



archividelnovecento
la memoria in rete

LE PAROLE DEL NOVECENTO – UN THESAURUS PER GLI ARCHIVI

Criteri metodologici

a cura del

Gruppo di lavoro sul Thesaurus

di *Archivi del Novecento*

Roma, gennaio 2007

Sommario

Introduzione	3
1. Peculiarità del contesto archivistico e riflessi sull'indicizzazione semantica.....	5
2. Informatica e indicizzazione	10
3. Concetti preliminari sul thesaurus.	11
4. Scelta del sistema di indicizzazione.	12
5. Specificità dei descrittori e livello archivistico.	15
6. Elementi da inserire o escludere dal campo descrittori (enti, antroponomi, toponimi, tipologie archivistiche).	17
7. I criteri per l'elaborazione dei descrittori: 1. Il rapporto con la natura delle basi dati. <i>La core-area</i>	18
8. I criteri per l'elaborazione dei descrittori: 2. Il rapporto con il tipo di utenza.	20
9. Metodo di sviluppo del thesaurus.	23
10. Criteri per il controllo numerico, morfologico (relativo a categorie grammaticali nella forma del nome) e semantico (relativo all'univocità dei termini); la relazione di equivalenza.	24
11. Struttura semantica e criteri per la costruzione delle relazioni gerarchiche (genere/specie, parte/tutto) e associative.....	28
12. Criteri per la classificazione dei descrittori nella presentazione sistematica.	31

Introduzione

Dal gennaio 2004 è stato avviato dal Consorzio Baicr Sistema cultura uno studio sperimentale dedicato all'ipotesi di un vocabolario controllato come strumento per raffinare la ricerca archivistica sulle basi dati degli istituti che aderiscono alla rete *Archivi del Novecento*¹. La sperimentazione, condivisa nei suoi nodi concettuali e nei suoi obiettivi dagli organismi competenti del Ministero per i beni e le attività culturali, ha potuto contare sul sostegno e sul contributo dell'allora Ufficio studi e pubblicazioni dell'Ufficio centrale per i beni archivistici e successivamente del Dipartimento per i beni archivistici e librari, entrambi in tempi diversi diretti da A. Dentoni-Litta.

Il problema della soggettazione o indicizzazione semantica dei documenti archivistici ha conosciuto una fase di grande interesse intorno agli anni '90 suscitando iniziative, anche a livello internazionale, e sollecitando alcune importanti riflessioni, che hanno condotto a chiarire, in gran parte, gli aspetti problematici legati all'accesso tematico agli archivi.

La consapevolezza con cui, in quella fase, ci si concentrò sulle peculiarità del contesto archivistico rispetto a quello biblioteconomico, la novità di un simile approccio tematico alle fonti e l'emergere di numerosi nodi problematici resero allora difficile la prosecuzione sul cammino concreto della sperimentazione. La discussione si arenò lasciando dietro di sé numerosi dubbi sulla reale possibilità di adottare strumenti e metodologie nati in contesti affini a quello archivistico ma con problematiche e obiettivi diversi.

Numerosi sono i fattori che contribuiscono oggi a richiamare l'attenzione sul problema della indicizzazione delle fonti archivistiche: l'evoluzione dell'informatica e la sua applicazione nel contesto degli archivi correnti, di deposito e storici; la diffusione degli standard nel settore archivistico; la comunicazione interdisciplinare fra archivisti e bibliotecari; la spinta impressa al settore archivistico dalla logica dell'accesso ai documenti tramite Internet. Tutto ciò non può che stimolare un rilancio delle iniziative di confronto con il problema dell'integrazione fra ricerca per argomento e ricerca di tipo archivistico.

Fin dall'inizio si è sempre sottolineato come l'accesso tematico costituisse semplicemente un complemento rispetto alla ricerca di tipo storico-istituzionale e archivistica; una ricerca cioè che muove anzitutto dal soggetto produttore della documentazione - piuttosto che dal

¹ Il gruppo di progetto è composto da Sabrina Auricchio, Patrizia Gabrielli, Simona Luciani e Cristiana Pipitone. I risultati del lavoro sono sistematicamente discussi e revisionati con la supervisione scientifica di

suo contenuto - e che si avvale, in secondo luogo, degli strumenti di corredo tradizionali che permettono di orientarsi nella struttura dell'archivio. Da queste premesse si è partiti per rilanciare la discussione sullo sviluppo di vocabolari controllati per l'indicizzazione di strumenti di ricerca archivistici.

Il progetto Archivi del Novecento, per la sua particolare natura e storia, è un contesto applicativo particolarmente favorevole allo sviluppo di una simile sperimentazione. Fin dalla sua primitiva ideazione Archivi del Novecento ebbe come obiettivo, e suo maggiore punto di forza rispetto ad altre esperienze di applicazione informatica agli archivi, l'integrazione delle fonti all'interno di una rete archivistica, con la possibilità di ottenere la ricostituzione virtuale dei fondi prodotti da uno stesso soggetto produttore ma conservati presso istituti diversi². Fra le dichiarazioni d'intenti dei suoi promotori c'era inoltre l'idea "di far uscire da un circuito esoterico il sapere archivistico per restituire al sociale, in maniera fondata, quanto la società ha prodotto"³.

Questo appare ancora oggi un punto di vista di grande attualità, poiché la tendenza generale a rendere fruibili su Internet inventari, banche dati e documenti in formato digitale è proprio quello di allargare a una utenza più vasta la conoscenza dei patrimoni archivistici e, insieme, la conoscenza storica.

L'evoluzione del compito degli archivisti deve tenere presenti le necessità che nascono dalla trasformazione dei mezzi di comunicazione del sapere storico e dall'incremento degli interlocutori di riferimento. Per far ciò gli stessi archivisti sono chiamati a far evolvere la loro professionalità per aumentare le possibilità di andare incontro a nuovi tipi di utenza stimolata dalle evoluzioni tecnologiche.

Proprio la rete di Archivi del Novecento - che fa comunicare archivi appartenenti a un arco cronologico ampio ma circoscritto e che arricchisce il proprio valore dagli intrecci documentari prodotti dalle relazioni fra fondi diversi - è un contesto in cui si è fatta al tempo stesso urgente e inderogabile la predisposizione di strumenti supplementari per l'accesso alle carte.

Il lavoro di oggi non fa dunque che portare a compimento un percorso avviato fin dalle origini del progetto Archivi del Novecento: già allora aveva preso corpo la volontà di

referee nei diversi ambiti disciplinari: Gabriella Nisticò, Marisa Trigari, Lucia Zannino, Leonardo Musci, Antonella Mulè, Mauro Tosti Croce, Guido Melis, Gianfranco Crupi.

² GABRIELLA NISTICÒ, *Archivistica contemporanea e informatizzazione degli archivi storici: il Progetto Archivi del '900*, in *Gli archivi dei partiti politici*, Roma, Ministero per i Beni Culturali, Ufficio studi e pubblicazioni, 1996.

³ *ibidem*.

fornire alla rete archivistica gli strumenti tecnici per la consultazione delle schede tramite parole chiave. Il software Gea, utilizzato dalla rete Archivi del Novecento, nelle sue prime versioni conteneva, infatti, un modulo denominato *Thesaurus*, che ancora oggi attende di essere sviluppato e reso operante. Inoltre nel software le schede descrittive di tutti i livelli dispongono, all'interno dell'area riservata agli strumenti di indicizzazione, di un campo *Descrittori* offerto fino a oggi come terreno di sperimentazione libero agli archivisti, ma senza tuttavia che la sua compilazione sia stata supportata da indicazioni di orientamento metodologico o di normalizzazione.

Gli stessi tentativi liberamente intrapresi dagli archivisti della rete in maniera sporadica e non coordinata hanno portato a maturazione la consapevolezza della necessità di un rilancio dell'idea del thesaurus, o comunque hanno reso evidente l'opportunità di dettare criteri almeno per l'elaborazione dei descrittori.

1. Peculiarità del contesto archivistico e riflessi sull'indicizzazione semantica.

Nel dibattito aperto sull'indicizzazione per gli archivi si è sempre sottolineata la differenza esistente fra il mondo archivistico e quello delle biblioteche, quasi a ribadire l'assoluta necessità che gli archivisti procedessero a darsi regole proprie, diverse da quelle sviluppate dai bibliotecari. Questo aspetto ha un suo fondamento se si pensa all'ordinamento e alla descrizione dei reciproci oggetti di lavoro, attività che, nei rispettivi settori, seguono logiche completamente diverse proprio in considerazione delle profonde difformità che riguardano la natura stessa di tali beni culturali⁴. Il libro è un bene assoluto, definito come oggetto, laddove il documento archivistico – pur avendo un contenuto informativo autonomo – è un bene relativo, ha bisogno del suo contesto, di legami obbligatori e correlazioni che lo legittimino. Queste differenze però non sembrano entrare in gioco nella fase di elaborazione delle parole chiave e nella costruzione di un dizionario controllato; ovvero si può dire che le peculiarità della documentazione archivistica non influiscano tanto sull'elaborazione di standard distinti, o sull'uso di criteri metodologici differenziati nella progettazione e realizzazione di un vocabolario controllato di termini, quanto piuttosto nelle modalità di utilizzo del vocabolario stesso.

