FABIO CIOTTI

*Tematologia e metodi digitali: dal markup alle ontologie*

In

*I cantieri dell'italianistica. Ricerca, didattica e organizzazione agli inizi del XXI secolo*.
Atti del XVII congresso dell'ADI – Associazione degli Italianisti (Roma Sapienza, 18-21 settembre 2013), a cura di B. Alfonzetti, G. Baldassarri e F. Tomasi, Roma, Adi editore, 2014
Isbn: 9788890790546

 

FABIO CIOTTI

# *Tematologia e metodi digitali: dal markup alle ontologie*

*La diffusione delle nuove tecnologie digitali negli studi letterari ha reso disponibile molta parte della tradizione testuale in formato digitale. Queste collezioni possono costituire la base per avviare un programma sistematico di annotazione tematica digitale. Il perseguimento di questo programma richiede di affrontare diversi problemi teorici, metodologici e strumentali: una definizione rigorosa del concetto di tema e motivo, e della relazione tra tema/motivo in quanto entità astratte e loro manifestazioni discorsive; la realizzazione di un repertorio sistematico dei temi e dei motivi letterari; la sistematica correlazione tra temi/motivi e testi. Le tecnologie e i formalismi del Web Semantico forniscono un apparato strumentale idoneo alla creazione di un repertorio tematico che si organizzi per multiple stratificazioni tipologiche ed esprima una ricca rete di relazioni tra classi e istanze di temi e motivi. La disponibilità di una ontologia tematica consente di annotare le manifestazioni discorsive dei temi nei testi digitali codificati, producendo progressivamente una mappatura tematica della tradizione letteraria. La realizzazione di questo programma richiede la cooperazione della comunità di studiosi e lo sviluppo di una apposita infrastruttura.*

In questo intervento ci proponiamo di esplorare le potenzialità offerte dai metodi e dai procedimenti computazionali per l'analisi e lo studio dei temi letterari. Questo ambito di studi, dopo la lunga *damnatio memoriae* cui era stato soggetto, a partire dagli anni '80 del secolo scorso è tornato progressivamente alla ribalta del dibattito teorico e critico-letterario. Le cause di questa rinnovata attenzione sono state indicate da molti studiosi nella reazione alla vulgata semiotico-strutturalista che individuava l'essenza della letterarietà nella forma pura – autonoma dai contenuti ed esente da processi evolutivi – e al contemporaneo emergere degli studi culturali e del neo-storicismo, scuole in cui analisi e genealogia dei temi assume una funzione centrale. Non è questa la sede per descrivere con dettaglio motivi e sviluppi di questo rinnovato interesse, che peraltro sono stati ben analizzata da Matteo Lefevre,[1] né per discutere le diverse e spesso contrapposte visioni che critici e teorici hanno espresso nei confronti di questo ritorno – si veda per questo gli atti dei tre convegni parigini *Pour une thématique* (I, 1984; II, 1986; III, 1988),[2] che, sotto la spinta intellettuale di C. Bremond e T. Pavel, hanno segnato l'avvio di questo rinnovato interesse, la fondamentale miscellanea a cura di W. Sollors[3] e per il panorama italiano più recente il numero 58 di Allegoria, che dedica al tema la sezione "La critica tematica oggi" a cura di R. Luperini.

Diverse voci hanno osservato come, collocato in questa temperie intellettuale di fine/inizio secolo, il ritorno alla critica tematica/tematologica, soprattutto in Italia, ha finito spesso per coincidere con il ritorno all'impressionismo soggettivista, alla mancanza o persino al rifiuto programmatico di una fondazione teorica del metodo:

> Guidata di volta in volta dalle predilezioni del singolo studioso, e molto spesso da un empirismo di buon senso, la ricerca sui temi appare come un filone eteroclito, non come una coerente metodologia.[4]

Il primo problema da affrontare per collocare utilmente l'analisi dei temi nel contesto delle Digital Humanities è proprio questo: alla base di qualsiasi uso intrinsecamente scientifico dei metodi computazionali (che non si limiti dunque alla semplice disseminazione dei prodotti della ricerca) c'è una definizione formale di termini, oggetti e processi analitici, attività che si riassume

[1] M. LEFEVRE, *Tema e motivo nella critica letteraria*, «Allegoria», XV (2003), 45, 5-22.
[2] Per gli atti: *Du thème en littérature*, «Poétique», 64 (1985); *Variations sur le thème*, «Communications», 47 (1988); *Pour une thématique*, «Strumenti critici», 60 (1989).
[3] W. SOLLORS, *The Return of Thematic Criticism*, Cambridge (MA) and London, Harvard University Press, 1993.
[4] P. PELLINI, *Critica tematica e tematologia: paradossi e aporie*, «Allegoria», XX (2008), 58, 61-83: 61.

nel termine modellizzazione.[5] Si tratta dunque di separare l'analisi tematologica da qualsiasi tendenza soggettivista, che è certo lecita nella critica militante, e di collocarla nel contesto di una metodologia formale e intersoggettiva.

