'ambito del progetto per un *Inventario Linguistico dell'Italiano delle Origini*.

La pubblicazione delle concordanze della *Divina Commedia*, a cura di Carlo Tagliavini, aprì la strada a ulteriori sperimentazioni per la lingua italiana: l'IBM affidò a Tagliavini la redazione del *Lessico di frequenza della lingua italiana contemporanea*.

Le *Concordanze Dante/IBM* sono state realizzate in tempi eccezionalmente brevi.

Negli anni successivi le iniziative digitali dedicate a Dante si sono moltiplicate.

Tra le caratteristiche che distinguono questo filone di ricerca, andrebbe comunque segnalata la singolare longevità di alcuni tra i progetti più importanti, come il *Dartmouth Dante Project* e il *Princeton Dante Project*; iniziative italiane: es. Società Dantesca Italiana → *Hypermedia Dante Network* che si propone di estendere alla *Commedia* per un'enciclopedia dantesca digitale.

4.4.2 PlutoVD per il Vocabolario Dantesco

Il *Vocabolario Dantesco* aggiunge a questa tradizione un dizionario elettronico *on-line* progettato sul modello del *TLIO*.

A differenza di Pluto, l'implementazione gemella per il VD non ha dovuto fare i conti con una redazione già avviata e con il problema della riconversione dei dati preesistenti.

Pluto VD ha quindi portato in rete per la prima volta quella linea di ricerca dell'OVI che risale alla procedura COVIRE.

Il lavoro sui *corpora* ha previsto la creazione di due nuovi archivi in GATTO: il *Corpus dei testi volgari di Petrarca e Boccaccio* e il *Corpus dei Testi fiorentini in prosa del sec. XIII*.

Questo meccanismo ha consentito di rendere dinamica la sezione 'corrispondenze'.

Presentazione del *Vocabolario Dantesco* → 2018.

La presenza di entrambe le piattaforme in rete ha permesso di attivare alcune funzioni e API dimostrative per sondare le possibili interazioni tra banche dati lessicografiche.

Nelle prime sperimentazioni è stato possibile anche agganciare il testo della *Commedia*, caricato su un blog esterno al VD.

4.4.3 Il Vocabolario Dantesco Latino (VDL)

Il *Vocabolario Dantesco Latino* si propone di costruire il primo sistematico strumento di conoscenza dedicato al lessico latino dantesco, attraverso strumenti e procedure analoghe a quelle messe in atto dal *Vocabolario Dantesco*.

Il VDL beneficia dell'indipendenza del gestore dei contesti dal CQS, perché sceglie di usare come CQS il performante *DanteSearch*, che rende disponibile *on-line* il *corpus* lemmatizzato di tutte le opere di Dante. L'interazione con *DanteSearch* è stata possibile nell'abito della collaborazione con l'Istituto di Scienza e Tecnologie dell'Informazione.

Il redattore può inserire la chiave di ricerca direttamente nella maschera di redazione, che provvederà a contattare la API in tempo reale e a presentare un'anteprima dei dati al redattore.

Non è necessario collegarsi all'indirizzo del CQS, né generare un *file* intermedio di esportazione: l'interrogazione per lemmi su *DanteSearch* diventa, in sostanza, disponibile nella stessa maschera di redazione di Lexicad.

Una seconda API ha permesso di effettuare dalla maschera ricerche per lemmi nel *Corpus Corporum*. Per mezzo di un secondo *script* di importazione, i risultati delle ricerche effettuate sul *Corpus Corporum* potranno essere copiati in un campo di testo libero corredato da un *editor* visuale.

La voce è sottoposta a una sequenza ordinata di passaggi di stato attraverso 5 fasi. Terminato il lavoro, ogni membro della redazione sottopone la voce ai responsabili del livello successivo, facendone avanzare lo stato e perdendo i privilegi di accesso.

