Introduzione ai thesauri

S. Spinelli*

Definizioni degli standard:

Guidelines for the establishment and development of monolingual thesauri. ISO (International Organization for Standardization), 2788/1986; trad. it. UNI/ISO 2788: 1993. *Linee guida per la costruzione e lo sviluppo di thesauri monolingue*:

> il th è il vocabolario di un "linguaggio di indicizzazione" controllato, organizzato in maniera formale, in maniera cioè da rendere esplicite le relazioni "a priori" fra i concetti

Guidelines for thesaurus structure, construction and use. ANSI (American National Standard Institute), Z39.19-1974:

> il th è un insieme di parole e frasi che rappresentano relazioni di equivalenza e associative che forniscono un vocabolario standardizzato per sistemi di archiviazione e recupero dei documenti

Altre definizioni:

Roget's International Thesaurus, 1977 (1. ed. 1852):

> il th è uno strumento basilare per trasformare le idee in parole

F.W. Lancaster, *Vocabulary control for information retrieval*, 1972:

> il th è uno strumento di controllo terminologico nei sistemi postcoordinati

ISO 2788/1974:

> in termini di funzione è uno strumento terminologico usato per tradurre il linguaggio naturale dei documenti, degli indicizzatori o degli utenti in un linguaggio di sistema, più strutturato, detto anche linguaggio documentario o linguaggio di informazione. In termini di struttura il th è un vocabolario controllato e dinamico di termini semanticamente correlati che coprono un determinato ambito disciplinare

Commenti:

- l'ANSI non cita le relazioni gerarchiche, non chiarisce che le relazioni investono i concetti rappresentati dai termini e non i termini stessi, pone l'accento sullo scopo di archiviazione e recupero;
- il Roget, che non è un th nel senso tecnico da noi adoperato del termine, ma una sorta di suo 'antenato', un dizionario tematico di sinonimi, contrari, ecc., ad uso di giornalisti, pubblicisti, scrittori ecc., con una frase suggestiva pur se dal sapore vagamente pubblicitario (trasformare le idee in parole...), enuclea il problema principe non solo dei th ma di qualsiasi sistema di indicizzazione, cioè la rappresentazione dei concetti attraverso codici linguistici;

*Coordinatore dell'Area Biosfera del Sistema Bibliotecario d'Ateneo, Biblioteca Biomedica Centrale dell'Università di Bologna

- la definizione di Lancaster enuncia semplicemente scopo e ambito del th, ambedue in senso stretto; l'ambito dei sistemi postcoordinati, tuttavia, non è più considerato valido in quanto i th odierni possono anche costituire la parte lessicale di sistemi precoordinati, dotati di sintassi;
- l'ultima definizione, l'ISO del 1974, descrive il th in maniera un po' prolissa sia in base alla sua funzione che in base alla sua struttura, ed è legato al principio, in seguito variamente rivisitato e discusso, dell'ambito disciplinare.

Soffermiamoci ora sulla definizione ISO 2788 del 1986, quella cui d'ora in poi faremo costantemente riferimento, poiché la standard internazionale più recente:

> il th è il vocabolario di un "linguaggio di indicizzazione" controllato, organizzato in maniera formale, in maniera cioè da rendere esplicite le relazioni "a priori" fra i concetti

La norma circoscrive innanzi tutto il th alla sola parte lessicale (semantica) di un linguaggio d'indicizzazione e di ricerca, al quale, onde ottenere un codice documentario completo, va abbinato il *corpus* di norme (sintassi) che regolano i rapporti sintagmatici tra gli elementi di un enunciato di soggetto (che esprime quindi rapporti fra termini non aprioristici ma dipendenti dal documento, ad es. l'ordine di citazione dei termini, l'impiego di simboli esprimenti determinate relazioni, gli accorgimenti tipografici o la punteggiatura fra i termini, e così via). Questo "vocabolario" è **alfabetico** in quanto adopera termini che, benché sottoposti a **controllo**, appartengono al più vasto insieme della lingua naturale, a differenza degli schemi di classificazione, detti **artificiali puri**, nei quali i concetti sono rappresentati da notazioni numeriche o alfanumeriche non portatrici di significato se non all'interno del proprio stesso sistema.
Il concetto di **controllo** si traduce invece in un requisito fondamentale a garantire l'incontro fra lessico dell'indicizzatore e lessico del ricercatore, e cioè la **relazione biunivoca** fra termine e concetto, fra significante e significato: ciò significa che in un th un termine esprime sempre uno ed un solo concetto, e che un concetto è sempre espresso da uno ed un solo termine. Poichè questa condizione è tutt'altro che vera nel linguaggio naturale, per nostra fortuna affetto da ridondanze, ambiguità, polisemie, omonimie, omografie ed altre terribili disfunzioni che ne garantiscono ricchezza ed espressività, il raggiungimento del controllo, cioè dell'**univocità semantica,** nella costruzione e manutenzione del th comporta due tipi di accorgimenti:

