How to reference this article
Ferrari, G. (2020). Digital Humanities: alcuni tratti caratteristici. *Italica Wratislaviensia*, 11(1), 11–29.
DOI: http://dx.doi.org/10.15804/IW.2020.11.1.01

Saggio introduttivo

Giacomo Ferrari
Università del Piemonte Orientale "A. Avogadro" – Alessandria, Novara, Vercelli, Italy
giacomo.ferrari@uniupo.it
ORCID: 0000-0002-6501-7845

DIGITAL HUMANITIES: ALCUNI TRATTI CARATTERISTICI

DIGITAL HUMANITIES: SOME DISTINCTIVE FEATURES

Abstract: Rather than reporting on original research, this paper seeks to define the complex and rather diffuse domain of digital humanities by examining the historical and technological origins of the discipline. The distinction between the practice of the computer-mediated storage and retrieval of data relevant to human artefacts and the creative building of 'digital culture' draws a rough dividing line across the objectives of digital humanists. A historical outline of the distant origins of digital humanities suggests that the discipline is foundationally and intrinsically linked to computational linguistics and the development of linguistic resources. The boundaries of the discipline have been shifting concomitantly with the broadening of the scientific horizon and the evolution of dedicated technologies. Text mark-up (stemming from text annotation) and the multimodal facilities offered by ordinary browsers are the two basic techniques which have promoted the progressive development and expansion of digital humanities. These two techniques are closely interconnected as the language operated by the http protocol (HyperText Transfer Protocol) derives from the same source as that used for text mark-up. Hypertext and multimodality allow extending the uses of the computer to store and access humanities data of various kinds, including images, videos and sound recordings. Finally, the declaration of entities, as a further development of mark-up, makes it possible to apply semantic web techniques to carry out advanced research studies. The field of creative digital culture is very large, and there are abundant software applications that support such creative pursuits. Consequently, several forms of art have largely profited from technological advancement. Given this, the paper also addresses technological obsolescence as a serious problem in digital humanities.

Keywords: digital humanities, computational linguistics, linguistic resources, mark-up, technology for humanities

Received: 25/01/2020; Accepted: 15/06/2020; Published: 31/07/2020
ISSN 2084-4514 e-ISSN 2450-5943

INTRODUZIONE

Il campo delle Digital Humanities è estremamente vasto e ben lontano dall'essere definito con precisione. Lo statuto dell'Associazione per l'Informatica Umanistica e la Cultura Digitale (AIUCD) indica che lo scopo è "promuovere e diffondere la riflessione metodologica e teorica, la collaborazione scientifica e lo sviluppo di pratiche, risorse e strumenti condivisi nel campo dell'informatica umanistica e nell'uso delle applicazioni digitali in tutte le aree delle scienze umane, nonché favorire inoltre la riflessione sui fondamenti umanistici delle metodologie informatiche e nel campo delle culture di rete"[1]. Si tratta di una definizione assai vaga da cui emerge una molteplicità di obiettivi che è difficile ricondurre ad unità. Da un lato si rinvia all'utilizzo di applicazioni digitali in tutte le aree delle scienze umane, dall'altro si promuove la riflessione su ciò che viene definito "cultura di rete". Nel primo caso, quindi, la definizione è rinviata ai singoli campi disciplinari, che abbiano avuto una qualche forma di intersezione con le tecnologie informatiche e digitali. Nel secondo, sembra piuttosto di confrontarci con la reazione degli umanisti e dei produttori di cultura, abituati, spesso, a tecniche individuali, di fronte alle potenzialità di una tecnologia che ha un impatto significativo sia sui metodi di lavoro sia, più generalmente, sul modo di "fare cultura". Questa stessa dicotomia appare chiaramente in un interessante articolo di Patrik Svensson[2], in cui, dopo aver compiuto un'interessante rassegna delle imprese umanistiche digitali più rilevanti, focalizza la specificità della "metodologia digitale" e della sua capacità di creare cultura in due aspetti complementari, l'apertura dei dati all'intera comunità scientifica e il metodo cooperativo di ricerca.

Obiettivo di questo articolo è identificare alcune linee ideali che unificano le diverse tendenze che vanno sotto l'etichetta "digital humanities", "informatica umanistica", "umanistica digitale" e numerose altre, spesso ideate per cercare di creare diversificazioni più sottili di quanto sarebbe necessario. In particolare, una breve incursione nelle origini

[1] Citato sul sito dell'Associazione stessa (www.aiucd.it).

[2] Svensson, 2010, completo di ampia bibliografia sull'argomento.

dell'uso del calcolatore in domini umanistici può aiutarci a comprendere sia la nozione di metodo cooperativo e di risorsa comune, sia a identificarne le radici tecnologiche e la loro influenza sulle metodologie.