1. Redazione → la scheda viene compilata dal redattore che assegna la voce al proprio revisore di riferimento.
2. Revisione → si effettua una correzione sostanziale sulla scheda.
3. Approvazione → il Comitato di redazione effettua una revisione formale della voce.
4. Approvata → in attesa del *placet* del Consiglio scientifico per la pubblicazione.
5. Pubblicata → dove diventa visibile ai lettori sulla piattaforma.

4.5 Il Vocabolario storico-etimologico del veneziano (VEV)

Il *Vocabolario storico-etimologico del veneziano* si discosta leggermente dalle procedure di analisi già previste dal modello Lexiad.

Il VEV si serve di diversi tipi di fonti:

1. *Corpus VEV*
2. *Corpus lessicografico*
3. *Altre fonti* (raccolta)
4. *Corpus delle fonti lessicografiche generali*
5. *Testi in veneziano*

Per il primo punto è stato possibile procedere come di consueto; per gli altri punti è stata sviluppata un'applicazione su Lexicad.

Nel progetto VEV l'aspetto tecnico viene ridimensionato, e viene esperito un approccio lessicografico meno radicale. Quattro volumi raccolgono voci tematicamente correlate:

- Venezianismi dell'italiano
- Ingiurie
- Improperi
- Contumelie
- Istituzioni della Serenissima
- Giochi e passatempi

Con la piattaforma del VEV Lexicad si è parzialmente staccato dal rigido approccio del *database* lessicale fortemente strutturato.

Il risultato ha prodotto una maschera di redazione che può essere considerata un *editor* lessicografico evoluto per la pubblicazione *on-line* delle voci in forma ipertestuale.

5. SPERIMENTAZIONI, APPLICAZIONI PROSPETTIVE

5.1 La macrostruttura nel dizionario elettronico

5.1.1 Il *The Random House Dictionary of the English Language*

Sebbene Busa sia riconosciuto come il fondatore dell'informatica umanistica e della linguistica computazionale, l'atto di nascita della lessicografia elettronica si posticipa alla pubblicazione del *The Random House Dictionary of the English Language*.

Né lessicografi nel corso della redazione, né i lettori hanno mai avuto esperienza diretta del calcolatore, amministrato esclusivamente da tecnici specializzati.

Con il *RHE* si verifica un cambiamento nel processo lessicografico: sette tipi di unità elementari discrete:

1. Illustrazioni
2. Intestazioni delle voci
3. Definizioni
4. Varianti
5. Etimologie
6. Lemmi senza definizione
7. Informazioni supplementari.

Le definizioni del *RHE* sono state poi riclassificate su oltre 150 soggetti tematici, utilizzati per distribuire il lavoro a specialisti della materia e consulenti esterni.

Nel periodo in cui è stato pubblicato il *RHE* l'ordinamento dinamico non avrebbe potuto esprimere ancora il suo pieno potenziale.

Pfister stima i vantaggi e i costi del trattamento elettronico rapportandoli a un'unica operazione di ordinamento delle schede, tant'è che approva l'informatizzazione del *TLF* proprio perché, nel caso del dizionario allora diretto da Paul Imbs, le schede sono state riclassificate una seconda volta per il calcolo delle frequenze.

5.2.1 *Der Kampf um die Gestaltung des Wörterbuchs*

La "tradition-sanctioned orthographic supremacy" si affermi prima dell'invenzione della stampa. L'ordinamento razionale dell'universo e la conoscenza sopravviveranno nei dizionari enciclopedici, che nel Rinascimento italiano si affermano con il nome di 'fabbrica del mondo'.

L'ordinamento semasiologico ha basato il suo successo sulla presunta efficacia come criterio convenzionale di accesso all'informazione.

Baldinger afferma che l'ordine alfabetico non è realmente più efficace di quello concettuale ma che si tratta di organizzazioni macrostrutturali che rispondono a domande diverse.

Nonostante la posizione consolidata dell'impostazione semasiologica, la competizione tra i due modelli di macrostruttura non si è spenta ed ha acceso uno scontro tra difensori dell'organizzazione semasiologica e sostenitori di quella onomasiologica.

