10.

Elaborazione e gestione di (meta) dati terminologici

Federica Vezzani*, Giorgio Maria Di Nunzio**
* Dipartimento di Studi Linguistici e Letterari,
Università di Padova
** Dipartimento di Ingegneria dell'Informazione
e Dipartimento di Matematica, Università di Padova

Abstract: The optimal organisation of terminological (meta)data is an indispensable practice in the design and implementation of language resources. In this paper, we describe a methodology for the structural standardisation of terminological resources based on the application of *de jure* standards developed by the ISO/TC 37/SC 3 in order to ensure the FAIRness of terminological data. In this regard, we describe a project, recently launched by the University of Padua, which adopts the proposed paradigm in order to create the CAMEO multilingual terminological database for the commercial domain. This resource aims to be a valid standardised linguistic support for two categories of text professionals (technical communicators and specialised translators) dealing with monolingual and multilingual commercial product documentation.
Keywords: terminological resources, FAIRness, structural standardisation, ISO standards, commercial terminology

1. Introduzione

L'implementazione e la messa a disposizione di una banca dati terminologica sono attività che richiedono un grande sforzo di progettazione al fine di organizzare e gestire in modo ottimale i (meta)dati terminologici.

> Gérer les données, c'est s'assurer que celles-ci sont correctement sélectionnées, décrites, préservées et rendues accessibles pour un traitement et/ou une réutilisation, et ce, bien au-delà du projet de recherche qui les a fait naître et les a exploitées au premier chef.
> (Calderan e Millet 2015: 92)

Calderan e Millet (2015) definiscono in questo modo l'insieme dei compiti coinvolti nel processo di *data curation*. Questa nozione identifica, in generale, l'insieme delle buone pratiche per l'organizzazione ottimale dei dati della ricerca (Palmer *et al*. 2013), la cui responsabilità ricade inevitabilmente sul ricercatore che li produce (McLure *et al*. 2014, Corti *et al*. 2019). La sensibilizzazione sull'importanza della cura dei dati ha condotto alla nascita di numerose linee guida e norme dettagliate che regolano le azioni necessarie (De Matos *et al*. 2004, Eaker 2016, Erkimbaev *et al*. 2019). In questo senso, il rispetto di questi requisiti può garantire non solo la correttezza ma anche il continuo arricchimento del valore qualitativo dei dati scientifici. A questo proposito, una serie di linee guida è stata pubblicata da Wilkinson *et al*. (2016) nel quadro della piattaforma europea European Open Science Cloud[1] (EOSC) per promuovere la *FAIRness* dei dati della ricerca. Queste linee guida sottolineano la necessità di mettere a disposizione dati trovabili, accessibili, interoperabili e riutilizzabili[2] (*Findable, Accessible, Interoperable, Reusable*). I principi FAIR si riferiscono a tre tipi di entità: il dato (l'oggetto digitale d'interesse), i metadati (le informazioni sull'oggetto digitale) e le infrastrutture. Tutte le componenti del processo di ricerca dovrebbero beneficiare,

1 https://eosc-portal.eu (consultato il 25/01/2022)
2 https://www.go-fair.org/fair-principles/ (consultato il 25/01/2022)