1. I PRIMORDI

1.1. Difficoltà di una definizione

Nel 1966 Joseph Raben fondò un giornale intitolato *Computer and the Humanities* che aveva lo scopo di pubblicare studi e ricerche nell'ambito dell'applicazione di metodi computazionali allo studio delle discipline umanistiche (*humanities*). Nella presentazione della rivista, Raben dà una definizione della disciplina:

> We define humanities as broadly as possible. our interests include literature of all times and countries, music, the visual arts, folklore, the non-mathematical aspects of linguistics, and all phases of the social sciences that stress the humane. when, for example, the archaeologist is concerned with fine arts of the past, when the sociologist studies the non-material facets of culture, when the linguist analyzes poetry, we may define their intentions as humanistic; if they employ computers, we wish to encourage them and to learn from them. [Prospect, 1966: 1, citato da Terras et al., 2013][3]

La definizione, quindi, è condizionata, fin dalle origini di questa disciplina, dalla definizione di *humanities,* un settore molto vasto e differenziato sia negli oggetti di studio, sia nelle metodologie d'indagine. In ogni caso, il testo è fortemente orientato verso lo sviluppo di tecniche di utilizzo del calcolatore nelle tematiche umanistiche, ignorando, invece, la prospettiva della "cultura di rete". In questa definizione originaria è difficile identificare elementi unificanti, se non l'uso del calcolatore che s'interseca con i problemi ed i metodi caratteristici delle scienze umane. Un confronto tra le diverse definizioni che vengono date di questo settore (vedi Lugmayr & Teras, 2015) confermano questo punto. Col tempo, l'evolvere del campo di ricerca e delle tecnologie disponibili ha conferito al settore una sua specificità sempre più identificabile.

[3] La minuscola dopo il punto è una scelta editoriale dell'autore.

1.2. Dal calcolatore al testo

Fin dalla prima apparizione del *computer*, nei tardi anni 40, molti studiosi ritennero che uno dei campi di applicazione più promettenti fosse quello del linguaggio naturale, ma con altrettanta prontezza il campo si divise in due settori. Da un lato, molti studiosi seguirono gli spunti teorici del matematico inglese Alan Turing secondo cui il lavoro dell'intelligenza umana può essere rappresentato dal funzionamento di un automa, l'*automa di Turing* (Turing, 1937, 1938); cercarono, quindi, di usare il computer per simulare le prestazioni dell'intelligenza umana, includendo la comprensione del linguaggio naturale. Dall'altro lato, Padre Roberto Busa SJ[4] intuì da subito le potenzialità del computer come memorizzatore e manipolatore di dati; dopo un lavoro di decenni produsse l'indice complessivo delle opere di San Tommaso (*Index Thomisticus,* oggi accessibile al sito https://www.corpusthomisticum.org/it/index.age). Quest'impresa costituì l'atto di nascita della lessicografia computazionale, che produsse in seguito molti risultati come la produzione di *concordanze* (cioè la lista delle parole di un testo, ciascuna associata a tutti i contesti in cui appare) o i lavori statistici[5]. La caratteristica di questo settore è che il calcolatore viene usato non come simulatore, ma come strumento per la memorizzazione e la manipolazione di dati linguistici e testuali, tecnicamente dati *alfanumerici*. Fu su questa linea che l'Accademia della Crusca cominciò ad utilizzare il computer per elaborare quei dati che dovevano confluire nel grande dizionario,

[4] Curiosamente sia la prima idea di utilizzare il calcolatore per produrre il lessico delle opere di San Tommaso sia quella di usarlo per tradurre testi da una lingua ad un'altra risalgono allo stesso anno, il 1949. In quell'anno, infatti, Padre Busa propose a Thomas Watson, presidente della IBM, il progetto dell'*Index Thomisticus* (Busa, 1949), mentre Warren Weaver propose, nel *Memorandum on Translation* indirizzato alla Rockefeller Foundation (Weaver, 1949), l'idea della traduzione automatica.

[5] Né le concordanze né la statistica linguistica sono necessariamente associate all'uso del computer. Le prime concordanze furono quelle bibliche sia latine che ebraiche che risalgono al basso medioevo (Ferrari, 2010), mentre i lavori statistici affondano le radici nelle ricerche di Herdan (1956, 1960), compiute in modo totalmente manuale. Naturalmente il calcolatore ha aggiunto la capacità di compiere tali operazioni in tempi brevi e per grandi masse di dati.

che una simile decisione fu adottata dalla Real Academia Española, che nacque il *Trésor de la Langue Française informatisé*[6]. L'evoluzione dei mezzi tecnologici ha poi stimolato la creazione di grandi *corpora* come il *British National Corpus* o il *Corpus of Contemporary American English (COCA)*.

1.3. Dal testo al corpus

La nozione di *corpus* si differenzia dalle prime imprese lessicografiche per l'approccio tecnologico e gli obiettivi. La lessicografia computazionale utilizza un insieme finito di testi da cui estrae un vocabolario di struttura tradizionale; il *corpus* da cui vengono tratte le schede lessicografiche resta generalmente in ombra, se non del tutto inaccessibile. Questo approccio ha fornito importanti prodotti come dizionari fondamentali, cioè dizionari estratti da campioni rappresentativi della lingua standard[7], dizionari di frequenza[8], lessici settoriali[9], ma il tratto fondamentale che unifica tutte le elaborazioni è la metodologia di raccolta del materiale. Fin dagli inizi gli studiosi furono coscienti che, dato il dispendio di risorse e di energie per la memorizzazione dei materiali testuali, sarebbe stato necessario concepire tale operazione come la creazione di un deposito di dati da cui estrarre il prodotto lessicografico, ma anche da mettere a disposizione di future ricerche, anche non previste nel progetto originale. S'impose, quindi, la necessità di scegliere una metodologia il più possibile neutrale rispetto alla rappresentazione dei dati. Si fa strada, così, l'idea che, al di là del prodotto lessicografico, il materiale testuale stesso sia una risorsa messa a disposizione di tutta la comunità scientifica. Questo passaggio dà origine alla nozione di *corpus*, introdotta sopra. Negli esempi citati, come in altri, il *corpus* diviene un insieme

---

[6] Grande impresa lessicografica promossa principalmente da Bérnard Quémada.