1. la raccolta e la collazione dei sinonimi, dei quasi sinonimi (cioè termini non sinonimi in senso proprio, ma considerabili tali ai fini della rappresentazione dei concetti dell'ambito coperto dal th) e degli antonimi (cioè degli opposti, o meglio dei termini collocati in diversi punti dello stesso continuum semantico: non solo *calore* e *freddo*, ma anche *tepore, fresco* e così via) atti a descrivere il medesimo concetto e la scelta di uno solo di questi termini: il termine prescelto diventa nel vocabolario **termine preferito**, **TP** (o **PT**, **preferred term**), **descrittore**, cioè termine abilitato a descrivere quel determinato concetto; tutti gli altri termini (**termini non preferiti**, **NPT**, **non descrittori**), che non possono essere assegnati ai documenti per esprimerne il contenuto concettuale, possono però vantaggiosamente entrare nel vocabolario controllato come punti di accesso che rinviano al termine preferito;

2. la riduzione del contenuto semantico del termine preferito ad un solo significato, di solito il più tipico nell'ambito disciplinare del th, per cui ad esempio in un th di ornitologia il termine *gru* esprimerà il concetto corrispondente all'"uccello dei Ralliformi" e non alla "macchina dal braccio girevole per sollevare pesi", né al "carrello mobile per riprese cinematografiche".

La definizione dello standard ISO 2788 pone inoltre in rilievo i due elementi strutturali fondamentali del th:

1. che le relazioni da esso esplicitate sono *formalizzate*, cioè che le relazioni fra i termini sono definite da una relazione esplicita e formalizzata; questo significa che ogni termine è inserito in una rete relazionale che ne chiarisce ulteriormente il contenuto semantico, e che mostra la distanza semantica fra i termini stessi;

2. che le relazioni trattate sono *a priori*, cioè sono relazioni che pertengono all'ambito semantico, del significato, dei termini, e pertanto sono sempre valide in qualsiasi contesto.

Termini d'indicizzazione

I concetti rappresentati dai termini di un th possono appartenere a diverse categorie:

1. **entità concrete**
   (a) oggetti e loro parti fisiche
   (b) materiali
2. **entità astratte**
   (a) azioni e avvenimenti
   (b) entità astratte e proprietà degli oggetti, dei materiali o delle azioni
   (c) discipline o scienze
   (d) unità di misura
3. **entità individuali o "classi di uno" analoghe a nomi propri**.

In fase di costruzione del th è di fondamentale importanza il controllo dell'appartenenza dei termini a queste categorie, poiché esse possono influenzare determinate procedure, come ad esempio la scelta del plurale o del singolare, o verificare la validità delle gerarchie (non può esistere rapporto gerarchico fra termini appartenenti a categorie diverse).

Struttura semantica e formalizzazione delle relazioni

La struttura relazionale consente al th di diventare una sorta di "mappa" dei significati espressi da un certo linguaggio di indicizzazione, che consente sia all'indicizzatore in fase di attribuzione dei descrittori al documento, sia al ricercatore in fase di costruzione del profilo di ricerca, di scandirne la rete semantica percorrendo nei sensi desiderati relazioni e strutture classificatorie.
Abbiamo già differenziato e definito i componenti lessicali del th in due categorie fondamentali, quella dei **TP** e quella dei **TNP**: tra queste due categorie di termini si instaura la prima fondamentale relazione semantica di un th, cioè la relazione **preferenziale** o **sinonimica** o **di equivalenza**.
La **relazione preferenziale** è quella deputata a rapportare uno o più termini non preferiti ad un termine preferito che esprime lo stesso concetto o un concetto molto simile, che sarà rappresentato sempre univocamente dal TP. Il gruppo di termini che si assume rappresentino lo stesso concetto, che si considerano, ai fini dell'indicizzazione, equivalenti, e fra i quali viene scelto il termine preferito, si definisce **gruppo di equivalenza**.
Le relazioni thesauriche vengono abitualmente esplicitate e rappresentate da un corredo di simboli o sigle, fra cui le più usate sono quelle suggerite dallo standard. In particolare, per la relazione preferenziale, il rinvio dal TNP al TP viene indicato dal simbolo **USE**:

```
tesauri
  USE Thesauri
```

mentre il rapporto reciproco, cioè la segnalazione dei TNP nel **corredo semantico** (o **blocco-parola**) del TP è indicato dal simbolo **UF**:

```
Thesauri
  UF tesauri
```

