



UNIVERSITÀ DI PISA

**Corso di Laurea Specialistica
in Informatica Umanistica**

La fruizione telematica dei carteggi d'artista.

**Una proposta metodologica e *Carlo Finelli Corpus*,
un corpus epistolare esemplificativo.**

Candidato: *Federico Giannini*

Relatore: *Cinzia Maria Sicca*

Relatore: *Chiara Renso*

Anno Accademico 2009-2010

Indici

Indice generale

Introduzione.....	10
1. Carteggi ed epistolari.....	14
1.1 L'importanza dello studio dei carteggi.....	14
1.2 Breve storia dell'epistolografia e dei carteggi: l'antichità classica.....	17
1.3 Breve storia dell'epistolografia e dei carteggi: Medioevo, Umanesimo e Rinascimento.....	18
1.4 Breve storia dell'epistolografia e dei carteggi: il Seicento e il Settecento.....	21
1.5 Breve storia dell'epistolografia e dei carteggi: dall'Ottocento ai giorni nostri.....	23
1.6 I carteggi d'artista.....	25
1.7 Problemi legati alla consultazione dei carteggi.....	27
2. La fruizione telematica dei carteggi.....	30
2.1 Informatica per l'archivistica e importanza della fruizione telematica dei carteggi. .	30
2.2 Problematiche relative alla fruizione telematica dei carteggi.....	31
2.3 Problemi: le edizioni critiche.....	32
2.4 Problemi: i disegni.....	36
2.5 Problemi: la lingua.....	38
2.6 Problemi: la standardizzazione.....	41
3. I corpora testuali.....	43
3.1 Che cos'è un corpus?.....	43
3.2 Breve storia della linguistica dei corpora.....	45
3.3 Tipi di corpora.....	48
3.4 L'annotazione dei corpora.....	51
3.5 Il linguaggio XML.....	53
3.6 Due codifiche standard per XML.....	58
4. Carteggi d'artista in rete: lo stato dell'arte.....	61
4.1 Introduzione.....	61
4.2 Il portale FO.S.C.A.....	61
4.3 The History of the Accademia di San Luca.....	64
4.4 Il portale della Fondazione Memofonte.....	66
4.5 Il Corpus Informatico Belloriano.....	68
4.6 Medici Archive Project.....	70
4.7 Vincent Van Gogh – The Letters.....	73
4.8 Google Books.....	75
5. Una proposta metodologica: fasi preliminari.....	77
5.1 Introduzione.....	77
5.2 Acquisizione delle lettere.....	77

5.3 Tokenizzazione.....	80
5.4 Lemmatizzazione e annotazione morfologica.....	82
5.5 Trasformazione dei testi lemmatizzati e annotati in file XML.....	85
5.6 Costruzione della base di dati MySQL.....	91
5.7 Trasformazione dei file XML in record MySQL.....	92
5.8 La presentazione delle lettere.....	97
6. Una proposta metodologica: i tipi di ricerca.....	101
6.1 Introduzione.....	101
6.2 La ricerca per forma e per lemma.....	101
6.3 La ricerca per categoria grammaticale.....	108
6.4 La ricerca per cooccorrenze	109
6.5 La ricerca di stringhe di testo.....	112
6.6 La ricerca nei contenuti (destinatario, luogo, data ecc.).....	115
7. Una proposta metodologica: la fruizione per il pubblico.....	118
7.1 Implementazione del sito web: linee-guida.....	118
7.2 Il sito del corpus: contenuti aggiuntivi.....	123
7.3 Usabilità.....	125
7.4 Accessibilità.....	127
7.5 Usability test e Website evaluation.....	131
8. Carlo Finelli Corpus.....	133
8.1 Obiettivi.....	133
8.2 Descrizione del progetto	134
8.3 Applicazione della metodologia proposta.....	136
8.4 La struttura del sito.....	138
9. Possibili utilizzi.....	148
9.1 Introduzione.....	148
9.2 Per i musei.....	148
9.3 Per gli organizzatori di mostre.....	150
9.4 Per i linguisti	151
9.5 Per docenti e studenti.....	152
9.6 Per gli appassionati	153
Conclusioni.....	154
Bibliografia.....	157
Siti web.....	163

Indice delle illustrazioni

2.1: l'incipit della <i>Divina Commedia</i> nell'edizione critica curata da Giorgio Petrocchi (Mondadori, 1996).....	33
2.2: confronto tra quattro edizioni critiche di una lirica di Arnaud Daniel nel portale RIALTO.....	34
2.3: pagina web con alcune righe tratte da un manoscritto autografo di Friedrich Nietzsche per la stampa della nuova edizione de <i>La Gaia Scienza</i> , contenente il Prologo, il V Libro, le “Canzoni del principe Vogelfrei”.....	35
2.4: incipit della lettera inviata da Vincent van Gogh ad Anton van Rappard l'8 marzo 1884.....	36
2.5: una pagina del sito <i>Vincent Van Gogh – The Letters</i>	37
4.1: la home page del portale FO.S.C.A.....	63
4.2: la <i>Raccolta di lettere</i> di Bottari e Ticozzi visualizzata su FO.S.C.A.....	63
4.3: home page del progetto <i>Documents from the Archivio di Stato di Roma</i>	66
4.4: la pagina di un documento in <i>Documents from the Archivio di Stato di Roma</i>	66
4.5: home page del sito della Fondazione Memofonte.....	68
4.6: una lettera da Giorgio Vasari a Pietro Aretino nel sito della Fondazione Memofonte. .	68
4.7: dichiarazione di conformità e report della validazione della home page di Fondazione Memofonte (13 marzo 2010).....	69
4.8: prima pagina del <i>Corpus Informatico Belloriano</i>	70
4.9: la visualizzazione KWIC del motore di ricerca del <i>Corpus Informatico Belloriano</i>	70
4.10: home page di <i>Medici Archive Project</i>	72
4.11: dati relativi a una lettera inviata da Cosimo I de' Medici al Bronzino in <i>Medici Archive Project</i>	72
4.12: home page di <i>Vincent Van Gogh – The Letters</i>	75
4.13: una lettera di <i>Vincent Van Gogh – The Letters</i> con la caratteristica doppia visualizzazione: a sinistra l'originale in olandese e a destra la traduzione.....	75
8.1: home page di <i>Carlo Finelli Corpus</i>	136
8.2: la pagina “Le lettere” di <i>Carlo Finelli Corpus</i>	136
8.3: i contesti KWIC della forma “presente” in <i>Carlo Finelli Corpus</i>	138

Indice delle tabelle

2.1: lemmatizzazione di una lettera di Lorenzo il Magnifico eseguita con un analizzatore morfologico per l'italiano contemporaneo.....	38
2.2: lemmatizzazione di una lettera di Giuseppe Raimondi eseguita con un analizzatore morfologico per l'italiano contemporaneo.....	39
3.1: un file XML correttamente formattato.....	54
3.2: la DTD del file XML di esempio presentato in TAB. 3.1.....	57

3.3: un'ipotetica codifica XML secondo gli standard TEI del carteggio Canova-Quatremère de Quincy.....	58
5.1: tokenizzazione di una frase tratta da una lettera di Carlo Finelli.....	81
5.2: lemmatizzazione di una frase tratta da una lettera di Carlo Finelli.....	82
5.3: lemmatizzazione e disambiguazione di una frase tratta da una lettera di Carlo Finelli.....	83
5.4: DTD per i token.....	84
5.5: il file annotato dopo la “pulizia”.....	85
5.6: codice in linguaggio Perl per la trasformazione dei file annotati in file XML.....	85
5.7: DTD per le lettere.....	87
5.8: formattazione per l'inserimento dei collegamenti alle immagini dei manoscritti.....	88
5.9: file XML dei token formattato secondo la DTD.....	89
5.10: file XML della lettera formattato secondo la DTD.....	89
5.11: codice per trasformare i file XML in record MySQL.....	92
5.12: codice per cancellare i token dalla base di dati MySQL.....	95
5.13: query per ottenere i dati della lettera richiesta.....	96
5.14: codice per la presentazione della lettera richiesta.....	97
5.15: la funzione “formatodata”.....	98
6.1: implementazione della ricerca per forma.....	102
6.2: switch per le categorie grammaticali.....	103
6.3: codice per la visualizzazione dei contesti KWIC.....	104
6.4: il select per le categorie grammaticali.....	107
6.5: ricerca delle cooccorrenze nella base di dati.....	109
6.6: query per i contesti KWIC nella ricerca di cooccorrenze.....	110
6.7: la ricerca di stringhe nel testo.....	112
6.8: evidenziazione dei risultati della ricerca nella pagina della lettera.....	113
6.9: ricerche nei contenuti.....	115
7.1: contenuti aggiuntivi dei siti analizzati nel CAP. 4.....	123
7.2: le WCAG 2.0 del progetto WAI del W3C.....	128
8.1: struttura tipica di una pagina HTML.....	138
8.2: DOCTYPE ed etichetta “html” con indicazione della lingua.....	138
8.3: <i>head</i> delle pagine di <i>Carlo Finelli Corpus</i>	139
8.4: struttura delle pagine di <i>Carlo Finelli Corpus</i>	141
8.5: le due funzioni per la rotazione delle immagini.....	143
8.6: definizione dell'elemento “body” in “stile.css”.....	144
8.7: definizione dell'elemento “testo” in “stile.css”.....	145

Introduzione

I progressi nel campo delle telecomunicazioni uniti alle tecnologie *web-based* hanno cambiato in modo notevole l'approccio allo studio, alla ricerca e alla didattica. Molte cose impensabili fino a pochi anni fa sono adesso possibili grazie all'aiuto offerto dalla rete, che consente di avere molte informazioni gratuite e a portata di mano, utili per le proprie ricerche.

Tuttavia, a seguito della crescita esponenziale della mole di informazioni presenti sul web, è diventato più che lecito domandarsi se le ricerche sulla rete diano luogo a risultati attendibili, se le fonti reperite possano essere definite autorevoli, se davvero è possibile trovare ciò che si cerca: si pone, in sostanza, il problema di avere un web di qualità.

Se il cosiddetto “Web 2.0” ha di fatto “allargato la base” della rete, consentendo a chiunque di poter gestire pagine web anche senza avere le competenze tecnologiche e informatiche richieste prima dell'invenzione di blog, social network e piattaforme simili, dall'altro lato ha posto le premesse per uno sviluppo “anarchico” dell'informazione in rete. In un web dove tutti possono fare informazione, molto spesso senza avere le basi e altrettanto spesso senza neppure avvertire il bisogno di firmarsi o di dare notizie sul proprio conto, diventa sempre più difficile verificare la qualità di ciò che si è trovato.

Quindi negli ultimi tempi è più che mai necessario che chi produce contenuti si ponga il problema della qualità e dell'attendibilità di ciò che immette in rete. Il progetto presentato in questo lavoro va proprio in questo senso: una proposta metodologica per realizzare pubblicazioni telematiche di carteggi d'artista di qualità, utili, innovative per la storia dell'arte.

La storia dell'arte è una disciplina nella quale è particolarmente importante lo studio delle fonti: tra queste ultime è doveroso includere i carteggi prodotti dagli artisti. Lo studio dei carteggi è imprescindibile per trarre informazioni utili per attribuire le opere in modo corretto, per conoscere con precisione la loro data di realizzazione e per sapere quali sono state le fasi che ne hanno accompagnato la creazione, per conoscere meglio il pensiero di un autore e per avere notizie sulla sfera privata di un artista.

Esistono in rete diverse pubblicazioni di carteggi d'artista, delle quali non si vuole mettere in dubbio la qualità e che verranno adeguatamente analizzate nel corso della trattazione. Si ha però l'impressione che molti enti e molti istituti, anche accademici, realizzino progetti ragionando più da “progettatori” e da “addetti ai lavori” che da “utenti”: talvolta le soluzioni scelte non sono le più intuitive e facili da usare, e spesso si è constatato che alcuni progetti non rispettano standard anche minimi di accessibilità.

La proposta qui presentata è stata ideata partendo da un presupposto ben preciso: che cosa cercano studiosi, ricercatori, docenti, studenti e appassionati d'arte in una pubblicazione di un corpus digitale di carteggi d'artista? In altri termini, quali sono le informazioni e gli strumenti di cui bisogna dotare le sopraccitate categorie in modo tale da fornire loro un prodotto di qualità, il più possibile completo e allo stesso tempo facile da usare e il cui utilizzo non richieda tempi lunghi per poter essere imparato? Ed è fattibile la realizzazione di un sistema che abbia queste caratteristiche?

Con il presente lavoro si cerca di rispondere proprio a queste domande: l'obiettivo della proposta metodologica è quello di permettere la realizzazione di pubblicazioni telematiche di carteggi d'artista di qualità, complete, utili, usabili e accessibili.

Per elaborare tale proposta è stato necessario unire tre diverse sfere di competenza: le competenze storico-artistiche, quelle linguistiche e quelle informatiche. Si tratta di un aspetto di cui tener conto, perché quando si realizzano progetti informatici per le risorse umanistiche è necessario che le diverse competenze si fondano e si completino in modo armonico e solidale. Altrimenti, il rischio è quello di produrre strumenti che non tengano conto da una parte delle abilità dei programmatori, e dall'altra delle esigenze degli utilizzatori finali. Anche in questo sta il senso dell'informatica umanistica: far dialogare tra loro sfere di competenza che alcuni, a torto, ritengono inavvicinabili.

Un altro scopo di questo lavoro, infatti, è da un lato far vedere che è possibile avvicinarsi all'informatica con una mentalità da “umanista”, e dall'altro dimostrare a studiosi e docenti di prevalente formazione storico-artistica che la tecnologia, in particolare quella *web-based*, non è un qualcosa dal quale tenersi lontani, ma è una grandissima e potente opportunità di studio, di approfondimento, di educazione.

Infine, l'ultimo proposito che si vuole perseguire è quello di presentare una proposta metodologica che possa di fatto diventare uno standard al quale attenersi quando si vorrà realizzare un corpus digitale di carteggi d'artista. Per documentare le potenzialità di questa proposta si presenterà anche il progetto *Carlo Finelli Corpus* (<http://carlofinelli.losnaweb.com>), realizzato per conto di Fondazione Cassa di Risparmio di Carrara e Losna – Soluzioni per il Web: si tratta della prima pubblicazione telematica di un carteggio d'artista realizzata secondo gli standard della proposta metodologica che qui si vuole discutere. Carlo Finelli (1782 – 1853) è stato un importante scultore neoclassico di Carrara e i suoi carteggi sono stati pubblicati in questo corpus telematico che vuole quindi assumere la

funzione di progetto pilota, per mostrare cosa si può realizzare seguendo le linee dettate dalla proposta metodologica.

La trattazione è stata suddivisa in nove capitoli. Durante il primo, si introdurranno le opportune terminologie (“carteggi”, “epistolari”, “epistolografia” e altre) alle quali si farà riferimento durante tutto il corso dell'esposizione. Inoltre, si discuterà dell'importanza dello studio dei carteggi e si delinea una breve storia dell'epistolografia e dei carteggi, per dimostrare l'importanza che questi ultimi hanno rivestito nel corso della storia, della storia della letteratura e della storia dell'arte; un paragrafo sarà proprio dedicato alla storia dei carteggi d'artista. Si discuteranno infine le spinose problematiche legate alla consultazione dei carteggi: del resto tali problemi costituiscono uno dei motivi per i quali si rendono necessarie le pubblicazioni *on-line*.

Il secondo capitolo vuole invece essere una dimostrazione di quale sia l'importanza della fruizione telematica dei carteggi, e allo stesso tempo vuole proporre soluzioni ai problemi che potrebbero nascere mentre si cerca di realizzare un progetto i cui contenuti, normalmente fruibili in versione “cartacea”, siano invece presentati in forma “digitale”.

Il terzo capitolo cerca di rispondere a una domanda ben precisa, che in genere viene rivolta a chi si occupa di linguistica dei corpora: che cos'è un corpus? In particolare, che cos'è un corpus in formato digitale? Si discuterà quindi dell'importanza dei corpora per la linguistica, si parlerà dei diversi tipi di corpora, di come questi ultimi possano essere annotati e infine si farà un'introduzione al linguaggio XML, indispensabile al fine di comprendere bene la proposta metodologica.

Nel corso del quarto capitolo si analizzerà invece lo stato dell'arte, ossia si condurrà una critica sui carteggi d'artista in rete già esistenti: il capitolo vuole evidenziare punti deboli e punti di forza delle pubblicazioni già esistenti, in modo tale da inquadrare meglio la proposta metodologica qui discussa.

Con il quinto capitolo si introdurrà la proposta metodologica stessa, e in particolare se ne illustreranno le fasi preliminari: dall'acquisizione delle lettere passando attraverso la “tokenizzazione” e la “lemmatizzazione” dei testi, fino ad arrivare alla costruzione dei file XML contenenti i testi annotati delle lettere e alla trasformazione di questi ultimi in record di una base di dati MySQL, che di fatto conterrà il corpus. Si daranno quindi le opportune giustificazioni al fatto che con la presente proposta si cambia il tradizionale formato di interscambio dei dati nella linguistica dei corpora: quindi non XML, ma MySQL.

Il sesto capitolo presenta i vari tipi di ricerca che la proposta permette di implementare: per forma, per lemmi, per categorie grammaticali, e così via. Il settimo capitolo invece fornisce linee guida per la costruzione del sito che ospiterà l'eventuale corpus realizzato in accordo con gli standard della proposta qui discussa. All'interno della stessa sezione si discuterà di come realizzare una grafica esteticamente gradevole e di come rendere il sito usabile e accessibile.

L'ottavo capitolo introdurrà il progetto *Carlo Finelli Corpus*: la trattazione sarà quindi arricchita con un esempio concreto e tangibile di cosa si possa ottenere con la proposta metodologica. Si tratta di un aspetto molto importante, perché un conto è fornire indicazioni teoriche, ma ben altra cosa è poterle dimostrare con un esempio reale, già disponibile e interamente consultabile in rete.

Infine, il nono e ultimo capitolo, a chiusura dell'esposizione, vuole chiarire in che modo le pubblicazioni telematiche dei carteggi d'artista possano essere utili per diverse categorie di utilizzatori: si tratta di una sezione importante perché se si mostra quali possono essere le applicazioni di un tale progetto, quest'ultimo acquista senza dubbio un valore aggiunto.

1. Carteggi ed epistolari

1.1 L'importanza dello studio dei carteggi

Spesso, quando si fa riferimento alla produzione epistolare di un artista, di un letterato, di un uomo politico, di un intellettuale, si usano in modo indistinto diversi termini; *epistolografia*, *epistolario*, *carteggio* sono i più diffusi e usati. È però necessario sottolineare che, benché tali termini possano apparire del tutto simili, recano invece qualche differenza che è bene evidenziare al fine di evitare fraintendimenti.

Il termine *epistolario* fa riferimento a una raccolta di lettere di personaggi illustri che abbiano intenti letterari¹, mentre si parla di *carteggio* quando la raccolta riguarda lettere scritte e ricevute senza intenti letterari². L'*epistolografia* invece è lo studio degli epistolari³, e da molti viene considerata come un genere letterario autonomo, praticato fin dall'antica Grecia soprattutto in campo filosofico.

È necessario quindi distinguere due tipi di lettere: quelle pensate per fini letterari, e quindi per la pubblicazione (ovvero le lettere raccolte negli epistolari), e quelle invece pensate senza intenti di pubblicazione (quelle raccolte nei carteggi, e tali lettere possono essere sia pubbliche che private).

I carteggi possono essere considerati come fonti di informazioni dirette per quanto concerne lo studio della storia e della storia dell'arte: le *fonti dirette* possono essere definite “frammenti del passato che esistono o che possono essere osservati direttamente dallo storico”⁴. Questo è particolarmente vero per le lettere, in quanto costituiscono un “frammento del passato” che può essere studiato e analizzato così com'è stato prodotto in un particolare contesto storico e da un particolare personaggio storico. Le lettere poi, oltre a poter essere considerate fonti dirette, possono anche configurarsi come *fonti indirizzate*: queste ultime sarebbero “non soltanto frammenti del passato che testimoniano azioni umane o, eventualmente, oggetti del mondo naturale che possono servire come base per la comprensione del passato [...]; ma si tratta di fonti che risalgono a un autore e sono destinate a un uditorio”⁵. Nella fattispecie, tale uditorio può essere collocato nel passato: è il caso del destinatario della lettera.

1 Giuliana Saporì, *Manuale di applicazione delle regole di catalogazione italiane al Servizio Bibliotecario Nazionale* (Segrate: Cilea, 2009). E-book disponibile all'indirizzo <http://manualesapori.cilea.it>.

2 *Ibid.*, <http://manualesapori.cilea.it/index.php?id=281>.

3 *Ibid.*

4 Jerzy Topolski, *Narrare la storia: nuovi principi di metodologia storica* (Milano: Mondadori, 1997), 53.

5 *Ibid.*, 54.

1. Carteggi ed epistolari

Da questa premessa risulta quindi particolarmente evidente la grande importanza che lo studio dei carteggi riveste per lo studio della storia in generale e, nel nostro caso, della storia dell'arte in particolare. I carteggi sono importanti per poter avere una conoscenza più approfondita dell'artista, del suo pensiero e del contesto entro il quale ha operato, e questa conoscenza può essere di notevole aiuto per fornire un quadro più completo di quello che si evince semplicemente dalle biografie di un artista. Nei carteggi si può riconoscere una fonte unica per condurre al meglio non solo lo studio del contesto culturale del tempo, ma anche per scoprire notizie sui personaggi illustri che difficilmente possono essere riscontrate attraverso altre fonti⁶. Un caso lampante può essere rappresentato dallo stesso Carlo Finelli, dal momento che grazie ai carteggi rinvenuti è stato possibile ricostruire molte delle sue vicende biografiche, ma non solo: è stato possibile anche conoscere con esattezza la datazione di alcune sue opere ed è stato possibile avere notizie circa sue opere andate perdute⁷.

Il ruolo della lettera risulta dunque fondamentale per diversi motivi. Mireille Bossis riconosce alla lettera quattro diverse funzioni: la lettera come *documento ritenuto "vero"*, la lettera come *oggetto di scrittura*, la lettera come *discorso* e infine la lettera come *gesto*⁸.

Della lettera come *documento ritenuto "vero"* si è già in parte discusso, ma è interessante aggiungere come la lettera possa anche essere una testimonianza che appartiene più spesso alla sfera del privato⁹. La lettera inoltre è un *oggetto di scrittura* che appartiene al dominio della rappresentazione, e in quanto tale non è possibile fare economia di un passaggio attraverso l'immaginario di chi l'ha scritta¹⁰: pertanto, la lettera è strettamente legata al suo autore, alle sue concezioni e anche al suo modo di scrivere, tanto che spesso (in particolar modo nelle corrispondenze degli artisti), è possibile imbattersi in errori di ortografia, oppure in forme che non sono delle migliori¹¹.

6 *Ibid.*

7 Barbara Musetti, *Carlo Finelli (1782 – 1853)* (Milano: Silvana Editoriale, 2002).

8 Mireille Bossis, "Introduction" in *La lettre à croisée de l'individuel et du social*, a cura di Mireille Bossis (Parigi: Éditions Kimé, 1994), 9 – 13.

9 *Ibid.*

10 *Ibid.*

11 È il caso, per esempio, di alcune lettere scritte dall'importante pittore secentesco Guido Cagnacci (1601 – 1663) a un suo corrispondente veneziano, Francesco Gionima. Uno dei primi biografi di Cagnacci, il pittore riminese Giovanni Battista Costa (1697 – 1767), nelle sue *Lettere varie e documenti autentici intorno le opere e vero nome e cognome e patria di Guido Cagnacci pittore*, riporta un'epistola di un suo corrispondente, Giampietro Zanotti, secondo il quale le lettere scritte da Cagnacci sarebbero "scritte pessimamente". Zanotti aggiunge nella sua lettera inviata a Costa il 28 luglio del 1742 che "se avesse così dipinto guai a lui". Per un approfondimento, Giovanni Battista Costa, "Lettere varie e documenti autentici intorno le opere e vero nome e cognome e patria di Guido Cagnacci" in *Raccolta d'opuscoli scientifici e filologici XLVII* (1752), 117-161 e *Guido Cagnacci. Protagonista del Seicento tra Caravaggio e Reni*, catalogo della mostra, a cura di Antonio Paolucci e Daniele Benati (Milano: Silvana Editoriale, 2008).

1. Carteggi ed epistolari

La lettera poi si configura come *discorso* che, sebbene scritto in base a uno stile personale, è tributario di rappresentazioni collettive della sua epoca e del sistema retorico che tale epoca suscita¹²; quindi è necessario, quando si studia un carteggio, tenere conto del contesto storico (e storico-artistico, nel caso della storia dell'arte) all'interno del quale le lettere sono state scritte.

Infine, la scrittura della lettera è un *gesto* in quanto azione comportamentale che mira a stabilire o a mantenere una comunicazione con l'altro, quindi una relazione intersoggettiva per abolire distanze di qualsiasi ordine¹³.

Le quattro funzioni della lettera non possono essere scisse e separate l'una dall'altra, ma convivono e si esercitano in modo simultaneo; la lettera diventa così un mezzo privilegiato per ricavare informazioni sia sull'individuo che sul contesto storico o storico-artistico.

In conclusione, è possibile aggiungere che la lettera può accogliere anche un'infinità di tematiche, ed è possibile affermare che la lettera di per sé non ne esclude alcuna¹⁴. Questo aspetto è rilevante anche se lo si abbina al fatto che le lettere riguardano la sfera privata del personaggio illustre, quindi spesso è possibile venire a conoscenza, tramite i carteggi, di aspetti legati al pensiero di un autore che altrimenti non potrebbero emergere dalle opere pensate per la pubblicazione. Cicerone, scrivendo a Trebonio (*Epistulae ad familiares*, 15, 21, 4), distingueva tra lettere riservate e lettere pensate per un'ampia cerchia di lettori¹⁵: lo stesso autore, scrivendo ad Attico (*Epistulae ad Atticum*, 16, 5, 5) chiedeva all'amico di farsi restituire esemplari di alcune missive in modo tale da rivederle e correggerle¹⁶. Il risultato è che le lettere private di Cicerone che ci sono giunte spesso non corrispondono all'immagine che l'oratore voleva dare di sé, così come non corrispondono alle sue intenzioni editoriali. Le lettere private erano scritte con un linguaggio quotidiano e ricche di vocaboli della vita di tutti i giorni, in uno stile molto lontano da quello che Cicerone usava per gli scritti destinati alla pubblicazione, e in più è del tutto lecito pensare che l'autore avesse voluto eliminare dalla corrispondenza privata alcune affermazioni particolarmente compromettenti sul piano personale e politico¹⁷, ma anche considerazioni che avrebbero nociuto alla cura dei suoi

12 *La lettre à croisée de l'individuel et du social*, a cura di Mireille Bossis, 9 – 13.

13 *Ibid.*

14 *Ibid.*

15 Alberto Caverzere, "Letteratura epistolare di Roma" in *Alla lettera. Teorie e pratiche epistolari dai Greci al Novecento*, a cura di Adriana Chemello (Milano: Guerini, 1998), 11-32.

16 *Ibid.*

17 *Ibid.*

1. Carteggi ed epistolari

rapporti con gli altri e, come si osservava poco sopra, alla propria immagine¹⁸.

1.2 Breve storia dell'epistolografia e dei carteggi: l'antichità classica

Secondo la tradizione, la prima lettera dell'antichità classica è testimoniata da Omero (*Iliade* VI, 170 ss.): nell'episodio di Bellerofonte, il poeta fa riferimento a una comunicazione privata, avente carattere di assoluta riservatezza, inviata da Preto, re di Tirinto, a Iobate, re della Licia¹⁹. Si può ipotizzare che la lettera in Grecia abbia origine nel IV secolo a.C.²⁰, e all'interno della cultura greca sarebbe possibile individuare quattro tipi di epistole: quelle contenenti messaggi personali, quelle scritte da personaggi di rilievo (sulla cui autenticità esistono però dubbi) e rivolte a un pubblico ampio (un esempio è costituito dalle lettere di Platone), quelle utilizzate a scopo propagandistico per la diffusione del pensiero filosofico (come quelle di Epicuro o di Dionigi di Alicarnasso) e quelle scritte per essere pubblicate, che quindi perderebbero la loro funzione di missiva per configurarsi come un genere letterario autonomo, avente regole proprie²¹. La maggior parte delle lettere arrivateci appartiene agli ultimi tre tipi.

Per ciò che riguarda invece la cultura romana, prima di Cicerone non esistono né epistolari, intesi come raccolte di lettere pensate per la pubblicazione, né tanto meno raccolte di lettere private²². Le uniche lettere delle quali si ha testimonianza sono le lettere ufficiali, funzionali alla comunicazione tra Roma e le province, pervenute tramite trattati di carattere divulgativo. Con il già citato Cicerone possiamo avere una testimonianza sia delle missive riservate scritte dall'oratore, sia delle epistole pensate per il pubblico²³: le sue lettere da una parte pongono le basi per una teorizzazione del genere epistolare, e dall'altra costituiscono un significativo *exemplum* di missiva privata²⁴. Da ricordare sono inoltre gli epistolari di Cesare e di Sallustio (benché sull'autenticità di quest'ultimo si nutrano dubbi), coevi a quello di Cicerone (I sec. a.C.), mentre per trovare un epistolario in età imperiale sarà necessario attendere Seneca (I sec. d.C.). Importanti sono anche le lettere di Plinio (I-II sec. d.C.) a causa dell'enorme mole

18 Sandra Citroni Marchetti, "Lo spazio straniato. Percorsi psicologici e percezione del tribunale nelle orazioni di Cicerone 'pro Fonteio', 'pro Q. Roscio comoedo', 'pro Cluentio'" in *Materiali e discussioni per l'analisi dei testi classici* XXXVI (1996), 33-71.

19 *Antichità classica*, a cura di Luciana Aigner Foresti (Milano: Jaca Book, 1994).

20 Elisa Avezzù, "Alle origini dell'epistolografia. I Greci e la lettera" in *Alla lettera*, 1-10.

21 *Antichità classica*, a cura di Luciana Aigner Foresti, 124.

22 *Ibid.*

23 Cfr. *supra* PAR. 1.1.

24 *Antichità classica*, a cura di Luciana Aigner Foresti, 125.

1. Carteggi ed epistolari

di informazioni sulla vita romana dell'epoca che da esse è possibile trarre, nonché quelle di Frontone (II sec. d.C.), che offrono al lettore interessanti spaccati di quotidianità romana. L'ultimo epistolario della letteratura latina pagana è opera di Simmaco (IV sec. d.C.), autore di una raccolta pensata per essere pubblicata e caratterizzata da uno stile molto ricercato. Pensati per la pubblicazione sono anche gli epistolari degli autori cristiani, come Ambrogio (IV sec. d.C.), Girolamo e Agostino (IV – V sec. d.C.)²⁵.

1.3 Breve storia dell'epistolografia e dei carteggi: Medioevo, Umanesimo e Rinascimento

Nel Medioevo la lettera diventa più di un veicolo per la trasmissione di informazione: la lettera inizia a essere concepita più per un pubblico vasto che per un solo destinatario, in virtù del fatto che essa poteva assumere più connotazioni del semplice scambio di informazioni riservate²⁶. I modelli degli scrittori medievali, nell'ambito dell'epistolografia, sono i grandi epistolari della cultura classica (su tutti quelli di Cicerone e Seneca), ma molti guardano anche agli autori cristiani.

Sono pochi gli epistolari appartenenti al periodo di transizione tra la tarda antichità e il Medioevo che ci sono pervenuti: tra questi è possibile ricordare le lettere di alcuni scrittori come san Patrizio (V secolo d.C.), Colombano (VII sec. d.C.) e sant'Adelmo di Sherborne (VII - VIII sec. d.C.). Durante il cosiddetto “Rinascimento carolingio”, l'arte dello scrivere lettere si sviluppa tra gli studiosi e i teologi della corte di Carlo Magno²⁷, principalmente come mezzo per promuovere la cultura. Durante il pontificato di Gregorio VII (1020 – 1085), la lettera inizia a diventare un potente strumento politico, dal momento che il papa usava le missive così come usava i suoi legati, cioè come estensioni del suo potere personale e strumenti per influenzare la politica²⁸. Lo stesso ragionamento vale anche per l'imperatore Enrico IV (1056 – 1106) e per Bernardo di Chiaravalle (1090 – 1153): quest'ultimo in particolare era solito inviare lettere alle maggiori figure di spicco del suo tempo (al papa stesso, per esempio) per commentare le vicende dell'epoca²⁹.

²⁵ Roberta Antognini, *Il progetto autobiografico delle Familiars di Petrarca* (Milano: LED, 2008).

²⁶ Julian Haseldine, “Epistolography” in *Medieval Latin: an introduction and bibliographical guide*, a cura di Frank Anthony Carl Mantello e A.G. Rigg (Washington, D.C.: Catholic University of America Press, 1996).

²⁷ *Ibid.*, 652.

²⁸ *Ibid.*, 653.

²⁹ Si veda a tal proposito *San Bernardo e l'Italia: atti del convegno di studi, Milano, 24-26 maggio 1990*, a cura di Paolo Zerbi (Milano: Vita e Pensiero, 1993).

1. Carteggi ed epistolari

La scrittura di lettere fiorisce nel XII secolo: basterà citare le lettere di sant'Anselmo, per il quale la stesura di epistole era una parte integrante della propria attività teologica e religiosa, nonché il celeberrimo scambio di lettere d'amore tra Abelardo ed Eloisa (benché si dibatta ancora sulla sua autenticità)³⁰.

Forse però è possibile considerare come la più importante opera epistolare del Medioevo, il *Familiarium rerum liber* di Francesco Petrarca (1304 – 1374), importante epistolario scritto dall'autore aretino a seguito della scoperta di alcune lettere di Cicerone³¹, benché tra i suoi modelli si possano annoverare anche Seneca e sant'Agostino. Il progetto di Petrarca era quello di realizzare una “autobiografia raccontata attraverso una sequenza di lettere”³². La stesura dell'epistolario occupa lo scrittore per ventun anni, poiché i lavori iniziano nel maggio del 1345, a seguito della scoperta delle lettere di Cicerone, e durano fino alla fine del 1366. È quindi possibile considerare la raccolta di Petrarca come un'opera letteraria a pieno titolo.

Cicerone è stato un notevole modello anche per gli umanisti³³, che vantavano di aver riportato allo splendore l'antica eloquenza latina del grande oratore. Di fatto è proprio Francesco Petrarca l'iniziatore del culto di Cicerone, un culto che, “reso più consapevole sul piano teorico e storico da Leonardo Bruni [...], toccò il massimo fulgore nei primi decenni del Cinquecento in seno alla curia papale, quando Pietro Bembo e il Sadoleto [...] introdussero il modello ciceroniano a canone ufficiale delle scritture epistolari latine”³⁴. Tra i più importanti autori umanisti è possibile citare Coluccio Salutati (che in una sua lettera offre un significativo parallelo tra eloquenza ed epistolografia), Gaspare Barzizza (che adatta la tecnica oratoria di Cicerone allo stile epistolare), e Francesco Filelfo (autore anche di una notevole quantità di epistole in greco). La lettera umanistica, oltre a essere un documento letterario di notevole importanza, può rivestire anche la funzione di documento storico (benché nasca con esclusivo intento letterario), in quanto “riflette [...] le varie stagioni della cultura borghese dalla fine del Tre agli inizi del Cinquecento italiano”³⁵ e perché “un'ampia

30 La questione sull'autenticità dell'epistolario di Abelardo ed Eloisa è diffusamente discussa in Paola De Santis, *I sermoni di Abelardo per le monache del Paracletto* (Leuven: Leuven University Press, 2002).

31 È bene però notare che in realtà le lettere di Cicerone erano già note, ma nessuno prima di Petrarca se ne era mai curato: le lettere di Cicerone erano fino allora ignote o sfogliate soltanto da piccoli eruditi locali. Cfr. Giuseppe Billanovich, *Lo scrittoio del Petrarca* (Roma: Edizioni di storia e letteratura, 1947).

32 Roberta Antognini, *Il progetto autobiografico delle Familiars di Petrarca*

33 Per i modelli latini degli epistolari umanistici, si veda anche Jacob Burckhardt, *La civiltà del Rinascimento in Italia*, tradotto in italiano da Diego Valbusa (Firenze: Sansoni, 1975)1860.

34 Claudio Griggio, “Dalla lettera all'epistolario. Aspetti retorico-formali dell'epistolografia umanistica” in *Alla lettera*, 83-107. Citazione a pag. 83.

35 Alessandro Perosa e Paolo Viti, *Studi di filologia umanistica* (Roma: Edizioni di Storia e di Letteratura, 2000), 14.

1. Carteggi ed epistolari

conoscenza degli epistolari umanistici [...] renderà un gran servizio alla storia della cultura del '400, mettendo in rilievo la concretezza di idee e la pienezza di vita di quella civiltà”³⁶. Non bisogna inoltre dimenticare che l'Umanesimo vede anche la nascita di trattati di epistolografia, utili per gli studenti di retorica ai quali fornivano regole ed esempi: sono da citare il *De Compositione* di Gaspare Barzizza, il *Modus epistolandi* di Francesco Nigro, nonché il *De conscribendis epistolis* di Erasmo da Rotterdam³⁷.

La lingua utilizzata dagli umanisti era il latino: per trovare lettere scritte in volgare è necessario aspettare il Cinquecento, quando l'epistolografia si impone come genere di grande successo della prosa in volgare. L'invenzione del nuovo genere letterario (la raccolta di lettere in volgare) si deve a Pietro Aretino³⁸, che nel 1538 fa uscire il primo epistolario con il quale l'autore “lancia” sul mercato il nuovo genere, “consistente nel dare al pubblico le lettere private, riunite in volume”³⁹. Quello del “divin Pietro” è un vero *exploit*: una novità alla quale pochi altri rispondono, ma senza riscontrare il successo dell'Aretino⁴⁰. Nel volgere di poco tempo però il genere comincia a fiorire, e vedono la luce altre raccolte, tra le quali, giusto per citare le più importanti, quelle di Anton Francesco Doni (1544), Girolamo Parabosco (1545), Bernardo Tasso (1549) e Pietro Bembo (1552, postuma). Per quanto riguarda quest'ultimo autore, è interessante notare come egli stesso avesse tentato di impedire, prima della sua scomparsa, un'edizione delle sue lettere: questo perché non riteneva il genere dell'epistola in volgare degno di imitazione, ma reputava necessaria la creazione di un modello che “potesse essere agevolmente seguito da coloro che volevano scrivere un volgare *alto*”⁴¹.

La rivoluzione a cui Pietro Aretino dà inizio coinvolge anche lo stile: nelle lettere iniziano a comparire “fenomeni come la parodia, la battuta, e una certa qual forma di trascuratezza [...] che mai sarebbe stata ammessa dagli scrittori di epistole secondo le normative più rigorose dettate dai formulari”⁴². Il nuovo genere riscontra un grandissimo successo di pubblico, tanto che i libri di lettere cominciano anche a dividersi in gruppi: lettere di un

36 *Ibid.*

37 Ruggero Cattaneo, “O stilu i kulturnom značenju Marulić eve Poslanice papi Adrijanu VI” [“Sullo stile e la rilevanza culturale dell'Epistola a Papa Adriano VI di Marco Marulić”], *Colloquia Marulliana* XVII (17 aprile 2008): 91. Abstract in italiano disponibile in rete all'indirizzo http://hrcak.srce.hr/index.php?show=clanak&id_clanak_jezik=37658.

38 *Il Cinquecento: la normativa e il contrario (1533 – 1573): le nuove regole e l'estensione dell'analogia*, a cura di Giovanni Da Pozzo (Padova: Piccin, 2006), 1178.

39 *Ibid.*

40 Amedeo Quondam, “Dal 'Formulario' al 'Formulario': cento anni di 'Libri di lettere’” in *Le “carte messaggere”. Retorica e modelli di comunicazione epistolare: per un indice dei libri di lettere del Cinquecento*, a cura di Amedeo Quondam (Roma: Bulzoni, 1981), 13 – 157.

41 *Il Cinquecento: la normativa e il contrario*, 1181.

42 *Ibid.*, 1179.

1. Carteggi ed epistolari

singolo autore, o ancora lettere di uno stesso argomento (facete, amoroze, spirituali...), o epistole scritte da gruppi di persone tra loro omogenee (per esempio, donne).

1.4 Breve storia dell'epistolografia e dei carteggi: il Seicento e il Settecento

La grande stagione degli epistolari cinquecenteschi si chiude sul finire del secolo, e durante i cento anni successivi si passerà dai “libri di lettere” ai “libri sulle lettere”, segno che agli inizi del nuovo secolo il genere epistolare conosce una crisi. Il Seicento è il secolo dei *Segretari*, ovvero “mega-antologie di lettere che [...] corredano la teoria, in funzione didattica ed esemplare”⁴³. La fine del Cinquecento quindi vede il fiorire di opere di impegno teorico, come i *Segretari* scritti da Giulio Cesare Capaccio (1589) e Angelo Ingegneri (1607), opere che riscontrano anche un certo successo.

Tuttavia il genere dell'epistolario non si estingue nel Seicento, anzi: anche durante il XVII secolo ci sono raccolte di una certa importanza. Tra queste è necessario segnalare l'epistolario di Tommaso Costo, pubblicato nel 1604. La raccolta conteneva lettere redatte da Costo durante la sua attività di segretario ed era stata realizzata secondo chiari intenti letterari, dal momento che nella dedica ad Alberico Cybo Malaspina, principe di Massa e Carrara, l'autore si rivolge apertamente “a chi legge” chiarendo le ragioni dell'opera e illustrandone l'organizzazione⁴⁴. Durante il Seicento, il modello di riferimento è sempre Cicerone, per due ragioni sostanziali: in primo luogo perché le *Epistulae ad familiares* rappresentano un esempio di fedeltà e di continuità epistolare, e in secondo luogo perché costituirebbero la perfezione dello stile⁴⁵.

L'epistolografia riacquista vigore durante il Settecento: la lettera “ha un ruolo rilevante nella storia della cultura settecentesca, non solo come scambio di notizie fra intellettuali, ma come mezzo di diffusione e di confronto delle idee”⁴⁶: le lettere, durante il XVIII secolo, sono di fondamentale importanza per la costituzione della cosiddetta “repubblica delle lettere”. Si

43 Elisabetta Selmi, “Fra «Negotio» e «Parole»: per una «Institutio» retorica dei «Libri del Segretario»”. *La svolta degli anni Novanta in Alla lettera*, 173 – 227. Citazione a pag. 179.

44 Stefania Capuozzo, “L'epistolario di Tommaso Costo” in *Gli scrittori d'Italia. Il patrimonio e la memoria della tradizione letteraria come risorsa primaria* (Grottammare: Graduus, 2008). Il testo del saggio è disponibile online all'indirizzo <http://www.italianisti.it/FileServices/23%20Capuozzo%20Stefania.pdf>.

45 Geneviève Haroche, “Familier comme une épître de Cicéron: Familiarité dans la lettre au tournant du XVII et XVIII siècle” in *La lettre à la croisée de l'individuel et du social*, 17 – 24.

46 Giorgio Ronconi, “Aspetti della lettera familiare nel Settecento. La corrispondenza tra Egidio e Marco Forcellini e la prima edizione dell'epistolario zeniano” in *Alla lettera*, 229 – 243. Citazione a pag. 229.

1. Carteggi ed epistolari

intende con quest'ultimo termine una sorta di “istituzione”⁴⁷ sovranazionale, che non coincideva con un organismo realmente esistente, ma che per la partecipazione richiedeva soltanto “l'amore per le arti e per le scienze e il desiderio di recare un contributo al loro accrescimento”. Centro focale di questa “repubblica” poteva considerarsi Parigi, visto anche che il propulsore di questi scambi di idee fu l'Illuminismo: tutti i grandi autori italiani del XVIII secolo, dai fratelli Verri per arrivare ad Alessandro Manzoni, soggiornano nella capitale francese e frequentano l'ambiente parigino.

Grazie alla sua nuova funzione culturale, la lettera smette anche di essere etichettata all'interno di rigidi schemi retorici: il suo compito diventa quello di favorire lo scambio di informazioni tra intellettuali, di notizie su argomenti di attualità, di pareri sulla filosofia, sulla religione, sulla storia, sulla politica.

Nel Settecento inoltre si “torna” ai carteggi (le raccolte non pensate per la pubblicazione), spesso pubblicati in seguito alla scomparsa dei rispettivi autori: è proprio durante questo secolo che la lettera torna, contrariamente a quanto era avvenuto nelle epoche immediatamente precedenti, ad acquistare un carattere riservato, privato e familiare, intimo e confidenziale.

Per rimanere in ambito italiano, rilevanti sono le produzioni epistolari di Pietro e Alessandro Verri, di Cesare Beccaria e di Romualdo De Sterlich, mentre allargando la visuale non si possono non citare i maggiori filosofi francesi, tutti impegnati in intense corrispondenze: Voltaire, Denis Diderot, Jean d'Alembert.

Da non dimenticare inoltre i carteggi dei cosiddetti “libertini”, tra cui è possibile annoverare Lorenzo Da Ponte e Giacomo Casanova, le cui opere potevano contare anche su un buon successo di pubblico. Ancora, nel Settecento cominciano ad acquistare un peso notevole i carteggi femminili: le donne rivestivano un ruolo di primo piano nei salotti letterari settecenteschi e per loro il carteggio costituisce un momento importante per la “costruzione di una identità e di una memoria / opinione del e sul salotto, e sulla donna”⁴⁸. Le lettere costituiscono inoltre uno strumento di consolidamento del salotto letterario, e in questi salotti alcune importanti intellettuali come Anne Marguerite de Lignéville, Teresa Ciamagnini, Maria

47 Claudio Donati, “Aspetti istituzionali della Chiesa di Roma tra XVII e XVIII secolo” in *Politica, vita religiosa, carità: Milano nel primo Settecento*, a cura di Marco Bona Castellotti, Edoardo Bressan e Paola Vismara, 107. Milano: Jaca Book, 1997.

48 *Salotti e ruolo femminile in Italia*, a cura di Maria Luisa Betri ed Elena Brambilla (Venezia: Marsilio, 2004), citato in Emanuela Boccassini, “I carteggi femminili: opere letterarie e di informazione” in *Ripensandoci*, n. 5 (Maggio 2009).

1. Carteggi ed epistolari

Maddalena Morelli e Madame de Staël riescono a dare vita a scambi tanto intensi e copiosi quanto vivaci e acuti.

1.5 Breve storia dell'epistolografia e dei carteggi: dall'Ottocento ai giorni nostri

Durante l'Ottocento il ruolo della lettera acquista un'importanza notevole e sono moltissimi i carteggi che possiamo leggere, appartenenti a una grandissima varietà di stili, generi, modi di scrittura. La lettera è il mezzo con il quale è nata e si è sviluppata la cosiddetta “polemica classico-romantica”, alla quale partecipano tutti i più grandi intellettuali dell'epoca. Tale polemica, scoppiata in seguito a un articolo scritto da Madame de Staël sulla rivista “Biblioteca italiana” inizia nel 1816 e si protrae, tra fervidi scambi epistolari, fino al 1818: vi prendono parte, tra gli altri, Giacomo Leopardi e Pietro Giordani (tra i classicisti), Ludovico di Breme, Giovanni Berchet e Pietro Borsieri (tra i romantici).

Un altro degli autori di cui possiamo ricostruire molti dei carteggi è Ugo Foscolo: grazie allo studio delle sue lettere possiamo conoscere la sua “propensione all'amore passionale, la disposizione ad infiammarsi”⁴⁹, il suo carattere problematico con “i suoi sbalzi d'umore, i suoi risentimenti repentini e violenti”⁵⁰.

