

UNIVERSITA' DEGLI STUDI DI BOLOGNA

FACOLTA' DI LETTERE E FILOSOFIA

Corso di laurea in **Filosofia**

**“Gli OPAC e la ricerca umanistica sul web”**

Tesi di laurea in **Informatica per le Scienze Umane**

Relatore  
Prof. Dino Buzzetti

Presentata da  
Fumagalli Antonella Carolina

Sessione Seconda  
Anno Accademico 2007/2008

A mio padre  
Carletto Fumagalli

maestro elementare  
(30.06.1924-05.07.1976)

Medaglia d'oro della  
Pubblica Istruzione

## INDICE

1. Premessa: perché questa tesi	1
2. Gli OPAC	3
2.1. Dal tavolino di Ranganathan agli OPAC	3
2.2. Breve storia degli OPAC: cosa sono, a cosa servono, come funzionano (breve cenni)	6
2.3. Il problema del recupero dell'informazione da dati digitali in generale (i motori di ricerca)	11
2.4. Quando nasce una nuova esigenza	16
3. Dal WWW al Semantic Web	17
3.1. Il World Wide Web	17
3.2. Il Semantic Web	20
3.3. Le ontologie (due teorie)	27
4. La ricerca umanistica sul web	39
4.1. L'Informatica Umanistica	39
4.2. Il ruolo dei filosofi (o degli umanisti in generale)	43
5. Conclusioni	50
6. Bibliografia	I
7. Ringraziamenti	III

## 1. Premessa: perché questa tesi

Più di venti anni fa (settembre 1984), quando mi sono iscritta per la prima volta al Corso di Laurea in Filosofia della Facoltà di Lettere e Filosofia dell'Università di Bologna, i computer iniziavano ad apparire nelle scuole e università solo negli uffici ed eravamo ancora lontani dagli albori del web in Italia.

Per trovare un libro era necessario recarsi di biblioteca in biblioteca per consultare il catalogo cartaceo per autore o per soggetto. E quando si trovava bisognava iscriversi alla biblioteca dove si era reperito il testo per poter avere o la possibilità di consultazione o la possibilità di accedere al prestito.

Ovviamente tutto ciò prendeva tempo per gli spostamenti a volte tra città diverse, e denaro per le conseguenti trasferte.

Consultare il web e mandare e-mail diventa una realtà a partire dagli anni '90, ma all'inizio solo per pochi iniziati. Dai primi anni '80 del Novecento le biblioteche decidono di dotarsi di cataloghi informatizzati prima per uso interno, poi anche da pubblicare sul web per consentire la ricerca dei testi anche da Internet, con conseguente risparmio di tempo e denaro da parte dell'utenza.

Nel settembre 2003, all'atto della mia nuova iscrizione al corso di laurea in Filosofia (in precedenza era subentrata la decadenza dal corso di Laurea per decorrenza dei termini di iscrizione dovuta ai miei pressanti impegni di lavoro) scopro<sup>1</sup> che non solo la ricerca dei testi è facilitata dal fatto che le biblioteche si sono riunite sotto alcuni OPAC (Online Public Access Catalogue) per cui la ricerca di un testo si effettua direttamente con un'unica interrogazione su diverse biblioteche e su diverse città, ma anche che le diverse università si sono dotate di portali sul web in cui offrono i più svariati servizi agli studenti, dalla consultazione della propria carriera scolastica fino alla prenotazione degli esami online.

Scopro anche con piacere che le Università umanistiche solitamente hanno portali molto più ordinati, chiari, comprensibili e consultabili che non le Università Tecnico-Scientifiche, segno di una progettazione più vicina alle esigenze degli studenti e dei docenti, e magari hanno anche servizi più semplici ma più utili per gli utenti e che anche i Docenti di estrazione umanistica utilizzano in modo costante la posta elettronica per comunicare con i propri studenti.

Allora mi domando quale sia la correlazione tra Facoltà umanistica e creatori del portale, se non vi sia non solo collaborazione tra tecnici del web (i cosiddetti webmasters) e i ricercatori umanisti, ma anche una certa sovrapposizione di ruoli, dove l'umanista non solo utilizza il web ma ne crea e ne modifica l'aspetto

---

<sup>1</sup> Avevo fino ad allora utilizzato Internet e i motori di ricerca solo per scopi lavorativi, tutti inerenti alle problematiche hardware e software del mio lavoro e sempre solo per necessità, per una sorta di diffidenza personale verso quell'oceano di informazioni caotico e poco consultabile che era ed è il web.

e i contenuti e viene da esso modificato nel suo lavoro e nel suo atteggiamento verso le nuove tecnologie.

Ovviamente anche l'umanista dovrà subire l'influenza del web sul suo lavoro e soprattutto dovrà, nell'impostazione di strumenti informatici che possano servire al suo lavoro, operare delle scelte per rispettare la formalizzazione e la modellazione che il mezzo informatico richiede.

Ricordo che di solito le nuove tecnologie e i computer in particolare erano, prima degli anni '90 del Novecento, guardati con sospetto e diffidenza dalle persone che vivevano ed esercitavano la loro professione in ambito umanistico, tra cui anche la sottoscritta.

Poi il mio lavoro<sup>2</sup> mi ha portato ad avvicinarmi in modo differente al mezzo informatico, scoprendone la potenzialità e la versatilità.

Allora perché questa tesi in filosofia?

Per capire, da informatico quale sono nella vita lavorativa:

- quali sono gli strumenti informatici messi a disposizione dal web al ricercatore in ambito umanistico;
- quali sono le limitazioni che il web impone alla ricerca;
- quali sono le prospettive future offerte alla ricerca umanistica dal web.

Per capire, da laureanda in filosofia<sup>3</sup>:

- quali sono le competenze richieste all'umanista per effettuare una ricerca sul web;
- quali invece le competenze che deve acquisire l'umanista per poter impostare in modo corretto gli strumenti del web in modo da consentire una ricerca che sia il più rispondente alle esigenze dei ricercatori;
- quale è il futuro della ricerca umanistica.

Per capire in generale qual'è la reale fruibilità delle informazioni sul web e quale futuro ha lo sviluppo del Semantic Web<sup>4</sup>.

Questa tesi non vuole affrontare le problematiche tecniche relative all'implementazione di linguaggi e programmi per il web, ma vuole affrontare da un punto di vista filosofico quali sono le tematiche di discussione e di sviluppo delle nuove potenzialità offerte dai mezzi informatici alla ricerca umanistica e quindi ai filosofi stessi.

---

<sup>2</sup> Sono Assistente Tecnico di Informatica con compiti di Amministratore di rete presso l'Istituto Tecnico Industriale Statale "Janello Torriani" di Cremona dal 1993 ([www.itistorriani.it](http://www.itistorriani.it)).

<sup>3</sup> Al momento in cui scrivo sono laureanda in Filosofia presso la Facoltà di Lettere e Filosofia dell'Università degli Studi di Bologna (<http://www.lettere.unibo.it/Lettere/default.htm>).

<sup>4</sup> Termine introdotto da Tim Berners Lee nel 2001, da "The Semantic Web" in "Scientific American".

## 2. Gli OPAC

### 2.1. Dal tavolino di Ranganathan agli OPAC

L'indiano Shiyali Ramamrita Ranganathan (India 1892-1972), formatosi come matematico, diventa uno dei più grandi bibliotecari del ventesimo secolo. Si avvicina alle biblioteche del suo paese in modo casuale, ma in occasione di un viaggio in Europa rimane colpito dalla realtà delle biblioteche occidentali.

Introduce idee di grande importanza per la biblioteconomia moderna:

- A. Le Cinque Leggi della biblioteconomia
- B. Lo spirito del *reference*
- C. La classificazione a faccette
- D. Una concezione umanistica della biblioteca

Vediamole velocemente nel dettaglio.

**A. Le Cinque Leggi della Biblioteconomia** sono cinque principi semplici:

1. I libri sono fatti per essere usati
2. Ad ogni lettore il suo libro
3. Ad ogni libro il suo lettore
4. Non far perdere tempo al lettore
5. La biblioteca é un organismo che cresce

Egli considera, tramite queste leggi che pone a fondamento della biblioteconomia, come centro della biblioteca l'uomo che se ne serve, il lettore: è un principio a suo modo rivoluzionario. Ogni operazione che si svolge in biblioteca deve tenere conto dell'utilità che avrà per le persone che la frequentano, quindi anche la catalogazione e la burocrazia delle procedure dovranno sempre tener conto dell'utente finale.

#### **B. Lo spirito del *reference***

*Reference* significa riferimento, relazione o consultazione.

Per *servizio di reference* in biblioteca si intendono tutte le attività dei bibliotecari volte ad aiutare gli utenti a trovare quello che stanno cercando: quali sono le fonti più appropriate da consultare, le migliori

strategie per usare i cataloghi e i repertori e come sviluppare un percorso che li porti all'acquisizione delle conoscenze che li interessano.

Nella realtà l'utente si trova molto spesso disorientato al suo ingresso in una biblioteca, e anche se si attiene alla prassi non è sicuro di seguire la via giusta per arrivare alle informazioni che cerca.

Questo provoca situazioni in cui il patrimonio della biblioteca viene poco sfruttato il che per l'utente si tramuta in una possibile perdita di informazioni.

I cataloghi e le procedure delle biblioteche per un utente novizio sono sempre uno strumento complesso.

La funzione dei bibliotecari di reference è quella di fare da punto di riferimento per gli utenti, di guidarli e assisterli nelle loro ricerche, facendo da tramite tra la mole di informazioni racchiuse nei libri e il bisogno di conoscenza delle persone. Oggi questo servizio è considerato una parte fondamentale della biblioteconomia moderna, e si è esteso anche nella vita di tutti i giorni (anche i grandi store<sup>5</sup> hanno al loro interno un incaricato del servizio di reference).

### **C. La classificazione a faccette**

Uno schema di classificazione, per Ranganathan, deve essere utilizzato in biblioteca in modo integrato: sia per realizzare un catalogo da consultare, sia per disporre i volumi negli scaffali secondo un ordine adeguato che permetta agli utenti, ma anche al bibliotecario stesso, la localizzazione di quei volumi che interessano o si stanno cercando.

Ranganathan studiò un sistema meno rigido e più articolato di quelli già esistenti da decenni (come la tradizionale Classificazione Decimale Dewey che risale al 1876 – CDD).

Una faccetta, in inglese facet, è un particolare aspetto sotto il quale un argomento viene trattato: le faccette di qualsiasi classe si possono ricondurre a cinque categorie fondamentali:

- *personalità* (l'oggetto centrale di un discorso)
- *materia* (i componenti e le proprietà dell'oggetto)
- *energia* (la caratteristiche dinamiche dei processi che lo interessano)

---

<sup>5</sup> Come possono essere Rizzoli a Milano, Feltrinelli nei suoi svariati punti vendita (tra cui anche quello di Cremona, mia città natale), o anche i grossi Ipermercati e Centri commerciali, ovviamente in questi casi per ragioni economiche.

- *spazio* (i suoi elementi geografici o spaziali)
- *tempo* (le sue fasi cronologiche)

Con questa classificazione il contenuto del documento può essere descritto analiticamente nei suoi diversi aspetti, che sono poi espressi tutti insieme secondo una sequenza determinata da regole di funzionalità: per questo Ranganathan definisce questa classificazione di tipo *analitico-sintetico*.

La classificazione a faccette è nota come Colon Classification, per la caratteristica frequenza con cui ricorre nella sua notazione il simbolo di due punti (in inglese “colon”).

Si tratta di un sistema raffinato e complesso, ma con una grandissima importanza teorica: ripreso da autorevoli studiosi ha fornito le basi per lo sviluppo di sistemi avanzati di indicizzazione: thesauri, classificazioni a faccette speciali e generali (tra cui la seconda ed. della Classificazione Bliss: 1935÷1953) e del PRECIS<sup>6</sup>, i cui principi hanno dato forma ai lavori del Gruppo Italiano di ricerca sull’indicizzazione per soggetto (GRIS<sup>7</sup>).

#### **D. Una concezione umanista della biblioteca**

La funzione ultima della biblioteca è quella di permettere l’accesso alla conoscenza, e quindi lavorare perché questo si realizzi effettivamente: la possibilità per ciascuna persona di beneficiare della conoscenza disponibile.

L’opera di Ranganathan ci propone una concezione umanista della biblioteca e la inquadra in una prospettiva più ampia.

Dopo l’introduzione dei computers negli uffici e nelle biblioteche nasce l’esigenza di trasferire il catalogo cartaceo su supporto elettronico per garantirne la facilità di aggiornamento, la conservazione su supporti informatici e soprattutto per poter fornire servizi più rapidi ed efficienti di consultazione, ricerca e prestito.

All’inizio ogni biblioteca crea il suo catalogo elettronico per uso interno. Si tratta in genere di una base di dati (DataBase) organizzata in record formati da campi, in cui ogni testo o documento è registrato sulla base della scheda catalografica (livello bibliografico, tipo di

---

<sup>6</sup> Nuovo soggettario della lingua italiana.

<sup>7</sup> Attivo dal 1990, e fra il dicembre 2003 e il 2005 confluito nella Commissione nazionale Catalogazione e indicizzazione, il GRIS si è ricostituito nel 2006 come gruppo di studio autonomo.

documento, titolo, pubblicazione, descrizione fisica, numeri, nomi, soggetti, classificazione, paese di pubblicazione, lingua, ecc..).

Nasce poi l'esigenza, con l'avvento del web, di pubblicare in rete i cataloghi e quindi di stabilire delle specifiche quanto più possibili comuni per l'indicizzazione dei cataloghi.

## **2.2. Breve storia degli OPAC: cosa sono, a cosa servono, come funzionano (brevi cenni)**

Il catalogo elettronico noto come OPAC (Online Public Access Catalogue, ossia catalogo in linea accessibile pubblicamente), consente di interrogare le biblioteche aderenti per sapere quale ente possiede un determinato documento o titolo, utilizzando chiavi di interrogazione basate sugli elementi della stringa catalografica.

Gli OPAC sono veri e propri cataloghi dotati di descrizioni bibliografiche complete e non sono meri elenchi alfabetici di titoli; essi, sfruttando le potenzialità della loro natura informatica, sono interrogabili automaticamente, digitando direttamente i termini cercati.

Gli OPAC sono principalmente sistemi di recupero dell'informazione adottati dalle biblioteche per dare accesso alla informazione in linea agli utenti mettendo a disposizione la conoscenza del proprio patrimonio documentario; sono quindi gli strumenti principali con cui attuare un'efficace politica di diffusione delle informazioni bibliografiche e servizi più efficienti di circolazione dei documenti sul territorio.

Tecnicamente un OPAC consiste di una base di dati strutturata interrogabile attraverso un apposito linguaggio. Ogni documento è registrato in unità dette *record*. Ogni record è la descrizione di un'opera: ne indica le caratteristiche bibliografiche (titolo, autori, curatori, numero dell'edizione, luogo e data di pubblicazione, editore, caratteristiche fisiche) e inoltre fornisce informazioni sulle copie dell'opera possedute dalla biblioteca in questione (numero di esemplari, collocazione, disponibilità o meno al prestito). Tutte queste informazioni sono organizzate in *campi*, cioè diverse parti del record, ciascuna contenente una porzione della descrizione dell'opera. L'insieme di tutti i record del catalogo, ciascuno suddiviso in campi,

forma virtualmente una grandissima griglia di informazioni, chiamata *database*, cioè base di dati o banca dati. La struttura dei record è di solito leggermente diversa a seconda che si tratti di opere monografiche, oppure di periodici; inoltre di opere singole oppure divise in più volumi o più parti.