È nella fase di indicizzazione, ovvero nell'applicazione concreta delle parole chiave alla base dati, che le operazioni acquistano una specificità per il settore archivistico. I bibliotecari

⁴ In proposito cfr. le osservazioni contenute in PAOLA CARUCCI, *Le fonti archivistiche: ordinamento e conservazione*, Roma, Nis, 1983.

compiono un'operazione di rappresentazione a partire dal documento primario (monografia o periodico), traendo i soggetti dall'analisi concettuale del volume (basata su titolo, indice, IV di copertina, introduzione). Invece gli archivisti – ed è uno dei punti fermi acquisiti dalla sperimentazione per il progetto “Le parole del Novecento” - all'interno di una base dati possono indicizzare il contenuto delle schede descrittive, e non il contenuto dei documenti. Sono infatti le stesse Isad a prevedere che l'individuazione dei punti di accesso siano basati sugli elementi di descrizione; specificando inoltre che il valore dei punti di accesso viene aumentato attraverso l'applicazione del controllo di autorità⁵. La differenza è sostanziale non solo per motivazioni di ordine metodologico ma anche di natura operativa: l'indicizzazione tematica costituisce una sintesi informativa che viene elaborata su una fonte che, a sua volta, è una informazione mediata, filtrata e rappresentata sinteticamente nella scheda descrittiva dall'archivista. Quindi a differenza del bibliotecario, l'archivista per indicizzare tematicamente non parte dal documento primario, bensì da un lavoro di mediazione concettuale già effettuato.

Si tenga presente che all'interno di una base dati archivistica come quella di Archivi del Novecento convivono descrizioni di insiemi profondamente diversi (fondi, livelli, fascicoli, singoli documenti). Ne deriva che una stessa voce d'indice potrà essere quindi assegnata a oggetti disomogenei fra loro, ovvero a schede che descrivono un singolo documento o un intero fondo (ovvero anche centinaia di migliaia di documenti).

Da ciò derivano alcune conseguenze:

- 1) La ricerca tematica in una base dati archivistica avviene con un passaggio in più rispetto a quella degli opac bibliotecari, nei quali i termini di indice (che possono essere stringhe di soggetto o descrittori) rinviano al singolo “oggetto-libro” (o periodico, o estratto, ecc.). Per Archivi del Novecento il rapporto non è parola chiave/documento, ma parola chiave/scheda descrittiva che, come detto, può contenere la sintesi del contenuto di un documento, come di quello di un fascicolo con potenzialmente centinaia di argomenti, che necessariamente devono essere già stati mediati dall'archivista.
- 2) Il ruolo affidato alla descrizione come fonte dell'indicizzazione tematica comporta che le parole chiave, inserite nel campo *Descrittori* presente in ogni tipologia di scheda, devono sempre rinviare a dei concetti espressi in linguaggio naturale in uno degli altri campi a testo libero (*Contenuto*, *Ordine del giorno*, *Note*, *Allegati*, ecc.). Ciò significa che l'indicizzazione

⁵ “Access points are based upon the elements of description. The value of access points is enhanced through authority control”, in Isad G, II editions point I.14.

tematica è concepita come uno strumento di recupero delle informazioni presenti in altri campi della scheda di descrizione archivistica e non come integrazione o completamento autonomo della stessa. Questa scelta metodologica riflette l'impostazione adottata dagli utenti di Archivi del Novecento relativamente a qualsiasi voce d'indice (quindi valida anche per nomi di persona, di enti, di luoghi). In questo modo, una volta raggiunta la scheda archivistica attraverso la ricerca per parola chiave, dall'analisi del contenuto si potrà dedurre in che contesto il descrittore è stato utilizzato. Tale impostazione rappresenta una differenza rispetto alla catalogazione bibliotecaria; infatti la soggettazione di libri non rimanda ad alcuna informazione mediata (e non sempre il titolo è indice del contenuto), quindi l'indicizzazione tematica è l'unica fonte di segnalazione dell'oggetto.

3) In questo processo diventa evidente come il ruolo degli archivisti autori degli inventari informatizzati sia centrale, sia che indicizzatore ed archivistica siano la stessa persona, sia che l'indicizzatore operi a posteriori su una descrizione archivistica preesistente.

Nel primo caso, l'archivista/indicizzatore semantico compie in parallelo una sintesi degli elementi essenziali identificativi del contenuto, un'operazione quindi che presiede sia all'elaborazione della descrizione archivistica, sia alla costruzione di un index semantico; nel secondo caso l'indicizzatore, senza tornare all'unità archivistica primaria, utilizza la descrizione ed il parziale controllo semantico già effettuati dall'archivista, per procedere alla costruzione di un index in linguaggio controllato.

L'importanza del concetto per cui vada indicizzata la scheda e non il contenuto diretto dei documenti (metodo che viene adottato infatti di solito in progetti di descrizioni a livello esclusivamente documentale) è evidente soprattutto nel caso di descrizioni di fascicoli, serie, fondi, ovvero di unità complesse, che possono contenere un numero indefinibile di unità documentarie, con tanti potenziali argomenti.

Ad es. si può immaginare una scheda che descriva un fascicolo dal titolo "Problemi di governo", composto da 100 documenti. Se l'archivista compila per tale fascicolo il seguente campo *Contenuto*: "Testi di discorsi relativi alla crisi del Mezzogiorno; un opuscolo relativo alla politica economica; rassegna stampa relativa alla disoccupazione giovanile", significa che ha già operato una sintesi descrittiva del contenuto, riportando le tipologie dei documenti in connessione ai molteplici aspetti della politica economica trattati dai documenti. Su questa sintesi saranno elaborate le parole chiave utilizzando le regole adottate: *questione Meridionale / politica economica / disoccupazione*. Queste parole chiave slegate dalle

schede non restituiscono il reale contenuto informativo del fascicolo. Per ottenere una piena comprensione del contenuto sarà necessario leggere la scheda descrittiva.

Diventa così evidente la stretta interconnessione tra qualità della descrizione e qualità dell'index.

Nella fase di sperimentazione del progetto è emerso chiaramente un dato: l'operazione concettuale volta alla elaborazione dei descrittori richiede uno sforzo di identificazione che giova anche alla formulazione della descrizione archivistica. Se nell'elaborazione di quest'ultima ci si pone anche l'obiettivo di estrapolare parole chiave, diventa necessaria una maggiore chiarificazione del contenuto, che avrà una ricaduta positiva sulla qualità e sulla efficacia della descrizione stessa.

Un grave ostacolo a una efficace indicizzazione semantica si può infatti presentare nei casi in cui la descrizione archivistica tenda a essere "economica" ed essenziale.

È perciò importante che chi descrive lo faccia con la consapevolezza della futura indicizzazione. Da questo punto di vista sarebbe meglio che le operazioni avvenissero contestualmente; ma ciò non è strettamente necessario. Ne deriva che la figura dell'archivista, con la sua capacità di analisi e di sintesi, diventa centrale anche nel lavoro di indicizzazione tematica: sono l'ordinamento e la descrizione dei fondi a costituire la base per la successiva indicizzazione. Quanto detto fin qui non deve tuttavia far pensare che l'indicizzazione archivistica sia una operazione semplice né che essa sia sempre possibile.

Nei casi in cui la descrizione è troppo generica (per es. quando una descrizione presenta solo un titolo generico come: "Corrispondenza 1957"), o quando il suo contenuto pur essendo descritto puntualmente è troppo vario, risulta impossibile assegnare delle parole chiave che ne sintetizzino il contenuto. Anche in questi casi va sottolineato che la descrizione archivistica deve essere già pensata in funzione dell'indicizzazione o, ancora meglio, essa andrebbe effettuata contestualmente alla descrizione. In alternativa, in tutti i casi in cui non è possibile indicizzare per mancanza di elementi, sarebbe opportuno almeno marcare tutte le schede non indicizzate con un identificativo come "n.i." (= non indicizzabile) o altra annotazione, che consenta all'utente di sapere che alcune schede sono rimaste escluse dall'accesso tematico, e far sì che queste possano essere richiamate tutte insieme ed esaminate dagli utenti.

4) Va inoltre aggiunto che la distinzione determinante tra i due ambiti è data dalla portata informativa degli indici tematici che per gli archivi è strettamente connessa con il contesto.

Come si è già detto, la ricerca per parole chiave in ambito archivistico, in particolare in una base dati complessa e variegata (in cui cioè non viene inventariata informaticamente un'unica tipologia documentaria) deve servire per approdare alle schede di descrizione archivistica. Una volta ottenuta una lista di occorrenze legate a un descrittore, per poterne valutare pienamente il significato sarà comunque necessario ricondurre le parole chiave al contesto di produzione delle fonti archivistiche. Su questo aspetto occorre dire che le possibilità di utilizzare al meglio gli strumenti di indicizzazione sono ampliate in maniera decisiva dall'efficienza del modulo di gestione informatico della ricerca tramite thesaurus. Partendo dalla navigazione del thesaurus il sistema dovrebbe consentire di selezionare i termini prescelti e di lanciare direttamente la ricerca sulla base dati. Quindi, una volta ottenuta la lista dei risultati di ricerca, dovrebbe essere possibile aprire ogni scheda all'interno del suo contesto documentario.

L'accesso alla documentazione archivistica può seguire due percorsi: quello storico-istituzionale e quello tematico; l'uno non sostituisce l'altro, ma ciascuno risponde a una propria logica. Il primo impiega come strumento principale di ricerca l'inventario, il secondo invece utilizza gli indici di argomenti.