Ma l'Ottocento, oltre a carteggi prodotti da importanti letterari, offre anche una consistente produzione epistolare d'uso quotidiano, familiare, privato, intimo. Un esempio è la corrispondenza tra Lorenzo Borsini, un giovane prete spretato senese trasferitosi a Palermo, e la moglie Gaetana: particolarmente interessanti sono le lettere di quest'ultima, lettere “senza alcuna ambizione di andare al di là dell'immediata funzione comunicativa e destinate al solo intestatario, senza alcuna prospettiva di più larga circolazione, private, anzi privatissime”⁵¹. Le lettere di Gaetana Borsini sono degne di nota anche perché la donna è analfabeta e demanda il compito di trascrivere il suo dettato a uno scrivano, a quel tempo il “mestiere [...] veramente al gradino più basso della scala lungo cui si disponevano coloro che praticassero la professione delle lettere”⁵².

L'Ottocento è anche il secolo delle lettere di viaggio, tra le quali giova ricordare quelle di Niccolò Tommaseo. All'interno di esse l'autore “alterna uno sguardo aperto sul di fuori, sul

49 Raffaele Morabito, *Lettere e letteratura. Studi sull'epistolografia volgare in Italia* (Alessandria: Edizioni dell'Orso, 2001), 134.

50 *Ibid.*

51 *Ibid.*, 159.

52 *Ibid.*

1. Carteggi ed epistolari

volto paesaggistico e umano di luoghi anche solo attraversati”⁵³: per Tommaseo il viaggio ha una funzione educatrice, per cui l'autore era solito viaggiare con foglietti e taccuini all'interno dei quali annotava le sue impressioni sui territori visitati, i suoi stati d'animo, le descrizioni dei luoghi. Le sue missive, inviate a importanti intellettuali tra i quali Gian Pietro Vieusseux, si risolvevano “interamente in un resoconto del «visto» e del «sentito»”⁵⁴.

Il secolo si chiude con un prolifico scrittore di lettere, Gabriele D'Annunzio: particolarmente interessante è la dimensione erotica delle sue missive, come quelle inviate a Barbara Leoni (una delle tante amanti del poeta), durante una relazione che inizia nel 1887 e si protrae fino al 1892. Le missive dannunziane sono scritte in modo appassionato, e ci restituiscono un'intimità descritta fin nei minimi dettagli⁵⁵.

Per quanto riguarda il Novecento, suscita notevole interesse lo scambio epistolare tra Luigi Pirandello e la sua attrice Marta Abba: uno scambio che permette di fare luce sul rapporto intercorso tra i due, protrattosi dal 1926 al 1936⁵⁶. Il carteggio è curioso anche per il fatto che da una parte si ha uno scrittore, che ovviamente scrive in maniera letterariamente impeccabile, e dall'altra invece si ha una donna la cui forma “talvolta è inappropriata” e la cui sintassi è “non sempre irreprensibile”⁵⁷, tanto che nelle lettere della Abba sono frequenti errori di grammatica e sintassi. Questo aspetto però contribuisce a dare un ritratto sincero della personalità dell'attrice.

Rimanendo nell'ambito della letteratura, si possono citare i carteggi tra Aldo Palazzeschi e Marino Moretti e tra Giovanni Papini e Ardengo Soffici.

Una delle raccolte epistolari più importanti e celebri del Novecento è costituita dalle *Lettere dal carcere* di Antonio Gramsci, non destinate alla pubblicazione da parte dell'autore e scritte tra il 1926 e il 1937. Si tratta di un insieme di lettere inviate per la maggior parte ai familiari (le rimanenti agli amici intimi: nessuna lettera viene invece indirizzata ai compagni di partito, per scelta di Gramsci), incompleto (perché molte lettere sono andate disperse) ma singolare per diversi motivi. Innanzitutto perché quasi tutte le missive “provengono da un luogo di costrizione dove la comunicazione con l'esterno diventa, per chi scrive, condizione

53 Laura Diafani, “Unsentimental travellers. La lettera di viaggio in Leopardi e in Tommaseo” in *Scrivere lettere. Tipologie epistolari nell'Ottocento italiano*, a cura di Gino Tellini (Roma: Bulzoni, 2002), 153 – 177. Citazione a pag. 167.

54 *Ibid.*

55 Carla Doni, “Le lettere erotiche di Gabriele D'Annunzio a Barbara Leoni” in *Scrivere lettere*, 259 – 277.

56 Giorgio Pullini, “Il Dramma di Pirandello nell'epistolario di Marta Abba” in *Alla lettera*, 391 – 420.

57 *Ibid.*, 395.

1. Carteggi ed epistolari

primaria per la difesa e la conservazione dell'identità biopsicologica”⁵⁸ e perché trasmettono “un impegno serrato a stabilire forme e forze di controllo per difendersi da un degrado fisiologico e psicologico che poteva trarre con sé anche un [...] temuto degrado intellettuale”.

In questo periodo sono soprattutto i carteggi politici a essere tenuti in maggior considerazione: degni di nota in tal senso sono le corrispondenze degli intellettuali antifascisti, tra i quali sono senz'altro da citare Giovanni Amendola (che intrattiene rapporti epistolari anche con Benedetto Croce), i fratelli Rosselli, Gaetano Salvemini, Filippo Turati e Piero Gobetti

Queste corrispondenze “rappresentavano il solo mezzo di comunicazione tra persone legate da sentimenti profondi”⁵⁹ e “consentono di cogliere le esitazioni, le espressioni di turbamento, così come le incertezze o le contraddizioni di quegli antifascisti che si apprestavano, di volta in volta, a raccontare e a raccontarsi alle persone più care”⁶⁰.

Successivamente, a seguito del successo di diversi mezzi di comunicazione (il telefono, la radio, i giornali), la lettera è passata in secondo piano, ma negli ultimi anni si è assistito a “un singolare rinnovamento del genere epistolografico, tutte le volte che scriviamo (digitiamo) e scambiamo messaggi di posta elettronica (e-mail) privati o ufficiali, o anche, a un più basso ma non meno vitale livello stilistico, messaggi SMS, che somigliano talora a mini-lettere, o più spesso epigrammi”⁶¹. I nuovi modi di comunicare forse hanno aperto anche nuove strade alla corrispondenza epistolare, tanto che negli ultimi anni è cresciuto il numero di libri aventi per oggetto scambi di e-mail tra due o più personaggi: l'ultimo in ordine di tempo è il romanzo epistolare (dove per “epistola” si intende la più moderna e-mail) *Le ho mai raccontato del vento del Nord?*, scritto dall'austriaco Daniel Glattauer.

1.6 I carteggi d'artista

Un discorso separato meritano i carteggi d'artista. Il primo caso noto in cui gli artisti appaiono come destinatari, ma in un manuale di corrispondenza (come tanti dell'epoca), risale al 1484, quando come ipotetici destinatari di una possibile epistola artistica sono indicati

58 Saveria Chemotti, “Oltre l'«Hortus conclusus»: le *Lettere dal Carcere* di Antonio Gramsci” in *Alla lettera*, 421 – 494. Citazione a pag. 423.

59 Sara Galli, “Antifascisti tra carcere, clandestinità ed esilio. La 'riscoperta' dei carteggi” in *Storicamente* 1 (2004), http://www.storicamente.org/04_comunicare/rassegne/galli.htm.

60 *Ibid.*

61 Ruggero Cattaneo, “O stilu”, 91.

1. Carteggi ed epistolari

Andrea Mantegna e Melozzo da Forlì nella categoria dei pittori e Leon Battista Alberti nella categoria degli architetti⁶². In materia epistolografica i nomi degli artisti circolano più che altro nei manuali, e i pittori, gli scultori e gli architetti prediletti sono quelli che, secondo gli intellettuali del tempo (come Flavio Biondo e Bartolomeo Facio)⁶³ godevano anche di una certa fama letteraria, oltre che artistica.

Il Rinascimento maturo offre anche lettere d'artista, e quelle più note sono sicuramente quelle dei più importanti geni del periodo, a cominciare da Leonardo (che avrebbe scritto lettere “eminente tecniche”, che escluderebbero un loro “riciclaggio letterario”⁶⁴; notazione di cui tener conto in un periodo in cui l'epistola era vista più come esercizio letterario che come strumento di comunicazione), passando attraverso Raffaello (la cui fama “vive negli epistolari grazie soprattutto alla lettera al Castiglione, a lungo edita nelle sillogi venete [...] anche con l'attribuzione a Pietro Aretino”⁶⁵), arrivando fino a Tiziano, amico nonché abituale corrispondente dell'Aretino, e a Michelangelo, i cui carteggi sono conosciutissimi e sono stati pubblicati a più riprese. Molto importante anche il fitto carteggio di Giorgio Vasari. Ma sono anche altri i destinatari delle lettere nel Cinquecento: tra i più grandi è possibile citare Francesco Salviati, Taddeo Zuccari (destinatari di missive presenti nel carteggio di Annibal Caro), Giulio Romano, Sebastiano Serlio, Bartolomeo Ammannati, il Sodoma, Baccio Bandinelli, Lorenzo Lotto, Paris Bordon, il Tintoretto, Iacopo Sansovino (corrispondenti di Pietro Aretino, che ebbe un vivace carteggio anche con il Vasari).

Cambiando secolo, si possono citare i carteggi di Cassiano dal Pozzo e del cardinale Leopoldo de' Medici; il primo (noto per l'attribuzione del nome di “Gioconda” al capolavoro di Leonardo), ha uno stretto rapporto con Nicolas Poussin e si rivolge anche ad artisti come Artemisia Gentileschi e Pietro da Cortona, mentre il secondo corrisponde con Ciro Ferri e con il Volterrano⁶⁶.

Contemporaneamente cominciano a svilupparsi le raccolte di lettere d'artisti: uno dei primi esperimenti è realizzato da Carlo Cesare Malvasia, che realizza una *Raccolta di lettere pittoriche*. L'impresa dell'autore della *Felsina pittrice* viene imitata anche nei secoli successivi, a cominciare da Giovanni Bottari, che nel Settecento comincia a raccogliere tutti i

62 Paola Barocchi, “Fortuna della epistolografia artistica” in *Metodologia ecdotica dei carteggi*, a cura di Elio D'Auria (Firenze: Le Monnier, 1989), 104 – 133.

63 *Ibid.*, 105.

64 *Ibid.*, 106.

65 *Ibid.*

66 *Ibid.*, 115.

1. Carteggi ed epistolari

carteggi cinquecenteschi e secenteschi dei maggiori artisti (tutti quelli citati in precedenza): la raccolta di Bottari ha un notevole successo e di fatto stimola numerose nuove raccolte, come quelle ottocentesche di Stefano Ticozzi (che continuò l'impresa di Bottari) e di Johann Gaye. Operazioni siffatte sono state svolte anche durante il Novecento, per esempio con una raccolta dei carteggi degli Impressionisti pubblicata nel 1939, e continuano fino ai giorni nostri.

Tornando ai carteggi d'artista, importanti sono, per quanto riguarda il Settecento, quelli, decisamente fitti, di Antonio Canova, che tra i suoi corrispondenti poteva annoverare Pietro Giordani e Antoine Quatremère de Quincy, nonché quelli di Giambattista Tiepolo, assiduo corrispondente di Francesco Algarotti. Passando al secolo successivo è doveroso citare i carteggi di Giovanni Maria Benzoni, di Adeodato Malatesta, di Giovanni Fattori, di Giovanni Boldini e, ovviamente, quello di Carlo Finelli. Per quanto riguarda invece il Novecento, degni di nota sono gli epistolari degli artisti futuristi, spesso riuniti in raccolte collettive⁶⁷.

1.7 Problemi legati alla consultazione dei carteggi

La consultazione dei carteggi, qualora non pubblicati in rete, dà luogo ad alcuni problemi che possono comprometterne la fruizione da parte di un pubblico non “addetto ai lavori”. Gli originali delle lettere sono conservati presso Archivi Storici di istituzioni, enti, biblioteche, accademie e l'accesso a tali fondi non è consentito se non per motivi di ricerca o di studio (e spesso se non si possiede lo status di ricercatore, l'accesso viene impedito), previa autorizzazione da parte dell'ente all'interno del quale si vuole svolgere la propria ricerca. Alcuni esempi:

- Archivio Storico dell'Accademia di Belle Arti di Venezia: “per accedere alla sala di studio è necessario presentare una richiesta nella quale si dichiara, oltre ai dati personali, l'argomento e lo scopo della ricerca. La domanda può essere inoltrata via mail o per posta al Direttore dell'Accademia”⁶⁸;
- Archivio Storico dell'Accademia di Brera: “l'accesso è riservato al personale interno o a ricercatori autorizzati dalla direzione”⁶⁹;

⁶⁷ Per esempio, *Futurismo: scritti, carteggi, testimonianze*, a cura di Paolo Buzzi (Milano: Biblioteca Comunale di Milano, 1981) o *Carteggio futurista*, a cura di Francesco Muzzioli (Roma: Officina Edizioni, 1987).

⁶⁸ <http://www.accademiavenezia.it/archivio.php>

⁶⁹ http://www.accademiadibrera.milano.it/ritait_patrim.asp?sel=1

1. Carteggi ed epistolari

– Archivio Storico del Comune di Ferrara: “L'accesso all'Archivio Storico è consentito a studiosi, studenti, ricercatori ecc. PER MOTIVI DI STUDIO, previa compilazione di domanda su modulo in carta libera, disponibile c/o l'archivio”⁷⁰;

– Archivio Storico dell'Accademia delle Scienze di Torino: “All'Archivio storico dell'Accademia delle Scienze sono ammessi i Soci e gli studiosi muniti di una lettera di presentazione (con firma leggibile) di un Socio che, sottoscrivendola, si rende garante del loro comportamento: la presentazione vale per l'anno accademico ed è rinnovabile”⁷¹;

– Biblioteca dell'Accademia dei Lincei: “l'accesso è consentito ai Soci dell'Accademia e agli studiosi che si rechino in biblioteca per motivi di ricerca”⁷².

Risulta quindi evidente che spesso (lampante è il caso dell'Archivio Storico dell'Accademia delle Scienze di Torino) si incontrano difficoltà qualora si voglia accedere a fondi che conservano documenti antichi.

Un altro problema legato alla consultazione può essere quello della “distanza fisica” dai manoscritti: uno studioso che voglia analizzare i manoscritti, qualora questi ultimi non siano pubblicati in rete, può essere costretto a lunghe e dispendiose trasferte. Questo aspetto è particolarmente importante soprattutto se si pensa agli studiosi indipendenti, che non sono supportati da istituzioni accademiche o enti privati e quindi non possono sobbarcarsi l'onere del viaggio. È tuttavia necessario considerare che l'eliminazione dei costi di gran parte delle trasferte può essere un'interessante fonte di risparmio anche per università ed enti privati.

Ancora, molti manoscritti versano in delicate condizioni di conservazione: la fragilità dei supporti cartacei dei manoscritti e la perdita di inchiostro dovuta al ripetuto uso sono fattori che possono causare il rapido deterioramento dei manoscritti originali, motivo per il quale spesso la consultazione di diversi documenti non viene consentita.

Per ovviare alle problematiche di cui sopra è possibile ricorrere alle pubblicazioni cartacee dei carteggi, ma anche in questo caso si pongono problemi di diversa natura. In primo luogo, è necessario riuscire a reperire una pubblicazione del carteggio desiderato, e spesso l'operazione non è facile: se esistono ormai pubblicazioni, anche economiche, dei carteggi dei più grandi autori, ciò non è altrettanto vero per artisti di fama minore. In più, molto spesso le pubblicazioni sono molto costose e gli studiosi, soprattutto se indipendenti, non possono permettersi l'acquisto. Problemi di natura economica nascono anche qualora la pubblicazione

70 <http://servizi.comune.fe.it/index.phtml?id=3383>

71 <http://www.accademiadelle scienze.it/archivio/regolamento>

72 <http://www.lincci.it/modules.php?name=Content&pa=showpage&pid=12>.

1. Carteggi ed epistolari

sia disponibile in una biblioteca lontana dal luogo di residenza dello studioso, che è così costretto a una trasferta per il suo studio.

A causa di tutti questi motivi, si comprende quanto possa essere necessaria la fruizione telematica dei carteggi.

2. La fruizione telematica dei carteggi

2.1 Informatica per l'archivistica e importanza della fruizione telematica dei carteggi

La molteplicità delle opportunità offerte dall'informatica ha apportato negli ultimi anni un notevole contributo in materia di fruizione telematica non solo dei carteggi, ma dei documenti storici in generale⁷³: le richieste di fruizione dei beni culturali, tra i quali rientrano a pieno titolo i documenti storici, hanno spinto gli archivisti a valutare le possibilità dell'informatica essenzialmente per tre motivi, ovvero la conservazione del materiale, l'ordinamento e l'inventariazione, la consultazione⁷⁴. Benché l'avvento dell'informatica abbia dato luogo a numerosi dibattiti⁷⁵, è necessario che gli archivisti si aprano alle novità e valutino come sfruttare al meglio i grandi mutamenti che l'informatica ha portato all'interno delle discipline umanistiche.

Federico Valacchi⁷⁶, che ha dedicato gran parte della sua attività allo studio dei rapporti tra informatica e archivistica, sostiene che l'informatica sia in grado di offrire tre diverse soluzioni legate alla conservazione e alla fruizione del patrimonio documentaristico: “software gestionali”, “sistemi informativi” e “risorse telematiche come strumenti integrativi e di potenziamento delle capacità di valorizzazione e fruizione”. I “software gestionali” sarebbero quelli “orientati alla descrizione e al riordino di fondi archivistici secondo modalità molto vicine al tradizionale lavoro degli archivisti” e rappresenterebbero “il primo anello della catena tecnologica in ambito archivistico”, dal momento che “la loro diffusione agevola il passaggio verso la creazione di più sofisticati sistemi di accesso alle fonti archivistiche”. Sono sostanzialmente strumenti ideati per agevolare il lavoro dell'archivista e pensati e realizzati secondo le tecniche della sua disciplina.

I “sistemi formativi archivistici” e le “risorse telematiche” sono invece strumenti pensati per la fruizione da parte del pubblico. I “sistemi informativi archivistici” sarebbero strumenti che “utilizzando o, per meglio dire, inglobando processi di digitalizzazione totali o parziali

73 A tal proposito si vedano Maria Guercio, *Archivistica informatica: i documenti in ambiente digitale* (Roma: Carocci, 2002) e Federico Valacchi, *La memoria integrata nell'era digitale. Continuità archivistica e innovazione tecnologica* (Corazzano: Titivillus, 2006).

74 *Consegnare la memoria: manuale di archivistica ecclesiastica*, a cura di Emanuele Boaga, Salvatore Palese e Gaetano Zito (Firenze: Giunti, 2003).

75 In particolare, si osserva (forse per tradizione, per sfiducia, per pregiudizio o per chiusura) una certa resistenza all'avvento dell'informatica nell'ambito dell'archivistica. Si veda a tal proposito Federico Valacchi, “Applicazioni tecnologiche agli archivi storici. Le risorse per la descrizione, il riordino e la utilizzazione delle fonti archivistiche” in *La Gestione dell'Archivio nelle Organizzazioni Non Profit*, a cura di Domenico Muscò (Firenze: CESVOT, 2002), 67 – 82.

76 Federico Valacchi, “Applicazioni tecnologiche agli archivi storici”.

2. La fruizione telematica dei carteggi

delle fonti cartacee si rivelano particolarmente congeniali alla ricerca archivistica attraverso la rete”. Attraverso il sistema informatico archivistico sarebbe possibile compiere “il passaggio successivo a quello del riordino, armonizzando, nell'ottica della fruizione, le diverse componenti informative che caratterizzano il materiale archivistico e consentendo percorsi di ricerca flessibili e capaci di ricostruire tale complessità”. È proprio in questa direzione che volge il lavoro della presente tesi: cercare di offrire una metodologia il più possibile completa e aggiornata, che tenga conto di tutte le problematiche poste dall'acquisizione dei carteggi d'artista e che possa permettere ricerche flessibili e approfondite.

Tutto ciò però non sarebbe possibile (o meglio, sarebbe possibile ma fortemente limitante per il pubblico) senza la terza soluzione informatica per l'archivistica proposta da Valacchi, ovvero le “risorse telematiche per la ricerca archivistica”. L'uso della rete è visto come “strumento di amplificazione del lavoro svolto su singoli fondi o su complessi documentari conservati nei diversi istituti” nonché come “possibile opportunità per la creazione di sistemi informativi archivistici integrati”. L'ampiezza del pubblico che si può raggiungere attraverso la rete è tale che un archivio, al fine di promuovere la conoscenza dei documenti in esso contenuti, dovrebbe considerare seriamente l'impatto che le risorse telematiche potrebbero avere su quanti siano interessati a svolgere ricerche mediante l'ausilio di internet.

La rete può inoltre configurarsi come autonomo strumento di ricerca per gli studiosi: questo aspetto rappresenta una rivoluzione nel campo dell'archivistica, soprattutto se si pensa al fatto che gli addetti ai lavori “non hanno ancora del tutto superato l'idea che la rete costituisca un gadget e che le informazioni offerte dai siti di natura archivistica siano al massimo una sorta di supplemento agli strumenti di ricerca 'tradizionali’”. Ciò è particolarmente grave perché la ricerca *on-line* deve essere vista come una evoluzione della ricerca tradizionale, ed è necessario acquistare la consapevolezza che le risorse telematiche possono offrire più di un aiuto in diversi ambiti.

2.2 Problematiche relative alla fruizione telematica dei carteggi

Benché siano enormi le potenzialità offerte dalla rete nell'ambito dello studio dei carteggi d'artista, bisogna precisare che la realizzazione di edizioni telematiche di carteggi non sia esente da problematiche di varia natura. Tali problematiche sono individuabili:

- nella difficoltà di riprodurre in versione telematica l'edizione critica di un carteggio

2. La fruizione telematica dei carteggi

pubblicato in formato cartaceo;

- nella presenza, soprattutto nei carteggi d'artista, di disegni che servono per meglio chiarire o per esemplificare un concetto spiegato all'interno della missiva;
- nella difficoltà di lemmatizzare adeguatamente un carteggio, dal momento che ogni epoca storica presenta linguaggi e modi di scrivere diversi;
- nella difficoltà di proporre un metodo standard.

Nei prossimi paragrafi si cercherà di analizzare singolarmente ognuna di queste problematiche e di proporre soluzioni magari non definitive ma che possano consentire comunque la realizzazione di prodotti di qualità e scientificamente validi.

2.3 Problemi: le edizioni critiche

L'edizione critica di un testo è “un'impresa scientifica fondata su procedure di pubblicazione relativamente formalizzate, che aspirano all'oggettività e che, storicamente, si riconoscono in scuole e metodi diversi”⁷⁷. Realizzare un'edizione critica di un testo significa affrontare e porre in atto tecniche che possano garantire che la pubblicazione sia più corretta e rispettosa possibile nei confronti dell'originale.

L'edizione critica di un testo è un'operazione molto delicata e faticosa, soprattutto qualora non sopravvivano gli originali e si debba quindi provvedere a ricostruirli nel modo più fedele possibile (*constitutio textus*), cercando di discernere, all'interno delle copie (magari anche di diverse epoche successive: è il caso dei testi dell'antichità classica) quali possano essere gli errori di trascrizione (voluti o meno), in quali casi sia avvenuta una contaminazione (ovvero un errore trasmesso da una copia a un'altra che non è una sua discendente), quali sono le varianti di una copia rispetto ad altre.

Il filologo, al termine del suo lavoro di ricostruzione, doterà il testo di un “apparato critico”, che dovrà contenere la giustificazione scientifica dell'edizione: il filologo, all'interno dell'apparato, dovrà quindi dare conto al pubblico dei criteri seguiti, delle scelte fatte, delle varianti scartate, di quali sono i “testimoni” (termine con il quale in filologia si indicano le copie di un testo del quale non si possiede l'originale) di cui si è tenuto conto durante le fasi di lavoro, e in generale dovrà comunicare al meglio le operazioni eseguite nella realizzazione dell'edizione critica.

⁷⁷ Bruno Bentivogli, Paola Vecchi Galli, *Filologia italiana*, (Milano: Mondadori, 2002), 41.

2. La fruizione telematica dei carteggi

di fruizione del testo⁷⁹.

L'edizione critica, quando trasposta sul computer, può innanzitutto “mettere a disposizione i vari stati del testo, visualizzando le immagini dei manoscritti ma anche gestendo in modo straordinariamente affidabile i fenomeni dei testimoni quali la lingua, la grafia, le varianti, le cancellature, le correzioni, le lacune, il processo evolutivo (riscritture, autocitazioni, autocommenti, chiose)”. Le possibilità date dagli ipertesti e in particolare dai linguaggi di marcatura risultano pressoché sconfinite: è però necessario che la codifica avvenga in modo rigoroso.

Grazie alle potenzialità degli ipertesti è stato possibile produrre edizioni elettroniche all'interno delle quali si è potuto rappresentare l'intero processo di formazione dell'opera letteraria⁸⁰, oppure ancora realizzare siti che presentassero i risultati di edizioni critiche realizzate da diversi filologi. È il caso di RIALTO (*Repertorio Informatizzato dell'Antica Letteratura Trobadorica e Occitana*, www.rialto.unina.it), un portale realizzato dall'Università di Napoli con l'intento di “immettere in rete, in edizioni critiche affidabili, l'intero corpus letterario occitano medievale”⁸¹. La figura di seguito mostra una pagina che propone un confronto tra quattro diverse edizioni critiche di una lirica di Arnaut Daniel. La pagina è dotata di note esplicative che propongono all'utente un esauriente commento delle edizioni prese in considerazione:

The screenshot shows a web browser window titled "Arnaut Daniel (29.7)". The page content is as follows:

Canello	Toja	Perugi	Ensebi
D'otra guiza e d'otra razon m'aven a chantar que no sol, e nous cugetz que de mon dol esper a far bona chanson, mas mestiers m'es qu'ieu fassa merceiar a mans, chantan leis que m'encolp'a tort, qu'ieu n'ai lezer qu'estiers non parl'ab [tres.	D'otra guiz'e d'otra razon m'aven a chantar que no sol, e no us cugetz que de mon dol esper a far bona chanson, mas mestiers m'es qu'ieu fassa merceiar a mans, chantan leis que m'encolp'a tort, qu'ieu n'ai lezer, qu'estiers non parl'ab [tres.	D'otra guiza e d'otra razo m'aven a chantar que no sol, e no us cuges que de mon dol esper a far bona chanso, mas mestier m'es qu'ieu fassa merceiar a mans, chantan leis que m'encolp'a tort, qu'ieu n'ai lezer qu'estiers non parl'a [tres.	D'otra guiz'e d'otra razo m'aven a chantar que no sol, e no us cugetz que de mon dol esper a far bona chanso, mas mestiers m'es qu'ieu fassa merceiar a mans, chantan leis qui m'encolp'a tort, qu'ieu n'ai lezer, qu'estiers non parl'ab [tres.
Merce dei trobar e perdon si l dreit usatzes nom destol tal que de merceiar nom tol. Ja salvet merces lo lairon quez autre bes no l podia salvar; ieu non ai plus vas ma vida cofort que, si l dreitz qu' ai nom val, vaillam [merces.	Merce dei trobar e perdon si l dreit usatzes no m destol tal que de merceiar no m tol. Ja salvet merces lo lairon quez autre bes no l podia salvar; ieu non ai plus vas ma vida cofort que, s' il dreitz qu' ai no m val, vailha m [merces.	Merce dei trobar e perdo si l dreitz uzatzes no m destol tal que de merceiar no m col: ja salvet Merce lo lairo que autre be no l podia salvar; ieu non ai plus vas ma vida cofort que, si l dreit c' ai no m val, vailla m [Merces.	Merce dei trobar e perdo si l dreitz uzatzes no m destol tal que de merceiar no m tol: ja salvet Merces lo lairo que autre be no l podia salvar; ieu non ai plus ves ma vida cofort que, si l dreitz qu' ai no m val, valha m [Merces.
Donc ha hom dreg en amor? Non, mas cudiarion so li fol, qu' elaus encolpara, si s vol, quar li Frances no son Guascon e quar la naus frais ans que fos a Bar. Las! per aital colpa sui pres de mort, que d' als, per Crist, no sai qu' anc tort [l'agues.	Donc ha hom dreg en amor? Non, mas cujarion so li fol: qu' el' us encolpara, si s vol, quar li Frances no son Guascon e quar la naus frais ans que fos a Bar. Las! per aital colpa sui pres de mort, que d' als, per Crist, no sai qu' anc tort [l'agues.	15 Donc a hom dreg en Amor? No, mas cujarion-s' o li fol: qu' ela us encolpara, si s vol, car li Frances no son Gasco e car la naus frais ans que fos a Bar: las, per aital colpa soi pres de mort, que d' als, per Crist, no sai qu' ieu tort [l'agues.	Donc a hom dreg en Amor? No, mas cujarion-s' o li fol, qu' ela us encolpara, si s vol, quar li Frances no son Guasco e quar la naus frais ans que fos a Bar: las! per aital colpa sui pres de mort, que d' als, per Crist, no sai qu' ieu tort [l'agues.

FIG. 2.2: confronto tra quattro edizioni critiche di una lirica di Arnaut Daniel nel portale RIALTO.

79 *Ibid.*, 24.

80 Andrea Bozzi, “Edizione elettronica e filologia computazionale” in *Fondamenti di critica testuale*, a cura di Alfredo Stussi (Bologna: Il Mulino, 2006), 210-211.

81 <http://www.rialto.unina.it/limen.htm>.

2. La fruizione telematica dei carteggi

La grande opportunità offerta dall'informatica è anche quella di poter consentire di osservare, all'interno del sito che ospita l'edizione critica, le immagini dei testimoni utilizzati per la sua realizzazione: la stessa operazione pensata in formato cartaceo sarebbe poco probabile. Inoltre “la disponibilità delle immagini [...] rende più veloce, più comoda e probabilmente anche più sicura la comparazione paleografica volta a risolvere problemi di datazione e/o localizzazione grazie al raffronto con altri documenti disponibili in formato digitale e collocati in precisi ambiti spazio-temporali”⁸²: l'informatica diventa così anche uno strumento aggiuntivo per il filologo, che può eseguire il suo lavoro avvalendosi dei mezzi che la tecnologia mette a sua disposizione per non dover agire sulle fonti, che spesso presentano spinosi problemi di conservazione.

Di seguito si propone un'immagine tratta dal progetto *Nietzsche Source*⁸³ (www.nietzschesource.org), un sito “dedicato alla pubblicazione di contenuti scientifici riguardanti la vita e l'opera di Friedrich Nietzsche”⁸⁴, una delle imprese editoriali più significative nell'ambito dell'edizione elettronica di scritti d'autore:

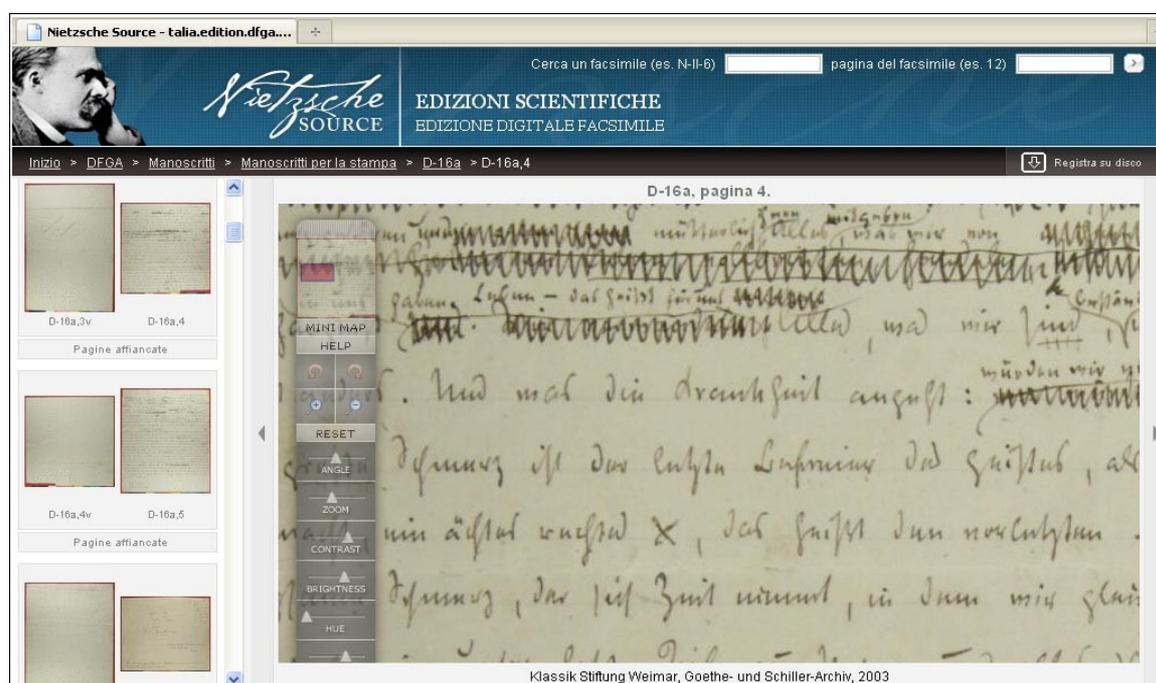


FIG. 2.3: pagina web con alcune righe tratte da un manoscritto autografo di Friedrich Nietzsche per la stampa della nuova edizione de *La Gaia Scienza*, contenente il Prologo, il V Libro, le “Canzoni del principe Vogelfrei”.

82 Andrea Bozzi, “Edizione elettronica e filologia computazionale”, 209-210.

83 Il progetto è stato realizzato da *Nietzsche Source Organization*, già *Association HyperNietzsche*, un'associazione senza scopo di lucro (con sede presso l'École normale supérieure di Parigi) il cui intento è formare una comunità internazionale di studiosi di Friedrich Nietzsche.

84 <http://www.nietzschesource.org/documentation/it/home.html>

2. La fruizione telematica dei carteggi

Il progetto contiene sia l'edizione critica dei testi del filosofo tedesco, sia le immagini dei manoscritti, che si possono consultare gratuitamente e in modo integrale

Il sistema realizzato dal progetto *Nietzsche Source* dà all'utente la possibilità di vedere i testi da vicino, ingrandirli a piacimento e regolare alcune caratteristiche come luminosità e contrasto al fine di ottenere una lettura ottimale.

L'informatica mette quindi a disposizione del filologo diversi mezzi con i quali poter riprodurre fedelmente in rete l'edizione critica di un testo, con in più il valore aggiunto che può derivare dall'utilizzo degli ipertesti e dallo sfruttamento della multimedialità, grazie ai quali è possibile fare in modo, come si è visto, che si possano vedere i manoscritti autografi di un autore stando comodamente seduti davanti al proprio personal computer.

2.4 Problemi: i disegni

È possibile che i carteggi degli artisti possano presentare, assieme al testo, anche disegni esplicativi realizzati per chiarire meglio un concetto espresso durante la lettera. È il caso, per esempio, di questa lettera scritta da Vincent van Gogh a un suo amico, il pittore Anton van Rappard, l'8 marzo del 1884:

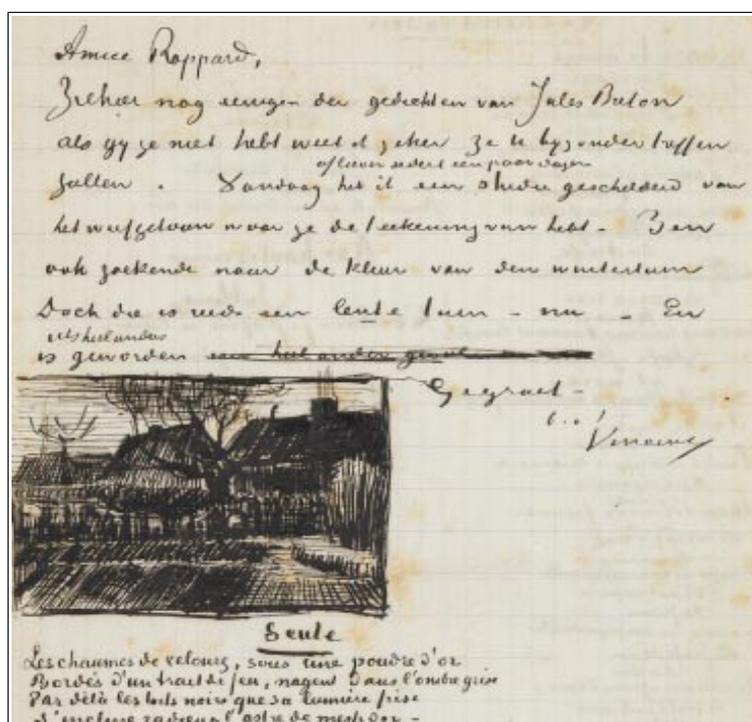


FIG. 2.4: incipit della lettera inviata da Vincent van Gogh ad Anton van Rappard l'8 marzo 1884.

2. La fruizione telematica dei carteggi

In casi come questi è necessario trovare una soluzione per far sì che si possa garantire all'utente di fruire in modo pieno della lettera. Il sito *Vincent Van Gogh – The Letters* (www.vangoghletters.org)⁸⁵ fornisce alcune efficaci risposte al problema che i disegni contenuti nei carteggi d'artista possono presentare.

La prima soluzione è la più intuitiva e in parte se ne è già parlato durante il paragrafo precedente: si tratta di fornire, assieme alla trascrizione, anche una copia virtuale dei manoscritti autografi, magari inserendo, in una parte della pagina che ospita la lettera, un collegamento (o più collegamenti, se i fogli sono più d'uno) che rimanda all'immagine del manoscritto. L'immagine potrà essere ospitata nello stesso dominio sul quale sono presenti le trascrizioni.

Altrimenti è possibile inserire dei link alle singole immagini nei punti della lettera in cui il disegno è presente. Il sito *Vincent Van Gogh – The Letters* offre la prima delle due opportunità. La pagina che ospita la trascrizione della lettera è divisa in due parti, e l'utente può scegliere cosa visualizzare in ambedue le parti. Di seguito si mostra un'immagine di una pagina del sito all'interno della quale è possibile vedere sulla parte sinistra la trascrizione della lettera, e sulla destra il corrispettivo originale:

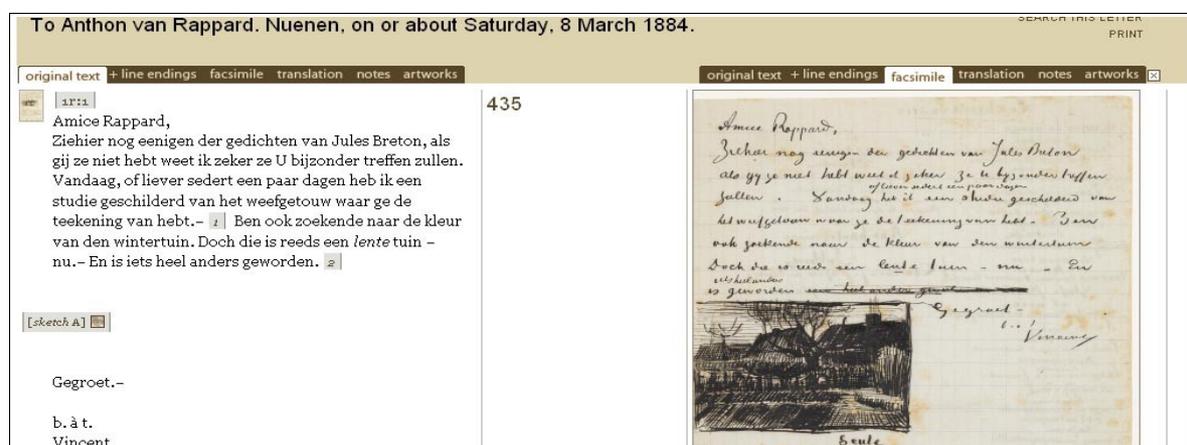


FIG. 2.5: una pagina del sito *Vincent Van Gogh – The Letters*

Risulta comunque evidente che per una più comoda analisi del manoscritto è necessario avvalersi degli strumenti di zoom al fine di poterlo leggere nel migliore dei modi. La doppia visualizzazione però risulta particolarmente utile perché al posto della visuale del manoscritto, all'interno di *Vincent Van Gogh – The Letters* è possibile anche leggere la traduzione in inglese dei carteggi: una soluzione molto favorevole per quanti non conoscono l'olandese.

⁸⁵ Per un'analisi più approfondita del sito cfr. *infra* PAR. 4.6.

2. La fruizione telematica dei carteggi

Qualora quindi non si decida di visualizzare contemporaneamente la trascrizione e il manoscritto, gli autori del progetto hanno pensato di inserire, nei punti in cui compare il disegno, un collegamento all'intero foglio del manoscritto dove è possibile trovare il disegno stesso.

2.5 Problemi: la lingua

I carteggi degli artisti non sono tutti scritti nella stessa lingua: in ogni epoca si scrive utilizzando una lingua diversa, e uno dei problemi più delicati, quando si realizza un corpus epistolare lemmatizzato, è utilizzare un programma che possa eseguire una lemmatizzazione adeguata per una certa varietà della lingua (l'italiano del Trecento, l'italiano del Cinquecento, l'italiano contemporaneo). Per “lemmatizzazione” si intende l'operazione con la quale a ogni occorrenza del carteggio si associa un lemma⁸⁶. Una “occorrenza” non è altro che la singola comparsa di una forma all'interno del corpus, mentre un “lemma” è una forma-base alla quale viene riportato un insieme di forme che si distinguono tra loro solo per l'assetto grafico (varianti grafiche, con o senza valore di varianti fonetiche)⁸⁷, e infine una “forma” è una singola parola, distinta dalle altre esclusivamente in base all'assetto grafico, che può comparire un qualsiasi numero di volte nel corpus⁸⁸. Un programma di lemmatizzazione pensato per varianti moderne dell'italiano potrà quindi dare ottimi risultati qualora riceva in input un testo scritto in italiano moderno, mentre al contrario i risultati potrebbero non essere quelli sperati se il testo sottoposto appartiene a una variante della lingua troppo distante da quella per la quale il programma di lemmatizzazione è stato pensato.

Nelle tabelle seguenti si mostrano i risultati della lemmatizzazione di due lettere utilizzando MAGIC, un programma di analisi morfologica per l'italiano di oggi (con il quale è stato lemmatizzato anche *Carlo Finelli Corpus*). La prima è una lettera scritta da Lorenzo il Magnifico e inviata al duca di Ferrara Ercole I d'Este il 13 giugno 1485⁸⁹, e la seconda è una lettera inviata dal poeta Giuseppe Raimondi al pittore Primo Conti nel 1918⁹⁰.

86 *Glossario in Guida a GattoWeb* (Gestione degli Archivi Testuali del Tesoro delle Origini). L'indirizzo web della risorsa è <http://gattoweb.ovi.cnr.it/%28S%28bk1d01acrpnfzj3waj042uym%29%29/HelpGattoWeb/Glossario.html>.

87 *Ibid.*

88 *Ibid.*

89 Pubblicata in *Lettere di Lorenzo de' Medici detto il Magnifico nell'Archivio Palatino di Modena*, a cura di Antonio Cappelli (Modena: Vincenzi, 1863).

90 Pubblicata in *Primo Conti – Giuseppe Raimondi: Carteggio*, a cura di Patrizia Mania (Roma: Edizioni di Storia e Letteratura, 2001).

2. La fruizione telematica dei carteggi

Prieghola PRIEGHOLA#SP@NN#
quanto QUANTO#B@# QUANTO#CC@# QUANTO#D@MS# QUANTO#P@MS# QUANTO#S@MS#
piu' PIU'#A@FP@FS@MP@MS# PIU'#B@# PIU'#E@# PIU'#S@MP@MS#
amorevolmente AMOREVOLMENTE#B@#
posso POTERE#V@S1IP#
che CHE#CC@# CHE#CS@# CHE#D@FP@FS@MP@MS# CHE#P@FP@FS@MP@MS#
ad AD#E@#
mia MIA#S@FS# MIO#A@FS# MIO#P@FS#
instantia
voglia VOLERE#V@S1CP@S2CP@S3CP# VOGLIA#S@FS#
commettere COMMITTERE#V@F#
a A#E@# A#S@FP@FS#
qualch'
uno UNO#N@MS# UNO#P@MS# UNO#RI@MS# UNO#S@FP@FS@MP@MS#
delli
suoi SUO#A@MP# SUO#P@MP#
che CHE#CC@# CHE#CS@# CHE#D@FP@FS@MP@MS# CHE#P@FP@FS@MP@MS#
se SE#CS@# SE#PQ@FP3@FS3@MP3@MS3# SE#S@MP@MS#
informi INFORMARE#V@S1CP@S2CP@S3CP# INFORME#A@FP@MP#
quello QUELLO#D@MS# QUELLO#P@MS#
che CHE#CC@# CHE#CS@# CHE#D@FP@FS@MP@MS# CHE#P@FP@FS@MP@MS#
et
nell' IN#E@FS@MS#
una UNA#N@FS# UNA#P@FS# UNA#RI@FS# UNA#S@FP@FS#
et
altra ALTRO#A@FS# ALTRO#D@FS# ALTRO#P@FS#
causa CAUSARE#V@S2MP@S3IP# CAUSA#S@FS#
soporta SOPPORTARE#V@S2MP@S3IP#
la LO#RD@FS# LA#PQ@FS3# LA#S@MP@MS#
iustitia

TAB. 2.1: lemmatizzazione di una lettera di Lorenzo il Magnifico

eseguita con un analizzatore morfologico per l'italiano contemporaneo

Come si può ben notare della tabella, anche senza conoscere il tagset (ovvero l'insieme di etichette, come “SP@NN”, utilizzate per la codifica), del quale comunque si parlerà in modo più dettagliato nel prosieguo della trattazione⁹¹, il programma non ha riconosciuto diverse forme (“instantia”, “qualch”, “delli”, “et”, “iustitia”) e ha riconosciuto in modo errato altre forme: per esempio la prima, “Prieghola”, è stata riconosciuta come un nome proprio (SP).

I risultati riscontrati nell'analisi della lettera di Giuseppe Raimondi sono invece ben diversi:

Caro CARO#A@MS# CARO#S@MS# CARO#SP@NN#
amico AMICO#A@MS# AMICO#S@MS# AMICARE#V@S1IP#
grazie GRAZIA#S@FP# GRAZIE#I@#
del DI#E@MS#
libro LIBRARE#V@S1IP# LIBRO#S@MS#
. #@@#
Senta SENTIRE#V@S1CP@S2CP@S3CP# SENTA#SP@NN#
: #@@#
sinceramente SINCERAMENTE#B@#

91 Cfr. *infra* PAR. 5.3.

2. La fruizione telematica dei carteggi

```
,#@@#  
lei LEI#PQ@FS3#  
e' ESSERE#V@S3IP#  
uno UNO#N@MS# UNO#P@MS# UNO#RI@MS# UNO#S@FP@FS@MP@MS#  
dei DI#E@MP# DIO#S@MP#  
migliori MIGLIORARE#V@S1CP@S2CP@S2IP@S3CP# MIGLIORE#A@FP@MP#  
MIGLIORE#S@FP@MP#  
scrittori SCRITTORE#S@MP# SCRITTORIO#A@MP# SCRITTORIO#S@MP#  
che CHE#CC@# CHE#CS@# CHE#D@FP@FS@MP@MS# CHE#P@FP@FS@MP@MS#  
ci CI#B@# CI#PQ@FP1@MP1#  
siano ESSERE#V@P3CP#  
li' LI#B@#  
nel IN#E@MS#  
gruppo GRUPPO#S@MS#  
futurista FUTURISTA#A@FS@MS# FUTURISTA#S@FS@MS#  
.#@@#
```

TAB. 2.2: lemmatizzazione di una lettera di Giuseppe Raimondi

eseguita con un analizzatore morfologico per l'italiano contemporaneo

La soluzione migliore per ovviare a questo problema sarebbe disporre di analizzatori morfologici diversi a seconda della varietà della lingua. Uno dei più importanti programmi per quel che riguarda l'italiano antico è GATTO (*Gestione degli Archivi Testuali del Tesoro delle Origini*), un software creato da Domenico Iorio-Fili dell'Istituto Opera del Vocabolario Italiano (OVI) del CNR⁹². GATTO è un software “nato come strumento finalizzato alla costruzione, gestione ed interrogazione del corpus di testi che è alla base del Vocabolario Storico della Lingua Italiana, in corso di realizzazione presso l'OVI”⁹³. Con GATTO è possibile gestire corpora testuali: si possono aggiungere e modificare testi, svolgere ricerche lessicografiche nonché lemmatizzare le forme presenti nel corpus.

