Terminologia, lessici specialistici e strutture tassonomiche nel dominio dell'efficienza energetica e dell'applicazione di fonti rinnovabili agli usi finali civili

Elisabetta Oliveri[1], Concetta Baroniello[1], Antonietta Folino[2], Rossella Scaioli[1]

[1]ITC-CNR, San Giuliano Milanese, Milano
{formazione, elisabetta.oliveri, rossella.scaioli}@itc.cnr.it

[2]Università della Calabria, Rende, Cosenza
antonietta.folino@unical.it

Abstract

Obiettivo della presente proposta è la descrizione degli sviluppi di un progetto per la costruzione di un centro di documentazione sui temi dell'efficienza energetica e dell'applicazione di fonti rinnovabili agli usi finali civili.

Tra gli scopi del progetto rientra la realizzazione di strutture di accesso e di organizzazione della conoscenza al fine di sopperire all'assenza in Italia di strumenti omogenei di classificazione e di controllo terminologico e di gestire fenomeni quali l'evoluzione linguistica, tecnica e normativa e la presenza di varianti locali. In particolare, si tratta di costruire un lessico specialistico condiviso tra i soggetti della filiera e un thesaurus che controlli l'uso dei termini e che consenta l'accesso all'informazione e la classificazione univoca dei concetti.

La prima fase per realizzare tali strumenti ha previsto la raccolta dei documenti dai quali estrarre i termini e la loro organizzazione in corpora. Si tratta di testi in lingua italiana, la cui copertura temporale è riferita all'ultimo quinquennio, affinché sia garantita la presenza di termini relativi alle tecnologie e ai materiali più innovativi, e appartenenti a diverse tipologie, quali articoli scientifici, riviste, leggi, norme, ecc., e a diversi contesti regionali. La loro eterogeneità ha permesso di estrarre termini specialistici e termini appartenenti al linguaggio comune, tra i quali, nel thesaurus, saranno stabilite delle relazioni di equivalenza. I documenti sono stati raggruppati in sottocorpora affinché possa essere valutata la frequenza dei termini e il rapporto tra termini specialistici e comuni in ciascuno di essi. L'estrazione terminologica è stata realizzata tramite il software T2K e risultati interessanti emergono anche dal confronto tra le liste terminologiche ottenute stabilendo soglie di frequenza diverse per lo stesso corpus. La valutazione manuale dei candidati a termine restituiti dal tool permette di ritenere solo i termini rappresentativi del dominio e di organizzarli nel thesaurus tramite un set di categorie e tramite relazioni semantiche standard.

In riferimento ad alcune scelte terminologiche, ci si propone di procedere parallelamente alle ricerche condotte dall'UNI e finalizzate alla codifica univoca di opere, attività e risorse nel settore dell'edilizia. Nella norma, la denominazione univoca prevede l'assegnazione di un nome complesso, costituito da un codice, dal nome comune e dalle caratteristiche identificative dell'oggetto di interesse, mentre la descrizione prevede la definizione di caratteristiche non identificative dell'oggetto stesso. Un possibile parallelismo potrebbe essere stabilito con la gestione dei termini composti, anche in riferimento alla loro introduzione nel thesaurus.

Keywords: centro di documentazione, thesaurus, terminologia, efficienza energetica

1. Introduzione

Nell'ambito del progetto "Nuove tecnologie e strumenti per l'efficienza energetica e l'utilizzo delle fonti rinnovabili negli usi finali civili", frutto di un Accordo Quadro tra la Regione Lombardia e il Dipartimento Sistemi di Produzione (DSP) del Consiglio Nazionale delle Ricerche (CNR) per l'attuazione di programmi di ricerca e sviluppo, il servizio Formazione e Informazione Tecnica dell'Istituto per le Tecnologie della Costruzione (ITC) è responsabile della realizzazione di un centro di documentazione e informazione scientifica sui temi del progetto.

Con il presente articolo ci si propone di illustrare gli sviluppi registratisi rispetto a quanto già presentato in (Folino et al., 2009) soprattutto in termini di predisposizione di strumenti di accesso all'informazione contenuta nel centro di documentazione. Nello specifico, si tratta della costruzione di un lessico specialistico e di una struttura tassonomica alla quale potranno essere assegnate funzioni di navigazione, indicizzazione, espansione dei risultati delle ricerche e recupero più efficiente dell'informazione. Quest'ultimo punto è particolarmente interessante se si considera che i potenziali utenti del centro di documentazione apparterranno a profili professionali differenti[1], i cui bisogni informativi saranno espressi per mezzo di termini più o meno specialistici. L'esistenza di relazioni per la gestione della sinonimia permetterà di recuperare l'informazione ricercata indipendentemente dal termine inserito per effettuare la ricerca libera.

Al centro di documentazione saranno attribuite, soprattutto sul lungo periodo, le seguenti funzioni: raccogliere e organizzare informazioni e documenti sul dominio di riferimento, fornire servizi di informazione tecnica e formazione, rendere fruibili i risultati relativi ai restanti work package del suddetto progetto. In particolare, si ritiene di fondamentale importanza fornire agli utenti uno strumento informativo che funga al tempo stesso da repository di documenti, all'interno della quale gli utenti possano trovare informazioni aggiornate in un contesto di continua evoluzione tecnologica e da fonte di servizi innovativi. A tal proposito, oltre a monitorare in maniera costante l'evoluzione normativa e tecnologica che interessa il dominio di riferimento[2], si stanno predisponendo servizi di informazione tecnica che rispondano ai potenziali bisogni di alcune tra le categorie di utenti individuate[3].