quindi, dell'applicazione di queste linee guida al fine di garantirne la loro trasparenza, riproducibilità e riusabilità. La gestione ottimale dei dati non si configura dunque come un obiettivo in sé, ma rappresenta piuttosto il mezzo primario che porta alla scoperta della conoscenza e all'innovazione, così come all'integrazione e al riutilizzo dei dati da parte della comunità scientifica.
Nell'ambito dell'attività terminografica, l'Organizzazione internazionale per la standardizzazione[3] (ISO) e, in particolare, il Comitato tecnico ISO/TC 37[4] (Lingua e Terminologia) forniscono le norme specifiche per la progettazione e la realizzazione di risorse linguistiche e terminologiche strutturalmente omogenee. Tuttavia, in questo contesto, i dati della ricerca sono ancora lontani dall'essere FAIR (Forkel *et al*. 2018). Le risorse linguistiche sono spesso codificate in un formato eterogeneo e sviluppate in modo isolato le une dalle altre (Cimiano *et al*. 2020), con il rischio di rendere la loro scoperta, riutilizzo e integrazione un compito difficile e macchinoso. In questo senso, si notino gli sforzi compiuti dall'infrastruttura di ricerca europea Common Language Resources and Technology Infrastructure[5] (CLARIN), che consente ai ricercatori in scienze sociali e umane di accedere alle risorse e alle tecnologie linguistiche disponibili a livello europeo e mira a fornire un'architettura di dati conforme ai principi FAIR (De Jong *et al*. 2018).
In questo panorama, il nostro interesse è volto verso l'analisi dell'eterogeneità delle risorse terminologiche disponibili per il dominio commerciale, alla luce dell'esigenza conclamata da parte delle aziende produttrici di disporre di strumenti linguistici strutturati al fine di gestire in modo ottimale e coerente la terminologia relativa al prodotto (Warburton 2015). Nella fattispecie, delimitiamo il campo di indagine commerciale a quattro settori specifici identificati sulla base di statistiche nazionali. Il nostro caso di studio è rappresentato dalle risorse linguistiche disponibili per la documentazione della terminologia impiegata per le aree di maggior produttività, in termini di esportazione, della regione Veneto. Secondo i dati forniti da *Il Sole 24 Ore* in un articolo del 2018 intitolato "Al Veneto il primato delle esportazioni" (Sole 24 Ore 2018), le imprese locali sono responsabili di circa il 13,7% in valore di tutte le esportazioni italiane. In particolare, con riferimento alle stime del Sistema Statistico Regionale (2019) del triennio 2017-2019, il Veneto si configura come una delle regioni leader nei settori: agroalimentare, tessile, conciario e del vetro.
Il nostro oggetto d'indagine è dunque circoscritto alle risorse terminologiche messe a disposizione per questi quattro sottodomini commerciali. L'analisi condotta mira, in primo luogo, a osservare la qualità di questi strumenti in termini di normalizzazione strutturale e FAIRness dei dati terminologici. Inoltre, valutiamo anche il contenuto di queste risorse in termini di ricchezza di informazioni linguistiche fornite per due categorie di professionisti del testo. In particolare, le risorse oggetto di indagine si configurano come potenziali strumenti a supporto di 1) redattori / comunicatori tecnici che si occupano della descrizione del prodotto e 2) traduttori specializzati incaricati della traduzione della documentazione relativa al prodotto al fine di diffonderne la commercializzazione all'estero. A partire da questi due versanti di analisi, questo contributo porta sulla descrizione di un recente progetto avviato dall'Università di Padova che mira all'implementazione della risorsa CAMEO[6] (CommerciAl terMinology rEsOurce) per il dominio commerciale. In particolare, il nostro obiettivo principale consiste nella definizione di una metodologia di normalizzazione strutturale delle risorse terminologiche basata sull'applicazione di standard *de jure* prodotti in seno al comitato ISO/TC 37/SC 3[7] al fine di assicurare la FAIRness dei (meta)dati

3 https://www.iso.org/home.html (consultato il 25/01/2022)
4 https://www.iso.org/committee/48104/x/catalogue/ (consultato il 25/01/2022)
5 https://www.clarin.eu (consultato il 25/01/2022)
6 https://purl.org/cameo (consultato il 25/01/2022)
7 https://www.iso.org/committee/48136/x/catalogue/ (consultato il 25/01/2022)

della ricerca terminologica (Vezzani e Di Nunzio 2020a, 2020b). A questo proposito, descriviamo l'applicazione di questa metodologia alla banca dati terminologica multilingue CAMEO, la quale mira a configurarsi come un valido supporto linguistico per i professionisti del testo che si occupano della documentazione monolingue e multilingue del prodotto.
Il presente articolo è organizzato come segue: nella sezione 2 forniamo un panorama delle risorse attualmente disponibili per i settori identificati e valutiamo la loro qualità in termini strutturali e di ricchezza dei dati terminologici forniti. La sezione 3 è dedicata alla presentazione del progetto CAMEO e mira a descrivere la metodologia adottata per l'organizzazione e la gestione dei (meta)dati terminologici della risorsa multilingue. In particolare, ne descriviamo il meta-modello strutturale (sezione 3.1), il modello di scheda terminologica adottato, le sue categorie di dati e il repertorio che le contiene (sezione 3.2) e il formato di implementazione del database (sezione 3.3). La sezione 3.4 sarà dedicata a una descrizione sullo stato attuale della risorsa, sui corpora di lavoro e sugli strumenti di estrazione terminologica impiegati. Infine, nella sezione 4 forniamo le nostre considerazioni finali e le prospettive future.