Per l'italiano moderno, esiste il già citato MAGIC, sviluppato dall'Istituto di Linguistica Computazionale del CNR di Pisa: il programma prende un testo “tokenizzato” (ovvero ridotto in “token”: un “token” può essere definito come un'unità minima del testo, quindi un'occorrenza ma anche un segno di punteggiatura) e restituisce un'associazione tra l'occorrenza e tutti i suoi possibili lemmi e categorie grammaticali. Le due tabelle riportate poco sopra sono proprio un esempio di un testo analizzato attraverso MAGIC.

Un altro analizzatore morfologico recentemente sviluppato da Emanuele Pianta e Roberto Zanolini del Centro per la Ricerca Scientifica e Tecnologica della Fondazione Bruno Kessler di Trento è *TagPro*, vincitore dell'edizione 2007 di Evalita⁹⁴ (una competizione riservata a

92 www.ovi.cnr.it

93 <http://www.ovi.cnr.it/index.php?page=informazioni-generalis>

94 <http://evalita.fbk.eu>

2. La fruizione telematica dei carteggi

programmi per il trattamento automatico della lingua⁹⁵: nel caso di Evalita, l'italiano) nella sezione *PoS-Tagging*⁹⁶. *TagPro* è in grado di eseguire analisi molto precise, dal momento che ha dimostrato percentuali di accuratezza tra il 94% e il 98%.

Questi sono soltanto alcuni degli analizzatori morfologici più importanti: quindi il problema della lingua si risolve utilizzando il programma più adatto per le proprie esigenze, cercando di ottenere i risultati più precisi possibili.

2.6 Problemi: la standardizzazione

Tutti i diversi progetti realizzati nell'ambito dell'edizione telematica di carteggi d'artista hanno dimostrato di aver seguito metodi e schemi di codifica diversi. Inoltre, ognuno ha seguito linee guida di usabilità e accessibilità diverse, spesso poco chiare e del tutto discutibili.

Uno dei problemi che si hanno nel realizzare progetti come corpora digitali di carteggi d'artista è riuscire a creare dei prodotti che oltre a essere di qualità, possano anche essere rispettosi degli standard. Di fatto ogni progetto adotta la strategia che ritiene migliore: starà poi ai singoli autori spiegare quali sono state le scelte effettuate in fase di progettazione.

Rispettare uno standard significa anche garantire “portabilità” al sito: con “portabilità” in informatica si intende quella particolare caratteristica che “permette di scrivere un unico programma e di eseguirlo su piattaforme anche diverse, senza dover apportare modifiche al sorgente, purché le piattaforme possiedano il corretto interprete”⁹⁷. Questo è particolarmente vero se si pensa all'implementazione del codice dei siti-web: qualora non vengano rispettati gli standard, si rischia di creare siti difficilmente visualizzabili nello stesso modo su piattaforme diverse (nel caso dei siti, i browser) e quindi si rischia di compromettere tutto il lavoro eseguito in precedenza.

Risulta quindi evidente l'importanza di un lavoro eseguito a regola d'arte anche per quanto concerne il codice: nella sezione all'interno della quale si discuterà lo stato dell'arte, si noterà come alcuni dei siti che verranno presi in esame siano stati realizzati fuori dal rispetto degli

95 Spesso ci si riferisce al trattamento automatico del linguaggio usando la corrispettiva espressione in inglese, “Natural Language Processing” (NLP).

96 “Part of Speech Tagging”: è un'espressione con la quale si indica l'analisi morfologica di un testo, al quale vengono associate delle etichette (“tag”) recanti informazioni sulle categorie morfologiche. Cfr. *infra* PAR. 3.4.

97 Marco Padula e Amanda Reggiori, *Fondamenti di informatica per la progettazione multimediale. Dai linguaggi formali all'inclusione digitale* (Milano: Franco Angeli, 2006), 78.

2. La fruizione telematica dei carteggi

standard.

Inoltre, creare un sito rispettoso degli standard significa anche creare un progetto che possa venire incontro a diverse tipologie di utenti, tra i quali ci sono anche gli utenti diversamente abili: gli standard infatti sono stati pensati in modo tale da soddisfare le esigenze di tutti gli utenti. Ma anche di questo si parlerà più diffusamente nel prosieguo della trattazione.

3. I corpora testuali

3.1 Che cos'è un corpus?

Poiché la proposta metodologica qui discussa è pensata anche per corpora digitali e poiché lo stesso *Carlo Finelli Corpus* è un corpus testuale (tanto che il termine “corpus” è presente nel nome del progetto), si è ritenuto necessario inserire nella presente trattazione una sezione dedicata ai corpora, per capire che cosa è un corpus e per comprendere qual è la grande utilità dei corpora nell'ambito degli studi linguistici.

Un “corpus” (plurale “corpora”, letteralmente “corpo”) in linguistica può essere definito come una “collezione di materiale testuale, raccolto secondo precisi criteri” (“a special collection of textual material collected according to a certain set of criteria”)⁹⁸, “memorizzata su un computer” (“usually stored on a computer”)⁹⁹: questi testi possono “caratterizzarsi per la lingua in cui sono codificati, per il periodo in cui sono stati prodotti o anche per l'ambito settoriale a cui fanno riferimento”¹⁰⁰. Ciò significa che un corpus potrebbe raccogliere opere letterarie del Cinquecento italiano, o ancora traduzioni in inglese, oppure leggi medievali, e via dicendo.

A questo punto è lecito chiedersi quale differenza può esserci tra un corpus e un archivio di testi o un database. La differenza esiste ed è suggerita da Geoffrey Leech, uno dei pionieri della linguistica dei corpora: “la differenza tra un archivio e un corpus consiste nel fatto che il secondo è designato per una particolare funzione 'rappresentativa’” (“the difference between an archive and a corpus must be that the latter is designed for a particular 'representative' function”¹⁰¹). Un archivio quindi è semplicemente una raccolta di testi, normalmente non strutturata e non organizzata secondo precisi criteri.

Un corpus ha quattro caratteristiche principali: la rappresentatività, la dimensione finita, il formato “machine-readable”, l'essere un riferimento standard¹⁰².

La rappresentatività è la capacità di un corpus di essere un valido campione del contesto al quale fa riferimento: questo significa, per esempio, che nel realizzare un corpus

98 Christopher D. Manning, Heinrich Schütze, *Foundations of statistical natural language processing* (Cambridge, Massachusetts: MIT Press, 1999), 119.

99 Anne O'Keeffe, Michael McCarthy e Ronald Carter, *From corpus to classroom: language use and language teaching* (Cambridge: Cambridge University Press, 2007), 2.

100 Bruno Osimo, *Traduzione e nuove tecnologie* (Milano: Hoepli, 2001), 37.

101 Geoffrey Leech, “The state of the art in corpus linguistics” in *English Corpus Linguistics: Studies in Honour of Jan Svartvik*, a cura di Karin Aijmer e Bengt Altenber (Londra: Longman, 1991), 11.

102 Tony McEnery e Andrew Wilson, *Corpus linguistics: an introduction* (Edinburgo: Edinburgh University Press, 2001).

3. I corpora testuali

rappresentativo del linguaggio epistolare del Cinquecento italiano, non sarà sufficiente selezionare soltanto lettere di Pietro Aretino, ma sarà necessario operare una selezione più ampia e quindi più rappresentativa possibile della lingua.

I corpora devono avere anche una dimensione finita, perché deve poter permettere precise analisi quantitative (oltre che qualitative). Nonostante ciò, possono esistere alcune eccezioni: un esempio è il “monitor corpus” *Bank of English* realizzato all'Università di Birmingham, una “collezione di testi” (come preferisce chiamarla John Sinclair, il direttore del progetto) in continuo aggiornamento per cercare nuove parole o per verificare come nel corso del tempo i significati delle parole possano cambiare¹⁰³.

Il termine “corpus” implica, oggi, la caratteristica di essere “machine-readable”: i corpora che possono essere analizzati attraverso il computer hanno notevoli vantaggi nei confronti di corpora stampati, in quanto possono essere manipolati e analizzati con rapidità e facilità¹⁰⁴. Infine, un corpus deve costituire un riferimento standard per la varietà di linguaggio che vuole rappresentare.

Lo studio dei corpora ha fatto sì che si sviluppasse un particolare approccio alla linguistica, al quale ci si riferisce denominandolo “linguistica dei corpora”: quest'ultima può essere definita come una metodologia che può essere utilizzata in quasi tutte le aree della linguistica¹⁰⁵, anche se bisogna prestare attenzione in quanto non delimita un'area della linguistica¹⁰⁶.

Infatti i corpora testuali (e soprattutto i corpora testuali informatici, tanto che negli ultimi anni quando si parla di “corpus” si fa riferimento soprattutto ai corpora informatici) hanno fatto segnare grandi progressi negli studi linguistici, sia a livello descrittivo che a livello teorico.

Gli studi del linguaggio possono essere divisi in due aree principali: gli studi della struttura e gli studi dell'uso di una lingua¹⁰⁷. Per tradizione, la linguistica si è focalizzata più sullo studio della struttura (forme, frasi, categorie grammaticali ecc.) che dell'uso (e cioè analizzare come i parlanti e gli scriventi di una certa lingua utilizzano le strutture che la lingua mette a loro disposizione)¹⁰⁸. Ambedue i tipi di studi sono particolarmente importanti, e un approccio

103 *Ibid.*, 30

104 *Ibid.*, 31

105 *Ibid.*, 2

106 *Ibid.*

107 Douglas Biber, Susan Conrad e Randi Reppen, *Corpus linguistics: investigating language structure and use* (Cambridge: Cambridge University Press, 1998).

108 *Ibid.*

3. I corpora testuali

“corpus-based” alla linguistica può offrire un notevole aiuto. Questo per una serie di fattori¹⁰⁹:

- attraverso i corpora è possibile condurre analisi su collezioni di testi che possono essere largamente rappresentative di una lingua, di un genere, di una varietà;
- i corpora informatici permettono di fare calcoli, ricerche e modelli utilizzando diversi programmi, il tutto con grande facilità e rapidità;
- la codifica può aiutare a identificare le diverse parti dei testi;
- la possibilità di creare modelli statistici e probabilistici attraverso l'utilizzo del calcolatore permette di studiare meglio gli sviluppi di una lingua, di un genere o di una varietà;
- l'affidabilità garantita dall'uso del calcolatore non può essere assicurata in altri modi¹¹⁰;
- attraverso l'approccio “corpus-based” sono state introdotte tecniche di ricerca in precedenza non considerate in quanto ritenute difficilmente realizzabili¹¹¹.

Ancora, un corpus può costituire una risorsa essenziale in lessicografia, in quanto attraverso ricerche “corpus-based” è possibile documentare i diversi significati di una parola e come essi si sono sviluppati nel corso del tempo. Infine, non bisogna dimenticare che i corpora possono anche essere messi in rete e condivisi quindi con un sempre maggior numero di studiosi (ma anche di appassionati).

3.2 Breve storia della linguistica dei corpora

L'uso di studiare la lingua analizzando collezioni di dati derivanti da fonti dirette esisteva già prima della moderna linguistica dei corpora: un esempio è costituito dall'operato di Franz Boas, che nel 1940 inizia a studiare la lingua dei nativi americani analizzando collezioni di dati, benché l'autore non si riferisca ai suoi studi utilizzando il termine “linguistica dei corpora”¹¹². Anche gli studiosi della grammatica degli inizi del XX secolo, come il danese Otto Jespersen, facevano largo uso di testi autentici.¹¹³ Quindi i corpora esistevano già prima

109 Douglas Biber, Ulla Connor e Thomas A. Upton, *Discourse on the move: using corpus analysis to describe discourse structure* (Amsterdam: John Benjamins Publishing Co, 2007).

110 Douglas Biber e Susan Conrad, *Lexical Bundles in Conversation and Academic Prose in Out of corpora: studies in honour of Stig Johansson*, a cura di Hilde Hasselgård e Signe Oksefjell (Amsterdam: Rodopi, 1999), 181.

111 *Ibid.*

112 Tony McEnery e Andrew Wilson, *Corpus linguistics*, 2-3.

113 Jan Svartvik, “*Corpus linguistics 25+ years on*” in *Corpus linguistics 25 years on*, a cura di Roberta Facchinetti (Amsterdam: Rodopi, 2007), 13.

3. I corpora testuali

dell'avvento dei computer, e la stessa cosa si può dire per gli studi condotti attraverso l'analisi di collezioni di testi.

Negli anni Cinquanta lo studio dei corpora conosce una battuta d'arresto a seguito delle critiche mosse da Noam Chomsky nei confronti di questo tipo di approccio alla linguistica: lo studioso contestava il fatto che i corpora non fossero sufficientemente rappresentativi di una lingua. Questo perché secondo il noto approccio generativo di Chomsky, esisterebbe nel linguaggio una creatività che permetterebbe di generare continuamente nuove frasi. Ciò implicherebbe il fatto che l'analisi del linguaggio non debba essere condotta soltanto su insiemi di parole, ma sulle strutture e sulle regole possedute dai parlanti. Secondo Chomsky, la grammatica di un linguaggio sarebbe quindi il meccanismo per il quale tutte le componenti grammaticali di una lingua possono essere generate¹¹⁴. Il linguaggio pertanto sarebbe infinitamente produttivo per il fatto che i parlanti sono abili di produrre un infinito numero di frasi¹¹⁵. Per fare un classico esempio, si possono considerare le stesse frasi che compaiono nel presente paragrafo: chi le legge riconosce in esse strutture grammaticalmente valide, benché è assolutamente probabile che sia la prima volta che le legge.

Un approccio siffatto si pone chiaramente in contrasto con lo studio dei corpora, perché quest'ultimo si fonda su dati empirici, oggettivi e finiti. In più, critiche vengono mosse da altri studiosi, perché prima dell'avvento dei calcolatori le ricerche sui corpora erano lunghe e dispendiose: si trattava di cercare tra milioni di occorrenze senza l'aiuto di macchine! Quindi, a causa del successo delle teorie di Chomsky e delle critiche mosse da altri linguisti, lo studio dei corpora va incontro a un periodo di impopolarità.

Le critiche tuttavia non fermano lo studio dei corpora: quest'ultimo in particolare prende nuovo vigore grazie alla nascita del computer, che ha permesso di eseguire con facilità ricerche prima di allora impensabili.

Il primo uomo a realizzare un corpus elettronico è un italiano, Roberto Busa¹¹⁶: questi, un sacerdote gesuita nato a Vicenza, dopo essersi laureato presso la Pontificia Università Gregoriana, nel 1949 contatta l'allora amministratore delegato della IBM, Thomas J. Watson Sr, e si reca a New York per convincerlo a supportare una sua ricerca¹¹⁷. Tale ricerca consisteva nella realizzazione di un corpus elettronico delle opere di San Tommaso d'Aquino.

114 Matt Carter, *Minds and computers: an Introduction to the Philosophy of Artificial Intelligence* (Edinburgo: Edinburgh Press University, 2007), 149.

115 *Ibid.*

116 Tony McEnery e Andrew Wilson, *Corpus linguistics*, 20-21.

117 *Ibid.*

3. I corpora testuali

Busa, all'epoca insegnante presso l'Aloisianum di Gallarate, inizia a trasferire le opere del filosofo medievale su schede perforate a partire dal 1949, e nel 1967 il progetto finisce con la creazione dell'*Index Thomisticus*, un corpus di 10.600.000 di occorrenze¹¹⁸, in seguito pubblicato anche a stampa e su CD-Rom.

Contemporaneamente viene realizzato il *Brown Corpus*, il primo corpus elettronico di riferimento per la lingua inglese (nonché il primo corpus elettronico progettato per lo studio di una lingua), completato nel 1964 da Henry Kučera e W. Nelson Francis della Brown University di Providence (Stati Uniti)¹¹⁹: la pubblicazione del loro lavoro *Computational Analysis of Present-Day American English* di fatto dà il via alla linguistica dei corpora così come la intendiamo oggi.

Un'altra tappa importante nella storia della linguistica dei corpora è stata la realizzazione, a opera di Sir Randolph Quirk e di Jan Svartvik, del *London-Lund Corpus* (LLC) negli anni Settanta: quest'ultimo è nient'altro che la digitalizzazione del *Survey of English Usage*, un corpus realizzato da Quirk, indicato come il primo corpus di testi moderno¹²⁰.

I lavori di questi primi pionieri ispirano la realizzazione di diversi altri corpora, che vengono sviluppati negli anni Settanta e Ottanta in centri creati appositamente per lo studio sui corpora¹²¹: tra i corpora prodotti in questo periodo bisogna citare il *Lancaster-Oslo-Bergen Corpus* (LOB), un corpus di riferimento per il “British English”, che costituisce la controparte inglese del *Brown Corpus*, in quanto elaborato secondo gli stessi criteri.

Gli anni Novanta vedono lo sviluppo di un altro tipo di corpus, in cui i dati raccolti si sviluppano su una scala diacronica (ovvero i testi raccolti provengono da epoche diverse)¹²²: il primo di questi corpora è l'*Helsinki Corpus of English Texts*, importante soprattutto perché poteva permettere ricerche di tipo storico e consentiva di esaminare i cambiamenti storici dell'inglese¹²³.

Gli ultimi anni hanno visto lo sviluppo di corpora di riferimento per la lingua parlata, contenenti registrazioni audio, nonché di corpora costruiti sulla base di testi prelevati dal web, tanto che secondo alcuni studiosi il web può essere considerato come un grande corpus di testi

118 *Ibid.*

119 *Ibid.*, 22.

120 Victorina González-Díaz, *English adjective comparison: a historical perspective* (Amsterdam: John Benjamins Publishing Co, 2008), 3.

121 Tony McEnery e Andrew Wilson, *Corpus linguistics*, 22-23.

122 Raimond Hickey, *Corpus presenter: software for language analysis* (Amsterdam: John Benjamins Publishing Co, 2003), 2-3.

123 *Ibid.*

3. I corpora testuali

adatti a studi linguistici¹²⁴. Sulla base di quest'ultima osservazione, nel 2001 è iniziato lo sviluppo del progetto WebCorp, un motore di ricerca costruito per permettere ricerche di tipo linguistico su internet. Gli ultimi sviluppi della linguistica dei corpora infatti riguardano proprio la rete: il dibattito si sviluppa intorno alla possibilità di considerare il web come un corpus. La questione presenta vantaggi ma anche svantaggi: tra i vantaggi, il fatto che il web possa essere enormemente più grande di qualunque corpus progettato “a mano”¹²⁵ nonché il fatto che possa essere una sterminata fonte di qualsiasi tipo di testo, mentre tra gli svantaggi si può considerare il fatto che non si conosce l'esatta grandezza del web (e ciò rende relative tutte le stime statistiche condotte su di esso)¹²⁶, nonché la mancanza di controllo sui testi, che possono sparire dalla rete da un momento all'altro¹²⁷.

3.3 Tipi di corpora

Esistono diversi tipi di corpora: la loro classificazione dipende dal materiale che contengono, dai principi applicati alla collezione di dati, dalla loro organizzazione interna, dalla lingua e così via.

In questa sede si proporrà una classificazione, ideata da Lenci, Montemagni e Pirrelli¹²⁸, basata sui seguenti parametri:

- generalità;
- modalità;
- cronologia;
- lingua;
- integrità dei testi;
- codifica digitale dei testi.

Per quanto riguarda il primo parametro, la generalità, è possibile distinguere in corpora

124 Cfr. Barry Morley, “WebCorp: a tool for online linguistic information retrieval and analysis” in *The changing face of corpus linguistics*, a cura di Antoinette Renouf e Andrew Kehoe (Amsterdam: Rodopi, 2006).

125 Tadeusz Piotrowski, “The Translator and Polish-English Corpora” in *Incorporating corpora: the linguist and the translator*, a cura di Gunilla Anderman e Margaret Rogers (Clevedon: Multilingual Matters, 2007), 129.

126 *Ibid.*

127 *Ibid.*

128 Alessandro Lenci, Simonetta Magni, Vito Pirrelli, *Testo e computer. Elementi di linguistica computazionale*, Carocci, 2005, 27 – 34.

3. I corpora testuali

generali (o “trasversali”¹²⁹) e corpora specialistici (o “verticali”¹³⁰): i primi sono corpora che raccolgono testi provenienti da diversi ambiti, mentre i secondi sono corpora i cui testi appartengono a settori ben precisi (per esempio, corpora con testi giuridici, medici, giornalistici, sportivi ecc.) o che vogliono descrivere un particolare linguaggio (come il linguaggio infantile o il linguaggio dei parlanti non madrelingua). Appartengono ai corpora generali, per esempio, il *Brown Corpus* e il *British National Corpus* (<http://www.natcorp.ox.ac.uk>). Un esempio di corpus specialistico invece è CHILDES (*Child Language Data Exchange System*, <http://childes.psy.cmu.edu>), che contiene trascrizioni di dialoghi di bambini.

Se si tiene conto della modalità, è possibile distinguere tra corpora di lingua scritta, corpora di lingua parlata e corpora misti: è facilmente intuibile il fatto che i primi contengano soltanto testi provenienti da fonti scritte, i secondi soltanto testi prodotti oralmente e i terzi ambedue le tipologie di testo. Possono esistere anche corpora con registrazioni audio, che si differenziano dai corpora di lingua parlata per il fatto che questi ultimi contengono trascrizioni di fonti orali. Tra i corpora di lingua scritta si possono annoverare il già citato *Brown Corpus* mentre, tra i corpora di lingua parlata, esempi sono lo stesso CHILD e l'italiano LIP (*Lessico di frequenza dell'Italiano Parlato*, http://languageserver.uni-graz.at/badip/badip/20_corpusLip.php), e infine, tra i corpora misti, si può citare il *British National Corpus*.

Passando al terzo parametro, la cronologia, si è fatto accenno nel paragrafo precedente al fatto che a partire dagli anni Novanta ai corpora *sincronici* (ovvero quelli in cui sono presenti testi che appartengono a una precisa epoca di una lingua) si sono affiancati i corpora *diacronici*, cioè quelli in cui i testi provengono da diverse varietà temporali di una lingua. Un ipotetico corpus diacronico dell'italiano potrebbe contenere contemporaneamente testi del Trecento, del Cinquecento e del Novecento, mentre un corpus sincronico dell'italiano conterrebbe soltanto testi o del Trecento, o del Cinquecento o del Novecento (e forse anche di archi temporali più ridotti). È sincronico *Brown Corpus* in quanto contiene testi pubblicati esclusivamente nel 1961, mentre è diacronico il corpus di riferimento del TLIO (*Tesoro della Lingua Italiana delle Origini*, il vocabolario dell'italiano antico sviluppato dall'OVI: www.ovi.cnr.it), in quanto contiene testi dell'italiano antico dalle origini fino a Boccaccio.

Se si classificano i corpora in base alla lingua, si avranno invece corpora monolingue oppure corpora multilingue. All'interno di questi ultimi è possibile fare un'ulteriore

129 *Ibid.*, 27.

130 *Ibid.*

3. I corpora testuali

distinzione, tra corpora *comparabili* e corpora *paralleli*: i primi contengono testi originali in lingue diverse, mentre i secondi contengono gli stessi testi tradotti in più lingue. Corpora monolingue sono tutti quelli citati in precedenza, mentre tra i corpora multilingue è possibile il BoLC (*Bononia Legal Corpus*, http://corpora.dslo.unibo.it/bolc_ita.html), che contiene testi giuridici in italiano e in inglese.

Per quanto riguarda l'integrità dei testi è invece possibile distinguere tra corpora che contengono testi interi o corpora che contengono frammenti di testi. Appartiene al primo tipo, per esempio, il corpus di riferimento del TLIO, mentre sono corpora costituiti di soli frammenti il *Brown Corpus* e il suo omologo britannico *Lancaster-Oslo-Bergen*, ambedue contenenti cinquecento testi di duemila occorrenze ciascuno, per un totale di un milione di occorrenze.

Infine, è possibile classificare i corpora in base alla codifica digitale dei testi in essi raccolti: ci saranno quindi corpora *ad alto livello*¹³¹, arricchiti con etichette che servono per marcare il testo dando informazioni di tipo strutturale o grammaticale (corpora *annotati*), ed esisteranno invece corpora a basso livello in cui tali informazioni non saranno presenti.

A questa classificazione è possibile aggiungere un ulteriore parametro, e cioè la finitezza della dimensione del corpus: come si accennava sopra¹³², ci saranno corpora a dimensione finita e corpora in continua espansione (detti anche “monitor corpora”).

Infine, è necessario evidenziare che i corpora possono variare molto in base alle loro dimensioni. Esistono quindi corpora con milioni di occorrenze, come il *British National Corpus* o il CORIS (*Corpus di Riferimento per l'Italiano Scritto*, http://corpora.dslo.unibo.it/coris_ita.html), che contengono oltre cento milioni di occorrenze, ma esistono anche corpus di dimensioni più ridotte, come lo stesso *Carlo Finelli Corpus* (che è composto da 5.364 occorrenze) oppure il corpus *Penelope* (disponibile all'indirizzo http://www.parlaritaliano.it/parlare/visualizza_d_s.php?id=5), una raccolta di circa 30.000 occorrenze composta “da campionature molto strette – piccoli prelievi di pochi periodi per testo o per autore – che cercano di coprire il numero più alto possibile di varietà d'uso della lingua”¹³³. Spesso ci si riferisce ai corpus di dimensioni più ridotte utilizzando il termine *mini-*

131 *Ibid.*, 32.

132 Cfr. *supra* PAR. 3.1.

133 Maggi Rombi e Gianna Policarpi, *Il corpus Penelope*, 2002,

http://www.parlaritaliano.it/parlare/dati_e_strumenti/5/Presentazione_Penelope.pdf. Secondo gli autori del corpus “l'italiano sembra più realisticamente descritto analizzando un ridotto numero di occorrenze prelevate però da un elevato numero di contesti d'uso piuttosto che un maggior numero di occorrenze concentrate su pochi contesti d'uso”.

3. I corpora testuali

corpus, ma si tratta di un concetto molto relativo. In generale, si può comunque dire che le dimensioni di un corpus sono strettamente legate alle finalità del corpus stesso: nel caso di *Carlo Finelli Corpus* si è voluto creare un corpus che raccogliesse l'intera produzione epistolare dello scultore, e il risultato è stato quindi un mini-corpus di 5.364 occorrenze.

3.4 L'annotazione dei corpora

Si è già accennato nel paragrafo precedente all'importanza dell'annotazione per i corpora. Annotare un corpus significa dotarlo di etichette che possano rendere esplicite le informazioni strutturali e grammaticali riferite ai testi. L'annotazione serve per rendere utile il corpus per diversi scopi (estrarre informazioni, condurre ricerche, analizzare i testi in modo approfondito, creare modelli statistici, ecc.).

Nonostante l'utilità dell'annotazione, sono state mosse nei confronti di questo tipo di operazione alcune critiche, che McEnery, Xiao e Tono riassumono in quattro punti, fornendo per ognuno di essi una risposta appropriata¹³⁴:

1. l'annotazione produce corpora disordinati: in particolare, secondo alcuni studiosi, i ricercatori dovrebbero essere in grado di analizzare il testo puro, privo delle etichette prodotte dalle operazioni di annotazione. I tre studiosi di cui sopra rigettano questa critica per il semplice fatto che l'annotazione non rende per nulla difficile la leggibilità del testo, in quanto esistono appositi strumenti che permettono agli utenti di leggere comunque il testo puro (quindi tale critica è diretta più agli strumenti di ricerca di un corpus che all'annotazione di per sé)¹³⁵;

2. un testo può essere annotato secondo diverse interpretazioni: la pluralità di interpretazioni che un testo può presentare rende l'annotazione soggettiva. In questo caso, la critica è inconsistente per il fatto che non annotare il corpus non significa che esso non sia comunque soggetto a interpretazioni diverse¹³⁶;

3. l'annotazione rende i corpora ingombranti e meno prontamente accessibili, aggiornabili ed espandibili. Tale critica è futile per il fatto che esistono appositi strumenti che rendono i corpora pienamente e velocemente accessibili nonostante l'annotazione, mentre per quanto riguarda gli aggiornamenti, si può obiettare sostenendo che sono soltanto pochissimi i

134 Tony McEnery, Richard Xiao e Yukio Tono, *Corpus-based language studies: an advanced resource book* (Londra: Routledge, 2006), 30 – 32.

135 *Ibid.*

136 *Ibid.*

3. I corpora testuali

“monitor corpora” che necessitano di essere aggiornati di continuo, mentre gli altri sono tutti di dimensione finita (per cui non è necessario che vengano espansi o aggiornati in maniera continuativa¹³⁷);

4. l'annotazione può non essere accurata. In questo caso, si obietta rispondendo che una combinazione di analisi automatica e di analisi manuale può ridurre al minimo le possibilità d'errore, benché ci sia da notare che anche il miglior linguista a volte commette errori¹³⁸.

L'annotazione di un corpus può essere condotta a diversi livelli e può assumere diverse forme: per esempio, si può annotare un corpus a livello fonologico, a livello morfologico, a livello lessicale, a livello sintattico, a livello stilistico, oppure combinando assieme i diversi livelli.

Possono esserci quindi diversi tipi di annotazione, e di seguito si riportano i principali, secondo una classificazione proposta da Aston e Burnard¹³⁹:

1. *PoS Tagging (Part-of-Speech)*: significa assegnare a ciascuna forma presente nei testi la propria categoria grammaticale, ovvero “sostantivo”, “aggettivo”, “verbo” ecc. (per esempio, alla forma “tavola” viene assegnata l'etichetta “sostantivo femminile singolare”);

2. *Lemmatizzazione*: significa associare a ciascuna forma il proprio lemma (per esempio, a “tavola” e “tavole” viene assegnato il lemma “tavola”);

3. *Word-sense* (annotazione semantica): si annotano le forme distinguendo i diversi significati della stessa parola (“tavola” può essere annotato come “asse di legno” oppure come “tabella”);

4. *Syntactic-role* (annotazione sintattica): in questo caso viene descritta la funzione sintattica degli elementi del testo (“il falegname ha costruito un tavolo”: in caso di annotazione sintattica della frase semplice, “un tavolo” viene annotato come “complemento oggetto”, e in caso di annotazione sintattica del periodo, la frase potrà essere classificata come “proposizione principale”);

5. Annotazione pragmatica: viene generalmente condotta su corpora di lingua parlata e serve per annotare le funzioni delle diverse parti del discorso (esclamazione, saluto, conferma, domanda, risposta ecc.).

A questi tipi si possono aggiungere inoltre l'annotazione stilistica (usata soprattutto per testi letterari, serve per assegnare alle diverse forme del testo la loro funzione stilistica) e

¹³⁷ *Ibid.*

¹³⁸ *Ibid.*

¹³⁹ Guy Aston, Lou Burnard, *The BNC Handbook* (Edinburgo: Edinburgh Press University, 1998), 25 – 27.

3. I corpora testuali

l'annotazione fonetica.

Per poter annotare un corpus è necessario disporre di un linguaggio e di uno schema di codifica: il linguaggio più diffuso e popolare per annotare i corpora è XML (*eXtensible Markup Language*). Si usa XML per diverse ragioni¹⁴⁰: perché è facile da imparare, da capire e da utilizzare, perché i file annotati con XML possono essere scambiati con facilità, perché con altrettanta facilità è possibile passare da XML ad altri formati. Per esempio, per *Carlo Finelli Corpus* XML è stato fondamentale per il passaggio allo schema di codifica finale basato su MySQL.

3.5 Il linguaggio XML

XML è “uno standard per la gestione dei documenti proposto dal World Wide Web Consortium”¹⁴¹. Si tratta della forma semplificata di un altro linguaggio di marcatura, SGML (*Standard Generalized Markup Language*), utilizzato fin dagli anni Ottanta. Si è pensato di passare da SGML a XML perché il primo è pesante e complicato, soprattutto per lo scambio di documenti via web¹⁴². In questo paragrafo si forniranno alcune nozioni di base su XML, necessarie per comprendere in che modo è avvenuta la codifica dei testi di *Carlo Finelli Corpus*.

L'unità base di un file XML è l'*elemento*, che deve essere marcato con la seguente sintassi:

```
<forma>Parola</forma>
```

dove `<forma>` e `</forma>` sono due *etichette*: la prima di apertura e la seconda di chiusura. Il testo contenuto tra le due etichette, “Parola”, è considerato parte dell'elemento stesso¹⁴³. Gli elementi possono avere *attributi*, che servono per specificare alcuni aspetti dell'elemento stesso e devono necessariamente essere specificati all'interno dell'etichetta di apertura. Il *valore* degli attributi deve essere racchiuso tra virgolette, come nel seguente caso:

```
<forma categoria="sostantivo">Parola</forma>
```

140 Cfr. Graham Wilcock, *Introduction to Linguistic Annotation and Text Analytics* (Londra: Morgan & Claypool, 2009).

141 Robert Eckstein, *XML* (Sebastopol, California: O'Reilly 1999), trad. it. *XML* (Milano: Hops, 2000), traduzione di Eugenia Franzoni, 1.

142 *Ibid.*

143 *Ibid.*, 3.

3. I corpora testuali

Gli elementi possono essere *nidificati*, ovvero un elemento può racchiuderne altri:

```
<testo><forma>Parola</forma></testo>
```

L'ordine di apertura e di chiusura delle etichette deve rispettare l'ordine degli elementi: quindi se `<testo>` precede `<forma>`, quest'ultima non può essere chiusa dopo il primo, altrimenti si incorre in un errore.

Oltre agli elementi, in un file XML possono essere inseriti dei commenti, che devono essere collocati al di fuori degli elementi. Al loro interno si può scrivere qualsiasi cosa, e la sintassi da adottare per inserire un commento è la seguente:

```
<!-- Questo è un commento. -->
```

Ogni documento XML deve iniziare con una dichiarazione XML, che può essere scritta nel modo seguente:

```
<?xml version="1.0" encoding="iso-8859-1"?>
```

L'attributo “version” indica la versione XML che si sta adottando, mentre l'attributo “encoding” serve per specificare il set di caratteri da utilizzare. In seguito alla dichiarazione XML è possibile specificare una DTD (*Document Type Definition*), che è un file che indica le regole per la definizione e la correlazione degli elementi, nonché i tipi di attributi. La DTD si dichiara nel modo seguente:

```
<!DOCTYPE esempio SYSTEM "esempio.dtd">
```

dove “esempio” è l'*elemento di root* (“radice”) del file ed “esempio.dtd” è il file in cui è contenuta la DTD.

Queste nozioni sono già sufficienti per creare un documento XML completo e correttamente formattato: un esempio è il file contenuto nella tabella seguente. Si tratta di una semplice annotazione della frase “il gatto rincorre il topo”:

<pre><?xml version="1.0" encoding="iso-8859-1"?> <!DOCTYPE testo SYSTEM "esempio.dtd"></pre>
--

3. I corpora testuali

```
<testo>
<soggetto>
<forma categoria="articolo">Il</forma>
<forma categoria="sostantivo">gatto</forma>
</soggetto>
<predicato>
<forma categoria="verbo">rincorre</forma>
</predicato>
<complemento>
<forma categoria="articolo">il</forma>
<forma categoria="sostantivo">topo</forma>
</complemento>
</testo>
```

TAB. 3.1: un file XML correttamente formattato

Come si può ben vedere, esiste un elemento base (“testo”), che contiene tre elementi (“soggetto”, “predicato” e “complemento”) i quali a loro volta possono contenere più occorrenze dell'elemento “forma”. Questo file, oltre a essere un esempio di un file XML correttamente formattato, è anche un esempio di PoS Tagging (a ogni forma viene associata la categoria grammaticale) e di analisi sintattica, in questo caso analisi sintattica della frase semplice.

Un discorso più dettagliato merita la DTD. Come si è accennato, la DTD è il file che indica in che modo gli elementi devono essere definiti e correlati, nonché quali possono essere gli attributi che gli elementi possono avere: la DTD quindi non è altro che il “regolamento” del file XML. Al suo interno, gli elementi devono essere dichiarati con un nome e una regola, in questo modo:

```
<!ELEMENT nome (regola)>
```

La regola dichiarata nell'elemento può indicare diverse situazioni. Le più semplici sono le regole ANY e PCDATA: la prima indica che un elemento può contenere qualsiasi cosa (sia altri elementi che testo), mentre la seconda indica che un elemento può contenere soltanto testo (PCDATA sta per *Parsed Character Data*). Nelle dichiarazioni, PCDATA deve essere sempre preceduto da un cancelletto:

```
<!ELEMENT forma (#PCDATA)>
```

3. I corpora testuali

Un elemento dichiarato con regola PCDATA non può contenerne altri. Se si vuole scrivere una regola per la quale un elemento ne possa contenere un altro, la forma da adottare sarà la seguente:

```
<!ELEMENT sostantivo (forma)>
```

Ciò significa che l'elemento “sostantivo” deve contenere un elemento di tipo “forma”. È possibile anche fare in modo che un elemento possa contenerne più di uno: in tal caso, gli elementi contenuti dovranno essere separati da una virgola, e nel file XML dovranno rispettare l'ordine dichiarato all'interno delle due parentesi:

```
<!ELEMENT libro (titolo, autori)>
```

In questo caso, significa che l'elemento “libro” deve contenere un elemento “titolo” immediatamente seguito da un elemento “autori”. Si può anche dichiarare che i due elementi “titolo” e “autori” siano mutuamente esclusivi (ovvero, se compare uno non può comparire l'altro), utilizzando la barra verticale al posto della virgola:

```
<!ELEMENT libro (titolo | autori)>
```

Esistono inoltre tre *operatori di ricorrenza* che possono essere associati a un elemento: il primo è il punto interrogativo, il secondo è il più e il terzo è l'asterisco. Il punto interrogativo indica che un elemento può comparire 0 o 1 volte, il più indica che può comparire 1 o più volte, l'asterisco indica che può comparire un qualsiasi numero di volte. Ecco un esempio:

```
<!ELEMENT persona (nome, cognome, indirizzo+, telefono*, email?)>
```

In questo caso, gli elementi “nome” e “cognome” possono comparire un'unica volta, l'elemento “indirizzo” può comparire una o più volte, “telefono” un qualsiasi numero di volte, “email” può non comparire o può comparire una sola volta.

In una DTD possono essere dichiarate anche le *entità*, ovvero oggetti che servono per sostituire stringhe di carattere. Le entità sono necessarie per il fatto che all'interno del testo possono essere inserite sequenze che potrebbero essere male interpretate da XML. Un

3. I corpora testuali

esempio è la parentesi uncinata (“<”): se nel testo ce ne fosse una, XML potrebbe pensare che si tratti dell'inizio di un'etichetta. Per evitare ciò si ricorre alle entità. In XML esistono cinque entità predefinite, e servono per sostituire i seguenti caratteri: &, <, >, “, '.

Un'entità si dichiara nel seguente modo:

```
<!ENTITY nome "caratteri" >
```

Per fare in modo che nel testo l'entità venga sostituita con l'opportuna sequenza di caratteri, sarà necessario fare riferimento al nome dell'entità preceduto da una “e” commerciale e seguito da un punto e virgola. Supponendo di voler sostituire la stringa “marchioregistrato” con il carattere ®, sarà necessario dichiarare l'entità nel modo seguente:

```
<!ENTITY marchioregistrato "&#xae;">
```

Questo significa che sarà sufficiente digitare la sequenza &marchioregistrato; ogni volta che nel documento XML si vorrà far apparire il simbolo del marchio registrato.

Infine, la DTD prevede che siano dichiarati anche gli attributi, e la sintassi per farlo è la seguente:

```
<!ATTLIST elemento nome_attributo tipo_attributo modificatore>
```

Bisogna per prima cosa specificare l'elemento di riferimento dell'attributo, quindi il nome dell'attributo, il suo tipo e infine il modificatore. I tipi di attributo sono nove, ai quali si può aggiungere un elenco di valori impostati dall'autore della DTD. I principali tipi di attributo sono CDATA (caratteri), ID (identificatore univoco), ENTITY (un'entità dichiarata nella DTD).

I modificatori sono tre: #REQUIRED (l'attributo deve essere specificato), #IMPLIED (l'attributo può essere omesso), #FIXED (l'attributo è costante e non può essere modificato: il valore deve essere indicato subito dopo la dichiarazione).

Un attributo può essere quindi dichiarato nel seguente modo:

```
<!ATTLIST forma categoria CDATA #REQUIRED>
```

3. I corpora testuali

Questo significa che l'elemento “forma” ha un attributo chiamato “categoria”, il quale può contenere soltanto sequenze di caratteri (CDATA) e la cui specificazione è obbligatoria (#REQUIRED).

Quindi, un'ipotetica DTD del file XML di esempio presentato nella tabella 3.1 potrebbe essere la seguente:

```
<!ELEMENT testo (soggetto*, predicato*, complemento*)>
<!ELEMENT soggetto (forma*)>
<!ELEMENT predicato (forma*)>
<!ELEMENT complemento (forma*)>
<!ELEMENT forma (#PCDATA)>

<!ATTLIST forma categoria CDATA #REQUIRED>
```

TAB. 3.2: la DTD del file XML di esempio presentato in TAB. 3.1.

3.6 Due codifiche standard per XML

Adottare degli standard di codifica, quando si annota un corpus con il linguaggio XML, può essere importante per il fatto che uno schema di annotazione condiviso da più autori di corpora può garantire una maggiore intercambiabilità dei testi: “un'opera di codifica che utilizzi un *metalinguaggio di marcatura standard* e uno *schema standard di codifica e annotazione* trasforma il testo digitale in una *risorsa di informazione* pronta all'uso”¹⁴⁴. Testi annotati secondo schemi di codifica standard, creati da consorzi internazionali, possono rendere i corpora più facilmente leggibili e manipolabili da parte di ricercatori interessati a fare ricerche su corpora non creati da loro.

In questo paragrafo si parlerà di due modelli standard per la codifica in XML: il primo è il modello XML-TEI, pensato per tutti i tipi di testi digitali, e il secondo è il progetto DALF, appositamente ideato per la codifica di lettere.

Il consorzio TEI (*Text Encoding Initiative*, <http://www.tei-c.org/index.xml>) è nato nel 1988 con il preciso scopo di creare uno standard di codifica per i testi digitali. Infatti sin dagli anni Ottanta “in campo umanistico si era avvertita l'esigenza di rispondere adeguatamente ai problemi di intercambiabilità e portabilità creati dalla proliferazione dei linguaggi di codifica, e di definire uno standard per la rappresentazione di testi su supporto digitale”¹⁴⁵:

144 Alessandro Lenci, *Testo e computer*, 92.

145 Giuseppe Gigliozzi, *Introduzione all'uso del computer negli studi letterari* (Milano: Mondadori, 2003), 78.

3. I corpora testuali

proprio per questi motivi ha preso il via il progetto TEI, al quale hanno partecipato studiosi provenienti da ogni parte del mondo.

La prima pubblicazione delle specifiche del sistema di codifica risale al 1991, quando il consorzio TEI dà alla luce *Guidelines for Electronic Text Encoding and Interchange* (TEI P1), per il linguaggio SGML. Negli anni seguenti escono diverse revisioni del primo lavoro, fino a giungere alla versione TEI P5 rilasciata nel 2007.

Le specifiche TEI prevedono che un testo sia diviso in due parti: la cosiddetta *TEI header* (la cui etichetta è `<teiHeader>`) e il testo (`<text>`). La *TEI header* fornisce alcune informazioni relative al testo, ovvero titolo, autore, modalità di codifica, descrizione, storia delle revisioni e così via. Il testo può essere diviso in tre parti: una sorta di “premessa” (`<front>`), il corpo del testo (`<body>`) e una appendice (`<back>`). A sua volta il testo può essere diviso in capitoli e paragrafi (oppure strofe se si tratta di una poesia e scene se si tratta di un'opera teatrale, e così via).

Ovviamente, anche la codifica TEI ha una DTD, che è molto modulare e si presta a personalizzazioni e modifiche. Nella tabella seguente è riportato un esempio di un ipotetico file XML che codifica il carteggio Canova-Quatremère de Quincy secondo lo standard TEI:

```
<TEI.2>

  <teiHeader>

    <fileDesc>
      <titleStmt>
        <title>Il carteggio Canova-Quatremère de Quincy</title>
        <author>Antonio Canova</author>
        <author>Antoine Chrysostome Quatremère de Quincy</author>
      </titleStmt>
    </fileDesc>

  </teiHeader>

  <text>
    <front>premessa... premessa... </front>
    <body>testo... testo... </body>
    <back>appendice... appendice... </back>
  </text>

</TEI.2>
```

TAB. 3.3: un'ipotetica codifica XML secondo gli standard TEI

del carteggio Canova-Quatremère de Quincy

3. I corpora testuali

Il progetto DALF (*Digital Archive of Letters in Flanders*, <http://www.kantl.be/ctb/project/dalf/>) è stato realizzato dal *Centrum voor Teksteditie en Bronnenstudie* (“Centro per l'edizione dei testi e gli studi dei documenti”) della *Koninklijke Academie vor Nederlandse Taal- en Letterkunde* (“Accademia Reale di Lingua e Letteratura Olandese”) di Gent (Belgio).

Si tratta di una iniziativa, nata nel 2002, volta a creare uno schema di codifica conforme agli standard TEI, inizialmente pensata per le lettere di scrittori e compositori fiamminghi del XIX e del XX secolo¹⁴⁶, ma comunque utile per autori di ogni luogo ed epoca.

Lo schema di base è quello visto in precedenza, ma il modello DALF propone in più diversi elementi per codificare al meglio una lettera, con apposite etichette per la busta (<envelope>), l'indirizzo (<address>), i saluti (<salute>), la firma (<signed>), il post scriptum (<ps>) e così via.

Bisogna precisare che la metodologia proposta nel presente lavoro non si avvale di nessuno dei due schemi di codifica. Questo per un semplice motivo: il formato di interscambio pensato per i corpora prodotti con la metodologia qui proposta è MySQL e non XML. Quest'ultimo rappresenta infatti soltanto un passaggio intermedio. Si parlerà in seguito dei motivi che hanno spinto a usare MySQL come formato di interscambio piuttosto che XML¹⁴⁷.

146 Dal sito ufficiale del progetto.

147 Cfr. *infra* PAR. 5.5.

4. Carteggi d'artista in rete: lo stato dell'arte

4.1 Introduzione

In questa sezione si discuterà dei più importanti progetti presenti in rete volti a proporre al pubblico carteggi d'artista in formato digitale (benché alcuni di essi contengano anche altri tipi di documenti). Tali progetti non sono comunque molti e sono tutti nati di recente, e nei prossimi paragrafi si discuterà della loro struttura, della loro organizzazione, dei loro vantaggi, dei loro punti deboli. Il tutto in modo tale da rendere anche esplicite le differenze rispetto alla metodologia proposta nella presente trattazione. I progetti che saranno discussi sono i seguenti:

- Il portale FO.S.C.A.;
- *The History of the Accademia di San Luca: Documents from the Archivio di Stato di Roma*;
- Il portale della Fondazione Memofonte;
- Il *Corpus Informatico Belloriano*;
- *Medici Archive Project*;
- *Vincent Van Gogh – The Letters*;
- *Google Books*.

4.2 Il portale FO.S.C.A.

Il portale FO.S.C.A. (*Fonti per la Storia della Critica d'Arte*, <http://www.fosca.unige.it>) è stato realizzato nell'ambito di un PRIN (*Progetto di Ricerca di Interesse Nazionale*) che ha coinvolto le Università di Genova, Perugia, Pisa, Urbino e Napoli. Il portale “aspira a diventare un luogo di discussione dei testi e dei percorsi artistici che si possono desumere dalle guide, biografie, carteggi, taccuini e letteratura di viaggio dall'età moderna fino al primo Novecento, per sviluppare il dibattito e la crescita della cultura storiografica che ha come argomento le varie realtà artistiche locali”¹⁴⁸.