Dal punto di vista architetturale il prototipo attualmente realizzato per il centro di documentazione è basato su un database relazionale e talune funzioni di gestione

[1] Professionisti del settore, studenti, enti di ricerca, utenti privi di conoscenze specialistiche, ecc.

[2] Anche attraverso l'iscrizione a servizi di news, blog e simili, i quali rimandano ad informazioni che necessitano una successiva fase di valutazione e di organizzazione.

[3] Un esempio è rappresentato dal confronto della normativa regionale in materia di certificazione energetica degli edifici sulla base di parametri predeterminati relativi ai requisiti dei soggetti certificatori, alle caratteristiche dell'attestato di certificazione, al metodo di calcolo delle prestazioni di un edificio, ecc. Le differenze riscontrate, dovute al fatto che alcune regioni italiane hanno legiferato in maniera del tutto indipendente e prima che l'Italia recepisse la direttiva Europea in materia di certificazione (Direttiva 2002/91/CE del Parlamento Europeo e del Consiglio del 16 Dicembre 2002 sul rendimento energetico nell'edilizia) con le Linee Guida Nazionali (DECRETO 26 giugno 2009, Linee guida nazionali per la certificazione energetica degli edifici), sono di notevole importanza soprattutto per coloro che vogliano o debbano certificare edifici in una regione diversa da quella in cui hanno ottenuto l'abilitazione di certificatore. Il lavoro si sta ora orientando anche verso l'analisi dei requisiti analizzati e dei metodi adottati per la definizione della classe di efficienza energetica di un edificio, poiché è del tutto possibile che ad una stessa struttura venga attribuita una classe diversa a seconda della regione in cui il calcolo viene eseguito.

documentale, quali l'indicizzazione e la ricerca, sono supportate dalla descrizione standardizzata dei documenti utilizzando il set di metadati Dublin Core[4].

2. Contesto e motivazioni

Se si considera l'ambito di interesse del suddetto progetto, l'efficienza energetica e l'applicazione delle fonti rinnovabili agli usi finali civili, il contesto internazionale è caratterizzato da poche iniziative orientate alla rappresentazione della conoscenza di dominio. Si tratta spesso di sistemi dalla copertura semantica più ampia, ai quali è possibile fare riferimento per l'organizzazione di categorie e termini propri del dominio di interesse. Possono essere citati il GEMET Thesaurus[5], la cui themes list contiene temi rilevanti per lo scopo della presente ricerca quali Energia, Politica Ambientale, Edilizia e il JOINT Thesaurus[6] (International Nuclear Information System (INIS), Energy Technology Data Exchange (ETDE)) che comprende, tra i comparti di interesse, anche le fonti di energia rinnovabili e l'efficienza energetica.

Il settore dell'edilizia, invece, è caratterizzato dalla presenza di numerose iniziative nel contesto internazionale, mentre il panorama italiano si dimostra carente di risorse terminologiche e/o di classificazione, se si escludono alcune tipologie di documenti tecnici, quali i prezzari regionali delle opere edili, nei quali le informazioni sono organizzate in forma di classificazione, pur non essendo questo il loro obiettivo primario; inoltre essi presentano notevoli differenze e incongruenze nelle scelte alla base dell'organizzazione dei concetti (Folino et al., 2009).

Nel panorama internazionale, particolare importanza riveste la norma ISO 12006-2, "Organization of information about construction works – Part 2: Framework for classification of information", la quale affianca ad una lista di definizioni dei concetti fondamentali del dominio delle costruzioni uno schema che stabilisce le relazioni tra questi e che è alla base di tabelle di classificazione che, tuttavia, non prevedono un contenuto dettagliato. Ciascuna di esse contiene, infatti, solo possibili intestazioni generiche e l'assegnazione di istanze concrete ad ogni tavola non viene in alcun modo esplicata o esemplificata. È chiaramente sottolineato nel testo della stessa norma che tale sistema funge da linee guida per gli operatori del settore, ma che la sua concreta implementazione può variare e adattarsi alle esigenze specifiche di ciascun contesto.

Per quanto riguarda le scelte di classificazione, in particolare l'identificazione di tre grandi classi (construction result, construction process e construction resource), è possibile stabilire un parallelismo tra la ISO 12006-2 e la recente norma UNI 11337 del novembre 2009, "Criteri di codificazione di opere e prodotti da costruzione, attività e risorse. Identificazione, descrizione e interoperabilità", che identifica appunto la tripartizione in Opere, Attività e Risorse. Tuttavia, è importante sottolineare come, a differenza della norma internazionale, la UNI 11337 non abbia pretese di classificazione, ma si proponga la sola attribuzione "orizzontale" dei concetti alle classi identificate.

Tale norma mira alla creazione di un sistema univoco di codifica e al raggiungimento di un elevato grado di condivisione nel riconoscimento di opere, attività e risorse del dominio delle

[4] <http://dublincore.org/>

[5] Tra le lingue previste rientra anche l'italiano. Si tratta di un thesaurus poligerarchico. <http://www.eionet.europa.eu/gemet>

[6] Il thesaurus, in inglese, si presenta come una lista alfabetica di termini, per ciascuno dei quali sono rese esplicite le relazioni gerarchiche, di equivalenza e associative con gli altri termini.

costruzioni, per mezzo di un'informazione normalizzata, condivisa e interoperabile e prevede alcune specifiche tecniche finalizzate alla denominazione, descrizione, raccolta e archiviazione dell'informazione tecnica e all'interoperabilità tra applicazioni software.