Nella lessicografia tradizionale, la prossimità grafematica e quella concettuale hanno costituito i termini d'un'inconciliabile opposizione binaria. L'approccio sperimentale del *RHE* produce per la prima volta uno strappo nell'antico ordine alfabetico.

5.2 La macrostruttura in Lexicad / Pluto

5.2.1 La macrostruttura come dispositivo di accesso

La macrostruttura rientra tra gli elementi del vocabolario dipendenti dal *medium*: tavole e indici tradizionali vanno intesi come espedienti messi in atto dal libro per permettere strategie di accesso alternative rispetto a un'organizzazione principale. Nei dizionari sviluppati su Lexicad/Pluto la macrostruttura così intesa non ha motivo di esistere.

Tutti i dispositivi di accesso alle voci costituiscono solo delle proiezioni virtuali del dizionario elettronico.

Micro- e macrostruttura in un *database* lessicale sono entrambi concetti dinamici, ma la macrostruttura si definisce solo a partire da campi appartenenti alla microstruttura attraverso:

1. l'interrogazione di una base di dati con un linguaggio di interrogazione strutturato (SQL - *Structured Query Language*)
2. Il possibile intervento di uno o più algoritmi implementati sulla piattaforma lessicografica.

Con lemmario sarebbe opportuno intendere:

1. L'insieme dei *record* contenuti nella tabella 'voce' della base di dati individuati da un identificativo univoco → qui i *record* non contengono la microstruttura nella sua interezza, ma costituiscono l'unità lessicografica e non modificabile. Il lemma è registrato nella tupla della voce ma quest'ultima può contenere più lemmi alternativi.
2. Una specifica interfaccia di accesso che fornisce un indice di voci.

Le entità identificate da queste due accezioni forniscono insiemi di entrate che non coincidono → il lemmario come indice può escludere dall'elenco elementi che non sono vere e proprie voci ma che comunque dispongono di un *record* nella tabella delle voci.

Il sistema Lexicad dispone di diverse interfacce già pronte tra cui un'API che è diventata lo strumento più usato dagli stessi moduli che compongono Pluto/Lxicad.

Nel pannello di redazione, l'interfaccia principale è costituita dalla pagina di gestione della voce. Quando è necessario introdurre nuova informazione funzionale alla macrostruttura, il modello Lexicad privilegia ugualmente l'introduzione di nuovi attributi a livello di microstruttura.

Solo in casi particolari si preferirà il ricorso a pagine statiche che non interferiranno con la capacità del *database* di generare infiniti percorsi di accesso alternativi a partire dalle info del *database* lessicale.

5.2.2 Il flusso redazionale e l'esempio del Diccionario Histórico de la Lengua Española (DHLE)

La riorganizzazione della macrostruttura può essere utile anche e soprattutto in fase di redazione. L'avanzamento della stesura per gruppi di parola affini, oltre a rendere più flessibile e sostenibile il processo redazionale, conferisce coerenza alle definizioni.

Dopo i due progetti dall'impostazione tradizionale la RAE (Real Academia Española) ha rilanciato l'impresa come dizionario *born-digital*. Ha scelto di non procedere in ordine alfabetico, ma sulla base delle relazioni semantiche tra lemmi.

Per assicurare l'omogeneità del trattamento, la descrizione semantica dei lessemi prevede una sequenza di fasi, finalizzate a impostare una struttura definitoria di riferimento per ciascun raggruppamento → vengono prima predisposti alcuni schemi di definizione provvisori; questi vengono

successivamente utilizzati per elaborare le prime definizioni dei lessemi che rientrano nel raggruppamento semantico.

Le strutture del dizionario elettronico hanno una natura pluridimensionale.

L'organizzazione del lavoro con una scansione concettuale potrà conferire ai dati del vocabolario una dimensione onomasiologica, che lo renderebbe esplorabile per percorsi iperonimici o iponimici e per ambiti tematici. Un'organizzazione per campo semantico si può ottenere su Lexicad/pluto in due modi:

1. Etichettatura preventiva delle entrate
2. Conferendo i privilegi di scrittura alle credenziali di accesso di un determinato utente.