Cominciamo dunque con un chiarimento a priori sul dominio di studi cui ci riferiamo. In questo contesto ci occuperemo del versante diacronico comparatista e intertestuale dell'analisi dei temi, quello che per primo P. Van Tieghem negli anni '30 del secolo scorso ha definito *thématologie*,[6] e verso il quale con maggiore acume si sono appuntate le critiche radicali della tradizione idealista e formalista statunitense prima e strutturalista francese poi (la "critica del cavallo", lo definì R. Jakobson), relegandolo al dominio degli studi folcloristici ed etnografici. La rinnovata fortuna degli studi sui temi ha restituito dignità di metodo critico anche a questa pratica, sebbene in funzione strumentale al lavoro ermeneutico vero e proprio, come osserva M. Domenichelli:[7]

> si può […] analizzare un tema come ricorrente attraverso variazioni significative in tutta la tradizione, come per esempio capita se si vuole scrivere una voce di dizionario tematico: in questo caso il problema che ci si pone è quello dell'informazione, ma senza rinunciare a un disegno ermeneutico che non riguarda i singoli testi, ma il macrotesto della tradizione, pur nella consapevolezza che si sta mettendo a punto uno strumento d'analisi più che un'analisi in sé.

Non possiamo in questa sede soffermarci oltre su questo dibattito, peraltro affrontato a sufficienza da molti tra i maggiori studiosi dei fatti letterari di questi ultimi decenni. Ci basta osservare come l'approccio tematologico sia di fatto rientrato a pieno titolo nell'apparato analitico degli studi letterari.

*Studi letterari e Digital Humanities*

Nel corso dell'ultimo ventennio l'Informatica umanistica, o *Humanities Computing* nella formulazione anglosassone, si è venuta progressivamente liberando della stimmate di disciplina di nicchia, riuscendo al contempo a ottenere una presenza rilevante nella didattica offerta dalla facoltà umanistiche (e questo anche in Italia, nonostante ritardi, ritrosie culturali e crisi dell'università in generale abbiano senza dubbio rappresentato e ancora rappresentino fattori di ostacolo); a conseguire importanti risultati sul piano della ricerca teorica e metodologica; a promuovere e consolidare infrastrutture e organizzazioni per la cooperazione scientifica a livello nazionale e internazionale che raccolgono e coordinano un numero ormai grandissimo di studiosi a livello planetario, organizzano convegni mastodontici e pubblicano monografie e periodici autorevoli.[8]

La recente e rapida diffusione della formula *Digital Humanities* sancisce, sul piano linguistico, il successo di questo processo di consolidamento e generalizzazione, e indica l'ambizione totalizzante di questo vasto e proteico campo di studi, i cui confini interni con il campo delle scienze umane *tout court* sono sempre più sfumate, come osservano - con buona dose di ottimismo - gli autorevoli autori del recente volume *Digital Humanities*[9]

---

[5] Su questo si veda G. GIGLIOZZI, F. CIOTTI, *Introduzione all'uso del computer negli studi letterari*, Milano, Bruno Mondadori, 2003; W. MCCARTY, *Humanities Computing*, Basingstoke,Palgrave Macmillan, 2005; T. ORLANDI, *Informatica testuale: teoria e prassi*, Roma-Bari, Laterza, 2010.

[6] P. VAN TIEGHEM, *La Littérature Comparée*, Paris, Colin, 1933.

[7] M. DOMENICHELLI, *Lo scriba, il mondo, la storia. Considerazioni in margine a «L'incontro e il caso. Narrazioni moderne e destino dell'uomo occidentale» di Romano Luperini*, «Intersezioni», XXVIII (2008), 1, 135-146: 138.

[8] Per avere un quadro della dimensione del fenomeno si può consultare il sito della *Alliance of Digital Humanities Organizations* (ADHO, http://adho.org), che federa e coordina numerose associazioni territoriali in tutto il mondo e organizza il convegno internazionale annuale 'DH'.

[9] A. BURDICK, J. DRUCKER, P. LUNENFELD, T. PRESNER, J. SCHNAPP, *Digital_Humanities*. Cambridge (MA), MIT Press, 2012, vii.

> We live in one of those rare moments of opportunity for the humanities, not unlike other great eras of cultural-historical transformation such as the shift from the scroll to the codex, the invention of moveable type, the encounter with the New World, and the Industrial Revolution. Ours is an era in which the humanities have the potential to play a vastly expanded creative role in public life.[…]
>
> Digital Humanities represents a major expansion of the purview of the humanities, precisely because it brings the values, representational and interpretive practices, meaning-making strategies, complexities, and ambiguities of being human into every realm of experience and knowledge of the world. It is a global, trans-historical, and transmedia approach to knowledge and meaning-making.

Tuttavia, nonostante la ricerca umanistica digitale abbia ormai una lunga storia di successi alle spalle, si deve rilevare come di rado sia riuscita a stabilire una relazione proficua con la generalità della comunità scientifica umanistica e letteraria in particolare (diverso in parte il discorso per la linguistica e in parte anche per la storiografia), tanto che a distanza di oltre dieci anni resta ancora valida la provocatoria osservazione di J. Unsworth:[10]

> We need (we still need) to demonstrate the usefulness of all the stuff we have digitized over the last decade and more – and usefulness not just in the form of increased access, but specifically, in what we can do with the stuff once we get it: what new questions we could ask, what old ones we could answer.