In particolare, il sistema di codifica unico dovrà garantire l'interoperabilità tra le varie basi di dati che attualmente strutturano l'informazione senza alcun criterio di uniformazione e la possibilità di scambio non ambiguo dei dati che esse contengono. L'applicazione del quadro normativo si renderà necessaria nel tentativo di identificare in modo univoco l'informazione contenuta in schede tecniche e in documenti di descrizione e presentazione, economici, di prescrizione, ecc. Ciascun soggetto della filiera dovrà essere in grado di identificare in maniera non ambigua gli elementi di conoscenza del dominio attraverso una terminologia e una simbologia condivise.

Nello specifico, il riconoscimento univoco di un soggetto, oggetto o attività deve avvenire attraverso l'attribuzione di un nome complesso, che prevede la definizione di alcune caratteristiche qualitative e quantitative indispensabili per l'identificazione del concetto. Ulteriori proprietà non necessarie a tale scopo saranno utilizzate per finalità descrittive all'interno di un'apposita scheda tecnica.

Le regole stabilite da tale norma per la definizione univoca di opere, attività e risorse costituiscono un punto di riferimento nell'identificazione dei termini per il lessico e per il thesaurus: è interessante valutare il rapporto tra l'uso del nome complesso previsto dalla norma e l'utilizzo dei termini composti nel thesaurus così come prescritto dalla normativa sulla costruzioni dei vocabolari controllati. Le regole alla base della combinazione dei concetti appartenenti alle diverse faccette di un thesaurus possono inoltre essere utilizzate per associare ai termini semplici le caratteristiche considerate fondamentali nell'identificazione non ambigua dei concetti.

3. Approccio metodologico

La predisposizione dei suddetti strumenti terminologici prevede l'articolarsi delle fasi di seguito descritte, alcune delle quali ancora in corso.

3.1 Costituzione del corpus e estrazione terminologica

Il primo momento è stato consacrato alla costituzione di un corpus documentale che, sulla base dei principi della linguistica dei corpora, risponda il più possibile al criterio di rappresentatività, secondo il quale il campione di documenti facenti parte del corpus deve avere dimensioni tali e deve essere sufficientemente equilibrato da poter render conto di come i fenomeni oggetto di interesse si manifestano nell'uso reale della lingua[7].

Nel caso specifico, il set di documenti deve essere rappresentativo dei termini effettivamente utilizzati nei diversi contesti d'uso al fine di poterne estrarre i candidati a termine per la costruzione del lessico e del thesaurus. Si tratta di un corpus dello scritto, specialistico per il contenuto dei documenti, pluridisciplinare, con diversi livelli di specializzazione, eterogeneo per tipologie testuali in esso contenute, e monolingue.

Il corpus costituito consta attualmente di circa 300 documenti e si è optato per un'ulteriore loro organizzazione in sottocorpora rappresentativi, al tempo stesso, delle tipologie testuali,

[7] La linguistica dei corpora può essere così definita: "The study of language based on examples of 'real life' language use" (McEnery, Wilson, 2001)

del registro di linguaggio utilizzato e dell'obiettivo di ciascuna di esse (es. divulgativo per gli opuscoli). Tale scelta permette di analizzare e di mettere a confronto – in un'ottica di corpus comparison - la distribuzione terminologica in ciascun set di documenti, che, sulla base degli scopi della presente ricerca, permette di compiere una prima distinzione tra termini preferiti e non preferiti attraverso un'analisi non indifferenziata delle fonti documentali.

I sottocorpora oggetto di analisi allo stadio attuale sono quattro e sono costituiti dagli articoli e dalle rendicontazioni scientifiche, dalle riviste di settore, dalle leggi e dalle norme, dagli opuscoli e dalle linee guida.

Sui corpora così costituiti è stata eseguita una fase di estrazione terminologica semi-automatica per mezzo del prototipo software dedicato T2K (Text to Knowledge), specificatamente concepito per il trattamento della lingua italiana[8] e sviluppato dall'ILC (Istituto di Linguistica Computazionale) di Pisa.

L'assunto principale sul quale si basa T2K, così come la maggior parte dei software di estrazione terminologica, è che i concetti rilevanti di un testo siano veicolati dai termini che vi ricorrono in maniera più frequente. Il software integra sistemi di analisi linguistica automatica, algoritmi statistici per l'identificazione di cluster concettuali, tool di annotazione o di marcatura dei testi e dati strutturati a supporto dell'indicizzazione terminologico - concettuale dei documenti (Dell'Orletta et al., 2008). Tra le tecniche linguistiche rientrano la tokenizzazione, la lemmatizzazione, la Part of Speech Tagging con la conseguente creazione di una stop list, il chunking di supporto al riconoscimento e all'estrazione di unità polirematiche.

Il risultato dell'analisi linguistico-statistica è costituito da un vocabolario terminologico corredato da informazioni sulla frequenza di ciascun termine, oltre che da informazioni semantiche e concettuali sui termini stessi, che costituiscono il valore aggiunto dell'output ottenuto.

La frequenza viene calcolata a livello di lemma e non di forma, affinché le informazioni relative alla rilevanza di un termine all'interno del documento e del corpus non risultino alterate e quindi poco significative, mentre per i termini composti viene calcolata la frequenza di co-occorrenza delle unità semplici che lo compongono: se le parti costituenti ricorrono in maniera significativamente frequente rispetto alla frequenza con cui compare ciascuna di esse, la probabilità che si tratti di un'unità polirematica è molto elevata. Il valore di rilevanza del termine è ottenuto tramite il calcolo della funzione TF*IDF (Term Frequency*Inverse Document Frequency) e a ciascun termine viene attribuito un indice posizionale, che permette di tenere traccia del o dei documenti all'interno dei quali esso occorre e di verificarne le co-occorrenze.