Possiamo ulteriormente distinguere varie sottospecie di relazioni preferenziali, determinate ad esempio dal tipo di relazione fra TNP e TP, **univoca** se si tratta di un rapporto 1:1, cioè se ad un TNP corrisponde un solo TP, **biunivoca** se si tratta di un rapporto 1:2, cioè se ad un TNP rappresentante un concetto complesso corrispondono due distinti TP rappresentanti suoi concetti costitutivi più semplici, che devono essere usati obbligatoriamente insieme:

```
sideremia
  USE Ferro AND Sangue
```

il cui reciproco si esprime con:

```
Ferro
  UF+ sideremia
```

e:

```
Sangue
  UF+ sideremia
```

All'interno della relazione di **preferenza univoca** distinguiamo ancora i due casi di:

1. **sinonimia assoluta o accentuata**
2. **sinonimia relativa e convenzionale**

Il primo si verifica se, indipendentemente dall'area semantica, dal grado di analiticità del th e da quale dei due termini viene definito come preferito e quale come non preferito, tra TP e TNP esiste sempre un rapporto sinonimico (si potrebbe perciò dire che il rapporto sinonimico è tale *a priori*). Rientrano in questo caso diverse tipologie:

- **sinonimia vera** (*regola* e *norma*)
- **variante ortografica** (*psicopedagogia* e *psico-pedagogia*)
- **sigle e acronimi** (*CNR* e *Centro Nazionale Ricerche*)
- **preferenza linguistica** (a. fra termine straniero e italiano: *week end* e *fine settimana*, *mountain bike* e *bici da montagna*; b. fra termine attuale e antico: *bicicletta* e *velocipede*; c. fra termine corrente e scientifico: *zucca* e *Cucurbita Maxima*, *mal di testa* e *cefalea*; d. fra termini di origini linguistiche differenti: *poliglotta* e *multilingue*; e. fra nomi comuni e marche commerciali: *penna a sfera* e *biro*; f. fra varianti esprimenti nozioni, oggetti, discipline nuovi o emergenti: *calcolatori* e *cervelli elettronici*, *telefoni portatili*, *telefonini* e *telefoni cellulari*)

Il secondo caso, quello della **sinonimia convenzionale**, si verifica quando la relazione tra due termini di significato vicino, appartenenti alla stessa area semantica, non è sinonimica in senso stretto, non verrebbe considerata tale nel linguaggio naturale, non è sempre considerata tale in tutti i th. In linea di massima si può dire che tale relazione viene risolta in un rapporto preferenziale (cioè un termine assume il ruolo di TP e l'altro di TNP) se i due termini si collocano in un'area di secondario rilievo per il th (la cosiddetta **fringe**, **frangia**, **area marginale**), in due diversi Termini Preferiti se questi si collocano entro l'area centrale del th (**core**, **nocciolo**, **nucleo dell'area**).
Anche in questo caso possiamo distinguere diverse tipologie:

- **quasi-sinonimia** (*punizione* e *sanzione*)
- **upward posting** o **rinvio al superiore gerarchico** (*microformati* e *microfilm*, *microformati* e *microfiche*); si tratta di una tecnica che accomuna nello stesso "insieme di equivalenza" termini in realtà non sinonimi ma appartenenti a diversi gradini della scala gerarchica. Viene scelto come TP quello di livello superiore, più generico
- **antinomia** (*pace* e *guerra*, *malattia* e *salute*, *secchezza* e *umidità*); è improprio collocare questo caso fra le sottospecie della sinonimia, in quanto ovviamente due termini opposti, o, come detto sopra, due termini collocati sullo stesso continuum semantico, valori diversi della stessa variabile, non sono

certo sinonimi. Si assume tuttavia che lo siano ai fini dell'indicizzazione e della ricerca in quanto un documento che tratta di un certo aspetto di un problema tratterà anche presumibilmente del suo opposto, o comunque potrà essere descritto concettualmente utilizzando il termine che indica il suo opposto.