Tuttavia molti tentativi di introduzione di strumenti di accesso per materia hanno confuso i due approcci e i due strumenti, che invece vanno tenuti accuratamente separati, poiché essi possono fornire i migliori risultati se utilizzati in modo distinto ma integrato.

Una ragione di questa confusione può essere rintracciata nella sovrapposizione fra due oggetti apparentemente simili in quanto entrambi costituiti da strutture gerarchiche di voci, ma in realtà profondamente diversi: uno appartenente alla tradizione archivistica (titolario) e l'altro a quella biblioteconomica-documentaristica (thesaurus).

Dall'analisi di esperienze di indicizzazione precedenti è invece emerso che, laddove sono stati utilizzati sistemi di indicizzazione *precoordinati* (in particolare con l'uso di stringhe di soggetto)⁶, le relazioni tra i termini della stringa riflettevano soprattutto il contesto archivistico e ne rispecchiavano i legami gerarchici. La stringa non mirava tanto a descrivere il contenuto, quanto a formare dei collegamenti tra il termine di indice e il contesto gerarchico; non fornendo quindi un accesso diverso e ulteriore.

⁶ Si intende la “formulazione di soggetti nei quali i diversi descrittori siano giustapposti a costituire un'unica stringa che esprima il contenuto del documento, piuttosto che rimanere indipendenti e associabili in una specifica combinazione solo nella fase di interrogazione (postcoordinazione)”, cfr. CLAUDIO GNOLI, *Coordinazione, ordine di citazione e livelli integrativi in ambiente digitale*, <http://www-dimat.unipv.it/~gnoli/coord.htm> (controllato il 19 maggio 2005).

L'utilizzo del metodo preordinato è fondato sul presupposto che esso garantisca una maggiore precisione nell'individuare i documenti. Le stringhe di soggetto, se possono funzionare come strumento di ricerca per singoli fondi archivistici, hanno lo svantaggio di non dimostrarsi efficaci come accesso a insiemi documentari ampi e composti. Se è vero che i sistemi di indicizzazione pre-coordinati hanno maggiore precisione, allo stesso tempo risultano più limitati: una lista di stringhe di soggetto permette di individuare in maniera esatta un argomento molto specifico, ma è piuttosto dispersiva e difficile da scorrere, non offre un decisivo orientamento, né suggerimenti su come allargare o completare la ricerca relativa a un determinato tema.

Viceversa in un sistema *post-coordinato*⁷ la scelta dei “concetti-termini” da coordinare viene fatta dall'utente nel momento della ricerca e non al momento dell'indicizzazione. Il ricercatore insomma è libero di porre domande e il thesaurus lo aiuta a interrogare la base dati traducendo le sue parole nel linguaggio dell'indicizzatore.

La lettura tematica va dunque concepita come una lettura orizzontale, “trasversale” agli insiemi archivistici, che dovrebbe essere in grado di raggruppare di volta in volta intorno a uno stesso tema quanto i soggetti produttori hanno prodotto separatamente su uno stesso argomento.

2. Informatica e indicizzazione

L'apporto che lo sviluppo informatico può dare alla gestione degli indici è un argomento di grande rilievo che si intende mettere in evidenza. Il thesaurus è uno strumento per sua natura complesso, che spesso risulta ostico perfino agli addetti ai lavori, e che, pur offrendo grandi opportunità per migliorare la ricerca, rischia, a causa della scarsa semplicità, di non essere utilizzato appieno. Un'interfaccia amichevole del thesaurus può aiutare a superare la diffidenza degli utenti, migliorandone il grado di utilizzazione.

Lo sviluppo informatico interviene anche per un altro aspetto legato alla interoperabilità e riusabilità dei dati del thesaurus. Il software adottato per la costruzione del thesaurus⁸ consente, oltre all'output in HTML, un output in formato RDF (Resource Description Framework), specifica versione del formato XML, elaborato dal Consorzio W3C per la rappresentazione di informazioni su Web.

⁷ Nel sistema postcoordinato i singoli concetti sono rappresentati da singoli termini.

⁸ Si tratta del software Tms, sviluppato dall'ingegnere informatico Antonio Ronca per l'Indire.

RDF presenta una sintassi astratta ed essenziale per descrivere informazioni strutturate. Per le sue caratteristiche RDF/XML si presta dunque naturalmente ad essere utilizzato nell'ambito della descrizione e gestione di thesauri in ambiente elettronico e garantisce:

- potenziale esportabilità e interoperabilità fra contesti applicativi diversi;
- capacità di gestire thesauri multilingui basati sulla descrizione dei concetti e non dei termini.

L'utilizzo di simili standard è divenuto oggi un fattore decisivo tanto da costituire un presupposto irrinunciabile per lo sviluppo di un progetto di lunga durata.

3. Concetti preliminari sul thesaurus.

Prima di descrivere le caratteristiche e la genesi del lavoro svolto, sembra opportuno fornire alcune definizioni preliminari relative al thesaurus monolingua che si presenta, utilizzato per l'indicizzazione e il reperimento dei documenti descritti nella base dati di Archivi del Novecento.

Con la parola **indicizzazione** si intende la tecnica che utilizza i termini che compongono il thesaurus per estrarre concetti racchiusi nelle descrizioni archivistiche. Un **thesaurus**, secondo gli Standard Iso, è “un vocabolario di un linguaggio di indicizzazione controllato”, relativo a un campo specifico della conoscenza, composto da termini correlati gerarchicamente e semanticamente⁹. I termini che compongono il thesaurus sono definiti **descrittori** o **termini preferiti** da utilizzarsi sia per l'indicizzazione che per la ricerca. Per maggiore chiarezza va precisato che i descrittori possono essere costituiti da una sola parola o da più parole, ovvero da sintagmi. Ai descrittori si affiancano i **non descrittori**, o **termini non preferiti**: potenziali sinonimi o quasi sinonimi dei termini preferiti presenti nel thesaurus, utilizzati come chiavi di accesso che rinviano al descrittore.

Un passo fondamentale per la realizzazione di un thesaurus è la costruzione di una rete di relazioni fra i termini per guidare l'utente fra i concetti che ruotano intorno al significato di un singolo termine e ad accedere in maniera mirata all'informazione. Tutte le voci che compongono il thesaurus sono quindi legate da un sistema di relazioni semantiche (di equivalenza e preferenza), gerarchiche (genere/specie e parte/tutto) e associative (correlazioni)¹⁰. Le relazioni sono indicate dalle seguenti sigle in lingua inglese:

⁹ MARISA TRIGARI, *I thesauri*, dispense per il Master docente bibliotecario, Terzo modulo, unità 3, p. 31.

¹⁰ Per una spiegazione più dettagliata vedi oltre, paragrafo 10.

equivalenza:

rinvio diretto USE (use)

rinvio inverso UF (use for)

gerarchiche:

termine più generale BT (broader term)

termine più specifico NT (narrower term)

termine di testa TT (top term)

associative:

termine correlato RT (related term)

Il thesaurus è quindi uno strumento che unisce un insieme organizzato di termini; tali termini sono utilizzati per rappresentare sinteticamente il contenuto concettuale espresso da una scheda di descrizione archivistica (in questo caso dalle schede fornite dal software Gea). In un sistema come Archivi del Novecento, in cui il thesaurus poggia su una base dati documentaria, i descrittori rappresentano un punto di accesso alla descrizione dei documenti che integra la ricerca a testo libero, la ricerca per nome di ente o di persona e naturalmente la ricerca per contesto archivistico.

Il linguaggio controllato assolve due funzioni: è utilizzato dagli operatori per indicizzare le schede di descrizione archivistica, cioè per descriverne il contenuto in maniera sintetica e controllata; è impiegato dagli utenti per recuperare i documenti di interesse su un certo argomento. La prima funzione si traduce nell'assegnare uno o più descrittori a una scheda descrittiva, in modo da sintetizzarne i concetti principali; la funzione di reperire informazioni su argomenti di interesse avviene interrogando il sistema informativo mediante i termini del thesaurus e di ottenere da esso, attraverso le relazioni tra termini, suggerimenti circa la possibilità di raggiungere contenuti collegati all'oggetto della ricerca.

4. Scelta del sistema di indicizzazione.

Una scelta preliminare investe il sistema di indicizzazione: preordinato o postordinato. Nel primo caso, come si è detto, “la ricerca avviene tramite termini forniti in forma già completa sulla base di formule che hanno una configurazione rigida e definitiva”¹¹; nel secondo caso i termini sono isolati e l'espressione di concetti specifici si otterrà in fase di

¹¹ *Documentazione e biblioteconomia*, a cura di MARIA PIA CAROSELLA e MARIA VALENTI, Milano, Angeli, 1982, p. 147.

ricerca con la combinazione dei descrittori attraverso gli operatori logici (AND, OR, NOT, AND NOT).

Per effettuare questa scelta, il progetto “Le parole del Novecento” si è basato anzitutto sull’analisi di esperienze di indicizzazione concretamente realizzate e di iniziative rimaste, per diversi motivi, in fase di studio di fattibilità¹².

Sulla scorta delle esperienze esaminate sembra utile e adeguato alle esigenze della base dati di riferimento orientarsi verso l’uso di descrittori espressi in linguaggio naturale ma controllato e affidarsi a un sistema di indicizzazione di tipo post-coordinato, in cui cioè la composizione degli elementi che formano un soggetto avviene in fase di ricerca e non viene fissata a priori dall’indicizzatore.