I termini sono organizzati in base a relazioni semantiche quali la relazione gerarchica di iperonimia/iponimia definita a partire dalla loro struttura linguistica interna, ovvero sulla base del principio di inclusione lessicale, e la relazione di affinità semantica (Dell'Orletta et al, 2008). Nel primo caso le unità polirematiche condividono la medesima testa lessicale o focus, che corrisponde all'unità monorematica rispetto alla quale si trovano in posizione iponimica, ma la condivisione può riguardare anche la testa e uno o più modificatori o differenze (Es. impianti, impianti di climatizzazione, impianti di climatizzazione invernale), mentre nel secondo caso, la relazione viene stabilita tra i termini considerati affini dal punto di vista semantico e quindi potenziali varianti, quasi sinonimi o termini correlati. L'identificazione dei termini

[8] L'insieme di tool per il trattamento automatico della lingua italiana va sotto il nome di AnIta.

semanticamente affini avviene sulla base dei contesti lessicali nei quali essi occorrono: se la distribuzione di due termini coincide in maniera frequente, gli stessi possono essere reciprocamente sostituibili e utilizzati più o meno indifferentemente in taluni contesti d'uso e hanno, perciò, un significato affine. Per ciascun termine vengono identificati molteplici potenziali termini correlati, che insieme costituiscono una sorta di rete o mappa concettuale intorno al termine stesso, delineandone lo spazio semantico[9].

Il risultato del processo di estrazione, soprattutto in termini di quantità di unità estratte, sia mono- che polirematiche, dipende dai valori assegnati ad alcuni parametri prima che il processamento dei dati venga lanciato. Tra questi parametri rientra il valore della soglia massima di frequenza al di sotto della quale il termine non viene presentato nella lista fornita in output dal sistema. Il valore da assegnare a questa soglia deve essere ben valutato per evitare che termini importanti per il dominio, ma meno frequenti nel corpus, non vengano inseriti nel thesaurus, soprattutto se le dimensioni del corpus sono modeste.

I risultati ottenuti hanno dimostrato che spesso specificare una soglia pari a 3, che corrisponde al valore definito per default, comporta una perdita di termini rilevanti, che compaiono nel corpus con una frequenza inferiore. La tabella di seguito riportata dimostra come la scelta di una soglia relativamente elevata possa far perdere termini assolutamente rilevanti:

Articoli Scientifici	
Soglia 3	Termini estratti: 1313
Soglia 1	Termini estratti: 3685
Termini non estratti con soglia 3	- Recupero di edifici esistenti - Coibentazione - Laterizio - …

La scelta dei termini candidati a diventare voci del lessico specialistico e termini del thesaurus viene effettuata, oltre che sulla base di criteri statistici, quale il valore del TF*IDF restituito da T2K, anche con il supporto di esperti del dominio, che possono rendere conto dell'utilizzo di un determinato termine, e considerando la fonte dalla quale i candidati a termini vengono estratti. La valutazione di più criteri permette di non basarsi solo sulla frequenza, che potrebbe non essere sufficiente nel caso di termini altamente specialistici e quindi meno frequenti o di termini che indicano innovazioni tecnologiche ancora poco presenti in letteratura, o solo sugli esperti, le cui indicazioni potrebbero essere contrastanti, limitate ad uno specifico contesto, ecc.

L'output generato dall'applicazione di T2K al corpus documentale necessita di una fase di validazione manuale, rendendo di fatto l'estrazione terminologica semiautomatica e riducendo

[9] Ad esempio, come potenziali termini correlati di calore il software restituisce indice, potenza e energia.

la quantità di termini selezionati, come dimostrato dalla figura seguente[10]. Tuttavia esso fornisce un supporto significativo alla successiva costruzione di prodotti terminologici, quale il thesaurus, basato principalmente su relazioni di tipo gerarchico, associativo e di equivalenza.

Articoli scientifici	
Termini estratti con T2K	Candidati termini ritenuti
3685	850