[7] Come il LIF (Lessico dell'Italiano Fondamentale) creato da Tagliavini, Bortolini e Zampolli nel 1972.

[8] Alphonse Juilland produsse tra il 1964 e il 1973 dizionari di frequenza dello Spagnolo, del Romeno, del Francese e dell'Italiano.

[9] Sia nel senso di dizionari terminologici di un determinato settore, sia nel senso di lessici di un particolare diasistema, come il LIP (Lessico dell'Italiano Parlato) di De Mauro *et al*., 1993.

aperto, costantemente aggiornabile ed utilizzabile dai ricercatori che ne facciano richiesta, senza che ne venga derivato un prodotto lessicografico. In questo modo la raccolta di dati linguistici diviene una "risorsa linguistica".

2. RISORSE (LINGUISTICHE)

La nozione di "risorsa linguistica" viene finalmente codificata negli anni 90 e formalizzata nel corso della prima International Conference on Language Resources and Evaluation (LREC), tenutasi a Granada nel 1998. Nell'invito alla conferenza si forniva la seguente definizione:

> The term language resources (LR) refers to sets of language data and descriptions in machine readable form, used specifically for building, improving or evaluating natural language and speech algorithms or systems, and in general, as core resources for the software localization and language services industries, for language studies, electronic publishing, international transactions, subject-area specialists and end users. Examples of linguistic resources are written and spoken corpora, computational lexicons, grammars, terminology databases, basic software tools for the acquisition, preparation, collection, management, customization and use of these and other resources.

Il tratto originario della produzione di tali materiali di studio è l'attenzione che non è concentrata solo sul dato come insieme di contenuti, ma anche sulla forma in cui tale dato si presenta e al contesto in cui si colloca. L'idea di memorizzare dei testi in una maniera che li rendesse accessibili, ma anche disponibili per iniziative editoriali, risale, probabilmente, alla Text Encoding Initiative, nata durante una riunione tenutasi presso il Vassar College nel 1987, con lo scopo di armonizzare le diverse tecniche di registrazione dei testi. Il sito ufficiale dell'organizzazione (https://tei-c.org) fornisce indicazioni che devono servire come standard e che evidenziano l'attenzione non solo al testo, ma anche ai suoi dati strutturali ed esterni.

Infatti, l'approccio "umanistico" assegna un ruolo importante anche alla struttura formale del testo (divisone in capitoli, in canti, in versi ecc.), la sua collocazione spazio-temporale (edizione o edizioni, diverse

versioni e letture, altri aspetti filologici) e le sue versioni (lavoro tipicamente filologico).

Questa "svolta" metodologica costituisce la radice e il banco di prova di due nozioni che divengono fondamentali nelle *digital humanties*, quella di risorsa, intesa come raccolta di dati aperti ad una intera comunità scientifica per ricerche particolari, che possono includere anche l'arricchimento ed il completamento della risorsa stessa, e quella più limitata di rappresentazione fedele della forma e della tradizione del testo stesso. Da quest'ultima idea si sviluppa l'intera branca della Filologia Computazionale[10].

3. ESTENSIONE DELLE RISORSE

È difficile stabilire con sicurezza se l'idea di creare vasti archivi di dati relativi alle scienze umane sia dovuta ad un estendersi dei paradigmi elaborati nell'ambito della creazione di risorse testuali o si sia sviluppata nei diversi settori autonomamente, giungendo progressivamente a convergenza. Certo è che nell'ambito dei diversi settori applicativi si diffondono, poi, metodologie specifiche di elaborazione. L'obiettivo finale di tutti i settori applicativi è la memorizzazione di dati relativi ai più disparati settori delle scienze umane e, soprattutto, la creazione di sistemi di accesso il più possibile aperti all'intera comunità di studiosi e flessibili; è questo il passo più significativo verso le *digital humanities* intese come fonte di una nuova cultura.

Oggi, infatti, è difficile separare la nozione di *digital humanities* dalla nozione di risorsa, cioè di raccolta dinamica di dati aperti alla comunità dei ricercatori. Un'evoluzione dei lavori di natura testuale e lessicografica sono le biblioteche di lingue classiche[11], mentre un caso particolare e relativamente ben radicato è rappresentato dai cataloghi

[10] Vedi Ferrari (2019).

[11] Ad esempio nell'ambito degli studi classici merita ricordare il sito che racchiude una grande quantità di opere latine (http://www.thelatinlibrary.com), quello che presenta molte opere sanscrite (http://www.sanskrit-linguistics.org), greche (https://www.perseus.tufts.edu) o in genere opere religiose orientali in diverse lingue classiche (https://www.sacred-texts.com/sbe/index.htm).

artistici ed archeologici, che includono schedari relativi ad oggetti antichi, ad opere d'arte, ma anche agli antichi cataloghi, come evidenziato dall'Istituto Centrale per il Catalogo e la Documentazione (http://www.iccd.beniculturali.it/it/home). Infatti, nella schedatura degli oggetti antichi e d'arte non è solo importante memorizzare le immagini, la descrizione e la bibliografia, ma anche le catalogazioni che sono state fatte precedentemente, soprattutto nel caso di acquisizione di collezioni private. Da queste iniziative, in genere dirette agli specialisti, si sviluppano le diverse idee di museo virtuale, che permette la visita, sempre virtuale, anche a normali visitatori.