5.2.3 *Dinamismo degli indici: rinvii, derivati e composti*

Nella ricerca su Lexicad/Pluto, una prima sollecitazione all'elaborazione dinamica dei lemmari è stata fornita dai tipi particolari di voci. Il caso elementare sono i rinvii previsti da tutti i vocabolari costruiti sul modello del *TLIO*.

In Lexicad il rinvio è concepito come una voce al pari di tutte le altre → la scheda nel *back-end* è in grado di accogliere tutte le informazioni e le annotazioni previste dalla maschera di redazione.

La differenza nei comportamenti tra voce e rinvio viene stabilita registrando l'identificativo univoco di una voce alla quale si intende rinviare.

È stato predisposto un campo dinamico 'Rinvia a', che, se viene compilato dal redattore utilizzando la casella di ricerca istantanea, trasforma virtualmente la voce in rinvio. Da questo momento, tutte le informazioni presenti nella scheda vengono considerate rilevanti esclusivamente ai fini della redazione.

La visualizzazione del rinvio negli indici del *front-end* è stata resa facoltativa e l'utente può decidere di disattivarla, intervenendo su un controllo dell'interfaccia.

Il *VEV* invece prevede anche rapporti di derivazione con voci speciali, marcate con l'etichetta 'derivato o composto semplice'.

Un aspetto singolare delle entrate di solo derivato o composto è che, mentre sul *database* e nell'ambiente di redazione vengono viste come voci autonome, nelle strutture lessicografiche di consultazione possono comparire esclusivamente all'interno di altre voci con cui si trovano in relazione.

Un elemento formalmente autonomo del lemmario può essere incapsulato all'interno della microstruttura di una o più voci di cui finisce per far parte.

5.2.4 *Applicazioni alla varia lectio della Commedia per il Vocabolario Dantesco e il lemmario per frequenze*

Il *Vocabolario Dantesco* si propone di studiare il lessico di Dante, tenendo conto anche di una selezione della *varia lectio* presente negli apparati delle edizioni di riferimento.

L'apparato *Petrocchi* affianca al *Corpus VD* nello spoglio.

Nei principi di schedatura degli apparati del *VD*, si parla di criterio di selezione filologico delle varianti. Per essere accolte nel vocabolario, le varianti dovranno anche essere latrici di un lemma o di un significato non altrimenti attestato nel testo dell'edizione *Petrocchi*.

Sulla base di quest'ultimo criterio, l'apporto delle varianti può concretizzarsi in due modi:

1. Le varianti possono dare luogo a un lemma non altrimenti attestato nel *VD*;

2. Le varianti possono contribuire all'informazione lessicale dal punto di vista semantico o grammaticale.

Il rapporto tra i due lemmi costituisce una relazione di tipo molti-a-molti perché un lemma a testo può avere più varianti in apparato, e la stessa variante può occorrere più volte in corrispondenza di diversi lemmi a testo.

Nel sistema, la «variante-nuovo lemma» non è esplicitamente marcata come tale, ma viene riconosciuta dinamicamente dagli algoritmi della piattaforma in quanto voce che soddisfa due condizioni: non è una voce di rinvio e la somma delle sue occorrenze nella *Commedia* è pari a zero.

Una seconda sperimentazione sui lemmari del *VD* è stata permessa dall'impiego del gestore dei contesti nella redazione.

La facilità e velocità d'importazione dei contesti e dei relativi metadati, unite al numero gestibile dei materiali oggetto di spoglio, hanno consentito di conservare sulla piattaforma i contesti importati e associati contestualmente alla voce.

Queste frequenze, nella rappresentazione di Pluto *VD*, sono conservate su undici campi: due per ciascuna cantica e cinque per altre opere volgari. Sulla base di questi dati puntuali è stato progettato un 'lemmario di frequenze' che fornisce un prospetto sintetico del rango di ogni lemma nella *Commedia*.

Questo lemmario presenta le entrate per rango decrescente sull'intera opera e fornisce all'utente la libertà di scegliere l'ordinamento.