In particolare si è scartata la possibilità di indicizzazione tramite l’uso di stringhe di soggetto sulla base delle seguenti considerazioni:

- Un’indicizzazione tramite i descrittori di un thesaurus consente di esprimere in modo agile e funzionale più soggetti principali e secondari, una necessità frequente in rapporto ad unità documentarie complesse come sono quelle archivistiche; in tale situazione non sarebbe impossibile introdurre più stringhe di soggetto, ma ciò determinerebbe un alto tasso di ridondanza, anche se il contenuto sarebbe obiettivamente descritto in maniera più specifica.

Per es., con riferimento ad un fascicolo contenente “circolari relative all’organizzazione di congressi e della propaganda di partito”, utilizzando un’indicizzazione analitica con thesaurus sarebbe sufficiente introdurre tre descrittori separati: *partiti / congressi / propaganda*; scegliendo la soggettazione, bisognerebbe introdurre due stringhe di soggetto: *Partiti – Congressi* e *Partiti – Propaganda*, il che comporterebbe la ripetizione, all’interno della stessa scheda, di alcuni termini di indicizzazione).

- La rigidità caratteristica della stringa di soggetto, che attribuisce una priorità logica al *focus* rappresentato dalla testa di soggetto, ha senso piuttosto in un catalogo cartaceo ad

¹² Oltre a uno sguardo complessivo sui thesauri esistenti, sono state analizzate nello specifico alcune esperienze per verificarne l’efficacia e l’eventuale applicabilità di alcune metodologie adottate al contesto di Archivi del Novecento, in particolare: 1) indice dei descrittori della *Guida alle fonti per la storia dei movimenti in Italia*, a cura di M. Grispigni e L. Musci; 2) indicizzazione attraverso descrittori dell’Archivio storico della Cgil; 3) indicizzazione, attraverso stringhe di soggetto, del progetto *Archifirenze* del Archivio storico del Comune di Firenze; 4) descrittori applicati al progetto di schedatura del Copialettere commerciale di Vieusseux, del Gabinetto G.B. Vieusseux di Firenze; 5) thesaurus della banca dati dell’Istituto Luce; 6) thesaurus TESEO del Senato della Repubblica, inquadrato nella classificazione decimale universale (CDU); 7) Thesaurus regionale toscano promosso dalla Giunta regionale toscana; 8) thesaurus *Linguaggiadonna*, a cura del Centro di studi storici sul movimento di liberazione della donna in Italia.

accesso lineare che in ambiente di *information retrieval* in cui ogni elemento della stringa è raggiungibile in maniera indipendente e slegata dalla testa di soggetto.

- Nello stesso settore delle biblioteche si sta studiando la possibilità di abbandonare progressivamente la logica della soggettazione per aprirsi sempre di più alla utilizzazione di sistemi come i thesauri. La BNCF, per esempio, ha elaborato uno *Studio di fattibilità* per il nuovo “soggettario” che prevede tra l'altro una possibile trasformazione del *Soggettario*, per il tramite del sistema PRECIS, nel modello del thesaurus¹³.
- La stringa di soggetto richiede un grado di specializzazione professionale di ambito bibliotecario che gli archivisti difficilmente possiedono. Il grado di complessità nell'operazione di elaborazione della stringa di soggetto implica un alto margine di errore. L'esperienza della soggettazione di SBN rende evidente che anche in quel contesto è difficile raggiungere un alto tasso di omogeneità e coerenza nella formulazione delle stringhe.
- La soggettazione pecca di rigore e coerenza nell'uso del linguaggio documentario, in quanto il Soggettario della Biblioteca nazionale centrale di Firenze (BNCF), largamente utilizzato come lista d'autorità piuttosto che come modello di riferimento, esercita un controllo forte sulla testa di soggetto, ma debole o nullo sui termini di suddivisione.
- A differenza del soggettario, il thesaurus offre la possibilità di ricerca per categorie, grazie alla macroclassificazione che si affianca normalmente alla microclassificazione dei descrittori. La classificazione implicita nella strutturazione dei termini consente anche la produzione di repertori organizzati tematicamente sulla base di diversi livelli di classificazione.

Accanto a questi vantaggi, la mancanza di una sintassi nella stringa di soggetto in un'indicizzazione con i descrittori di un thesaurus è certamente un limite, che però può essere parzialmente corretto con strategie per indicare il grado di rilevanza dei descrittori in un index.

Ad. es. in un fascicolo intitolato “Fronte democratico popolare” con la seguente descrizione del contenuto: “Schema di comizio, appunti sulla campagna elettorale; testo della conferenza radio; corrispondenza relativa a comizi, all'organizzazione della campagna elettorale; opuscoli sulle ingiustizie

¹³ Cfr. <http://www.bncf.firenze.sbn.it/progetti/Nuovo%20Soggettario/> (controllato il 15 gennaio 2007).

sociali *Documenti presentati alla 1. assemblea nazionale (feb. 1948), Le terre in Italia sono divise così [...]*”.

Nel campo *Descrittori* potrebbe essere inserito: *elezioni politiche; campagna elettorale; lotte sociali; questione agraria*.

Dall'analisi dei descrittori però non è possibile rilevare il fatto che i primi due rappresentano l'argomento predominante della documentazione, mentre il terzo e il quarto coprono argomenti secondari.

Naturalmente il problema si risolve nel momento in cui si accede alla descrizione archivistica, che rende esplicito il contesto nel quale i descrittori utilizzati per la ricerca sono stati adottati in fase di indicizzazione.

Dal punto di vista informatico, tale limite potrebbe essere superato creando un sistema di descrittori maggiori (focus) e minori (contesto o soggetti secondari). Il descrittore riferito al focus del soggetto potrebbe essere indicato con un asterisco * (in riferimento all'es. precedente: *elezioni politiche**; *lotte sociali*) mentre per l'inserimento potrebbe essere introdotti campi diversi per descrittori maggiori e minori. Per non appesantire eccessivamente la consultazione, la scheda in modalità visualizzazione potrebbe presentare i descrittori in un unico campo in cui il focus è distinto semplicemente dall'asterisco.

Un descrittore che indica il concetto focus in una scheda potrebbe coprire un soggetto secondario in un'altra. In fase di ricerca, si potrà chiedere di consultare solo le schede in cui un determinato descrittore è presente come descrittore “maggior” - che rappresenta l'argomento principale della documentazione - oppure estendere il campo anche alle unità archivistiche che lo contengono come descrittore minore.

5. Specificità dei descrittori e livello archivistico.

Da più parti è stata sollevata la questione del rapporto tra specificità dei descrittori e livello archivistico. Apparentemente sembra naturale, e anzi quasi scontato, che la schedatura di “oggetti” archivistici diversi tra loro (fondi, serie, fascicoli, singoli documenti) conduca a un diverso grado di genericità del descrittore. Si è quindi posto il problema metodologico di stabilire se fosse necessario adottare criteri diversi per modulare l'uso dei descrittori a seconda che si indicizzi una scheda fondo, una scheda serie, ecc.

Come si è detto bisogna partire dal presupposto che l'indicizzazione tematica non sostituisce e non deve cercare di riprodurre la struttura archivistica, ma deve basarsi strettamente sulla descrizione: il livello di analicità dei descrittori sarà quindi strettamente

dipendente da quello delle schede. Ciò che deve essere omogeneo è il rapporto tra descrizione e descrittore e non quello tra descrittori e livelli archivistici.

Questo problema assume un notevole rilievo se si pensa di indicizzare insieme archivistici allargati e non omogenei riguardo al livello di descrizione, come è per Archivi del Novecento.

La prospettiva, infatti, cambia se si vuole indicizzare un singolo fondo, più fondi all'interno di un singolo istituto, o più fondi conservati in istituti diversi.

Il quadro generale in cui viene elaborato un progetto di indicizzazione ne condiziona direttamente alcuni criteri metodologici perché, se in un ambito limitato alcuni descrittori possono essere omessi in quanto validi a ogni livello, in un contesto come quello di una rete archivistica ogni scheda deve essere autonoma e raggiungibile comunque, perché indicizzata secondo criteri che potremmo definire "assoluti". Essa diviene pertanto un'"isola" nella lettura tematica, che si rapporta di volta in volta a un arcipelago di altre schede solo sulla base dei descrittori che le accomunano.

La base dati adottata come terreno di sperimentazione (www.archividelnovecento.it) è in gran parte costituita da inventari che descrivono fascicoli, pur presentando casi di descrizioni più generiche (livello fondo) e altre estremamente analitiche (livello documento). Ciò ha consentito di valutare anche le differenze di metodo e di misurarsi con la prospettiva di una indicizzazione tematica generica, nel caso di descrizioni archivistiche che si fermano a livello della scheda fondo, o spinta a livello del singolo documento.

Se l'inventariazione si ferma alla scheda del fascicolo, nel relativo campo *Contenuto* vi sarà una sintesi, operata dall'archivista, del contenuto complessivo del fascicolo. L'indicizzazione tematica sarà, come già detto più volte, basata su tale descrizione.