Probabilmente la grande influenza che ancora oggi hanno gli approcci post-strutturalisti negli studi letterari gioca un ruolo importante in questa distanza: la 'Teoria' senza aggettivi, come la definisce J. Culler,[11] non si presta facilmente a interagire con il formalismo delle strutture dati e dei modelli informatici. Ma si deve ammettere che i metodi computazionali per l'analisi e l'edizione dei testi e i relativi risultati in termini di analisi ed edizioni hanno spesso deluso le aspettative (salvo ambiti ristretti come gli studi sull'attribuzione) e di rado sono riusciti ad acquisire un sufficiente riconoscimento nell'ambito delle discipline letterarie e filologiche tradizionali.

Nonostante la chiara consapevolezza teorica più volte enunciata, la predisposizione degli strumenti di rappresentazione e analisi concreti non ha finora risolto in modo sodisfacente il nodo della specificità e della complessità degli oggetti e delle procedure di analisi tipiche della ricerca letteraria. L'investimento degli stessi cultori delle Digital Humanities nella definizione di nuovi modelli e linguaggi per la rappresentazione ed elaborazione formale dei complessi oggetti culturali cui si applicano è infatti stato piuttosto ridotto. Più comunemente si sono ereditati e applicati modelli e linguaggi elaborati dall'informatica per finalità e domini diversi.

Paradigmatico il caso del linguaggio XML. Esso ha assunto un ruolo centrale nella modellizzazione dei dati in ambito testuale, per numerose buone ragioni.[12] Ma è ben noto che XML da una parte impone l'adozione di un modello di dati ad albero che non sempre si adatta alla natura strutturale degli oggetti da rappresentare, dall'altra non è in grado di rappresentare adeguatamente i numerosi e complessi livelli semantici che caratterizzano un testo letterario.

Se questo è il quadro, quali sono le prospettive che si possono aprire per lo sviluppo dell'informatica umanistica e di quella letteraria in particolare? Senza dubbio, consolidare e se possibile estendere i risultati acquisiti è una missione validissima e anzi irrinunciabile. Gli archivi testuali vanno preservati ed implementati, la trascrizioni ed edizioni digitali basate sui

---

[10] J. UNSWORTH, *Tool-Time, or 'Haven't We Been Here Already?': Ten Years in Humanities Computing*, «Transforming Disciplines: The Humanities and Computer Science», Washington, D.C., Jan.18, 2003, <http://www.iath.virginia.edu/~jmu2m/carnegie-ninch.03.html>

[11] J. D. CULLER, *Teoria della letteratura : una breve introduzione*, Roma, Armando, 1999.

[12] F. CIOTTI, *La rappresentazione digitale del testo: il paradigma del markup e i suoi sviluppi*, in L. Perilli, D. Fiormonte (a cura di), *La macchina nel tempo: studi di informatica umanistica in onore di Tito Orlandi*, Firenze, Le lettere, 2011.

formalismi attualmente disponibili moltiplicate, gli standard mantenuti, applicati e diffusi. Ma è giunto il momento di individuare nuove linee di ricerca, di esplorare le tendenze innovative che potrebbero fornire un ulteriore salto di qualità e una più ampia giustificazione scientifica (ma anche istituzionale) all'incontro tra informatica e studi umanistici.

Tra i numerosi campi di indagine emersi in questi ultimi anni, due si segnalano come i più promettenti e generalizzabili:

1) *Big Data* umanistici: l'applicazione di strumenti per l'analisi automatica delle ingenti masse di risorse testuali/documentali e di dati digitali prodotte finora, attraverso metodi probabilistici e tecnologie di *text mining* e *knowledge extraction*;

2) *Web Semantico* e *Linked Open Data* per gli oggetti culturali: la sperimentazione dei nuovi linguaggi e modelli di dati per l'annotazione semantica delle risorse informative e per la relativa elaborazione basata su sistemi inferenziali.

Per quanto riguarda l'applicazione delle tecnologie di *Big Data analysis* in ambito umanistico, ci limitiamo a ricordare che si tratta di un ambito di ricerca che consiste nell'applicazione di tecniche ed euristiche di *data mining* per la ricerca di regolarità e schemi ricorrenti impliciti e non osservabili a priori all'interno di grandi moli di dati strutturati e non strutturati.

La ricerca di tali pattern e regolarità si basa su complessi algoritmi statistici e probabilistici, i più noti dei quali sono derivati dall'analisi probabilistica bayesiana, che studia la probabilità di eventi non quantificabili a priori (ad esempio la probabilità che in un insieme di soggetti prevalga una certa aspettativa, o che un testo sia categorizzabile in base a un dato aspetto prevalente). Quando questi algoritmi sono applicati a dati testuali si parla più specificamente di *text mining*. In questa direzione si sono indirizzati alcuni importanti progetti di ricerca nell'ambito delle Digital Humanities, tra cui ricordiamo in particolare, per l'eco che hanno avuto, le ricerche condotte presso lo Stanford Literary Lab fondato e diretto da F. Moretti.[13]

Lo stesso Moretti ha teorizzato come queste tecnologie di ricerca possano essere il fondamento di un vero e proprio nuovo metodo di studio dei fenomeni letterari, che ha definito *distant reading* (giocando sulla opposizione con il *close reading* introdotto nella critica letteraria dal *New Criticism*).