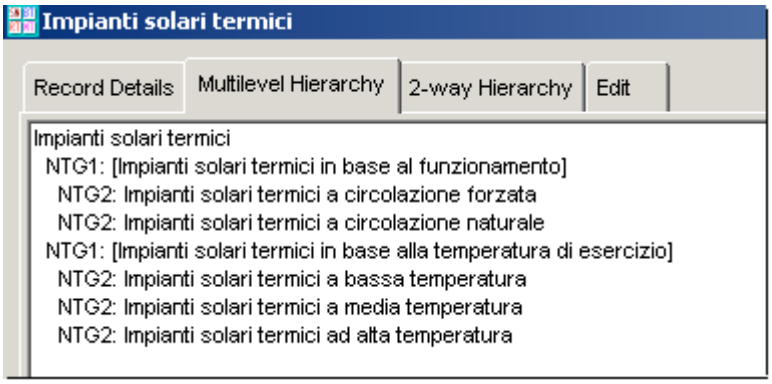
3.2 Costruzione del thesaurus

A partire dai termini così ottenuti, si è passati alla fase di normalizzazione e organizzazione degli stessi all'interno del thesaurus, sulla base di quanto prescritto dalla più recente normativa in materia di costruzione di vocabolari controllati (ISO 25964-1, Thesauri and interoperability with other vocabularies — Part 1: Thesauri for information retrieval, 2009), soprattutto per quanto riguarda la scelta della forma dei termini, l'inserimento di termini composti, la gestione di relazioni poligerarchiche, e simili. In riferimento alla struttura di tale strumento, si è optato per la costruzione di un thesaurus a faccette, la cui definizione comporta, a partire dai principi della classificazione a faccette ideata da Ranganathan (Ranganathan, 1988), la scomposizione di un dominio di conoscenza in categorie, dette appunto faccette, rappresentative dello stesso, secondo un approccio analitico-sintetico[11]. Rispetto alle tradizionali strutture gerarchiche, tale forma di organizzazione presenta vantaggi in termini di pluridimensionalità, scalabilità, persistenza, flessibilità (Rosati, 2003)

La costruzione del thesaurus, ancora in corso, ha quindi richiesto un'analisi del dominio di interesse, anche a partire dai termini estratti, per la definizione delle faccette oltreché di criteri di suddivisione, identificati a partire dalle proprietà intrinseche dei concetti e introdotti sottoforma di etichette di snodo per l'ulteriore organizzazione dei termini all'interno di ciascuna faccetta. La figura seguente riporta un esempio di tale organizzazione:

[10] "[…] though the stop word list can be adjusted by the user, it seems that even after eliminating the stop words, the most frequent words from a specialized corpus are not all true terms but include many general words used across a wide range of subjects". (Chung, Nation, 2004)

[11] L'approccio della classificazione a faccette viene detto analitico-sintetico, poiché a seguito dell'identificazione delle faccette e delle loro suddivisione, le stringhe di soggetto possono essere costruite liberamente attraverso delle regole sintattiche per la combinazione dei termini.



La definizione dello schema di classificazione, la cui struttura attuale è ancora suscettibile di subire delle modifiche a seguito del continuo inserimento di nuovi termini, è avvenuta attraverso l'adozione complementare di un approccio deduttivo, secondo il quale solo a partire dall'analisi dei termini di cui si dispone è possibile definire una struttura classificatoria che li organizzi in maniera coerente, e di un approccio induttivo volto alla preliminare definizione delle categorie all'interno delle quali organizzare i termini (Broughton, 2008): mentre da un lato è importante poter far riferimento ad uno schema, anche se a livello prototipale, attraverso il quale organizzare i termini, dall'altro è necessario poter liberamente apportare modifiche di varia entità allo schema ideato per mezzo dei termini analizzati in momenti successivi.

Per quanto concerne il settore dell'efficienza energetica e delle energie rinnovabili, l'assenza di schemi di classificazione, anche in contesto internazionale, ha richiesto un'analisi approfondita del dominio di interesse e ha permesso una maggiore libertà organizzativa, mentre per i termini appartenenti al settore dell'edilizia, si è deciso di adattarsi il più possibile a quanto esiste nel panorama internazionale e ai tentativi di definizione di alcune macroaree promossi dall'UNI attraverso la norma precedentemente presentata. Da un punto di vista più prettamente metodologico, a supporto della definizione dello schema di faccette, si è fatto riferimento essenzialmente al set di categorie definito dal CRG (Classification Research Group) e al sistema di categorie/faccette adottato dal Nuovo Soggettario della Biblioteca Nazionale di Firenze (Biblioteca Nazionale Centrale di Firenze, 2006)

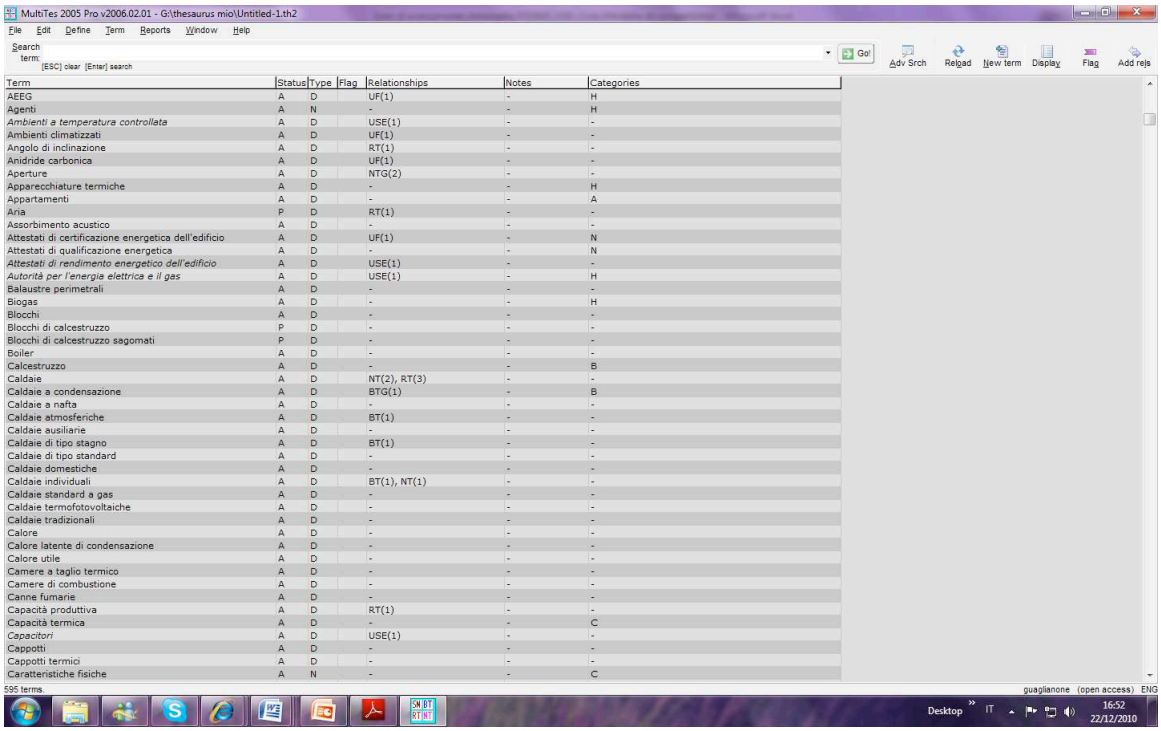
Il prospetto seguente illustra le scelte di classificazione adottate e permette di metterle a confronto da un lato con le categorie del Soggettario e con le faccette del CRG su un piano prettamente metodologico, dall'altro con la proposta di organizzazione della norma UNI 11337 e con lo schema classificatorio esemplificato nella norma ISO 12006-2, in qualità di fonti di riferimento per la strutturazione della conoscenza relativa al sottodominio dell'edilizia .