Ne fanno parte importante anche le raccolte di materiale antropologico, come "I granai della memoria", collezione memorizzata presso l'Università di Scienze Gastronomiche di Pollenzo (https://www.granaidellamemoria.it/index.php/it), una raccolta antropologica di memorie, documenti, immagini relative alle tradizioni di alcune zone italiane e non, completate con memorie di vita di persone che vivono o hanno vissuto nelle stesse aree; in essa, quindi, convivono testi, musiche, filmati e immagini fisse.

A questa categoria appartengono anche le raccolte di canti popolari, come quella offerta da RAITeche (http://www.teche.rai.it/archivio-del-folclore-italiano), che affondano le radici in iniziative anteriori all'introduzione dei calcolatori[12].

Ma il settore principale rimane la raccolta di testi, la biblioteca digitale con tutte le sue derivazioni. Le ragioni di questa priorità sono due, il fatto che il trattamento dei testi è all'origine dello sviluppo del settore, e che la maggioranza delle informazioni che si trattano nell'Internet compaiono in forma testuale. Infatti, a dispetto delle dichiarazioni definitorie, le iniziative registrate presso le diverse associazioni di settore contano una notevole maggioranza di progetti di natura testuale, sia di memorizzazione sia di catalogazione[13].

---

[12] Una delle prime raccolte di canti popolari fu quella dei canti piemontesi di Costantino Nigra, raccolta iniziata nel 1855 e pubblicata nel 1888.

[13] Si veda, a titolo di esempio, il sito dell'Associazione Italiana di Informatica Umanistica e Cultura Digitale (AIUCD – http://www.aiucd.it).

4. EVOLUZIONI TECNOLOGICHE

Le applicazioni umanistiche hanno ricevuto un importante impulso da due innovazioni tecnologiche, l'uso di linguaggi di *markup* e la multimedialità offerta dalla struttura ipertestuale.

4.1. Il markup

Un linguaggio di *markup* è costituito da un insieme di "etichette" che permettono di inserire all'interno di un documento una serie d'indicazioni classificatorie che possono essere "lette" da un *browser* ed interpretate come indicazioni operative di vario livello, dalla formattazione alle informazioni grammaticali e oltre. Il primo linguaggio, *sgml* (Smith 1988), trova la sua applicazione più importante nella Text Encoding Initiative introdotta sopra. Un derivato dell'*sgml* è l'*html*, il linguaggio con cui si implementavano le pagine del web, successivamente evoluto in *shtml* o *xhtml*. Nel filone del mark-up di testi è stato poi sviluppato l'xml[14] che costituisce oggi lo standard di tutte le iniziative di annotazione testuale. Il markup, infatti, è l'aspetto implementativo dell'attività detta "annotazione", ma si estende presto al trattamento di dati di ogni genere.

Alle origini della disciplina, la preparazione di un testo prevedeva la memorizzazione del testo stesso con tutte le indicazioni contestuali chiare. Un programma successivo produceva i contesti[15] necessari per produrre le concordanze da cui si poteva poi produrre un lessico o un lessico statistico; infine si poteva procedere alla "lemmatizzazione", cioè il processo di riconduzione di una parola alla sua forma di dizionario. Il markup permette di realizzare, appunto, quella che si chiama *annotazione*. Possiamo illustrare con un esempio questa importante innovazione. Dato il primo verso della Divina Commedia *Nel mezzo del*

[14] Si veda www.w3schools.com.xml.

[15] I contesti venivano definiti in maniera rigida, cioè prendendo un certo numero di parole a destra e a sinistra della parola considerata; in alcune applicazioni più sofisticate si procedeva ad "aggiustare" il contesto in base ai segni d'interpunzione.

cammin di nostra vita, ci si aspetta che nelle concordanze appaia, ad esempio, la seguente linea

Cammino "nome maschile"
Nel mezzo del cammin di nostra vita

Utilizzando le vecchie procedure questo risultato poteva essere ottenuto attraverso la seguente serie di passi:

– estrazione della lista di parole (con o senza frequenze) dal testo
– lemmatizzazione, fase in cui si assegna il codice "nome maschile" e il lemma *cammino*
– lista dei contesti
– assegnazione di ogni contesto al lemma corrispondente; in questo modo il termine *cammin* insieme al suo contesto viene assegnato all'esponente *cammino*.

Al contrario con l'introduzione del *markup* si potrà indicare nel testo la categorizzazione delle parole, come segue:

<w1 cat=prepart lem=in>*nel*</w1><w2 cat=nome lem=mezzo>*mezzo*</w2><w3 cat=prepart lem=di>*del*</w3><w4 cat=nome lem=cammino>*cammin*</w4><w5 cat=prep lem=di>*di*</w5><w6 cat=poss lem=nostro>*nostra*</w6><w7 cat=nome lem=vita>*vita*</w7>.

In questo esempio sono state introdotte le etichette arbitrarie[16] w1, w2 ecc. per indicare l'unità "parola", la proprietà "cat", per la categoria lessicale con i valori "prepart" (preposizione articolata), "prep" (preposizione), "nome", "poss" (possessivo), e la proprietà "lem" per indicare l'esponente di dizionario.