Nascono alla fine degli anni '80 del Novecento, e all'epoca gli unici strumenti per l'accesso al computer remoto erano i sistemi di emulazione terminale (telnet) tramite il protocollo Z39.50, che fu sviluppato per far interagire un database e un modulo di ricerca senza conoscere la particolare sintassi di ricerca del database. La ricerca però era comunque ristretta ad una cerchia di esperti perché i comandi di interrogazione del database avvenivano tramite riga di comando e bisognava comunque conoscere un minimo di sintassi di ricerca per poter ottenere dei risultati soddisfacenti.

Attualmente tramite il browser web<sup>8</sup> è possibile interrogare uno o più database contemporaneamente grazie ad una interfaccia sicuramente più amichevole (users-friendly).

Interrogare un OPAC consiste essenzialmente nel verificare se una o più parole sono contenute in qualcuno dei record che lo costituiscono. L'utente dovrà quindi inserire le parole desiderate negli appositi spazi visibili sullo schermo; il sistema risponderà visualizzando quei record che contengono le parole richieste, i quali descrivono i volumi posseduti dalle biblioteche interrogate.

Esistono diverse modalità di ricerca:

- una *ricerca per campi*: in questo tipo di ricerca si richiede che determinati termini siano presenti all'interno di determinati campi; in questo modo si ottiene in risposta direttamente l'insieme dei documenti che soddisfano le condizioni specificate (campo autore = Leopardi Giacomo, si ottiene la lista di tutti i documenti scritti da Leopardi)
- una *ricerca per liste*: qui si richiede che il contenuto di un determinato campo inizi con dei termini specificati, e si ottiene in risposta un elenco di voci in ordine alfabetico contenente quella che risponde meglio ai requisiti; a partire dalla lista, in cui viene comunque mostrato anche un certo

---

<sup>8</sup> Programma che consente la navigazione nel web e tra documenti ipertestuali.

numero di voci precedenti e successive, si possono poi visualizzare le descrizioni dei documenti corrispondenti.

Se una ricerca produce una grande quantità di risultati, la maggior parte dei quali non sembrano interessanti, piuttosto che scorrerli tutti conviene effettuare un'altra ricerca utilizzando dei termini più specifici: scegliere parole meno comuni per i titoli e i soggetti, specificare l'autore (eventualmente anche il nome, se il cognome è troppo comune), specificare un intervallo di date, utilizzare più campi contemporaneamente.

Gli OPAC offrono grandi vantaggi: possono essere consultati da qualsiasi computer in rete, in tutto il mondo; permettono di cercare lo stesso documento in più biblioteche, anche contemporaneamente e di recuperare l'informazione in breve tempo (come si dice in gergo informatico: nel tempo di un click).

In più consentono di superare alcune limitazioni del catalogo cartaceo:

- il problema della *scelta* dell'intestazione della scheda catalografica: il record catalografico è interrogabile a partire dai diversi campi che costituiscono la scheda e quindi non è più necessario definire a priori le chiavi di accesso;
- il problema della *forma* da assegnare all'intestazione, o alle chiavi di accesso: viene superato individuando le forme accettate e collegando a queste le forme varianti, tramite l'impiego di rimandi invisibili all'utente. Tutto ciò è possibile grazie agli *authority files* cioè agli archivi standardizzati di informazioni (per autore, per soggetto, per titolo,..) collegati tramite rimandi e rinvii alle altre forme ipoteticamente utilizzabili nella ricerca;
- la divisione del catalogo per autori da quello per soggetti e per materie;
- l'estensione delle chiavi di accesso all'informazione, con l'aggiunta di *key words* (parole chiave) desunte da titolo, introduzione, sommario o da altre zone significative del documento, in uno specifico campo della registrazione catalografica. La ricerca dell'informazione ha così la certezza del recupero.

L'uso degli OPAC consente all'utente di reperire non solo un determinato titolo, ma anche tutte le opere di un determinato autore o verificare quali pubblicazioni condividono lo stesso argomento.

L'evoluzione dei servizi offerti dagli OPAC si può ricondurre ad una serie di requisiti ormai indispensabili per un OPAC:

- interrogazione del catalogo in varie modalità
- connessione coi dati informativi/posseduto
- disponibilità, prestito, prenotazione, fotocopie
- scarico/stampa/e-mail dei risultati delle ricerche
- connessione a banche dati bibliografiche e full-text<sup>9</sup>
- cattura dei dati (catalogazione derivata, bibliografie).

Il Servizio Bibliotecario Nazionale (SBN) è la rete informatizzata di servizi nazionali alla quale sono collegate biblioteche dello Stato, degli Enti locali, delle Università, di accademie pubbliche e private che contribuiscono alla creazione del catalogo collettivo nazionale in linea. Gli istituti bibliotecari sono organizzati in poli che gestiscono un catalogo collettivo locale il quale confluisce poi nell'Indice SBN, il catalogo unico nazionale che è gestito dall'ICCU (Istituto Centrale per il Catalogo Unico delle biblioteche italiane e delle informazioni bibliografiche: [www.iccu.sbn.it](http://www.iccu.sbn.it)). Il progetto risalente al 1982, è entrato in fase operativa nel 1992.

Questo servizio viene alimentato tramite la catalogazione partecipata. Un documento viene catalogato solo dalla prima biblioteca; le altre si limitano a catturare la copia della scheda catalogografica e ad aggiungervi la propria localizzazione (tramite un codice identificativo della biblioteca). Le banche dati sono suddivise e organizzate per tipologie di documenti:

- *Libro moderno*, cioè il catalogo dei testi a stampa;
- *Libro antico*, cioè il catalogo dei testi editi dall'invenzione della stampa fino al 1830;
- *Beni musicali, edizioni e manoscritti musicali*
- *Manoscritti*, cioè descrizioni di testi in alfabeto latino
- *Anagrafe delle biblioteche italiane*, fornisce informazioni su quindicimila biblioteche italiane

---

<sup>9</sup> E' possibile la consultazione dei sommari delle riviste attraverso Servizi Table of Contents (ToC) offerti da Società, Consorzi di Biblioteche, Centri Specializzati, Case Editrici.

- *Censimento delle edizioni italiane del XVI secolo*, ha lo scopo di documentare la produzione italiana a stampa del XVI secolo e di effettuare una ricognizione del patrimonio posseduto a livello nazionale;
- *letteratura grigia e spoglio periodici*, basi di dati specializzate;
- *Discoteca di Stato*, che comprende documenti sonori registrati e registrazioni sonore.

Il miglior repertorio di OPAC italiani è reperibile sul sito dell'AIB (Associazione Italiana Biblioteche, [www.aib.it](http://www.aib.it)), la quale ha anche realizzato il MetaOPAC Azalai<sup>10</sup> Italiano (MAI, <http://azalai.cilea.it/mai>), che permette di selezionare in anticipo quali cataloghi interrogare e fornisce una maschera in cui è possibile specificare i termini della ricerca. Permette una ricerca *multithreading* su più OPAC remoti in contemporanea. Lanciando delle *query* (interrogazioni) in parallelo, riporta all'utente, in una pagina di risposta unica, le informazioni dislocate in punti differenti della rete: unica è anche l'interfaccia di riferimento, dove l'utente esegue la sua ricerca su più OPAC differenti.

Esistono diversi MetaOPAC che differiscono l'uno dall'altro per diversi aspetti:

- i raggruppamenti di OPAC che le MetaInterfacce vanno ad interrogare, creano cataloghi Collettivi Virtuali differenti per contenuto (area geografica, ambito disciplinare, set preconfezionato disponibile);
- i software utilizzati sono diversi da prodotto a prodotto e possono permettere l'interrogazione di cataloghi simili per peculiarità tecniche (stessa tipologia) oppure di cataloghi eterogenei;
- le MetaInterfacce comprendono talvolta anche cataloghi di altri insiemi informativi (banche dati, cataloghi editoriali) e non solo di biblioteche;

---

<sup>10</sup> Azalai era il nome della Carovana del Sale dai Mille Cammelli, parola che nell'antico linguaggio Tuareg ("Tifinagh" o "Tifinar") significava "separarsi per poi ricongiungersi di nuovo". L'Azalai impiegava nove mesi lungo il deserto per trasportare il suo carico di sale che scambiava con l'oro. Questa metafora (il valore del sale della conoscenza scambiato al pari dell'oro) delinea il faticoso percorso intrapreso dal settembre 1998 dai membri dell'AIB-WEB, tragitto durato in tutto esattamente nove mesi, per portare l'Azalai alla sua presentazione ufficiale a Roma il 18 maggio 1999 al Congresso Nazionale AIB.

- differiscono per approccio di interrogazione, per grafica e per ciò che sta intorno alla MetaInterfaccia (servizi di supporto);
- le modalità di ricerca e la velocità di risposta sono riconducibili alle potenzialità del motore di ricerca che ci sta dietro.

Esistono anche MultiOPAC che si distinguono dai MetaOPAC in quanto consentono di interrogare Cataloghi singoli o collettivi attraverso un'unica interfaccia, ma non contemporaneamente.

Solitamente è possibile contrassegnare o selezionare da un menu a tendina il Catalogo su cui si vuole effettuare la ricerca. L'interfaccia unica di riferimento per più Cataloghi, agevola l'utente nell'accesso all'informazione.

### **2.3. Il problema del recupero dell'informazione<sup>11</sup> da dati digitali in generale (i motori di ricerca)**

Tutti ormai utilizziamo il Web per trovare documenti utili alla nostra vita professionale o personale. Per far ciò sono disponibili quattro strade: seguire i link da una pagina all'altra fino a trovare quello che cercavamo, servirci di un motore di ricerca per ottenere una lista di link tra i quali scegliere quelli che ci interessano, usare una directory o un catalogo di risorse web oppure digitare direttamente l'indirizzo della risorsa che ci interessa nel browser.

Vantaggi e svantaggi dei quattro metodi:

- seguire i link ipertestuali è sicuramente un processo cognitivamente più ricco: tuttavia tale processo può essere dispendioso in termini di tempo e difficile da avviare se non sappiamo da dove partire;
- usare i motori di ricerca ha il vantaggio di richiedere poca informazione in partenza: ma i motori di ricerca non coprono tutto il contenuto del web (si parla di un 80% di "hidden web" o web nascosto formato da contenuto non indicizzato o non indicizzabile per motivi tecnici) e la ricerca per parole chiave può risultare deprimente per l'elevato numero di "falsi positivi" (pagine che contengono la parola da noi definita ma

---

<sup>11</sup> In inglese "Information retrieval": letteralmente "recupero di informazioni (disperse)".

che non parlano di quello che ci interessa) e di “falsi negativi” (nel risultato non compaiono pagine che sarebbero state di nostro interesse, ma che contengono non la parola da noi inserita ma un suo sinonimo);

- nelle directory il mondo viene strutturato in categorie e sottocategorie e scegliendo il settore che ci interessa abbiamo a disposizione solo quei siti che i recensori hanno individuato per noi. Ciò funziona se dobbiamo cercare una risorsa molto nota, come un sito istituzionale, ma è inadatto se stiamo cercando informazioni su un particolare tema generico o generale;
- digitare il link della risorsa nel browser è possibile solo se si conosce tale link, ma allora non stiamo più effettuando ricerche ma semplicemente stiamo visitando un sito conosciuto.

In più la ricerca sul web restituisce sempre documenti e non esattamente l'informazione che stavamo cercando.

Spesso diverse parti dell'informazione che ci interessa sono contenute in documenti differenti, disponibili magari su siti diversi. Nulla ci permette di combinare informazione proveniente da fonti diverse in un piano che risolva il nostro problema: si parla in questo caso di mancanza di *integrazione di informazione*.

Infine il web attuale non permette la *cooperazione* tra programmi e tra programmi ed utenti per risolvere problemi complessi. La maggior parte dei siti web è progettata come contenitore di informazioni che possono essere estratte a richiesta ma non fornisce servizi ad altri servizi. Molte applicazioni richiederebbero che i siti di varie organizzazioni potessero interagire in modo flessibile e dinamico, e che gli utenti potessero intervenire nel processo interagendo con i programmi dei vari siti.

Ma come funzionano i motori di ricerca??

I motori di ricerca (*search engine*) usano degli agenti software detti *spiders* o *crawlers* che navigano per la rete e riportano tutte le risorse che incontrano, privilegiando quelle modificate più di recente. Nei *repository* dei servers avviene una catalogazione che consente ai motori di ricerca di costruire una sorta di “fotografia” del web,

realizzando un indice inverso di tutto quello che trovano. Utilizzano solo strumenti di tipo sintattico, andando ad analizzare i tags (etichette) inseriti nei files HTML pubblicati nel web, i quali descrivono il contenuto del documento. Essi indicizzano il web in modo automatico utilizzando un applicativo e quindi non operano nessun tipo di selezione qualitativa delle risorse.

I motori di ricerca sono quindi costituiti da:

- un *programma* che interroga il web e indicizza le pagine estraendo parole;
- un *database* che raccoglie e archivia le parole estratte inserendole nel contesto della pagina in cui occorrono (si tratta perciò di una raccolta strutturata di dati omogenei);
- un'*interfaccia utente* con la quale è possibile interrogare tale database di parole e quindi di pagine.

La validità di un motore di ricerca è valutata in base a due parametri:

- *precisione*, del risultato dell'interrogazione in relazione al rapporto "totale pagine trovate/totale pagine rilevanti";
- *recall*, cioè la capacità di un motore di ricerca di trovare il maggior numero di pagine rilevanti rispetto a quelle trovate.

Esiste anche un altro parametro che è il *limite di utilità (futility point)* che consiste nel numero di risultati oltre il quale l'utente si stanca di esaminarli uno per uno. Secondo David C. Blair<sup>12</sup> è circa 30. Esso dipende dalla motivazione, dal tempo e dagli scopi dell'utente.

Dobbiamo anche considerare:

- che esiste un *deep web*, un web profondo o meglio nascosto (di cui ho già accennato in precedenza) che i motori di ricerca non possono interrogare;
- le pagine non più aggiornate o non più esistenti rimangono comunque nel database del motore di ricerca se non esiste una procedura di *refresh* (aggiornamento) dei dati in archivio;

---

<sup>12</sup> Professore di Business Information Technology presso la Stephen M. Ross School of Business University of Michigan.

- la *query* (interrogazione) con parole chiave non è detto che restituisca l'argomento cercato: esistono problemi di polisemia del linguaggio naturale e di sinonimia delle parole usate come chiavi di interrogazione che ostacolano non poco la precisione del risultato;
- in aggiunta, l'utilizzo di un linguaggio formale nella query, tramite gli operatori di ricerca (operatori logici AND, OR, NOT, oppure le notazioni di troncamento virgil\*, o le virgolette “parola e parola ” per cercare una frase esattamente come viene scritta) rende a volte molto difficile porre una query che sia esaustiva ma non ridondante.