CRG[1]	Nuovo Soggettario	UNI 11337	ISO 12006-2	Schema Thesaurus
Cose	Cose Forme Materia Oggetti Spazio Strumenti Strutture	Opere: Edificio od opera di ingegneria civile o, comunque, il risultato di un insieme di lavori, che di per sè esplichi una funzione economica o tecnica [...]. Prodotto risultante della produzione edilizia e dell'ingegneria civile	Construction Result: construction object which is formed or changed in state as the result of one or more construction processes utilizing one or more construction resources Property/Characteristic: members of all the classes defined [...] have properties and characteristics [...]	Opere
Tipi				
Parti				Parti di opere
Proprietà				Proprietà
Materiali				Materiali
Processi				Processi
Attività	Azioni Attività Discipline Processi	Attività: aggregazione organizzata di una o più risorse in termini di lavori, forniture e servizi.	Construction Process: process which transforms construction resources into construction results	Attività
Pazienti				
Prodotti				Prodotti da costruzione (v. Parti di Opere)
Prodotti intermedi				Semilavorati
Agenti/Strumenti	Agenti Organismi Organizzaz. Persone e gruppi	Risorse: qualsiasi soggetto, oggetto o azione che costituisce fattore produttivo in un lavoro, una fornitura od un servizio.	Construction Resource: construction object used in a construction process to achieve a construction result	Risorse
Spazio				Elementi spaziali (v. Parti di Opere)
Tempo				+ Forme di energia, Tecnologie, Documenti, Discipline

L'intervento degli esperti di dominio si rende necessario anche a supporto della scelta e dell'organizzazione dei termini e della validazione di scelte terminologiche e di classificazione.

Per quanto riguarda il software per la costruzione del thesaurus, la scelta è ricaduta sul tool Multites[12], che risponde ad un insieme di requisiti fondamentali per garantire una corretta gestione di tale strumento. Si tratta di un software non specificatamente concepito per la creazione di thesauri a faccette, ma l'analisi delle funzionalità di cui lo stesso dispone ha permesso di adattarne e personalizzarne l'utilizzo finalizzato alla creazione della struttura prevista per la soluzione che si intende ottenere. Di seguito un print screen relativo all'utilizzo di Multites:

[12] <www.multites.com>



Term	Status	Type	Flag	Relationships	Notes	Categories
AEEG	A	D		UF(1)	-	H
Agenti	A	N		-	-	H
Ambienti a temperatura controllata	A	D		USE(1)	-	-
Ambienti climatizzati	A	D		UF(1)	-	-
Angolo di inclinazione	A	D		RT(1)	-	-
Anidride carbonica	A	D		UF(1)	-	-
Aperture	A	D		NTG(2)	-	-
Apparecchiature termiche	A	D		-	-	H
Appartamenti	A	D		-	-	A
Aria	P	D		RT(1)	-	-
Assorbimento acustico	A	D		-	-	-
Attestati di certificazione energetica dell'edificio	A	D		UF(1)	-	N
Attestati di qualificazione energetica	A	D		-	-	N
Attestati di rendimento energetico dell'edificio	A	D		USE(1)	-	-
Autorità per l'energia elettrica e il gas	A	D		USE(1)	-	H
Balaustre perimetrali	A	D		-	-	-
Biogas	A	D		-	-	H
Blocchi	A	D		-	-	-
Blocchi di calcestruzzo	P	D		-	-	-
Blocchi di calcestruzzo sagomati	P	D		-	-	-
Boiler	A	D		-	-	-
Calcestruzzo	A	D		-	-	B
Caldaie	A	D		NT(2), RT(3)	-	-
Caldaie a condensazione	A	D		BTG(1)	-	B
Caldaie a nafta	A	D		-	-	-
Caldaie atmosferiche	A	D		BT(1)	-	-
Caldaie ausiliarie	A	D		-	-	-
Caldaie di tipo stagno	A	D		BT(1)	-	-
Caldaie di tipo standard	A	D		-	-	-
Caldaie domestiche	A	D		-	-	-
Caldaie individuali	A	D		BT(1), NT(1)	-	-
Caldaie standard a gas	A	D		-	-	-
Caldaie termofotovoltaiche	A	D		-	-	-
Caldaie tradizionali	A	D		-	-	-
Calore	A	D		-	-	-
Calore latente di condensazione	A	D		-	-	-
Calore utile	A	D		-	-	-
Camere a taglio termico	A	D		-	-	-
Camere di combustione	A	D		-	-	-
Canne fumarie	A	D		-	-	-
Capacità produttiva	A	D		RT(1)	-	-
Capacità termica	A	D		-	-	C
Capacitori	A	D		USE(1)	-	-
Cappotti	A	D		-	-	-
Cappotti termici	A	D		-	-	-
Caratteristiche fisiche	A	N		-	-	C

4 Aspetti terminologici

Nella presente sezione vengono illustrati gli aspetti terminologici più significativi emersi dall'analisi dei risultati dell'estrazione terminologica.