L'assegnazione delle diverse categorizzazioni e dei lemmi corrispondenti può essere realizzata con una serie di programmi automatici. Nel nostro esempio, l'assegnazione delle categorie lessicali e del lemma vengono, in genere, eseguite da programmi di "POS tagging" (Part Of Speech tagging) che possono seguire diverse metodologie di assegna-

[16] Le etichette dell'xml possono essere totalmente arbitrarie in quanto ne vengono definiti il valore e le proprietà in una dichiarazione a cura dell'estensore (detta DTD).

zione[17]. Man mano che incontra le indicazioni, dette *tags*, un browser opportunamente istruito potrà compiere ogni sorta di operazioni che siano associate a tali codici, come assegnare un colore diverso per ogni categoria, costruire una lista di lemmi con i relativi contesti, calcolare le frequenze[18]. Naturalmente l'esempio è molto semplificato ma il meccanismo funziona come descritto. La prima conseguenza evidente è che mediante i *tags* si possono rappresentare tutti i tipi di informazione, come indicazioni di tipo grammaticale, sintattico o semantico, ma anche di tipo editoriale, come il *font* da utilizzare. La notazione si rivela, quindi, molto potente e capace di rappresentare tutte le caratteristiche del testo, sia relative alla sua forma, alla sua storia ed alle sue stratificazioni, sia relative ai suoi contenuti. Allo scopo di sottolineare la validità e l'importanza di dati testuali ben rappresentati in modo generale, vale la pena di menzionare che dall'Index Thomisticus è stata derivata una *treebank*, cioè un repertorio di alberi sintattici derivati dal testo originario[19].

4.2. La multimedialità

Il protocollo ipertestuale *http* permette di far convivere ed interagire *file* di tipo testuale, ma anche immagini, filmati o *file* audio, prestandosi così pienamente al trattamento di informazioni multimediali. È possibile, così, andare oltre nella rappresentazione di ogni genere di dato umanistico. Si possono rappresentare i manoscritti (attraverso immagini ad altissima risoluzione) a fianco della loro trascrizione, fotografie di oggetti d'arte o archeologici associate alle schede descrittive[20] e, come accade in molte aree, anche alle schede man mano prodotte durante la storia del

---

[17] Vedi ad es. De Rose (1988), Church (1988, 1990), Charniak (1997).

[18] L'interpretazione che il browser deve assegnare ai *tags* viene definita in un altro file separato, detto *stylesheet xsl* (https://www.w3schools.com/xml/xsl_intro.asp) ed interpretata da un programma in linguaggio php, asp o altro.

[19] Vedi Passarotti 2015. La *treebank* è un altro tipo di risorsa linguistica che consiste in una raccolta di alberi sintattici derivati da un testo; esempi illustri sono la PennTreebank (Taylor *et al.*, 2003, pp. 5–22) o la Prague dependency treebank (Bejček *et al*., 2013), ma ne esistono ormai molte.

[20] O anche repertori di epigrafi antiche (http://www.edr-edr.it/it/Link_it.php).

collezionismo. Questa struttura ha dimostrato la sua piena utilità nella costruzione, ad esempio, dei *Granai della memoria* (si veda il § 3).

4.3. Il semantic web

Una conseguenza significativa dell'uso dei linguaggi di *markup* è il *semantic web*. Si tratta della materializzazione dell'intuizione, già proposta da Barners-Lee (1999), che il calcolatore possa divenire "capable of analyzing all the data on the Web – the content, links, and transactions between people and computers". Gli antecedenti concettuali devono essere ricercati in tutte le teorie di rappresentazione della conoscenza, in particolare nella famiglia delle reti semantiche[21]. Tuttavia la realizzazione di quel sogno è stato reso possibile dall'introduzione, nella sfera dei linguaggi di *markup*, dell'annotazione dei meta-dati, fino all'introduzione di specifici linguaggi come RDF (Resource Description Framework) e OWL (Web Ontology Language), tutti utilizzabili nell'ambito della codifica xml e gestiti dal consorzio W3C[22] in un'unica struttura. In termini semplici, i meta-dati permettono di assegnare singoli dati, o classi di dati, a tipologie specifiche, creando così una gerarchia di concetti. RDF permette di connettere tra di loro diverse informazioni, dette "risorse", mentre OWL offre una tecnologia per stabilire relazioni concettuali simili a quelle delle reti semantiche[23]. Senza entrare nei dettagli tecnici, che sono ben rintracciabili sul sito del consorzio W3C[24], queste estensioni della codifica web permettono di condurre ricerche nelle varie raccolte rese disponibili seguendo dei percorsi che non sono soltanto di ricerca formale, ma anche concettuale.

---

[21] Vedi Berners-Lee *et al*., 2001; Hendler, van Harmelen, 2008.

[22] World Wide Web (W3) Consortium, è il gestore e garante degli standard del Web e dei servizi che rende disponibili.

[23] Sulla nozione informatica di ontologia si veda Gruber, 1993 o Guarino, Musen, 2015.

[24] Le descrizioni tecniche, complete di esercitazioni, sono presenti sui siti del W3C, https://www.w3schools.com, per quello che riguarda HTML, XML ed annessi, e http://w3schools.sinsixx.com/rdf/rdf_owl.asp.htm per quello che riguarda RDF e OWL.