Se la schedatura è fatta a livello documentale (per ogni documento viene creata un'apposita scheda, che oltre a possedere campi identificativi come *Autore*, *Destinatario*, *Estremi cronologici*, avrà compilato anche un campo *Contenuto*), il contenuto della scheda del fascicolo che contiene i singoli documenti dovrebbe essere compilato solo genericamente, mentre l'indicazione dei temi specifici e quindi la relativa indicizzazione tematica deve "scendere" al livello più basso. Una ricerca condotta attraverso descrittori specifici punterà in questo caso direttamente alle schede relative ai singoli documenti; per vederne il contesto archivistico si ricorrerà ai mezzi tradizionali della gerarchia archivistica, ripetendo che questo non può e non deve essere il compito di un'indicizzazione tematica.

In sostanza il rapporto fra livello archivistico e specificità dei descrittori è un rapporto naturale perché scaturisce dalle informazioni presenti nelle rispettive schede e non richiede l'adozione di una metodologia o di criteri di analisi differenti.

6. Elementi da inserire o escludere dal campo descrittori (enti, antroponimi, toponimi, tipologie archivistiche).

Uno dei problemi generali individuati sulla base delle esperienze esaminate (cfr. nota 12) è quello della scelta delle informazioni da includere negli indici tematici. Alcuni progetti, infatti, accolgono al loro interno non solo descrittori di argomento ma anche voci relative ai nomi di persona, enti, luoghi e tipologie documentarie.

Nell'elaborazione dei criteri metodologici di questo progetto sono state prese in considerazione due possibilità:

- escludere dall'indice dei descrittori e quindi dal thesaurus queste informazioni in considerazione dell'esistenza nel software utilizzato dalla rete di Archivi del Novecento di tre campi separati predisposti per accogliere tali voci di indice (campi *Antroponimi*, *Toponimi*, *Enti*);
- inserire le voci relative ai nomi di persone, luoghi ed enti di cui si parli nei documenti in un unico campo, collegato a quello dei descrittori tematici, che raccolga quelli che potrebbero essere definiti *identificatori*, ferma restando la permanenza di tutti i nomi propri nei campi deputati del record documentario. Questa seconda soluzione avrebbe il vantaggio di distinguere le voci di indice quando sono rappresentative dell'argomento della scheda archivistica indicizzata, dai nomi di persona, enti e luoghi che compaiono ad altro titolo nella descrizione (corrispondenti, autori, destinatari, ecc.).¹⁴

Per es., se la descrizione del contenuto di un'unità archivistica dovesse essere:

“corrispondenza con X relativa alla promozione di Y”, X sarebbe indicizzato nel campo *Antroponimi*, Y verrebbe invece riportato sia nel campo *Antroponimi* sia nel campo *Identificatori*, in quanto argomento della corrispondenza.

Compatibilmente con lo sviluppo di un software adeguato, questa seconda ipotesi appare la migliore. Consente, infatti di individuare immediatamente se una persona o un ente sia **autore** o **oggetto** della documentazione. La ripetizione dei nomi propri 'oggetto' nell'area dedicata all'indicizzazione semantica, lungi dall'essere un'inutile ridondanza, consente di

¹⁴ Per questo argomento, cfr. BUREAU OF CANADIAN ARCHIVISTS, *Subject indexing for archives. The report of the Subject Indexing Working Group*, Ottawa, Bureau of Canadian Archivists, 1992, pp. 30 e sgg.

mantenere ai campi *Enti, Antroponimi e Toponimi* il loro scopo originario di riportare tutte le informazioni presenti nel corpo della scheda (a qualsiasi titolo esse vi si trovino).

Alcuni progetti hanno proposto di includere le tipologie documentarie tra i descrittori (*circolari, lettere, volantini, ecc.*), coerentemente con una prassi della soggettazione bibliotecaria che ammette l'inserimento della tipologia del documento nella stringa di soggetto. Sembra più adeguato escludere queste informazioni dall'indice generale, dato che esse sono parte integrante della descrizione archivistica e pertanto facilmente recuperabili con un'eventuale ricerca compiuta in altri campi della descrizione.

7. I criteri per l'elaborazione dei descrittori: 1. Il rapporto con la natura delle basi dati. *La core-area.*

I contesti archivistici si differenziano anzitutto sulla base della loro natura giuridica e particolarmente in riferimento alla differenza fra istituzioni di natura pubblica o privata. Gli archivi pubblici (e in parte anche alcuni tipi di privati) hanno una organizzazione burocratica che fissa organigrammi, competenze e procedure amministrative dei soggetti che producono documentazione. Sebbene la strutturazione degli archivi privati presenti a volte un'ampia analogia con le strutture organizzative di enti strutturati, il riflettersi ad esempio delle competenze degli organismi burocratici di partito sull'impianto e sul contenuto degli archivi appare più labile. Ciò significa che in questo contesto non ha molto senso ragionare partendo dal rapporto fra competenze degli uffici e contenuto della documentazione, osservazione che invece ha stimolato riflessioni all'interno di archivi di istituzioni pubbliche. Esiste infatti una tendenza ad utilizzare come base per l'indicizzazione tematica il titolare che, facendo riferimento alle funzioni degli uffici piuttosto che alla loro organizzazione, fornisce una griglia tematica in relazione alle competenze dell'ente.

Gli schemi di classificazione possono essere considerati quasi alla stregua di vocabolari controllati "delle funzioni" di un ente; la sinergia fra i due strumenti (titolario e vocabolario controllato) è infatti alla base dei più evoluti sistemi di ricerca inclusi nei progetti di informatizzazione degli archivi correnti. Essi però non sono assimilabili ai thesauri se non per l'esistenza di relazioni fra le voci. Inoltre la loro validità è limitata ad uno specifico archivio e non applicabile a una base dati come quella di Archivi del Novecento.

In generale la natura del contesto archivistico appare determinante e condiziona direttamente alcune scelte di metodo. Il progetto "Archivi del Novecento – la memoria in rete", coinvolge una rete di istituti culturali che conservano importanti fondi archivistici.

Nel suo insieme la rete ha come obiettivo l'individuazione e la valorizzazione delle fonti per la storia italiana del Novecento. La base dati descrive gran parte del patrimonio archivistico degli istituti aderenti, costituito per lo più da fondi personali di esponenti politici, da carte di partiti e movimenti politici e sindacali e di personalità di grande rilievo del mondo della cultura novecentesca. Accanto ad essi sono anche presenti nuclei documentari rappresentativi di altre aree (editoriale, scientifica, artistica).

Attualmente, il settore tematico prevalente all'interno della base dati archivistica di Archivi del Novecento è quello legato alla storia politica di partiti e movimenti.

Considerata la natura dei fondi archivistici, il loro contesto politico istituzionale e l'arco cronologico delle carte conservate, la *core area* del thesaurus non può che fare riferimento a un ambito tematico di carattere storico, politico, sociale e culturale del Novecento.

È fondamentale, nell'ambito di costruzione di un thesaurus, individuare un'area concettuale specifica, anche se non rigidamente delimitata, tenendo presente la distinzione terminologica tra campo concettuale, che “determina i significati dei termini che ad esso appartengono in modo univoco e su base extralinguistica”, e campo semantico, che invece “copre tutti i possibili significati di una parola polisema nei più vari contesti concreti e metaforici”¹⁵. Un descrittore non copre tutti i campi semantici che a un termine fanno riferimento: è proprio l'univocità di senso dei descrittori una delle caratteristiche necessarie dei thesauri. Tale univocità è data dalla scelta di un campo concettuale delimitato che permette di eliminare potenziali ambiguità nell'interpretare il significato di un termine.

Ad es. al termine *picchettaggio*, secondo il vocabolario Treccani, possono essere attribuiti due significati: “operazione di rimescolamento della massa del carbone nei gassogeni a griglia fissa ...” e “funzione di vigilanza e controllo esercitata durante gli scioperi da gruppi di lavoratori o da rappresentanti sindacali ...”. All'interno della *core area* storico-politica della base dati di Archivi del Novecento, a qualsiasi utente che troverà il descrittore *picchettaggio* verrà naturale attribuire la seconda definizione; se ci si trovasse invece in una *core area* di tipo scientifico sarebbe naturale pensare al primo significato; in un thesaurus che volesse abbracciare l'intero scibile sarebbe invece indispensabile specificare in quale delle due accezioni si sta utilizzando tale termine, generando però il rischio di ambiguità.

¹⁵ MARISA TRIGARI, *Come costruire un thesaurus*, Modena, Panini, 1992, p. 23.

Selezionare la propria terminologia all'interno di un campo concettuale diventa quindi essenziale. Al contrario, se si cercasse di inglobare in un thesaurus ogni campo concettuale, il risultato condurrebbe inevitabilmente ad ambiguità terminologiche e associative in rapporto ai contesti oltre che ad una rapida obsolescenza.

Ad es. il descrittore *scissione* inserito nel contesto dei partiti assume il significato della “divisione di un gruppo conseguente a un contrasto politico”; se ci fosse anche il contesto scientifico o naturale sarebbe necessaria una specificazione, anche delle sue eventuali relazioni.

Si è ritenuto necessario, dopo aver individuato una o più aree centrali e delle aree marginali, procedere a un'organizzazione del thesaurus (anche raccomandato dallo Standard Iso-2788/1986) con il conseguente sviluppo di micro-thesauri come sottosettori specialistici (vedi oltre, paragrafo 11). Si avranno quindi microthesauri legati alla *core area* (*Lavoro e occupazione; Stato e pubblica amministrazione; Politica; Gruppi, movimenti, organizzazioni; Politica internazionale; Individuo e società*) e microthesauri delle aree minori, ma dotati di una coerenza interna, e che permettono eventuali allargamenti dei contesti semantici di riferimento (*Cultura, arte; Diritto, legislazione, giustizia; Economia; Ambiente fisico e infrastrutture; Educazione, informazione e comunicazione; Scienze e tecnologia*).