2. Risorse terminologiche: il dominio commerciale

Numerosi studi sono stati condotti al fine di documentare le proprietà terminologiche delle quattro attività commerciali oggetto di indagine: agroalimentare, tessile, conciario e vetro (Moretti 2002, Cardillo *et al.* 2011, Chessa *et al.* 2015). Abbiamo condotto un'indagine volta a osservare – dal punto di vista strutturale e contenutistico – gli strumenti linguistici attualmente disponibili in rete per questi sottodomini commerciali visto il nostro interesse sugli aspetti implementativi di risorse terminologiche. La piattaforma Lexicool[8] ci permette, in questo senso, di raccogliere una lista delle principali risorse multilingui disponibili a livello internazionale e ordinate per categorie, quali: cibo[9], bevande[10], settore tessile[11] (abbigliamento e cuoio) e vetro[12]. Spostandoci sul versante nazionale e, in particolare, sul territorio regionale veneto, esistono numerose risorse che si presentano sotto forma di glossari monolingui e multilingui disponibili sui siti delle aziende produttrici. Per citare qualche esempio:

8 https://www.lexicool.com/dizionari-online-per-categoria.asp (consultato il 25/01/2022)
9 https://www.lexicool.com/dizionario-online.asp?FSP=C18&FKW=cibo (consultato il 25/01/2022)
10 https://www.lexicool.com/dizionario-online.asp?FSP=C18&FKW=bevande (consultato il 25/01/2022)
11 https://www.lexicool.com/dizionario-online.asp?FSP=C30&FKW=tessile (consultato il 25/01/2022)
12 https://www.lexicool.com/dizionario-online.asp?FSP=C02 (consultato il 25/01/2022)

— l'azienda Follador propone un glossario del vino monolingue consultabile online[13];
— l'azienda regionale per i settori agricolo, forestale e agroalimentare, Veneto Agricoltura, fornisce un vocabolario in lingua italiana, inglese e tedesca (Gartner *et al.* 2012);
— il sito *Venezia e le sue lagune*[14] mette a disposizione un glossario multilingue (italiano, francese, spagnolo, inglese) su tematiche locali eterogenee;
— l'azienda Conceria La Veneta[15] fornisce una lista di termini in italiano[16];
— per la fabbricazione del vetro, la Regione del Veneto mette a disposizione sul suo portale una lista dei termini tecnici in lingua italiana[17].

In termini di fruibilità, queste risorse si configurano come potenziali strumenti di supporto linguistico per due categorie specifiche di professioni del testo: 1) redattori / comunicatori tecnici e 2) traduttori specializzati. I prodotti manifatturieri sono generalmente accompagnati da una descrizione delle loro proprietà e funzioni. La scrittura di questa tipologia testuale rientra nell'attività di redazione della documentazione tecnica, come i manuali d'uso, le schede tecniche, l'etichettatura (Muzii 1995). In questo contesto, si fa spazio la figura del redattore / comunicatore tecnico al quale viene assegnato il compito di "spiegare" un prodotto nella sua interezza. L'Associazione italiana per la comunicazione tecnica Com&Tec[18] si fa portavoce, a livello nazionale, di numerose iniziative finalizzate a favorire e supportare la formazione di questa figura professionale. Tra le sue competenze linguistico-culturali rientrano un'ottima conoscenza della terminologia tecnica e una buona padronanza delle lingue. In questo senso, non solo la redazione ma anche la traduzione tecnica acquisisce un ruolo fondamentale per la diffusione e l'internazionalizzazione del prodotto commerciale.
Senza la pretesa di completezza, la raccolta delle risorse sopracitate ci permette di constatare che questi strumenti linguistici presentano dei limiti in termini di ricchezza e normalizzazione strutturale dei dati terminologici forniti. I glossari, infatti, raccolgono informazioni minime (per lo più termine e definizione) che risultano insufficienti nel sostenere il lavoro di redazione e/o traduzione di testi specialistici[19]. Al fine di decodificare e, successivamente, transcodificare le informazioni trasmesse nella documentazione tecnica di un prodotto, il professionista del testo necessita di una panoramica completa del comportamento morfologico, sintattico, semantico e fraseologico del termine (Scarpa 2001). In questa prospettiva si colloca l'iniziativa del Centro di Ricerca in Terminologia Multilingue dell'Università di Genova[20] (CeRTeM) che porta alla realiz-