Da un primo confronto tra i glossari terminologici relativi ai sottocorpora costituiti dagli articoli scientifici e dagli opuscoli, due tipologie documentali che presentano significative differenze in termini di obiettivi, struttura, pubblico di riferimento, livello di linguaggio impiegato, ecc., è emerso quanto già sostenuto da (Zanola, 2008) nel contributo relativo all'analisi della terminologia del dominio delle energie rinnovabili. In questo articolo si sostiene che molti termini propri del lessico dell'energia non appartengono più in maniera esclusiva al linguaggio specialistico, ma entrano progressivamente a far parte del linguaggio comune, come conseguenza probabilmente di un massiccio fenomeno di divulgazione.

Tra i termini oggetto di tale fenomeno, che, nel caso specifico, non si riferiscono esclusivamente alle energie rinnovabili, ma anche agli altri sottodomini di interesse, rientrano ad esempio: moduli fotovoltaici, coibentazione, collettori solari, ecc.

Tuttavia, la constatazione di tale fenomeno non impedisce che vi siano alcune importanti differenze tra il lessico specialistico e il linguaggio comune:

- Sinonimia o quasi sinonimia:
    - Radiatore e termosifone appartengono rispettivamente al linguaggio specialistico e a quello comune, così come coibentazione e isolamento termico;
    - Pompe di calore e raffrescatori appartengono al lessico specialistico e identificano oggetti aventi funzione diversa, ma principi di funzionamento simili. Nel linguaggio comune è frequente utilizzare un unico termine, condizionatori, per far riferimento ad entrambe le entità, determinando, quindi, una generalizzazione

concettuale. Il contrario accade invece con il termine climatizzazione[13]: nel linguaggio comune tale termine viene associato solo agli impianti di raffrescamento, mentre, da quanto si evince dalla norma UNI 10339-1995, nel linguaggio specialistico tale termine si riferisce anche al processo di riscaldamento. In questo caso il linguaggio comune restringe il significato del concetto.

Di seguito sono presentati alcuni dei fenomeni che caratterizzano il lessico del dominio di interesse:

- Terminologizzazione: "processo per cui un termine o un'espressione di uso generale o comune viene trasformato in un termine che designa un concetto particolare in un linguaggio speciale"[14] , definizione ripresa in (Bertaccini et al., 2008). Nell'esempio di seguito riportato il termine cappotto, appartenente alla lingua comune ha subito tale processo acquisendo un significato speciale nel lessico dell'edilizia. Tuttavia, si può notare come nel passaggio dal linguaggio comune al linguaggio specialistico, il termine abbia conservato alcuni dei suoi tratti semantici originari:

    - Cappotto : Il sistema "a cappotto" serve per isolare in modo sicuro e continuo pareti costituite anche da materiali diversi[15];

- Elevata presenza di unità polirematiche formate spesso da più di due elementi: l'elevata presenza di termini composti è un fenomeno peculiare dei lessici specialistici, probabilmente per la necessità di esprimere concetti complessi attraverso la specializzazione spesso di unità monorematiche appartenenti al linguaggio quotidiano;

- Presenza di termini che appartengono ai lessici specialistici di diverse discipline, quali l'edilizia e le energie rinnovabili in coesistenza con termini relativamente nuovi legati alla certificazione energetica, alla domotica, alla sostenibilità ambientale:

    - Edilizia: involucro edilizio, coperture, serramenti, ecc.

    - Energia fotovoltaica: pannelli fotovoltaici, moduli fotovoltaici, impianti solari, ecc.

    - Certificazione energetica: classi di efficienza energetica, diagnosi energetica, targa energetica, ecc.

- Considerevole influenza della lingua inglese:

    - Presenza di Acronimi di termini inglesi accanto a termini italiani

        - Microcogenerazione o MCHP (Micro Combined Heat and Power), Fotovoltaico o PV (photovoltaic)

---

[13] "La climatizzazione consiste nel controllo simultaneo e interdipendente di quattro condizioni ambientali fondamentali: la temperatura, l'umidità, il movimento e la qualità dell'aria".
APAT (Agenzia per la protezione dell'ambiente e per i servizi tecnici), Climatizzazione e ventilazione, impianti di climatizzazione e qualità dell'aria interna
<http://www.indoor.apat.gov.it/site/_contentfiles/00000000/52_Climatizzazione%20e%20ventilazione.pdf>

[14] <http://www.assiterm91.org/it/index.php?option=com_content&task=view&id=29&Itemid=25>

[15] <http://www.edilportale.com/upload/prodotti/prodotti-3632-cat01.pdf>

- Coesistenza di acronimi di termini inglesi e italiani:
  - SGV - Sonda geotermica verticale / BHE - Borehole Heat Exchanger
- Prestiti, anche integrali:
  - Vetro float, boiler, impianto grid connected, impianto stand-alone, Solar cooling