8. I criteri per l'elaborazione dei descrittori: 2. Il rapporto con il tipo di utenza.

Nella metodologia di costruzione del thesaurus è necessario tener conto del tipo di utenza di riferimento. Negli istituti culturali, infatti, è prevalente un accesso alla documentazione che si caratterizza come utenza di tipo specialistico. Diverso è il caso di una consultazione tramite web, in cui l'accesso può avvenire da parte dei soggetti più vari; ci si trova quindi di fronte a una utenza difficilmente prevedibile. È evidente che il tipo di utenza condiziona direttamente alcuni aspetti dell'indicizzazione, soprattutto per ciò che attiene alla scelta del registro linguistico da adottare per i termini preferiti: specialistico, attualizzato ecc.

In fase di elaborazione e soprattutto in fase di controllo terminologico dei descrittori, si è presentato il problema di gestire termini del gergo politico, spesso riferiti ad un determinato periodo storico e termini di lingua speciale, molto frequenti in una base dati come quella di Archivi del Novecento.

Ad es. *diffusori* o *diffonditrici* (coloro che si occupavano, tra la fine degli anni Quaranta e i primi anni Sessanta rispettivamente, della diffusione della stampa di partito o di associazioni femminili), *giunte anomale, doroteismo, scala mobile* ecc.,

sono tutti concetti espressi secondo una terminologia propria dell'ambito politico.

All'interno di una lista di descrittori la questione fondamentale è consentire un accesso anche a questa tipologia di espressioni. Sia che si scelga il termine gergale o di lingua speciale, sia che si preferisca il termine equivalente nel linguaggio comune, è opportuno valersi delle relazioni di equivalenza (*Use* e *Use for* – vedi oltre) che permettono generare e collegare tra loro i cosiddetti *termini preferiti* con quelli *non preferiti*. Diventa quindi fondamentale nella fase di indicizzazione individuare dei criteri di preferenza tra sinonimi, ovvero i termini da scegliere come preferiti e quelli da relegare al ruolo di non descrittori.

In linea di massima, se si intende favorire una utenza generica, si tenderà ad aggiornare i termini “storici” o datati e a trasformare in termini di linguaggio comune quelli di lingua speciale.

Ad es. il termine molto usato negli anni Cinquanta *ricreazione* sarà un non preferito e si avrà quindi: *ricreazione* USE *tempo libero*.

Si è ritenuto però essenziale nel nostro caso usare con discrezione tale orientamento, sia per la natura dell'utenza potenziale, che si presume esperta e relativamente specializzata, sia per “salvaguardare” e valorizzare la tipicità di una banca dati come quella di Archivi del Novecento. La scelta di utilizzare con larghezza i termini storici o specialistici come termini preferiti ha tenuto anche conto della frequenza dell'uso degli stessi termini nella fase di indicizzazione libera o in fase di ricerca nell'ambito degli archivi oggetto della sperimentazione. Se un concetto si è rivelato essere identificato con maggiore frequenza con un dato termine (sia in fase di indicizzazione libera che di ricerca) si è optato per mantenere quest'ultimo come termine preferenziale, corredandolo con una **scope note** che servisse a precisare l'ambito semantico del termine. È infatti importante dare la preferenza a quei termini che hanno maggior probabilità di essere scelti dagli utenti.

Ad es., su queste basi, è stato inserito il termine *centrosinistra* come descrittore, corredato dalla *Scope note*: “Alleanza politica tra partiti di centro e di sinistra, nonché la maggioranza di governo negli anni Sessanta in Italia”; è stato conservato come non-descrittore *apertura a sinistra* (*apertura a sinistra* Use *centrosinistra*); meno usato, ma utilizzato comunque come punto di accesso.

Il gergo politico pone ulteriori problemi. Molti termini gergali infatti sono doppiamente ambigui, perché assumono nel linguaggio della politica significati diversi da quelli originari,

che persistono nell'uso comune. Si tenderà a non sceglierli come descrittori se non quando il significato "gergale" si è sovrapposto a quello primitivo quasi obliterandolo.

Ad es. la locuzione *franchi tiratori* verrà utilizzata per identificare i deputati della maggioranza che votano in modo contrario alla linea del governo, mentre l'originario significato darà luogo al rimando *franchi tiratori (militare)* USE *cecchini*.

In linea di massima si è scelto di abbondare nell'uso dei **non descrittori**, soprattutto nell'area concettuale principale, al fine di offrire un'ampia apertura verso il linguaggio libero dell'utente e quindi maggiori punti di accesso in fase di ricerca (tenendo presente un limite costituito dalla necessità di non appesantire eccessivamente il thesaurus, la cui efficacia è legata anche alla maneggevolezza). Si è cercato quindi di non tralasciare quei termini che non possono essere utilizzati come **termini preferiti**, in quanto troppo connotati "gergalmente", ma che presumibilmente hanno ampie probabilità di essere ricercati dagli utenti.

Es. *maggio francese* (non descrittore)

USE contestazione + movimento studentesco

oppure *politica dell'albero* (non descrittore)

USE forestazione

Un'ulteriore scelta metodologica si è resa necessaria relativamente alla possibilità di esprimere e rappresentare attraverso descrittori alcuni eventi storici. Questa possibilità è stata ammessa relativamente ad eventi di particolare rilevanza nell'ambito della base dati e universalmente noti con una certa denominazione, tanto da poter essere considerati termini sincategorematici, secondo l'accezione dello Standard ISO 2788, che ne esclude la scomposizione.

Sono stati dunque introdotti termini quali *Guerra mondiale (1939-1945)*, *marcia su Roma (1922)*, *governo di unità nazionale (1978)*, evitando la necessità di macchinose e poco significative post-coordinazioni di termini generici in sede di indicizzazione e ricerca.

L'ambito documentario di riferimento ha anche determinato il criterio con cui selezionare gli eventi: solo quelli effettivamente presenti nei documenti sono stati presi in considerazione, senza una copertura in astratto di tutti gli eventi che si potrebbe presumere di incontrare.

9. Metodo di sviluppo del thesaurus.

La metodologia di sviluppo del thesaurus adottata è stata di tipo induttivo. Ciò significa che si è data priorità alla pratica documentaria: i termini sono stati inizialmente raccolti attraverso un'indicizzazione libera di un campione significativo di descrizioni archivistiche tratti dalla base dati di riferimento (www.archividelnovecento.it). Sulla base di una lista provvisoria di termini così costituita, in una seconda fase si è fatto riferimento più in astratto all'ambito concettuale della *core area* individuata, dando una strutturazione ai termini che superasse la pura empiricità. È infatti apparso necessario unire le due modalità di sviluppo (quella empirica e quella astratta): un thesaurus costruito esclusivamente con metodo induttivo avrebbe presentato difficoltà nella strutturazione dei termini, lacune terminologiche ed eccessivi squilibri.

Ad es. con il metodo induttivo erano stati inseriti descrittori quali *criminalità organizzata*, *mafia*, *terrorismo*, tratti direttamente dall'indicizzazione delle descrizioni archivistiche. Si è ritenuto necessario completare il quadro inserendo in astratto altri descrittori, come *criminalità* e *camorra*, che potevano costituire utili termini di testa ai fini della ricerca, nel caso di *criminalità*, o ritenuti utili in prospettiva, nel caso di *camorra*.

Questa metodologia si è rivelata utile soprattutto per definire le relazioni gerarchiche tra termini; solo dopo l'intervento "in astratto" si è potuto avere un reticolo di relazioni omogeneo:

Es:

criminalità

.. NT1 criminalità organizzata

... NT2 camorra

... NT2 mafia

.. NT1 terrorismo.

Una volta creata una lista considerata significativa di descrittori (circa 400 termini), si è passati alla fase del controllo volta ad assicurare al linguaggio di indicizzazione quella regolarità e univocità semantica.

10. Criteri per il controllo numerico, morfologico (relativo a categorie grammaticali nella forma del nome) e semantico (relativo all'univocità dei termini); la relazione di equivalenza.

Il controllo numerico.

Relativamente al controllo numerico del thesaurus è stato necessario trovare un giusto equilibrio nell'utilizzo della **modularità**, ovvero nel maggiore o minor tasso di scomposizione di descrittori composti.

Di norma nella formulazione dei descrittori è preferibile un livello alto di scomposizione o **tasso di modularità**, coerentemente con la logica di post-coordinazione nella ricerca connessa all'uso di un thesaurus.

Ad es.:

ricercatori universitari + energia nucleare e non ricercatori nucleari

oppure

monopolio + energia elettrica e non monopoli elettrici

In qualche caso tuttavia questa regola generale appare difficilmente adottabile nell'ambito di questa base dati nella formulazione di descrittori riferiti a concetti largamente identificati e conosciuti con sintagmi nominali, anche estesi, che è necessario perciò mantenere inalterati (es.: *Rivoluzione d'ottobre, festa della donna, paesi in via di sviluppo* ecc.). L'altro limite a questa regola generale è costituito dalla necessità di non spezzare alcuni concetti al fine di evitare un eccessivo "rumore" in fase di ricerca, in special modo nell'area nucleo della base dati da indicizzare. In linea di massima, si è scelto di adottare descrittori composti sostanzialmente nella *core area*, mentre è stato aumentato il grado di modularità (scomposizione) nelle aree periferiche.