Google™ ([www.google.it](http://www.google.it))<sup>13</sup> rappresenta sicuramente il motore di ricerca più utilizzato dagli utenti. Le sue caratteristiche principali sono: la *velocità* dei tempi di risposta all'interrogazione dell'utente, il *numero di pagine* archiviate nel database e la tecnica matematica del *relevance ranking* nella modalità di selezione dei risultati restituiti all'utente. Ogni pagina viene valutata da Google™ in base a una serie di parametri, tra i quali il numero di link che puntano alla risorsa: più un sito o una risorsa sono linkati da altri, più in alto sarà nella lista dei risultati forniti dal motore. Per questa valutazione viene considerata anche l'autorevolezza dei siti che linkano una pagina o un altro sito. Google™ ha trovato quindi un modo per tenere conto della struttura sociale delle connessioni tra le pagine.

Ai fini del *pagerank*<sup>14</sup> è anche considerata da tutti i motori di ricerca la presenza del termine usato nell'interrogazione dell'utente all'interno delle parole utilizzate dal creatore per definire i metadati (dati che parlano dei dati ed esprimono quello che c'è da sapere su un certo insieme di dati).

Altre funzionalità di Google™ sono: la *ricerca avanzata*, la *ricerca di immagini*, la *ricerca di libri* (che permette di effettuare una interrogazione sul testo pieno o completo, detto full-text, di numerosi libri digitalizzati, provenienti da biblioteche ed editori di tutto il mondo).

---

<sup>13</sup> Realizzato nel 1997 da due studenti californiani, Sergey Brin e Lawrence Page. Dal 2000 è diventato uno dei più grandi portali sul web.

<sup>14</sup> Il **PageRank** è un algoritmo di analisi che assegna un peso numerico ad ogni elemento di un collegamento ipertestuale d'un insieme di documenti, come ad esempio il World Wide Web, con lo scopo di quantificare la sua importanza relativa all'interno della serie.

Vale la pena di evidenziare che esiste anche il problema della reperibilità delle informazioni multimediali (o *multimedia information retrieval*, o MMIR) dovuto alla constatazione di un processo di rivoluzione dei documenti digitali e dei processi informativi e comunicativi.

Fino a non molto tempo fa l'*information retrieval* si è prospettato come un dominio privilegiato dei bibliotecari, dei documentaristi e degli specialisti dell'informazione, che avevano adottato un'interpretazione *user-centered* focalizzando l'attenzione sui modi testuali con cui gli utenti interpretano e utilizzano l'informazione, in opposizione alle modalità automatiche di strutturazione, immagazzinamento e recupero dei dati.

Il delinarsi di nuove modalità di comunicazione e condivisione di informazioni ha dato luogo a quella che è stata definita la "convergenza al digitale".

Le multimedia digital library sono caratterizzate da una gestione e un accesso integrati per documenti tipologicamente eterogenei, attraverso l'impiego di specifici sistemi per l'indicizzazione, la ricerca e l'estrazione automatica di dati rappresentativi del complesso contenuto dei documenti multimediali (immagini, audio, video), che si vanno ad affiancare ai tradizionali sistemi di analisi e indicizzazione terminologica dei documenti testuali o audiovisivi.

Si possono attualmente distinguere:

- sistemi di *information retrieval term-based*, basati su informazioni testuali (termini estratti dal linguaggio naturale, schemi di classificazione e soggettazione, thesauri) per la ricerca di documenti testuali;
- sistemi di *information retrieval content-based* o *multimedia information retrieval* suddivisibili in:
  - sistemi di *visual retrieval*, in cui i file di immagini sono cercati e recuperati tramite dati visivi interni ai file, quali colore, forma, orientamento e distribuzione spaziale, ecc;
  - sistemi di *video retrieval*, dove per il recupero di documenti audiovisivi si utilizzano elementi di ricerca ricavati dalle immagini del filmato, dal movimento degli oggetti nelle inquadrature, dall'analisi degli stacchi di montaggio o della traccia sonora;

- sistemi di *audio retrieval*, nei quali l'informazione sonora è ricercata in misure di suoni, ricavando i dati dall'analisi dei volumi, delle sonorità, dei ritmi o delle melodie.

Tuttavia per ottenere un buon livello di precisione nel recupero dei documenti da una base multimediale sembra auspicabile la compresenza dei due sistemi: *information retrieval term-based* e *information retrieval content-based*; il primo infatti può costituire un ottimo metodo preliminare di selezione dei documenti e può consentire di centrare la ricerca in base agli ambiti di appartenenza, le tipologie, le classi, i titoli e gli autori e successivamente può essere un sistema di ripulitura finale dal rumore specifico di un'interrogazione content-based. In tutto ciò i due procedimenti possono operare in armonia e in interazione costante, in un'unica interfaccia utente.

## 2.4. Quando nasce una nuova esigenza

L'assetto del web dà l'idea, per le ragioni già sopra esposte, di un luogo effimero e caotico dove si ha una spiacevole sensazione di disorientamento all'interno del materiale sovrabbondante, eccessivo e perciò dispersivo che vi è pubblicato. Quello del web è un universo precario, instabile, soggetto a continue revisioni ma senza memoria storica. Quello che oggi c'è, domani potrebbe non esserci, e non ce n'è più traccia.

Significativo è il tentativo compiuto da Brewster Kahle<sup>15</sup> che col motore di ricerca Alexa ha istituito una organizzazione senza fini di lucro per creare un archivio di tutte le pagine web: attualmente quelle accumulate ammontano all'incirca a 100miliardi di pagine, corrispondenti a circa 100terabyte di materiale.

Nasce allora l'esigenza di un passo avanti per il web, di una evoluzione che apporti un maggior quantitativo di conoscenza che sia di qualità. Nasce l'esigenza del Web Semantico<sup>16</sup>(o Web 2.0).

---

<sup>15</sup> Fondatore di «Internet archive».

<sup>16</sup> Il termine "semantica" (dal greco *semantikos*, significato, derivato da *sema*, segno) è usato in molte accezioni, tutte attinenti al concetto di "significato di un messaggio", in senso ampio. La semantica è quella parte della linguistica che studia il significato delle parole (semantica lessicale), degli insiemi di parole, delle frasi (semantica frasale) e dei testi. Essa considera il rapporto tra l'espressione e la realtà extralinguistica. Altre specifiche accezioni del termine: in informatica (e in particolare in programmazione) la semantica di un linguaggio (di programmazione o di un programma) indica il significato operativo delle istruzioni e dei costrutti del linguaggio stesso; in logica matematica dare una semantica ad un sistema formale significa assegnare opportune classi di modelli alle formule del sistema.

## 3. Dal WWW al Semantic Web

### 3.1. Il World Wide Web

Anche se la rete Internet esiste già a partire dalla fine degli anni '60 del Novecento<sup>17</sup>, solo al principio degli anni '90 del Novecento il World Wide Web<sup>18</sup> nasce grazie al progetto di alcuni ricercatori, tra cui Tim Berners Lee (Londra, 1955) del CERN di Ginevra, di rendere pubblicamente accessibili documenti di testo interconnessi tra loro. Nasce così un insieme di documenti in formato ipertestuale<sup>19</sup>. La grande diffusione del Web è dovuta in larga parte anche alla nascita dei Personal Computer che si diffusero capillarmente nel corso degli anni '90. Vorrei qui ricordare che il Web è solo uno dei servizi offerti da Internet (e-mail, chat, forum, newsgroup, blog, wiki ed altro).

Caratteristiche fondamentali del Web sono:

- la *decentralizzazione delle risorse* (i dati sono residenti su macchine diverse, sparse in tutto il mondo);
- l'*universalità di accesso* (possibilità per qualunque macchina di accedere ai dati a prescindere dall'hardware e dal software in uso).

Risale al 1994 la nascita di un importante consorzio, il World Wide Web Consortium (W3C), impegnato nella creazione di specifiche (le cosiddette Recommendations o Guides Lines, raccomandazioni o linee guida) destinate alla definizione delle caratteristiche dei linguaggi e dei protocolli standard che possono essere condivisi fra le comunità del Web.

L'uso di questi standard garantisce l'*interoperabilità tecnica*, intesa come possibilità di consentire a sistemi diversi, a livello di hardware e software di dialogare tra loro.

Le specifiche del W3C raccomandano:

---

<sup>17</sup> Nasce per esigenze militari, in pieno periodo di guerra fredda, grazie agli sforzi dell'Agenzia ARPA (Advanced Research Project Agency) e del MIT (Massachusetts Institute of Technology). I primi nodi della rete ARPANET entrarono in funzione nel 1969. Solo nel 1972 prenderà il nome di Internet. Il protocollo TCP/IP (Transmission Control Protocol/Internet Protocol) che definiva le regole generali attraverso cui le macchine potevano comunicare tra di loro (inventato da Vinton Cerf e Robert Kahn) fu sperimentato nel 1977 ed entrò in funzione nel gennaio 1983.

<sup>18</sup> Una nuova frattura si inserisce tra i paesi sviluppati e quelli arretrati, il digital divide, che passa attraverso i diversi strati sociali e culturali delle società.

<sup>19</sup> Il linguaggio standard per la creazione e pubblicazione degli ipertesti è HTML (HyperText Markup Language). Ci saranno poi varie evoluzioni di questo linguaggio (XML, XMS, OWL).

- l'*indipendenza dal software*, importante per prevenire il rischio di un controllo monopolistico della rete;
- l'adozione di *standard per la codifica dei caratteri*<sup>20</sup>, che consentano la presenza sul Web delle diverse lingue naturali;
- la creazione e la diffusione di standard per l'accessibilità, che tutelano chi è svantaggiato in termini di capacità fisiche e chi dispone di macchine con tecnologie obsolete.

Lo scenario attuale è quello di un enorme insieme di testi collegati tra di loro. Ma i testi sono creati ad uso e consumo dei soli utenti umani, gli unici in grado di comprendere i contenuti delle pagine che visitano. I collegamenti fra pagine sono di due classi:

- i *collegamenti sintattici* che sono legati al funzionamento di un qualche codice di programmazione: sono piuttosto solidi<sup>21</sup>;
- i collegamenti che descrivono il *significato* di un collegamento: sono piuttosto deboli (generici e vaghi), e oltre a portare ad una determinata risorsa dovrebbero descrivere la risorsa verso cui portano (capacità semantica).

Solo gli utenti umani riescono ad orientarsi nel Web grazie all'esperienza di navigazione e alla capacità di evocazione di parole chiave significative. Nessuna applicazione è in grado di interpretare il contenuto delle pagine. I computers non sanno come affrontare il Web almeno per quanto attiene ai suoi contenuti. Essi non fanno altro che spostare informazioni da un luogo all'altro e mostrarcele sullo schermo: ma tutto il lavoro di comprensione, combinazione, interpretazione e giudizio è lasciato esclusivamente all'utente umano<sup>22</sup>.

Il Web non è quindi identificabile con la fisionomia di una biblioteca digitale in quanto difetta di organizzazione e di una struttura organica e consapevole che un insieme di documenti digitali richiede per poter essere definito come biblioteca. Si è avuto negli anni un

---

<sup>20</sup> Come standard è stato assunto dal 1991 il codice UNICODE, che codificato a 16 bit consente di rappresentare ben 65.536 caratteri, permettendo di contenere in una singola tabella tutti gli elementi necessari per la maggior parte delle lingue del mondo. E' in grado di gestire tutti gli alfabeti (dal greco antico, all'arabo, al cinese, ecc) e consente di rappresentare caratteri speciali utili nel campo dello studio delle scritture antiche (paleografia); permette dunque di distinguere i diversi tracciati della stessa lettera (assegnando a ogni glifo un diverso codice) e rappresentare caratteri dei manoscritti medievali e umanistici (come i segni abbreviativi).

<sup>21</sup> Un link localizza ed identifica una risorsa attraverso un URL (Uniform Resource Locator) univoco, anche se si pone il problema dell'updating (aggiornamento) dei link.

<sup>22</sup> E' un male questo?

dilatarsi incontrollato dei documenti pubblicati che ha decretato un guazzabuglio informatico e un senso di smarrimento e disorientamento negli utenti. Esso infatti mette a disposizione risorse di ogni genere sul nostro “tavolo” di casa o sulla “scrivania” di lavoro, ma la sovrabbondanza di materiale confonde e ci fa disperare di trovare quello che realmente stiamo cercando.

Il Web, con le sue peculiarità e le sue convergenze, influenza l'equilibrio esistenziale della specie umana, così come hanno fatto le tecnologie che lo hanno preceduto (vedasi ad esempio l'impatto della Rivoluzione Industriale sulla struttura sociale della società). E' il vettore privilegiato della comunicazione umana nel nuovo millennio e per questo è andato a ridisegnare, non solo la strutturazione formale della conoscenza, ma anche il modo di approcciarsi cognitivo e culturale alla medesima, forgiando così le capacità inferenziali delle nuove generazioni.

Con la sua costituzione dinamica ha suggestionato l'approccio degli internauti<sup>23</sup>, spostandolo verso caratterizzazioni procedurali focalizzate alla spigliatezza dell'agire, ad una fervida immaginazione virtuale, alla rinascita dell'immagine come pedagogia e alla delocalizzazione della divulgazione privata e pubblica tramite e-mail, chat, newsgroup. A ciò si aggiungono anche corollari inattesi, come il progressivo depauperarsi dell'attenzione rivolta al testo in favore della cornice multimediale e l'approssimazione e la disinformazione che si accompagnano alla diffusione del sapere e dei soggetti che vogliono contribuirvi.

La struttura stessa di un ipertesto<sup>24</sup>, che si compone di sezioni tematiche (microtesti che costituiscono di norma spazi informativi autonomi ed interrelati fra loro da legami e rimandi detti link), fa decadere quei requisiti, quelle fissità, come un inizio ed una conclusione ben definiti, che inscrivano il testo cartaceo di una coerenza interna e deduttiva.

---

<sup>23</sup> Internauti sta per i navigatori in Internet, per analogia a cosmonauti che sono i navigatori nel cosmo.

<sup>24</sup> L'ipertesto è tipicamente il formato di un documento pubblicato nel web. E' l'esempio più recente di rimediatazione delle forme della scrittura tradizionale realizzata con la scrittura digitale. L'ipertesto introduce una nuova dimensione testuale aperta, flessibile, interattiva, multimediale, allargata alla cooperazione con la comunità dei lettori e diffusa nelle pratiche comuni di comunicazione attraverso l'ipertesto globale del World Wide Web.

### 3.2. Il Semantic Web

Nel maggio del 2001, Tim Berners Lee traccia le coordinate teoriche su cui instradare lo sviluppo del Web.

Egli dice<sup>25</sup> : “Il Web diventa un mezzo ..... potente per favorire la *collaborazione tra i popoli... la collaborazione si allarga ai computer* ..... tutti i dati sul web .....verranno gestiti da macchine che parleranno alle macchine lasciando che gli uomini pensino soltanto a fornire *l’ispirazione e l’intuito* ..... il Web sarà un luogo in cui l’improvvisazione dell’essere umano e il ragionamento della macchina coesisteranno in *una miscela ideale e potente.*”

Con queste parole Tim Berners Lee presentava la sua nuova visione del Web, una realtà diversa da quella preesistente ma non una sua alternativa, bensì una evoluzione.

L’architrave della costruzione teorico-tecnologica di Tim Berners Lee è che le informazioni diventino processabili e comprensibili direttamente alle macchine.

Il W3C ha inserito il Web Semantico tra i suoi sette obiettivi strategici primari di medio-lungo termine. Gli altri sono: Accesso universale, Fiducia, Interoperabilità, Capacità evolutiva, Decentralizzazione, Multimedia più coinvolgente.