148 Dalla home page del sito ufficiale.

4. Carteggi d'artista in rete: lo stato dell'arte

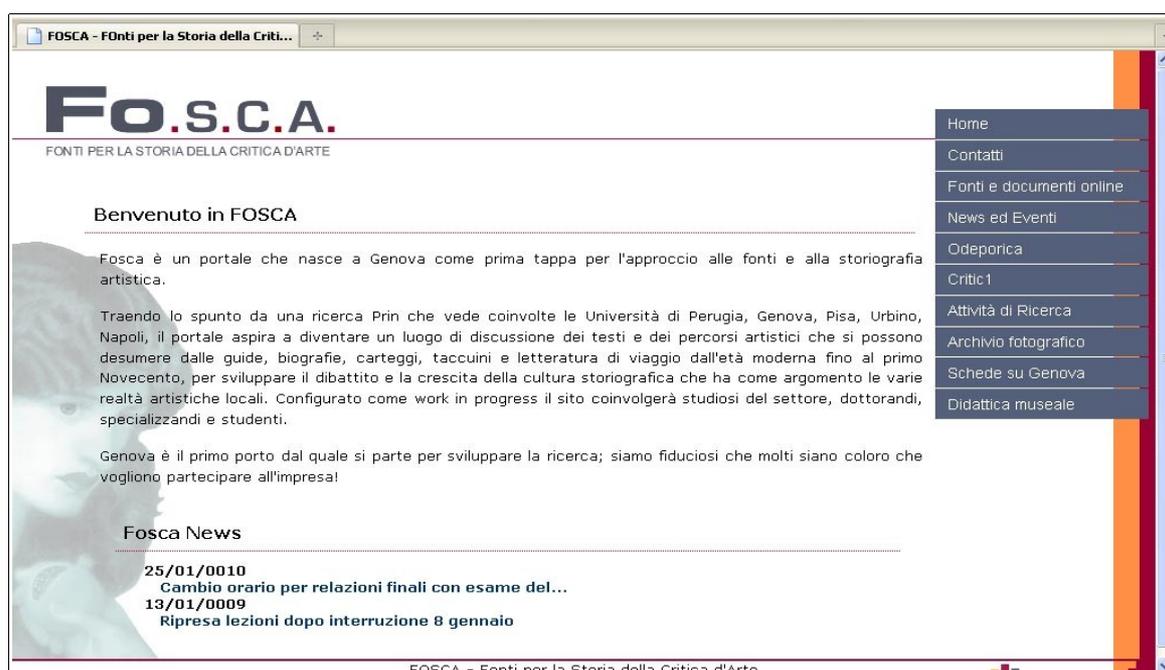


FIG. 4.1: la home page del portale FO.S.C.A.

Al gruppo di lavoro, guidato dalla Professoressa Maurizia Migliorini dell'Università di Genova, partecipano docenti e ricercatori di tutte le università coinvolte.

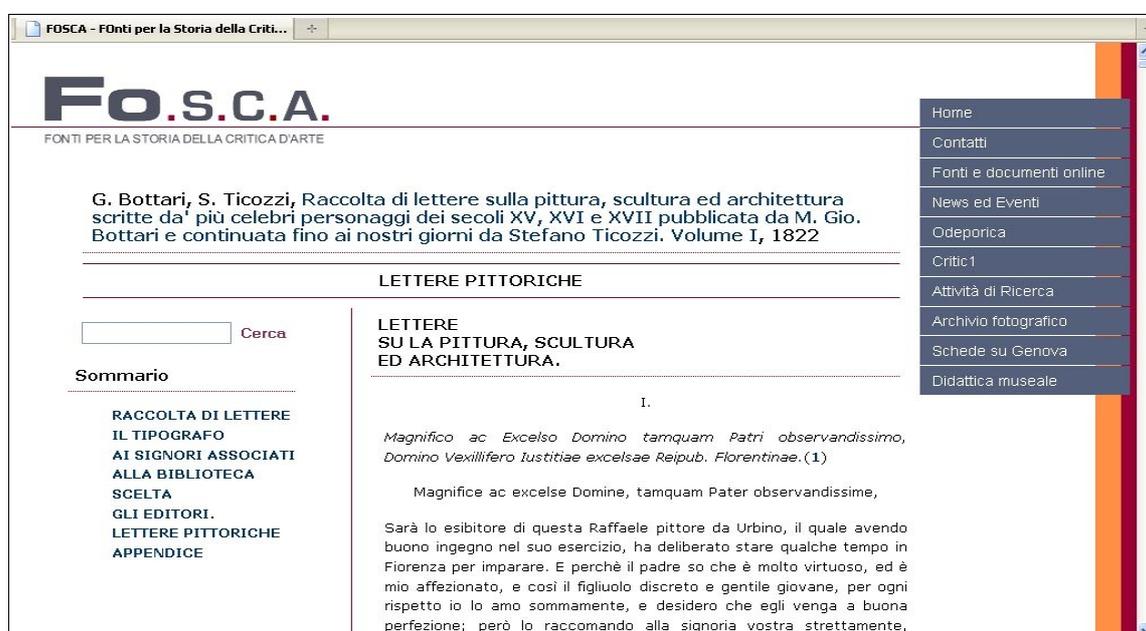


FIG. 4.2: la Raccolta di lettere di Bottari e Ticozzi visualizzata su FO.S.C.A.

Inoltre, il sito sembrerebbe aperto alle collaborazioni in quanto si dichiara che il sito è “configurato come work in progress”, “coinvolgerà studiosi del settore, dottorandi,

4. Carteggi d'artista in rete: lo stato dell'arte

specializzandi e studenti” e infine i responsabili si auspicano “che molti siano coloro che vogliono partecipare all'impresa”¹⁴⁹.

Il sito si dimostra organizzato in più sezioni, e la più interessante ai fini della presente trattazione è sicuramente “Fonti e documenti online”. Per quanto riguarda i carteggi, tra i documenti è presente anche la *Raccolta di lettere sulla pittura, scultura ed architettura* di Bottari e Ticozzi alla quale si accennava sopra¹⁵⁰.

Cliccando sull'apposito collegamento si accede alla pagina all'interno della quale sono conservati i documenti.

Facendo clic sul titolo dell'opera è possibile accedere alla raccolta. Di quest'ultima è presente la trascrizione, divisa seguendo all'incirca l'impostazione dell'opera originale: sono quindi presenti sezioni per il titolo, per la presentazione del tipografo e degli autori, per il testo dell'opera e per l'appendice.

Il sito ha un semplice motore di ricerca per i documenti, a dire il vero piuttosto inutile: cercando un qualsiasi pattern infatti non si viene rimandati al punto nel testo in cui è presente il pattern, ma viene indicato in modo molto generico soltanto il link alla sezione all'interno della quale il pattern è presente. Per esempio, cercando la stringa “Vasari”, si apre una pagina con tre collegamenti: la presentazione degli autori, il testo dell'opera e l'appendice. Una scelta piuttosto infelice se si tiene conto del fatto che la stringa “Vasari” all'interno del solo testo dell'opera compare trentanove volte e che il testo stesso è composto da quasi ottantamila occorrenze!

Ancora, nel testo sono presenti note con collegamenti: facendo clic su di essi si viene rimandati alla nota corrispondente. Non per tutti i documenti però il sistema funziona, e in alcuni casi i collegamenti sembrano interrotti. Il problema consiste nel fatto che nel sito è presente un errore nel codice PHP che fa sì che non si possa accedere direttamente alle note qualora le si voglia leggere, ma è necessario scorrere a mano tutto il testo fino alla fine per trovare la nota di proprio interesse (sono invece funzionanti in altri testi presenti in FO.S.C.A.).

Non sono presenti le immagini dei manoscritti originali e non sono presenti neppure indicazioni sul modo in cui sono stati codificati i testi, ma comunque il sito è stato realizzato in modo conforme agli standard del W3C (*World Wide Web Consortium*, www.w3c.org: è il consorzio, fondato da Tim Berners Lee nel 1994, che definisce gli standard per il web e si

¹⁴⁹ *Ibid.*

¹⁵⁰ Cfr. *supra* PAR. 1.6.

4. Carteggi d'artista in rete: lo stato dell'arte

occupa di migliorare il potenziale della rete): ciò dimostra che i realizzatori del portale hanno avuto particolare cura nel progettare la struttura HTML. Si tratta però, in definitiva, di un progetto che ha dei margini di miglioramento molto ampi, ed è necessario ripensare di sicuro il sistema di ricerca in modo da renderlo utile.

4.3 *The History of the Accademia di San Luca*

Il portale *The History of the Accademia di San Luca: Documents from the Archivio di Stato di Roma* (<http://www.nga.gov/casva/accademia>) è stato realizzato dal “Center for Advanced Study in the Visual Arts” della National Gallery di Washington in collaborazione con l'Archivio di Stato di Roma e l'Accademia Nazionale di San Luca. Il progetto “presenta organicamente documenti di tipo notarile sinora in gran parte inediti” dell'Archivio di Stato di Roma, “molti dei quali erano stati ritenuti perduti”, “riguardanti la storia istituzionale dell'accademia”¹⁵¹.

Come per il portale FO.S.C.A. quindi è possibile trovare all'interno del sito non soltanto carteggi, ma qualsiasi tipo di documento. La banca dati del sito consente di accedere alle trascrizioni dei documenti succitati (appartenenti a un periodo compreso tra il 1590 e il 1635) e di poter visualizzare anche l'immagine del documento originale.

Come il sito precedente è un archivio e non un corpus, in quanto realizzato senza intenti di rappresentatività. Il sito, anch'esso come FO.S.C.A. organizzato in modo piuttosto intuitivo, si compone di due sezioni: una denominata “ricerca” e una “immagini”. La prima sezione consente di svolgere ricerche all'interno dell'archivio: è presente un tipo di ricerca “semplice”, attraverso il quale è possibile ricercare un qualsiasi pattern all'interno del sito, ed è presente una ricerca detta “guidata”, che consente all'utente di svolgere ricerche in base al nome citato, ai luoghi, alle parole chiave, al tipo di documento, al nome del notaio, all'anno del documento.

In questo caso i motori di ricerca sono utili in quanto restituiscono i collegamenti ai file all'interno dei quali è presente la stringa ricercata (nel caso della ricerca semplice) e un elenco di documenti legati ai criteri di ricerca selezionati, nel caso della ricerca guidata. Quest'ultima consente di svolgere anche ricerche combinate.

Le pagine dei documenti contengono un registro, la trascrizione del documento e le immagini degli originali. Inoltre dalla pagina del documento è anche possibile scaricare il

151 http://www.nga.gov/casva/accademia/intro_ita.shtml

4. Carteggi d'artista in rete: lo stato dell'arte

relativo file XML codificato secondo lo schema TEI.

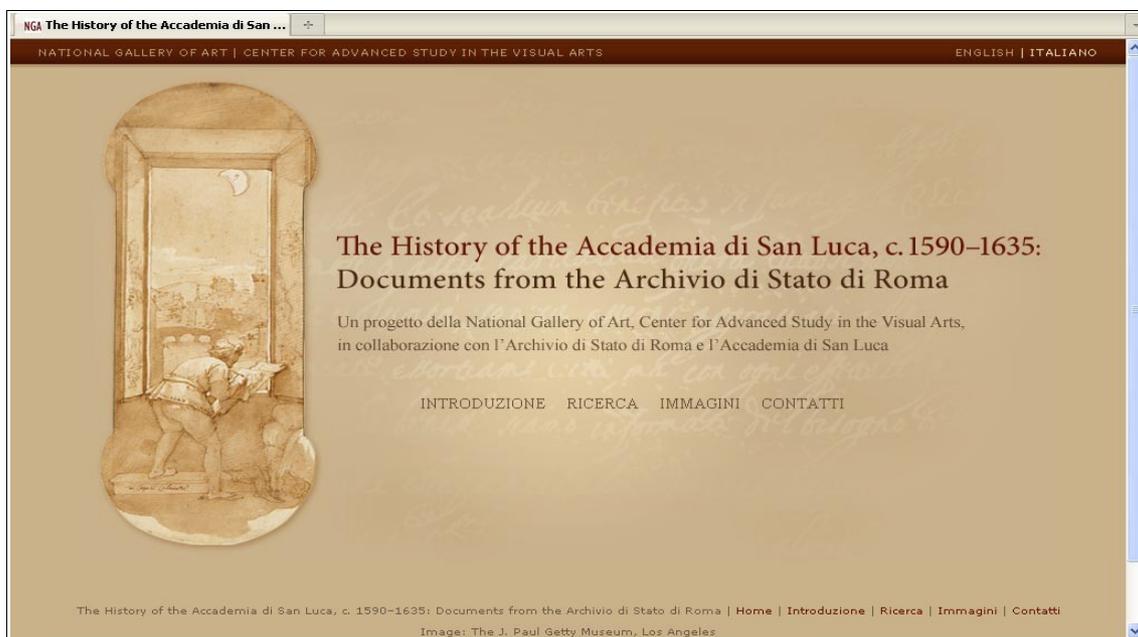


FIG. 4.3: home page del progetto *Documents from the Archivio di Stato di Roma*

La sezione “immagini” invece presenta immagini di dipinti, disegni, incisioni, sculture realizzate da artisti legati in qualche modo all'Accademia di San Luca.

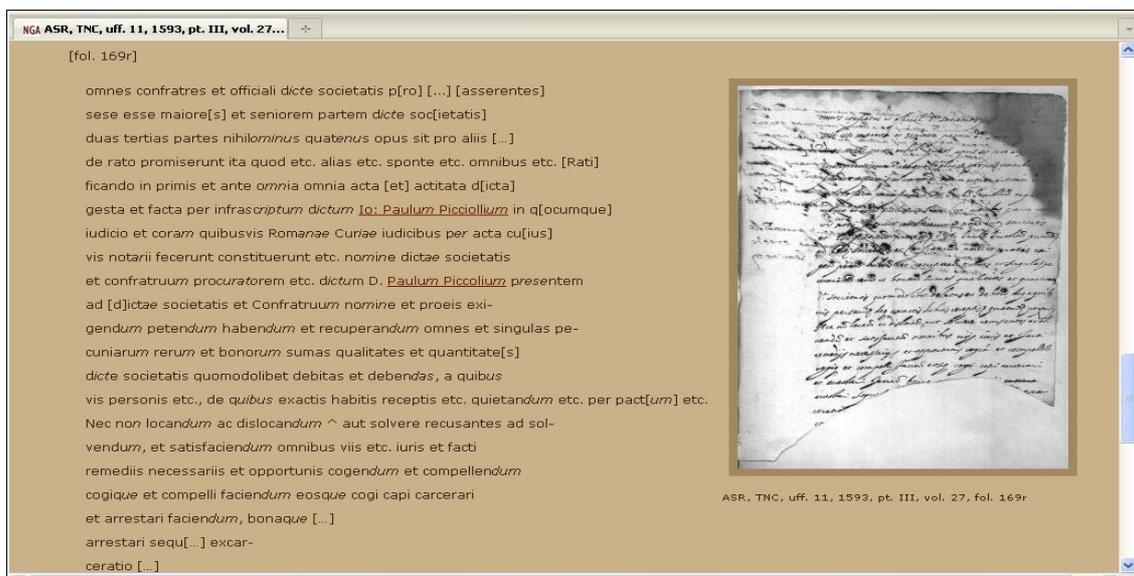


FIG. 4.4: la pagina di un documento in *Documents from the Archivio di Stato di Roma*

Il sito si dimostra senza dubbio ben concepito e ben strutturato, con una grafica semplice ma molto chiara e intuitiva, benché ci siano comunque alcuni punti deboli.

4. Carteggi d'artista in rete: lo stato dell'arte

Questi ultimi riguardano soprattutto l'aderenza agli standard del W3C, dal momento che il portale non è conforme: questo potrebbe dimostrare poca cura in fatto di usabilità e accessibilità. Per esempio, le immagini mancano dell'attributo "alt", ovvero una descrizione testuale alternativa alle immagini, necessaria per i programmi di lettura per non vedenti: è un errore che gioca a sfavore degli utenti diversamente abili, ed è facilmente risolvibile. Si tratta tuttavia di un portale di indubbia qualità e che fornisce strumenti molto utili, ma che potrebbe però essere migliorato soprattutto per quel che riguarda la conformità agli standard.

4.4 Il portale della Fondazione Memofonte

La Fondazione Memofonte è stata fondata come associazione nel 2000 a Firenze da Paola Barocchi. Il suo scopo è quello di pubblicare in rete fonti e documenti di non facile consultazione relativi alla storia dell'arte, in un periodo compreso tra il XV e il XX secolo. Divenuta fondazione nel 2006, Memofonte offre, all'interno del proprio sito ufficiale (www.memofonte.it) un buon numero di documenti.

Il sito, nonostante la ricchezza del materiale contenuto, non è strutturato in maniera molto organica. I contenuti appaiono nella home page, in una colonna collocata a destra nella pagina e sono ordinati in due modi: in base agli autori e in base al tipo di documento o alla collezione di riferimento.

Il sito della Fondazione presenta inoltre un motore di ricerca che però non conduce la ricerca all'interno dei soli documenti, bensì in tutto il sito: questa caratteristica rende il lavoro di ricerca piuttosto pesante.

I documenti sono tutti disponibili in formato PDF (questa caratteristica è particolarmente utile per poter eseguire il download e la stampa dei documenti in modo facile e rapido), eccezion fatta per i carteggi, che sono stati inseriti in database all'interno dei quali è possibile condurre ricerche. Queste ultime possono essere condotte in base a diversi criteri: si possono filtrare le lettere in base alla data o a stringhe contenute nel testo, nell'intestazione e nella segnatura della missiva.

Il motore di ricerca interno alle lettere, molto ben organizzato, è particolarmente interessante perché come quelli dei siti analizzati in precedenza restituisce i collegamenti ai documenti che contengono la stringa cercata, ma a differenza di essi evidenzia tutte le occorrenze della stringa cercata nel documento: si tratta di una caratteristica molto importante

4. Carteggi d'artista in rete: lo stato dell'arte

che rende il motore di ricerca di fatto più utile, perché l'utente non è costretto a cercare di individuare all'interno del testo della missiva la stringa da lui cercata.



FIG. 4.5: home page del sito della Fondazione Memofonte

Mancano gli originali ma in compenso la grafica risulta curata ed elegante e il sito può essere navigato con relativa facilità.

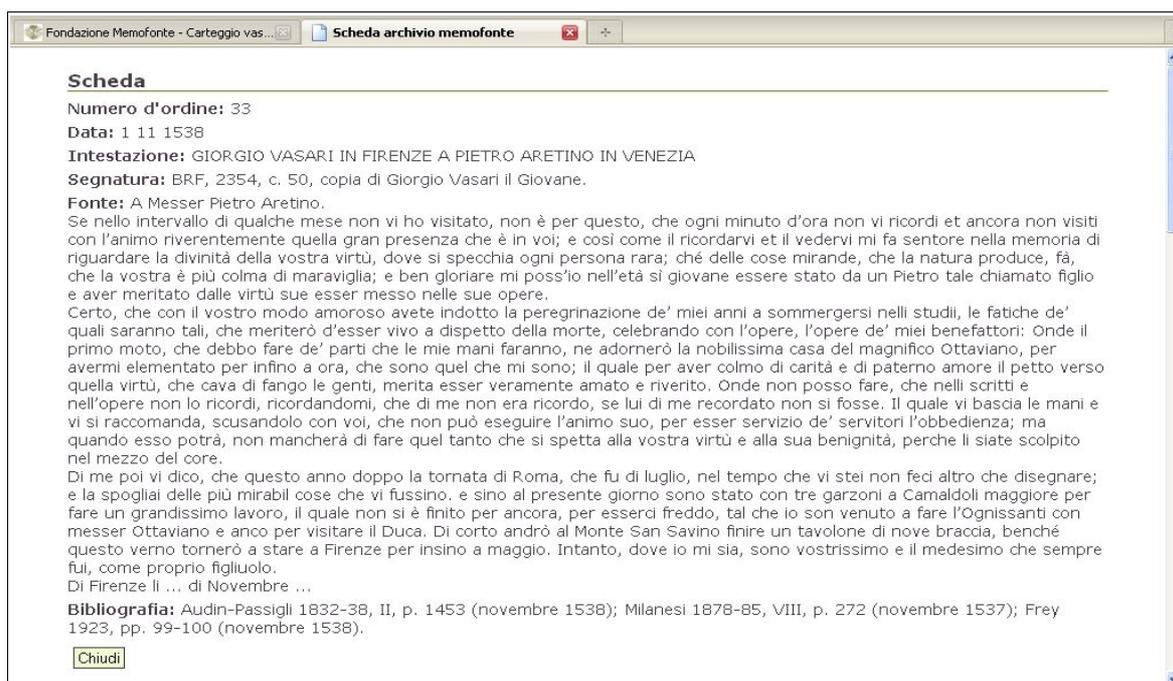


FIG. 4.6: una lettera da Giorgio Vasari a Pietro Aretino nel sito della Fondazione Memofonte

4. Carteggi d'artista in rete: lo stato dell'arte

Anche in questo caso però il sito ha dimostrato di non essere conforme agli standard W3C: il validatore (<http://w3.validator.org>), in data 13 marzo 2010, ha segnalato addirittura trentacinque errori nella home page. Decisamente troppi per un sito così importante.

La gravità della cosa è inoltre accresciuta dal fatto che nel footer del sito compaiono due icone (una per XHTML e una per CSS) che indicano la conformità del sito agli standard: la Fondazione deve rivedere il sito e ristrutturarlo secondo gli standard, perché dichiarare una conformità che in realtà non esiste è indice di scarso rispetto nei confronti di tutti gli utenti. Questo anche perché l'aderenza agli standard è importante per proporre al pubblico un prodotto che sia usabile, ma che sia anche accessibile¹⁵².

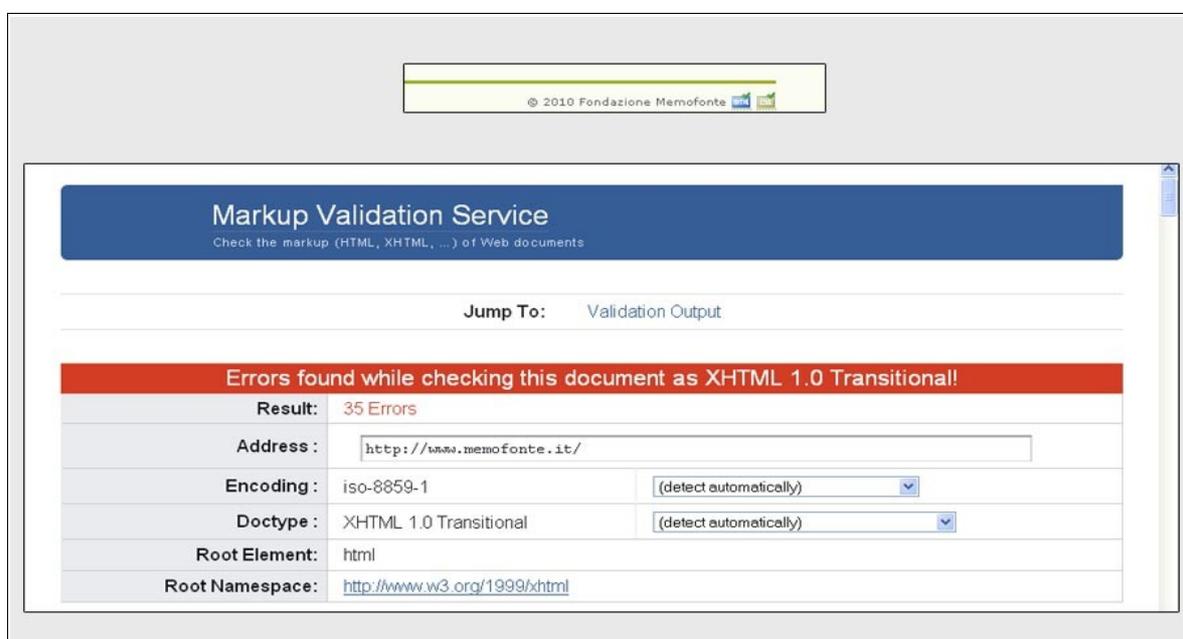


FIG. 4.7: dichiarazione di conformità e report della validazione della home page di Fondazione Memofonte (13 marzo 2010).

4.5 Il *Corpus Informatico Belloriano*

Il *Corpus Informatico Belloriano* (www.biblio.signum.sns.it/bellori), realizzato dal Centro Ricerche Informatiche per i Beni Culturali della Scuola Normale Superiore di Pisa, “prevede la memorizzazione di tutti gli scritti noti (editi ed inediti) di Giovan Pietro Bellori, e del relativo corredo iconografico, valorizzando appieno la possibilità tecnica di una consultazione

152 Cfr. *infra* PAR. 7.3, 7.4.

4. Carteggi d'artista in rete: lo stato dell'arte

simultanea del testo e delle immagini”¹⁵³.

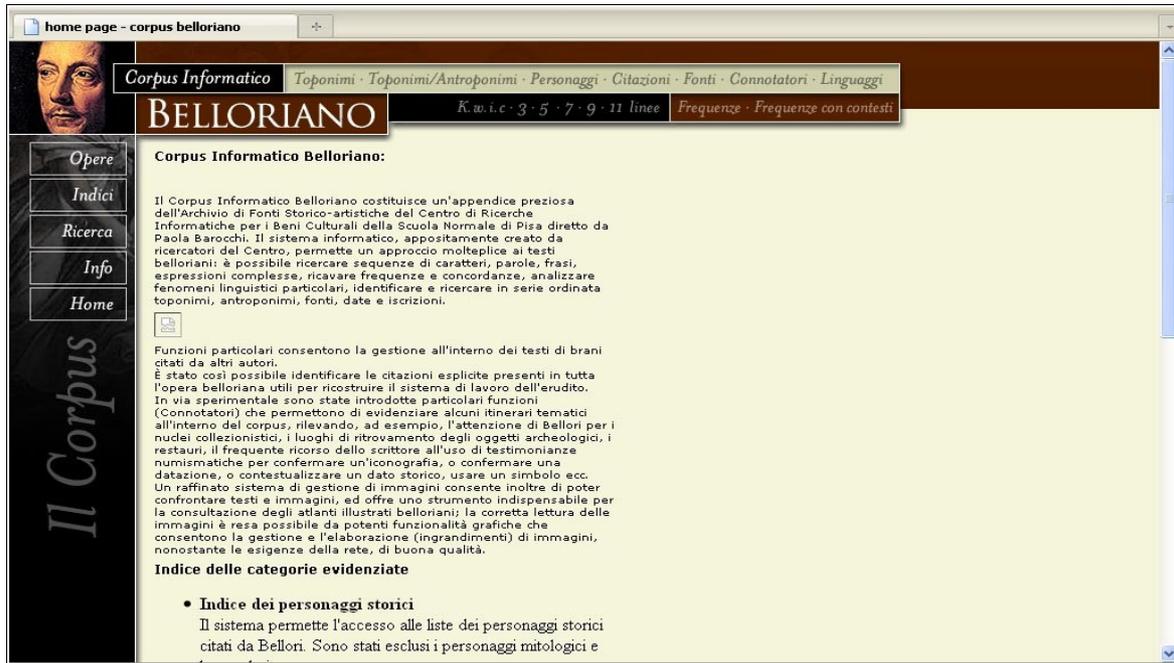


FIG. 4.8: prima pagina del *Corpus Informatico Belloriano*

Il notevole punto di forza del *Corpus Informatico Belloriano* è il suo motore di ricerca che restituisce risultati visualizzati come contesti KWIC (*Key Word In Context*): significa che la forma ricercata viene posta al centro della pagina e prima e dopo di essa vengono visualizzate un tot di occorrenze del contesto.

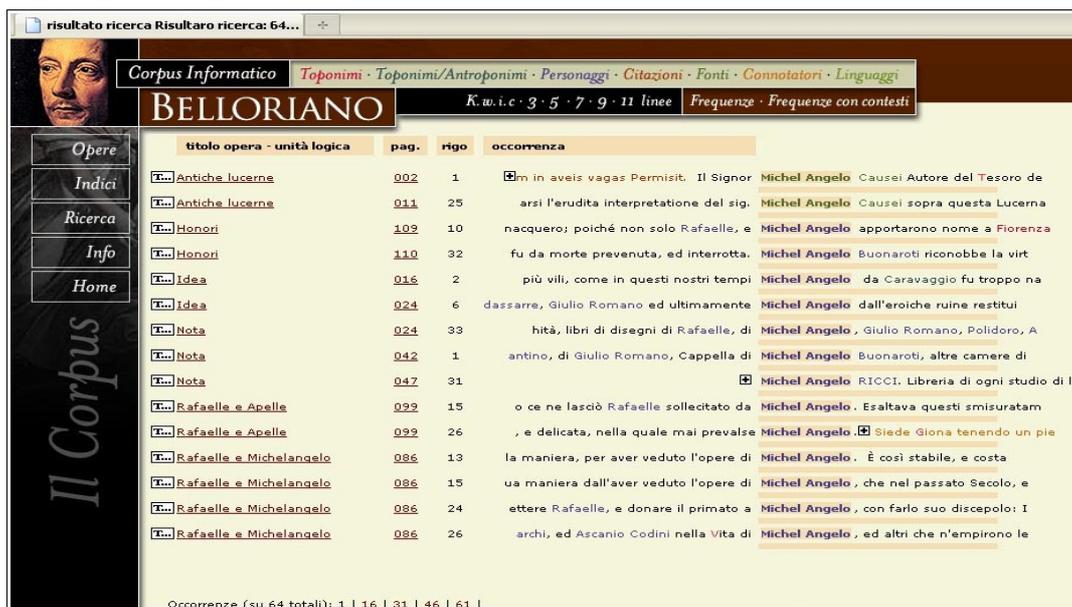


FIG. 4.9: la visualizzazione KWIC del motore di ricerca del *Corpus Informatico Belloriano*

153 Dal sito del progetto.

4. Carteggi d'artista in rete: lo stato dell'arte

Facendo clic sui collegamenti che si presentano vicini alla forma è inoltre possibile accedere alla fonte che contiene l'occorrenza cercata.

Però, come si nota dall'immagine 4.8, la prima pagina si presenta con un collegamento a un'immagine inesistente (l'icona visibile tra il primo e il secondo capoverso del testo): di certo non un buon biglietto da visita per il sito.

Il motore permette inoltre di condurre ricerche all'interno di un'opera ben precisa oppure di eseguire ricerche in base alla lingua, ai nomi di persona, ai toponimi e alle citazioni contenute nei documenti.

Il punto debole del sito è costituito dalla sua mancanza di completezza (alcuni testi non sono completi, in più mancano molti indici), dalla sua scarsa intuitività e dalla difficoltà di navigazione. Il menù in alto (per evidenziare toponimi, citazioni, personaggi ecc.) si attiva soltanto quando si naviga all'interno dei testi: sarebbe stato più intelligente far sì che apparisse soltanto mentre l'utente sfoglia le opere e non in tutte le pagine del sito. L'utente, vedendolo disattivato nella home page, è portato a credere che non sia funzionante e quindi non lo terrebbe in considerazione durante il resto della navigazione. Ancora, le icone che permettono di visualizzare il corredo iconografico delle opere non sono di immediata intuitività, e la mancanza dell'attributo "alt" che possa descriverne la funzione è penalizzante in questo caso non soltanto per i diversamente abili, ma anche per gli utenti che non possono sapere cosa indichi l'icona.

I testi sono stati codificati in linguaggio XML secondo lo schema TEI e le ricerche sono possibili grazie a un motore di ricerca basato su XML e sviluppato dallo stesso Centro Ricerche Informatiche per i Beni Culturali in collaborazione con il Dipartimento di Informatica dell'Università di Pisa. Però, anche in questo caso, il sito non è conforme agli standard del W3C.

4.6 *Medici Archive Project*

Medici Archive Project (www.medici.org) è “una fondazione senza scopo di lucro con sedi presso l'Archivio di Stato di Firenze ed in USA”¹⁵⁴. All'interno del sito sono indicati i diversi obiettivi che la fondazione persegue: tra questi, “permettere l'accesso online dei materiali contenuti nell'archivio Granducaale dei Medici tramite un database facilmente consultabile”¹⁵⁵.

154 Dal sito ufficiale.

155 La missione della fondazione è disponibile in italiano all'indirizzo <http://www.medici.org/italian>.

4. Carteggi d'artista in rete: lo stato dell'arte

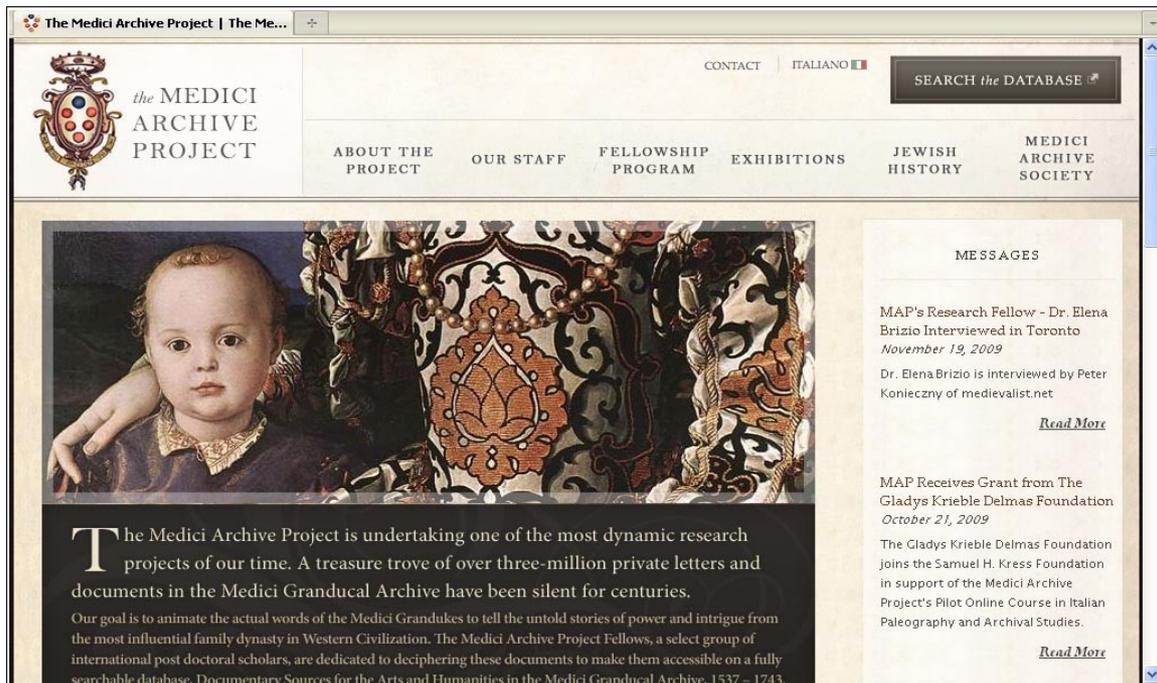


FIG. 4.10: home page di Medici Archive Project

Il progetto, realizzato anche grazie al supporto di altri enti e fondazioni (come la Fondazione Monte dei Paschi di Siena), prevede la pubblicazione online del database delle fonti documentarie per l'arte e le discipline umanistiche contenute nell'*Archivio Mediceo del Principato*.



FIG. 4.11: dati relativi a una lettera inviata da Cosimo I de' Medici al Bronzino in *Medici Archive Project*

4. Carteggi d'artista in rete: lo stato dell'arte

L'*Archivio Mediceo del Principato* contiene circa tre milioni di lettere inviate e ricevute dai membri della famiglia Medici, dai loro segretari, cortigiani e corrispondenti in un periodo che va dal 1537 al 1743. Queste lettere sono incluse in 6.429 volumi e gli originali sono conservati presso l'Archivio di Stato di Firenze. Il database nella primavera del 2009 (periodo dell'ultimo aggiornamento delle FAQ) conteneva circa 360 volumi e 20.000 documenti, e il suo completamento, data l'enorme quantità del materiale, è previsto per il 2030¹⁵⁶. Il database occupa un sottodominio del sito e il suo indirizzo è <http://documents.medici.org>.

Non si tratta di un vero corpus digitale di carteggi, ma è una base di dati nel vero senso del termine: una raccolta di informazioni sui documenti conservati presso l'Archivio di Stato di Firenze. Quindi non sono presenti i testi completi delle lettere, ma soltanto brevi estratti, come illustrato in figura 4.11.

Si tratta sicuramente di una risorsa molto utile per cercare preziose informazioni per le proprie ricerche all'interno dei documenti dell'Archivio dei Medici, anche per il fatto che per molti anni è stato difficile accedere alle informazioni contenute nei documenti: le lettere erano infatti organizzate in base a mittenti e destinatari e catalogati sulla base dell'anno in cui sono state scritte, quindi chi avesse voluto condurre ricerche si sarebbe trovato in difficoltà perché non avrebbe saputo dove guardare¹⁵⁷.

Il sistema offre la possibilità di condurre diversi tipi di ricerca: è possibile cercare documenti in un preciso intervallo di tempo, conoscere le fonti che riguardano un personaggio ben specificato, cercare documenti che riguardano un certo argomento (sono quarantadue gli argomenti inseriti tra le opzioni di ricerca) e molto altro ancora. Mancano però i testi integrali, quindi coloro che vogliono esaminare nello specifico i documenti ritenuti rilevanti per i propri studi, devono affrontare una trasferta a Firenze per consultare dal vivo le fonti. In un'ottica di accesso libero, facile e universale si tratta di una lacuna di grandi proporzioni.

Il sito è comunque ben organizzato, e soprattutto dispone di una grafica molto raffinata, che riprende l'estetica dei documenti della Firenze medicea. Ci sono però diverse lacune a livello di implementazione del codice, tanto che il sito non passa l'esame del validatore W3C e pertanto non è conforme agli standard. Anche in questo caso non sono stati rispettati i requisiti per l'accessibilità: quindi ancora poca cura nei confronti degli utenti diversamente abili.

156 La pagina delle FAQ (*Frequently Asked Questions*) che contiene queste informazioni è disponibile all'indirizzo <http://documents.medici.org/faq.cfm>.

157 <http://www.theflorentine.net/articles/article-view.asp?issuetocId=1471>

4. Carteggi d'artista in rete: lo stato dell'arte

4.7 Vincent Van Gogh – The Letters

Si è già fatto accenno in precedenza¹⁵⁸ al progetto *Vincent Van Gogh – The Letters*, ma in questo paragrafo si condurrà un'analisi più dettagliata.

Il progetto (www.vangoghletters.org), realizzato da un gruppo di ricercatori diretti da Chris Stolwijk del Van Gogh Museum di Amsterdam, raccoglie tutte le lettere scritte e ricevute da Van Gogh per un totale di 902 missive.

Le lettere, tutte numerate, possono essere cercate in base al periodo in cui sono state scritte, al commento, al luogo dal quale sono partite e in base alla presenza o meno di disegni al loro interno.

È inoltre presente un pratico motore di ricerca per cercare un pattern all'interno delle lettere oppure nomi di persone, opere di letteratura o di arte, date, luoghi e molto altro. I risultati vengono visualizzati in una nuova pagina e contengono il collegamento alla lettera all'interno della quale è presente l'oggetto della ricerca: il contesto viene visualizzato facendo scorrere il mouse sul collegamento, ma qualora si decida di entrare nel documento, l'oggetto della ricerca non sarà evidenziato.

La visualizzazione dei documenti è divisa in due colonne: all'interno di ciascuna di esse è possibile scegliere se visualizzare la trascrizione in olandese, la stessa trascrizione con indicati i numeri delle righe, le immagini degli originali, la traduzione in inglese, le note o le immagini delle opere d'arte citate. Questa caratteristica è particolarmente interessante per il fatto che si possono visualizzare nello stesso momento la trascrizione in olandese e la traduzione.

Il sito inoltre è completato da alcuni contenuti extra: sono presenti saggi sulle lettere di Van Gogh, informazioni sui suoi corrispondenti, notizie di carattere storico e biografico, la storia della pubblicazione delle lettere, una bibliografia, un testo che illustra le fasi del progetto e altro ancora.

Le lettere sono state codificate in XML in base a uno schema creato appositamente per il progetto a partire dagli schemi TEI e DALF (anche se sul sito viene specificato che sono state usate anche alcune etichette non standard)¹⁵⁹.

158 Cfr. *supra* PAR. 2.4.

159 http://vangoghletters.org/vg/about_6.html#intro.VI.6.1.

4. Carteggi d'artista in rete: lo stato dell'arte

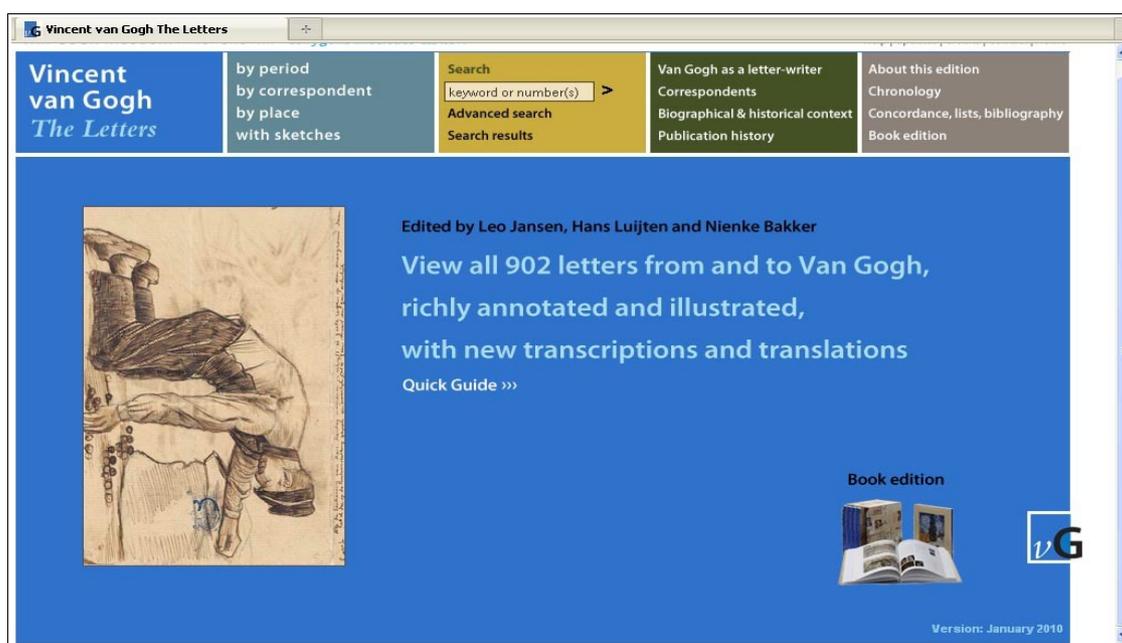


FIG. 4.12: home page di *Vincent Van Gogh – The Letters*

Si tratta di un sito completo, esauriente non soltanto in relazione ai documenti ma anche per quanto riguarda i contenuti aggiuntivi, nonché costruito in maniera ottima dal punto di vista grafico.

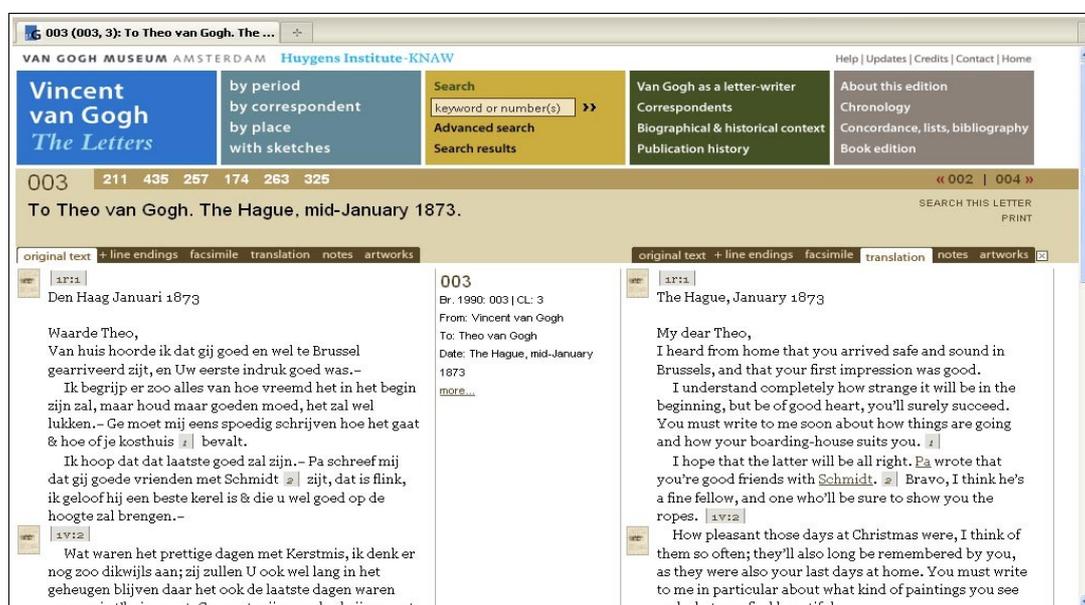


FIG. 4.13: una lettera di *Vincent Van Gogh – The Letters* con la caratteristica

doppia visualizzazione: a sinistra l'originale in olandese e a destra la traduzione

Anche un progetto così completo però dimostra di non rispettare gli standard del W3C,

4. Carteggi d'artista in rete: lo stato dell'arte

benché gli errori siano comunque minori per numero e per gravità rispetto a quelli di altri siti analizzati in precedenza.

4.8 *Google Books*

È possibile considerare *Google Books* come fonte importante per fare ricerche sui carteggi d'artista? *Google Books* (<http://books.google.com>) è un progetto nato nel 2004 che prevede la digitalizzazione di libri antichi e contemporanei: attualmente Google dichiara di avere all'interno delle pagine del progetto circa sette milioni di testi¹⁶⁰. L'acquisizione di un così elevato numero di testi è stata possibile grazie agli accordi stipulati da Google con case editrici e università.

Fare una ricerca all'interno di *Google Books* è estremamente semplice: è sufficiente inserire il titolo del libro nell'apposito campo per ottenere i risultati. È possibile anche cercare pattern presenti nel testo: il motore di ricerca rimanderà l'utente direttamente alla pagina del libro all'interno del quale è contenuto il pattern cercato.

Tra i libri presenti su *Google Books* è possibile reperire anche alcune edizioni di carteggi d'artista: per esempio, una selezione di lettere di Michelangelo pubblicata nel 2002.

Viste tali premesse, è quindi possibile utilizzare *Google Books* per le proprie ricerche sui carteggi d'artista? Allo stato attuale, il portale può essere un utile strumento di ricerca, ma per quanto riguarda i carteggi d'artista è piuttosto povero. Il materiale presente sul sito infatti è decisamente scarso, e la maggior parte delle volte è incompleto: questo significa che non tutte le pagine di un libro sono digitalizzate, e capita spesso che mentre si sta leggendo un testo interessante, la lettura debba essere interrotta perché mancano le pagine successive a quella che si sta leggendo. Ancora, molto spesso il testo non è selezionabile perché le pagine sono visualizzate come se fossero immagini (sono pochi i casi in cui *Google Books* permette di visualizzare il libro in formato immagine e in formato testo: si tratta di libri il cui copyright è scaduto). Infine, spesso i libri non sono digitalizzati in maniera ottimale: capita quindi spesso di imbattersi in libri con le pagine capovolte o che presentano abrasioni e cancellature. Si tratta quindi di uno strumento che per alcuni tipi di ricerca può essere di aiuto, ma allo stato attuale è poco utile per i carteggi d'artista.

Quindi, per ricerche approfondite, è meglio utilizzare strumenti pensati per scopi specifici

¹⁶⁰ <http://books.google.com/intl/it/googlebooks/agreement/>

4. Carteggi d'artista in rete: lo stato dell'arte

e realizzati da esperti in storia dell'arte, i quali sicuramente comprendono meglio di Google le esigenze di studiosi di arte, studenti ma anche di appassionati.