Con questa eccezione, si è rispettato il criterio di un buon tasso di modularità cercando di equilibrare i vantaggi di scarso rumore in fase di risposta, propri di una bassa modularità, con i vantaggi di un *corpus* del vocabolario sufficientemente agile e flessibile, propri di un'alta modularità.

Ad es. se si utilizzasse una bassa modularità per esprimere il concetto di riforme elettorali, costituzionali o di patti agrari si avrebbero tre descrittori: *riforma elettorale, riforma costituzionale, riforma dei patti agrari*, a cui naturalmente si andrebbero ad aggiungere i descrittori "semplici" *sistema elettorale, costituzioni, patti agrari*: per un totale di sei descrittori. Con un tasso di modularità più alto si avrà invece *sistema elettorale + riforme, costituzioni + riforme, patti agrari + riforme*:

per un totale di quattro descrittori. Considerate tutte le possibilità di combinazioni tra termini, appare chiaro quanto questa scelta possa incidere sul numero complessivo dei descrittori presenti nel thesaurus.

Mantenere un numero controllato di termini è importante per ottenere un thesaurus gestibile sia dal punto di vista della consultazione, sia da quello della implementazione dei termini.

Il controllo morfologico.

Il controllo morfologico è una delle funzioni fondamentali per assicurare una corretta gestione del linguaggio di indicizzazione. Assolve infatti il compito di garantire l'omogeneità dei termini dal punto di vista formale.

Le aree in cui trova applicazione il controllo sulla forma dei termini sono quella del numero e del genere, dei termini composti e dell'omografia.

Per conseguire la regolarità morfologica nella scelta del numero e del genere dei nomi ci si è orientati verso le soluzioni comunemente adottate nei thesauri di area anglosassone, riprese nell'esperienza italiana dalle regole del Gris¹⁶. In sostanza, il principio di base per la scelta della forma singolare o plurale è quello della numerabilità: si adotta il plurale per i concetti numerabili; il singolare per concetti astratti.

Per es.:

operai e **non operaio** (categoria numerabile di persone);

consigli di fabbrica e non **consiglio di fabbrica** (categoria numerabile di strutture);

totalitarismo e **non totalitarismi** (categoria singolare di credenze e sistemi ideologici);

ricerca scientifica e **non ricerche scientifiche** (categoria singolare di attività, discipline, processi).

Si è però deciso di non adottare rigidamente questa regola, ma di adeguarla alla recente tradizione italiana di thesauri, dizionari ed enciclopedie, orientata verso un uso più esteso del singolare.

Rientra nell'ambito del controllo morfologico anche la scelta delle forme grammaticali. Coerentemente con le raccomandazioni degli Standard Iso, sono stati preferiti quanto più

¹⁶ GRUPPO DI RICERCA DI INDICIZZAZIONE PER SOGGETTO - GRIS, *Guida all'indicizzazione per soggetto*, Roma, Aib, 1996; il testo della normativa GRIS è disponibile anche alla pagina www.aib.it/aib/gris.htm (controllato il 26 maggio 2005).

possibile sostantivi e sintagmi nominali, escludendo verbi all'infinito e avverbi; si è cercato di limitare al massimo l'uso di aggettivi isolati: l'aggettivo di norma è preordinato con un sostantivo in un termine composto.

Ad es. si utilizzeranno i termini *lavoratori atipici* e *lavoratori stagionali* e non il termine *lavoratori* combinato con due termini autonomi come *atipici* e *stagionali*.

Una scelta metodologica va applicata alle varianti di lingua: nel caso di termini presi in prestito da altre lingue, si è preferito adottare la forma italiana, utilizzando come *non descrittore* la forma straniera. Fanno eccezione i casi in cui la forma italiana sarebbe risultata non corrispondente all'uso prevalente.

Ad es., nel primo caso, *devolution* rinvierà a *federalismo*; nel secondo *elaboratore elettronico* rinvierà a *computer*.

Il controllo semantico.

Una caratteristica importante del thesaurus consiste nella univocità semantica dei descrittori, cioè nella rappresentazione di ogni concetto con un solo termine, salva restando la possibilità di mantenere come chiavi di accesso sinonimi o quasi sinonimi in qualità di termini non preferiti.

Il controllo semantico è dunque indirizzato principalmente alla selezione di un solo significante, il più chiaro possibile, per uno stesso concetto. Ciò comporterà la scelta del termine meno ambiguo tra vari sinonimi possibili, una gestione accurata dei non descrittori, un uso delle *scope note* per la precisazione dell'ambito semantico in caso di ambiguità, ma soprattutto una adeguata gerarchizzazione che collochi il termine tra i suoi sovraordinati ed i suoi sottordinati. La struttura di per se stessa è il più importante strumento per la disambiguazione del significato del termine.

L'utilizzo del termine preferenziale e il suo collegamento al termine non preferito si effettua attraverso la relazione di sinonimia tra termini, espressa dalle sigle USE (usa) e UF (usato per).

Ad es.:

- 1) *lager* Use *campo di concentramento* (relazione di sinonimia);
- 2) e nel caso di quasi sinonimi *stipendio* Use *salario* (relazione di quasi-sinonimia).

Le relazioni di preferenza è anche stata adottata nel caso di varianti grafiche dello stesso termine (*filosofia medievale* e *filosofia medioevale*); mentre si è scelto di non utilizzarla nel caso di termini che possano prevedere l'uso di trattini: in questo caso è stata generalmente adottata la forma priva di trattino (*centrosinistra* e non *centro-sinistra*, *controcultura* e non *contro-cultura*), più semplice ed immediatamente identificabile anche senza rinvio.

Di fronte a termini che coesistono in aree concettuali diverse è importante - per garantire l'univocità - la scelta del termine, che privilegerà ancora una volta la maggiore frequenza d'uso nei rispettivi contesti e sarà accompagnata da un uso delle *scope note* per chiarire ulteriormente la scelta.

Termini polisemi e omografi possono essere disambiguati anche con l'uso di opportune qualificazioni che seguono il descrittore, denominate "qualificatori" e consistenti in una breve definizione inserita tra parentesi uncinate.

Ad es. si utilizzerà il descrittore *franchi tiratori* per intendere i rappresentanti di uno schieramento politico che, in votazioni segrete, votano in modo diverso da quanto concordato ufficialmente dal partito di appartenenza; il termine *franchi tiratori (militare)*, con un qualificatore, sarà introdotto per indicare il guerrigliero che opera contro forze regolari.

In realtà, si è cercato di limitare l'uso di qualificatori per precisare il significato di termini polisemi, in quanto pesante e poco naturale, preferendo un uso esteso delle *scope note* (o note d'ambito), modificando – ove possibile – la forma del nome e contando in ogni caso sul valore autoesplicativo della struttura gerarchica.

Ad es. si è scelto di utilizzare il descrittore *colonie*, con la *scope note* "Territorio dominato da uno Stato estero e abitato da popolazioni indigene le quali non godono degli stessi diritti civili dei gruppi etnici provenienti dallo Stato dominante"; mentre per indicare gli istituti benefici che procurano ai bambini di famiglie meno abbienti un soggiorno estivo, si è scelto il sintagma *colonie estive*. In tutti e due i casi la collocazione dei termini nella struttura è di per se stessa semanticamente significativa.

Le **note d'ambito** (SN) sono state utilizzate con diverse funzioni, sempre nel rispetto degli Standard Iso:

- nel caso di termini polisemi per chiarire in quale accezione un termine sia stato impiegato;

caduti

MT 07. Politica internazionale

SN: “Chi è morto in combattimento. Per estensione, chi rimane vittima in un conflitto, in una lotta (anche ideale), o cade nell'adempimento del proprio dovere”¹⁷.

- nel caso di termini specialistici o gergali che si è ritenuto necessario adottare in quanto fortemente ricorrenti per definirne il significato;

centrosinistra (1963-)

MT 05. Politica

SN: Alleanza politica tra partiti di centro e di sinistra, nonché la maggioranza di governo negli anni Sessanta in Italia.

- per dare una definizione del termine se esso ha avuto oscillazioni di significato nel tempo;

sicurezza sociale

MT 08. Individuo e società

SN: Complesso di interventi pubblici volte all'erogazione di beni e servizi ai cittadini.

BT politica sociale

NT assistenza

NT previdenza sociale

- per delimitare l'ambito d'uso di un termine 'storico';

alleati

MT 07. Politica internazionale

SN: Alleanza militare che si oppose all'Asse durante la Seconda Guerra Mondiale.

11. Struttura semantica e criteri per la costruzione delle relazioni gerarchiche (genere/specie, parte/tutto) e associative.

Si è detto che i descrittori, oltre ad essere sottoposti alle norme relative al controllo numerico, morfologico e semantico, sono inquadrati all'interno di una struttura gerarchica e associativa di per se stessa semanticamente autoesplicativa. Tale struttura è una potente

¹⁷ Cfr. *Il Vocabolario Treccani*, a cura dell'ISTITUTO DELLA ENCICLOPEDIA ITALIANA FONDATA DA G. TRECCANI, Roma, 1997.

guida alla ricerca del termine più adatto alla rappresentazione dei contenuti dei documenti, sia in fase di indicizzazione, sia in fase di ricerca.