L’idea del Web Semantico nasce dalla necessità di estendere l’attuale Web in modo da favorire lo scambio di informazioni oltre che tra esseri umani, anche tra programmi per computer (quindi lo scambio che era da “uomo a uomo” diventerà “da uomo a macchina, da macchina a macchina, da macchina a uomo”) tramite una rappresentazione che questi siano in grado di utilizzare e in certo modo comprendere.

L’idea di un Web in cui sia possibile far effettuare alla macchina un’*elaborazione semantica dei contenuti dei documenti*, richiede di sviluppare una tecnologia che consenta una descrizione semantica delle pagine Web accessibile all’elaboratore elettronico tramite software agenti in grado di *lavorare sui significati delle risorse*.

Il progetto di un Web semantico si presenta come un luogo in cui esprimere vere e proprie affermazioni non ambigue che riportano relazioni esistenti tra oggetti, risorse o fatti del mondo reale.

---

<sup>25</sup> Da un articolo pubblicato da Tim Berners Lee in collaborazione con James Hendler e Ora Lassila, sulla rivista “Scientific American” dal titolo “The Semantic Web”.

Allora lo scopo in generale sarà quello di favorire la conoscenza e le possibilità di conoscenza.

Interrogando Wikipedia<sup>26</sup> troviamo la seguente *definizione di Web*: una rete di risorse di informazioni, basata sull'infrastruttura Internet [...] che si basa su tre meccanismi per rendere queste risorse prontamente disponibili al più vasto insieme di utenti:

- una schema di denominazione uniforme per localizzare le risorse (URL: Uniform Resource Locator);
- protocolli per accedere alle risorse (HTTP: HyperText Transmission Protocol);
- ipertesto, per una facile navigazione tra le risorse (HTML: HyperText Markup Language).

Le pagine Web sono collegate *sintatticamente* mediante indici che localizzano la URL della pagina e tali collegamenti consentono di identificare le pagine in modo univoco. Uno dei principali limiti di questa impostazione risiede nell'assenza di significato dei collegamenti: in altre parole questo sistema manca di *capacità semantica*. I collegamenti dovrebbero non solo condurci in un luogo, ma anche descrivere il luogo in cui ci conducono.

La realizzazione del Web Semantico consentirà di potenziare le capacità di ricerca dei motori attuali e di incrementare sensibilmente le potenzialità del web, fornendoci servizi considerati ancora un'utopia.

Il WS<sup>27</sup> non implica forme di intelligenza da parte delle macchine che siano paragonabili a quella di cui è dotata la mente umana, ma solo un'abilità delle macchine a risolvere problemi ben definiti compiendo operazione ben definite su dati ben definiti esistenti. Questo comporterà uno sforzo in più in fase di progettazione e realizzazione di risorse per il Web.

Il Web attuale (o Web 1.0) è caratterizzato dal fatto di essere “leggibile dalle macchine” (machine-readable) ma non “comprensibile alle macchine” (machine-understandable). Il termine “semantico”, cioè “che ha a che fare col significato”, assume allora la valenza di

---

<sup>26</sup> Wikipedia (www.wikipedia.org) è un'enciclopedia sul web, in cui gli utenti oltre a trovare informazioni, possono aggiungere voci e contenuti. Un Wiki è un sito web (o comunque una collezione di documenti ipertestuali) che permette a ciascuno dei suoi utilizzatori di aggiungere contenuti, ma anche di modificare quelli già inseriti da altri. E' un esempio di collaborazione collettiva, con tutti i limiti riguardo la qualità dell'informazione, la serietà e l'affidabilità scientifica dei contenuti.

<sup>27</sup> D'ora in avanti indicherà il Web Semantico.

machine-processable (elaborabile dalla macchina). Il WS sarà pertanto un ambiente caratterizzato dalla presenza di strutture di collegamento più espressive di quelle già esistenti.

Costruire un sito web o una pagina web seguendo logiche semantiche vuol dire consentire ad alcuni software di interpretare il contenuto di un testo imitando il meccanismo di lettura e comprensione proprio degli esseri umani. L'obiettivo del WS è di rendere il web che conosciamo più funzionale, non solo nelle relazioni tra uomini e macchine, ma anche e soprattutto tra macchina e macchina. Nel WS quindi le informazioni saranno codificate in una forma che le renderà direttamente accessibili alle applicazioni software. La possibilità di scambio di informazioni tra programmi aumenta indirettamente ed esponenzialmente la quantità di informazioni che possono utilizzare gli esseri umani.

Perché possa esprimere tutte le sue potenzialità il WS deve essere popolato di informazioni. Mancano inoltre i programmi, detti *agenti intelligenti*<sup>28</sup>, in grado di scambiarsi informazioni autonomamente, al fine di collezionare le informazioni sul web, in modo da poter rispondere alle nostre richieste. Una parte di questi programmi saranno disponibili solo dopo che si saranno definiti tutti gli standard per mediare e scambiare informazioni su vasta scala.

Mancano anche gli strumenti per determinare la *provenienza* e la *certezza delle informazioni*, ovvero le condizioni relative a tempi, modi e luoghi di origine dell'informazione e prove e sicurezza che la conoscenza codificata sia realmente tale.

Gli ambiti applicativi possibili sono il commercio elettronico, l'e-Government e i motori di ricerca che verrebbero potenziati perché la ricerca si baserebbe su concetti anziché su parole chiave.

Sul piano teorico, il WS si può realizzare utilizzando gli strumenti della logica descrittiva; sul piano tecnologico, tutto deve essere reso compatibile con le tecnologie del web nel loro complesso.

Il WS non è un servizio, ma è un modo di registrare informazioni di un testo e renderle disponibili in maniera più fruibile per sviluppare servizi.

---

<sup>28</sup> Ruolo degli agenti intelligenti nel WS è quello fornire più vaste capacità di inferenza.

Limiti e potenzialità.

*Potenzialità:* molti servizi che oggi richiedono l'intervento umano potrebbero essere completamente automatizzati, e altri che richiederebbero comunque l'intervento umano potrebbero essere razionalizzati, consentendo agli utenti di accedere soltanto alle informazioni pertinenti e filtrando le informazioni che non interessano.

*Limiti:* sul piano teorico, i linguaggi attualmente proposti come standard per il web sono poco espressivi (certe informazioni interessanti non possono essere rese disponibili semplicemente perché non si sa come rappresentarle); sul piano tecnologico, il mondo delle tecnologie basate su XML sta nascendo "dal basso" e diventa perciò molto farraginoso.

L'architettura del WS risulta stratificata su tre livelli:

- i dati: definiti in modo strutturato tramite XML<sup>29</sup>, XML-Schema<sup>30</sup> e Name Space<sup>31</sup>;
- i metadati<sup>32</sup>: "informazioni sui dati" gestite tramite RDF<sup>33</sup> e RDF-Schema<sup>34</sup>;
- le ontologie<sup>35</sup>: rappresentazione semantica di dati e metadati tramite specifici linguaggi.

Questi tre livelli sono centrali nel problema della rappresentazione della conoscenza, problema che assume un ruolo centrale nello sviluppo del WS.

Una volta definiti i dati e i metadati (ossia regole di ragionamento su cui possano appoggiarsi i software agente per districarsi tra le

---

<sup>29</sup> XML (eXtensible Markup Language) è un meta-linguaggio di markup: fornisce un insieme di regole sintattiche (dette specifiche) per modellare la struttura di documenti e dati.

<sup>30</sup> XML-Schema fornisce un metodo per comporre vocabolari XML definendo regole relative alla struttura e al contenuto di un documento XML.

<sup>31</sup> Name Space non è altro che un insieme di nomi di elementi e/o attributi identificati in modo univoco da un identificatore.

<sup>32</sup> I metadati possono essere divisi in tre macrocategorie: *descrittivi* (servono per l'identificazione e il recupero di dati digitali; sono costituiti da descrizioni dei documenti fonte o di quelli nati in formato digitale); *amministrativi e gestionali* (evidenziano le modalità di archiviazione e manutenzione degli oggetti digitali nel sistema di gestione dell'archivio digitale; hanno importanza ai fini della conservazione permanente degli oggetti digitali: possono documentarne processi tecnici di archiviazione, fornirne informazioni sulle condizioni e i diritti di accesso, certificarne l'autenticità e l'integrità del contenuto, identificarli in maniera univoca); *strutturali* (collegano le varie componenti di una risorsa, attraverso la mappatura di schemi di metadati diversi (dati di identificazione e localizzazione del documento, come codice identificativo, indirizzo del file sul server, archivio digitale di appartenenza e indirizzo Internet).

<sup>33</sup> RDF (Resource Description Framework) fornisce un insieme di regole per definire informazioni descrittive sui dati, più precisamente sugli elementi descrittivi di un documento web: queste asserzioni sono realizzate tramite triple (soggetto - predicato - oggetto) che legano tra loro gli elementi in una relazione binaria.

<sup>34</sup> RDF-Schema fornisce un metodo per combinare le descrizioni di RDF in un singolo vocabolario.

<sup>35</sup> Sono strumenti per sviluppare vocabolari specifici per un dato dominio di conoscenza.

plurime desinenze significative degli ipertesti) e le relazioni tra questi, il passaggio successivo fondamentale è rappresentato dall'attribuzione della capacità semantica a questa struttura.

Lo strumento individuato nell'ambito del WS per risolvere questo problema è rappresentato dalle ontologie per la cui descrizione il W3C ha promosso lo sviluppo di OWL (Web Ontology Language): è un linguaggio di markup per rappresentare esplicitamente significato e semantica di termini con vocabolari e relazioni tra termini.

Per utilizzare le basi di conoscenza formalizzate secondo standard RDF è necessario un linguaggio per interrogarle. Questo linguaggio è SPARQL (Simple Protocol And RDF Query Language): è un linguaggio che interroga sistemi che definiscono *asserzioni* RDF. Esso è candidato a divenire un raccomandazione del W3C.

Con il WS possiamo aggiungere alle pagine web un senso compiuto, un significato che va oltre le parole scritte, una personalità che può aiutare ogni motore di ricerca ad individuare ciò che stiamo cercando. Tutto questo in virtù di una marcatura dei documenti, di un linguaggio gestibile da tutte le applicazioni e dell'introduzione di vocabolari specifici, ossia insiemi di frasi alle quali possano associarsi relazioni stabilite fra elementi marcati. Il WS per poter funzionare deve poter disporre di informazione strutturata e di regole di deduzione per gestirla, in modo da accostare quelle informazioni che un'interrogazione ha richiesto. Tim Berners Lee ha sottolineato che uno degli elementi fondamentali del WS sarà la compresenza di più ontologie. Se si vuole un sistema dinamico e in grado di raffinarsi su scala universale, bisognerà pagare il prezzo di una certa dose di incoerenza almeno iniziale.

Il WS crea un mondo nel mondo attraverso un sistema di informazione automatizzata, decidendo della qualità e utilità propria delle informazioni e permettendo un'interazione tra vari programmi e vari utenti: una grande famiglia multimediale della conoscenza.

La piena realizzazione dei principi del WS è ancora lontana da una sua realizzazione e gli ostacoli maggiori si incontrano proprio al livello ontologico della sua architettura. Altri grossi ostacoli sono:

- l'onerosità della mappatura delle risorse;
- la piena interoperabilità tra i diversi linguaggi utilizzati per la descrizione dei dati e le relazioni tra essi;

- i cambiamenti profondi, anche culturali, che si richiedono in fase di progettazione e realizzazione dei documenti destinati al web;
- un adeguamento sociale e tecnologico che fin dagli inizi Tim Berners Lee aveva indicato come chiave del cambiamento.

Alla base del progetto del WS vi è una duplice assunzione:

- l'idea che sia possibile aggiungere , in maniera semplice, coerente, pertinente e sufficientemente standardizzata, metadati semantici a gran parte dell'informazione primaria inserita in rete;
- l'idea che questi metadati semantici siano a loro volta suscettibili di essere gestiti, analizzati e aggregati in maniera utile e funzionale attraverso l'impiego intelligente di appositi agenti software.

Entrambe queste assunzioni costituiscono vere e proprie sfide per la comunità degli utenti della rete. Occorre in primo luogo che gli utenti che inseriscono informazione in rete siano abituati ad usare i metadati, a descrivere semanticamente l'informazione primaria da essi prodotta, a considerare la presenza di metadati semantici adeguati come una *conditio sine qua non* per l'inserimento efficace di informazione in rete.

In un contesto di rete, la discussione sui metadati è diventata una discussione su schemi standard condivisi per far sì che un'indicizzazione con uno schema di metadati omogeneo consenta l'interoperabilità anche tra vari tipi di risorse (testi, audio, video, ecc..) e l'integrazione dei vari sistemi informativi, sia all'interno che all'esterno dei diversi sistemi locali.

Vi è allora una terza sfida: la formazione degli utenti.

Il WS ha bisogno, per potersi sviluppare, di conquistare consensi e può farlo solo se è capace di trovare le proprie "killer applications": situazioni in cui un insieme sufficientemente ampio di fornitori di informazione sia spinto a integrare in maniera semplice e consistente metainformazione semantica nelle risorse informative prodotte, e in cui la disponibilità di tale metainformazione offra vantaggi immediati agli utenti delle risorse informative in questione. Secondo Gino

Roncaglia<sup>36</sup>, un settore promettente è quello rappresentato dai weblog e dal mondo del giornalismo e dell'informazione online.

L'esplosione del weblog è strettamente legata all'evoluzione tecnica e stilistica di due strumenti:

- le pagine web personali;
- i programmi CMS (Content Management System) nati per semplificare l'inserimento, la strutturazione e l'impaginazione dei siti Web.

I siti personali hanno una forma abbastanza costante: articoli caratterizzati da una data di pubblicazione e da un titolo, ordinati dal più recente al meno recente. Spesso questi articoli contengono collegamenti ad altre pagine Web.

I weblog si arricchiscono di ulteriori funzionalità: innanzitutto un meccanismo di archiviazione (dopo una certa permanenza sulla home page gli articoli vengono trasferiti in pagine d'archivio); ogni articolo ha però di norma fin dal suo inserimento un proprio indirizzo specifico, che non cambia col tempo ed è denominato *permalink*. I rimandi da un weblog all'altro sono frequentissimi e danno vita ad una vera e propria ragnatela di riferimenti incrociati. Il mondo dei weblog si trasforma in uno spazio condiviso, popolato da utenti che dispongono di strumenti simili (o compatibili) e li utilizzano non solo per scambiarsi informazioni ma anche per approfondirle collaborativamente e per discuterle.

A questo spazio condiviso è stato dato il nome di *blogosfera*. Essa rappresenta un ambiente ideale per la sperimentazione di strumenti legati all'idea del WS. Si tratta infatti di un sottoinsieme caratterizzato da:

- una relativa uniformità strutturale dei contenuti;
- una relativa semplicità strutturale dei contenuti;
- l'uso diffuso di un gruppo abbastanza ristretto di strumenti avanzati per la creazione e gestione dei contenuti e che nella grande maggioranza fanno uso di XML;
- la presenza di una comunità di utenti molto attiva; fortemente interconnessa, interessata allo scambio di contenuti, e assai consapevole dell'importanza degli strumenti di gestione e classificazione semantica dell'informazione;

---

<sup>36</sup> Nato a Roma l'11 gennaio 1960 è Docente al Dipartimento di Scienze Umane, presso l'Università degli Studi della Tuscia.