5. Una proposta metodologica: fasi preliminari

5.1 Introduzione

Nelle prossime sezioni verrà illustrata la proposta metodologica per creare edizioni *on-line* di carteggi d'artista di qualità. La proposta è strutturata in tre sezioni (fasi preliminari, organizzazione delle ricerche, implementazione del sito), a loro volta articolate in diverse fasi. Nella presente sezione vengono illustrate le fasi preliminari, che sono:

- l'acquisizione delle lettere (originali e trascrizione);
- la tokenizzazione dei testi;
- la lemmatizzazione e l'annotazione morfologica;
- la trasformazione dei testi annotati in file XML;
- la costruzione della base di dati MySQL;
- la trasformazione dei file XML in record MySQL;
- la presentazione delle lettere.

5.2 Acquisizione delle lettere

Prima di iniziare ad acquisire le lettere (cioè a trasferire i carteggi dal supporto cartaceo a quello digitale), è necessario sapere se si dispone dei manoscritti originali, se si hanno pubblicazioni cartacee dei carteggi, se si è in possesso della strumentazione necessaria.

Nel caso in cui si posseggano i manoscritti e si debba procedere alla loro *digitalizzazione* (ovvero alla loro trasformazione in immagini digitali) è necessario scegliere quale tipo di strumento utilizzare per poter ottenere le copie digitali dei manoscritti.

Innanzitutto è bene precisare che un'immagine elettronica non è altro che un insieme di *pixel* (*picture elements*: sono le unità minime dell'immagine, assimilabili a dei piccoli puntini): maggiore è il numero di pixel, maggiore è la *risoluzione* (assimilabile alla qualità) dell'immagine. Durante la trasformazione della fonte cartacea in formato digitale, gli strumenti per la digitalizzazione eseguono “mappe di bit”, ovvero creano una sorta di “griglia” dei pixel che rappresentano l'immagine, e a ognuno di essi viene assegnato un certo numero di bit per ottenere il colore (per esempio, un'immagine a sedici colori avrà bisogno di quattro bit per ogni pixel, perché ogni bit può assumere due valori, cioè 0 o 1, e il logaritmo in

5. Una proposta metodologica: fasi preliminari

base due di sedici è proprio quattro).

Per avere un'idea di cosa sia una “mappa di bit” (*bitmap*) è possibile pensare a un mosaico: ogni tessera rappresenta un pixel dell'immagine.

In genere ci sono due modi per digitalizzare manoscritti: o si esegue una scansione attraverso uno *scanner*, o si opta per l'utilizzo di fotocamere digitali. Prima di procedere all'acquisizione delle immagini però è necessario seguire alcune regole precise:

- i manoscritti devono essere maneggiati con cura (gli operatori dovrebbero indossare guanti di lattice);
- è necessario verificare che i manoscritti si trovino in uno stato di conservazione che possa consentire la digitalizzazione;
- i manoscritti devono essere preparati nel migliore dei modi per due ragioni, ovvero per far sì che l'intervento possa essere meno invasivo possibile e per ottenere un risultato di alta qualità;
- è necessario valutare la qualità degli strumenti che si possiedono in modo che i manoscritti non vengano danneggiati.

Se i manoscritti sono facilmente deperibili, si dovrà escludere l'utilizzo di uno scanner piatto, perché prevede la pressione sui fogli, e si dovrà ricorrere a uno scanner planetario oppure a una fotocamera digitale.

Lo scanner piatto permette di acquisire l'immagine del documento facendo scorrere una matrice (sulla quale sono posti dei sensori ottici che permettono di rilevare la mappa dell'immagine) sulla superficie da acquisire. Con questo strumento, il foglio da digitalizzare deve essere posizionato “a faccia in giù” e pressato sul piano d'appoggio dello scanner in modo da ridurre gli effetti di ombra e di curvatura delle immagini.

Lo scanner planetario è uno strumento molto più sofisticato dei normali scanner piatti e permette di digitalizzare manoscritti e volumi anche molto spessi in modo molto meno invasivo rispetto agli scanner piatti. Lo scanner planetario infatti esegue la scansione dall'alto, per cui non c'è bisogno di fare pressione sui documenti. Inoltre, le componenti di uno scanner planetario sono soggette a una minore usura rispetto a quelle di uno scanner piatto: questo fa sì che gli interventi di manutenzione siano ridotti al minimo e la vita dello scanner si allunghi in maniera consistente.

Altrimenti è possibile acquisire le immagini con l'utilizzo di fotocamere digitali ad alta qualità: sarà sufficiente riprendere i manoscritti con l'obiettivo per ottenere le immagini

5. Una proposta metodologica: fasi preliminari

digitali.

Gli originali, una volta acquisiti, dovranno essere opportunamente catalogati e inseriti nella cartella che ospiterà il progetto. Più in dettaglio, saranno inseriti in una sottocartella che sarà chiamata “immagini”. Nel caso in cui il corpus contenga carteggi di più artisti, sarà necessario creare una cartella diversa per ogni artista, denominata “NomeCognome”. I fogli andranno numerati tenendo conto dell'id (identificatore) progressivo della lettera e del numero del foglio. L'id può essere assegnato secondo diverse ragioni: per *Carlo Finelli Corpus* gli id sono stati assegnati in ordine cronologico (la lettera contrassegnata dal numero uno è quella più antica). Quindi, le immagini dei manoscritti avranno come nome del file, per esempio, 1-3.jpg (lettera 1, foglio 3), 6-2.jpg (lettera 6, foglio 2), 12-5.jpg (lettera 12, foglio 5) e così via.

Il percorso dei file quindi dovrà essere il seguente: “immagini/NomeCognome/idlettera-idfoglio.jpg”, per esempio “immagini/CarloFinelli/1-1.jpg” o “immagini/AntonioCanova/3-6.jpg”.

Anche qualora il corpus sia composto da lettere di più artisti, l'identificatore dovrà essere *univoco*, ovvero non ci dovranno essere due lettere con lo stesso numero (quindi non sarà possibile avere una lettera numero 1 di Carlo Finelli e una lettera numero 1 di Antonio Canova). Questo aspetto è particolarmente importante affinché il computer possa evitare fraintendimenti. Per esempio, come si vedrà in seguito¹⁶¹, in fase di ricerca il sistema restituisce i contesti basandosi non solo sui pattern cercati dall'utente, ma anche sugli id delle lettere: se ci fossero due lettere con lo stesso id, i risultati della ricerca sarebbero falsati in quanto verrebbero restituiti anche contesti errati.

Si potrebbe obiettare dicendo che è possibile assegnare id non univoci pur garantendo l'assenza di fraintendimenti, per esempio associando a ogni id il nome dell'autore. Si è deciso invece di assegnare id univoci per rendere più semplici le operazioni: la restituzione del contesto in fase di ricerca, in caso di id univoco, avviene controllando soltanto l'id della lettera, mentre in altri casi bisognerebbe controllare non solo l'id ma anche, per esempio, l'autore. In questo esempio quindi l'id univoco consente di risparmiare tempo prezioso durante le fasi di ricerca.

Dopo aver catalogato in modo opportuno le immagini, sarà necessario eseguire la trascrizione delle lettere. Per questo è importante disporre anche di un'edizione a stampa dei carteggi: ciò consentirà di rendere le operazioni più automatiche e quindi più veloci. Infatti ci

¹⁶¹ Cfr. *infra* PAR. 6.2.

5. Una proposta metodologica: fasi preliminari

sono in commercio programmi di OCR (*Optical Character Recognition*, “riconoscimento ottico dei caratteri”), che permettono di eseguire la trascrizione automatica dei caratteri presenti in un'opera stampata e di “ottenere buoni risultati su pagine stampate di recente, uniformemente inchiostrate e senza danneggiamenti al supporto cartaceo”¹⁶². Ovviamente però anche tali programmi sono soggetti a errori, e sarà quindi necessario un controllo finale per correggere i caratteri riconosciuti in modo errato dal programma. Nel caso in cui si disponga soltanto dei manoscritti, sarà necessario procedere a una trascrizione manuale. La trascrizione manuale è preferibile soprattutto se il numero di lettere da acquisire è esiguo: si evita l'utilizzo dello scanner (con conseguente preparazione delle impostazioni e dei testi da sottoporre a digitalizzazione) e si evitano le fasi di controllo delle trascrizioni automatiche.

Le trascrizioni saranno salvate in file di tipo *txt*, ai quali sarà dato come nome l'id della lettera, e saranno inserite in una cartella chiamata “lettere”. Anche in questo caso, se si vuole creare un corpus con lettere di più artisti, sarà necessario creare apposite cartelle.

5.3 Tokenizzazione

Come si è accennato sopra¹⁶³, “tokenizzare” un testo significa ridurlo in “token”, ovvero in unità minime, pronte per essere analizzate a livello approfondito. Benché apparentemente semplice, la tokenizzazione di un testo richiede alcune importanti considerazioni.

Si potrebbe, per esempio, far coincidere i token con i singoli grafemi presenti nel testo. Si prenda come esempio questa frase tratta da una lettera di Carlo Finelli: “Resto con augurarvi le buone feste”. Nel caso in cui si decida di far corrispondere i token ai singoli grafemi, “resto”, “con”, “augurarvi”, “le”, “buone” e “feste” costituirebbero token unici. Dal punto di vista della grammatica però la parola “augurarvi” è costituita da due forme ben distinte, ovvero l'infinito presente del verbo “augurare” e il pronome complemento di termine enclitico “vi”.

I token potrebbero quindi essere considerati come unità divise da uno spazio. Ma in questo caso si pone il problema dei segni di interpunzione, perché questi ultimi verrebbero associati alle forme che li precedono. Nel caso della frase di Carlo Finelli “la accludo alla presente con maggior sollecitudine.” i token sarebbero “la”, “accludo”, “alla”, “presente”, “con”, “maggior” e “sollecitudine.”. Però in questo caso il punto verrebbe incluso in un token unico

162 Andrea Bozzi, “Edizione elettronica e filologia computazionale”, 231.

163 Cfr. *supra* PAR. 2.5.

5. Una proposta metodologica: fasi preliminari

assieme alla forma “sollecitudine”, e questo non è possibile. Il *tokenizzatore* (ovvero il programma che esegue la tokenizzazione) dovrà quindi tener conto anche dei segni di interpunzione e far sì che essi vengano riconosciuti come token unici. Il punto però può essere utilizzato anche per altri scopi, come nell'espressione “Stimatissimo Sig. Segretario”, nella quale il punto serve per abbreviare la parola “signore”: il tokenizzatore dovrà tenere conto anche di queste particolari espressioni.

Altri problemi possono essere dati dalle forme composte (un esempio è la parola “Capo d'Anno” presente in una lettera di Finelli), dalle polirematiche (ovvero espressioni composte da più grafemi ma dal significato unico, per esempio “di tanto in tanto”), dalle date (“2 ottobre 1803”), dalle misure (“palmi 2,8”).

Le lettere di *Carlo Finelli Corpus* sono state tokenizzate attraverso l'utilizzo di *Token-IT*, tokenizzatore sviluppato dall'Istituto di Linguistica Computazionale del CNR di Pisa. *Token-IT* si avvale dell'uso di repertori (acquisiti da dizionari e da testi)¹⁶⁴ per riconoscere correttamente i token tenendo conto delle specificità alle quali si è fatto cenno.

Token-IT riceve in input un file di tipo *txt* e restituisce una pagina con la tokenizzazione del testo. Prendendo come esempio la frase “Stimatissimo sig. Camillo dal gentilissimo suo fratello ho ricevuto il di lei foglio gratissimo, nel quale ho inteso tutte le ragionevoli correzioni”, *Token-IT* restituisce il seguente output:

```
Stimatissimo
sig.
Camillo
dal
gentilissimo
suo
fratello
ho
ricevuto
il
di
lei
foglio
gratissimo
,
nel
quale
ho
```

¹⁶⁴ Alessandro Lenci, “Le tecnologie della lingua per il Knowledge Management nelle piatte forme di e-learning” in *Wbt.it Web Magazine*, Dicembre 2004, <http://www.giornalelearning.it/httpdocs/index.php?risorsa=tecnologie_della_lingua>

5. Una proposta metodologica: fasi preliminari

```
inteso
tutte
le
ragionevoli
correzioni
```

TAB. 5.1: tokenizzazione di una frase tratta da una lettera di Carlo Finelli

Dopo aver eseguito correttamente la tokenizzazione ed aver inserito i file risultati in un'apposita cartella, chiamata “filetokenizzati”, è possibile passare alla fase successiva, che consiste nella lemmatizzazione e nell'annotazione del testo.

5.4 Lemmatizzazione e annotazione morfologica

“Lemmatizzare” un testo significa associare a ogni forma il proprio lemma¹⁶⁵, mentre annotarlo a livello morfologico significa associare a ciascuna occorrenza la propria categoria grammaticale: queste informazioni abbinate alle forme presenti in un testo sono utili per svolgere ricerche, per estrarre informazioni, per costruire modelli probabilistici, per meglio comprendere il modo di scrivere di un autore.

La lemmatizzazione e l'annotazione morfologica sono due operazioni distinte ma sono strettamente legate l'una all'altra: pertanto, molti programmi che eseguono la lemmatizzazione eseguono anche l'annotazione morfologica.

Questo vale anche per i due PoS-tagger di cui si è parlato in precedenza, e cioè *MAGIC* e *TagPro*: ambedue infatti eseguono sia la lemmatizzazione che l'analisi morfologica.

Per la lemmatizzazione e l'annotazione morfologica di *Carlo Finelli Corpus* è stato utilizzato *MAGIC*: il programma riceve in input un testo tokenizzato e restituisce il testo lemmatizzato e annotato a livello morfologico. Si prenda come esempio la frase di Carlo Finelli “Astimone ricondotto a Crise da Ulisse per ordine di Agamennone è il soggetto di questo bassorilievo.”. Dopo aver correttamente eseguito la tokenizzazione, il testo viene passato a *MAGIC* che restituisce questo risultato:

```
Astimone ASTIMONE#SP@NN#
ricondotto RICONDURRE#V@MSPR#
a A#E@# A#S@FP@FS#
Crise CRISE#SP@NN#
da DA#E@#
```

165 Cfr. *supra* PAR. 2.5.

5. Una proposta metodologica: fasi preliminari

```
Ulisse ULISSE#SP@FS@MS# ULISSE#SP@NN#
per PER#E@#
ordine ORDINE#S@MS#
di DI#E@#
Agamennone AGAMENNONE#SP@FS@MS# AGAMENNONE#SP@NN#
e' ESSERE#V@S3IP#
il IL#RD@MS#
soggetto SOGGETTO#A@MS# SOGGETTO#S@MS#
di DI#E@#
questo QUESTO#D@MS# QUESTO#P@MS#
bassorilievo BASSORILIEVO#S@MS#
. #@@#
```

TAB. 5.2: lemmatizzazione di una frase tratta da una lettera di Carlo Finelli

Come si può ben notare, a ogni occorrenza viene associato il lemma, scritto interamente con caratteri maiuscoli. Per comprendere invece quale sia il significato delle ulteriori annotazioni inserite da MAGIC, è necessario fornire qualche delucidazione sul *tagset* utilizzato. Per *tagset* si intende un gruppo di etichette (“tag”) da utilizzare per l'annotazione morfologica.

MAGIC utilizza il tagset EAGLES-ILC, un tagset sviluppato dall'Istituto di Linguistica Computazionale del CNR di Pisa, conforme alle linee guida del progetto EAGLES (*Expert Advisory Group on Language Engineering Standards*, il cui sito internet è ufficiale è <http://www.ilc.cnr.it/EAGLES/home.html>), un'iniziativa della Commissione Europea nata con l'obiettivo di “accelerare la definizione di standard per: risorse linguistiche su vasta scala (quali corpora testuali, lessici computazionali e corpora di parlato); metodologie per il riutilizzo di tali conoscenze mediante formalismi di linguistica computazionale, linguaggi di mark up e vari strumenti software; metodologie per la valutazione di risorse, strumenti e prodotti”¹⁶⁶.

Lo schema di annotazione previsto da EAGLES¹⁶⁷ prevede l'utilizzo di dodici categorie morfosintattiche principali: aggettivo, articolo, avverbio, congiunzione, determinatore, interiezione, preposizione, nome, numerale, verbo, pronome, residuale. A ogni occorrenza vengono associati anche i tratti morfologici opportuni: per esempio, “maschile singolare” per un aggettivo, “indicativo presente terza persona singolare” per un verbo, e così via.

Lo schema EAGLES-ILC utilizza sigle diverse per ognuna delle categorie: A (aggettivi), B (avverbi), C (congiunzioni), D (determinatori), E (preposizioni), I (interiezioni), N (numerali),

166 <http://www.ilc.cnr.it/viewpage.php/sez=ricerca/id=59/vers=ita>

167 Per un'analisi più approfondita di EAGLES e di altri schemi di annotazione cfr. Giulia Venturi, *Rassegna comparativa degli schemi di annotazione morfosintattica per la lingua italiana* (Roma: Università di Roma Tre, 2009). Testo disponibile online all'indirizzo http://host.uniroma3.it/laboratori/triple/Risorse_Resources_files/RRT1_Venturi_xsito.pdf.

5. Una proposta metodologica: fasi preliminari

P (pronomi), R (articoli), S (nomi), V (verbi). Quindi a ogni forma viene associata la propria categoria di riferimento, oltre a ulteriori etichette per specificare i tratti della forma: per esempio, V@S1IP indica un verbo alla prima persona singolare del presente indicativo, S@MS un sostantivo maschile singolare, e così via.

Nel caso della frase di esempio, “Astimone” viene riconosciuto come nome proprio (SP) invariabile (NN), “ricondotto” come participio passato (PR) maschile singolare (MS) di un verbo (V), mentre per “a” vengono proposti due abbinamenti: preposizione oppure sostantivo.

Il programma di annotazione infatti spesso non riesce a capire da solo quale sia la categoria corretta all'interno della quale inserire una forma. Altrettanto spesso, viene individuata correttamente la categoria, ma il programma non sa quali tratti associare: è il caso di “Ulisse”, che viene correttamente riconosciuto come nome proprio, ma il programma non sa decidere se si tratti di un nome invariabile, di un maschile singolare o di un femminile singolare (e, tra l'altro, il tagset EAGLES-ILC non prevede la distinzione in maschile e femminile per i nomi propri¹⁶⁸).

Si pone quindi necessaria un'operazione di *disambiguazione*: dopo aver completato la lemmatizzazione e l'analisi morfologica, è fondamentale ricontrollare il testo al fine di cercare ambiguità e risolverle manualmente. Solo in questo modo sarà possibile passare alla fase successiva.

Pertanto, la frase precedente lemmatizzata correttamente e disambiguata darà questo risultato:

Astimone ASTIMONE#SP@NN# ricondotto RICONDURRE#V@MSPR# a A#E@# Crise CRISE#SP@NN# da DA#E@# Ulisse ULISSE#SP@NN# per PER#E@# ordine ORDINE#S@MS# di DI#E@# Agamennone AGAMENNONE#SP@NN# e' ESSERE#V@S3IP# il IL#RD@MS# soggetto OGGETTO#S@MS# di DI#E@# questo QUESTO#D@MS# bassorilievo BASSORILIEVO#S@MS# .#@#@#
--

TAB. 5.3: lemmatizzazione e disambiguazione di una frase tratta da una lettera di Carlo Finelli

168 Cfr. la pagina web del tagset EAGLES-ILC, <http://webilc.ilc.cnr.it/~lenci/Tagset%20ILC.htm>.

5. Una proposta metodologica: fasi preliminari

Dopo aver disambiguato correttamente il testo ed aver inserito i file in una cartella chiamata “filelemmatizzati”, è possibile passare alla fase successiva.

5.5 Trasformazione dei testi lemmatizzati e annotati in file XML

L'operazione successiva è la trasformazione dei testi in file XML che siano conformi alla seguente DTD:

```
<!ELEMENT testo (id, token+)>
<!ELEMENT id (#PCDATA)>
<!ELEMENT token (forma, lemma, categoria, info)>
<!ELEMENT forma (#PCDATA)>
<!ELEMENT lemma (#PCDATA)>
<!ELEMENT categoria (#PCDATA)>
<!ELEMENT info (#PCDATA)>
```

Tab. 5.4: DTD per i token

L'elemento radice è l'elemento “testo”, il quale deve contenere un solo elemento “id” (l'identificatore) e almeno un elemento “token”. Quest'ultimo invece può contenere un solo elemento “forma”, un solo elemento “lemma”, un solo elemento “categoria” (la categoria grammaticale) e un solo elemento “info” (le note per specificare i tratti morfologici della forma). Questi quattro elementi possono contenere soltanto testo (“#PCDATA”).

Come è possibile trasformare il testo nella tabella 5.3 in un file XML? La prima operazione da compiere è “pulire” il file annotato, eliminando i cancelletti e le chioccioline, in modo tale da avere, in ogni riga, quattro elementi (oppure tre, nel caso non siano presenti i tratti morfologici, come per le preposizioni semplici) separati da un unico spazio. Nel caso dei segni di interpunzione, la notazione “#@@#” verrà sostituita da tre trattini. Si tratta di operazioni che si possono svolgere in modo automatico e rapido anche attraverso l'utilizzo di un semplice editor di testi:

Astimone ASTIMONE SP NN
ricodotto RICONDURRE V MSPR

5. Una proposta metodologica: fasi preliminari

```
a A E
Crise CRISE SP NN
da DA E
Ulisse ULISSE SP NN
per PER E
ordine ORDINE S MS
di DI E
Agamennone AGAMENNONE SP NN
e' ESSERE V S3IP
il IL RD MS
soggetto OGGETTO S MS
di DI E
questo QUESTO D MS
bassorilievo BASSORILIEVO S MS
. . ---
```

TAB. 5.5: il file annotato dopo la “pulizia”

È necessario prestare particolare attenzione alle forme che presentano enclitici. Per esempio, alla forma “dicendomi” viene associata la seguente etichetta: DIRE#V@G<+MI#PQ@MS1#>#. In questi casi, la notazione sarà pulita in questo modo: DIRE V G+MI-PQ (MS1). Si elimineranno quindi le parentesi uncinate, il primo cancelletto sarà sostituito con un trattino e i tratti dell'enclitico saranno inseriti tra parentesi tonde.

A questo punto il file di testo è pronto per essere trasformato in un file XML. Per poter fare ciò è necessario disporre di un codice in linguaggio *Perl*. Quest'ultimo è un linguaggio di programmazione molto potente, usato in particolar modo per trattare file di testo, come in questo caso. Grazie a Perl e soprattutto grazie alle espressioni regolari (una “sorta di linguaggio interno al Perl”¹⁶⁹) è possibile automatizzare operazioni che, se eseguite manualmente, potrebbero richiedere tempistiche improponibili. In questa sede non si discuterà nel dettaglio il linguaggio Perl, ma si proporrà e si analizzerà unicamente il codice (realizzato attraverso le espressioni regolari di Perl) utile per poter trasformare i file di testo annotati in file XML. Il codice pensato per eseguire tale operazione è il seguente:

```
while (<>) {

s/^(([\s]+) ([^\s]+) ([^\s]+) (.+)\$ /
<token><forma>\1<\ /forma><lemma>\2<\ /lemma><categoria>\3<\ /cate
goria><info>\4<\ /info><\ /token>/g;

s/^(<.*'"$) / \1 \\> /g;

print

}
```

TAB. 5.6: codice in linguaggio Perl per la trasformazione dei file annotati in file XML

169 Edoardo Valsesia, *Guida a Perl in Html.it*, <<http://programmazione.html.it/guide/leggi/37/guida-perl/>>

5. Una proposta metodologica: fasi preliminari

Nella riga

```
s/^( ([^\s]+) ([^\s]+) ([^\s]+) (.+) )$/
```

“^” è il simbolo di inizio riga, mentre “\$” di fine riga. I caratteri `[^\s]` identificano tutto ciò che non è uno spazio (i caratteri “\s” nelle espressioni regolari di Perl si utilizzano proprio per cercare gli spazi): sono state inserite tre di queste sequenze (tra parentesi tonde e seguita da un più) per identificare i primi tre elementi di una riga, separati da uno spazio (e cioè la forma, il lemma e la categoria grammaticale). Le parentesi tonde servono per raggruppare e memorizzare i pattern, mentre il più indica di considerare un intervallo di caratteri che va da uno a infinito (come il più in XML). L'ultima sequenza è “(.+)” che prende tutto ciò che si trova dopo l'ultimo spazio fino alla fine della riga (quindi i tratti morfologici, qualora ci siano), dal momento che il punto è l'operatore che considera qualsiasi carattere.

Memorizzati i quattro pattern, si dà il via alla sostituzione con le etichette XML:

```
<token><forma>\1<\form><lemma>\2<\lem><categoria>\3<\cate  
goria><info>\4<\info><\token>/g;
```

La sostituzione è garantita dal carattere “g”: viene creato un file con le etichette XML (“token”, “forma”, “lemma”, “categoria”, “info”) all'interno delle quali vengono inseriti i quattro pattern precedentemente memorizzati, ognuno con il proprio numero progressivo (uno, due, tre e quattro).

Il risultato sarà un file XML i cui elementi saranno inseriti all'interno dell'elemento “testo” (come da DTD). Inoltre al file risultante si dovranno aggiungere l'intestazione, l'indicazione della DTD e l'identificatore di riferimento, nell'apposita etichetta. Fatto questo, si sostituiranno gli apostrofi con gli accenti dove necessario (per esempio, la forma “e” sarà sostituita con la forma “è”). È assolutamente necessario che nell'intestazione del file XML venga specificato il tipo di codifica dei caratteri, ovvero ISO-8859-1: se manca questa specificazione, le lettere accentate non saranno correttamente riconosciute e il file restituirà un errore.

I file saranno numerati con l'identificatore univoco della lettera del quale si è parlato sopra¹⁷⁰: quindi i file saranno chiamati “1.xml”, “2.xml”, “3.xml”. I file serviranno per

¹⁷⁰ Cfr. *supra* PAR. 5.2.

5. Una proposta metodologica: fasi preliminari

l'inserimento nella base di dati MySQL e saranno inseriti in una cartella “token”, creata all'interno di una ulteriore cartella denominata “xml”.

All'interno di quest'ultima cartella, ne sarà creata anche una seconda chiamata “lettere”, che dovrà contenere i file XML con i dati e il testo puro delle lettere. Tali file dovranno rispettare la seguente DTD:

```
<!ELEMENT lettera (metadata, testo)>

<!ELEMENT metadata (autore, destinatario, raccolta,
conservazione, luogo, data, indirizzomit, indirizzodes,
originali, note)>

<!ELEMENT autore (#PCDATA)>

<!ELEMENT destinatario (#PCDATA)>

<!ELEMENT raccolta (#PCDATA)>

<!ELEMENT conservazione (#PCDATA)>

<!ELEMENT luogo (#PCDATA)>

<!ELEMENT data (#PCDATA)>

<!ELEMENT indirizzomit (#PCDATA)>

<!ELEMENT indirizzodes (#PCDATA)>

<!ELEMENT originali (#PCDATA)>

<!ELEMENT note (#PCDATA)>

<!ELEMENT testo (#PCDATA)>
```

TAB. 5.7: DTD per le lettere

Anche in questo caso i file XML prodotti in modo conforme a questa DTD serviranno per l'inserimento nella base di dati MySQL.

L'elemento radice, “lettera” sarà composto di due elementi, “metadata” e “testo”. Quest'ultimo conterrà il testo della lettera, formattato in linguaggio HTML (*HyperText Markup Language*: la formattazione in HTML è importante perché il sito su cui sarà pubblicato il corpus prenderà i dati proprio dalla base di dati MySQL: è importante quindi che nella base di dati sia inserito il testo già formattato).

La formattazione in HTML dovrà avvenire prestando attenzione alle parentesi uncinate, ai

5. Una proposta metodologica: fasi preliminari

doppi apici e in generale a tutti i caratteri che potrebbero non essere correttamente interpretati dal file XML: per esempio, le parentesi uncinata dovranno essere sostituite con le opportune sequenze di caratteri (“<” per la parentesi uncinata aperta e “>” per la parentesi uncinata chiusa). In caso di più autori o più destinatari, verranno specificati tutti nello stesso elemento utilizzando come spaziatore l'etichetta “
” del codice HTML. Si è preferito utilizzare questo sistema, piuttosto che l'annotazione con più etichette XML (una etichetta per ogni autore) in modo da rendere più agevole la trasformazione in MySQL e in modo da rendere più elegante il testo che sarà presentato all'utente.

La stessa cosa vale per tutte le altre etichette della DTD. Qualora un'informazione non sia presente (per esempio, qualora manchino gli indirizzi o gli originali), l'etichetta sarà riempita con la dicitura “Assente”.

La data inoltre sarà inserita in formato aaaa-mm-gg (per esempio, 1807-12-05), in modo che possa essere correttamente interpretata dalla base di dati: quando la lettera sarà presentata all'utente, ci sarà un'apposita funzione che “tradurrà” la data in un formato più elegante (per esempio: Roma, 5 dicembre 1807). Nel caso in cui la data non sia presente, si inserirà la sequenza “0000-00-00”.

Nell'etichetta “originali” andranno inseriti i collegamenti alle immagini, anch'essi formattati in linguaggio HTML. Saranno numerati progressivamente, come in questo esempio:

```
<a href="immagini/CarloFinelli/1-1.jpg" target="_blank">Primo
foglio</a><br />
<a href="immagini/CarloFinelli/1-2.jpg" target="_blank">Secondo
foglio</a><br />

[...]
```

TAB. 5.8: formattazione per l'inserimento dei collegamenti alle immagini dei manoscritti

Si è pensato di creare due DTD volutamente molto semplici in primo luogo, come si è già detto, perché il file XML rappresenta soltanto un passaggio intermedio, e l'interscambio avverrà quindi non con XML ma con MySQL. In secondo luogo, perché è ipotizzabile che anche operatori non esperti di linguaggi di marcatura o di scripting potranno essere coinvolti nella costruzione di corpora di lettere di artisti: a uno storico dell'arte, specie se piuttosto “ancorato” alla tradizione e quindi poco avvezzo all'uso delle tecnologie, non può essere

5. Una proposta metodologica: fasi preliminari

richiesta una conoscenza approfondita del linguaggio XML o di altri linguaggi (anche le operazioni di costruzione della base di dati sono ridotte al minimo, e verranno eseguite mediante l'utilizzo di appositi programmi come PhpMyAdmin).

Di seguito si forniscono due esempi di file XML, uno per i token e uno per una lettera, formattati in modo conforme alle due DTD.

```
<?xml version="1.0" encoding="iso-8859-1"?>
<!DOCTYPE testo SYSTEM "CorpusToken.dtd">
<testo>

<id>8</id>

<token><forma>Stimatissimo</forma><lemma>STIMATO</lemma><categoria>A</categoria><info>MSS</info></token>

<token><forma>signor</forma><lemma>SIGNORE</lemma><categoria>S</categoria><info>MS</info></token>

<token><forma>segretario</forma><lemma>SEGRETARIO</lemma><categoria>S</categoria><info>MS</info></token>

[...]

</testo>
```

TAB. 5.9: file XML dei token formattato secondo la DTD

```
<?xml version="1.0" encoding="iso-8859-1"?>
<!DOCTYPE lettera SYSTEM "CarloFinelliCorpus.dtd">
<lettera>
<metadata>
<id>25</id>
<autore>Carlo Finelli</autore>
<destinatario>Giuseppe Zanoia</destinatario>
<conservazione>Milano, Archivio Storico dell'Accademia di Brera</conservazione>
<luogo>Roma</luogo>
<data>1807-12-05</data>
</metadata>
<testo>
testo... testo... testo... testo...
</testo>
</lettera>
```

TAB. 5.10: file XML della lettera formattato secondo la DTD

5.6 Costruzione della base di dati MySQL

Prima di procedere è necessario fornire le ragioni per le quali si è deciso di memorizzare il corpus in una base di dati MySQL piuttosto che proporre di inserire sul sito direttamente i file XML e di implementare quindi un sistema di ricerca all'interno degli XML.

MySQL¹⁷¹ è un DBMS (*Data Base Management System*), ovvero un prodotto software in grado di gestire basi di dati. Le caratteristiche peculiari di questo tipo di prodotti sono¹⁷²:

- la capacità di gestire grandi quantità di dati;
- la possibilità di condividere i dati con più utenti e applicazioni;
- la possibilità di utilizzare sistemi di protezione e autorizzazione per l'accesso ai dati.

MySQL è un database di tipo *relazionale*, ovvero basato sulle “relazioni tra i dati, i quali vengono presentati in forma tabulare, cioè come un insieme di tabelle ciascuna composta da righe e colonne”¹⁷³: il linguaggio SQL (*Structured Query Language*) permette di operare in modo molto versatile su questi dati attraverso l'utilizzo di semplici parole chiave.

Quali sono poi gli ulteriori vantaggi forniti da MySQL? Innanzitutto, dal momento che si tratta di un software distribuito con licenza libera GNU GPL (*GNU General Public License*), non comporta alcun costo da parte di chi vuole sviluppare un sistema che sfrutti questo tipo di base di dati. Inoltre, MySQL gode di una eccezionale diffusione e quindi in molti riescono a comprendere il suo funzionamento e il suo linguaggio. Ancora, MySQL garantisce ottime e rapide prestazioni ed è molto semplice nell'utilizzo e nella configurazione. Infine, rispetto a XML è molto più leggero: l'intero database di *Carlo Finelli Corpus* ha infatti un peso complessivo di 254,4 KB, mentre i file XML hanno un peso di 547 KB, ovvero più del doppio! MySQL consente inoltre di risparmiare anche la costruzione di singole pagine web per le lettere: sarà infatti sufficiente un'unica pagina, in linguaggio PHP, contenente un codice che richieda la lettera alla base di dati solo quando è necessario. Ciò rende anche le operazioni e la navigazione sul sito molto più veloci per l'utente.

La base di dati del corpus contiene due tabelle: una chiamata “lettere” e una chiamata “token”. La prima deve contenere dodici campi: “id”, “autore”, “destinatario”, “raccolta”, “conservazione”, “luogo”, “data”, “indirizzomit”, “indirizzodes”, “originali”, “note”, “testo”.

Il primo, “id”, è l'identificatore univoco delle lettere (il campo deve essere di tipo INT,

171 Il sito ufficiale di MySQL è www.mysql.com.

172 Gianluca Gillini, *Guida a MySQL in Html.it*, <<http://database.html.it/guide/leggi/87/guida-mysql/>>

173 *Ibid.*

5. Una proposta metodologica: fasi preliminari

ovvero un numero intero). I successivi otto campi (da “autore” fino a “indirizzodes”), a eccezione di “data” (che deve essere di tipo DATE), devono essere di tipo VARCHAR (*VARIABLE CHARACTER*, un insieme di carattere dalla lunghezza variabile), mentre gli ultimi tre (“originali”, “note” e “testo”) devono essere di tipo TEXT (il tipo che serve per memorizzare grandi quantità di dati).

La tabella “token” contiene sei campi: “id”, “idlettera”, “forma”, “lemma”, “categoria” e “info”. “Id” è l'identificatore univoco dei singoli token e deve essere di tipo INT, così come “idlettera”, che è l'identificatore della lettera alla quale si riferisce il singolo token. Gli altri quattro campi devono essere tutti di tipo VARCHAR.

È assolutamente importante che la base di dati venga costruita con la stessa codifica dei file XML: in caso contrario, si verificheranno inconvenienti e alcuni caratteri non verranno riconosciuti.

Una volta costruita la tabella è possibile passare alla fase successiva, la trasformazione dei file XML in record MySQL. Con “record” si intende ogni singola riga della base di dati MySQL.

5.7 Trasformazione dei file XML in record MySQL

La trasformazione dei file XML in record MySQL è possibile grazie all'utilizzo del linguaggio PHP. Si tratta di un linguaggio di scripting distribuito con licenza open source e creato nel 1994 dal danese (nato in Groenlandia) Rasmus Lerdorf: l'acronimo inizialmente aveva il significato di *Personal Home Page* (dal momento che nacque proprio come linguaggio per il web), ma in seguito ha assunto il significato di *PHP Hypertext Processor*. Si tratta di un linguaggio lato server: significa che i codici sono eseguiti solamente dal server nel momento in cui il client richiede una pagina. Il linguaggio PHP quindi fornisce un aiuto per inserire istruzioni nelle pagine HTML per creare un contenuto dinamico¹⁷⁴, ovvero per fare in modo che i contenuti delle pagine vengano generati solo a seguito di specifiche richieste del client.

I file devono avere estensione “.php” e il codice deve essere compreso tra due marcatori, quello di apertura (“<?php>”) e quello di chiusura (“?>”): i due marcatori servono per fare in modo che il preprocessore possa interpretare in modo corretto il codice PHP. Nelle tabelle che

¹⁷⁴ Rasmus Lerdorf, *Php* (Sebastopol, California: O'Really, 2000), trad. it. *Php* (Milano: Hops, 2003), traduzione di Marina Sturino.

5. Una proposta metodologica: fasi preliminari

illustreranno i codici utilizzati dalla proposta metodologica che qui si presenta, i marcatori saranno omessi in quanto di volta in volta si mostreranno solo porzioni di codice e non le intere pagine.

È necessario sottolineare che con la proposta qui discussa soltanto i codici PHP vogliono essere standard: la grafica e la struttura HTML infatti possono essere completamente personalizzabili a seconda delle esigenze di coloro che realizzano il corpus (anche se si daranno delle linee guida al fine di realizzare un prodotto elegante e di qualità anche dal punto di vista grafico e strutturale)¹⁷⁵.

Ogni variabile in PHP deve essere preceduta dal simbolo “\$”. Il codice seguente è quello che permette la trasformazione dei file XML relativi ai token in record di una base di dati MySQL:

```
$oDOM = new DOMDocument();

$oDOM->loadXML(file_get_contents($file));

foreach ($oDOM->getElementsByTagName('testo') as $Testo) {

    $idlettera=mysql_real_escape_string($Testo->
getElementsByTagName('id')->item(0)->nodeValue);

    foreach($oDOM->
getElementsByTagName('token')as $TokenNode)

        {

            $forma = mysql_real_escape_string($TokenNode
->getElementsByTagName('forma')->item(0)->nodeValue);

            $lemma = mysql_real_escape_string($TokenNode
->getElementsByTagName('lemma')->item(0)->nodeValue);

            $categoria = mysql_real_escape_string($TokenNode
->getElementsByTagName('categoria')->item(0)->nodeValue);

            $info = mysql_real_escape_string($TokenNode
->getElementsByTagName('info')->item(0)->nodeValue);

            $query = "INSERT INTO token (id, idlettera, forma,
lemma, categoria, info) VALUES ('NULL', '$idlettera', '$forma',
'$lemma', '$categoria', '$info)";

            $result=mysql_query($query);
```

Tab. 5.11: codice per trasformare i file XML in record MySQL

175 Cfr. *infra* CAP. 7.

5. Una proposta metodologica: fasi preliminari

La prima variabile, denominata “\$oDOM”, crea un'istanza appartenente alla classe “DOMDocument”: quest'ultima serve per rappresentare interi documenti HTML o XML¹⁷⁶. Attraverso il metodo “loadXML” della classe, invocato sulla variabile appena creata, è possibile caricare il file XML da processare. Il metodo riceve come parametro la funzione “file_get_contents” (serve per leggere un file all'interno di una stringa) che a sua volta ha come parametro la variabile “\$file”: si tratta di una variabile precedentemente creata, che corrisponde al nome del file inserito dall'operatore attraverso un semplice form.

Per l'operatore che deve processare il file infatti la pagina si presenta in un modo molto semplice: una casella bianca all'interno della quale deve essere inserito il nome del file (per esempio, “1.xml”) e un pulsante per poterlo inviare alla base di dati. Come si è detto nel paragrafo precedente, è ipotizzabile che nei progetti di costruzione di corpora vengano coinvolte anche persone non esperte di informatica, quindi la proposta metodologica qui discussa è stata ideata anche pensando a operatori le cui conoscenze informatiche sono elementari o addirittura nulle.

È importante, soprattutto in caso di corpora di grandi dimensioni, che vengano caricati singoli file XML per ogni testo e non un unico XML con milioni di occorrenze, per due motivi: innanzitutto, perché i tempi di trasformazione per un unico file con milioni di occorrenze sono elevati (a causa delle grandi dimensioni che il file XML avrebbe), e in secondo luogo perché è più logico trasformare i singoli file una volta completata l'annotazione, invece che riunirli poi tutti in un unico XML (sarebbe solo una inutile perdita di tempo). C'è poi da considerare che anche i programmi per l'annotazione morfologica sono poco performanti in caso di file con migliaia di occorrenze, quindi anche in fase di lemmatizzazione e annotazione è più logico intervenire su file di dimensioni contenute.

Dopo aver caricato il file, il codice fa partire due cicli “foreach”. Il “foreach” è una cosiddetta *struttura di controllo iterativo* (come “for” e “while”): le strutture di controllo iterativo permettono di eseguire in modo ciclico alcune operazioni finché si verificano certe condizioni. In particolare, foreach attraversa un “array” (ovvero una collezione di elementi) e per ognuno degli elementi dell'array compie determinate istruzioni.

Nel primo dei due foreach del presente codice, gli elementi dell'array da passare in rassegna non sono altro che le etichette di XML (riconosciute tramite il metodo “getElementsByTagName”) chiamate “testo” (quindi, di fatto, questo foreach scorre un solo

¹⁷⁶ Cfr. la documentazione ufficiale della classe all'indirizzo <http://hr.php.net/manual/en/class.domdocument.php>.

5. Una proposta metodologica: fasi preliminari

elemento). All'etichetta, che in PHP diventa una normale variabile, viene dato un nome provvisorio (“\$Testo”).

All'interno del ciclo viene dichiarata una variabile, che come valore assume il valore dell'etichetta “id” trovata all'interno dell'elemento “testo”: per riconoscere gli elementi “id” si utilizza il solito metodo “getElementsByTagName”, quindi si invoca il metodo “item” impostando l'indice a zero (serve per trovare il primo degli elementi che soddisfa la richiesta espressa dal codice), e per reperire il contenuto (“valore”) degli elementi “id” si utilizza invece la proprietà “nodeValue”. Il metodo “mysql_real_escape_string” serve invece per aggiungere le *sequenze di escape* ai caratteri speciali nelle istruzioni SQL. Cosa sono le sequenze di escape e a che cosa servono? L'escape non è altro che il carattere “\” e si pone davanti a caratteri che potrebbero non essere correttamente interpretati da SQL, come gli apici o i doppi apici¹⁷⁷. Se infatti nel testo è presente un apostrofo, SQL potrebbe interpretarlo come un apice e quindi come un'istruzione: questo potrebbe dare luogo a errori e inconvenienti. Aggiungendo le sequenze di escape, si impedisce che questo avvenga.

Le stesse operazioni fin qui analizzate avvengono nel secondo ciclo foreach per gli elementi “forma”, “lemma”, “categoria” e “info”: ognuno di essi viene memorizzato in una variabile. Terminata la memorizzazione, il codice può iniziare a trasferire i valori degli elementi trovati nella base di dati, utilizzando il linguaggio SQL. Pertanto si dà il via a una “query” (ovvero un'interrogazione al database, nel codice chiamata semplicemente “\$query”), con la quale si chiede di inserire nella tabella “token” e nei campi “id”, “idlettera”, “forma”, “lemma”, “categoria” e “info” i valori delle variabili contenenti gli elementi reperiti all'interno del file XML. Nel campo “id” viene invece inserito il valore “NULL” in quanto tale campo incrementa in automatico di una unità ogni volta che viene inserito un record.

Infine, il metodo “mysql_query” invia la query precedentemente creata alla base di dati e di fatto inserisce i valori trovati nel file XML all'interno della base di dati.

Quindi, riassumendo:

- si crea un oggetto della classe “DOMDocument”, utile per rappresentare file HTML e XML;
- attraverso l'oggetto creato si carica il file XML, il cui nome (per esempio, “1.xml”) viene passato da un operatore attraverso un semplicissimo form HTML;
- si aprono due cicli foreach che servono per reperire i valori degli elementi del file

¹⁷⁷ In programmazione, gli “apici” e i “doppi apici” non sono altro che i caratteri che in scrittura si utilizzano per gli apostrofi e le virgolette.

5. Una proposta metodologica: fasi preliminari

XML;

- si apre una query SQL con la quale si inseriscono i valori trovati negli appositi campi;
- si invia la query alla base di dati.

Queste operazioni vengono effettuate per ciascuno degli elementi: significa che il ciclo viene ripetuto per ciascuno dei token finché non sono finiti. Nella stessa pagina PHP inoltre sono stati inseriti alcuni comandi per la sostituzione di caratteri che potrebbero essere male interpretati da SQL: si tratta soprattutto delle lettere accentate. Attraverso query simili a questa:

```
$query2 = "UPDATE token SET forma = REPLACE(forma, 'Ã', 'è');"
```

è possibile evitare che caratteri non corretti compaiano sul sito. Basta semplicemente aggiornare la tabella (“UPDATE”) impostando i campi in modo tale da sostituire (“REPLACE”) i caratteri non corretti con quelli corretti.

La procedura per l'inserimento delle lettere è del tutto analoga a quella appena discussa.

Queste operazioni devono essere svolte in locale e non sul sito sul quale sarà ospitato il corpus: questo per non dover caricare sul sito anche i file XML. Sarà quindi sufficiente eseguire le operazioni in locale e poi trasferire la sola base di dati sul sito che ospiterà il corpus.

Così come è possibile inserire elementi XML all'interno della base di dati MySQL, è altrettanto possibile cancellarli. Il codice pensato per permettere tale operazione è il seguente:

```
$query = "DELETE FROM token WHERE idlettera LIKE '%
$_POST[cancellatesto]%' ";

$result = mysql_query($query);
```

TAB. 5.12: codice per cancellare i token dalla base di dati MySQL

Si tratta di una semplice query che chiede di cancellare dalla base di dati tutti i token la cui lettera di riferimento ha come identificatore quello inserito dall'operatore in un form, in particolare in un campo denominato “cancellatesto”. L'operatore in questo caso non deve far altro che inserire nell'apposito campo il numero della lettera dalla quale cancellare i token. Tale operazione può essere eseguita anche sul sito.

5. Una proposta metodologica: fasi preliminari

5.8 La presentazione delle lettere

Per la presentazione all'utente delle lettere del corpus è stata realizzata un'unica pagina PHP. Uno dei vantaggi di tale linguaggio di programmazione consiste proprio in questa caratteristica: è possibile realizzare un'unica pagina per poter mostrare un numero potenzialmente infinito di lettere.

La visualizzazione dei dati della lettera è possibile grazie a una query SQL presente nel codice PHP della pagina (chiamata “lettera.php”):

```
$query = "SELECT * FROM lettere WHERE id = '$_GET[id]'";  
$result = mysql_query($query);
```

TAB. 5.13: query per ottenere i dati della lettera richiesta

La query seleziona tutti i dati (attraverso l'utilizzo dell'operatore asterisco) della tabella “lettere” che soddisfano una precisa condizione, ovvero che nella “query string” dell'indirizzo web sia presente l'id della lettera.

Infatti, per poter permettere agli utenti di accedere alle lettere, sarà necessario creare una pagina (denominata “lettere.php”) contenente i collegamenti alle lettere, scritti nella seguente forma:

```
www.sitodelcorpus.xx/lettera.php?id=1
```

La query string è una stringa che fa parte di un URL (*Uniform Resource Locator*: un indirizzo web) e che serve per passare al server dei dati, e si separa dall'URL attraverso il punto interrogativo: quindi in questo caso, la query string dell'indirizzo sopraccitato è ?id=1.

La pagina PHP, trovando nell'indirizzo la query string, comprende che la condizione `WHERE id = '$_GET[id]'` è soddisfatta, perché nell'URL indicato è di fatto presente l'id della lettera richiesta. Quindi, selezionerà dal database tutti i dati della lettera che ha come id “1”.