5.3 Tra lemmario e dizionario macchina

5.3.1 La forma del lemma e il caso della lessicografia del *moyen français*: il lemmatizzatore *LGeRM*

Con il *Godefroy* e il *Tobler-Lommatzsch*, i due dizionari di riferimento per l'antico francese, i lettori sono stati per molto tempo costretti a districarsi tra differenti criteri, comunque legittimi, per l'individuazione delle entrate.

Le prime edizioni elettroniche non hanno tentato di risolvere questo problema ma accettare solo la chiave di ricerca esatta.

Il *Dictionnaire du Moyen Français (DMF)* accoglie a lemma, quando disponibile, la forma del francese moderno; l'*Anglo-Norman Dictionary (AND)*, invece, sceglie di indicare la grafia più attestata nelle fonti. Dal punto di vista del *DMF* si ritiene che la casualità della grafia più frequente finisca per creare una nomenclatura rappresentativa ma graficamente inconsistente.

Il problema dell'accesso si risolve solo in parte con la scelta del lemma, perché il lettore non specialista potrebbe riscontrare difficoltà nel risalire alla grafia moderna a partire dalla forma di cui è a conoscenza. La variazione grafica nel *DMF* è stata trattata anche con lo sviluppo di *LGeRM*.

Il lemmatizzatore del *DMF* è stato integrato nel sistema di pubblicazione del vocabolario per guidare l'accesso alle voci. Un lemmatizzatore per la lessicografia dell'antico francese e del medio francese deve confrontarsi con una marcatura polimorfia.

Il primo elemento, il repertorio che contiene le forme grafiche ricondotte ai rispettivi lemmi, è stato costituito primariamente a partire dalle voci del *DMF*. Osservando la codifica in XML si può notare come le forme siano marcate con l'elemento <OCC>, rendendone possibile e affidabile l'estrazione automatica.

I dati sono stati ulteriormente arricchiti attingendo ai riferimenti del *DMF*, al *Tobler-Lommatzsch* e al *Godefroy*.

5.3.2 Considerazioni sull'approccio *thesaurus-based* e rappresentatività del corpus

Souvay/Pierrel dichiarano che la complessa soluzione degli algoritmi di trasformazione di *LGeRM* è stata sollecitata dai limiti di un approccio *thesaurus-based* sulle lingue antiche.

In nessuna risorsa lessicale si può sperare di trovare rappresentata l'intera variabilità grafica delle scritture del Medioevo, ma ci sono differenze sostanziali nelle risorse lessicografiche di cui disponiamo, che rendono diversa la situazione dell'italiano delle origini da quella del francese medio.

Le fonti primarie di *DMF* sono organizzate in tre archivi di natura eterogenea: i testi integrali, parziali e edizioni critiche.

Anche il *TLIO* ricorre a più fonti di informazione linguistica ma dispone di un nucleo principale costituito da documentazioni integrali.

Per il *TLIO* è ipotizzabile un rapporto lemma / forme molto diverso, a causa della maggiore variabilità grafica nell'italiano delle origini rispetto al francese moderno.

La consistente mole di documentazione raccolta sembra di per sé sufficiente a certificare una buona rappresentatività della collezione testuale.

È sembrato sentato costruire, sviluppando il *DWS*, una risorsa che raccoglie tutta l'informazione estraibile dal punto 0.1 delle voci del *TLIO*. L'obiettivo è fornire all'infrastruttura di consultazione di Pluto l'informazione necessaria per facilitare l'accesso alle voci.

5.3.3 Costruzione di un dizionario macchina del *TLIO* in XML-TEI

Dalle intestazioni delle schede del *TLIO* è possibile estrarre un dizionario macchina, cioè un insieme strutturato di coppie forma-lemma, più ricco e raffinato di quello del *Corpus*.

Tutta l'informazione è già virtualmente presente nella base di dati di Pluto ma, per il suo riutilizzo, si è provveduto a rappresentarla in un formato *machine-readable*, conforme a uno *standard* per la codifica dei dizionari. Il formato XML di *Morphalou* è sovrapponibile alla proposta di marcatura TEI sviluppata per il *VSM*.