- una popolazione di utenti e un insieme di strumenti relativamente giovani e in rapidissima evoluzione.

Non stupisce che l'interesse verso il *semantic blogging* cominci a produrre i primi frutti:

- *Live Topics* è in grado di semplificare l'associazione di metadati semantici ai post<sup>37</sup> di un blog;
- *k-collectors* (strumento lato server) in grado di utilizzare metadati per aggregare semanticamente in categorie i post di un weblog;
- *ecc..*

Il WS sarà un successo soprattutto se rimarrà invisibile. Tutta la tecnologia di cui ha bisogno deve restare “sotto la superficie”. La sola differenza che si dovrà notare sarà che la qualità dei risultati restituiti da un motore di ricerca sarà migliore rispetto al passato. Il motore di ricerca sarà in grado di determinare che ci sono magari due tipi di risultati e che dovrebbero essere mostrati separatamente, oppure chiedere all'utente quale dei due stava cercando.

Un altro aspetto importante sarà la personalizzazione: un utente vedrà uno stesso sito in modo differente da un altro utente, ovvero in relazione agli interessi dei due utenti il sito potrebbe essere ripulito e sfrondata di tutto quello che ad un utente non interessa. Al momento molte isole di WS sono in corso di sviluppo. A lungo termine è auspicabile che queste isole si uniscano e formino il vero WS: solo allora l'utente medio riscontrerà dei vantaggi in termini di ricerca e servizi.

Per gli studiosi di filosofia il WS potrebbe diventare un laboratorio in cui applicare e far nascere teorie filosofiche.

Se HTML e il Web hanno fatto sembrare tutti i documenti in rete come un unico immenso libro, il linguaggio RDF farà sembrare tutti i dati al mondo un'unica enorme banca dati.

### **3.3. Le ontologie (due teorie)**

Abbiamo detto precedentemente che il terzo livello dell'architettura del WS è composto dalle ontologie, cioè da strumenti in grado di sviluppare vocabolari specifici per un dato dominio di conoscenza.

---

<sup>37</sup> Non esiste una traduzione italiana di post: sono comunque i messaggi lasciati da altri utenti su un determinato blog.

Un'ontologia è una descrizione formale esplicita dei concetti di un dominio specifico. Una ontologia popolata di istanze e completata con regole di inferenza viene detta base di conoscenza.

Tramite un'ontologia si effettua una *definizione formale di un determinato campo del sapere umano*.

Idealmente essa dovrebbe consistere in un'unica tassonomia omnicomprensiva.

All'interno del WS essa è intesa come una base di conoscenza progettata con l'obiettivo di essere condivisa e riutilizzabile concretamente nel mondo reale: una sorta di vocabolario di termini con precise definizioni e assiomi relativi alla loro interpretazione e al loro uso.

I passi che portano alla creazione di una ontologia sono:

1. definire i *concetti* del dominio (classi);
2. organizzare i concetti in una *gerarchia tassonomica* (sottoclassi e superclassi);
3. specificare gli *attributi* dei concetti, le restrizioni su di essi e le *relazioni* tra concetti (proprietà o slot);
4. stabilire *istanze* dei concetti, popolando l'ontologia.

In *filosofia*, l'ontologia<sup>38</sup>, branca fondamentale della metafisica, è lo studio dell'essere in quanto tale, nonché delle sue categorie fondamentali. E' lo studio di ciò che esiste, di perché e come esiste, se esiste, se è pensabile, e dunque di ogni domanda circa il senso della vita, dal momento che l'esistenza è proprio ciò che contraddistingue ogni cosa senza distinzioni. Ogni domanda intorno al "soggetto", alla "relazione" e all' "oggetto", dunque tra "io", "mondo" e reciproca relazione, è una domanda di ordine ontologico. Ne consegue che la domanda sull'essere, cioè sull'insieme complessivo, e dunque massimamente astratto, di tutto ciò che è, ovvero che esiste, è il nucleo dell'ontologia. L'ontologia si interessa di determinare quali sono le *categorie dell'essere* fondamentali, e si chiede se, ed in che senso, si può dire che gli elementi di queste categorie esistono. Esiste una differenza terminologica tra *essere* ed *esistere* che ricorre nella

---

<sup>38</sup> Il termine ontologia fu coniato soltanto agli inizi del XVII sec. da Jacob Lorhard nella prima ed. della sua opera "Ogdoas Scholastica"(1606) e successivamente utilizzato da Rudolph Göckel per il suo "Lessico filosofico" (1613). Il suo primo uso nella letteratura informatica sembra che risalga al 1967, in un lavoro sui fondamenti del *data modelling* di S.H. Mealy (in un passaggio egli dice che "Il problema è l'ontologia, o la domanda su cosa esiste").

storia della filosofia in vari filosofi: mentre l'essere è in sé e per sé, e non ha bisogno di nient'altro, l'esistenza non ha l'essere in proprio ma lo riceve da qualcos'altro. Così se l'essere è qualcosa di assoluto, l'esistenza<sup>39</sup> è invece subordinata a un essere superiore dal quale dipende.

Platone (427 a.c.–347 a.c.) fu il primo a distinguere esplicitamente l'essere dall'esistere; in particolare egli attribuiva l'esistenza alla condizione umana, sempre in bilico tra essere e non-essere, sottoposta alla contingenza e al divenire, mentre l'essere è la dimensione ontologica più vera nella quale si trova il mondo delle idee, incorruttibile, immutabile ed eterno.

Anche Heidegger (1889-1976) ha ripreso la distinzione tra Essere ed esistere, con particolare riferimento alla condizione umana: l'uomo è un essere calato in una dimensione temporale e transitoria, un esserci che vive suo malgrado a contatto col non-essere.

L'uso del termine ontologia in informatica deriva ovviamente dal precedente uso dello stesso termine in filosofia, così come dalla filosofia sono tratti i concetti fondamentali di categoria e di relazione.

In *informatica* una ontologia è il tentativo di formulare uno schema concettuale esaustivo e rigoroso nell'ambito di un dato dominio; si tratta generalmente di una struttura di dati gerarchica che contiene tutte le entità rilevanti, le relazioni esistenti tra esse, gli assiomi e i vincoli specifici del dominio. E' formata da una parte tassonomica, che definisce i concetti e li mette in relazione gerarchica fra di loro, e da una serie di regole di deduzione che permettono al computer di manipolare i termini in modi significativi per l'utente.

Secondo Tom R. Gruber<sup>40</sup> “an ontology is a specification of a conceptualization”: egli pone l'accento dunque sul fatto l'ontologia è una descrizione di concetti e relazioni, un set di definizioni di un vocabolario formale, in modo tale che un insieme di agenti possa utilizzarle al fine di comunicare tra loro, oppure possa interrogare l'ontologia attraverso domande (query).

Per Nicola Guarino<sup>41</sup> una ontologia è “un artefatto ingegneristico, costruito con uno specifico vocabolario, usato per descrivere una certa

---

<sup>39</sup> Esistenza deriva dal composto latino *ex + sistentia*, che significa *essere da*, cioè *essere a partire da* qualcos'altro.

<sup>40</sup> della Stanford University.

<sup>41</sup> Nicola Guarino, nato a Messina nel 1954, lavora presso il Laboratorio per l'Ontologia Applicata (LOA), presso il Consiglio Nazionale delle Ricerche di Trento.

realtà, più un insieme di assunzioni esplicite sul significato inteso delle parole del vocabolario .... Nel caso più semplice, ... descrive una gerarchia di concetti correlati da relazioni di sussunzione; nei casi più sofisticati, sono aggiunti assiomi adatti a esprimere altre relazioni tra concetti e a restringere la loro interpretazione intesa”.

Una vera ontologia non deve limitarsi ad una gerarchia di concetti organizzati con la relazione di sussunzione (in inglese *is a*, cioè *è un*), andando a definire sottoclassi e sottotipi, ma deve includere anche altre relazioni semantiche che descrivono in che modo i concetti sono interrelati. Una delle relazioni più comuni è la relazione “parte di”.

Lo scopo di una ontologia computazionale è quello di creare una base di dati, che è un artefatto dell’uomo, contenente i concetti riferiti al dominio di indagine dell’ontologo, e che verrà impiegata per eseguire certi tipi di computazione. Diverse ontologie computazionali, sviluppate indipendentemente per scopi diversi, per lo stesso dominio di applicazione, possono risultare sensibilmente diverse tra di loro.

L’ontologia di cui si serve l’informatica è un’*ontologia descrittiva*: non cerca una spiegazione, ma dà una descrizione, non universale ma particolare, calandosi in un certo dominio, adottando sistemi di classificazione e legandosi al mondo esterno. Offre gli schemi concettuali sulla realtà, creando un senso comune utile alla comprensione dei dati e dei loro rapporti anche tra diversi domini, proponendo così una nuova rappresentazione della conoscenza. Si crea così un mondo logico nuovo e in continua evoluzione fatto di regole semantiche e sintattiche, una nuova semiologia.

Esistono due tipi di ontologie: le ontologie pesanti (Top-Level, di massimo livello, *costitutive* o superiori o *fondazionali*) e quelle leggere (*di dominio*).

Le prime riguardano aspetti più generici e astratti ravvisabili nella realtà intesa nella sua completezza; le seconde, che è possibile ricondurre a quelle di massimo livello, si occupano dell’analisi categoriale e relazionale che si specifica in una determinata porzione della realtà (in un dominio specifico).

La differenza è legata soprattutto all’uso di un linguaggio formale molto espressivo e alla presenza di numerose relazioni intracategoriali, che insieme consentono una migliore caratterizzazione assiomatica delle scelte ontologiche rispetto a quella

fornita dalle ontologie leggere. Queste infatti si presentano per lo più in forma di semplici tassonomie. Il significato dei termini in questo caso si suppone noto e condiviso all'interno di una comunità d'uso. Le ontologie fondazionali invece possono essere applicate nella definizione di una struttura terminologica trasversale alle comunità d'uso, candidandosi a svolgere il ruolo di *traduttore del significato inteso* nelle diverse comunità.

Una ontologia fondazionale ha la funzione di un'ontologia di base sia per gli utenti che per i programmi, influenzando la loro prospettiva dei dati e degli eventi.

Sia dal punto di vista teorico che applicativo, un *approccio monolitico* rivolto allo sviluppo di una teoria omnicomprensiva sembra difficilmente sostenibile: esistono infatti posizioni ontologiche incompatibili già a livello delle relazioni e delle categorie più basilari.

E' quindi plausibile prevedere un *approccio modulare* in cui i singoli moduli ontologici, ispirati da particolari punti di vista o posizioni, sono per quanto possibile integrati tramite relazioni formali<sup>42</sup>.

Le ontologie costitutive sono importanti per sviluppare, sulla base dei concetti fondanti e delle assiomatizzazioni che contengono, ontologie specializzate che mantengano un disegno integro e coerente.

Ma la creazione di una ontologia costitutiva fondante e totale risulta un'impresa titanica, che richiederà la conciliazione di moltissime esigenze e punti di vista diversi.

Questo per molte ragioni: la difficoltà di *negoziare* un modello condiviso che soddisfi le necessità di tutte le parti in gioco; l'impossibilità pratica di mantenere un tale modello in un ambiente altamente dinamico; il problema di trovare soddisfacenti relazioni tra preesistenti modelli locali dentro il modello globale (il fatto che debba essere l'agente a dover cambiare i propri modelli locali in funzione di

---

<sup>42</sup> In questo contesto teorico, l'ontologia fondazionale DOLCE (Descriptive Ontology for Linguistic and Cognitive Engineering), sviluppata nell'ambito del progetto WonderWeb, costituisce un primo modulo di riferimento di una libreria di ontologie fondazionali. DOLCE è un insieme di termini e di relazioni (caratterizzati formalmente in un linguaggio logico) con un chiaro orientamento cognitivo: si propone di catturare le categorie legate al linguaggio naturale e al senso comune. Cerca di costruire "scatole cognitive" in grado di cogliere dimensioni della realtà vincolate alla percezione umana, allo sfondo culturale e alle convenzioni sociali. Assume l'*approccio moltiplicativo*: entità distinte possono essere co-localizzate nella stessa regione spazio-temporale. Quattro sono le categorie ontologiche più generali: *endurant* (continuante), *perdurant* (occorrente), *quality* (qualità) e *abstract* (astratto). La relazione principale tra tali categorie è quella della *partecipazione*: un *continuante* vive cioè nel tempo in quanto *partecipante* ad un certo *occorrente*.

quello globale, e non viceversa, limita l'eterogeneità semantica che rappresenta una ricchezza nel web).

E' invece già possibile e praticata la creazione di molte ontologie, ognuna limitata ad un dominio ben preciso e persino ad un preciso punto di vista o scopo di quel dominio.

Le ontologie così create potrebbero poi, in caso di necessità, venire mappate le une sulle altre, sfruttando il meccanismo di importazione delle ontologie, in modo da farle interagire senza perdere la complessità e la particolarità di ciascuna.

I programmi nei computer possono utilizzare un'ontologia per una varietà di scopi, fra cui il ragionamento deduttivo, la classificazione, diverse tecniche di problem-solving (risoluzione di problemi), oltre che per facilitare la comunicazione e lo scambio di informazione tra sistemi diversi. Inoltre le ontologie possono essere utilizzate per *negoziare il significato*, ovvero per permettere una cooperazione effettiva tra molteplici agenti artificiali, o per *stabilire un consenso* in una società mista, dove gli agenti artificiali collaborano con gli esseri umani.

L'organizzazione delle conoscenze è una fase fondamentale per il loro utilizzo e il loro progresso.

Elaborazione (*kho hsüeh*), che è il termine cinese tradizionale per indicare la scienza, significa "classificazione della conoscenza".

Nella filosofia cinese si parla anche di *raddrizzamento dei nomi*. Confucio (551-479 a.c.) diceva: "Che il governante sia governante, il suddito suddito, il padre padre e il figlio figlio"<sup>43</sup>, intendendo che ogni persona doveva essere trattata e rispettata come si conveniva al suo nome.

Ma Hsün-tse (289-238 a.c.), un grande maestro appartenente alla Scuola dei Nomi, al raddrizzamento dei nomi oltre che un senso etico ne diede anche uno terminologico: "Quando i cinque sensi, avendo notato qualche cosa, non riescono a classificarla, e quando l'intelletto non riesce a identificarla e a darle un significato, allora non c'è conoscenza"<sup>44</sup>.

Dunque l'organizzazione semantica è un presupposto necessario per il progredire della conoscenza. Questa funzione è stata svolta

---

<sup>43</sup> Analecta 12.11

<sup>44</sup> Hsün-tse.22, citato in Storia della filosofia cinese, Fung Yu-lan.

specialmente dalla classificazione dei concetti (si pensi a Linneo<sup>45</sup>, Bacone<sup>46</sup>, Dewey<sup>47</sup>, ...).

Ciò che è importante non è il nome scelto, ma che una volta scelto si usi sempre lo stesso: “colui che indica dati e situazioni simili dovrebbe usare lo stesso nome. Non esistono nomi necessariamente appropriati. I nomi vengono fissati per convenzione. Ma quando la convenzione è stabilita ed è divenuta costume, allora il nome è appropriato.”<sup>48</sup>.