Quindi, attraverso la funzione “`mysql_fetch_array`”, i risultati della query vengono restituiti sotto forma di array, in modo tale da poterli trattare per l'operazione successive, ovvero la presentazione (“stampa”) su schermo. La seguente porzione di codice specifica meglio quanto appena illustrato:

5. Una proposta metodologica: fasi preliminari

```
while ($row = mysql_fetch_array($result, MYSQL_ASSOC)) {  
    echo "<b>Data</b>: $row[luogo]";  
    formatodata();  
    echo "<br /><br />";  
    echo "<b>Destinatario</b>: $row[destinatario]<br /><br />";  
    if ($row['raccolta'] != "Assente") {  
        echo "<b>Raccolta</b>: $row[raccolta]<br /><br />";  
    }  
    [...]  
}
```

TAB. 5.14: codice per la presentazione della lettera richiesta

Attraverso la funzione “mysql_fetch_array” vengono create alcune variabili, dette “\$row”, che rappresentano le “righe” della tabella “lettere” (quindi autore, destinatario, data, ecc.). Attraverso il comando “echo”, che serve per “stampare” (cioè scrivere) sullo schermo i dati, è possibile presentare all'utente tutte le informazioni relative alle lettere. Per esempio, la riga di codice

```
echo "<b>Destinatario</b>: $row[destinatario]<br /><br />";
```

scrive sullo schermo l'indicazione “Destinatario” in grassetto (l'etichetta serve proprio per questo scopo) e di fianco la data della lettera richiesta (\$row[data]).

Per stampare sullo schermo la data viene invocata invece una particolare funzione, chiamata “formatodata”. Questa funzione è definita nel file “functions.php”, che contiene funzioni utili per il sito e che viene incluso all'interno di ciascuna pagina del sito stesso attraverso il comando `include("includes/funzioni.php");` (significa che il sito deve includere il file “funzioni.php” presente nella cartella “includes”).

La funzione “formatodata” stampa la data in un formato elegante (per esempio, “5 dicembre 1807” invece di “1807-12-05”) ed è definita dal seguente codice:

```
function formatodata() {  
    global $row;
```

5. Una proposta metodologica: fasi preliminari

```
if ($row['data'] != "0000-00-00") {  
    $data = $row['data'];  
  
    $mesi = array('gennaio', 'febbraio', 'marzo', 'aprile',  
                'maggio', 'giugno', 'luglio', 'agosto',  
                'settembre', 'ottobre', 'novembre', 'dicembre');  
  
    list($anno, $mese, $giorno) = split('-', $data);  
  
    $mese = $mesi[$mese-1];  
    $data = $giorno. ' '. $mese . ' ' . $anno;  
  
    echo ", $data";  
}  
  
else {  
    echo "";  
}  
}
```

TAB. 5.15: la funzione “formatodata”

Il primo passo da compiere è dichiarare una variabile globale (ovvero che vale per tutta la pagina e non solo per la porzione di codice all'interno della quale viene dichiarata), di nome “\$row”: non è altro che l'array con i record della base di dati e viene dichiarata come variabile globale in modo da poter riconoscere i record correttamente.

A questo punto avviene un controllo (tramite blocco if) sull'elemento “data” di \$row: se è diverso da “0000-00-00”, viene inizialmente creata una variabile chiamata “\$data” che assume come valore la data della lettera richiesta. Successivamente viene creato un nuovo array denominato “\$mesi”, contenente i nomi dei mesi in italiano. Quindi la funzione seguente, “list”, crea tre nuove variabili (“\$giorno”, “\$mese”, “\$anno”) facendo uso del metodo “split”, che serve proprio per creare un array a partire da una stringa: lo split crea i tre elementi individuando, in questo caso, i trattini della data (“-”), che fungono da separatori. Il passo successivo consiste nell'assegnare alla variabile “\$mese” il corretto elemento dell'array “\$mesi”, ossia il nome del mese in italiano: per farlo è necessario decrementare l'elemento dell'array di una unità, in quanto il conteggio parte da zero (quindi nell'array gennaio non è 1 ma è 0, febbraio è 1 invece di 2, marzo è 2 e non 3, e così via). Fatto questo, viene impostato il nuovo valore della variabile “\$data”, con il giorno, il mese e l'anno scritti correttamente.

5. Una proposta metodologica: fasi preliminari

L'ultimo passaggio è la stampa della data in formato elegante tramite il comando “echo”.

Se la data è uguale a “0000-00-00”, il codice non stampa alcunché.

Altri tipi di informazioni (la raccolta di riferimento, l'indirizzo del mittente, l'indirizzo del destinatario, i collegamenti alle immagini dei manoscritti originali, le note) sono stati inseriti in “blocchi if”: questi ultimi permettono di eseguire certe operazioni solo se è verificata una certa condizione. In questo caso, il blocco if fa sì che i dati sopraccitati (raccolta, indirizzo del mittente ecc.) siano mostrati all'utente solo nel caso in cui siano presenti. Attraverso l'istruzione

```
if ($row['raccolta'] != "Assente") {  
  echo "<b>Raccolta</b>: $row[raccolta]<br /><br />";  
}
```

si comunica al sistema di scrivere i dati relativi alla raccolta delle lettere solo nel caso in cui il contenuto del campo sia diverso (“!=") dalla stringa (“Assente”). Come si è detto in precedenza, l'operatore, nel file XML, dovrà infatti inserire la dicitura “Assente” nel caso in cui un particolare dato non sia disponibile.

6. Una proposta metodologica: i tipi di ricerca

6.1 Introduzione

La grande originalità della proposta metodologica qui discussa consiste nel fatto che fino a oggi nessuna pubblicazione telematica di carteggi d'artista presenta sistemi di ricerca sofisticati e completi come quelli che qui si vogliono presentare. Il sistema infatti prevede diversi tipi di ricerca:

- ricerca per forma;
- ricerca per lemma;
- ricerca per categoria grammaticale;
- ricerca per cooccorrenze;
- ricerca di stringhe di testo;
- ricerca nei contenuti (destinatari, date, luoghi, indirizzi ecc.).

In questo capitolo si vedrà in che modo sono stati implementati i vari tipi di ricerca.

6.2 La ricerca per forma e per lemma

Attraverso un apposito form, l'utente può eseguire ricerche per forma (per esempio, cercare tutti i contesti in cui appare la forma “arte”) o per lemma (trovare tutte le forme del lemma “arte” e visualizzarne i contesti).

I due tipi di ricerca sono stati implementati in modo analogo, quindi si prenderà come riferimento soltanto la ricerca per forma, evidenziando comunque le differenze rispetto alla ricerca per lemma.

Dopo che l'utente ha compilato il form per la ricerca, si presenta una tabella dei risultati che mostra la forma cercata, il suo lemma, la sua categoria grammaticale, il numero di volte che compare nel corpus e i collegamenti ai contesti, in questo modo:

arte	ARTE	n.c.	4	Vedi contesti...
------	------	------	---	------------------

Cliccando su “Vedi contesti...” sarà possibile accedere alla seconda parte della ricerca: la visualizzazione dei contesti.

6. Una proposta metodologica: i tipi di ricerca

Per quanto riguarda la ricerca per lemma, la tabella dei risultati sarà analoga, ma le colonne per il lemma e per la forma saranno invertite:

ARTE	arte	n.c.	4	Vedi contesti...
ARTE	arti	n.c.	3	Vedi contesti...

Il codice seguente è quello che permette la prima parte della ricerca per forma:

```
$formacercata = $_POST['ricerca'];

echo "Ecco i risultati della tua ricerca:<br /><br />";

$query = "SELECT forma, lemma, categoria, COUNT(*)
AS tot FROM token
WHERE LOWER(forma) LIKE LOWER('$_POST[ricerca]')
GROUP BY (forma)";

$result = mysql_query($query);
$numero = mysql_num_rows($result);

if ($numero == 0) {
echo "Nessun risultato. ";
}

else {
echo "<table border='1'>";

while ($row = mysql_fetch_array($result, MYSQL_ASSOC)) {

echo "<tr>";

echo "<td width='200'>". $row['forma']. "</td>";

echo "<td width='200'>". $row['lemma']. "</td>";

[...]

echo "<td width='30' align='center'>". $row['tot']. "</td>";

echo '<td width="100" align="center">
<a href="ricercaforma1.php?step=2&cerca='.$row['forma'].'">Vedi
contesti...</a></td>';

echo "</tr>";

}

}
```

6. Una proposta metodologica: i tipi di ricerca

```
echo "</table><br /><br /><br />";  
}
```

TAB. 6.1: implementazione della ricerca per forma

All'inizio si dichiara una variabile “\$formacercata”, che corrisponde alla forma cercata dall'utente (immessa attraverso un form). Quindi si crea una query che seleziona dalla tabella “token” forma, lemma, categoria e numero di occorrenze (COUNT(*)): a quest'ultimo viene dato il nome “tot”. Ovviamente i dati selezionati faranno riferimento esclusivamente alla forma cercata dall'utente (WHERE LOWER(forma) LIKE LOWER('\$_POST[ricerca]')). Il comando “LOWER” serve per ignorare le maiuscole: se questo comando non ci fosse, l'utente potrebbe cercare la stringa “Arte” al posto di “arte” e il sistema non restituirebbe alcun risultato nel caso in cui nel corpus non ci siano occorrenze della parola “arte” con la lettera iniziale maiuscola.

Attraverso il metodo “mysql_num_rows” invocato sul risultato della query, si contano i record trovati: nel caso non ce ne fossero, sulla pagina comparirà la scritta “Nessun risultato”, per avvisare l'utente dell'esito negativo della sua ricerca. In caso contrario, si procederà a creare la tabella dei risultati con le apposite etichette HTML.

Nella tabella 6.1 non è stata presentata la parte di codice relativa alla visualizzazione della categoria grammaticale. Come si è visto in precedenza¹⁷⁸, il corpus viene annotato con etichette che potrebbero non essere facilmente comprensibili da parte degli utenti (nel caso di *Carlo Finelli Corpus* è stato usato il tagset EAGLES-ILC). Diventa quindi necessario creare codice per “tradurre” il tagset in notazioni che siano comprensibili da parte degli utenti, come nella seguente porzione di codice:

```
switch ($row['categoria']) {  
  
    case "A": echo "agg.";  
    break;  
    case "B": echo "avv.";  
    break;  
    case "C": echo "cong.";  
    break;
```

178 Cfr. *supra* PAR. 5.4.

6. Una proposta metodologica: i tipi di ricerca

```
[...]  
}
```

TAB. 6.2: switch per le categorie grammaticali

In questo caso è stato realizzato uno “switch”, ovvero un comando che controlla il valore di una certa variabile ed esegue determinate istruzioni a seconda del valore che tale variabile può assumere.

Nel codice qui presentato, lo switch controlla il valore della variabile “\$row['categoria]”, ovvero della categoria grammaticale della forma. Nel caso in cui il valore sia “A” (aggettivo), questo viene tradotto con “agg.”, nel caso in cui il valore sia “B” (avverbio), viene tradotto con “avv.”, e così via.

La seconda parte del codice, realizzata in una pagina a parte, è quella relativa alla presentazione dei contesti della ricerca, visualizzati sotto forma di contesti KWIC¹⁷⁹. Il codice seguente è quello che ne permette la visualizzazione:

```
$formacercata2 = $_GET['cerca'];  
  
echo "<table width='600'>";  
  
$query1 = "SELECT * FROM token WHERE LOWER(forma) LIKE  
LOWER('$_GET[cerca]')";  
  
$result1 = mysql_query($query1);  
while ($row1 = mysql_fetch_array($result1, MYSQL_ASSOC)) {  
  
    $idricerca = $row1['id'];  
    $formascelta = $row1['forma'];  
  
    $query2 = "  
SELECT * FROM  
(SELECT * FROM token WHERE id < '$row1[id]'  
AND idlettera = '$row1[idlettera]'  
ORDER BY id DESC LIMIT 8) AS t  
ORDER BY t.id ASC";  
    $result2 = mysql_query($query2);  
  
    $query3 = "  
SELECT * FROM token  
WHERE id > '$row1[id]' AND idlettera = '$row1[idlettera]'  
ORDER BY id ASC LIMIT 8";
```

179 Cfr. *supra* PAR. 4.5.

6. Una proposta metodologica: i tipi di ricerca

```
$result3 = mysql_query($query3);

echo "<tr>";

echo "<td width='410' align='right'>";

while ($row2 = mysql_fetch_array($result2, MYSQL_ASSOC)) {

    $risultato = "&nbsp;$row2[forma]";

    [...]

    echo $risultato;

}

echo "</td>";
echo "<td width='50'>";
echo "<b>$formascelta</b>";
echo "</td>";
echo "<td width='410'>";

while ($row3 = mysql_fetch_array($result3, MYSQL_ASSOC)) {

    $risultato2 = "&nbsp;$row3[forma]";

    [...]

    echo $risultato2;

}

echo "<td width='30'>";

echo "<a href=lettera.php?id=$row1[idlettera]><img
src='grafica/freccia.gif' alt='Leggi la lettera' style='border:
0' /></a>";

echo "<br /><br />";
echo "</td>";
}
echo "</td>";

}
echo "</tr>";
echo "</table>";
```

TAB. 6.3: codice per la visualizzazione dei contesti KWIC

La prima parte della ricerca, attraverso la tabella di presentazione dei risultati, crea link del tipo `www.nomecorpus.xx/ricercaformal.php?step=2&cerca=arte`. In questo

6. Una proposta metodologica: i tipi di ricerca

modo, nella query string sarà presente la forma della quale visualizzare i contesti (in questo caso, “arte”).

La pagina “ricercaformal.php” imposta una variabile “\$formacercata” inserendo come valore la forma reperita proprio attraverso la query string. Quindi, attraverso la prima query, seleziona tutti i token del corpus corrispondenti a tale forma, memorizzando identificatore e, appunto, forma in due variabili (“\$idricerca” e “\$formascelta”).

A questo punto con due query, denominate “\$query2” e “\$query3”, si cercano rispettivamente gli otto token precedenti e gli otto token successivi. I token precedenti dovranno essere ordinati in modo crescente attraverso il comando `ORDER BY t.id ASC`, dove “t” è il nome dato alla tabella provvisoria dei token precedenti: per rinominare la tabella in modo provvisorio è necessario un *select annidato*, ovvero una sorta di query all'interno di un'altra query. L'ordinamento crescente è necessario perché per poter selezionare i record che precedono la forma cercata sarà necessario fare una ricerca a partire dall'id di quest'ultima verso il basso (`WHERE id < '$row1[id]'`), quindi di default saranno ordinati in modo decrescente: per una corretta visualizzazione non è possibile lasciarli in questo ordine.

Un esempio renderà più chiara questa precisazione. Si supponga di avere il seguente contesto, tratto da una lettera di Carlo Finelli:

“... cere e bassorilievi dei tempi dell'**arte** antica. Nelle ore notturne mi occupa la...”

Nel caso in cui i token che precedono la forma “arte” non vengano ordinati in modo crescente, il risultato della ricerca sarà il seguente:

“... dell' tempi dei bassorilievi e cere **arte** antica. Nelle ore notturne mi occupa la...”

Il comando `AND idlettera = '$row1[idlettera]'` serve per selezionare solo i token appartenenti alla lettera nella quale si trova la forma cercata. Questo serve per evitare, qualora per esempio il token selezionato sia uno degli ultimi della lettera, che vengano visualizzati anche i token appartenenti alla lettera successiva. Il comando `LIMIT 8` serve invece per selezionare soltanto gli otto token immediatamente precedenti e immediatamente successivi.

Per visualizzare i contesti in formato KWIC viene quindi creata una tabella (non visibile da

6. Una proposta metodologica: i tipi di ricerca

parte dell'utente perché presenta bordi azzerati): nella prima colonna saranno inseriti i token precedenti allineati a destra, mentre nella seconda colonna sarà inserita la forma scelta con i token successivi, con allineamento a sinistra. Le variabili “\$risultato” e “\$risultato2” contengono i token precedenti e successivi, ai quali è stata anteposta la sequenza “ ”, ovvero lo spazio. Senza questa specificazione, sullo schermo i token apparirebbero tutti attaccati, in questo modo:

“... cereebassorilievideitempidell'**arte**antica.Nelleorenotturemioccupala...”

Allo stesso modo sono presenti nel codice alcuni comandi per rendere più elegante il testo (per esempio per far comparire i segni di interpunzione attaccati alle parole ma seguiti da uno spazio), non illustrati nella tabella 6.2 in quanto ripetitivi. Un esempio di tali comandi è il seguente:

```
$risultato=str_replace("&nbsp;"," ", "$risultato");
```

Con questo comando, si ordina di sostituire la sequenza spazio-virgola con la sola virgola. Comandi simili sono stati preparati per tutti gli altri segni di interpunzione.

I contesti invece saranno quindi visualizzati in questo modo:

mio piacere di restare per il vantaggio dell' **arte** non per altra cosa, ma non posso
nella composizione e tutto ciò che riguarda l' **arte**. Ella è veramente sufficiente se mi
crederà

cere e bassorilievi dei tempi più floridi dell' **arte** antica. Nelle ore notturne mi occupa la
più sode e costume più semplice alla scultura **arte** più delle altre severa; più mollezza e

Ma il sistema in realtà produce una tabella di questo tipo:

mio piacere di restare per il vantaggio dell'	arte non per altra cosa, ma non posso
nella composizione e tutto ciò che riguarda l'	arte . Ella è veramente sufficiente se mi crederà
cere e bassorilievi dei tempi più floridi dell'	arte antica. Nelle ore notturne mi occupa la
più sode e costume più semplice alla scultura	arte più delle altre severa; più mollezza e

6. Una proposta metodologica: i tipi di ricerca

Vicino a ogni contesto sarà presente un'icona, cliccando sulla quale si sarà rimandati alla lettera nella quale è possibile trovare la frase visualizzata.

Giova specificare che l'utente può cercare non soltanto forme e lemmi precisi: attraverso l'operatore “%” potrà cercare anche sequenze di caratteri. Per esempio, immettendo nel campo di ricerca per forma la stringa “mar%”, il sistema cercherà tutte le forme che iniziano con la sequenza “mar” (“mare”, “marmo”, “marmi” nel caso di *Carlo Finelli Corpus*), mentre immettendo la stringa “%dare” saranno cercate tutte le forme che terminano con la sequenza “dare” (“andare”, “mandare”, “raccomandare”, “ritardare”). Infine, immettendo la stringa “%ive%” saranno cercate tutte le forme che contengono la sequenza “ive” (“arriverà”, “diverse”, “diversi”, “diversità”, “divertimenti” ecc.).

6.3 La ricerca per categoria grammaticale

La ricerca per categoria grammaticale è del tutto simile alla ricerca per forma e per lemma, con una sola eccezione: non è una ricerca libera ma è una ricerca svolta attraverso un *select*. Un *select* è un campo di un form in cui le opzioni sono predefinite: l'utente quindi ha una serie di voci tra le quali scegliere e non può inserire una stringa di testo. Il *select* appare all'utente come un cosiddetto “menù a tendina”. Nella tabella seguente viene illustrato il codice del *select* che permette la ricerca per categoria grammaticale:

```
<select size='1' name='ricgramm' id='ricgramm'>
  <option selected='selected' value='A'>Aggettivo</option>
  <option value='B'>Avverbio</option>
  <option value='C'>Congiunzione</option>
  <option value='DD'>Aggettivo dimostrativo</option>
  <option value='DE'>Aggettivo esclamativo</option>
  <option value='DI'>Aggettivo indefinito</option>
  <option value='DT'>Aggettivo interrogativo</option>
  <option value='E'>Preposizione</option>

  [...]

</select>
```

TAB. 6.4: il *select* per le categorie grammaticali

Il sistema si comporta poi allo stesso modo di quello che effettua la ricerca per forma e per lemma: una volta selezionata la voce desiderata (per esempio, un aggettivo), il sistema

6. Una proposta metodologica: i tipi di ricerca

restituisce tutti gli aggettivi presenti nel corpus, ognuno con il proprio lemma, il numero di volte in cui compare e il collegamento ai contesti KWIC. Facendo clic su tale collegamento si viene rinviiati a una pagina che visualizza i contesti KWIC nello stesso modo illustrato nel paragrafo precedente.

6.4 La ricerca per cooccorrenze

La ricerca per cooccorrenze è un potente strumento che permette di cercare contesti all'interno dei quali compaiano due occorrenze cercate dall'utente, in un intervallo di occorrenze stabilito dall'utente stesso. Infatti, per *cooccorrenze* si intendono le occorrenze che ricorrono insieme ad altre occorrenze. Un esempio: l'utente può cercare quante volte le forme “arte” e “scultura” appaiano a distanza ravvicinata in un intervallo di dieci occorrenze. Se invece l'utente imposta l'intervallo a due, attraverso la ricerca per cooccorrenze è possibile effettuare anche la ricerca di *bigrammi*: si intende con quest'ultimo termine una coppia di occorrenze contigue presenti in un testo.

La pagina si presenta all'utente con cinque campi: il primo serve per selezionare l'intervallo, mentre i successivi quattro servono per la ricerca delle cooccorrenze. I quattro campi sono suddivisi in due per la prima occorrenza e due per la seconda: di questi due, uno è per cercare una forma e l'altro è per cercare un lemma.

L'utente quindi a sua disposizione ha diverse combinazioni:

- cooccorrenza forma-forma;
- cooccorrenza forma-lemma (verranno cercate tutte le forme di un lemma scelto);
- cooccorrenza lemma-lemma;
- cooccorrenza di due forme individuate in base a sequenze di caratteri (utilizzando l'operatore “%”);
- cooccorrenza forma-forma/lemma (utilizzando tre campi per la ricerca invece di due);
- cooccorrenza forma/lemma-forma/lemma (utilizzando tutti i quattro campi);

E così via. Si consideri che per ognuna delle combinazioni sopra elencate (e per le altre non citate) è possibile selezionare in modo del tutto libero l'intervallo di occorrenze all'interno del quale svolgere la ricerca di cooccorrenze, e inoltre è possibile individuare forme e lemmi in base a sequenze di caratteri (per esempio, digitando “ar%” si troveranno tutte le forme o i lemmi che iniziano con la sequenza “ar”): si intuisce quindi la grandissima versatilità di

6. Una proposta metodologica: i tipi di ricerca

questo importante strumento.

La ricerca per cooccorrenze è simile a quella per forma e per lemma, ma ovviamente il controllo avviene attraverso l'utilizzo di due query, una per la prima occorrenza e una per la seconda:

```
$periodo = $_POST['periodo'];

$formacercata1 = $_POST['ricerca'];
$formacercata2 = $_POST['ricerca3'];
$lemmacercato1 = $_POST['ricerca2'];
$lemmacercato2 = $_POST['ricerca4'];

echo "<table>";

$query1 = "SELECT * FROM token WHERE
LOWER(forma) LIKE LOWER('$formacercata1') OR
LOWER(lemma) LIKE LOWER('$lemmacercato1')";

$result1 = mysql_query($query1);

while ($row1 = mysql_fetch_array($result1, MYSQL_ASSOC)) {

$query2 = " SELECT * FROM (
(SELECT * FROM token WHERE id >= '$row1[id]'
ORDER BY id ASC
LIMIT $periodo) tabella WHERE
LOWER(forma) LIKE LOWER('$formacercata2') OR
LOWER(lemma) LIKE LOWER('$lemmacercato2')
ORDER BY id ASC";

$result2 = mysql_query($query2);
```

TAB. 6.5: ricerca delle cooccorrenze nella base di dati

Il codice inizialmente non fa altro che impostare cinque variabili, una per ognuno dei campi del form di ricerca (il periodo, la prima forma, il primo lemma, la seconda forma, il secondo lemma). A questo punto, attraverso la prima query vengono cercati, all'interno della tabella “token”, tutti i record che abbiano la forma o il lemma uguali a quelli immessi dall'utente nel form. Viene così creato un array per i record tramite il solito metodo “mysql_fetch_array”: senza tale creazione non sarebbe possibile cercare la seconda occorrenza, in quanto il controllo (attraverso l'operatore “WHERE”) viene svolto sull'id della prima occorrenza. Inoltre, l'inserimento della seconda query nel ciclo while fa sì che il controllo possa essere ripetuto per ognuna delle occorrenze trovate con la prima query (se così

6. Una proposta metodologica: i tipi di ricerca

non fosse, verrebbe controllata soltanto la prima occorrenza).

La seconda query presenta un select annidato: la query interna (`SELECT * FROM token WHERE id >= '$row1[id]' ORDER BY id ASC LIMIT $periodo`) cerca all'interno della tabella "token" tutti i record successivi alla prima occorrenza nell'intervallo scelto dall'utente (`LIMIT $periodo`). Per esempio, se la prima occorrenza è "arte" e l'intervallo scelto è di otto occorrenze, con la prima query il sistema cerca le otto occorrenze successive alla forma "arte".

A questa query viene dato il nome provvisorio di "tabella". La seconda query, quella esterna, non fa altro che cercare, all'interno della prima query, tutti i record la cui forma o il cui lemma corrispondano a quelli cercati dall'utente. Una volta fatto questo, i risultati vengono ordinati in modo crescente.

Eseguita questa operazione (e, quindi, trovati tutti i record necessari), il sistema fa partire altre tre query, che serviranno per permettere la visualizzazione dei contesti KWIC. Il meccanismo di queste query è simile a quello visto nella tabella 6.3:

```
$query3 = " SELECT * FROM
(SELECT * FROM token
WHERE id < '$row1[id]'
AND idlettera = '$row1[idlettera]'
ORDER BY id DESC
LIMIT 5) AS t
ORDER BY t.id ASC";
$result3 = mysql_query($query3);

$query4 = "SELECT * FROM token
WHERE id > '$row2[id]'
AND idlettera = '$row1[idlettera]'
ORDER BY id ASC
LIMIT 5";
$result4 = mysql_query($query4);

$query5 = "SELECT *FROM token
WHERE id > '$row1[id]' AND id < '$row2[id]'";
$result5 = mysql_query($query5);
```

TAB. 6.6: query per i contesti KWIC nella ricerca di cooccorrenze

Come nella ricerca per forma e per lemma, la prima di queste query cerca i record che precedono la prima occorrenza, e il select annidato è necessario per poterli poi ordinare in

6. Una proposta metodologica: i tipi di ricerca

modo crescente; la seconda query cerca i record che invece succedono la seconda occorrenza, e la terza query cerca tutti i record compresi tra la prima e la seconda occorrenza. I risultati vengono poi presentati all'utente in un modo simile a quello visto per la ricerca per forma e per lemma.

6.5 La ricerca di stringhe di testo

L'esperienza acquisita attraverso lo studio e l'analisi di progetti già esistenti in materia di carteggi d'artista¹⁸⁰ ha suggerito di implementare un sistema di ricerca di stringhe nei testi delle lettere, con la possibilità di vedere il risultato della propria ricerca evidenziato all'interno del testo stesso: quest'ultima è una caratteristica molto importante per un motore di ricerca ma, come si è visto, spesso non viene implementata.

La ricerca di stringhe all'interno dei testi delle lettere è stata realizzata in una sola pagina, e il codice che permette la ricerca è il seguente:

```
session_start();
$stringacercata = $_POST['ricerca'];
$_SESSION['stringa'] = $stringacercata;

echo "Ecco i risultati della tua ricerca:<br /><br />";

$query = "SELECT * FROM lettere
WHERE LOWER(testo) LIKE LOWER('%$stringacercata%')";

$result = mysql_query($query);
$numero = mysql_num_rows($result);

if ($numero == 0) {
echo "Nessun risultato. ";
}

else {

$i = 1;
while ($row = mysql_fetch_array($result, MYSQL_ASSOC)) {

echo $i;
echo ".&nbsp;";
echo "Data della lettera: ";
```

180 Cfr. *supra* CAP. 4.

6. Una proposta metodologica: i tipi di ricerca

```
echo"  
<a href='lettera.php?id=$row[id] '$row[luogo]";  
  
formatodata();  
  
echo "</a>";  
  
echo "<br /><br />";  
$i++;  
  
}  
  
}
```

TAB. 6.7: la ricerca di stringhe nel testo.

Il primo passo da compiere è aprire una *sessione* attraverso il metodo “`session_start`”: una sessione è semplicemente un mezzo per memorizzare dati, e servirà per poter evidenziare i risultati della ricerca all'interno del testo. Successivamente viene dichiarata una variabile “`$stringacercata`” alla quale viene assegnata la stringa cercata dall'utente attraverso l'apposito form, quindi tale stringa viene memorizzata all'interno della sessione attraverso l'utilizzo della variabile *superglobale* `$_SESSION`: una variabile superglobale è una particolare variabile che è disponibile per qualsiasi funzione senza che ci sia la necessità di dichiararla.

Eseguite queste operazioni preliminari, si dà inizio alla query che cerca all'interno della tabella “lettere” tutte le missive che, all'interno del testo, rechino la stringa cercata dall'utente. Nella query, la variabile “`$stringacercata`” è preceduta e seguita da due operatori percentuali per permettere anche ricerche non precise: quindi se l'utente cerca la sequenza “art”, il risultato sarà costituito da tutti i testi che contengono la sequenza “art”. Se non fossero presenti gli operatori percentuali, il risultato sarebbe costituito da testi nei quali è presente la forma precisa “art”.

La query viene inviata alla base di dati tramite la solita funzione “`mysql_query`”. Viene creata quindi una variabile, denominata “`$numero`”, che conta semplicemente il numero dei record trovati con la ricerca. Questa variabile andrà poi a costituire la condizione del successivo blocco if: se infatti “`$numero`” è uguale a zero, sullo schermo all'utente apparirà la scritta “Nessun risultato”.

Altrimenti, viene creato l'array per la visualizzazione dei record tramite il consueto metodo “`mysql_fetch_array`”, e successivamente viene dichiarata una variabile “`$i`”, impostata a 1: si tratta semplicemente di un contatore per numerare i risultati della ricerca. Ogni volta che il

6. Una proposta metodologica: i tipi di ricerca

ciclo while per la visualizzazione dei record viene ripetuto, il contatore viene incrementato ($\$i++$), in modo tale che i collegamenti ai testi che contengono la stringa cercata possano essere numerati in modo progressivo.

I collegamenti vengono presentati con il numero di cui si è appena parlato e con la data di invio della lettera, formattata in modo elegante tramite la funzione “formatodata”: cliccando sul collegamento si sarà rimandati alla pagina della lettera¹⁸¹, all'interno della quale la stringa cercata sarà evidenziata.

Come è possibile evidenziare la stringa cercata in un'altra pagina? Il codice che permette tale operazione è il seguente:

```
if (isset($_SESSION['stringa'])) {  
    $stringacercata = $_SESSION['stringa'];  
    $testo= $row['testo'];  
    $pattern = preg_quote($stringacercata);  
    $testo =  
    preg_replace("|($pattern)|i",  
    "<font color='#ffffff'><b>\1</b></font>",  
    $testo);  
    echo $testo;  
    echo "<br /><br />";  
    session_destroy();  
}  
else {  
    echo "$row[testo]";  
    echo "<br /><br />";  
}
```

TAB. 6.8: evidenziazione dei risultati della ricerca nella pagina della lettera.

Questo codice illustra anche il modo attraverso il quale nella pagina della lettera viene rappresentato il testo della missiva selezionata.

¹⁸¹ Cfr. *supra* PAR. 5.8.

6. Una proposta metodologica: i tipi di ricerca

Innanzitutto, un blocco if controlla che sia stata impostata la sessione “stringa” (e, qualora sia stata svolta la ricerca per stringa, la condizione è soddisfatta in quanto la sessione dura per tutta la navigazione nel sito finché non viene interrotta da un'apposita funzione). Quindi, viene dichiarata una variabile “\$stringacercata”, il cui valore sarà quello della stringa memorizzata nella sessione (la stringa cercata dall'utente).

Viene poi dichiarata una variabile per il testo e una variabile “\$pattern”, il cui valore non sarà altro che la stringa cercata dall'utente a cui saranno aggiunte le sequenze di escape (tramite la funzione “preg_quote”) per evitare problemi di riconoscimento di caratteri.

A questo punto avviene la messa in rilievo della stringa cercata, che si effettua con una sostituzione del testo: la funzione “preg_replace”, tramite l'uso delle espressioni regolari di PHP, cerca all'interno della variabile “\$testo” la variabile “\$pattern” (la “i” che segue la dichiarazione all'interno di “preg_replace” è un operatore delle espressioni regolari e serve per indicare la *case-insensitiveness*, ovvero si ordina alla funzione di trascurare le differenze tra lettere maiuscole e lettere minuscole) e la sostituisce sempre con la stessa variabile, ma opportunamente evidenziata (nel codice in tabella 6.8, la variabile è stata messa in grassetto e colorata di bianco).

A questo punto la sessione viene annullata tramite il metodo “session_destroy”, in modo tale che il sistema possa essere pronto per una nuova ricerca.

Qualora la condizione del blocco if non sia soddisfatta, la pagina “lettera.php” non farà altro che presentare all'utente il testo della lettera selezionata senza alcuna evidenziazione.

6.6 La ricerca nei contenuti (destinatario, luogo, data ecc.)

L'ultimo tipo di ricerca è quello all'interno delle informazioni aggiuntive, come destinatario, luogo, data, raccolta, luogo di conservazione e così via. Nella pagina per questi tipi di ricerca si prepareranno tanti form per quanti sono i tipi di ricerca, e ognuno rimanderà alla pagina “ricercacont.php” alla quale sarà aggiunta una query string chiamata “step” e numerata in modo progressivo. Per esempio, la ricerca per destinatario condurrà a “ricercacont.php?step=1”, la ricerca per luogo condurrà a “ricercacont.php?step=2” e così via.

Si tratta di ricerche del tutto simili tra di loro, pertanto si presenta di seguito solo il codice relativo a una di esse, nella fattispecie la ricerca in base al destinatario:

6. Una proposta metodologica: i tipi di ricerca

```
if ($_GET['step'] == 1) {

    echo "Ecco i risultati della tua ricerca:<br /><br />";

    $query = "SELECT * FROM lettere WHERE LOWER(destinatario)
    LIKE LOWER('%$_POST[ricdes]%)";

    $result = mysql_query($query);
    $numero = mysql_num_rows($result);

    if ($numero == 0) {
        echo "Nessun risultato. ";
    }

    else {

        $i = 1;
        while ($row = mysql_fetch_array($result, MYSQL_ASSOC)) {

            echo $i;
            echo ".&nbsp;";
            echo "Data della lettera: ";

            echo "<a href='lettera.php?id=$row[id] '$>$row[luogo]";
            formatodata();

            echo "</a>";

            echo "<br /><br />";

            $i++;

        }

    }

}
```

TAB. 6.9: ricerche nei contenuti.

Come si può notare, il codice è all'incirca lo stesso della tabella 6.6: la query seleziona le lettere in base al destinatario scelto, operando un controllo sul campo “destinatario” della tabella “lettere”, quindi viene impostata la variabile “\$numero” per dare il via al blocco if successivo, che in caso di esito negativo restituirà la dicitura “Nessun risultato”, e in caso contrario presenterà all'utente la lista dei collegamenti alle lettere che soddisfano la condizione cercata. I collegamenti saranno numerati progressivamente tramite l'aiuto del contatore (“\$i”) e rimanderanno alla pagina “lettera.php” che conterrà dati e testo della missiva che soddisferà i risultati della ricerca.

6. Una proposta metodologica: i tipi di ricerca

Leggermente diversa dalle altre è la ricerca per data: in questo caso viene chiesto all'utente di immettere la data iniziale e la data finale, e il sistema cercherà tutte le lettere scritte durante l'intervallo selezionato dall'utente.

7. Una proposta metodologica: la fruizione per il pubblico

7.1 Implementazione del sito web: linee-guida

Creare un sito web non è difficile: è invece difficile creare un buon sito web. Questa sezione cercherà di dare delle direttive per fare in modo che tutto il lavoro preliminare svolto (l'acquisizione delle lettere, la trascrizione, la costruzione del corpus, la sua annotazione ecc.) non sia vanificato con la realizzazione di un sito poco professionale, poco usabile, che non tenga conto delle esigenze di tutti i tipi di utenti (e non soltanto di alcuni).

Quali sono i principî che dovrebbero guidare la realizzazione di un sito web di qualità? Sono stati scritti molti libri a tal proposito, e in estrema sintesi le linee-guida per la realizzazione di un buon sito web sono riassumibili, in generale, nel modo seguente:

1. *Definire gli obiettivi del proprio sito*: una delle prime cose da fare quando si realizza un sito web è avere ben chiari quali sono i propri obiettivi. Un sito deve sempre essere creato in vista di un obiettivo ben preciso¹⁸², e cercare di far capire quali sono i propri obiettivi è una prassi utile sia per i realizzatori che per gli utenti. Per i realizzatori, in quanto obiettivi chiari permettono di progettare un sito in modo più razionale e anche in modo più veloce. Per gli utenti, in quanto la definizione degli obiettivi dimostra che dietro alla realizzazione di un sito c'è un progetto serio, credibile e unico. Priebe individua cinque obiettivi che ogni sito dovrebbe perseguire¹⁸³:

I. Fornire informazione: ci sono moltissimi siti web che forniscono informazioni scarse o addirittura nulle. I visitatori di un sito devono conoscere per quali ragioni il sito è stato realizzato, cosa è possibile trovare all'interno del sito stesso, in che modo è stato sviluppato;

II. Guadagnare credibilità: l'obiettivo è strettamente legato al primo, e un sito chiaro, che fornisce buone informazioni, sarà anche un sito credibile per il pubblico;

III. Garantire l'approccio da parte degli utenti: il pubblico del sito deve essere in grado di fornire feedback ai gestori del sito e di ricevere risposte, e per garantire questo basterà soltanto fornire un semplice indirizzo di contatto;

IV. Creare un beneficio per gli utenti: deve esserci una precisa ragione per il quale il pubblico dovrebbe visitare il sito, quindi gli utenti dovrebbero aspettarsi dal sito un prodotto

182 Tim Priebe, *Webifiable – Plan, create and maintain your web site the right way* (Edmond, Oklahoma: T&S Web Design, 2007), 8 – 13.

183 *Ibid.*, 10.

7. Una proposta metodologica: la fruizione per il pubblico

unico e originale, di qualità. Obiettivo di un buon sito dovrebbe quindi essere il “dare qualcosa in più” ai propri utenti;

V. Creare un beneficio per se stessi: maggiore è l'appetibilità di un sito per gli utenti, maggiore sarà anche il ritorno per i realizzatori.

Risulta quindi fondamentale l'importanza di avere chiari gli obiettivi ancor prima di progettare il proprio sito.

2. *Redigere con estrema cura il contenuto del sito*: si tratta forse dell'aspetto più importante della progettazione dei siti web perché, in fin dei conti, gli utenti visitano un sito web per il suo contenuto¹⁸⁴. Gli studi di usabilità sostengono che ciò che spinge gli utenti è la ricerca dei contenuti, e quando essi arrivano su una nuova pagina, il loro sguardo volge immediatamente nella parte di schermo dedicata ai contenuti: la pagina sarà scorsa alla ricerca di titoli e indicazioni relative all'argomento della pagina¹⁸⁵.

La redazione dei testi deve quindi avvenire con cura molto particolare, perché il pubblico della rete è molto diverso dal pubblico del “cartaceo”: in rete la navigazione avviene quasi per colpo d'occhio, “l'utente del web non *legge* le pagine, ma le *scorre*, un po' come se cercasse informazioni su una carta geografica”¹⁸⁶. Se l'utente non trova subito l'informazione che cerca, è probabile che passerà ad un'altra pagina. Quindi i testi devono essere organizzati con chiarezza, rigore e sinteticità, i menù devono essere presentati in modo intuitivo, e in generale i contenuti devono far sì che l'utente possa arrivare “al punto” nel modo più rapido possibile. È assolutamente necessario inoltre redigere i testi avendo ben presente il pubblico al quale si rivolgono, in modo da scegliere la forma e il linguaggio appropriati. In generale però è bene fornire informazioni complete utilizzando il minor numero di parole possibile, evitare periodi lunghi e complessi, usare un linguaggio sobrio, chiaro e diretto, centrare i punti salienti della discussione, esprimere i concetti in modo rapido e conciso.

Di fondamentale importanza risulta fornire al pubblico informazioni sulle persone, gli enti, gli istituti e in generale su chiunque sia stato coinvolto da vicino nella realizzazione del sito. Questo perché al giorno d'oggi è molto facile pubblicare in rete un sito, quindi gli utenti non hanno ben chiaro che peso dare alle informazioni reperite in rete¹⁸⁷. Nel progetto *Carlo Finelli*

184 Jakob Nielsen, *Designing Web Usability* (Basingstoke: Macmillan Computer Publishing, 2000), trad. it. *Web usability*, Apogeo, 2000, traduzione di Walter Vanini, 99.

185 *Ibid.*, 100.

186 Roberto Polillo, *Plasmare il web: road map per siti di qualità* (Milano: Apogeo, 2006), 177.

187 Jakob Nielsen, *Designing Web Usability*, 92.

7. Una proposta metodologica: la fruizione per il pubblico

Corpus è stato riservato un apposito spazio ai loghi delle aziende e degli istituti universitari che sono stati coinvolti nella realizzazione, e ogni logo è stato dotato di un collegamento facendo clic sul quale l'utente può accedere ai rispettivi siti.

L'inserimento di informazioni sulle persone coinvolte è fonte di credibilità e attendibilità per il proprio sito ed è strettamente legato alla definizione dei contenuti, che devono sempre occupare un ruolo centrale.

La proposta metodologica discussa in questa sede è stata ideata in modo tale che chi in futuro vorrà utilizzarla per una pubblicazione telematica di un carteggio d'artista, potrà avere uno standard di riferimento che tiene conto dei contenuti e fa sì che a questi ultimi vengano dati il giusto risalto e la giusta importanza. Si tratta infatti di una proposta metodologica pensata anche in vista della valorizzazione dei contenuti: la facilità d'uso, la rapidità della navigazione, i diversi tipi di ricerca e tutte le altre caratteristiche fin qui viste sono state implementate per conferire il massimo valore ai contenuti.

3. *Modellare l'espressione grafica del sito in modo chiaro ed elegante*: il sito web è uno strumento che offre servizi ma è anche uno strumento di comunicazione¹⁸⁸, e in quanto tale è necessario che sia dotato di una grafica attraente, che possa far presa sull'utente. Questo però non significa che il sito debba necessariamente avere una grafica complicata, e soprattutto in fase di progettazione non si deve assolutamente conferire più importanza alla grafica che ai contenuti: non si deve dimenticare che gli utenti navigano su internet soprattutto per cercare informazioni, e che quindi la grafica non deve essere “un surrogato dei contenuti”¹⁸⁹. Grafica e contenuti devono essere supplementari: una grafica molto artistica non supplisce a contenuti scadenti¹⁹⁰ (e spesso invece si utilizzano interfacce grafiche molto elaborate per mascherare la povertà dei contenuti), e viceversa contenuti di alta qualità presentati in modo poco chiaro e poco elegante non saranno attraenti per gli utenti.

È pertanto necessario realizzare un'interfaccia che possa valorizzare i contenuti, che sia raffinata ed elegante ma al contempo semplice, che sia bilanciata e ottimizzata, che garantisca piena leggibilità dei contenuti e che infine sia gradevole dal punto di vista estetico.

La grafica non deve essere pesante, non deve sovrastare i contenuti e deve essere pensata

188 Roberto Polillo, *Il check-up dei siti web* (Milano: Apogeo, 2004), 61 ss.

189 Mario Grasso, *Scrivere per il web. Annotazioni, considerazioni e suggestioni* (Milano: Franco Angeli, 2005), 39.

190 Lorenzo Cantoni, Nicoletta Di Blas e Davide Bolchini, *Comunicazione, qualità, usabilità* (Milano: Apogeo, 2003), 96.

7. Una proposta metodologica: la fruizione per il pubblico

in modo che nessun tipo di utente risulti penalizzato dall'interfaccia.

La proposta metodologica che si vuole presentare con questa tesi consente di modellare una grafica flessibile, perché il codice PHP è adatto a qualsiasi tipo di realizzazione: attorno ai codici presentati nei due capitoli precedenti è possibile sviluppare qualsiasi tipo di grafica, ma le caratteristiche della proposta qui discussa sono volte a portare i webdesigner alla realizzazione di una grafica che possa valorizzare il sito. Questo perché, come si è detto al punto 3 del presente elenco, la proposta metodologica è stata pensata in vista della valorizzazione dei contenuti, e quindi chi si troverà a progettare l'interfaccia grafica sarà condotto ad attribuire il giusto peso ai contenuti.

4. *Creare pagine formattate in modo leggero e coerente e ottimizzare le immagini*: le dimensioni e la complessità del markup di un sito determinano in gran parte la loro velocità di caricamento iniziale¹⁹¹. Pagine che presentano un codice pesante e complesso, soprattutto quelle strutturate con tabelle e che presentano eccessi di codice JavaScript, possono ritardare il caricamento delle pagine. Le tabelle infatti, oltre a essere antiestetiche e a rappresentare una scelta molto discutibile per la realizzazione del layout del sito, richiedono il doppio del tempo per essere caricate rispetto ai blocchi di contenuti: questo anche per il fatto che una tabella ha bisogno di molto più codice per essere realizzata. Allo stesso modo, il caricamento può essere ritardato da immagini troppo pesanti o da eccessi di elementi Flash.

Come è possibile quindi ridurre i tempi di caricamento della pagina agendo sul codice? Lo si può fare con un uso intelligente dei fogli di stile CSS (*Cascading Style Sheets*, “fogli di stile a cascata”): questi ultimi rappresentano uno standard per specificare le caratteristiche del codice HTML inserito nel sito. Il concetto di CSS fu proposto per la prima volta nel 1994 dal norvegese (allora ventinovenne) Håkon Wium Lie, che fu anche il responsabile del loro sviluppo presso il W3C. Attraverso i CSS è possibile specificare caratteristiche universali per le etichette del codice (ovvero, è possibile specificare in un'unica pagina CSS caratteristiche che altrimenti dovrebbero essere dichiarate ogni volta in ogni pagina del sito). Per esempio, si supponga di voler creare un sito con lo sfondo nero: senza il foglio di stile è necessario specificare in ogni pagina il colore dello sfondo, e qualora lo si voglia cambiare, per esempio facendolo diventare bianco, si dovranno aggiornare tutte le pagine del sito: un lavoro particolarmente fastidioso se la quantità di pagine è elevata. Viceversa, con i CSS è sufficiente

¹⁹¹ Andrew B. King, *Website optimization* (Sebastopol, California: O'Reilly Media, 2008), 155 ss.

7. Una proposta metodologica: la fruizione per il pubblico

specificare in un unico foglio di stile il colore dello sfondo, e in caso di aggiornamento si dovranno apportare modifiche soltanto a quell'unica pagina che ospita il foglio di stile.

Grazie ai CSS quindi è possibile rendere i documenti HTML indipendenti dal dispositivo usato per la loro visualizzazione¹⁹²: i CSS infatti permettono di separare il contenuto dalla sua presentazione, perché le etichette per la presentazione possono essere definite in un'unica pagina CSS, quindi il file HTML non dovrà contenere etichette con specifiche relative alla presentazione o alla grafica.

Un sito web organizzato in modo razionale non può prescindere dall'uso dei fogli di stile CSS, e i vantaggi derivanti dal loro utilizzo sono enormi: si separa il contenuto dalla presentazione, si risparmia tempo in fase di aggiornamento, si creano pagine HTML leggere e caricabili molto più velocemente, e in definitiva si realizza un sito armonico, coerente e ben strutturato.