Per pubblicare il lessico si è scelto di ricorrere a un'API in grado di dialogare con altri agenti informatici. L'API costruisce e percorre il lemmario escludendo le entrate di solo rinvio e le entrate di derivati e composti.

Per ottenere i lemmi a partire dalle forme è disponibile l'*endpoint* 'formelemmi' che accetta gli stessi parametri 'lemma' e 'filtro'.

Rispetto a *Mprphalou*, l'aspetto morfologico è chiaramente penalizzato; la specificazione dell'attributo *type* è limitata, perché nel formario del *TLIO* la natura della forma non è esplicitata e non sarà possibile distinguere tra varianti grafiche e forme flesse, secondo i valori suggeriti dalle *TEIP5-Guidelines*.

Questo lessico elettronico in XML può essere accostato al dizionario macchina di *GATTO* composto dai lemmi muti.

Le associazioni forma-lemma sulle voci meno recenti sono meno complete di quelle censite dalle versioni del *corpus* pubblicate dopo la redazione della voce.

Per favorire l'interoperabilità, l'identificativo usato nell'elemento <entry> non è l'*id* univoco del *database* di Pluto, ma corrisponde all'attributo *id* rilevato nell'XML del *TLIO* di ReddiX.

5.4 La dimensione onomasiologica

5.4.1 Un «lemmario interlinguistico» per lo studio dell'affettività nella lirica romanza

Lo studio diacronico del lessico su domini linguistici diversi e attraverso molteplici risorse digitali ha sollevato alcune questioni che hanno interessato direttamente il *TLIO*, il *Corpus TLIO* e *GATTO*. Il progetto *L'affettività lirica romanza: lemmi e temi*, ha elaborato una classificazione lessicale sulle emozioni, con una metodologia spiccatamente comparativa in ambito romanzo ed europeo.

Il progetto ha affidato la connessione tra i domini a un modello semantico che è stato indicato come chiave di volta per la comparazione, integrazione e interoperabilità di risorse eterogenee.

Due strumenti di riferimento valutati nel corso delle fasi iniziali:

1. *Wordnet*
2. *Begriffssystem* di Hallig/Wartburg.

Alla fine, il progetto ha optato per una struttura modellata sulle banche dati già esistenti, prendendo come base di partenza il modello teorico del *Geneva Affect Label Coder* di Sherer. Sulle cinque 'categorie emozionali' è stata costruita l'articolazione principale della classificazione semantica: i cinque settori sono stati articolari in 16 ulteriori nodi denominati *synset*.

Nell'ambito del progetto, l'OVI ha costruito il *LirIO*, un *corpus* di poesia italiana anteriore al Quattrocento con 77.490 occorrenze per circa mille lemmi. Il *corpus* è stato lemmatizzato esaustivamente e le categorie emozionali, i *synset* e i macrolemmi sono stati implementati con il sistema degli iperlemmi di *GATTO*, rispettivamente di terzo, secondo e primo livello.

Con l'applicazione della classificazione alle entrate del *TLIO* le voci vengono introdotte in una rete che permette un'organizzazione onomasiologica interna del lemmario. Questo metodo rischia tuttavia di dimostrarsi poco scalabile.

A livello tecnico non ci sarebbero impedimenti ad associare un lemma a più punti della gerarchia, ma il consistente numero di associazioni finirebbe per creare contraddizioni logiche che vieterebbero di sviluppare ulteriormente la rete semantica come un'ontologia.

A livello lessicografico, una corretta e più funzionale associazione della voce a una struttura semantica, adatta soprattutto all'applicazione su ambiti semantici più ampi, dovrebbe essere istituita a partire da ciascuna definizione presente nella voce e non dal lemma.