L'idea fondamentale di questi approcci è che esistono dei fenomeni di lunga durata o di vasta portata che sono inosservabili a 'occhio nudo' (il significato di occhio nudo va ovviamente contestualizzato caso per caso) e che giocano un ruolo esplicativo rilevante nel comprendere fenomeni letterari come l'evoluzione dei generi, l'affermazione di uno stile e la sua recezione, la presenza di cluster contenutistici ricorrenti in un dato intervallo temporale della storia letteraria ampia (cioè inclusiva di tutti i livelli e i generi della produzione testuale del periodo preso in esame).

Non è questa la sede per discutere ulteriormente pregi e difetti di questo approccio. Rileviamo solo due questioni critiche: gli algoritmi adottati, che sono in generale del tutto indipendenti dal contesto (si applicano cioè indifferentemente a serie decennali di dati delle transazioni finanziarie come a *very large textual corpora*) sono efficaci nell'identificazione di pattern rilevanti se il set di dati da analizzare è decisamente grande; è quantomeno dubbio che la dimensione dei set di dati testuali lo siano. In secondo luogo questi algoritmi individuano similarità rintracciando sequenze ricorrenti di dati atomici indipendentemente dalla loro semantica; ma in molti sensi la semantica dei dati fortemente strutturati è fissata a priori, e quindi le regolarità sono indirettamente fondate semanticamente; se si lavora su dati scarsamente strutturati, come sono i testi digitali non marcati, i dati atomici finiscono per essere i singoli caratteri che non giocano alcun ruolo semantico (o ne giocano uno molto limitato). Non diciamo che non si possano individuare tratti interessanti emergenti anche a questo livello, ma le conclusioni che si possono trarre da queste analisi escludono una grande quantità di fatti comunemente ritenuti rilevanti nell'analisi degli oggetti testuali, e letterari in particolare.

[13] Si veda il sito del laboratorio, http://litlab.stanford.edu/; e dello stesso Moretti l'ormai classico *Graphs, Maps, Trees: Abstract Models for a Literary History*, London, Verso, 2005.

*Web Semantico, ontologie formali e Linked Data*

Ai fini della nostra proposta per una tematologia digitale riteniamo più promettenti le potenzialità delle tecnologie del Web Semantico (WS) e dei Linked Data.[14] L'idea fondamentale del WS consiste nell'associare alle risorse informative sul Web una descrizione formalizzata di parti determinate del loro significato intensionale o estensionale. Tali descrizioni semantiche sono espresse mediante una famiglia di formalismi con diverse capacità espressive e complessità di trattamento computazionale. Alla base c'è *Resource Description Framework* (RDF) che consente di esprimere proprietà semantiche mediante semplici asserti predicativi. Ogni asserto ha una struttura soggetto − predicato − oggetto. Un asserto specifica una relazione predicativa tra soggetto e oggetto (in RDF sono consentite solo relazioni binarie). Gli asserti sono anche noti come triple e gli insiemi di asserti si possono rappresentare come grafi etichettati orientati aciclici.

RDF in quanto tale non fornisce un vocabolario predefinito di proprietà e di relazioni sotto cui sussumere e organizzare le risorse. Si tratta di un modello di dati semplice e rigoroso per specificare proprietà di risorse, qualsivoglia esse siano. In un contesto ampio ed eterogeneo come il Web possono esistere numerosi schemi e vocabolari semantici, basati su diverse concettualizzazioni di particolari domini, su diverse terminologie e lingue. In linea generale si può assumere che esistano anche concettualizzazioni mutuamente contraddittorie e/o mutevoli nel tempo. Al fine di rendere utilizzabili queste concettualizzazioni in modo computazionale (almeno in parte) è necessario conseguire un ulteriore livello di formalizzazione: quello delle ontologie formali.

La definizione classica di questo concetto è stata fornita da Gruber: «An ontology is an explicit specification of a conceptualization».[15] Il termine ontologia, ereditato dalla metafisica classica dove, sin dalla sistemazione aristotelica, denotava la teoria dell'essere e delle sue categorie, è oggi adottato a designare una ampia e diversificata classe di oggetti che vanno dai vocabolari controllati, ai thesaura fino alle ontologie formali vere e proprie. Queste, oltre a fissare una terminologia strutturata per gli enti di un dato dominio, ne fissano anche la semantica condivisa da una data comunità, in termini logico-formali:

> In the context of computer and information sciences, an ontology defines a set of representational primitives with which to model a domain of knowledge or discourse. The representational primitives are typically classes (or sets), attributes (or properties), and relationships (or relations among class members). The definitions of the representational primitives include information about their meaning and constraints on their logically consistent application.[16]

Esistono numerosi linguaggi formali per specificare ontologie formali. A un primo livello di complessità e capacità espressiva si pone *RDF Schema* (RDFS), che permette di definire formalmente classi, tipi di proprietà, relazioni tra classi e proprietà. I vincoli espressi da RDFS, tuttavia, non sono sufficienti per esprimere interamente i vincoli logici adeguati per formalizzare un dominio concettuale. Occorre un formalismo in grado di specificare le relazioni logico-semantiche (equivalenza, specificazione, generalizzazione, istanziazione, cardinalità, simmetria etc.) tra oggetti e proprietà di un medesimo schema e di schemi diversi. Nel contesto del WS il linguaggio deputato a conseguire questo secondo livello di formalizzazione è *Web Ontology Language* (OWL), a sua volta diviso in sottoinsiemi con diversa capacità espressiva.