Per quanto riguarda le immagini, quelle che servono per la grafica dovranno essere dichiarate all'interno del foglio di stile, mentre quelle che fanno parte dei contenuti dovranno essere scelte in maniera appropriata e coerente e soprattutto non dovranno appesantire troppo il caricamento della pagina. In che modo possono essere ottimizzate le immagini per il web? Ci sono vari espedienti per fare in modo che le immagini non appesantiscano le pagine¹⁹³, e i seguenti sono sicuramente i più importanti:

- inserire immagini delle esatte dimensioni che si vogliono utilizzare (molti webdesigner infatti inseriscono immagini di dimensioni spropositate e le rimpiccioliscono attraverso l'uso delle proprietà “height” e “width” dell'etichetta “”: questo però non influisce minimamente sul peso dell'immagine, che rimarrà inalterato);

- utilizzare i bordi dei CSS invece di creare bordi interni alle immagini;

- utilizzare i formati appropriati per le immagini. I tre formati universalmente supportati da tutti i browser sono GIF, JPEG e PNG: se l'immagine ha molti gradienti in diversi colori, è bene usare JPEG, mentre se l'immagine è una figura geometrica o ha diverse aree di un unico colore è preferibile l'uso dei formati GIF o PNG, e infine se l'immagine deve avere una certa forma, anche irregolare (e quindi deve avere sezioni trasparenti per adattarsi al sito), bisogna usare GIF e PNG¹⁹⁴.

192 Cfr. Håkon Wium Lie e Bert Bos, *Cascading Style Sheets: Designing for the Web – Third Edition* (Upper Saddle River: Addison-Wesley, 2005).

193 Cfr. Andrew B. King, *Website optimization*.

194 Christian Heilmann e Mark Norman Francis, *Web development solutions* (Berkeley, California: Friends of ED, 2007), 23 – 25.

7. Una proposta metodologica: la fruizione per il pubblico

5. *Usabilità e accessibilità*: si tratta di due aspetti fondamentali della realizzazione di un sito web, tanto importanti che in questa sede si è deciso di dedicare a usabilità e accessibilità due paragrafi a sé stanti¹⁹⁵. Un ottimo sito web infatti deve pensare a tutte le tipologie di utenti e in fase di progettazione è necessario comprendere che un sito web è uno strumento di comunicazione, quindi è rivolto a un pubblico, un pubblico spesso eterogeneo: per creare un prodotto che possa definirsi di qualità bisogna tener conto delle singole esigenze di ogni diversa tipologia di utente.

7.2 Il sito del corpus: contenuti aggiuntivi

Per far sì che il sito del corpus possa avere una maggior completezza, lo si può dotare di contenuti aggiuntivi per fornire informazioni in più agli utenti. Un sito che presentasse infatti soltanto il corpus del carteggio di artista, senza ulteriori contenuti, non risulterebbe uno strumento utile per tutti: se l'intento di una pubblicazione telematica dei carteggi d'artista è anche quella di avvicinare un certo tipo di pubblico all'opera di un artista, sarà necessario inserire informazioni aggiuntive sull'artista stesso. La seguente tabella mostra i contenuti aggiuntivi offerti dai siti analizzati nel capitolo 4 (a eccezione di *Google Books*):

Progetto	Contenuti aggiuntivi
FO.S.C.A.	<ul style="list-style-type: none">- Pagina delle news- Archivio fotografico (solo per utenti registrati)- Link ad altri progetti dell'Università di Genova
<i>Documents from the Archivio di Stato di Roma</i>	<ul style="list-style-type: none">- Immagini di opere degli artisti dell'Accademia di San Luca- Storia dell'Accademia di San Luca- Bibliografia sull'Accademia di San Luca
Memofonte	<ul style="list-style-type: none">- Pagina delle news- Informazioni sui progetti didattici- Rivista online "Studi di Memofonte", in formato pdf
<i>Corpus Informatico Belloriano</i>	Nessuno
<i>Medici Archive Project</i>	<ul style="list-style-type: none">- Pagina delle news- Rivista online "Il Gazzettino", in formato pdf- Fotografie e brevi biografie dello staff- Informazioni su mostre in cui la fondazione è implicata- Collegamenti a siti esterni

195 Cfr. *infra* PAR. 7.3, 7.4.

7. Una proposta metodologica: la fruizione per il pubblico

<i>Vincent Van Gogh – The Letters</i>	<ul style="list-style-type: none">- Saggio “Van Gogh come scrittore di lettere” (<i>Van Gogh as a letter-writer</i>)- Biografie dei corrispondenti- Storia delle pubblicazioni delle lettere- Cronologia- Bibliografia
---------------------------------------	--

TAB. 7.1: contenuti aggiuntivi dei siti analizzati nel CAP. 4.

Di seguito si propone un elenco di possibili contenuti aggiuntivi con i quali arricchire il sito che ospiterà il carteggio d'artista:

– *Biografia e informazioni sull'opera di un artista*: inserire notizie sulla vita e sull'opera dell'artista o degli artisti menzionati nel corpus è importante perché può rendere il sito del corpus anche un potente strumento di avvicinamento all'opera di un artista anche da parte di un pubblico non necessariamente esperto. Sarà poi compito dei realizzatori del corpus decidere se inserire notizie approfondite oppure brevi schede: la redazione dei testi deve essere pensata, come detto sopra, tenendo conto sia dello strumento che del pubblico;

– *Bibliografia*: una bibliografia è utile per ogni tipo di utente: tramite la bibliografia si suggeriscono importanti spunti di approfondimento sia all'utente esperto, che può allargare le proprie conoscenze, sia all'utente che si avvicina a un artista per la prima volta;

– *Saggi e articoli*: oltre ad arricchire il sito e a fornire informazioni aggiuntive, sia per gli esperti che per i non esperti, contribuiscono a fornire un maggiore spessore scientifico alla pubblicazione;

– *Immagini*: dal momento che i carteggi appartengono ad artisti, è auspicabile che nel sito compaiano immagini relative alle opere degli artisti che hanno prodotto i carteggi. Le immagini però non dovranno avere un ruolo preponderante: a esse saranno dedicati spazi opportuni e saranno ottimizzate secondo quanto descritto nel paragrafo precedente;

– *Link*: se lo si ritiene necessario, sarà possibile fornire dei link: se selezionati con cura e in modo razionale possono costituire un'importante fonte di approfondimento per il pubblico del sito;

– *Pagina delle news*: può essere interessante inserire nel sito web una pagina di news (che deve però essere aggiornata in modo costante), al fine di offrire al pubblico informazioni sulle attività degli istituti che si sono occupati della realizzazione del corpus e anche per

7. Una proposta metodologica: la fruizione per il pubblico

dimostrare in modo tangibile che il progetto è seguito, curato e aggiornato;

– *Storia della pubblicazione*: qualora la pubblicazione telematica segua una o più pubblicazioni cartacee dei carteggi, potrebbe essere interessante inserire sul sito la storia delle pubblicazioni cartacee.

In ogni caso, nel momento in cui si progettano i contenuti aggiuntivi, è bene tener presente delle natura *ipertestuale* e *multimediale* della rete.

L'*ipertestualità* fa riferimento al concetto di *ipertesto*: “un testo non sequenziale, ossia un testo in cui l'autore non identifica un unico possibile percorso di lettura, ma offre al lettore delle possibilità alternative, in vari punti specifici del testo”¹⁹⁶. È bene sottolineare che comunque il concetto di ipertesto esisteva già prima di internet, anzi la rete stessa è stata concepita come una sorta di ipertesto globale. Chi progetta il sito del corpus dovrà quindi tenere a mente che scrivere in vista dell'ipertestualità è diverso rispetto a scrivere per il cartaceo: i testi di un sito internet devono essere visti in un'ottica reticolare.

La *multimedialità* è “l'uso combinato di più linguaggi (testi, immagini, suoni) per trasmettere un messaggio”¹⁹⁷. Le opportunità della multimedialità sono enormi, perché tramite il suo adeguato sfruttamento è possibile offrire grande varietà di codici visivi e uditivi, e questa caratteristica è particolarmente stimolante per l'utente. Oltre ai testi quindi sarà possibile inserire immagini, come si è già detto, ma anche elementi video o audio (per esempio, video divulgativi oppure podcast) per presentare un panorama più completo possibile sugli artisti presentati nel sito.

7.3 Usabilità

L'usabilità è il “grado in cui un prodotto può essere usato da particolari utenti per raggiungere certi obiettivi con efficacia, efficienza e soddisfazione in uno specifico contesto d'uso”¹⁹⁸. Per quanto riguarda i siti web, l'usabilità può essere individuata come “una delle

196 Francesca Ricciardi, *Il viaggio dell'innovazione. Informatica, beni culturali e turismo* (Milano: Vita e Pensiero, 2005), 103.

197 Marco Pratellesi, *New journalism: teorie e tecniche del giornalismo multimediale* (Milano: Mondadori, 2004), 65.

198 È la definizione di “usabilità” secondo la norma ISO 9241, emanata dall'ISO (*International Organization for Standardization*, “Organizzazione Internazionale per le Standardizzazioni”, che si occupa di definire standard industriali e commerciali).

7. Una proposta metodologica: la fruizione per il pubblico

proprietà di un sito web che concorrono a definirne la *qualità in uso*¹⁹⁹, che fa riferimento alle già citate efficacia, efficienza e soddisfazione. Con *efficacia* si intende “l'accuratezza e la completezza con cui l'utente raggiunge il risultato voluto”²⁰⁰, mentre l'*efficienza* è la “quantità di risorse spese per ottenere questo risultato”²⁰¹ e infine per *soddisfazione* si intendono “il comfort e l'accettabilità del sistema da parte dell'utente”²⁰². Le tre variabili dell'usabilità possono essere misurate, ognuna con una metrica diversa²⁰³: la percentuale dei compiti portati a termine con successo dagli utenti per l'efficacia, il tempo medio impiegato dagli utenti per effettuare i compiti per quanto riguarda l'efficienza e infine il gradimento medio degli utenti per ciò che concerne la soddisfazione.

L'usabilità può anche essere definita in termini di facilità di apprendimento da parte di utenti principianti assoluti, ovvero quanto è facile per gli utenti svolgere i compiti richiesti dal sito durante il primo utilizzo²⁰⁴: quindi progettare un sito web usabile significa anche progettare per l'utente. Purtroppo però non tutti coloro che progettano siti internet pongono l'utente al centro delle loro attenzioni, quindi è possibile asserire che l'usabilità sia probabilmente “la caratteristica più trascurata nella pratica del Web, perché realizzare sistemi usabili è difficile e impone di prestare un'attenzione particolare all'utente”²⁰⁵.

Ma in che modo è possibile progettare un sito usabile? Si tratta di una domanda alla quale non è facile dare una risposta esauriente, ma è possibile fornire alcune linee guida:

- *coerenza*: il sito deve essere coerente in ogni suo aspetto, sia dal punto di vista della grafica che dal punto di vista dei contenuti, perché la confusione, oltre a essere esteticamente inaccettabile e a creare disorientamento per l'utente, garantisce una sicura perdita di credibilità;
- *estetica gradevole*: se si presenta un sito con un'interfaccia raffinata, elegante e gradevole si otterranno buoni risultati in termini di soddisfazione dell'utente, che sarà più propenso a tornare una seconda volta sul sito;
- *pulizia*: creare un prodotto esteticamente e artisticamente valido non significa abusare della propria creatività creando design troppo arzigogolati e pieni di orpelli inutili. Eleganza

199 Giorgio Brajnik ed Elio Toppano, *Creare siti web multimediali. Fondamenti per l'analisi e la progettazione* (Milano: Pearson Education Italia, 2007), 233.

200 Roberto Polillo, *Il check-up dei siti web*, 214.

201 *Ibid.*

202 *Ibid.*

203 *Ibid.*, 215-217.

204 Giorgio Brajnik ed Elio Toppano, *Creare siti web multimediali*, 234 ss.

205 Roberto Polillo, *Plasmare il web*, 25.

7. Una proposta metodologica: la fruizione per il pubblico

consiste anche nel fornire un prodotto chiaro, che ponga al centro i contenuti e non il design;

– *rapidità*: i tempi di caricamento delle pagine devono essere adeguati. Quindi è fondamentale ottimizzare codici, testi e immagini per creare pagine snelle, leggere e veloci da caricare, in modo da non costringere gli utenti ad attese estenuanti che potrebbero portarli a fare clic sull'icona di chiusura del browser e a non farli più tornare sul sito;

– *prevenzione degli errori*: il modo migliore per evitare un errore è fare in modo che non si presenti. Tuttavia, se il sistema prevede errori, è necessario avvisare l'utente con messaggi semplici e costruttivi, evitando tecnicismi che un utente, soprattutto se inesperto, potrebbe non comprendere;

– *indicazioni e vie d'uscita*: qualora il sito sia particolarmente complesso, è bene indicare all'utente in quale punto del sito si trova e permettergli di raggiungere in modo facile e intuitivo qualsiasi pagina, facendo utilizzo di mappe, barre di stato e ritorni alla pagina precedente;

– *facilità d'uso*: per molti “usabilità” è sinonimo di “facilità d'uso”, quindi è assolutamente fondamentale realizzare un sito semplice, familiare, intuitivo, che porti l'utente a eseguire operazioni in modo naturale, senza che ci sia richiesta di un eccessivo sforzo mentale.

7.4 Accessibilità

L'accessibilità di un sito web può essere definita come “la capacità di un sito web di essere acceduto efficacemente (alla sua interfaccia e al suo contenuto) da utenti diversi in differenti contesti. Rendere un sito web accessibile significa permettere l'accesso all'informazione contenuta nel sito anche a persone con disabilità fisiche di diverso tipo e a chi dispone di strumenti hardware e software limitati”²⁰⁶. Il sito accessibile quindi è quello che può “essere visitato da qualsiasi utente indipendentemente dal computer usato, dalla velocità del collegamento, dal browser, dall'interfaccia utente, dalle tecnologie assistive utilizzate (tastiera braille, joystick speciali, ecc.), e dove il contenuto dell'informazione e la sua presentazione siano sempre indipendenti l'uno dall'altra”²⁰⁷.

206 Definizione di “accessibilità” del progetto WAI (*Web Accessibility Initiative*) del W3C. La definizione si può trovare all'indirizzo <http://www.w3c.it/wai/>.

207 Cristina Delogu et al., “Accessibilità. L'apprendimento per tutti” in *Tecnologie per il web learning: realtà e scenari*, a cura di Cristina Delogu (Firenze University Press), 2007, 120.

7. Una proposta metodologica: la fruizione per il pubblico

Chi progetta un sito deve sapere che non tutti gli utenti sono uguali e non tutti i loro sistemi di visualizzazione sono identici. Un sito quindi può definirsi accessibile quando ogni suo contenuto può essere raggiunto da tutti gli utenti (compresi gli ipovedenti e i diversamente abili) attraverso qualsiasi sistema di visualizzazione (compresi i browser testuali come Lynx).

Per rendere un sito accessibile è necessario attenersi a diverse regole. Il progetto WAI (*Web Accessibility Initiative*) del W3C ha elaborato una serie di linee guida, esistenti in due versioni: la 1.0, risalente al 1999, e la 2.0, del 2008, la cui traduzione in italiano è uscita di recente (il 30 dicembre 2009). La versione WCAG 2.0 (*Web Content Accessibility Guidelines*) suggerisce dodici linee-guida ispirate a quattro principi²⁰⁸, che si presentano nella seguente tabella:

WCAG 2.0	
1.	Percepibile – Le informazioni e i componenti dell'interfaccia utente devono essere presentati agli utenti in modo che possano essere percepiti:
1.1.	Fornire alternative testuali per qualsiasi contenuto non di testo in modo che lo stesso possa essere trasformato in altre forme fruibili secondo le necessità degli utenti, come stampa a grandi caratteri, Braille, sintesi vocale, simboli o linguaggio più semplice (1);
1.2.	Fornire alternative per i tipi di media temporizzati ²⁰⁹ (2);
1.3.	Creare contenuti che possano essere rappresentati in modalità differenti (per esempio, con layout più semplici), senza perdite di informazioni o di struttura (3);
1.4.	Rendere semplice per gli utenti la visione e l'ascolto dei contenuti, separando i livelli di primo piano e di sfondo (4).
2.	Utilizzabile – I componenti e la navigazione dell'interfaccia utente devono essere utilizzabili:
2.1	Rendere disponibili tutte le funzionalità anche tramite tastiera (5);

208 È possibile reperire la documentazione completa delle WCAG 2.0 all'indirizzo <http://www.w3.org/Translations/WCAG20-it/>.

209 Per “media temporizzati” si intendono i contenuti audio e video.

7. Una proposta metodologica: la fruizione per il pubblico

2.2	Fornire agli utenti tempo sufficiente per leggere e utilizzare i contenuti (6);
2.3	Non sviluppare contenuti che possano causare attacchi epilettici (7);
2.4	Fornire all'utente funzionalità di supporto per navigare, trovare contenuti e determinare la propria posizione (8);
3.	Comprensibile – Le informazioni e le operazioni dell'interfaccia utente devono essere comprensibili:
3.1	Rendere il testo leggibile e comprensibile (9);
3.2	Creare pagine web che appaiano e funzionino in modo prevedibile (10);
3.3	Aiutare gli utenti a evitare gli errori e agevolarli nella eventuale correzione (11);
4.	Robusto – Il contenuto deve essere abbastanza robusto per essere interpretato in maniera affidabile mediante una vasta gamma di programmi utente, comprese le tecnologie assistive:
4.1	Garantire la massima compatibilità con i programmi utente attuali e futuri, comprese le tecnologie assistive.

TAB. 7.2: le WCAG 2.0 del progetto WAI del W3C

Ma in che modo è possibile realizzare nel concreto le linee guida proposte dal progetto WAI? Di seguito si elencano alcuni importanti accorgimenti proposti da Mark Pilgrim²¹⁰:

- indicare la DOCTYPE all'interno delle pagine del sito (un'etichetta che indica il tipo di documento, simile a quella di XML);
- identificare la lingua del sito attraverso l'apposita etichetta (per esempio, `<html lang="it">`);
- dare titoli significativi (ogni pagina del sito web dovrebbe avere un titolo esplicativo e significativo);
- presentare subito il contenuto principale;
- usare colori sicuri, soprattutto per i collegamenti;
- utilizzare link “veri”, ovvero evitare link del tipo “`javascript:`”, perché per utenti che hanno browser testuali come Lynx è come se tali collegamenti non esistessero. In

210 Cfr. Mark Pilgrim, *Dive into accessibility – 30 days to a more accessible web site*, 2002, online all'indirizzo <http://diveintoaccessibility.org/>. Traduzione italiana: *Affrontare l'accessibilità – Per un sito più accessibile in 30 giorni* a cura di Franco Carcillo, <http://www.francocarcillo.it/dive/index.html>.

7. Una proposta metodologica: la fruizione per il pubblico

generale, è bene evitare l'uso di JavaScript, perché i browser testuali non lo supportano e i browser grafici consentono di disabilitare l'esecuzione di codice JavaScript (quindi possono esserci molti utenti anche normodotati con la visualizzazione di JavaScript disabilitata);

- aggiungere titoli ai link;
- definire scorciatoie da tastiera (le “accesskeys”, per permettere una navigazione agevole con la tastiera);
- definire gli acronimi;
- fornire testi equivalenti per le immagini, in modo tale da poterne presentare il contenuto anche a coloro che non le possono visualizzare (non vedenti o utenti con browser testuali);
- evitare immagini spaziatrici.

Inoltre è sempre bene fare in modo che i testi possano essere ingranditi (in modo tale da facilitare gli ipovedenti) e in generale è sempre necessario separare il contenuto dalla sua presentazione.

Una volta finite le operazioni, sarà importante inserire una pagina all'interno della quale dichiarare quali sono state le linee seguite per garantire accessibilità al sito. Particolarmente utile sarà includere in tale pagina l'elenco delle “accesskeys” che servono per navigare all'interno del sito. La dichiarazione di accessibilità è importante anche per trasparenza nei confronti degli utenti del sito.

Infine merita una menzione, in fatto di accessibilità, la Legge 4 gennaio 2004. n.4, *Disposizioni per favorire l'accesso dei soggetti disabili agli strumenti informatici*, meglio nota come “Legge Stanca” dal nome dell'allora ministro proponente Lucio Stanca. La Legge²¹¹ intende per accessibilità “la capacità dei sistemi informatici, nelle forme e nei limiti consentiti dalle conoscenze tecnologiche, di erogare servizi e fornire informazioni fruibili, senza discriminazioni, anche da parte di coloro che a causa di disabilità necessitano di tecnologie assistive o configurazioni particolari” e obbliga le pubbliche amministrazioni e le aziende a prevalente capitalizzazione pubblica a stipulare contratti per la realizzazione e la modifica di siti internet che rispettino i requisiti di accessibilità stabiliti dal Decreto Ministeriale 8 luglio 2005²¹², attuativo della Legge. Tale Decreto stabilisce, nei suoi allegati, i requisiti tecnici di accessibilità dei siti internet e le metodologie e i criteri di valutazione per la verifica

211 Il testo è disponibile all'indirizzo http://www.pubbliaccesso.gov.it/normative/legge_20040109_n4.htm.

212 Il testo del Decreto e dei suoi allegati è disponibile all'indirizzo <http://www.pubbliaccesso.gov.it/normative/DM080705.htm>.

7. Una proposta metodologica: la fruizione per il pubblico

dell'accessibilità dei siti stessi.

I requisiti per i siti sono ventidue e sono stati stabiliti sulla base delle raccomandazioni del progetto WAI di cui si è parlato poco sopra, degli standard definiti nella Sezione 508 del *Rehabilitation Act* degli Stati Uniti d'America (la Sezione stabilisce i requisiti di accessibilità dei siti della pubblica amministrazione statunitense), degli standard ISO e delle “esperienze acquisite nell'ambito della Pubblica Amministrazione”.

I ventidue requisiti fanno esplicitamente riferimento alle raccomandazioni WAI (in particolare, la versione 1.0 delle WCAG): tra di essi, si impone la conformità alle specifiche del W3C per HTML e XHTML, il divieto di utilizzo di frame, l'obbligo di utilizzare alternative testuali a qualsiasi oggetto non testuale (immagini, audio, video ecc.), il divieto di utilizzare grafiche che possano provocare disturbi da epilessia fotosensibile, la separazione dei contenuti dalla presentazione mediante l'utilizzo di fogli di stile, la completa fruibilità del sito anche in caso di script, applet o altri oggetti di programmazione.

Si tratta di una legge importante perché introduce il tema dell'accessibilità nella pubblica amministrazione, e l'imposizione di vincoli forti ai siti della pubblica amministrazione può offrire un'occasione per un netto miglioramento della qualità dei servizi forniti ai cittadini. E, oltre che un obbligo per le pubbliche amministrazioni, la Legge può essere un invito anche per molti siti che sono particolarmente mancanti in fatto di accessibilità.

7.5 Usability test e Website evaluation

Per monitorare il sito in modo adeguato e per verificare quali siano i punti di forza e i punti deboli del sito stesso nonché le aspettative degli utenti, i loro motivi di soddisfazione o, viceversa, per conoscere quali siano gli aspetti per loro da migliorare, è necessario condurre apposite ricerche. Due strumenti in particolare possono rivelarsi molto utili allo scopo: lo *usability test* e la *website evaluation*.

Il primo permette di condurre ricerche di tipo *qualitativo* (ovvero ricerche che cercano di illustrare in modo articolato e approfondito l'oggetto d'indagine, descrivendone la sua complessità), mentre il secondo fa sì che possano essere svolte ricerche di tipo *quantitativo* (ossia ricerche il cui obiettivo è misurare un certo oggetto d'indagine).

Lo *usability test* è uno strumento che “permette di ottimizzare il disegno e le funzionalità sviluppate all'interno di uno specifico sito attraverso il coinvolgimento in test 'di laboratorio'

7. Una proposta metodologica: la fruizione per il pubblico

di un ristretto numero di utenti prima del lancio ufficiale in rete (*early usability test*) o ad attività già avviate, per valutare come concretamente i navigatori utilizzano uno specifico sito nell'ambito del proprio contesto abituale di fruizione (*contextual inquiry*)²¹³. Ovviamente gli utenti selezionati per effettuare questo tipo di test dovranno essere rappresentativi del pubblico che dovrà visitare il sito.

Lo *usability test* può essere determinante in quanto, chiedendo al campione selezionato di navigare sul sito, è possibile “evidenziare eventuali funzionalità da aggiungere, togliere o semplicemente ripensare”²¹⁴, comprendendo quindi quale può essere l'atteggiamento degli utenti nei confronti dei contenuti, della grafica, della facilità d'uso, dei tempi di caricamento e individuare quindi su quali aree del sito intervenire per poter offrire un prodotto migliore.

La *website evaluation* invece si configura come un “questionario via web o email volto a sondare grado di ricordo e gradimento”²¹⁵ nei confronti del sito web. In particolare, “è la forma di ricerca volta a mettere a fuoco le funzionalità del sito maggiormente gradite e quelle invece potenzialmente in grado di allontanare gli utenti o comunque dagli stessi valutate come scarsamente rilevanti”²¹⁶. Attraverso la *website evaluation* è quindi possibile conoscere quali sono le sezioni del sito più apprezzate e viceversa quelle che riscontrano minore apprezzamento, e in più sarà anche possibile dotare il questionario di domande a risposta aperta in modo da “rendere le funzionalità proposte sempre più coerenti rispetto alle preferenze e alle attese degli utenti”²¹⁷.

213 Emanuela Prandelli, Gianmario Verona, *Marketing in rete. Oltre Internet verso il nuovo marketing* (Milano: McGraw-Hill, 2006), 219.

214 *Ibid.*, 220.

215 *Ibid.*, 222.

216 *Ibid.*

217 *Ibid.*

8. *Carlo Finelli Corpus*

8.1 Obiettivi

Carlo Finelli Corpus, corpus telematico dei carteggi di Carlo Finelli (1782 – 1853), importante scultore neoclassico di Carrara, è stato realizzato per conto della Fondazione Cassa di Risparmio di Carrara (www.fondazioneccarrara.com) e di Losna – Soluzioni per il Web (www.losnaweb.com), in collaborazione con il Corso di Laurea in Informatica Umanistica dell'Università di Pisa (<http://infouma.di.unipi.it>).

Il corpus è stato pubblicato sugli spazi forniti da Losna – Soluzioni per il Web ed è disponibile all'indirizzo <http://carlofinelli.losnaweb.com>.

Il corpus contiene ventuno lettere di Carlo Finelli mai pubblicate finora in rete (per un totale di 5.634 occorrenze) e consente di poter leggere i testi integrali delle missive inviate dall'artista e di poter svolgere al loro interno tutti i tipi di ricerca (per forma, per lemma ecc.) elencati nel capitolo 6 della presente trattazione.

Carlo Finelli Corpus, oltre che come strumento di studi e di approfondimento sull'arte dello scultore carrarese, è importante anche in quanto primo progetto realizzato in modo conforme alla proposta metodologica presentata in questa sede.

Diverse sono le motivazioni che hanno spinto a realizzare il progetto:

- promuovere l'arte di Carlo Finelli, poco conosciuta anche tra i suoi concittadini;
- pubblicare su internet lettere che finora erano appannaggio di studiosi o di appassionati dell'arte di Carlo Finelli e quindi mettere a disposizione di tutti un'importante risorsa per lo studio dell'arte di Carlo Finelli in particolare e dell'arte neoclassica in generale;
- far conoscere agli “addetti ai lavori” quali possono essere le potenzialità dell'informatica per i beni culturali;
- combinare studi di tipo linguistico a studi di tipo storico-artistico in una cornice possibile grazie agli strumenti che la tecnologia mette a disposizione;
- dare lustro alla città di Carrara con un progetto culturale innovativo;
- creare un progetto conforme agli standard e alla metodologia proposti nella presente trattazione, che possa fungere anche da progetto pilota per altri che vogliano adeguarsi a quanto descritto in questa sede.

8.2 Descrizione del progetto

Il sito di *Carlo Finelli Corpus* è composto di sei sezioni:

1. *Home Page*: al suo interno è presente una descrizione generale del progetto, che servirà per meglio guidare l'utente nella navigazione, per informarlo sui perché del sito, per fargli capire cosa può trovare al suo interno;

2. *Carlo Finelli*: la sezione contiene una breve biografia dello scultore, divisa in quattro parti. La biografia, oltre a contenere importanti notizie sulla vita di Carlo Finelli (molte delle quali desumibili dai carteggi), contiene anche alcune brevi note per comprendere le sue opere più importanti;

3. *Le lettere*: all'interno di questa sezione il visitatore può trovare tutte le lettere di Carlo Finelli, i cui testi sono stati pubblicati in modo integrale;

4. *Ricerca*: in questa sezione l'utente trova i collegamenti alle pagine dei vari tipi di ricerca. Sono state implementate per *Carlo Finelli Corpus* la ricerca per forma, per lemma, per categoria grammaticale, per cooccorrenze, per stringa contenuta nel testo e la ricerca all'interno dei contenuti (ricerca per destinatario, luogo, luogo di conservazione e data);

5. *Bibliografia*: contiene una bibliografia per poter approfondire l'arte di Carlo Finelli;

6. *Informazioni*: contiene informazioni di carattere generale sul progetto *Carlo Finelli Corpus* (sulle aziende che hanno contribuito alla realizzazione, sui copyright, sull'implementazione) nonché un indirizzo e-mail per contatti.

La parte destra del menù contiene, oltre ai loghi delle aziende, anche diverse immagini di opere di Carlo Finelli, mostrate a rotazione in modo casuale: facendo clic su di esse è possibile ingrandirle (tutte le immagini sono state fornite dalla Fondazione Cassa di Risparmio di Carrara). Sempre nella parte destra sono inoltre presenti il collegamento che rimanda alla dichiarazione di accessibilità e i loghi della validazione secondo gli standard XHTML 1.0 e CSS 2.1. del W3C²¹⁸.

L'header del sito (la testata) riporta il logo del progetto, ovvero la testa di una delle tre *Ore Danzanti* di Carlo Finelli in gesso, opera conservata presso l'Accademia di Belle Arti di Carrara (le *Ore Danzanti* in marmo sono invece all'Hermitage di San Pietroburgo). Al suo fianco, compare la scritta, mentre la *Ebe* del 1835 separa logo e scritta dal sottotitolo e dai loghi di Fondazione Cassa di Risparmio di Carrara, Losna – Soluzioni per il Web e

218 Cfr. *supra* PAR. 7.4.

8. Carlo Finelli Corpus

Informatica Umanistica. Di seguito si propongono due immagini tratte da *Carlo Finelli Corpus*:



FIG. 8.1: home page di *Carlo Finelli Corpus*



FIG. 8.2: la pagina "Le lettere" di *Carlo Finelli Corpus*

8.3 Applicazione della metodologia proposta

Durante le fasi di realizzazione di *Carlo Finelli Corpus* sono stati seguiti gli standard proposti nella presente trattazione, e lo stesso *Carlo Finelli Corpus* vuole dimostrare quali potenzialità possono avere le pubblicazioni telematiche di carteggi d'artista realizzate secondo la metodologia qui discussa.

La prima operazione è stata l'acquisizione delle lettere, avvenuta manualmente. Purtroppo, non essendo stato possibile avere a disposizione gli originali, il progetto deve per il momento fare a meno delle immagini delle lettere autografe di Carlo Finelli.

Le lettere sono state quindi tokenizzate attraverso il tokenizzatore *Token-It*²¹⁹, quindi lemmatizzate e analizzate morfologicamente tramite l'uso di *MAGIC*²²⁰. La lemmatizzazione ha posto qualche lieve problema di carattere linguistico: dal momento che *MAGIC* è un programma pensato per l'italiano contemporaneo, qualche occorrenza non è stata riconosciuta in modo corretto. Si è trattato però di una quantità limitata, tale da non giustificare il ricorso ad altri programmi per la lemmatizzazione e l'analisi morfologica. Le operazioni si sono quindi svolte in modo automatico, ma le ambiguità e le occorrenze non riconosciute sono state risolte manualmente.

Quindi, i testi finali, lemmatizzati e annotati, sono stati puliti in modo tale da poter essere trattati per essere trasformati in file XML mediante l'utilizzo del codice Perl descritto nel capitolo 5²²¹.

I file XML “grezzi” sono stati quindi resi conformi alla DTD appositamente pensata per i corpora realizzati seguendo le linee della metodologia qui presentata. L'operazione successiva è stata la creazione dei file XML delle lettere, anch'essi resi conformi alle DTD proposte in questa sede.

Il passo successivo è stato la creazione della base di dati MySQL, con due tabelle: una per i token e una per le lettere. Tramite l'utilizzo degli appositi codici PHP, i file XML sono stati trasformati in record MySQL e quindi resi pronti per l'utilizzo.

Una volta compiute le operazioni preliminari è stato possibile procedere con la realizzazione del template grafico e della struttura del sito: sono stati quindi integrati i codici per la presentazione delle lettere e per le ricerche. Di seguito si propone un'immagine dei

219 Cfr. *supra* PAR. 5.3.

220 Cfr. *supra* PAR. 5.4.

221 Cfr. *supra* PAR. 5.5.

8. Carlo Finelli Corpus

contesti KWIC della forma “presente”:

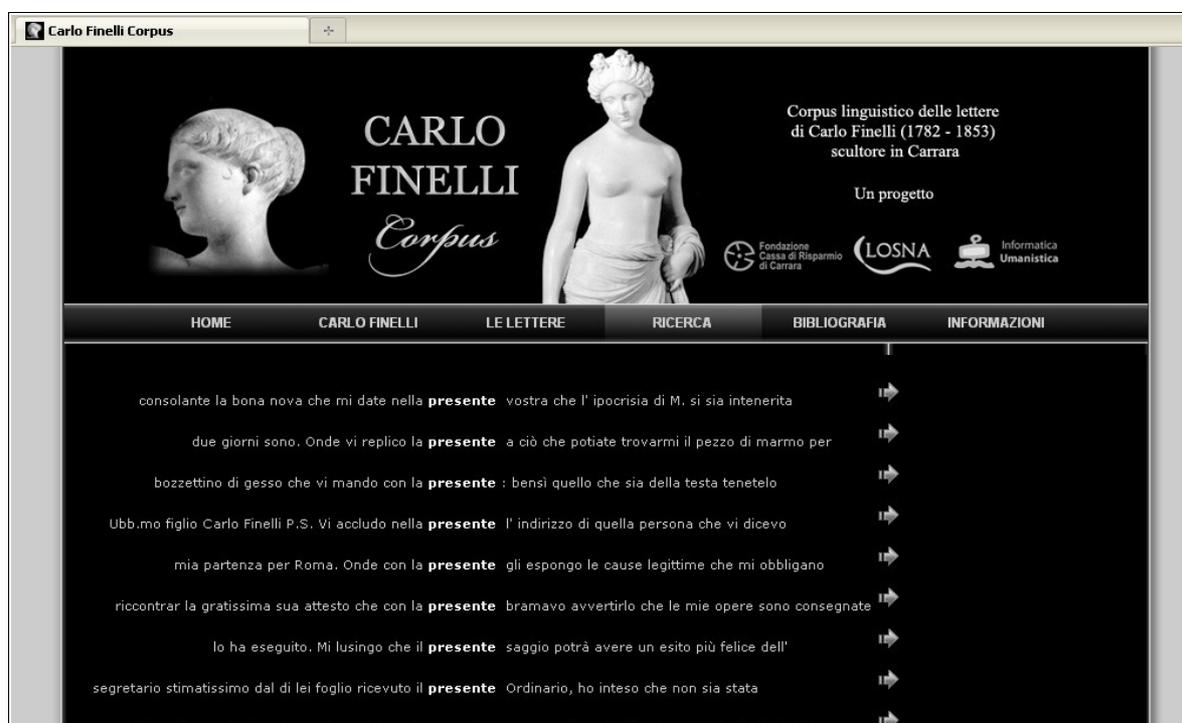


FIG. 8.3: i contesti KWIC della forma “presente” in *Carlo Finelli Corpus*

Le frecce accanto al contesto rimandano alla lettera che contiene la frase.

Una volta finito il tutto, il sito è stato pubblicato sul server. Le operazioni preliminari (tra cui anche la trasformazione dei file XML in record MySQL) sono state tutte eseguite in locale. Sul server è stata trasferita soltanto la base di dati già pronta: questo per non dover caricare sul server anche gli XML e quindi per non occupare spazio (come già ribadito più volte nel corso della trattazione, i file XML servono unicamente per la realizzazione della base di dati).

Si potrebbe obiettare sostenendo che un sistema pensato in questo modo (cioè pensato per essere sviluppato prima in locale e poi trasferito sul server remoto) non sarebbe adatto ad aggiornamenti. Si tratta di una critica che può essere rigettata con facilità, dal momento che, come detto sopra, i corpora devono avere dimensioni finite²²². Quindi, una volta creati, disambiguati, corretti e riveduti i file XML relativi a tutti i testi, non avrebbe più senso fare aggiornamenti ai testi.

La struttura grafica è stata realizzata in modo elegante e gradevole ed è basata sui colori

²²² Cfr. *supra* PAR. 3.1.

nero-bianco-grigio (per richiamare la scultura e il marmo: si è pensato quindi a una grafica coerente con il prodotto proposto). Inoltre, durante la realizzazione della grafica si è tenuto conto di principi di chiarezza e facilità d'uso: il menù è facilmente intuitivo, la grafica non è dispersiva e l'utente ha in una sola schermata tutto ciò di cui ha bisogno per navigare in modo rapido e facile. Le pagine più interne del sito sono state dotate di opportuni pulsanti per tornare alla pagina precedente.

Infine, ampio peso è stato dato all'accessibilità: le immagini sono state dotate di adeguate alternative testuali, sono state realizzate scorciatoie da tastiera, non si è fatto alcun uso di codice JavaScript, i collegamenti sono stati ben evidenziati, i caratteri sono stati resi ingrandibili, i contenuti sono stati redatti in modo chiaro e conciso. Il sito è stato quindi validato secondo gli standard XHTML 1.0 e CSS 2.1 del W3C: i loghi sono stati inseriti nel blocco destro del sito, così come il collegamento alla dichiarazione di accessibilità, facendo clic sul quale l'utente può trovare la lista delle "accesskeys" e constatare in quale modo è stata garantita la piena accessibilità a tutti i tipi di utenti.

Per avere un'ulteriore garanzia della sua accessibilità, *Carlo Finelli Corpus* è stato visitato anche attraverso *Lynx*, un browser testuale: ogni parte del sito (comprese le ricerche) è visibile in modo ottimale anche su questo tipo di sistema di visualizzazione.

8.4 La struttura del sito

La costruzione delle pagine del sito e dei fogli di stile CSS è avvenuta in modo sincronico. Sono stati creati tre fogli di stile: due per la presentazione su schermo (uno per i browser più recenti e uno per le versioni di Internet Explorer precedenti alla 7) e uno per la stampa.

Innanzitutto è importante specificare che un documento HTML inizia con la dichiarazione della DOCTYPE (è opzionale: le pagine, senza tale dichiarazione, funzioneranno ugualmente ma inserirla è una buona pratica, irrinunciabile per creare siti usabili e accessibili, anche perché la sua assenza non è conforme agli standard del W3C), e continua quindi con l'etichetta "html", all'interno della quale compariranno i contenuti.

Normalmente una pagina HTML è divisa in due parti: una *head* che contiene informazioni sulla pagina e un *body* che invece contiene i contenuti:

8. Carlo Finelli Corpus

```
<!DOCTYPE html PUBLIC "-//W3C//DTD XHTML 1.0 Transitional//EN"
"http://www.w3.org/TR/xhtml1/DTD/xhtml1-transitional.dtd">

<html>
<head>

Informazioni (titolo, autore, collegamenti ai fogli di stile
ecc.)

</head>

<body>

Contenuti

</body>

</html>
```

TAB. 8.1: Struttura tipica di una pagina HTML

Le pagine di *Carlo Finelli Corpus* si aprono con la dichiarazione della DOCTYPE e della lingua:

```
<!DOCTYPE html PUBLIC "-//W3C//DTD XHTML 1.0 Transitional//EN"
"http://www.w3.org/TR/xhtml1/DTD/xhtml1-transitional.dtd">

<html xmlns="http://www.w3.org/1999/xhtml" xml:lang="it"
lang="it">
```

TAB. 8.2: DOCTYPE ed etichetta "html" con indicazione della lingua

All'interno della *head* delle pagine (ovvero l'elemento che contiene informazioni aggiuntive) sono stati dichiarati il linguaggio dei contenuti, il tipo di contenuti e il tipo di codifica, il nome dell'autore, il titolo del progetto e i collegamenti ai fogli di stile:

```
<head>

<meta http-equiv="Content-Language" content="it" />

<meta http-equiv="Content-Type" content="text/html;
charset=WINDOWS-1252" />
```

8. Carlo Finelli Corpus

```
<meta name="author" content="Federico Giannini" />

<meta name="description" content="Carlo Finelli Corpus" />

<title>Carlo Finelli Corpus</title>

<link rel="stylesheet" media="screen, projection"
type="text/css" href="stile.css" title="1024" />

<link rel="stylesheet" type="text/css" media="print"
href="stampa.css" />

<link rel="icon" href="Grafica/favicon.ico" />

<!--[if lte IE 6]>
<link rel="stylesheet" type="text/CSS" href="stilevecchi.css">
<![endif]-->
</head>
```

TAB. 8.3: *head* delle pagine di *Carlo Finelli Corpus*

Le prime etichette che compaiono nella head sono le cosiddette “metatag”, ovvero elementi che servono per fornire informazioni aggiuntive sul sito (titolo, autore, lingua ecc.). Le metatag sono elementi “vuoti” (*empty tags*): significa che non hanno l’etichetta di chiusura. Altri esempi di elementi vuoti sono `` (per le immagini), `
` (per andare a capo), `<hr>` (per inserire una linea orizzontale), `<link>` (che definisce una relazione tra una pagina HTML e un file esterno, in questo caso i fogli di stile). Anche se gli elementi vuoti non prevedono l’etichetta di chiusura, questo non vuol dire che non debbano essere chiusi: per farlo è necessario aggiungere una slash prima della parentesi uncinata chiusa (per esempio: `
`). La chiusura degli elementi vuoti è imprescindibile per un codice ben formattato, anche perché il validatore W3C non permette documenti HTML che abbiano elementi aperti.

Si noterà che nella head mancano le “keyword”, ovvero le parole chiave per aumentare la rintracciabilità sui motori di ricerca: benché molti ancora si ostinino a indicare come una buona abitudine inserire l’etichetta “keywords” all’interno delle pagine del proprio sito, è necessario specificare in questa sede che le keyword ormai hanno un’influenza minima sul comportamento dei motori di ricerca²²³. Le tecniche per garantire una maggior rintracciabilità sui motori di ricerca sono altre: in questa sede sarà sufficiente dire che dare titoli significativi alle pagine ed esprimere i concetti chiave nei contenuti (soprattutto nei contenuti della home

223 Jennifer Grappone e Gradiva Couzin, *Search Engin Optimization: an hour a day* (Hoboken, New Jersey: Wiley, 2008), 57 ss.

8. Carlo Finelli Corpus

page) è già una pratica importante, che dà buoni risultati.

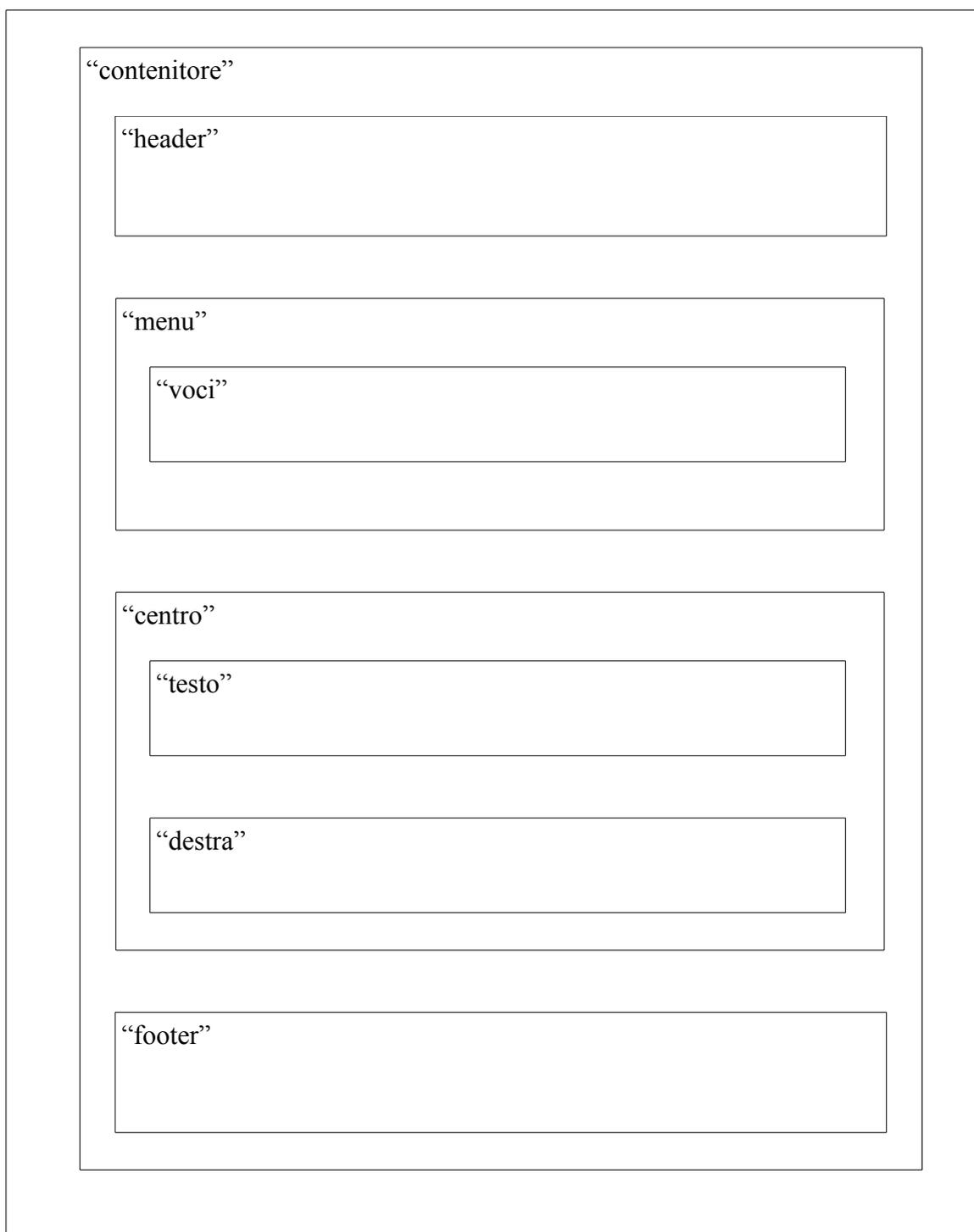
Come si può notare dalla tabella, i fogli di stile sono tre: il primo è “stile.css” ed è quello principale, per la presentazione del sito. Il secondo è “stampa.css” e serve per stampare i contenuti, mentre il terzo, “stilevecchi.css”, si attiva soltanto se il browser è una versione uguale o precedente alla 6.0 di Internet Explorer (<!-- [if lte IE 6]>).

Il contenuto del sito è stato strutturato in *div*, ovvero in “blocchi” che servono per contenere testo, immagini e quant'altro.

Ci sono otto *div* all'interno delle pagine principali di *Carlo Finelli Corpus* (diventano sette soltanto nelle pagine che mostrano i contesti KWIC):

- *contenitore*: è il *div* principale, che contiene tutti gli altri;
- *header*: la testa del sito, contiene l'immagine con il logo (impostata come background del *div*);
- *menu*: contenitore del menù;
- *voci*: contiene le singole voci del menù (la realizzazione di due *div* separati *menu* e *voci* è stata necessaria per motivi di grafica, dal momento che il *div* delle voci doveva occupare uno spazio di poco più ristretto di quello occupato dal *div menu*);
- *centro*: la parte centrale della pagina che contiene il blocco del testo e il blocco destro;
- *testo*: contiene i testi del sito;
- *destra*: il blocco destro, con le immagini a rotazione, i loghi linkati di Fondazione Cassa di Risparmio di Carrara, Losna e Informatica Umanistica, i loghi linkati degli standard W3C e la dichiarazione di accessibilità;
- *footer*: contiene il piede della pagina.