5.4.2 L'Historical Thesaurus of English (HTE)

Con l'*Historical Thesaurus of English (HTE)* è stato costruito un tesoro strutturato dei significati delle parole della lingua inglese, lungo cronologicamente dalle origini ai giorni nostri. Due sezioni:

1. Analisi dei significati dell'*OED*;
2. Dedicata al periodo 700-1150 d.C.

L'organizzazione del tesoro utilizza un'impalcatura gerarchica che si articola a partire da tre nodi sovraordinati, chiamati '*mega-catego-*': '*External World*', '*Mental World*' e '*Social World*'. Le tre sezioni primarie sono divise in 26 campi semantici maggiori, che si ramificano in ulteriori sottocategorie, procedendo dai concetti più generali ai più specifici. A ciascuna categoria sono associati un titolo e un codice numerico che viene generato ricomponendo il percorso che sulla tassonomia conduce a quel punto.

La profondità massima della gerarchia può arrivare a 7 livelli. Ciascuna accezione di una parola può avere una o più categorie d'appartenenza.

È la relazione di iperonimia/iponomia che regola l'organizzazione strutturale in questo tipo di tassonomie. I termini che si collocano all'interno di una singola categoria sono considerati co-ponimi o sinonimi.

I nodi terminali costituiti dai lemmi vengono presentati in ordine cronologico sulla base della data di prima attestazione; ciascuna entrata è poi corredata anche da eventuali marche dell'*OED*.

Nel corso dei lavori preparatori per il *VSM*, era stata manifestata la volontà di organizzare la redazione su quadri concettuali, più agili da pianificare e controllare, che sarebbero stati successivamente integrati dal *DWS*.

Le possibili applicazioni del *VSM* sono state successivamente approfondite in Arcidiacono dov'è stato individuato un nucleo di 50 lessemi riferibili all'edilizia e alle abilitazioni riscontrati in un inventario di inizi anni '50.

5.4.3 *DHistOntology*

L'integrazione tra la lessicografia storica per l'italiano e il web semantico, riconosciuta tra le attività di una collaborazione istituzionale tra la Real Academia Española (RAE) e l'OVI, è entrata in una concreta fase sperimentale con il progetto *DHistOntology*.

L'impatto di una ricerca su un dispositivo concettuale per la lessicografia storica si riflette su diversi aspetti dell'attività lessicografica dell'OVI e della RAE, con esiti che sono difficilmente prevedibili nel complesso ma che possiamo collocare su tre diverse prospettive.

Tra i risultati a breve termine, una tassonomia integrata nel sistema Pluto potrebbe migliorare le funzioni di navigazione e ricerca all'interno dei due grandi dizionari.

Estendendo le aspettative al medio termine, si mira a sviluppare pienamente la dimensione onomasiologica all'interno di entrambi i dizionari con un approccio teorico coerente.

Prospettiva più generale → l'osservazione dei significati associati a un medesimo concetto, permetterebbe di guadagnare una posizione di osservazione privilegiata per studiare il cambiamento semantico in diacronia.

Nella sua prima forma provvisoria, il modello concettuale realizzato da *DHistOntology* consiste in una tassonomia da applicare alle accezioni del *DHLE* e del *TLIO*, basata sul modello dell'*Historical Thesaurus of English*.

La ricerca pilota per lo sviluppo di *DHistOntology* si è c100 nodi e sottonodi che si riallacciano al segmento, sono stati presi in considerazione quelli che hanno un riscontro nel lemmario del *DHLE* e del *TLIO*.

Come prima attività, è stato compilato un censimento di voci integrabili nella tassonomia, corredate dalla prima documentazione di ogni parola, dalla definizione, dal numero del significato e dall'eventuale marca d'uso; la selezione si è rivelata più facile sul *DHLE* grazie al suo processo redazionale basato su criteri morfo-etimologici e sulle famiglie di parole.

Successivamente è stata avviata una serrata riflessione sulle difficoltà riscontrate al momento della classificazione nella struttura concettuale dell'*HTE* e le relative implicazioni onomasiologiche nella storia del lessico italiano e di quello spagnolo. La formalizzazione concettuale della tassonomia è poi stata arricchita e convertita nell'ontologia *DHistOntology* in Protégé.