---

[14] Entrambi i termini e la visione a cui alludono, sono stati proposti da Tim Berners-Lee, l'inventore del Web.

[15] T. R. GRUBER, *A translation approach to portable ontology specifications*, «Knowledge Acquisition», V (1993), 2, 199.

[16] T. R. GRUBER, *Ontology*, *Encyclopedia of Database Systems*, Springer-Verlag, 2009.

Nonostante lo sviluppo di numerosi linguaggi in grado di sostenere i requisisti del progetto, la originaria visione universalista del Web Semantico si è dimostrata tecnicamente e socialmente non realizzabile. Tuttavia l'applicazione dei metodi e delle tecnologie semantiche in domini ristretti e in contesti controllati e locali ha conseguito dei buoni esiti. In particolare si è rivelata vincente l'idea dei *Linked Data*, con cui lo stesso Berners-Lee ha inteso dare una versione praticabile del Web Semantico:[17]

> The term Linked Data refers to a set of best practices for publishing and interlinking structured data on the Web. These best practices were introduced by Tim Berners-Lee in his Web architecture note Linked Data and have become known as the Linked Data principles. These principles are the following:
>
> 1. Use URIs as names for things.
> 2. Use HTTP URIs, so that people can look up those names.
> 3. When someone looks up a URI, provide useful information, using the standards (RDF, SPARQL).
> 4. Include links to other URIs, so that they can discover more things.
>
> The basic idea of Linked Data is to apply the general architecture of the World Wide Web to the task of sharing structured data on global scale.

Intorno al nucleo originale costituito da DBPedia, la versione formalizzata del contenuto di Wikipedia, è andata crescendo una rete di dati e servizi, erogati in regime di *open content*, ormai vastissima (se ne può vedere la struttura sul sito http://lod-cloud.net). In questa rete un ruolo importante è svolto da fonti legate alla gestione del patrimonio culturale.[18]

*Ontologie e Linked Data per la tematologia*

Dato il contesto teorico e tecnico appena delineato, quale convergenza ci può essere tra le tecnologie del Web Semantico e gli studi umanistici e letterari in particolare? Torniamo così al nostro punto di partenza: il dominio della tematologia tradizionale costituisce uno dei candidati ideali per avviare questa sperimentazione. A nostro avviso le tecnologie ontologiche del Web Semantico forniscono un apparato strumentale idoneo alla creazione di un repertorio tematico che non sia una pura enumerazione – come nei pur validi repertori a stampa – ma che al contrario si organizzi per multiple stratificazioni tipologiche e al contempo permetta una ricca rete di relazioni orizzontali tra classi e tra istanze di temi e motivi.

Il lavoro condotto negli anni passati dalla comunità delle Digital Humanities (in particolare in area letteraria e linguistica) ha portato alla creazione di grandi archivi testuali: molta parte delle varie tradizioni testuali nazionali sono disponibili ormai in formato digitale; una buona percentuale sono disponibili in formati di codifica avanzati come XML /TEI, sebbene il livello di codifica di queste risorse sia perlopiù limitato alla rappresentazione della struttura editoriale.[19] Queste collezioni di risorse testuali digitali di qualità sono la base per avviare un programma di analisi e annotazione tematica della tradizione letteraria. La realizzazione di tale programma richiede una architettura complessa e modulare che integri molteplici livelli di modellizzazione ontologica e di codifica testuale al fine di:

---

[17] T. HEATH, C. Bizer, *Linked Data: Evolving the Web into a Global Data Space*, «Synthesis Lectures on the Semantic Web: Theory and Technology», I (2011), 1, 1–136.

[18] Si veda il sito del network *Linked Open Data in Library Archives and Museum* per un quadro generale del fenomeno, http://lodlam.net.

[19] Per il contesto italiano ricordiamo *Biblioteca Italiana* (http://www.bibliotecaitaliana.it), che contiene oltre 1700 opere del canone letterario italiano; *Memorata Poesis* (http://www.memoratapoetis.it/) che offre accesso al corpus completo della poesia latina; DigilibLT (http://www.digiliblt.unipmn.it), che contiene un canone degli autori e delle opere latine tardoantichi.