5. LA CULTURA DIGITALE

È difficile definire gli aspetti culturali più generali delle *digital humanities* in quanto, come si è visto, non disponiamo di una definizione che ci aiuti. Abbiamo detto che nella letteratura internazionale si sottolinea la capacità di cooperazione e la larga accessibilità. Quest'ultima caratteristica emerge dal progredire degli aspetti tecnologici, quali la multimedialità e il *semantic web*, che permettono una grande flessibilità di accesso e la capacità di adattare gli algoritmi di ricerca alle esigenze di ogni singolo. La cooperazione, che si fonda anch'essa sulla base tecnologica, ha modificato anche la mentalità del ricercatore. La possibilità di accedere a risorse condivise ha prodotto, fin dagli anni 90, l'abitudine alla collaborazione sulle risorse stesse, concretizzatasi principalmente nell'annotazione condivisa, estesa poi a diversi settori (co-authoring, didattica ecc.). Questa consiste nel far sì che più di una persona possa operare su una stessa collezione di dati in modo condiviso e contemporaneo, sia in senso reale che in senso concettuale.

Resta dubbio se l'uso della tecnologia digitale come supporto alla creatività artistica possa considerarsi un atto di costruzione di cultura digitale. Il campo si estenderebbe, quindi, dall'uso di software dedicati alla creazione musicale (Digital Audio Workstation), ai supporti tecnologici per la grafica, e addirittura plastica (stampa 3D). A proposito della grafica, c'è da chiedersi se la rimarchevole crescita della produzione di *graphic novels* e *graphic journalism*, nonché l'evoluzione qualitativa dei fumetti, non sia il frutto di una felice convergenza tra uno straordinario progresso tecnologico e un'evoluzione significativa dei consumi. Portando la nozione di cultura digitale ai suoi confini estremi potremmo anche includere le diverse iniziative di creazione di film di animazione; questi, infatti, hanno completamente rivoluzionato la stessa concezione del film e della sua produzione, portando, in casi estremi all'uso marginale di attori, cui rubare il volto e il corpo per animarli, con tecniche diverse, senza ricorso alla loro presenza fisica[25]. L'uso di tecnologie di

[25] Dopo le prime sperimentazioni, condotte con diverse metodologie (si vedano i film Polar Express o Ratatouille) il settore si è espanso ad un punto tale che cercare di

supporto a forme d' arte tradizionali finisce col creare forme nuove di arte o, almeno, la ricerca di forme nuove, come accade per la musica elettronica o per la costruzione di componimenti letterari, originali o tradotti, usando i *social* o mediante *emoji*. Non sono forme espressive intrinseche al mondo digitale, in quanto non sono legate alle esigenze di calcolo, ma sono nate in quell'ambito ed hanno esportato un modo di vivere la creatività. È in quest'area di sovrapposizione che si crea la nuova cultura.

6. GUARDANDO AVANTI

Se allineiamo le caratteristiche che sono emerse durante la storia di questo complesso campo, si può chiarire la funzione di questo settore scientifico e provare anche a guardare al futuro della disciplina.

La caratteristica primaria è la fedeltà al dato originario e l'accessibilità concessa alla più ampia comunità di scienziati. La prima caratteristica si può descrivere, utopisticamente, come la possibilità per ogni utente (scientifico o comune) di accedere al dato come se si presentasse nella sua forma reale. Ad esempio, un epigrafista dovrebbe essere in grado non solo di vedere il testo trascritto di un'iscrizione, ma di vedere l'epigrafe stessa come se l'avesse davanti. In un futuro non tanto lontano, si potrebbe anche immaginare l'uso di realtà virtuale per consentire all'epigrafista di "toccare" con le sue dita l'epigrafe stessa, alla ricerca di trascrizioni diverse da quelle fornite dai suoi colleghi[26]. La fedeltà al dato permette di parlare di musei virtuali, in cui si può visitare un'area, un oggetto, un testo come se si fosse presenti di persona. Lo stesso vale per le biblioteche virtuali, nelle quali dovrebbe essere possibile sfogliare

---

darne uno stato dell'arte richiederebbe un lavoro enorme, anche soltanto per esaminare le diverse tecnologie. In questo settore sono nati colossi come la Pixar (https://www.pixar.com) o la Sony Pictures (http://www.sonypicturesanimation.com) e da alcuni anni si svolge a Torino la View Conference che ne costituisce un'importante rassegna (https://www.viewconference.it).

[26] Sono comunque già in uso tecniche di miglioramento della visione di un'epigrafe o di un manoscritto, usando luci radenti di varia frequenza, che renderebbero inutile toccare, sia pure virtualmente, il manufatto.

un libro in tutte le sue edizioni, ma, allo stesso tempo, compiere operazioni di confronto tra un'edizione e l'altra.

Il limite estremo di questa tendenza alla digitalizzazione potrebbe essere quanto descritto nel racconto di Borges "La mappa dell'impero", in cui l'imperatore richiede una mappa del suo impero in scala 1:1[27]. Nel testo letterario, la mappa finisce col coprire il territorio, soffocarlo e condannarlo alla carestia e alla morte. Al contrario, i dati digitali sono più compressi rispetto a quelli cartacei e l'accessibilità e la possibilità di cooperazione costituiscono il valore aggiunto di una tale iniziativa.