La scienza e la tecnologia infatti sviluppano una *terminologia* precisa, che è necessaria per condividere e sviluppare le conoscenze.

Se dunque vogliamo organizzare le conoscenze per renderle meglio fruibili, dobbiamo evitare ambiguità terminologiche e dedicarci al raddrizzamento dei nomi.

Esiste dunque un problema di *ambiguità semantica* o problema della *disambiguazione del significato*, come già accennato in precedenza: esso può essere risolto solo con un processo di *negoiazione del significato*.

In biblioteconomia e documentazione sono state messe a punto numerose tecniche per l'organizzazione delle conoscenze (KOS: knowledge organization system), che differiscono per il grado di sofisticatezza:

- indicizzazione derivata:
  - testi completi digitali ricercabili (full-text);
  - liste invertite di parole (inverted files);
  - parole-chiave (key-words).
- indicizzazione assegnata
  - intestazioni per soggetto (subject headings);
  - thesauri, tassonomie, ontologie;
  - classificazioni;
  - sintesi, riassunti (abstract, summaries).

L'indicizzazione può essere inoltre:

- *pre-coordinata*: le relazioni fra termini sono indicate in fase di indicizzazione;

---

<sup>45</sup> (1707-1778) biologo svedese, padre della moderna classificazione scientifica degli organismi viventi.

<sup>46</sup> (1561-1626) filosofo, politico e saggista inglese.

<sup>47</sup> (1851-1931) bibliotecario statunitense, ideatore del moderno sistema di catalogazione bibliotecario.

<sup>48</sup> cfr. Hsün-tse.22

- *post-coordinata*: le relazioni fra termini sono indicate in fase di ricerca.

Più i sistemi sono avanzati e sofisticati, più sono ricchi di informazioni utili per la ricerca ma anche costosi in termini di tempo e competenze, e quindi di denaro.

La classificazione allora è un investimento nella qualità dell'informazione.

Attribuire efficacemente un indice o descrittore semantico a un documento richiede, oltre a una cultura generale sufficientemente intensa ed estesa, anche professionalità specifiche che non sono apprendibili con facilità né in poco tempo.

Indicizzare significa creare indici, cioè un'organizzazione sistematica di oggetti simbolici (parole, frasi, codici alfa-numeric) finalizzati a consentire a un utente di trovare l'informazione relativa a un documento ospitato in un determinato archivio.

Si distinguono due tipi di indicizzatori:

- l'*autore* del documento che indicizza la propria opera;
- l'*indicizzatore professionista* che indicizza l'opera di un autore;

che presentano fra loro *relazioni inverse* per quanto riguarda il rapporto: conoscenza dell'argomento / conoscenza della tecnica indicale.

Se l'autore è avvantaggiato dal fatto di conoscere in modo approfondito e dettagliato l'argomento trattato dal documento, ma anche le esigenze e le caratteristiche dell'utente finale cui esso è destinato, raramente egli possiede anche la professionalità necessaria per una buona indicizzazione del proprio lavoro.

Il professionista, al contrario, pur padroneggiando il proprio mestiere, è raramente in grado di conoscere altrettanto bene il destinatario finale così come il campo disciplinare e l'argomento specifico del documento che deve indicizzare.

Questa dicotomia nel campo dell'indicizzazione in biblioteconomia, si ripresenta quando parliamo del modello di sviluppo delle ontologie, a proposito del quale esistono due teorie differenti: quella propugnata da Tim Berners Lee e quella invece sostenuta dall'italiano Nicola Guarino.

Andiamo ora ad analizzarle.

Tim Berners Lee ritiene che lo sviluppo del WS e dunque anche lo sviluppo delle ontologie che andranno a renderlo funzionante, debba avvenire in modo aperto, democratico, su base sociale, spontaneamente e non rigidamente, secondo un modello di tipo Bottom-up, come già avvenuto per il Web. Uno sviluppo aperto del WS passa attraverso una definizione non rigida di un vocabolario di ontologie, senza dover definire un magazzino centrale di informazioni: il WS impara un concetto tramite contributi ripetuti da diverse fonti indipendenti. La convinzione di Berners Lee è basata sull'oggettiva difficoltà di produrre standard globali condivisi. I punti fondamentali della questione sono l'*interoperabilità* del Web e la sua *evolvibilità*. Attraverso una libera evoluzione della rete, che passi anche attraverso qualche errore o incomprensione, si otterrà la tanto agognata intercomunicabilità. L'evolvibilità è altrettanto cruciale: l'idea è che la rete cresca passo per passo, quasi naturalmente, senza dover essere riprogettata da zero. Le macchine dovranno poter raggiungere un massimo livello di *trust*, ossia di fiducia. Il progetto del WS deve ripercorrere le tappe del vecchio Web: non ne deve smarrire la base sociale che lo ha contraddistinto nelle fasi della sua esplosione nel 1989. Il W3C si pone come obiettivo di emanare delle Guidelines (Linee Guida), scrivere una parte del nuovo codice, coordinarne la scrittura, ma nulla più. Le ontologie che lo renderanno funzionante, queste specificazioni di regole logiche, saranno sviluppate in modo aperto e partecipativo. Secondo la previsione fiduciosa di Berners Lee saranno in molti ad occuparsi dello sviluppo di queste questioni, e sarà proprio l'aver permesso a chiunque di lavorarci la base dell'evoluzione e dell'interoperabilità. Le differenze non vanno tolte con rigidi linguaggi di ontologie, ma vanno invece valorizzate. Un software intelligente deve fare proprio il motto di Bateson "informazione è qualunque differenza che generi delle differenze". In sostanza, Tim Berners Lee sostiene che lo sviluppo del WS deve avvenire attraverso un processo empirico, di tipo Bottom-up, cioè lo sviluppo avviene liberamente, dal basso, attraverso l'intervento degli utenti.

Nicola Guarino al contrario sostiene un modello di sviluppo delle ontologie ben fondato, con regole ferree, di tipo Top-down. Esse devono essere sviluppate secondo standard rigidi e in modo rigoroso,

in un gioco di regole ben definito. Egli ritiene utopico il progetto di Berners Lee e non esita a definire “anarchico” un web dove tutti possono fare tutto. E’ un metodo valido, ma non utile ai fini dell’interoperabilità. Per sviluppare un WS all’altezza occorre percorrere la strada delle ontologie ben fondate, basate sulla logica e sulla linguistica. Bisogna disambiguare in modo preciso le ontologie: esse spesso non riescono ad essere interoperabili a causa dell’ambiguità che deriva dall’uso degli stessi termini per concetti diversi (ad esempio il termine “cane” ha svariati significati: l’animale, la parte di una pistola, ed è pure usato in senso metaforico dispregiativo “quell’uomo è un cane”). Per Guarino occorre una vera e propria semantica, dettagliata, rigida, ben fondata, che abbia regole ferree. Anche Barry Smith<sup>49</sup> sostiene che le tassonomie, se vogliono davvero essere utili alla creazione di una ontologia, devono essere ben formate ed evitare, nella rappresentazione ad albero, di creare doppioni. Guarino è perciò un sostenitore di uno sviluppo Top-down del WS: occorre definire ontologie ben fondate, da una generalissima a ontologie via via sempre più specifiche, per evitare ambiguità.

Esistono alcune ontologie già formate:

- Cyc è un’ontologia popolare, sviluppata a partire dal 1985 tramite un sistema proprietario<sup>50</sup>, che consiste in un’ontologia costitutiva e diverse ontologie specializzate per dominio (chiamate microteorie); un sottoinsieme di questa ontologia è stato rilasciato per uso libero col nome di OpenCyc;
- WordNet<sup>51</sup> è un database liberamente disponibile, progettato come una rete semantica basata sui principi della psicolinguistica: è stato espanso con l’aggiunta di definizioni ed è attualmente visto come un dizionario. Si qualifica come un’ontologia costitutiva perché include sia concetti di tipo generale, sia concetti con un maggior grado di specializzazione, collegati da relazioni di sussunzione e con

---

<sup>49</sup> Nell’articolo “Ontologie e sistemi informativi”.

<sup>50</sup> Si indica quel software che ha restrizioni sul suo utilizzo, sulla sua modifica, riproduzione o redistribuzione, solitamente imposti da un proprietario. Queste restrizioni vengono ottenute tramite mezzi tecnici o legali. Un sistema proprietario è un software fatto da un costruttore che ne detiene il controllo e i diritti, e quindi anche la possibilità di aggiornamento, in opposizione ai sistemi Open Source che invece sono condivisi da tutti gli utenti nel senso che ognuno può non solo utilizzarli in modo gratuito, ma anche contribuire al loro sviluppo andando a modificarli a seconda delle nuove esigenze della rete.

<sup>51</sup> Sviluppato dall’Università di Princeton definisce i concetti come grappoli (clusters) di termini chiamati co-insiemei (synsets). Consiste di qualcosa come 100.000 co-insiemei.

relazioni semantiche come quella parte-insieme e causa. Viene ampiamente usato nella ricerca sull'elaborazione del linguaggio.

- da segnalare il sito [schemaweb.info](http://schemaweb.info) che si propone come un catalogo per ontologie scritte con RDF, OWL e DAML+OWL.
- esistono poi ontologie estese per domini specifici: il settore biomedico ha sviluppato un'ontologia ampia e di alta qualità che descrive i termini medici e i nomi dei farmaci; anche l'industria automobilistica è piuttosto avanzata: la Daimler-Chrysler è persino membro del gruppo di lavoro del W3C.

Un'alternativa al linguaggio OWL (col quale è possibile scrivere ontologie che descrivono la conoscenza che abbiamo di un certo dominio, tramite classi, relazioni tra classi e individui appartenenti a classi), proposto dal W3C, è *Topics Maps*, standard ISO su sintassi XML. Esso è sempre finalizzato all'organizzazione e alla rappresentazione della conoscenza.

Secondo Federico Meschini, “Le *Topics Maps* forniscono funzionalità combinate di indici, glossari e thesauri, creando così potenti meccanismi di navigazione tra vaste collezioni di risorse digitali interconnesse tra loro”.

La *Topics Maps* è composta da *topics*, i quali sono divisibili per categorie a seconda della tipologia che varia dai requisiti funzionali alla finalità dell'applicazione a cui la *topics maps* viene destinata. Queste categorie sono chiamate *topic type* e costituiscono a loro volta dei *topics*.

Ogni *topic* ha tre proprietà:

- Nomi: i *topic* hanno una denominazione che rende l'identificazione immediata in fase di elaborazione;
- Occorrenze (*topic association*): sono i documenti e le risorse a cui un *topic* è collegato, con le quali ha una qualche forma di relazione, e vengono distinte a seconda del ruolo che svolgono (struttura a più livelli);
- Ruolo: riguarda il tipo di relazione che si viene a formare tra *topic* e documento.

I collegamenti associativi tra topiche si distinguono da quelli convenzionali perché in questi le ancore ipertestuali risiedono all'interno del documento (nel tag).

La mappa è quindi una struttura informatica che può esistere in modo totalmente indipendente dall'esistenza di risorse informatiche e documentarie ad essa allegate.

Il santo Graal è: metodi per generare automaticamente le ontologie.

Solo il livello dell'ontologia può garantire la creazione di una vera base di conoscenza: associazione fra concetti e oggetti delle classi, relazioni fra concetti, conoscenza derivata (espressa con regole logiche).

Ma rimane ancora non ben definito e stabilito come questa conoscenza sarà utilizzata: linguaggi logici, formule per la dimostrazione e reti della fiducia rappresentano la visione del nuovo Web. L'obiettivo è quello di realizzare sistemi in grado di formulare ogni principio logico e permettere ai computer di ragionare (per *inferenza*) usando questi principi.

Manca però un controllo sull'autenticità di ciò che si afferma: per questo sono state introdotte le *firme digitali*. Basate su lavori in matematica e in crittografia, esse attestano che una determinata persona ha scritto (o ritiene veritiero) un determinato documento o un'istruzione. Quindi firmando digitalmente le istruzioni, chi le incontrerà potrà essere sicuro della loro autenticità. Ogni utente fisserà il suo personale livello di fiducia e sarà il computer a decidere a cosa (e quanto) credere. E' però difficile avere fiducia in un gran numero di persone (per lo più sconosciute) e questo potrebbe limitare l'utilizzo del Web.

Si vuole allora costruire il *Web of Trust*: un utente comunica che ha fiducia in una persona, a sua volta questa persona ha fiducia in altre e queste ultime in altre. Tutte queste relazioni di fiducia si aprono a ventaglio e formano il Web of Trust.

## 4. La ricerca umanistica sul web

### 4.1. L'Informatica Umanistica

Padre Roberto Busa (Vicenza 1913), gesuita italiano, nel 1949 concepì e nei decenni successivi realizzò l'*Index Thomisticus*, un lessico elettronico di tutta l'opera di Tommaso d'Aquino.

La prima rivista relativa all'uso dei calcolatori per le ricerche umanistiche fu "Computer and the Humanities" pubblicata a partire dal 1966.

Le due principali associazioni del settore nacquero nel 1972 (Association for Literary and Linguistic Computing – ALLC) e nel 1978 (Association for Computer and the Humanities – ACH).

Nel 1987 videro la luce due attività congiunte a queste associazioni: la lista di discussione Humanist<sup>52</sup> e la TEI<sup>53</sup> (Text Encoding Initiative), il più importante progetto collettivo per la standardizzazione dei linguaggi di codifica nel campo umanistico.

In Italia, le ricerche sulle applicazioni del computer in campo umanistico presero l'avvio da ricerche di linguistica fino al consolidamento della linguistica computazionale ad opera di Antonio Zampolli<sup>54</sup>.

Il termine "*Informatica Umanistica*" nasce solo all'inizio del 1990, con il titolo di un famoso testo scritto da Tito Orlandi<sup>55</sup>.

In inglese la disciplina è più nota come *Humanities Computing*, o *Computers in the Humanities*.

L'obiettivo di questa disciplina era la definizione dei temi comuni ad ambiti come la linguistica computazionale e l'informatica applicata a tutte le scienze.

---

<sup>52</sup> <http://linux.lettere.unige.it/mailman/listinfo/idulist>

<sup>53</sup> La sua finalità è quella di definire uno standard di codifica specificamente orientato alla gestione dei dati umanistico-letterari, e realizzare una normalizzazione dei formati di memorizzazione dell'informazione testuale, al fine di consentire l'interscambio di documenti. All'inizio assume come linguaggio l'SGML, poi nel 2002 viene rilasciata la TEI P4 con lo scopo di implementare nelle linee guida il supporto a XML. Il modello descrittivo dei testi si fonda su di una normalizzazione delle convenzioni vigenti nell'ambito dell'organizzazione strutturale dei documenti (divisioni in parti, capitoli, paragrafi, ecc.). Sono anche previste strutture per la codifica di fenomeni testuali complessi (trascrizione di fonti manoscritte, ecdotica ovvero teoria e pratica dell'edizione dei testi letterari, l'analisi linguistica e strutturale del testo, la creazione di *corpora*, e la realizzazione di complesse strutture ipertestuali).

<sup>54</sup> Il Professor Antonio Zampolli, fondatore e direttore dell'Istituto di Linguistica Computazionale del CNR, è scomparso in un tragico incidente il 22 Agosto 2003.