13 https://www.folladorprosecco.com/storia-del-prosecco/glossario-del-vino-e-terminologia/ (consultato il 25/01/2022)
14 https://www.venicethefuture.com/schede/data/it/glossario.htm (consultato il 25/01/2022)
15 http://www.concerialaveneta.com/IMG/pdf/Glossario-industria-conciaria.pdf (consultato il 06/03/2021)
16 Per il settore pelletteria, si noti anche il glossario italiano-inglese "Il lessico della pelle" edito da UNIC – Concerie italiane e disponibile al link: https://www.lineapelle-fair.it/it/lineapelle-training/lessico-della-pelle (consultato il 25/01/2022).
17 http://www2.regione.veneto.it/cultura/itinerari/vetri-antichi/glossario/termini.htm (consultato il 25/01/2022)
18 https://www.comtec-italia.org (consultato il 25/01/2022)
19 A questo proposito, sottolineiamo che anche le risorse di più ampia diffusione e, di conseguenza, non specifiche per i quattro domini selezionati, come la banca dati terminologica europea IATE, presentano limiti in termini di quantità dei dati forniti. Consultando la scheda del termine "vino", ad esempio, si nota la presenza del solo dato relativo alla definizione.
20 http://www.lcm.unige.it/certem/ (consultato il 25/01/2022)

zazione di un glossario enologico[21], il quale si configura, a livello nazionale, come la risorsa più completa in termini di ricchezza dei dati e di varietà linguistica contenente la terminologia del vino come prodotto economico e culturale legato al territorio della Liguria e del Basso Piemonte (Piccardo 2020). La risorsa si presenta come una collezione di schede terminologiche multilingue che forniscono informazioni sulla definizione, contesto, trascrizione fonetica, etimologia, sinonimi / antonimi del termine e altri dati di tipo amministrativo come l'autore e la data di creazione della scheda. Rispetto ai glossari precedenti, il vantaggio evidente di questa risorsa è la ricchezza delle informazioni fornite necessarie al redattore e al traduttore tecnico al fine di contestualizzare il termine e disambiguarlo.
Da un punto di vista puramente strutturale, si può constatare, tuttavia, che tutte le risorse precedentemente citate non aderiscono agli standard internazionali ISO attualmente in vigore per la gestione efficace e la diffusione ottimale della terminologia. Benché perfettamente consultabili online, i dati terminologici forniti non risultano riutilizzabili automaticamente, ad esempio, in sistemi e/o applicazioni di traduzione assistita ostacolando, di conseguenza, la qualità del lavoro dei professionisti del testo (Naldi 2014). A questo proposito, l'ISO/TC 37/SC 3, in qualità di comitato tecnico per la gestione di risorse terminologiche, promuove l'adozione dello standard ISO 30042:2019 secondo il quale tutti i dati terminologici dovrebbero essere organizzati secondo il formato Term Base eXchange (TBX) al fine di garantirne l'interoperabilità e il riutilizzo. Questo formato rappresenta, infatti, la strutturazione di riferimento utilizzata dalla maggior parte dei software di gestione della terminologia e di traduzione assistita (Bowker 2015). Inoltre, per assicurare l'uniformità delle stesse categorie di dati contenute nelle risorse terminologiche, il comitato ISO/TC 37/SC 3 promuove l'adozione dello standard ISO 12620:2019 che mira all'armonizzazione della loro documentazione tramite la messa a disposizione di repertori di categorie di dati, come DatCatInfo[22]. La non conformità agli standard terminologici ISO in vigore per la gestione di risorse linguistiche ostacola, in questo senso, la FAIRness dei dati terminologici.

3. Progetto CAMEO: metodologia di organizzazione dei (meta)dati terminologici

Dall'analisi dello stato attuale è nata l'esigenza di formulare un progetto di ricerca[23] che prevede la progettazione e successiva implementazione di una risorsa multilingue standardizzata al fine di raccogliere e documentare in modo esaustivo la terminologia tecnica per i quattro domini identificati. In particolare, la risorsa è pensata per rispondere ai bisogni informativi degli addetti alla redazione e alla traduzione della documentazione commerciale. Sulla base del panorama

21 http://www.farum.it/glos_enol/intro.php (consultato il 06/03/2021)
22 http://datcatinfo.net (consultato il 25/01/2022)
23 https://purl.org/fairterm (consultato il 25/01/2022)

presentato, il progetto CAMEO mira ad apportare un contributo qualitativo in termini di:

1. Ricchezza: la risorsa proposta si distinguerà dalle altre attualmente disponibili in quanto progettata su un nuovo modello di scheda terminologica che fornirà informazioni su vari assi di analisi linguistica, quali l'aspetto formale, etimologico, semantico, fraseologico e pragmatico del termine, al fine di offrirne una panoramica completa a uso del fruitore dello strumento.
2. Standardizzazione: l'implementazione della banca dati terminologica seguirà gli ultimi standard ISO vigenti in materia di gestione terminologica. In particolare, sulla base delle direttive europee, il nostro obiettivo è la produzione di una risorsa trovabile, accessibile, interoperabile e riutilizzabile.

Il progetto, ufficialmente avviato nel gennaio 2021 dall'Università di Padova, vede la collaborazione tra il Dipartimento di Studi linguistici e letterari[24] (DiSLL) e il Dipartimento di Ingegneria dell'informazione[25] (DEI). La risorsa CAMEO che ne deriva è progettata e implementata secondo la metodologia strutturale descritta nelle sezioni 3.1-3.3. Il paradigma proposto si basa sull'applicazione degli standard *de jure* ISO/TC 37/SC 3 al fine di garantire la FAIRness dei (meta)dati della ricerca terminologica (Vezzani e Di Nunzio 2020a, 2020b).

3.1 META-MODELLO STRUTTURALE

La progettazione della struttura della risorsa CAMEO si basa sullo standard ISO 16642:2017 relativo al meta-modello strutturale Terminological Markup Framework (TMF). Questo standard internazionale promuove l'adozione di un modello comune per la rappresentazione di collezioni di dati terminologici in eXstensible Markup Language (XML) che dovrebbe essere uniformemente impiegato per facilitare l'interoperabilità, la condivisione e il riutilizzo dei dati. Il quadro TMF è strutturato su due livelli di astrazione. Il primo livello comporta una descrizione del meta-modello sottostante l'analisi, la progettazione e lo scambio di dati. Il meta-modello è quindi indipendente da qualsiasi implementazione o software specifico. Il secondo livello riguarda piuttosto le categorie di dati che possono essere associate ai vari livelli del meta-modello e che sono specifiche per ogni collezione terminologica. Il modello strutturale della risorsa CAMEO si basa dunque sul quadro TMF che adotta una prospettiva di tipo onomasiologico: un concetto è descritto in *n* lingue ed è designato da *m* termini per ogni lingua. In particolare, il modello strutturale che proponiamo si configura come segue:

— CAMEO è una collezione di dati terminologici contenente un numero qualsiasi di "schede terminologiche".
— Ogni scheda terminologica si riferisce a un unico concetto che può essere rappresentato in *n* "sezioni di lingua".
— Per ogni lingua, ci possono essere *m* "sezioni di termine" contenenti le unità lessicali che, per quella lingua specifica, designano il concetto.

24 https://www.disll.unipd.it/ (consultato il 25/01/2022)
25 https://www.dei.unipd.it/home-page (consultato il 25/01/2022)

— Ogni sezione di termine può contenere un numero qualsiasi di "sezioni di componenti del termine" che forniscono informazioni sulle parti di un termine come morfemi, fonemi o sillabe.

Le relazioni tra le quattro istanze del meta-modello strutturale (scheda terminologica, sezione di lingua, sezione di termine e sezione di componenti del termine) sono regolate da delle cardinalità, vale a dire dei valori che indicano il numero minimo e massimo (x, y) degli elementi delle istanze che sono in relazione tra di loro.
Nella figura 1 proponiamo il modello entità-associazione (Chen 1976) del quadro TMF per la rappresentazione di CAMEO, dove le entità (rettangoli) sono le istanze e le associazioni (rombi) sono le relazioni tra di loro. La collezione di dati terminologici contiene un numero qualsiasi di schede terminologiche (0, n). Una scheda terminologica deve contenere almeno una sezione di lingua (1, n). Una sezione di lingua deve contenere almeno una sezione di termine (1, n) e quest'ultima può contenere un numero qualsiasi di sezioni di componenti del termine (0, n). La gerarchia è assicurata dalle cardinalità (1,1) espresse tra i nodi del meta-modello TMF. Questa struttura adotta quindi il modello *hub and spoke* (Van Campenhoudt 2017) e distingue i livelli gerarchici (Romary 2001) ai quali possono essere associate le diverse categorie di dati: dati di tipo concettuale comuni a tutte le lingue; dati specifici della lingua di lavoro; dati specifici del termine oggetto di analisi.