La relazione sinonimica è l'unica relazione thesaurica che mette in rapporto tra di loro l'insieme dei termini preferiti e quello dei termini non preferiti: tutte le altre relazioni, di cui stiamo per occuparci, sono relazioni **fra termini preferiti**.
La relazione di base che distingue tipicamente un th da una semplice lista di termini non strutturati, e ne rivela le fondamenta sistematiche, è la relazione classificatoria per eccellenza, la **relazione gerarchica**. La relazione gerarchica esprime il concetto ed il grado di subordinazione o sovraordinazione fra termini appartenenti allo stesso albero gerarchico; in questo albero, il termine sovraordinato rappresenta una classe o un tutto, e il termine subordinato rappresenta un suo elemento o parte. La **sigla** che nel blocco parola di un termine individua i suoi sovraordinati è **BT** (**Broader Term**), alla quale può utilmente essere aggiunta una cifra indicante la distanza in "gradini" gerarchici fra i due termini legati dal rapporto:

```
Geometria iperbolica
  BT1 Geometria non euclidea
    BT2 Geometria
      BT3 Matematica
```

La sigla identificante il rapporto inverso, cioè i subordinati del termine dato, è **NT** (**Narrower Term**):

```
Geometria
  NT1 Geometria euclidea
  NT1 Geometria non euclidea
    NT2 Geometria iperbolica
    NT2 Geometria ellittica
```

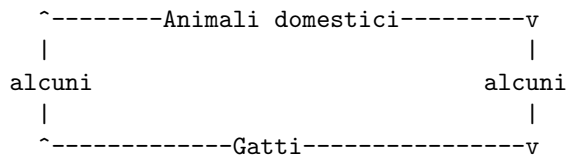
Rientrano nella categoria delle relazioni gerarchiche tre sottospecie di relazioni:

1. la **relazione generica o relazione genere-specie** (distinta eventualmente dalle sigle BTG e NTG)
2. la **relazione partitiva o relazione parte-tutto** (distinta eventualmente dalle sigle BTP e NTP)
3. la **relazione esemplificativa o classe-istanza o specie-esempio**

ognuna delle quali corrisponde ad una diversa situazione logica e conduce a gerarchie verificabili attraverso un test di verifica di tipo logico. Resta inteso che la relazione gerarchica può intercorrere solo fra termini che fanno riferimento a concetti appartenenti alla stessa categoria di nozioni (ad esempio oggetti, azioni, proprietà, discipline; per l'elenco completo delle categorie vedi il paragrafo *Termini d'indicizzazione*).
La **relazione generica** o **relazione genere-specie** identifica il legame che intercorre fra una classe o categoria ed i suoi elementi, membri o specie, ed è la tipica relazione delle classificazioni zoologiche o botaniche. Per soddisfare a questo tipo di relazione i termini non solo devono appartenere alla stessa categoria, ma rispondere anche alla condizione "tutti e sempre" in senso ascendente, ovvero alla condizione "alcuni/tutti" nei due sensi discendente e ascendente. Un esempio per chiarire:

```
        ^--------Felini---------v
        |                       |
tutti e sempre              alcuni
        |                       |
        ^---------Gatti---------v
```

alcuni elementi della classe Felini sono dei Gatti, tutti i Gatti sono sempre, per definizione e indipendentemente dal contesto, dei Felini. Viceversa lo schema:



chiarisce perché non è possibile instaurare un rapporto gerarchico genere-specie fra Gatti e Animali domestici: se è vero che alcuni Animali domestici sono Gatti, non è però vero che tutti i Gatti sono sempre Animali domestici (esistono infatti anche i gatti selvatici), e i due termini devono quindi appartenere a categorie diverse del th. Il test ha lo scopo di evitare che criteri soggettivi del costruttore influenzino la stesura delle strutture tassonomiche, tuttavia s'intende che per scopi o in ambiti particolari ci si possa sottrarre dal seguire rigidamente questa norma. Riprendendo l'esempio appena citato, se si sta costruendo un th specializzato sugli Animali domestici si potrà a buon diritto fare dei Gatti un NT di quelli, appartenente alla stessa categoria, dal momento che i Gatti selvatici non compariranno nel th e non rivestono alcun interesse per l'utente.
La **relazione partitiva** o **parte-tutto** *non* è considerata dallo standard una relazione gerarchica universalmente valida, ma è sottoposta ad una restrizione: essa è ritenuta valida solo nelle situazioni in cui il nome della **parte** implica il nome del corrispondente **tutto**, qualunque sia il contesto. In tal caso i termini possono essere strutturati gerarchicamente come BT (tutto) e NT (parte); lo standard elenca quattro casi rispondenti a questo requisito in un contesto generale:

1. sistemi e organi del corpo

   ```
   Sistema circolatorio
     NT1 Sistema vascolare
       NT2 Arterie
       NT2 Vene
   ```

2. luoghi geografici

   ```
   Canada
     NT1 Manitoba
       NT2 Winnipeg
   ```

3. discipline e campi di studio

   ```
   Scienze
     NT1 Chimica
     NT1 Biologia
       NT2 Botanica
   ```

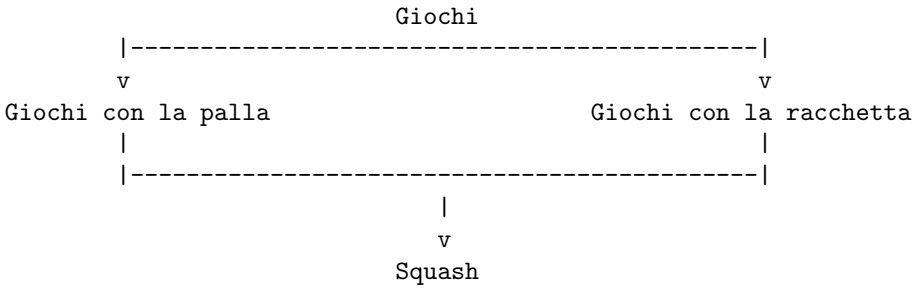
4. strutture sociali gerarchizzate

   ```
   Corpi d'armata
     NT1 Divisioni
       NT2 Reggimenti
   ```

mentre riconosce la possibilità di organizzare gerarchicamente anche altri tipi di termini solo in contesti particolarmente specializzati, in cui è la specializzazione del campo di azione a garantire che il nome del tutto sia univocamente implicato dal nome delle sue parti.
La **relazione esemplificativa** o **specie-esempio** identifica il legame che intercorre fra una **classe** o categoria generale di cose o avvenimenti, espressa da un nome comune, ed un suo **individuo**, rappresentato da un nome proprio, e costituente una "classe di uno". In realtà la cosa può complicarsi ulteriormente, poichè esistono classi di uno che possono avere in sottordine altre classi di uno:

```
Regioni montuose <classe>
         <rel. specie-es.>
  NT1 Alpi <individuo>
           <rel. parte-tutto>
    NT2 Alpi Graie <individuo>
```

L'esempio mostra come la prima relazione, quella classe-individuo, è una relazione esemplificativa, mentre la seconda, quella individuo-individuo, è una relazione partitiva. Possiamo concludere perciò che la relazione esemplificativa è sempre tra un BT di tipo classe e un NT di tipo individuo, mentre la relazione partitiva è o tra due classi o tra due individui.
Aggiungiamo ancora che nella maggior parte dei th la relazione esemplificativa è del tutto assente, per evitare che la presenza dei nomi propri sovraccarichi le categorie rendendole difficili da gestire. Una soluzione frequentemente adottata consiste nello stendere elenchi a parte in cui gli "individui" sono rappresentati da termini normalizzati nella forma detti **identificatori** (perché appunto identificano gli individui) per analogia con **descrittori** (cioè i termini che "descrivono" una classe). E' chiaro che se nel th non sono presenti gli individui come TP inseriti in relazioni esemplificative, l'indicizzatore dovrà attribuire al documento che tratti di un "individuo" anche il termine che indica la sua classe di appartenenza.
Per completare il discorso sulle relazioni gerarchiche, distinguiamo due tipi di th, il **th monogerarchico** e il **th poligerarchico**, a seconda che i termini possano appartenere, per ragioni logicamente fondate, ad una sola o a più d'una categoria.