1) tenere distinti i diversi livelli di astrazione sia dal punto di vista logico sia nella scelta dei relativi formalismi di rappresentazione; in questo modo a ogni livello verrà adottato il formalismo più efficiente senza aumentare la complessità gestionale immotivatamente;

2) ridurre al minimo la quantità di informazioni semantiche espresse direttamente a livello di inline markup privilegiando quanto più possibile strategie di stand-off markup;

3) facilitare la progressiva estensione e rimodulazione delle annotazioni tematiche (in prospettiva anche mediante modalità di lavoro in *crowdsourcing*);

4) facilitare l'interoperabilità con altri set e repository di dati rilevanti espressi in forma di Linked data.

Alla base, come detto, ci sono i testi ospitati nei vari repository e già codificati in formato XML/TEI. A questi viene associato un livello di annotazioni digitali con tecniche di stand-off markup (cioè markup esterno al documento digitale che contiene il testo stesso) che assegna ai testi e a loro porzioni una specificazione tematica basandosi su un repertorio terminologico/concettuale condiviso. Questo a sua volta è organizzato sulla base di una ontologia formale dei concetti tematologici.

La distinzione tra livello terminologico e livello ontologico si fonda sulla distinzione canonica nel contesto delle Description Logic (le logiche alla base del linguaggio di modellazione ontologica del Web Semantico OWL) tra *Terminological box* (Tbox) e *Assertion box* (Abox):[20]

> Description logics and their semantics traditionally split concepts and their relationships from the different treatment of instances and their attributes and roles, expressed as fact assertions.
>
> The concept split is known as the TBox (for terminological knowledge, the basis for T in TBox) and represents the schema or taxonomy of the domain at hand. The TBox is the structural and intensional component of conceptual relationships. It is this construct for which Structure Dynamics generally reserves the term "ontology".
>
> The second split of instances is known as the ABox (for assertions, the basis for A in ABox) and describes the attributes of instances (or individuals), the roles between instances, and other assertions about instances regarding their class membership with the TBox concepts. Both the TBox and ABox are consistent with set-theoretic principles.

Nel nostro caso la formalizzazione concettuale deve affrontare e risolvere un problema annoso nella tematologia, ovvero la definizione chiara e rigorosa dei termini teorici che si adottano. Problema complesso se ancora trenta anni dopo l'uscita del saggio "Tema/Motivo" di C. Segre,[21] il dibattito teorico sull'argomento deve registrare una impasse. Si veda ad esempio cosa dichiara esplicitamente R. Ceserani, curatore del maggiore repertorio tematico uscito di recente su volume a stampa, il *Dizionario dei temi letterari* della Utet:[22]

> Abbiamo evitato di dare una spiegazione esplicita delle scelte fatte; per questo non si trova nel nostro dizionario una motivazione ragionata della scelta fatta di non distinguere fra temi e motivi e tanto meno si trovano in esso, come avviene invece nel dizionario dei Daemmrich, delle vere e proprie voci teoriche su Motiv e Thema.[23]

Eppure proposte di sistematizzazione della materia, più o meno rigorose e consistenti, non mancano certo. Forse proprio la molteplicità di tali sistematizzazioni e le molteplici

[20] M. K. BERGMAN, *The Fundamental Importance of Keeping an ABox and TBox Split*, «AI3 Adaptive Information», <http://www.mkbergman.com/489/ontology-best-practices-for-data-driven-applications-part-2>.

[21] C. SEGRE, *Tema/motivo*, in *Avviamento all'analisi del testo letterario*, Torino, Einaudi, 1985, 331-356.

[22] R. CESERANI, M. DOMENICHELLI, P. FASANO, *Dizionario dei temi letterari*, Torino, UTET, 2006-2007.

[23] R. CESERANI, *Il punto sulla critica tematica*, «Allegoria», XX (2008), 58, 25-33:28.

contraddizioni tra l'una e l'altra, costituiscono un ostacolo a riproporne oggi una che unifichi la famiglia. Sia chiaro, la materia è proteica in sé, e il concetto di famiglia concettuale (per come lo intende Wittgenstein) sarebbe il più adatto a dare conto di tale materia. Tuttavia l'approccio alla computazione che abbiamo scelto ci chiama a fare delle scelte, a ridurre quanto possibile l'implicito e modellizzare in modo formale il resto. Consapevoli che la stessa modellizzazione formale sta dentro un processo ermeneutico complessivo che ci condurrà a ritornarvi sopra per modificarla e adattarla, *ad infinitum*. Ma ad ogni dato momento sincronico del processo la struttura deve fissarsi, rispondendo a regole di isomorfismo rispetto al dominio e di dipendenza dal punto di vista della comunità di interpreti che vi si applica. D'altronde la nozione stessa di ontologia formale nella concezione di uno dei maggiori teorici della disciplina, N. Guarino, implica quella di 'significato inteso' delle primitive concettuali.[24]

Tornando al nostro dominio, scopo di una ontologia formale del dominio tematologico è la definizione dei concetti fondamentali: tema, motivo e concetti correlati come stereotipo, *locus communis* o *topos*, immagine, carattere-tipo. La individuazione delle loro reciproche relazioni (gerarchia, similarità, istanziazione, etc.). La specificazione delle loro proprietà accessorie e tipologico/funzionali. Le diverse relazioni (manifestazione, esemplificazione, istanziazione) tra temi e motivi e le loro manifestazioni discorsive o espressive in generale (infatti una tale ontologia potrebbe essere facilmente generalizzabile a oggetti non linguistici).