<sup>55</sup> (Como, 1940) Ricercatore del Consiglio Nazionale delle Ricerche per «Papirologia copta» dal 1966 al 1976. Professore incaricato di *Lingua e Letteratura Copta* presso la Facoltà di Lettere, Università degli Studi di Roma, dal 1969 e titolare della medesima cattedra dal 1976. Dal 2001 inquadrato nel settore scientifico-disciplinare "Egittologia e antichità copte".

Da quando l'IU<sup>56</sup> è entrata nei piani di studio triennali dei corsi delle facoltà umanistiche è diventata necessaria una riflessione sulla sua didattica ai fini della progettazione di un curriculum adeguato.

Emergono così vari orientamenti:

- la semplice alfabetizzazione informatica degli studenti di facoltà umanistiche (*applicazione di strumenti*);
- lo studio dei metodi computazionali da adottare per le ricerche nelle singole discipline (*applicazione di metodi*);
- la ricerca sui metodi e le tecniche per la gestione digitale dell'informazione senza entrare nelle specifiche applicazioni di ogni settore (*sviluppo dei metodi*).

L'alfabetizzazione informatica dovrebbe essere oggetto della cultura generale di tutti gli studenti a livello di scuola secondaria superiore (si sono moltiplicati dal 1999 ad oggi gli Istituti Scolastici Superiori che accanto alla normale didattica offrono il servizio delle Certificazioni Informatiche di base ai propri studenti<sup>57</sup>). L'alfabetizzazione all'uso dei pacchetti applicativi più comuni è però destinata a invecchiare in tempi brevissimi, anche se la dimestichezza acquisita nell'uso dello strumento informatico può comunque facilitare la comprensione e la gestione di nuovi applicativi.

I fondamenti dell'informatica invece offrono agli studenti gli strumenti essenziali per approfondire la materia (hardware/software, l'uso dei sistemi operativi, il concetto di algoritmo, ecc...).

L'IU non dovrebbe trascurare nemmeno gli effetti delle ICT (Information and Communication Technologies) sull'accesso, la trasmissione e la modifica delle informazioni, fornendo anche un definizione di sapere collettivo (pubblicazione elettronica dei risultati di ricerca, newsgroup, chat, ricerca in rete, creazione di comunità di pratica tra gli studiosi sparsi nel mondo).

L'IU non possiede ancora uno statuto epistemologico stabile e condiviso, né una chiara collocazione accademica.

---

<sup>56</sup> D'ora in avanti indicherà l'Informatica Umanistica.

<sup>57</sup> Le certificazioni più diffuse e rilasciate sulla base di standard formalizzati e controllati a livello europeo sono l'ECDL (European Computer Driving Licence) e le Certificazioni Office (MSO) di Microsoft. Anche l'Istituto dove lavoro, trattandosi di un Istituto Tecnico Industriale, ha scelto dal dicembre 1999 di essere sede di esami per il rilascio dell'ECDL (di cui sono Esaminatore accreditato AICA, l'Associazione Italiana Calcolo Automatico e Informatica che ne garantisce qualità e standard in Italia) ai suoi studenti e agli iscritti provenienti dal mondo del lavoro, offrendo così una opportunità di alfabetizzazione informatica sul territorio di Cremona, in anni in cui se ne sentiva maggiormente il bisogno.

Il compito dell'IU dovrebbe però essere quello di fornire uno strumento utile alla costruzione di alcune competenze immancabili nella cassetta degli attrezzi base dell'umanista che voglia o debba accostarsi all'informatica per le proprie ricerche.

La prima questione che si pone è quella che vede contrapposti i sostenitori di una IU unica *trasversale* a tutte le scienze umanistiche, e quelli che sostengono l'esistenza di IU *specifiche*, che darebbero vita a discipline diverse a seconda della natura dell'oggetto a cui si applicano i metodi informatici.

Secondo Roncaglia e Cadioli (2002), esiste un ambito generale comune della IU che riguarda l'applicazione di alcune tecniche dell'informatica e della telematica a tutto lo spettro delle discipline umanistiche. Secondo loro, la riflessione sull'impatto che certi metodi hanno sul trattamento, l'innovazione, il trasferimento e la memorizzazione della conoscenza deve riguardare le discipline umanistiche in maniera trasversale.

Secondo Orlandi (2001-2002), la IU tende ad incidere sull'assetto teorico degli oggetti di ricerca nei quali si imbatte, modificando confini, concetti, approcci delle singole aree di studio con le quali entra in contatto.

L'idea di una IU unica è più adeguata anche ai fini della formazione dei formatori in questa disciplina, che finora sono stati per lo più umanisti autodidatti<sup>58</sup>, o informatici prestatati alle ricerche umanistiche.

Non esiste una posizione condivisa per quanto riguarda l'insieme dei concetti teorici su cui fondare le ricerche, né si è ancora costituita una base di pratiche generalmente accettate per la presentazione e il controllo dei risultati. La definizione epistemologica della disciplina non può avvenire una volta per tutte, né può fotografare e immobilizzare l'oggetto della ricerca, le tecniche adottate e i suoi confini.

Secondo Willard McCarty<sup>59</sup>, vi è una significativa somiglianza tra IU e le scienze sperimentali: entrambe sono concentrate sui dati, sulle apparecchiature, entrambe implicano la modellizzazione e tendono ad essere collaborative. Ciò permette di dedicarsi alle pratiche di ricerca

---

<sup>58</sup> Come penso sia stato il mio relatore per questa tesi, il Prof. Dino Buzzetti, Docente degli insegnamenti di Storia della Filosofia Medievale e di Informatica per le Scienze Umane del Corso di Laurea in Filosofia della Facoltà di Lettere e Filosofia dell'Università degli Studi di Bologna.

<sup>59</sup> Professore dell' Humanities Computing Centre presso Computing in the Humanities King's College di Londra.

anche in assenza di un paradigma teorico unificante, potendo ottenere risultati sperimentali rilevanti, a partire dai quali costruire un piano teorico di riferimento.

Ma la mancanza di riconoscimento sociale dello status epistemologico e accademico della disciplina ha effetti negativi nella pratica (mancanza di fondi per la ricerca<sup>60</sup> o difficoltà nell'ottenere un ruolo ufficiale per gli studiosi esperti in materia). E' necessario perciò trovare un equilibrio tra ricerca di una stabilizzazione accademica e l'accettazione del carattere fluido dei confini della disciplina.

Il paradosso è evidente: mentre cresce la richiesta di formazione nel campo dell'applicazione dell'informatica alle discipline umanistiche, manca un pieno riconoscimento istituzionale della disciplina.

L'IU può diventare un *progetto culturale*. Esso non consiste soltanto nell'insegnare agli studenti delle facoltà umanistiche un certo *corpus* di conoscenze tecniche per metterli in condizione di usarle nelle ricerche specifiche del loro settore: si tratta anche di un tentativo di cambiare la mentalità degli studiosi di materie umanistiche e degli informatici. Occorre creare una mentalità di dialogo tra competenze diverse, frutto del rispetto e della conoscenza dei reciproci risultati.

Lo sviluppo armonico della tecnologia non può avvenire senza la partecipazione vigile e attiva degli umanisti aperti al futuro.

Un problema importante che l'umanista si deve porre, per poter utilizzare al meglio le risorse del Web, è che l'inglese è divenuto lingua franca delle relazioni internazionali e deve perciò essere imparato dall'umanista, ma allo stesso tempo che l'italiano non deve essere relegato in una riserva indiana destinata a scomparire.

Un altro aspetto importante è il controllo e l'attenzione che un navigatore umanista deve porre alla natura e al controllo dell'informazione reperita. Chiunque, ma in particolare l'umanista deve essere in grado di distinguere la tipologia dell'informazione, e di capire da dove viene.

Ci sono alcuni elementi da prendere in considerazione per valutare l'attendibilità di un sito:

---

<sup>60</sup> Purtroppo alla luce della crisi corrente (la fine del 2008 sarà ricordata come il periodo in cui ebbe inizio la seconda recessione economica mondiale dopo quella terribile del 1929, almeno a quanto stanno riportando i mezzi di comunicazione di tutto il mondo in questi giorni in cui scrivo: siamo alla metà di ottobre 2008), questo problema è condiviso da tutte le Università, gli Istituti di Istruzione e gli Enti che fanno ricerca.

- documentarsi su chi sia o siano gli autori del sito (deve essere sempre chiaro all'interno del sito il responsabile delle notizie);
- notare il livello di interattività del sito (come fare per contattare il webmaster, l'indirizzo fisico di riferimento);
- cercare l'istituzione cui afferisce.

In rete, come altrove, le credenziali hanno un peso, anche se non sono tutto.

L'applicazione dell'informatica alla ricerca umanistica non è neutrale e comporta una sostanziale trasformazione dei metodi, dovuta in primo luogo al vincolo della formalizzazione delle procedure.

L'interesse dell'umanista è portato a rivolgersi verso la capacità del calcolatore di simulare il funzionamento semiotico dei propri oggetti di studio e di elaborare l'informazione che essi veicolano. Da qui deriva l'interesse e l'attenzione che l'IU rivolge ai problemi del WS e alle tecnologie sviluppate per implementare la capacità delle macchine di riconoscere ed elaborare il contenuto informativo delle pagine Web.

Ogni disciplina di area umanistica tradizionale ha sviluppato differenti strategie computazionali, a seconda delle esigenze del settore di competenza, ma quasi tutte le discipline condividono metodologie formali nella gestione automatica dei dati e concordano su un uso non esclusivamente tecnico dello strumento informatico. Questa serie di comuni metodologie percorre trasversalmente le discipline umanistiche e costituisce una base condivisa per le operazioni legate alla rappresentazione e alla conservazione delle fonti, alle modalità della manipolazione, ai criteri del trattamento e alle forme della disseminazione e distribuzione.

L'informatica non deve essere strumento ad uso dell'umanista, ma deve essere pensata come fondamento per una riflessione sui metodi della ricerca umanistica.

Un corso accademico di IU deve mirare a discutere dei metodi almeno nella stessa misura in cui fornisce abilità esclusivamente tecniche.

## **4.2. Il ruolo dei filosofi (o degli umanisti in generale)**

Il WS è un ambizioso e immane sforzo volto a definire una piattaforma concettuale e tecnologica per supportare su scala globale

processi comunicativi significativi uomo-uomo, uomo-macchina e macchina-macchina. Nonostante questa visione abbia una relazione evidente con diverse comunità di ricerca, il dibattito oggi si svolge alla luce scintillante (affascinante) ma abbagliante (inaccessibile) della tecnica e dell'ingegneria. Tra le comunità escluse da tale dibattito una è quella dei filosofi<sup>61</sup>, i quali potrebbero dare un contributo fondamentale.

Alcune domande fondamentali permeano il dibattito sul WS: quale livello di eterogeneità linguistico semantica è accettabile? qual è il criterio logico per giudicare la correttezza della descrizione di un dominio di conoscenza?

Il dibattito intorno al WS è oggi la manifestazione più palese di una serie di domande profonde e irrisolte la cui risposta è tutt'altro che scontata.

Oltre a fornire prospettive alternative rispetto alle questioni principali, i filosofi possono utilmente beneficiare di una contaminazione con i computers scientists, nella misura in cui possono pensare al WS come al più grande laboratorio sperimentale del linguaggio e del significato.

L'esplosione dei processi comunicativi ha già messo in evidenza un ovvio trade-off di sistema: all'aumentare della quantità di informazione presente in rete corrisponde una diminuzione della sua significatività.

Il primo problema evidente sul quale il WS si scontra è quello antico della Torre di Babele dei linguaggi e quello più moderno dell'eterogeneità linguistico semantica palesato dal fenomeno del Web. E' chiaro a tutti come il Web abbia evidenziato in modo forte la quantità dei linguaggi utilizzati per articolare contenuti, sistemi categoriali, tassonomie e ontologie. Tale eterogeneità è rilevante anche per il fatto che non tende a ridursi ma piuttosto tende ad esplodere. Dietro questa esplosione quantitativa dei linguaggi si cela un'esplosione qualitativa dei sistemi d'uso e d'interesse che su quei significati e linguaggi si fondano.

---

<sup>61</sup> Col termine filosofi, per brevità indicherò d'ora in avanti sia i filosofi sia tutti coloro che hanno a che fare con discipline di ambito umanistico.

Nasce allora, come visto in precedenza, la prima grande domanda del WS: a che livello l'eterogeneità semantica è accettabile e a che livello va eliminata tramite standardizzazione?

Vi sono oggi due anime del WS che propongono due risposte apparentemente diverse se non opposte, ma che in realtà postulano e legittimano la stessa assunzione di fondo: ovvero che il problema dell'eterogeneità linguistico semantica deve essere risolvibile, scegliendo un livello di eterogeneità che varia tra 0 (gli Ontologi) e 1 (i Relativisti).

Il problema del linguaggio e del significato è quello di definire la linea che separa ciò che possiamo assumere come “linguaggio stabile e condiviso” e ciò che invece è da considerarsi “linguaggio e significato che va negoziato di volta in volta”. Tale assunzione esclude l'aspetto processuale del significato, ovvero il modo in cui questa linea viene formulata e spostata nel tempo.

La prima fazione, quella degli Ontologi, afferma che l'eterogeneità va ridotta attraverso un processo di standardizzazione linguistica e semantica. Per la seconda fazione, quella dei Relativisti, lo sforzo di standardizzazione dei linguaggi è praticamente impossibile se non altro perché, lungi dall'essere fissati nel mondo, essi cambiano costantemente. Ma se il significato è sempre mutevole com'è possibile stabilire processi di comunicazione stabili ed efficaci? Se è vero che i sistemi linguistico semantici cambiano, è anche vero che le persone esprimono il bisogno di convergere su sistemi comuni stabili e affidabili, al fine di poter svolgere, in modo sicuro e dagli effetti prevedibili, transazioni critiche: di fatto il significato cambia ma in qualche modo si stabilizza.

Si nota così un *empasse*: da un lato l'esigenza di appoggiarsi a qualche fatto o referente esterno e condiviso per giustificare la stabilità e l'affidabilità di una comunicazione (gli Ontologi), dall'altro la necessità di spiegare che anche questo referente è in qualche modo soggetto a cambiamento (i Relativisti). E ciò è ancor più vero per l'universo linguistico che gravita intorno alla tecnologia, laddove l'interoperabilità è una necessità pressante e il mutamento è all'ordine del giorno.

Il WS dovrebbe indagare il processo attraverso il quale linguaggi e significati eterogenei vengono negoziati al fine di convergere verso

interpretazioni stabili e comuni e, di converso, come tali interpretazioni vengono nel tempo poste in discussione e rinegoziate.

Questo è esattamente il tipo di contributo che i filosofi potrebbero dare alla fondazione del WS.

La seconda grande domanda è: qual è il criterio logico per giudicare la correttezza della descrizione di un dominio di conoscenza?

Per giungere a un sistema linguistico semantico comune è necessario trovare il linguaggio e il significato idealmente corretto rispetto ad un determinato dominio. Ma dietro ogni sistema linguistico semantico si cela un mondo di interessi concreti e tangibili che da quel sistema dipendono. Negoziare un'ontologia non è un'operazione logica e astratta dal mondo: ogni cambiamento concettuale produce effetti reali e tangibili sul mondo. Effetti che tra l'altro mutando il valore di alcune risorse (nella misura in cui le rendono più o meno utili), spostano il valore da un soggetto all'altro (rendendo le risorse detenute da alcuni più utili delle risorse detenute da altri).