Mentre nei th poligerarchici un termine può appartenere a più di una gerarchia, nei monogerarchici ogni termine può avere uno ed un solo BT1, poiché può essere inserito in una sola catena gerarchica: con le gerarchie ulteriori esso può intrattenere solo un rapporto di tipo associativo; questo significa che, benché i legami fra i termini siano gli stessi, il th poligerarchico evidenzia un maggior numero di strutture rispetto al monogerarchico:

Poligerarchico

```
Organo
  BT1 Strumenti a fiato
  BT1 Strumenti a tastiera
    BT2 Strumenti

Strumenti a fiato
  BT1 Strumenti
  NT1 Organo
  NT1 Flauto

Strumenti a tastiera
  BT1 Strumenti
  NT1 Organo
  NT1 Pianoforte
```

Monogerarchico

```
Organo
  BT1 Strumenti a fiato
    BT2 Strumenti
  RT Strumenti a tastiera

Strumenti a fiato
  BT1 Strumenti
  NT1 Organo
  NT1 Flauto

Strumenti a tastiera
  BT1 Strumenti
  RT Organo
  RT Pianoforte
```

L'ultima delle relazioni thesauriche classiche, la **relazione associativa**, è una relazione che si illustra più facilmente analizzandone le caratteristiche negative che non definendola in positivo; non a caso è detta anche relazione "residuale", in quanto è in grado di collegare coppie di termini che non rientrano né nella casistica della relazione sinonimica (non fanno parte dello stesso "insieme di equivalenza"), né in quella della relazione gerarchica (non appartengono alla stessa catena gerarchica), ma sono tuttavia così fortemente associati che è necessario esplicitarne il legame all'interno del th in modo da poter suggerire all'indicizzatore o al ricercatore che acceda al primo dei due il secondo termine. La relazione associativa è reciproca, e viene indicata in ambedue i casi con la sigla **RT** (**Related Term**). Proprio perché non è sottoposta a rigidi test di verifica e ammette un certo grado di discrezionalità, è importante adottare metodi che garantiscano la relazione associativa dallo scatenarsi delle opinioni personali e dei giudizi soggettivi del costruttore del th: come regola generale, diciamo che uno dei due termini deve essere fortemente implicato dall'altro, ovviamente nel quadro di riferimento condiviso dagli utenti del th. Un esempio in particolare è il rapporto di dipendenza fra la disciplina e il suo oggetto, fattispecie del rapporto tipicamente associativo fra due termini, di cui uno è necessario alla spiegazione o definizione dell'altro. Ci sono due tipi di termini suscettibili di intrattenere rapporti associativi:

1. **quelli appartenenti alla stessa categoria**
2. **quelli appartenenti a categorie diverse**

Fra i termini appartenenti alla **stessa categoria** distinguiamo brevemente fra:

a) termini che hanno lo stesso termine sovraordinato, ed i cui significati hanno una zona di sovrapposizione, e che quindi, anche se nel th hanno una definizione che li distingue esattamente, potrebbero essere adoperati dagli utenti in maniera non rigorosa (e perciò quasi intercambiabile), per i quali è quindi necessario ricordare l'esistenza dell'altro quando l'utilizzatore impiega l'uno:

```
Barche
  BT Veicoli
  RT Navi

Navi
  BT Veicoli
  RT Barche
```

b) termini che rappresentano concetti legati da una relazione di tipo "familiare" o di tipo "derivato" (un concetto che deriva dall'altro). Ad esempio, poiché i Muli derivano dall'incrocio fra Asini e Cavalli (rispetto ai quali sono però situati sullo stesso gradino gerarchico, cioè come NT1 di Equini), si possono configurare i seguenti blocchi-parola:

```
Equini
  NT1 Asini
  NT1 Cavalli
  NT1 Muli

Asini
  BT1 Equini
  RT Cavalli
  RT Muli

Cavalli
  BT1 Equini
  RT Asini
  RT Muli

Muli
  BT1 Equini
  RT Asini
  RT Cavalli
```

Fra termini appartenenti a **categorie diverse**, sempre rispondenti al requisito dell'implicazione dell'uno dall'altro, si configurano diverse tipologie di rapporti che possono motivare una relazione associativa:

1. una disciplina e il suo oggetto di studio (*zoologia* e *animali*);
2. un processo od operazione e il suo agente o strumento (*termometro* e *misurazione della temperatura*);
3. una azione e il suo prodotto (*scrittura* e *document*i);
4. una azione e chi o cosa la subisce (*potatura* e *piante*; *pesca* e *pesci*);
5. oggetti e fenomeni e loro proprietà (*magneti* e *magnetismo*);
6. concetti e loro origini (*Tedeschi* e *Germania*);
7. concetti legati da rapporti causali (*inquinamento* e *sostanze inquinanti*);
8. una cosa e il suo antidoto (*piante* ed *erbicidi*);
9. un concetto e la sua unità di misura (*frequenza* e *hertz*);
10. locuzioni sincategorematiche (cioè termini composti) e loro nomi sottocategoriali (*piante fossili* e *piante*).