Rispetto all'inglese è interessante notare come anche in settori specialistici, quale quello dell'edilizia, l'ambiguità della lingua italiana sia molto elevata. L'esempio seguente mostra come al termine italiano "rivestimento" corrispondano quattro termini in inglese, con variabilità inferiore:

Rivestimento: materiale che ricopre una superficie per decorazione, protezione, isolamento, ecc. (cappa, foglio, mantello, membrana, telo)	**Covering**
Rivestimento: copertura non portante delle mura di un edificio (pannello di r., piastra di r., strato di r.)	**Cladding, Facing**
Rivestimento: isolante per caldaia	**Lagging, Racket**
Rivestimento: copertura di un tetto di un edificio	**Roof cladding, Roof covering**

5 Conclusioni e prospettive

Con la predisposizione dei prodotti terminologici oggetto del presente intervento ci si propone di fornire un tentativo di organizzazione dei termini relativi al dominio di riferimento, oltre che, più specificatamente, di garantire un accesso più strutturato all'informazione contenuta nel centro di documentazione, che rappresenta un caso di studio concreto per la loro sperimentazione e per il loro utilizzo.

In riferimento a tale applicazione, ci si propone di passare nel breve periodo dalla versione prototipale a quella definitiva con l'implementazione di ulteriori funzionalità di gestione documentale, mentre possibili prospettive di ricerca riguardano la definizione di tecniche di indicizzazione e di classificazione automatica dei documenti anche attraverso una fase di estrazione terminologica tematica che recuperi solo i termini rappresentativi del contenuto concettuale dei documenti e al tempo stesso del dominio di riferimento.

Di particolare interesse anche la prospettiva di multilinguismo finalizzata alla predisposizione di strumenti terminologici in lingue diverse dall'italiano, che non si limiti alla semplice traduzione, ma valuti le differenze concettuali legate ai contesti nazionali coinvolti.

Riferimenti bibliografici

BERTACCINI, F., et al., Processi di terminologizzazione e determinologizzazione nel dominio della diffusione e distribuzione del libro, «Aida Informazioni», n. 1-2, anno 26, pp. 47-61, gennaio-giugno 2008.

BIBLIOTECA NAZIONALE CENTRALE DI FIRENZE, Nuovo Soggettario, Editrice Bibliografica, Milano, 2006.

BROUGHTON, V., Costruire Thesauri: strumenti per indicizzazione e metadati semantici, (a cura di) P. CAVALERI, (traduzione di) L. BALLESTRA e L. VENUTI, Milano, EditriceBibliografica, 2008.

CHUNG, T.M., NATION, P., Identifying technical vocabulary, «System», n.32, Elsevier, pp. 251-263, 2004.

DELL'ORLETTA, F., et al., Dal testo alla conoscenza e ritorno, «Aida Informazioni», n. 1-2, anno 26, pp. 185-206, gennaio-giugno 2008.

FOLINO A., OLIVERI E., SCAIOLI R., Il Centro regionale di documentazione scientifica della Lombardia: un thesaurus e un lessico specialistico nel dominio dell'efficienza energetica degli edifici come tassonomie di accesso, Contributo alla Giornata Realiter 2009- V Giornata scientifica: Terminologia e plurilinguismo nell'economia internazionale, Milano, 9 Giugno 2009.

INTERNATIONAL NUCLEAR INFORMATION SYSTEM (INIS), ENERGY TECHNOLOGY DATA EXCHANGE (ETDE) JOINT Thesaurus, Part I + II, International Atomic Energy Agency (IAEA), Vienna, 2007

ISO 12006-2: 2001 Building construction - Organization of information about construction works – Part 2: Framework for classification of information, 2001.

ISO 25964-1: 2009, Information and documentation — Thesauri and interoperability with other vocabularies — Part 1: Thesauri for information retrieval, 2009.

MCENERY, T., WILSON, A., Corpus Linguistics, «Edinburgh Textbooks in Empirical Linguistics», Edinburgh University Press, Edimburgo, 2001.

RANGANATHAN, S.R., The Five Laws of Library Science, 2nd ed., Sarada Ranganathan Endowment for Library Science, Bangalore, India, 1988.

ROSATI, L., La classificazione a faccette fra Knowledge Management et Information Architecture (parte I), It Consult, 2003.
<http://www.itconsult.it/knowledge/articoli/pdf/itc_rosati_faccette_e_KM.pdf>

UNI 11337:2009, Edilizia e opere di ingegneria civile - Criteri di codificazione di opere e prodotti da costruzione, attività e risorse - Identificazione, descrizione e interoperabilità, 2009.

UNI 10339:1995, Impianti aeraulici a fini di benessere - Generalità, classificazione e requisiti. Regole per la richiesta d'offerta, l'offerta, l'ordine e la fornitura, 1995.

ZANOLA, M.T, Energie tradizionali e rinnovabili:proposte di interventi terminologici, Atti del Convengo Nazionale Ass.I.Term. I-TerAnDo, Università degli Studi della Calabria 5-7 giugno 2008.

<http://www.eionet.europa.eu/gemet>

<www.multites.com>

<http://www.indoor.apat.gov.it/site/_contentfiles/00000000/52_Climatizzazione%20e%20ventilazione.pdf>

<http://www.assiterm91.org/it/index.php?option=com_content&task=view&id=29&Itemid=25>

<http://www.edilportale.com/upload/prodotti/prodotti-3632-cat01.pdf>

<http://dublincore.org/>