Questa ricerca ha permesso di evidenziare come la prospettiva storica e l'impostazione lessicografica si riflettano direttamente sulle classificazioni. I nomi di malattie vengono definiti con una certa sistematicità, descrivendone sintomatologie e l'eziologia.

Quando ci spostiamo alla corrispondente definizione del *TLIO* non si rilevano analoghe nozioni eziologiche.

In alcuni casi la distribuzione di una stessa accezione tra diversi nodi può avere ragioni strutturali.

Per i nomi di molte malattie sul *TLIO* si è optato per una più efficace collocazione al quarto livello a causa della descrizione non sempre omogenea.

In altri casi il *TLIO* descrive la malattia con l'indicizzazione della parte anatomica interessata e in rapporto ad altre malattie.

5.5 Nota sulla compilazione dei lemmari

Nel pannello di redazione di tutte le implementazioni di Lexicad esiste una funzione di anteprima che consente ai redattori di visualizzare, impaginate nel *template* dell'interfaccia di consultazione, anche le voci non ancora pubblicate.

Per intervenire sulle voci, l'utente autenticato nel *DWS* deve essere riconosciuto con il ruolo di redattore e disporre di specifiche autorizzazioni.

5.5.1 Dizionari a lemmario controllato

Sebbene il problema della compilazione di una nomenclatura di massima per il *Vocabolario storico* sia stato affrontato già nel corso della direzione di Aldo Duro, nessuna delle attività messe in atto nel cantiere dell'OVI è riuscita a raccogliere un lemmario di riferimento prima dell'avvio della redazione del *TLIO*. Il lemmario è stato così costruito con il procedere della redazione.

Nella piattaforma collaborativa di Pluto / Lexicad, ciascun utente dispone:

1. delle proprie credenziali di accesso
2. di un livello di accesso e può far parte di →
3. uno o più gruppi di utenti.

Sulla base di questi tre attributi, per ogni risorsa o elemento il sistema può consentire o negare l'accesso in lettura ai contenuti presenti nell'archivio, la creazione di nuove entrate, la modifica di quelle già inserite e la loro eliminazione.

Il tipo più semplice di controllo è basato sul livello di *default*.

Per le voci, in generale, il livello è impostato sul valore 5, che corrisponde alla figura di redattore generico. È possibile sovrascrivere il permesso di *default* di una risorsa per ciascun utente registrato.

Il *software* prevede, per ciascuno dei 4 permessi, altrettanti metodi specifici implementati dalle classi che gestiscono ciascuna risorsa.

Il *VMS* si basa sul *Corpus ARTESIA* indicizzato solamente per forme → il *VMS* non può adottarne le forme, perché la nomenclatura di Scobar è spesso incoerente; non può neanche perseguire la maggiore conformità possibile alla lingua moderna, dato che l'evoluzione del siciliano non ha un punto di arrivo codificato in uno *standard*.

5.5.2 Dizionari a lemmario precompilato

Per pianificare e controllare attentamente il flusso di redazione, la condizione ideale per un vocabolario è quella in cui si dispone di un lemmario completo prima dell'avvio della redazione.

Il *TLIO* è finito per rientrare in questa categoria a partire dal dicembre 2014, quando Rossella Mosti ha completato e pubblicato *on-line* il *Lemmario generale*. In questi casi il sistema non permetterà di inserire nuove voci.

Il sistema permetterà agli amministratori di poter effettuare i fisiologici aggiustamenti in corso d'opera.

5.5.3 Dizionari a lemmario libero

Possiamo definire dizionari a lemmario libero tutti quei dizionari in cui il livello di autorizzazione richiesto per la creazione di nuove voci è pari a quello di un redattore.

Il *VDL* non ha caricato preventivamente le voci nel vocabolario; queste vengono inserite nel *back-end* dai redattori senza disporre di privilegi particolari. La libertà di questo processo riguarda i soli aspetti esecutivi e non implica una procedura redazionale realmente permissiva.