Naturalmente bisogna precisare che la decadenza del dato elettronico è più veloce di quella dei materiali originari (pergamena, carta, dipinti, marmo ecc.), sia in termini di deterioramento fisico che di obsolescenza digitale. Nessuna delle due cause è analizzata in modo esauriente, ma in genere s'ipotizza che il deterioramento fisico dei supporti digitali abbia un ciclo di un decennio o meno, il che impone di ricopiare i dati a cadenze fisse. L'obsolescenza, invece, è funzione delle innovazioni tecnologiche, cioè dell'evoluzione dell' *hardware* e del *software* che si utilizzano per accedere ai dati su supporto digitale. Il tempo in cui i vecchi *floppy disk* sono divenuti inaccessibili è abbastanza recente; è nota la fragilità dei *pendrive*. L'unica misura che può garantire una conservazione paragonabile a quella del dato originario è la costituzione d'infrastrutture *cloud*, che garantiscano la ritrascrizione di tutti i dati secondo cicli che ne garantiscano la perfetta conservazione; una soluzione che metterebbe l'integrità del dato (e della cultura che ne deriva) nelle mani di pochi gestori. Non mi spingo a immaginare enormi basi di dati umanistici memorizzati su calcolatori posti in caverne al riparo da catastrofi naturali e guerre, anche se questo è stato fatto, ad esempio, per il patrimonio mondiale di semi vegetali (*Svalbard globale frøhbvelv*).

[27] Forse questo è il vagheggiamento che stava alla base degli ambiziosi progetti della fondazione Paul Getty (https://www.getty.edu/foundation).

BIBLIOGRAFIA

Barners-Lee, T., & Fischetti, M. (1999). *Weaving the Web*. San Francisco: Harper.

Berners-Lee, T., Hendler, J., & Lassila, O. (2001). The Semantic Web. *Scientific American*, *May 2001*.

Bortolini, U., Tagliavini, C., & Zampolli, A. (1971). *Lessico di Frequenza della Lingua Italiana Contemporanea*. Milano: IBM.

Bejček, E., Hajičová, E., Hajič, J., Jínová, P., Kettnerová, V., Kolářová, V., Mikulová, M., Mírovský, J., Nedoluzhko, A., Panevová, J., Poláková, L., Ševčíková, M., Štěpánek, J., & Zikánová, Š. (2013). *Prague Dependency Treebank 3.0* (data/software). Praha: Univerzita Karlova v Praze, MFF, ÚFAL, Prague. Retrieved from *http://ufal.mff.cuni.cz/pdt3.0*.

Busa, R. SJ. (1949). *La terminologia tomistica dell'interiorità. Saggi di metodo per una interpretazione della metafisica della presenza*. Milano: Bocca.

Charniak, E. (1997). Statistical Techniques for Natural Language Parsing. *AI Magazine*, *18(4)*, 33–44.

Church, K.W. (1988). A stochastic parts program and noun phrase parser for unrestricted text. In N. Sondheimer, & B. Ballard (Eds.), ANLC '88: *Proceedings of the Second Conference on Applied Natural Language Processing. Association for Computational Linguistics* (pp. 136–143). Stroudsburg, PA. doi:10.3115/974235.974260.

De Mauro, T., Mancini, F., Vedovelli, M., & Voghera, M. (1993). *Lessico di frequenza dell'Italiano Parlato*. Milano: Etaslibri.

DeRose, S.J. (1988). Grammatical category disambiguation by statistical optimization. *Computational Linguistics*, *14(1)*, 31–39.

DeRose, S.J. (1990). *Stochastic Methods for Resolution of Grammatical Category Ambiguity in Inflected and Uninflected Languages.* Ph.D. Dissertation. Providence, RI: Brown University Department of Cognitive and Linguistic Sciences. Retrieved from *http://www.derose.net/steve/writings/dissertation/Diss.0.html*.

Ferrari, G. (2010). Considerazioni sulla Lessicografia antica e moderna. In G. Vanotti (Ed.), *Il lessico Suda e gli storici greci in frammenti. Atti dell'incontro internazionale, Vercelli 6–7 Novembre 2008* (pp. 477–494). Roma: Tored.

Ferrari, G. (2019). Filologia Computazionale, una terza via. In M.S. Corradini, & G. Ferrari (Eds.), *Percorsi di linguistica e di filologia computazionali* (pp. 121–124). Pisa: ETS.

Gruber, T.R. (1993). A translation approach to portable ontologies. *Knowledge Acquisition*, *5(2)*, 199–220.

Guarino, N., & Musen, M. (2015). Applied ontology: The next decade begins**.** *Applied ontology*, *10*, 1–4.

Hendler, J.A., & van Harmelen, F. (2008). The Semantic Web: webizing knowledge representation. In F. van Harmelen, V. Lifschitz, & B. Porter, *Handbook of Knowledge Representation* (pp. 821–839). Amsterdam: Elsevier.

Herdan, G. (1956). *Language as Choice and Chance*. Groningen: Noordhoff.

Herdan, G. (1960). *Type-Token Mathematics. A Textbook of Mathematical Linguistics*. The Hague: Moulton & Co.'S-Gravenhage.

Lugmayr, A., & Teras, M. (2015). Immersive Interactive Technologies in Digital Humanities: A Review and Basic Concepts. In T. Chambel, & P. Viana (Eds.), *ImmersiveME '15: Proceedings of the 3rd International Workshop on Immersive Media Experiences* (pp. 32–36). New York: Association for Computing Machinery.