La tabella nella pagina seguente mostra la struttura dei *div* delle pagine principali di *Carlo Finelli Corpus*.



TAB. 8.4: struttura delle pagine di *Carlo Finelli Corpus*

Il blocco destro e il footer sono stati realizzati in due documenti separati (“destra.php” e “footer.html”) che vengono richiamati all'interno della pagina del sito tramite la funzione “include”, la cui sintassi è:

8. Carlo Finelli Corpus

```
include("includes/destra.php");
```

I due documenti sono stati inseriti in una cartella “includes”, che contiene anche il file “config.php” e il file “funzioni.php”. Il file “config.php” contiene le informazioni necessarie per la connessione alla base di dati MySQL (indirizzo dell'host, nome utente della base di dati, nome della base di dati e password). Per la connessione alla base di dati si fa uso delle funzioni “mysql_connect” e “mysql_select_db”: la prima serve per la connessione al server, la seconda serve per selezionare la base di dati, e ambedue le funzioni vengono chiamate solo nelle pagine che necessitano di collegarsi alla base di dati (quindi le pagine che mostrano i testi delle lettere e le pagine per la ricerca).

La pagina “funzioni.php” viene invece inclusa in tutte le pagine del sito e contiene le funzioni: “formatodata”²²⁴, della quale si è già discusso, e le due funzioni necessarie per la rotazione delle immagini. Tali due funzioni sono definite nel modo seguente:

```
$destra[] = "immaginidestra/amore";
$destra[] = "immaginidestra/amorepsiche";
$destra[] = "immaginidestra/angelo";
$destra[] = "immaginidestra/ariosto";
$destra[] = "immaginidestra/aurora";
$destra[] = "immaginidestra/ebe";

[...]

$destra[] = "immaginidestra/venere";

$totale = count($destra) - 1;

$casuale = rand(0, $totale);

function rotazione()
{
    global $destra;
    global $totale;
    global $casuale;

    echo '<a href="' . $destra[$casuale] . '.1.jpg'
target="_blank"></a>';
}
```

224 Cfr. *supra* PAR. 5.8.

```
function associazione() {  
  
    global $casuale;  
  
    if ($casuale == '0') {  
        echo "<i>Amore che tormenta l'anima</i><br />1822, Marmo, h.  
127 cm<br />Bakewell, collezioni del Duca del Devonshire";  
    }  
    if ($casuale == '1') {  
        echo "<i>Amore sdegnato con Psiche</i><br />1830, Marmo, h.  
137 cm<br />Las Vegas, collezione privata";  
    }  
    if ($casuale == '2') {  
        echo "<i>Angelo del Giudizio</i><br />1829, Marmo, h. 150  
cm<br />Piacenza, San Giovanni in Canale";  
    }  
    if ($casuale == '3') {  
        echo "<i>Ludovico Ariosto</i><br />1815, Marmo<br />Roma,  
Protomoteca Capitolina";  
    }  
    if ($casuale == '4') {  
        echo "<i>L'Aurora</i><br />1824, Marmo, h. 53 cm<br  
>Carrara, collezione Marchini";  
    }  
    if ($casuale == '5') {  
        echo "<i>Ebe</i><br />1835, Marmo, h. 146,5 cm<br />Milano,  
collezione privata";  
    }  
  
    [...]  
  
    if ($casuale == '16') {  
        echo "<i>Venere che esce dalla conchiglia</i><br />1847,  
Marmo, h. 105 cm<br />San Pietroburgo, Hermitage";  
    }  
  
}
```

TAB. 8.5: Le due funzioni per la rotazione delle immagini

Le variabili “\$destra[]” sono elementi di un array, identificati in modo tale dalle parentesi quadre vuote. Gli elementi contengono i percorsi alle immagini (queste ultime sono state però private del suffisso “.jpg”, perché verrà aggiunto in seguito).

La variabile “\$totale” conta il numero degli elementi dell'array: il numero dovrà essere poi decrementato di una unità. Questo perché il primo elemento dell'array ha indice 0 (e non 1): poiché gli elementi dell'array sono diciassette, l'ultimo avrà indice 16. Se non si decrementasse la variabile “\$totale”, la variabile successiva, ovvero

8. Carlo Finelli Corpus

```
$casuale = rand(0, $totale);
```

assumerebbe un valore casuale (definito dal metodo “rand”) compreso tra 0 e 17. Ma nel momento in cui si richiama un elemento con indice 17 ci sarebbe un errore, in quanto l'ultimo elemento ha indice 16.

Dichiarate le variabili, è possibile definire la funzione “rotazione”: le variabili “\$destra”, “\$totale” e “\$casuale” devono essere definite globali in quanto dichiarate al di fuori della funzione (questo perché la variabile “\$casuale” viene richiamata anche dalla funzione “associazione”, che serve per associare a ogni immagine la corretta didascalia). La funzione non fa altro che stampare l'immagine definita in modo casuale attraverso le due variabili “\$destra” e “\$casuale”, nonché il rispettivo collegamento all'immagine ad alta definizione.

La successiva funzione “associazione” definisce come globale la variabile “\$casuale” e, attraverso opportuni blocchi if, stampa la corretta didascalia associata alle immagini (per esempio, se compare l'immagine che ha come indice 0, cioè *Amore che tormenta l'anima*, opera del 1822 conservata presso le collezioni del Duca del Devonshire, comparirà la rispettiva didascalia).

L'ultima analisi riguarda i fogli di stile CSS. Al loro interno, le dichiarazioni delle informazioni sono affidate a dei *selettori* che portano il nome dell'etichetta HTML alla quale si devono riferire gli attributi. Il primo selettore del foglio “stile.css” è quello relativo all'etichetta “body”:

```
body {  
  
background-color:#cccccc;  
color: #ffffff;  
font-family: arial;  
font-size:12px;  
  
}
```

TAB. 8.6: definizione dell'elemento “body” in “stile.css”

In questo caso viene specificato il colore dello sfondo (“background-color”) con il rispettivo codice esadecimale, quindi si dichiarano il tipo e la dimensione del carattere (nel caso di *Carlo Finelli Corpus*, un arial a dodici pixel). I fogli di stile sfruttano la struttura “a

cascata”, cioè gli “elementi-figli” riprendono tutte le caratteristiche degli elementi dai quali dipendono, se non viene specificato qualcosa di diverso. Quindi, a meno che non venga specificato un carattere differente, tutti gli elementi contenuti nel corpo della pagina avranno carattere arial a dodici pixel.

I div hanno tutti strutture più o meno simili. Di seguito si presenta il più complesso, ovvero il div “testo”:

```
#testo {  
  position:relative;  
  margin-top:0px;  
  margin-left:23px;  
  width:655px;  
  background-color:#000000;  
  text-align:justify;  
  z-index:3;  
  float:left;  
  padding:20px 15px 70px 20px;  
}
```

Tab. 8.7: definizione dell'elemento “testo” in “stile.css”

La prima proprietà, “position:relative” indica che le proprietà che specificano il posizionamento dell'elemento “testo” (come “margin-top”) devono essere considerate in relazione con l'elemento che gerarchicamente lo precede (quindi l'elemento “menu”): specificando quindi la proprietà “margin-top” (ovvero la distanza del bordo superiore dall'elemento precedente) a zero pixel, significa che il div “testo” sarà attaccato al div “menu”.

La proprietà “width” indica la larghezza dell'elemento (in questo caso, 655 pixel); “background-color” è, come visto poco sopra, il colore dello sfondo (qui specificato perché diverso dal colore indicato per il corpo del documento); “text-align” indica l'allineamento del testo (giustificato), “z-index” è l'ordine di sovrapposizione dell'elemento, “float” consente di rimuovere il div dal normale flusso del documento e di spostarlo, in questo caso, a sinistra, e infine “padding” indica la distanza in pixel che gli elementi contenuti nel div devono mantenere nei confronti dei bordi rispettivamente in alto, a sinistra, in basso e a destra.

Qualora si debba utilizzare un'immagine al posto del colore nello sfondo di un div, si utilizzerà la proprietà “background-image”, come nel caso del div “menù”, per il quale tale

8. Carlo Finelli Corpus

proprietà è stata impostate nel modo seguente:

```
background-image:url('grafica/sfondomenu.png');
```

Sono stati poi impostati specifici selettori anche per altri elementi (titoli, immagini, collegamenti). Nel caso dei titoli e dei collegamenti, sono state impostate proprietà per decorarli: per esempio i titoli sono stati evidenziati in grassetto ed è stata specificata una distanza di 25 pixel dagli elementi immediatamente successivi, mentre per i collegamenti è stata definita la marcatura in grassetto e il cambio di colore al passaggio del mouse. Invece, nel caso delle immagini, sono state definite delle classi per allineare le immagini a destra e a sinistra o per centrarle.

9. Possibili utilizzi

9.1 Introduzione

Benché siano evidenti i vantaggi offerti dalla proposta metodologica presentata in questo lavoro, potrebbe essere lecito domandarsi a che cosa possa essere utile e chi possa trarre benefici dalla pubblicazione telematica di un corpus di carteggi d'artista. In questa sezione si tenterà di rispondere a queste due domande, e per farlo si è deciso di suddividere il capitolo in base alle categorie di utilizzatori che potrebbero trarre vantaggi dalle pubblicazioni telematiche di carteggi d'artista. In particolare, sono state individuate cinque categorie per i quali la presente proposta potrebbe rivelarsi molto utile:

- per i musei;
- per organizzatori e curatori di mostre;
- per i linguisti;
- per docenti e studenti;
- per gli appassionati.

9.2 Per i musei

Le nuove tecnologie offrono ai musei “interessanti possibilità per l'utilizzo di strumentazioni avanzate che consentono di produrre e di rendere accessibili a chiunque dati di difficile reperibilità, sino ad ora riservati agli addetti, e servizi che dilatano le modalità di approccio al museo”²²⁵. È il caso dei carteggi e dei documenti storici: come si è detto sopra²²⁶, spesso non tutti possono accedere a documenti conservati presso archivi storici. Tali documenti però potrebbero essere una notevole fonte di approfondimento per inquadrare meglio la produzione di un'artista conservata all'interno di un museo. In particolare, la possibilità di fruire di un'edizione telematica o informatica dei carteggi di un artista può essere utile per la realizzazione di supporti didattici e informativi riguardanti i contenuti del museo.

Non bisogna poi trascurare l'aiuto che un'edizione telematica di un carteggio può fornire per la realizzazione di supporti pensati per utenti *off-line* (CD-Rom divulgativi multimediali) o per l'allestimento di realtà museali virtuali, ovvero quelle applicazioni che elaborano

225 Cecilia Prete, *Aperto al pubblico. Comunicazione e servizi educativi nei musei* (Firenze: Edifir, 2005).

226 Cfr. *supra* PAR. 1.7.

9. Possibili utilizzi

“itinerari e contesti espositivi inesistenti con materiali selezionati secondo le finalità prefissate dagli operatori”²²⁷ e che sono “potenzialmente capaci [...] di offrire illimitate possibilità di fruizione, un'assoluta libertà di scelta dei contenuti da rappresentare e una completezza di informazioni che un'esposizione reale non potrebbe mai eguagliare”²²⁸.

Le nuove tecnologie in ambito museale permettono altresì di creare “percorsi non confinati al perimetro interno del museo, né alla sola disciplina o al genere al quale appartiene. Vi devono essere possibili esplorazioni non solo di oggetti, ma di idee e personaggi, di luoghi ed eventi, con l'intervento di testimoni e di commentatori autorevoli, diversi livelli di approfondimento, possibilità di accesso a sussidi didattici, ecc.”²²⁹. È evidente quindi che un progetto volto a digitalizzare un carteggio d'artista può essere di grande aiuto nella realizzazione di quel “percorso non confinato al perimetro interno del museo” che può non solo offrire un notevole contributo per la promozione dell'opera di un artista e per una sua conoscenza più approfondita, ma può essere anche un importante mezzo di promozione del museo stesso: una politica culturale volta alla diffusione della conoscenza di uno o più artisti, della loro opera e della loro sfera privata, perseguita anche attraverso la diffusione (telematica o su supporti *off-line*) di un carteggio d'artista altrimenti difficilmente fruibile da parte del pubblico, non può che giovare all'immagine del museo.

Tra i progetti realizzabili da enti museali è possibile inoltre citare le “mostre virtuali”, ossia mostre fruibili attraverso il calcolatore: grazie a esse è possibile presentare, a scopo divulgativo e come invito a osservarle dal vivo, opere di artisti che nella realtà sarebbe difficile riunire assieme in un unico contesto²³⁰. Anche per l'allestimento di mostre virtuale possono rivelarsi utili le pubblicazioni telematiche di carteggi d'artista.

Da non trascurare inoltre l'aiuto che i carteggi possono fornire per allestire il percorso museale stesso: didascalie e pannelli sono parte integrante del percorso e attraverso lo studio dei carteggi di un artista possono essere allestiti con la maggior completezza possibile. I carteggi possono essere utili non solo per l'allestimento, ma anche per l'aggiornamento di didascalie e pannelli, dato che “la didascalizzazione e, ancora di più, i pannelli didattici sono

227 Cecilia Prete, *Aperto al pubblico*.

228 *Ibid.*

229 Paolo Galluzzi, “Nuove tecnologie e funzione culturale nei musei” in *I formati della memoria. Beni culturali e nuove tecnologie alle soglie del terzo millennio*, a cura di Paolo Galluzzi e Pietro Valentino (Giunti, 1997), 28.

230 Si vedano a tal proposito i progetti *Caravaggio: una mostra impossibile*, disponibile all'indirizzo <http://www.caravaggio.rai.it> e *Ugo Guidi: l'uomo, l'arte, le grandi opere*, disponibile all'indirizzo <http://mostra.ugoguidi.it>.

9. Possibili utilizzi

soggetti a variazioni [...] rapide [...] perché sono legati non solo all'avanzamento della ricerca, ma anche all'evolversi della domanda del pubblico e dei progetti di divulgazione del museo”²³¹. Come si è evidenziato nel corso della trattazione, i carteggi d'artista (e la scoperta di nuovi carteggi) possono essere di grandissimo aiuto per datare in modo corretto un'opera d'arte: dando ai musei la possibilità di accedere a una pubblicazione telematica di un carteggio (magari scoperto di recente), è possibile dar loro l'opportunità di aggiornare in modo rapido e senza costi aggiuntivi il percorso espositivo, e un aggiornamento rapido e costante è un importante biglietto da visita per il museo stesso nonché un buon investimento in termini di immagine.

9.3 Per gli organizzatori di mostre

Un discorso simile a quello affrontato nel paragrafo precedente può essere preso in considerazione anche per quanti organizzano e allestiscono mostre ed esposizioni temporanee.

Le mostre temporanee rappresentano un importante (e spesso unico!) motivo di approfondimento dell'opera di un artista, dato che “riunire, da collezioni pubbliche e private di tutto il mondo, un gran numero di opere dipinte da un artista nel corso della sua carriera ci permette di esaminare il suo sviluppo [...] con un'accuratezza che né lui né i suoi mecenati potevano permettersi”²³².

Non bisogna poi trascurare il fatto che le esposizioni temporanee rappresentano una grande risorsa economica, utile per la promozione non solo dell'arte, ma anche del turismo. Risulta quanto mai importante promuovere in maniera adeguata l'arte e il turismo culturale, visto che nel 2007 solo il 34% degli italiani al di sopra dei 15 anni di età ha visitato almeno un museo o una mostra²³³: un dato che pone l'Italia tra gli ultimi paesi dell'Unione Europea (ai primi tre posti figurano Danimarca, Paesi Bassi e Svezia con le percentuali di 65%, 62% e 62%, mentre la media dell'Unione Europea è del 41%).

Se per quanto riguarda l'allestimento del percorso espositivo l'aiuto offerto da un'edizione telematica di carteggi d'artista può essere simile a quello che può offrire per un museo,

231 Maria Laura Tomea Gavazzoli, *Manuale di Museologia* (Milano: Etas, 2003), 113 – 114.

232 Francis Haskell, *The Ephemeral Museum. Old Master Paintings and the Rise of the Art Exhibition* (New Haven, Connecticut: Yale University Press, 2000), trad. it. *La nascita delle mostre* (Milano: Skira, 2000), traduzione di Federica d'Amiraglio e Roberta D'Adda, 15.

233 Fonte: Commissione Europea, Eurobarometer, citato nell'indagine Istat “Aspetti della vita quotidiana”. La scheda “Visitatori di musei e mostre” è disponibile all'indirizzo [http://noi-italia.istat.it/index.php?id=7&user_100ind_pi1\[id_pagina\]=41](http://noi-italia.istat.it/index.php?id=7&user_100ind_pi1[id_pagina]=41).

9. Possibili utilizzi

bisogna considerare il fatto che le mostre implicano anche la redazione di un catalogo che possa tenere memoria dell'esposizione stessa, conservando saggi inerenti le materie della mostra nonché le schede delle opere esposte.

La realizzazione del catalogo della mostra comporta diverse fasi: è necessario stabilire un piano, individuare gli autori dei saggi, redigere le schede delle opere esposte, controllare la redazione del volume, nonché occuparsi di operazioni “pratiche” come scegliere l'editore e stipulare con esso il contratto, stabilire il numero di cartelle per i saggi e i termini di consegna, decidere il compenso economico per gli autori²³⁴. Disporre di una pubblicazione telematica dei carteggi è un'opportunità che può essere sfruttata non soltanto dagli autori dei saggi durante le fasi redazionali, poiché avrebbero a disposizione in rete materiale prezioso per le ricerche, ma anche per quanti siano incaricati di dover redigere le schede delle opere. Un carteggio a documentazione di un'opera è una fonte preziosa per dare al visitatore della mostra (e probabile lettore del catalogo) utili indicazioni storiche e storico-artistiche sulle opere esposte. Il riferimento è sempre Carlo Finelli: non sarebbe stato possibile datare con sicurezza alcune delle sue opere senza la scoperta dei suoi carteggi.

Quindi, una pubblicazione telematica di carteggi d'artista consente un notevole risparmio di tempo nonché un grande vantaggio economico per chi organizza mostre: avendo a disposizione in rete le lettere, magari con il valore aggiunto delle immagini dei manoscritti originali e di un'edizione critica a supporto, è possibile risparmiare onerose ricerche presso fondi e archivi, che spesso costringono a lunghe e durature trasferte.

9.4 Per i linguisti

Una pubblicazione telematica di carteggi d'artista può essere una risorsa in più per lo studio della lingua. Si è già dato ampio spazio all'importanza che i corpora rivestono negli studi linguistici²³⁵, ma può essere importante sottolineare come la metodologia qui proposta possa fornire un supporto in più anche per i linguisti.

È sempre utile avere a disposizione fonti in più per i propri studi, e disporre di pubblicazioni di carteggi d'artista in rete, facili da utilizzare e da studiare, complete, di qualità e gratuite non può che essere uno stimolo anche per i linguisti.

234 Alessandra Mottola Molfino, Cristiana Morigi Govi, *Lavorare nei Musei* (Torino: Umberto Allemandi & C., 2004), 123 – 139.

235 Cfr. *supra* CAP. 3.

9. Possibili utilizzi

Basti pensare a studiosi che vogliono condurre ricerche sulle varietà linguistiche, sul linguaggio tecnico utilizzato dagli artisti, sul loro stile di scrittura: le pubblicazioni telematiche di carteggi d'artista potranno essere quindi un importante strumento di analisi per i linguisti.

9.5 Per docenti e studenti

Nonostante siano chiari i vantaggi che le tecnologie (in particolare l'e-learning) possano offrire a docenti e studenti, si nota purtroppo che esistono ancora diverse resistenze da parte dei docenti all'impiego di supporti didattici che vadano al di là dei tradizionali libri di testo²³⁶. Sono diversi gli elementi che inducono i docenti a resistere al cambiamento: tra i fattori, i docenti riferiscono la mancanza delle necessarie competenze informatiche e di utilizzo di internet e la mancanza di modelli didattici²³⁷. Si notano altresì una mancanza di consapevolezza sulle possibilità offerte dalla tecnologia²³⁸ nonché difficoltà di adattamento (abitudine all'aula) e timore della novità²³⁹.

Si pensi però a quale può essere l'utilità di un efficace sistema di e-learning opportunamente integrato con contenuti che sarebbe difficile trattare in modo ampio a lezione: tra questi contenuti è possibile includere anche le i corpora digitali di carteggi d'artista.

La natura ipertestuale e multimediale della rete dà luogo a opportunità di studio inimmaginabili fino a pochi anni fa: attraverso la rete è possibile esplorare e trovare riscontro a quanto spiegato dal docente durante la lezione, è possibile approfondire quello che si è detto a lezione, è possibile far nascere percorsi diversi rispetto a quelli battuti nella classica lezione frontale in aula, è possibile anche sviluppare nuove conoscenze e nuovi interessi. Anche in quest'ottica vogliono porsi le pubblicazioni telematiche dei carteggi d'artista.

Si pensi a un docente di storia dell'arte che durante una lezione spiega un'opera di Antonio Canova ai suoi studenti: con uno strumento come quello che qui si vuole proporre, è possibile controllare in tempi rapidissimi se esistono riscontri all'interno dei carteggi dell'artista, è possibile conoscere più da vicino il pensiero dello scultore, è possibile scoprire in modo più

236 Cfr. *E-learning/università: esperienze, analisi, proposte*, a cura di Cesare Scurati (Milano: Vita e Pensiero, 2005).

237 Cesare Scurati et al., "eUniversity. Modelli di innovazione didattica a confronto" in *Ontologie, simulazione, competenze*, a cura di Nicola Paparella (Castrignano dei Greci: Amaltea, 2007), 167.

238 *Ibid.*

239 Filippo La Noce, *E-learning. La nuova frontiera della formazione* (Milano: Franco Angeli, 2002), 190.

9. Possibili utilizzi

dettagliato gli eventi che hanno dato luogo alla nascita di un'opera o che hanno accompagnato la sua creazione. Questo grazie anche all'aiuto degli strumenti di ricerca, che possono essere utili per trovare qualsiasi forma e qualsiasi sequenza all'interno delle lettere, e grazie ai contesti KWIC, che rappresentano uno dei notevoli punti di forza della metodologia proposta con il presente lavoro: attraverso il loro utilizzo è infatti possibile inquadrare in maniera ottimale le forme cercate nei rispettivi contesti e poter quindi iniziare l'esplorazione delle missive di un artista.

9.6 Per gli appassionati

Infine, è necessario sottolineare come una pubblicazione telematica di carteggi d'artista possa essere utile non soltanto per musei, curatori, storici dell'arte, docenti e studenti: può e deve essere anche una valida risorsa di approfondimento per tutte le persone appassionate d'arte. Si pensi a quante possono essere: fornendo loro un prodotto di qualità si offre anche l'opportunità di documentare meglio le proprie conoscenze, magari invogliandole a visitare musei e mostre per vedere dal vivo le opere d'arte di cui si parla nei carteggi, trovando, attraverso la visione dell'opera, riscontri a quanto letto in rete.

Un corpus digitale di carteggi d'artista può quindi avere anche un importante ruolo promozionale: la pubblicazione, se adeguatamente pubblicizzata, può spingere molti appassionati a documentarsi sugli artisti che hanno prodotti i carteggi e li può spingere a visitare le loro opere dal vivo. Del resto, uno degli scopi principali dell'informatica per i beni culturali è proprio questo: invitare le persone ad andare ai musei e alle mostre per osservare le opere. È necessario comprendere come l'informatica per i beni culturali possa costituire un nuovo linguaggio di comunicazione, che se sfruttato nel migliore dei modi può portare grandi risultati.

Infine, si pensi al grande vantaggio per tutti coloro che, non essendo specialisti, non avrebbero accesso ai manoscritti: una pubblicazione telematica dei carteggi d'artista è anche un mezzo per fare in modo che la cultura possa veramente giungere a tutti, in modo gratuito, senza discriminazioni e con piena libertà di accesso.

Conclusioni

Il presente lavoro è partito con un chiaro obiettivo: proporre uno standard per creare corpora di carteggi d'artista utili, di qualità, usabili e accessibili, adatti per studiosi, storici dell'arte, ricercatori, docenti ma anche per studenti e appassionati d'arte. Si trattava quindi di cercare di realizzare un sistema che unisse rigore e precisione a versatilità, rapidità e facilità d'utilizzo.

L'obiettivo può dirsi raggiunto, e ciò è dimostrato anche da *Carlo Finelli Corpus*, il primo corpus di carteggi d'artista realizzato secondo la proposta metodologica discussa in questo lavoro: *Carlo Finelli Corpus* è stato creato proprio per avere un esempio concreto di ciò che è possibile ottenere seguendo le linee dettate dal presente lavoro.

Il progetto è intuitivo, facile da usare e soprattutto è completo: permette la lettura integrale dei testi delle singole lettere e soprattutto permette di eseguire sul corpus diversi tipi di ricerca, offrendo di fatto una gamma di opportunità che vanno al di là di quanto realizzato finora nell'ambito della storia dell'arte.

Si tratta tuttavia di un corpus di ridotte dimensioni (5.634 occorrenze), quindi si potrebbe pensare che il sistema non sia performante con corpora di dimensioni maggiori. A puro scopo dimostrativo sono stati condotti due esperimenti, uno su circa un milione di occorrenze e l'altro su circa dodici milioni di occorrenze. La base di dati è stata poi installata sullo stesso server che ospita *Carlo Finelli Corpus*: i risultati sono stati ottimi e le ricerche, anche di migliaia di forme, sono state eseguite in pochi secondi, in tempi del tutto ragionevoli.

Il progetto vuole inoltre essere una chiara e tangibile dimostrazione di quello che l'informatica umanistica in generale e l'informatica per i beni culturali in particolare consentono di ottenere: si tratta di un messaggio rivolto in modo particolare a quei docenti e a quegli studiosi che ancora sottovalutano il potenziale delle nuove tecnologie, in particolare di quelle *web-based*. È necessario sperimentare diversi approcci alla ricerca e allo studio, cercando di non temere la novità.

L'esigenza di cercare di creare una proposta metodologica standard è nata dal fatto che nel campo della storia dell'arte ogni progetto è stato finora realizzato secondo schemi metodologici propri, e può essere quindi interessante disporre di uno schema condiviso che possa facilitare la costruzione e la realizzazione di corpora di carteggi d'artista.

Bisogna poi considerare un ulteriore aspetto importante, ovvero quello della fruibilità dei carteggi. Molti documenti sono accessibili soltanto a pochi studiosi, e con il lavoro qui

Conclusioni

proposto si vuole dare uno stimolo affinché un sempre maggior numero di fonti finora consultabili solo da pochi possa, attraverso la rete, “aprirsi” al pubblico e diventare patrimonio di tutti. E, dal momento che tutti devono avere accesso a tali fonti e devono farlo in modo facile e intuitivo, una fase importante del lavoro è stata la riflessione sul modo in cui realizzare un progetto che potesse dirsi utile a più categorie e che potesse risultare usabile e accessibile.

Anche per questo sono state indicate linee guida per creare un prodotto conforme agli standard di usabilità e di accessibilità: questo bisogno è stato avvertito anche per il fatto che molti dei siti analizzati hanno dimostrato di non curare a sufficienza questi aspetti. E si tratta di un errore, perché è doveroso che un sito, soprattutto se vuole fare cultura, non debba escludere nessuno dalla sua completa fruizione.

La realizzazione di questa proposta è stata poi particolarmente complicata in quanto è stato necessario adottare tre punti di vista diversi, ovvero quello dell'informatico, quello del linguista e quello dello storico dell'arte, cercando di far sì che nessuno dei tre prevalesse ma che, al contrario, potessero completarsi a vicenda.

Anche nell'approccio multidisciplinare è possibile trovare uno dei punti di forza del lavoro: è stato possibile utilizzare gli strumenti messi a disposizione dall'informatica, dalla linguistica e dalla storia dell'arte per creare un prodotto completo e di qualità. Una qualità che è garantita da diverse caratteristiche: il rigore con il quale sono state condotte le ricerche, la sequenzialità delle fasi di lavoro, la facilità con il quale è possibile utilizzare il sistema (sia da parte degli operatori che devono creare il corpus, ai quali viene richiesta una competenza minima in informatica, sia da parte degli utilizzatori finali), la completezza dei diversi sistemi di ricerca, la possibilità di costruire intorno al corpus un sito che non sia soltanto un “contenitore”, ma che possa anche informare, fare cultura, avvicinare i visitatori alle opere di un artista, invitarli a osservarle dal vivo.

Si è notato che molti dei progetti finora realizzati non hanno considerato in modo adeguato le esigenze della linguistica: questa mancanza emerge soprattutto dall'osservazione dei motori di ricerca interni ai siti, che spesso sono incompleti e poco utili. Pertanto anche l'analisi dei portali già esistenti è stata fondamentale per ideare una proposta che potesse colmare le lacune riscontrate.

Si tratta quindi, nel campo della storia dell'arte, di un lavoro senza dubbio originale in quanto mette a disposizione della storia dell'arte stessa gli strumenti tipici della linguistica in

un contesto informatico: nessun sito tra quelli finora realizzati è stato pensato ragionando in quest'ottica.

È doveroso inoltre sottolineare che quando si cerca di elaborare uno standard il lavoro non può mai dirsi veramente “finito”. Quindi si tratta a tutti gli effetti di un lavoro “aperto”, realizzato anche in vista di ulteriori miglioramenti che potranno e dovranno essere apportati a questa proposta metodologica nel corso del tempo, qualora essa possa riscontrare successo presso coloro che si occuperanno della realizzazione di pubblicazioni telematiche di carteggi d'artista. Si è trattato quindi di tracciare un percorso, che solleva alcune problematiche (l'approccio ai carteggi, la necessità di ricerche complete, la facilità d'uso per studiosi non avvezzi alle tecnologie *web-based*) e che potrà essere sicuramente arricchito.

Durante l'introduzione si è detto che uno degli obiettivi del lavoro è far sì che possa diventare uno standard al quale attenersi: è ancora presto per dire se questo obiettivo, particolarmente ambizioso, potrà essere raggiunto. Per il momento è sufficiente augurarsi che questo lavoro possa costituire un valido contributo per la ricerca nel campo delle fonti per la storia dell'arte e aspettarsi che possa essere uno stimolo per la realizzazione di progetti di qualità, che possano soddisfare diverse esigenze, ma soprattutto che possano aggiungere un piccolo tassello alla costruzione di una cultura che possa veramente dirsi di tutti.

Bibliografia

Alla lettera. Teorie e pratiche epistolari dai Greci al Novecento. A cura di Adriana Chemello. Milano: Guerini, 1998.

Antichità classica. A cura di Luciana Aigner Foresti. Milano: Jaca Book, 1994.

Antognini, Roberta. *Il progetto autobiografico delle Familiars di Petrarca.* Milano: LED, 2008.

Aston, Guy e Lou Burnard. *The BNC Handbook.* Edinburgo: Edinburgh Press University, 1998.

Barocchi, Paola. "Fortuna dell'epistolografia artistica". In *Metodologia ecdotica dei carteggi: atti del Convegno internazionale di studi, Roma 23, 24, 25 ottobre 1980*, a cura di Elio d'Auria, 104 – 133. Firenze: Le Monnier, 1989.

Bentivogli, Bruno e Paola Vecchi Galli. *Filologia italiana.* Milano: Mondadori, 2002.

Biber, Douglas, Ulla Connor e Thomas A. Upton. *Discourse on the move: using corpus analysis to describe discourse structure.* Amsterdam: John Benjamins Publishing Co, 2007.

Biber, Douglas e Susan Conrad. "Lexical Bundles in Conversation and Academic Prose". In *Out of corpora: studies in honour of Stig Johansson.* A cura di Hilde Hasselgård e Signe Oksefjell, 181 – 190. Amsterdam: Rodopi, 1999.

Biber, Douglas, Susan Conrad e Randi Reppen. *Corpus linguistics: investigating language structure and use.* Cambridge: Cambridge University Press, 1998.

Billanovich, Giuseppe. *Lo scrittoio del Petrarca.* Roma: Edizioni di storia e letteratura, 1947.

Boccassini, Emanuela. "I carteggi femminili: opere letterarie e di informazione". *Ripensandoci*, n. 5 (Maggio 2009), http://www.editaonline.com/ripensandoci/index.php?option=com_content&view=article&id=1:carteggi-donne&catid=1:pensiero-e-azione-politica-delle-donne&Itemid=58 (controllato nel mese di marzo 2010).

Bozzi, Andrea. "Edizione elettronica e filologia computazionale". In *Fondamenti di critica testuale*, a cura di Alfredo Stussi, 207 – 234. Bologna: Il Mulino, 2006.

Brajnik, Giorgio ed Elio Toppano. *Creare siti web multimediali. Fondamenti per l'analisi e la progettazione.* Milano: Pearson Addison Wesley, 2007.

Cantoni, Lorenzo, Nicoletta Di Blas e Davide Bolchini. *Comunicazione, qualità, usabilità.* Milano: Apogeo, 2003.

Bibliografia

Capuozzo, Stefania. “L'epistolario di Tommaso Costo”. In AA. VV., *Gli scrittori d'Italia. Il patrimonio e la memoria della tradizione letteraria come risorsa primaria, XI Congresso dell'ADI, Napoli 26-29 Settembre 2007*. Grottammare: Graduus, 2008.

Carter, Matt. *Minds and computers: an Introduction to the Philosophy of Artificial Intelligence*. Edinburgo: Edinburgh Press University, 2007.

Cattaneo, Ruggero. “O stilu i kulturnom značenju Marulić eve Poslanice papi Adrijanu VI” [“Sullo stile e la rilevanza culturale dell'Epistola a Papa Adriano VI di Marco Marulić”]. *Colloquia Marulliana XVII* (17 Aprile 2008): 91 – 115. Abstract in italiano all'indirizzo http://hrcak.srce.hr/index.php?show=clanak&id_clanak_jezik=37658 (controllato nel mese di Marzo 2010).

Cazalé Berard, Claude e Raul Mordenti. “La costituzione del testo e la 'comunità degli interpreti'”. In *Internet e le muse. La rivoluzione digitale nella cultura umanistica*, a cura di Patrizia Bellmann Nerozzi, 13 – 39. Milano: Mimesis Edizioni, 1997.

Citroni Marchetti, Sandra. “Lo spazio straniato. Percorsi psicologici e percezione del tribunale nelle orazioni di Cicerone 'pro Fonteio', 'pro Q. Roscio comoedo', 'pro Cluentio'. *Materiali e discussioni per l'analisi dei testi classici XXXVI* (1996): 33-71.

Consegnare la memoria: manuale di archivistica ecclesiastica. A cura di Emanuele Boaga, Salvatore Palese, Gaetano Zito. Firenze: Giunti, 2003.

Delogu, Cristina, Andrea Bernardini, Daniela D'Aloisi, Raffaele Nicolussi e Susanna Ragazzini. “Accessibilità. L'apprendimento per tutti”. In *Tecnologie per il web learning: realtà e scenari*, a cura di Cristina Delogu, 119 – 136. Firenze: Firenze University Press, 2007.

De Santis, Paola. *I sermoni di Abelardo per le monache del Paracletto*. Leuven: Leuven University Press, 2002.

Donati, Claudio. “Aspetti istituzionali della Chiesa di Roma tra XVII e XVIII secolo”. In *Politica, vita religiosa, carità: Milano nel primo Settecento*, a cura di Marco Bona Castellotti, Edoardo Bressan e Paola Vismara, 103 – 128. Milano: Jaca Book, 1997.

Eckstein, Robert. *XML*. Tradotto da Eugenia Franzoni. Milano: Hops, 2000.

E-learning/università: esperienze, analisi, proposte. A cura di Cesare Scurati. Milano: Vita e Pensiero, 2004.

Galli, Sara. “Antifascisti tra carcere, clandestinità ed esilio. La 'riscoperta' dei carteggi”. *Storicamente* 1 (2004), http://www.storicamente.org/04_comunicare/rassegne/galli.htm.

Galluzzi, Paolo. “Nuove tecnologie e funzione culturale nei musei”. In *I formati della memoria. Beni culturali e nuove tecnologie alle soglie del terzo millennio*, a cura di Paolo Galluzzi e Pietro Valentino, 3 – 39. Firenze: Giunti, 1997.

Bibliografia

- Gigliozzi, Giuseppe. *Introduzione all'uso del computer negli studi letterari*. Milano:en Mondadori, 2003.
- González-Díaz, Victorina. *English adjective comparison: a historical perspective*. Amsterdam: John Benjamins Publishing Co, 2008.
- Grappone, Jennifer e Gradiva Couzin. *Search Engin Optimization: an hour a day*. Hoboken, New Jersey: Wiley, 2008.
- Grasso, Mario. *Scrivere per il web. Annotazioni, considerazioni e suggestioni*. Milano: Franco Angeli, 2005.
- Guercio, Maria. *Archivistica informatica: i documenti in ambiente digitale*. Roma: Carocci, 2004.
- Haseldine, Julian. "Epistolography". In *Medieval Latin: an introduction and bibliographical guide*, a cura di Frank Anthony Carl Mantello e A.G. Rigg, 650 – 658. Washington, D.C.: Catholic University of America Press, 1996.
- Haskell, Francis. *The Ephemeral Museum. Old Master Paintings and the Rise of the Art Exhibition*. New Haven, Connecticut: Yale University Press, 2000. Traduzione italiana: *La nascita delle mostre: i dipinti degli antichi maestri e l'origine delle esposizioni d'arte*. Milano: Skira, 2008. Tradotto da Federica d'Ammiraglio e Roberta D'Adda.
- Heilmann, Christian e Mark Norman Francis. *Web development solutions*. Berkeley, California: Friends of ED, 2007.
- Hickey, Raimond. *Corpus presenter: software for language analysis*. Amsterdam: John Benjamins Publishing Co, 2003.
- Il Cinquecento: la normativa e il contrario (1533 – 1573): le nuove regole e l'estensione dell'analogia*. A cura di Giovanni Da Pozzo. Padova: Piccin, 2006.
- King, Andrew B. *Website optimization*. Sebastopol, California: O'Reilly Media, 2008.
- La lettre à croisée de l'individuel et du social*. A cura di Mireille Bossis. Parigi: Éditions Kimé, 1994.
- La Noce, Filippo. *E-learning. La nuova frontiera della formazione*. Milano: Franco Angeli, 2002.
- Le "carte messaggiera". Retorica e modelli di comunicazione epistolare: per un indice dei libri di lettere del Cinquecento*. A cura di Amedeo Quondam. Roma: Bulzoni, 1981.
- Leech, Geoffrey. "The state of the art in corpus linguistics". In *English Corpus Linguistics: Studies in Honour of Jan Svartvik*, a cura di Karin Aijmer e Bengt Altenberg, 8 – 29. Londra: Longman, 1991.

Bibliografia

Lenci, Alessandro. “Le tecnologie della lingua per il Knowledge Management nelle piattaforme di e-learning”. *Wbt.it Web Magazine*, Dicembre 2004. http://www.giornalelearning.it/httpdocs/index.php?risorsa=tecnologie_della_lingua (controllato nel mese di marzo 2010).

Lenci, Alessandro, Simonetta Montemagni e Vito Pirrelli. *Testo e computer. Elementi di linguistica computazionale*. Roma: Carocci, 2005.

Lerdorf, Rasmus. *Php*. Sebastopol, California: O'Reilly, 2000. Traduzione italiana: *Php*. Milano: Hops, 2003. Tradotto da Marina Sturino.

Lettere di Lorenzo de' Medici detto il Magnifico nell'Archivio Palatino di Modena. A cura di Antonio Cappelli. Modena: Vincenzi, 1863.

Lie, Håkon Wium e Bert Bos. *Cascading Style Sheets: Designing for the Web – Third Edition*. Upper Saddle River: Addison-Wesley, 2005.

Manning, Christopher D. e Heinrich Schütze. *Foundations of statistical natural language processing*. Cambridge, Massachusetts: MIT Press, 1999.

McEnery, Tony e Andrew Wilson. *Corpus linguistics: an introduction*. Edinburgo: Edinburgh University Press, 2001.

McEnery, Tony, Richard Xiao e Yukio Tono. *Corpus-based language studies: an advanced resource book*. Londra: Routledge, 2006.

Mottola Molfino, Alessandra e Cristiana Morigi Govi. *Lavorare nei Musei*. Torino: Umberto Allemandi & C., 2004.

Morabito, Raffaele. *Lettere e letteratura. Studi sull'epistolografia volgare in Italia*. Alessandria: Edizioni dell'Orso, 2001.

Morley, Barry. “WebCorp: a tool for online linguistic information retrieval and analysis” In *The changing face of corpus linguistics*, a cura di Antoinette Renouf e Andrew Kehoe. Amsterdam: Rodopi, 2006.

Musetti, Barbara. *Carlo Finelli (1782 – 1853)*. Milano: Silvana Editoriale, 2002.

Nielsen, Jakob. *Designing Web Usability*. Basingstoke: Macmillan Computer Publishing, 2000. Traduzione italiana: *Web Usability*. Milano: Apogeo, 2000. Tradotto da Walter Vanini.

O'Keeffe, Anne, Michael McCarthy e Ronald Carter. *From corpus to classroom: language use and language teaching*. Cambridge: Cambridge University Press, 2007.

Osimo, Bruno. *Traduzione e nuove tecnologie*. Milano: Hoepli, 2001.

Bibliografia

Padula, Marco e Amanda Reggiori. *Fondamenti di informatica per la progettazione multimediale. Dai linguaggi formali all'inclusione digitale*. Milano: Franco Angeli, 2006.

Perosa, Alessandro e Paolo Viti. *Studi di filologia umanistica*. Roma: Edizioni di Storia e di Letteratura, 2000.

Pilgrim, Mark. *Dive into accessibility – 30 days to a more accessible web site*, 2002. <http://www.diveintoaccessibility.org/> (controllato nel mese di marzo 2010).

Piotrowski, Tadeusz. “The Translator and Polish-English Corpora”. In *Incorporating corpora: the linguist and the translator*, a cura di Gunilla Anderman e Margaret Rogers, 117 – 132. Clevedon: Multilingual Matters, 2007.

Polillo, Roberto. *Il check-up dei siti web*. Milano: Apogeo, 2004.

Polillo, Roberto. *Plasmare il web: road map per siti di qualità*. Milano: Apogeo, 2006.

Prandelli, Emanuela e Gianmario Verona. *Marketing in rete. Oltre Internet verso il nuovo marketing*. Milano: McGraw-Hill, 2006.

Pratellesi, Marco. *New journalism: teorie e tecniche del giornalismo multimediale*. Milano: Mondadori, 2004.

Prete, Cecilia. *Aperto al pubblico. Comunicazione e servizi educativi nei musei*. Firenze: Edifir, 2005.

Priebe, Tim. *Webifiable – Plan, create and mantain your web site the right way*. Edmond, Oklahoma: T&S Web Design, 2007.

Primo Conti – Giuseppe Raimondi: Carteggio. A cura di Patrizia Mania. Roma: Edizioni di Storia e Letteratura, 2001.

Ricciardi, Francesca. *Il viaggio dell'innovazione. Informatica, beni culturali e turismo*. Milano: Vita e Pensiero, 2005.

Salotti e ruolo femminile in Italia: tra fine Seicento e primo Novecento. A cura di Maria Luisa Betri ed Elena Brambilla. Venezia: Marsilio, 2004

San Bernardo e l'Italia: atti del convegno di studi, Milano, 24-26 maggio 1990. A cura di Paolo Zerbi. Milano: Vita e Pensiero, 1993.

Sapori, Giuliana. *Manuale di applicazione delle regole di catalogazione italiane al Servizio Bibliotecario Nazionale*. Segrate: Cilea, 2009. <http://manualesapori.cilea.it> (controllato nel mese di marzo 2010).

Bibliografia

Scurati, Cesare, Paolo Ardizzone, Monica Parricchi, Floriana Falcinelli, Paolo Calidoni e Patrizia Ghislandi, “eUniversity. Modelli di innovazione didattica a confronto”. In *Ontologie, simulazione, competenze*, a cura di Nicola Paparella, 139 – 172. Castrignano dei Greci: Amaltea, 2007.

Svartvik, Jan. “Corpus linguistics 25+ years on”. In *Corpus linguistics 25 years on*, a cura di Roberta Facchinetti, 11 – 26. Amsterdam: Rodopi, 2007.

Scrivere lettere. Tipologie epistolari nell'Ottocento italiano. A cura di Gino Tellini. Roma: Bulzoni, 2002.

Tomea Gavazzoli, Maria Laura. *Manuale di Museologia*. Milano: Etas, 2003.

Topolski, Jerzy con la collaborazione di Raffaello Righini. *Narrare la storia: nuovi principi di metodologia storica*. Milano: Mondadori, 1997.

Valacchi, Federico. “Applicazioni tecnologiche agli archivi storici. Le risorse per la descrizione, il riordino e la utilizzazione delle fonti archivistiche”. In *La Gestione dell'Archivio nelle Organizzazioni Non Profit*, a cura di Domenico Muscò, 67 – 82. Firenze: CESVOT, 2002.

Valacchi, Federico. *La memoria integrata nell'era digitale. Continuità archivistica e innovazione tecnologica*. Corazzano: Titivillus, 2006.

Venturi, Giulia. *Rassegna comparativa degli schemi di annotazione morfosintattica per la lingua italiana*. Roma: Università di Roma Tre, 2009.

Wilcock, Graham. *Introduction to Linguistic Annotation and Text Analytics*. Londra: Morgan & Claypool, 2009.

Siti web

Tutti i siti sono stati controllati nel mese di marzo 2010.

Accademia delle Scienze di Torino, www.accademiadelle scienze.it

Accademia di Brera, www.accademiadibrera.milano.it

Accademia di Belle Arti di Venezia, www.accademiavenezia.it

Accademia Nazionale dei Lincei, www.lincci.it

Archivio Storico del Comune di Ferrara, <http://servizi.comune.fe.it/index.phtml?id=3383>

BoLC (*Bononia Legal Corpus*), http://corpora.dslo.unibo.it/bolc_ita.html

British National Corpus, www.natcorp.ox.ac.uk

Carlo Finelli Corpus, <http://carlofinellicorpus.losnaweb.com>

CHILDES (*Child Language Data Exchange System*), <http://childes.psy.cmu.edu>

CORIS (*Corpus di Riferimento per l'Italiano Scritto*),
http://corpora.dslo.unibo.it/coris_ita.html

Corpus Informatico Belloriano, www.biblio.signum.sns.it/bellori

Corso di Laurea in Informatica Umanistica, Università di Pisa – www.unipi.it

DALF (*Digital Archive of Letters in Flanders*), <http://www.kantl.be/ctb/project/dalf>

Documents from the Archivio di Stato di Roma, <http://www.nga.gov/casva/accademia>

Evalita, <http://evalita.fbk.eu>

Fondazione Cassa di Risparmio di Carrara, www.fondazioneccrcarrara.com

Fondazione Memofonte, www.memofonte.it

FO.S.C.A. (*Fonti per la Storia della Critica d'Arte*), www.fosca.unige.it

GattoWeb, <http://gattoweb.ovi.cnr>

Italian NLP – Strumenti per il Trattamento Automatico della Lingua,
<http://foxdrake.ilc.cnr.it/webtools/>

LIP (*Lessico di Frequenza dell'Italiano Parlato*),
http://languageserver.uni-graz.at/badip/badip/20_corpusLip.php

Siti web

Losna – Soluzioni per il Web, www.losnaweb.com

Medici Archive Project, www.medici.org

MySQL, www.mysql.com

Nietzsche Source, www.nietzschesource.org

Opera del Vocabolario Italiano, www.oivi.cnr.it

Parlare Italiano, www.parlareitaliano.it

PHP, <http://php.net>

RIALTO (*Repertorio Informatizzato dell'Antica Letteratura Trobadorica e Occitana*),
www.rialto.unina.it

TagPro, <http://textpro.fbk.eu/>

TEI (*Text Encoding Initiative*), www.tei-c.org

Vincent Vang Gogh – The Letters, www.vangoghletters.org

W3C, www.w3.org