Ci pare che ancora oggi un punto di partenza ottimale per il lavoro di formalizzazione concettuale sia la sistemazione proposta da C. Segre:[25]

> Tema e motivo sono dunque unità di significato stereotipe, ricorrenti in un testo o in un gruppo di testi e tali da individuare delle aree semantiche determinanti. Unità di significato: si può trattare infatti di parole, frasi e gruppi di frasi del testo; oppure di parafrasi di parti del testo istituenti un significato autonomo. Stereotipe: la stereotipia può esser prodotta dalla sola ripetizione, entro un testo, ma per lo più è il prodotto di un continuo riutilizzo culturale (ripetizione in una successione di testi considerati come testo complessivo). […] Individuazione delle aree semantiche determinanti: si tratti dei dintorni dell'azione, o di campi concettuali, temi e motivi posano su punti chiave costituiscono delle specie di falsarighe per parti (narrative o illustrative) più o meno ampie del testo.

Temi e motivi sono dunque entità di contenuto che possono manifestarsi linguisticamente in specifici sintagmi o enunciati, ma che possono altresì essere manifestate da vaste porzioni di testo (a limite un testo nella sua interezza) senza avere nessun correlato linguistico immediato. Si identificano in quanto temi e non generici contenuti concettuali in virtù della loro natura di stereotipi culturali che risiedono nella memora culturale collettiva a cui autori e lettori attingono, pur mutandone nel tempo e nello spazio i valori semantici connotativi. Essi sono dunque oggetti del processo della intertestualità. In taluni casi la stereotipizzazione può produrre delle vere unità convenzionali e stilizzate anche linguisticamente, denominate *topoi*, come il *locus amoenus*, o il 'servitore furbo'.

Venendo alla distinzione tra i due concetti principali i teme e motivo, risulta ancora prezioso il suggerimento di C. Segre (anche se grande è la variabilità di definizioni che si riscontra nella letteratura):[26]

> Si chiameranno temi quegli elementi stereotipi che sottendono tutto un testo o una parte ampia di esso; i motivi sono invece elementi minori, e possono essere presenti in numero anche elevato. Molte volte un tema risulta dall'insistenza di più motivi. I motivi hanno maggior facilità a rivelarsi sul piano del discorso linguistico, tanto che, se ripetuti, possono operare in modo simile a dei ritornelli; i temi sono perlopiù di carattere metadiscorsivo. I motivi costituiscono di solito risonanze discorsive della metadiscorsività del tema.

---

[24] N. GUARINO, *Formal ontology, conceptual analysis and knowledge representation*, «Int. J. Hum.-Comput. Stud. », 43 (1995), 5-6, 625-640, DOI=10.1006/ijhc.1995.1066.
[25] C. SEGRE, *Tema/motivo*…, 348.
[26] C. SEGRE, *Tema/motivo*…, 349.

Possiamo estrapolare dunque le seguenti relazioni: un tema di norma non ha manifestazione discorsiva immediata, e si applica all'intero testo o a porzioni molto ampie di questo (si pensi alla coincidenza tema/capitolo ne *La coscienza di Zeno*). Un motivo è una unità di contenuto più piccola che può comporre un tema (ma non necessariamente tutti i motivi di un testo sono funzionali al tema o ai temi che lo caratterizzano) ed è dotato di correlati discorsivi immediati.

Queste osservazioni sono già sufficienti per procedere a una prima provvisoria sistemazione formale, il cui dettaglio esula dagli scopi di questo lavoro. A esse possono essere affiancate le varie categorizzazione tipologiche di temi e motivi che gli studiosi hanno di volta in volta proposto, distinguendo temi primari o fondazionali (derivati dalla culturalizzazione di fatti ed eventi dell'esperienza biologica della specie, come la nascita, la morte, la malattia, la guerra, la paura dell'ignoto, la natura benevola o distruttrice), temi secondari o sociali (derivati dall'esperienza della socializzazione). O ancora temi e motivi situazionali, emozionali, spaziali, temporali, figure-tipo e così via.[27]

A parziale superamento dell'obiezione circa la natura proteica e sfuggente di questi concetti si deve osservare che i fondamenti logici dei linguaggi per ontologie formali come OWL (ad esempio la non monotonicità degli assiomi e l'assunzione del criterio del mondo aperto) rendono possibile una certa flessibilità nella definizione concettuale e nell'assegnazione di proprietà a classi e individui. Detto in altri termini, riconosciuta la difficoltà oggettiva della materia, il formalismo che la modellizza è in grado almeno parzialmente di rappresentarne la complessità e persino la plurivocità.

*Dall'ontologia tematica ai testi e ritorno*

Una vola disegnata l'ontologia generale dei concetti generali necessari all'analisi tematologica, si pone il problema del popolamento – come usa dire – di tale ontologia, istanziando le classi con esemplari di singoli temi e motivi, e definendo le relazioni tra di essi. La disponibilità di questa ontologia tematica consentirà di annotare le manifestazioni discorsive dei temi nei testi digitali codificati, producendo progressivamente una mappatura tematica della tradizione letteraria, a costituire una base empirica inestimabile sulla quale estendere in modo sistematico l'analisi tematica con strumenti informatici.