Juilland, A., & Chang-Rodriguez, E. (1964). *Frequency Dictionary of Spanish Words*. Den Haag/Paris: North-Holland.

Juilland, A., Edwards, P.M., & Juilland, I.I. (1966). *Frequency Dictionary of Rumanian Words*. London/Den Haag/Paris: North Holland.

Juilland, A., Broding, D., & Davidovitch, C. (1970). *Frequency Dictionary of French Words*. Den Haag/Paris: North Holland.

Juilland, A., & Traversa, V. (1973). *Frequency Dictionary of Italian Words*. Den Haag/Paris: North Holland.

Nigra, C. (1888). *Canti popolari del Piemonte*. Torino: Loescher.

Passarotti, M. (2015). What you can do with linguistically annotated data. From the Index Thomisticus to the Index Thomisticus Treebank. In P. Roszak, & J. Vijgen (Eds.), *Reading Sacred Scripture with Thomas Aquinas. Hermeneutical Tools, Theological Questions and New Perspectives* (pp. 3–44). Turnhout: Brepols.

Smith, J.M. (1988). *SGML: the User's Guide to ISO 8879*. Chichester: Horwood.

Svensson, P. (2010). The Landscape of Digital Humanities. *Digital Humanities Quarterly*, *4.1*. Retrieved from *www.digitalhumanities.org/dhq/vol/4/1/000080/000080.html*.

Taylor, A., Marcus, M., & Santorini, B. (2003). The Penn Treebank: An overview. In A. Abeillé (Ed.) *TREEBANKS. Building and using parsed corpora* (pp. 5–22). Berlin: Springer.

Terras, M., Nyhan, J., & Vanhoutte, E. (Eds.). (2013). *Defining Digital Humanities: A Reader*. London/New York: Routledge.

Turing, A. M. (1937). On Computable Numbers, with an Application to the Entscheidungsproblem. *Proceedings of the London Mathematical Society*, *2* (42**)**, 230–265 (republished in 1965: Ed. by M. David, Hewlett, NY: Raven Press).

Turing, A.M. (1938). Correction to: On Computable Numbers, with an Application to the Entscheidungsproblem. *Proceedings of the London Mathematical Society, 2* (43**)**, 544–546.

Weaver, W. (1949). *Memorandum on Translation*. Retrieved from http://www.mt-archive.info/Weaver-1949.pdf.

Sitografia

Associazione di Informatica Umanistica e cultura digitale:	www.aiucd.it
British National Corpus	https://www.english-corpora.org/bnc/
Corpus of Contemporary American English	https://www.english-corpora.org/coca/
Digital Corpus of Sanskrit	http://www.sanskrit-linguistics.org/
Granai della memoria	https://www.granaidellamemoria.it/index.php/it
Index Thomisticus	https://www.corpusthomisticum.org/it/index.age
Internet Scred Text Archive	https://www.sacred-texts.com/sbe/index.htm
Paul Getty Foundation	https://www.getty.edu/foundation/
Perseus Digital Library	https://www.perseus.tufts.edu/
Pixar	https://www.pixar.com/
RAITeche	http://www.teche.rai.it/archivio-del-folclore-italiano/
Sony Pictures	http://www.sonypicturesanimation.com/
Text Encoding Initiative	https://tei-c.org/
The Latin Library	http://www.thelatinlibrary.com/
Trésor de la Langue Française informatisé	http://atilf.atilf.fr/
View Conference	https://www.viewconference.it/

Riassunto: Questo articolo non presenta una ricerca originale, ma è piuttosto un tentativo di definire questo settore scientifico così complesso e piuttosto sfumato, facendo anche ricorso alle sue radici storiche e tecnologiche. Una prima linea di distinzione degli obiettivi di questa disciplina consiste nell'opposizione tra tecniche per memorizzare e recuperare mediante calcolatore dati rilevanti agli

artefatti umani e la costruzione creativa di una “cultura digitale”. Un breve ricostruzione storica delle prime origini delle *digital humanities*, fa supporre una connessione con la linguistica computazionale e con lo sviluppo delle risorse linguistiche come fondamenti della disciplina. I confini della disciplina sono evoluti con l’ampliarsi degli orizzonti scientifici e l’evoluzione delle tecnologie dedicate. Le tecnologie di base che rendono possibile lo sviluppo di aree più vaste di Digital Humanities sono due, l’annotazione testuale, con il conseguente *markup*, e le capacità multimediali offerte dai browser ordinari. Queste due tecniche sono strettamente legate, dal momento che il linguaggio che è utilizzato dal protocollo http (HyperText Transfer Protocol) ha origine comune con quello che viene utilizzato nel *markup* testuale. Ipertestualità e multimodalità permettono l’estensione dell’uso del computer nella memorizzazione e recupero di materiali umanistici di tipo diverso, includendo immagini, video e suoni. Alla fine, un ulteriore sviluppo del *markup*, cioè la dichiarazione delle *entities*, ha reso possibile l’uso di tecniche di *semantic web* per condurre ricerche avanzate. Il campo della cultura digitale creativa è vastissimo e la quantità di software disponibili per rendere possibile questa creatività è enorme; alcune forme di arte ne hanno largamente tratto vantaggio. Nella conclusione si discute il serio problema dell’obsolescenza tecnologica.

Parole chiave: digital humanities, linguistica computazionale, risorse linguistiche, markup, tecnologia per le scienze umane