Il thesaurus è definito dalle norme ISO come “vocabolario di un ‘linguaggio di indicizzazione’ controllato, organizzato in maniera formale, in maniera cioè da rendere esplicite le relazioni ‘a priori’ fra i concetti”¹⁸.

La relazione gerarchica lega ciascun descrittore a un termine di significato più generale e ad uno o più termini più specifici nell’ambito di una stessa categoria. La presenza di tali relazioni è – come si è già detto – di per se stessa una definizione del significato del termine e rende il thesaurus capace di migliorare l’indicizzazione e la ricerca, suggerendo termini alternativi o aggiuntivi.

Nella definizione delle relazioni gerarchiche si distinguono due tipologie di relazioni possibili: genere/specie e parte/tutto. Nel primo caso il termine sott’ordinato è un elemento della classe rappresentata del termine sovraordinato.

Ad es.

movimenti

.. NT1 movimenti femminili

.. NT1 movimenti religiosi

oppure

elezioni

.. NT1 elezioni amministrative

... NT2 elezioni regionali

Nelle relazioni di tipo parte/tutto invece il termine sott’ordinato rappresenta una parte di ciò che è espresso dal termine sovra ordinato.

Ad es.

forze armate

.. NT1 aeronautica militare

.. NT1 esercito

processi

.. NT1 dibattimenti

¹⁸ *Linee guida per la costruzione e lo sviluppo di thesauri monolingue : versione in lingua italiana della Norma ISO 2788*, Milano, UNI, 1993.

- . . NT1 inchieste giudiziarie
- . . NT2 indagini

Altro tipo di relazione è quella associativa: è usata per unire quei descrittori che, pur non legati da una relazione gerarchica o di equivalenza, coprono concetti fortemente correlati. La relazione, che potremmo anche definire ‘di implicazione’, è reciproca e si riferisce a concetti che hanno alta probabilità di ricorrere insieme nello stesso contesto. Essa suggerisce all’indicizzatore o al ricercatore punti di vista diversi o soggetti accessori a cui potrebbe non aver pensato.

A differenza delle relazioni gerarchiche, quella associativa ammette un certo grado di discrezionalità. Di norma si tende a non correlare termini appartenenti alla stessa catena gerarchica, vale a dire alla stessa categoria logica, per non appesantire troppo la struttura del thesaurus, ipotizzandosi che relazioni associative in quest’ambito ristretto siano agevolmente ricostruibili dall’utente finale in una presentazione sistematica dei termini.

Ad esempio non sarebbe raccomandabile la seguente relazione associativa:

antisemitismo RT *razzismo*

in quanto i due descrittori sono affiancati nella stessa gerarchia sotto il top term *ideologia* e la relazione tra loro è trasparente per l’utente che li incontra nello stesso contesto.

Diverso è il caso di:

movimenti

. . NT1 movimenti pacifisti

. . . . RT non violenza

in questo caso il descrittore *movimenti pacifisti* ha come top term *movimenti*, mentre il descrittore *non violenza*, ad esso associato, ha come top term *comportamento politico*.

Nella costruzione del thesaurus si è cercato di rispettare la correttezza formale delle relazioni gerarchiche, senza sostituirle, come spesso accade, con relazioni di tipo contenutistico, che possono risultare più intuitive, ma sono formalmente scorrette.

Il rispetto della correttezza formale rischiava però di rendere la presentazione **sistematica** del thesaurus molto dispersiva, moltiplicando i termini ‘isolati’ nella classificazione, quelli cioè che non potevano contare su di una organizzazione gerarchica di sotto- e sovraordinati.

Questo problema è molto frequente in thesauri con un numero relativamente ristretto di termini e nel nostro caso si faceva sentire maggiormente nella *core area* relativa al tema politico e in particolare alle organizzazioni politiche e sindacali.

Una soluzione originale, che ha consentito di ovviare ad una presentazione sistematica poco compatta e dispersiva, non amichevole nei confronti dell'utente finale, è offerta dal software usato per la costruzione del thesaurus: nel caso di termini 'isolati', legati ad altri termini del thesaurus dalla sola relazione associativa, il software consente di collocare il termine nello schema classificatorio in relazione agli altri termini, senza però presentarlo come isolato.

Ad es. sotto il top term *partiti* erano stati inizialmente collocati come NT *partiti di sinistra* e *militanti*. In realtà solo il primo degli esempi riportati era formalmente corretto (i partiti di sinistra sono una "specie" di partiti e dunque la relazione 'genere-specie' era rispettata), mentre i *militanti*, che non configurano né una specie, né una parte dei partiti, erano stati considerati NT al solo scopo di associare i due termini evitando una comparsa isolata di *militanti*, non appartenente ad alcuna altra gerarchia semantica.

La soluzione adottata è quella utilizzare l'*escamotage* consentito dal software TMS dell'Indire, utilizzato per la costruzione del thesaurus. Il software TMS consente, all'atto dell'input, di correlare i termini privi di relazioni gerarchiche, ma dotati di relazioni associative, con un tipo speciale di relazione associativa, definita RBT (related broader term) e RNT (related narrower term). In uscita il termine figurerà correttamente correlato in **relazione associativa** con i termini di riferimento, comparirà regolarmente nella presentazione alfabetica con tutti i suoi RT, ma non figurerà come **isolato** nella presentazione gerarchica. Ciò consente di rispettare il rigore delle relazioni semantiche e insieme la amichevolezza della presentazione.

12. Criteri per la classificazione dei descrittori nella presentazione sistematica.

Accanto alla strutturazione semantica costituita dalle relazioni fra i termini (BT, NT, RT) è possibile inserire le voci dell'indice all'interno di una struttura di livello alto organizzata in classi. È necessario tener presente che il fine di questo raggruppamento in classi è quello di un ulteriore aiuto nella ricerca dei descrittori più adeguati per l'indicizzazione e il reperimento dell'informazione.

Si ricorda sinteticamente che la classificazione dei descrittori di un thesaurus può essere di tipo tematico-disciplinare o per faccette. La prima risponde a un principio di suddivisione sostanzialmente storico-culturale: un tema o una disciplina viene diviso in sotto-temi o sotto-discipline che ne rappresentino una specificazione in senso contenutistico. La seconda risponde a un principio di divisione logico e astratto, valido indipendentemente dalle aree concettuali di riferimento, perché rispondente a una divisione per categorie universali (es. *atti, strutture, azioni, processi*, ecc.).

Per l'organizzazione del thesaurus si è preferito adottare una classificazione di tipo tematico-disciplinare che risulta più intuitiva rispetto alla divisione per faccette, pur essendo meno rigorosa formalmente e soggetta a un certo livello di "arbitrarietà".

Una volta identificato il campo concettuale di riferimento nelle sue aree centrali e marginali, è stato necessario individuare dei criteri di massima per la suddivisione delle categorie: sono state ipotizzate alcune aree-guida, in cui far confluire i descrittori, come "politica", "diritto", "amministrazione pubblica", "economia", "cultura", "occupazione e lavoro", ecc. Tale macroclassificazione ha permesso di avere una visione di insieme, utile a valutare lacune, ridondanze ed equilibri nell'uso dei termini. Naturalmente questa classificazione immaginata all'inizio del lavoro si presentava elastica e flessibile, in grado cioè di adattarsi alle spinte provenienti dei termini concretamente utilizzati per l'indicizzazione. Man mano che procedeva il lavoro di indicizzazione, di implementazione della lista di descrittori, della loro classificazione, la struttura macroclassificatoria tendeva ad assumere dei confini più stabili; si è passati quindi alla costituzione di "microthesauri".

Rispetto alle scelte metodologiche adottate durante la fase *in progress* ci sono state delle variazioni determinate dalle occorrenze di alcuni termini o dalla dimensione assunta da alcune aree, che hanno portato alla revisione dei criteri iniziali di classificazione¹⁹.

Al termine della sperimentazione si è giunti alla creazione di 14 Microthesauri²⁰, alcuni componenti la core area del thesaurus (01. Stato e pubblica amministrazione; 02. Diritto, legislazione, giustizia; 03. Economia; 04. Lavoro e occupazione; 05. Politica; 06. Gruppi, movimenti, organizzazioni; 07. Politica internazionale; 08. Individuo e società); altri relativi

¹⁹ Inizialmente il lavoro è stato basato su un metodo di tipo sistematico grafico, utilizzando un database in ambiente Access che permettesse per ogni descrittore di avere una visione di "quadro", articolata graficamente. Quando il numero dei descrittori ha raggiunto una certa consistenza si è abbandonato questo modello, di difficile gestione per un numero di termini elevato, passando a un'impostazione alfabetica, dove le relazioni sono date dalle sigle e codici. A questo scopo è stato utilizzando il software *TMS* dell'Indire fornito gratuitamente dall'Ente.

²⁰ Naturalmente tale strutturazione non è definitiva; è pensata per essere incrementabile in vista di eventuali ampliamenti dei campi concettuali rappresentati nella banca dati di Archivi del Novecento.

a campi concettuali più marginali, ma non meno essenziali, nell'economia della lista dei descrittori (09. Ambiente fisico e infrastrutture; 10. Cultura, arte; 11. Educazione, informazione e comunicazione; 12. Scienze e tecnologia; 13. Religione e chiese). A questi si aggiunge per ultimo il Microthesaurus 14. Mot outils, comprensivo di termini "di utilità" e generici (anche se si è cercato di limitarne l'uso).