Ovviamente la distinzione in tre momenti che qui proponiamo è logica ma solo in parte cronologica. Dal punto di vista operativo il popolamento dell'ontologia tematica andrebbe effettuato in parallelo con l'analisi tematica di testi e collezioni testuali, in modo progressivo e ricorsivo (prevedendo cioè la possibilità di tornare più volte su una data attribuzione tematica). Si tratta ovviamente di predisporre un *workflow* adeguato e diversi livelli di controllo che permettano a chi ha maggiore esperienza e conoscenza non solo di applicare un repertorio tematico predefinito ma anche di estenderlo, implementando l'ontologia nella A-box.[28]

L'annotazione tematica dei testi, visto l'elevato livello di granularità testuale che presuppone (soprattutto per l'identificazione dei motivi), deve affrontare anche i problemi tecnici posti dalla marcatura testuale. Infatti la segmentazione del testo mediante markup generico, al fine di assegnare a ciascun segmento il motivo rilevante, incorrere facilmente nel problema delle

[27] Cfr. T. WOLPERS, *Motif and Theme as Structural Content Units and "Concrete Universals"*, in W. Sollors, *The Return*…, 80-91.

[28] Questo modello operativo è in effetti in corso di sperimentazione nell'ambito di un progetto di analisi tematica forse meno complesso e ambizioso dal punto di vista teorico di quello che qui proponiamo. Si tratta del programma PRIN 2010/11 *Memoria poetica e poesia della memoria. Ricorrenze lessicali e tematiche nella versificazione epigrafica e nel sistema letterario*, che si propone di condurre su vasta scala una etichettatura tematica della tradizione epigrafica latina e volgare delle origini. In questo caso ci si è limitati a predisporre un repertorio terminologico sistematico di temi e non una vera ontologia, e la granularità dell'analisi testuale è assai grossa: il componimento nella sua interezza o al massimo le macrostrutture dello stesso. Ma si tratta del primo tentativo sistematico di procedere a una annotazione semantica di risorse digitali letterarie sui vasta scala.

"gerarchie sovrapposte" che caratterizza XML (e ogni altro linguaggio di markup basato su grammatiche non contestuali trattabili).[29]

Riteniamo che una strategia equilibrata per eseguire questa marcatura debba adottare sia tecniche tradizionali di *inline markup* con XML/TEI, sia tecniche di *stand-off markup*, in base al dettaglio dell'analisi e alla granularità delle unità discorsive. La scelta di tecniche di annotazione a molteplici livelli è resa praticabile da un nuovo formalismo di recente proposto per questo genere di trattamenti. Si tratta di *Open Annotation Data Model* (OA), un framework basato sul data model RDF finalizzato all'espressione di annotazioni di risorse digitali in modo interoperabile:

> An annotation is considered to be a set of connected resources, typically including a body and target, and conveys that the body is related to the target. The exact nature of this relationship changes according to the intention of the annotation, but most frequently conveys that the body is somehow "about" the target. Other possible relationships include that the body is an identifier for the target, provides a representation of the target, or classifies the target in some way […].[30]

Questo modello ha anche il vantaggio di rendere possibile l'attribuzione di una meta-tipologia e di dettagliati metadati di provenienza (responsabilità, livello di certezza, data) alle annotazioni stesse.

La realizzazione di una simile architettura è ovviamente assai complessa tecnicamente e altrettanto onerosa in termini di tempo e risorse (e tuttavia, se venti anni fa qualcuno avesse detto che oggi avremmo avuto a disposizione in formato digitale, opportunamente codificate in XML/TEI, intere tradizioni letterarie, non si sarebbe forse manifestato il medesimo scetticismo?).

La storia e l'evoluzione del Web ha dimostrato che non solo è possibile costruire sistemi, anche di enorme complessità, attraverso un processo pubblico incrementale e cooperativo, ma che tale strategia si dimostra assai più efficiente ed efficace di quelle private, monolitiche e centralizzate. Moltissimo lavoro nella costruzione della macro-architettura che qui proponiamo potrebbe essere condotto usando sistemi di cosiddetto *crowdsourcing* guidato, posto che esistano le opportune infrastrutture abilitanti. Il modello del *social tagging*, opportunamente corretto mediante sistemi basati su ontologie che ne orientino e controllino l'applicazione, permetterebbe di coinvolgere studiosi esperti ma anche giovani ricercatori e cultori nel costruire e popolare le ontologie. Un simile sforzo intellettuale e tecnologico non potrebbe che essere condotto in questo modo. E il prodotto di un tale impresa non potrà che essere un bene comune, un contenuto aperto e disponibile per tutta la comunità degli studi letterari.

[29] Su questo rimandiamo di nuovo a F. CIOTTI, *La rappresentazione digitale*….

[30] R. SANDERSON, P. CICCARESE, H. VAN DE SOMPEL, (a c. di), *Open Annotation Data Model*, W3C, 2013, <http://www.openannotation.org/spec/core/20130208/index.html>.