La natura economica di questo processo è stata di recente analizzata da alcuni economisti del linguaggio.

L'idea di fondo è che la tecnologia in uso viene considerata la migliore (giudizio *ex-post*), mentre non è vero che la tecnologia migliore venga effettivamente usata (giudizio *ex-ante*). Se le persone definiscono una situazione come reale, essa è reale nelle sue conseguenze.

Il linguaggio mostra una dinamica simile: intuitivamente quanto più è condiviso, tanto più aumenta il suo valore perché consente di comunicare con più persone. L'adozione di un linguaggio non dipende dalla sua bontà astratta (difficilmente giudicabile *ex-ante*), ma solo *ex-post* dalla sua diffusione e, indirettamente, dalla disponibilità di una serie di servizi complementari e collegati ad esso (traduttori, servizi formativi, riviste, libri, ecc.).

Ma questo stesso aspetto positivo cela quello negativo: nel momento in cui entriamo a far parte di una comunità linguistica facciamo una serie di investimenti il cui valore viene congelato in quel linguaggio. Sui linguaggi quindi investiamo risorse che se da una lato ci consentono di sfruttarli, dall'altro divengono esse stesse una barriera all'uscita, che ci disincentiva ad abbandonare quel linguaggio:

avviene un effetto *lock-in* (bloccaggio dentro), dovuto ad un costo affondato. Quanto più le persone investono in un linguaggio, tanto più tenderanno a confermarne la validità nel tempo rendendolo sempre più istituzionalizzato ed assumendo i significati che esso veicola alla stregua di oggetti del mondo.

Emerge, in primo luogo, come la negoziazione del significato non sia tanto un processo astratto tendente alla correttezza logica, quanto piuttosto una vera e propria negoziazione nel senso economico del termine: implica la considerazione di costi e benefici (valori) nelle posizioni di partenza e dell'impatto che i cambiamenti producono sulle posizioni finali dei partecipanti al "gioco linguistico".

Il WS appare dunque come un'enorme arena della modernità in cui, dietro le dichiarazioni ideali, il tema del significato si manifesta come un processo (o un gioco) tutt'altro che astratto, concretamente calato sia nelle dinamiche sociali e culturali dei partecipanti sia nei loro interessi economici.

La filosofia può dare un contributo decisivo nella scelta (e nell'analisi) delle cosiddette primitive ontologiche, nell'individuazione degli oggetti di un certo dominio, e nella rappresentazione adeguata dei concetti più generali, a partire da quelli di spazio, tempo, soggetto e oggetto, ecc..

Un secondo ambito di collaborazione è quello della comunicazione tra agenti con diverse "teorie del mondo". Infatti il WS offre un esempio concreto di come diverse comunità umane sviluppino diverse rappresentazioni persino dello stesso dominio. La soluzione di questo problema richiede un armamentario concettuali di cui i filosofi sono forniti. Si tratta infatti di definire protocolli di comunicazione volti a trovare un accordo sui significati dei termini (problema del *coordinamento* e della *negoziazione semantica*), sulla traduzione da un linguaggio all'altro, di dare una definizione di livello di comprensione accettabile.

In terzo luogo, l'accento posto dal WS sul tema della cooperazione tra agenti umani e software pone interessanti sfide rispetto ad altri temi di rilievo per la filosofia, quali quelli del rapporto tra conoscenza/credenza e azione, nonché quello della definizione di un modello di agente in cui gli stati interni dell'agente (credenza, desideri, intenzioni) determinano il suo comportamento.

Un'altra questione è quella del ruolo dell'interfaccia nella diffusione delle innovazioni. La tecnologia è tanto più utile quanto più ci permette di evitare di confrontarci con le sue difficoltà. In questo senso il contributo di umanisti, psicolinguisti, psicologi e artisti è stato fondamentale per rendere accessibili a tutti le procedure delle macchine. Il successo (anche commerciale) di dispositivi ad alta tecnologia è decretato dagli esperti che ne rendono semplice l'uso.

Il WS si pone quindi come terreno di collaborazione fortemente multi-disciplinare. Esso infatti richiede la collaborazione anche di altre discipline: la sociologia, l'economia, il diritto, la psicologia, ecc....

Il WS è un progetto ambizioso in quanto si propone come una nuova famiglia di strumenti in cui gli esseri umani possono vivere, lavorare e apprendere insieme. Un progetto ambizioso nel quale i filosofi dovrebbero giocare un ruolo da protagonisti.

La formazione umanistica svolge un ruolo insostituibile nella costruzione anche di alcune professionalità:

- produzione web e multimediale: la realizzazione di un sito web, di una intranet aziendale o di un DVD riguarda la produzione editoriale: in questo contesto l'umanista può impiegare proficuamente le proprie professionalità, coordinando la realizzazione di contenuti testuali e gestendo l'intera organizzazione dell'opera secondo i parametri di usabilità e accessibilità
- didattica a distanza (e-learning): l'invecchiamento sempre più rapido delle conoscenze richiede la realizzazione di percorsi formativi continui in qualsiasi ambito professionale perciò aumenta la necessità di competenze di didattica, tradizionalmente proprie dell'umanista
- studi sul knowledge management: in questo campo che si occupa dei metodi di gestione della conoscenza in ogni contesto lavorativo, viene riconosciuto che il capitale intellettuale presente in azienda offre il principale vantaggio competitivo e ne garantisce la sostenibilità; la padronanza degli strumenti di ricerca propria della formazione umanistica, la capacità di gestione del team work per porre alla persona giusta le domande adatte alla soluzione del nuovo problema in esame sono preziose qualità.

Il confronto con i testi, l'allenamento al ragionamento astratto, le competenze linguistiche, l'osservazione artistica, l'analisi delle fonti della storia connesse con le competenze tecnologiche sono le capacità che garantiscono il successo nei nuovi mestieri. Del resto anche i campi propri dello studente di Facoltà umanistiche, come l'archivista, il bibliotecario, il docente si rinnovano nell'incontro con la tecnologia della comunicazione e offrono nuove e più interessanti possibilità di lavoro.

## 5. Conclusioni

Alla luce della trattazione sopra esposta, e della premessa che apre tale trattazione, possiamo arrivare alla conclusione che ancora molte sfide ci aspettano prima di arrivare ad una vera collaborazione consapevole tra macchina e macchina, e che ancora tanto c'è da lavorare e da formalizzare per permettere agli OPAC di divenire un vero strumento di ricerca umanistica, in grado di portare direttamente conoscenza sul tavolo di casa nostra, e non essere semplicemente dei cataloghi per le reperibilità dei documenti da analizzare poi in modo del tutto privato.

Nonostante questo, io mi auguro che, anche se il WS si svilupperà in modo completo, esso lasci ancora qualche angolo di ricerca inesplorato dalle macchine, in modo che l'uomo possa ancora usare il proprio intuito e la propria capacità di inferenza unica al mondo per poter portare contributi particolari e originali alla conoscenza.

# 1. Bibliografia

- T. Numerico, A. Vespignani : “Informatica per le scienze umanistiche”, Il Mulino, 2003
- F. Tomasi: “Metodologie informatiche e discipline umanistiche”, Carocci editore, 2008
- [www.wikipedia.org](http://www.wikipedia.org)
- [www.uib.no/acohum](http://www.uib.no/acohum): aco\*hum, Working Group on Formal methods in the Umanities, Chapters 2, “European studies on formal methods in the humanities”, <http://gandalf.aksis.uib.no/AcoHum/fm/fm-chapter-final.html>
- [www.swif.uniba.it/lei/ai/networks](http://www.swif.uniba.it/lei/ai/networks):
  - Networks 2:
    - i-viii, 2003 (P.Bouquet, R. Ferrario: “Il Semantic Web”);
    - 7-13, 2003 (F. van Harmelen: “Il futuro del motore di ricerca: ‘cerca e troverai’. Un’intervista con Frank van Harmelen”);
    - 14-24, 2003 (A. Oltramari, S. Borgo, C. Catenacci, R. Ferrario, A. Gangemi, N. Guarino, C. Masolo, D. Pisanelli: “Il ruolo dell’ontologia nella disambiguazione del significato”);
    - 25-32, 2003 (E. Franconi: “Description Logics for Conceptual Design, Information Access, and Ontology Integration: Research Trends”);
    - 33-46, 2003 (L.Serafini, S. Zanobini: “Coordinamento semantico”);
    - 47-56, 2003 (G. Roncaglia: “Blogosfera e Feed RSS: una palestra per il Semantic Web?”);
    - 57-66, 2003 (M. Bonifacio: “Alcune domande che il Semantic Web non si fa: il processo e il valore economico del linguaggio e del significato”);
  - Networks 6:
    - 137-164,2006 (Barry Smith: “Ontologia e sistemi informativi”).
- C. Gnoli: “L’indicizzazione semantica dalla biblioteca al Web”, Supporto didattico, ver. 1.4: 2006.02, <http://www-dimat.unipv.it/gnoli/semantica/>
- C. Gnoli: “Gli opac. Una guida per il pubblico all’utilizzo dei cataloghi in linea”, in AIB-WEB. Contributi, <http://www.aib.it/aib/contr/gnoli1.htm>
- C. Gnoli: “Il tavolino di Ranganathan”, <http://www.geocities.com/Athens/Agora/7070/ranga.htm>
- C. Gnoli: “Opac in Italia: una panoramica delle tipologie e delle modalità di consultazione”, <http://www2.spbo.unibo.it/bibliotime/num-ii-1/gnoli.htm>
- R. Raieli, P. Innocenti : “Multimedia Information Retrieval: stato dell’arte e prospettive di applicazione” su [www.aidainformazioni.it/pub](http://www.aidainformazioni.it/pub), trimestrale elettronico, anno 21, numero 4, Ottobre-Dicembre 2003
- S. Spinelli: “Introduzione all’indicizzazione”, Macerata, 2006, <http://biocfarm.unibo.it/~spinelli/indicizzazione/>
- A. Robbio: Seminario AIB-WEB-2, 18 maggio 1999, “L’evoluzione della specie: dagli OPAC al MetaOPAC. Presentazione del MAI MetaOPAC Azalai Italiano”, <http://www.aib.it/aib/congr/co99metaopac.htm>
- H. Saber: “La pubblica amministrazione e il web semantico: un approccio possibile”, <http://tecnologie.forumpa.it/story/41892/la-pubblica-amministrazione-e-il-web-semantico-un-approccio-possibile>
- F. Di Giammarco: “Il Web Semantico”, <http://www.culturadigitale.it/articoli/PDF/Il%20WebSemantico.pdf>
- A. Dorati e S. Costantini: “Approcci al Web Semantico”, articolo, <http://www.websemantico.org/articoli/approciwebsemantico.php>

- Tecnoteca.it: “Massimo Martinelli: il web semantico”, intervista, <http://www.tecnoteca.it/interviste/martinelli>
- “Cos’è e a cosa serve il web semantico”, <http://tecnologie.forumpa.it/story/41884/cose-e-a-cosa-serve-il-web-semantico>
- “L’ABC del Web semantico”, intervista a M. Colombetti, <http://tecnologie.forumpa.it/story/41885/labc-del-web-semantico>
- “La rappresentazione della conoscenza nel web semantico: Introduzione”, [http://www.elearninglab.eu/studying/sw/sw\\_intro.html](http://www.elearninglab.eu/studying/sw/sw_intro.html)
- D. Bogliolo: “Metadati: parte 1 sintesi del problema e bibliografia essenziale; parte 3 indicizzazione semantica”, <http://www.uniroma1.it/documentation/metadati.html>
- Vari post nell’area dell’Insegnamento di Informatica per le Scienze umane - Attività pratica associata al corso- tenuto dal Prof. D. Buzzetti presso l’Università di Bologna (<http://antonietta.philo.unibo.it/blog/>), tra cui:
  - A. Carano: “Web semantico: una questione di paradigmi gnoseologici”, Marzo 2008
  - M. Bravi: “Il web barocco e la sua ontologia informatica”, Dicembre 2007
  - E. D’Alessandro: “Il web semantico”, Dicembre 2007
  - I. Lazzarin: “Il computer tra psicologia, linguaggio e conoscenza”, Novembre 2007
  - F. Nipoti: “Usabilità e web semantico”, Novembre 2007
  - E. Bergianti: “Berners-Lee e Guarino: due diversi approcci allo sviluppo del Web Semantico”, Novembre 2007
  - T. Castelli: “I limiti del web semantico”, Maggio 2007
  - L. Sagripanti: “Entità astratte”, Maggio 2007
  - F. Pontiroli: “Il futuro del web: web semantico”, Marzo 2007
  - D. Iori: “Il Web semantico ed i suoi linguaggi: un’introduzione”, Gen 2007
  - J. Nuvolari: “Web semantico: una nuova lingua franca”, Aprile 2006

## 2. Ringraziamenti

Vorrei ringraziare:

- mia madre Clara che, nei primi anni di iscrizione all'Università si è sobbarcata, non senza difficoltà, il carico economico del mio mantenimento a Bologna per ben tre anni, oltre che i costi dell'iscrizione;
- mio marito Guido che, negli anni, ha vissuto con me ansie, delusioni e soddisfazioni, e che durante la preparazione degli esami e di questa tesi ha tollerato con pazienza i miei frequenti sbalzi d'umore;
- le mie amiche Cristina, Ida e Lina che, con entusiasmo, mi hanno sostenuto emotivamente per tutto il corso dei miei studi;
- la mia amica Daniela, per il minuzioso lavoro di traduzione di alcuni testi dall'inglese (che io ancora devo imparare) all'italiano, e per il prezioso supporto nella preparazione dell'esame di abilitazione alla lingua francese, senza il quale avrei avuto molte difficoltà;
- i miei colleghi di lavoro più stretti Gino M. e Gino A. (più amichevolmente chiamati Gino1 e Gino2 in base all'anzianità d'ingresso nel nostro reparto di lavoro), per avermi sopportata in questi anni e aiutata, sostituendomi egregiamente all'occorrenza sul posto di lavoro;
- i Dirigenti Scolastici che si sono succeduti, durante il mio percorso accademico, alla guida dell'Istituto Tecnico Industriale Statale di Cremona dove lavoro, nelle persone del Dott. Guido Lazzarini e della Dott.ssa Maria Paola Negri, e il Direttore dei Servizi Generali Amministrativi, nella persona della Sig.ra Maria Silvia Gozzoli, per la loro disponibilità alla concessione dei permessi di studio che mi hanno permesso di arrivare a questo traguardo;
- il relatore di questa tesi Prof. Dino Buzzetti, per aver voluto fortemente questa tesi, vedendo nel mio doppio ruolo di tecnico informatico e di filosofo un'opportunità per svolgerla in modo armonico e senza pregiudizi di sorta, in anni in cui la filosofia, purtroppo, guarda ancora con sospetto verso l'informatica e viceversa;
- il personale della Segreteria di Facoltà e quello dell'Ufficio Didattico del Dipartimento di Filosofia, sempre gentile e puntuale nel risolvere in modo semplice i miei dubbi e le mie perplessità;
- la città di Bologna e i suoi cittadini che mi hanno accolto come una figlia, con la loro simpatia e disponibilità;
- tutti quelli che mi hanno sostenuto e appoggiato, e che involontariamente ho dimenticato di nominare, e anche quelli che magari senza volerlo mi hanno ostacolato, perchè così facendo mi hanno dato quella grinta in più che mi è servita a portare a termine questo progetto personale.