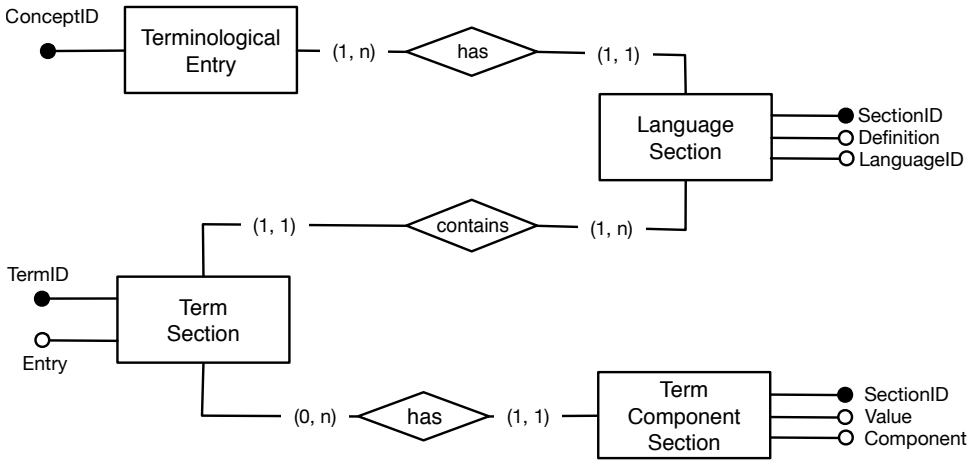

Figura 1: Meta-modello strutturale della risorsa CAMEO

Nella risorsa CAMEO, un concetto unico ed esclusivo per una scheda terminologica, definita dal suo identificatore, può essere espresso in *n* lingue. Le sezioni di lingua hanno un proprio identificatore e comprendono la definizione che esprime il concetto associato, così come il codice della lingua definito dalla norma ISO 639-1:2002. Per ogni lingua, ci sono *m* sezioni di termine contenenti ciascuna i termini (sinonimi tra di loro) che designano tale concetto e i quali sono corredati da un vasto numero di informazioni linguistiche. Infine, le descrizioni delle componenti del termine sono raggruppate nella sezione dedicata che costituisce l'ultima entità del modello.

3.2 CATEGORIE DI DATI TERMINOLOGICI

La risorsa CAMEO è progettata per rispondere in modo quanto più esaustivo ai bisogni informativi di redattori / comunicatori e traduttori tecnici incaricati della descrizione monolingue e/o multilingue del prodotto commerciale. Il modello di scheda terminologica multilingue che proponiamo mira, dunque, a fornire un panorama completo sul comportamento morfologico, sintattico, semantico e fraseologico del termine analizzato e dei suoi traducenti. Le lingue di lavoro attualmente contemplate nel progetto sono l'italiano, il francese, l'inglese e lo spagnolo[26]. La tabella 1 riassume le categorie di dati disponibili per ciascun termine nel modello di scheda terminologica progettato:

26 La struttura della scheda terminologica permette di includere, in futuro, altre lingue sulla base delle esigenze dei fruitori della risorsa terminologica. Nel breve termine, prevediamo l'inserimento delle seguenti lingue di lavoro: tedesco e russo.

Asse di analisi	Categorie di dati
Morfologia	Parte del discorso, genere grammaticale, numero grammaticale, forme derivate
Fonetica	Trascrizione IPA
Etimologia	Derivazione, composizione
Variazione	Variante ortografica, abbreviazione, forma estesa, acronimo
Semantica	Definizione, analisi semica, sinonimo, quasi-sinonimo, iponimo, iperonimo, meronimo, olonimo
Fraseologia	Unità fraseologica, collocazione
Pragmatica	Contesto d'utilizzo
Registro	Nome popolare, nome scientifico
Dominio	Dominio, sottodominio

Tabella 1: Categorie di dati in CAMEO

Inoltre, per garantire la tracciabilità dei dati forniti vengono riportate anche le categorie di dati relative alle fonti consultate[27] e alle informazioni circa l'autore e la data di creazione e modifica della scheda in oggetto.
In termini di standardizzazione FAIR, le stesse categorie di dati fornite all'interno della scheda terminologica sono soggette a normalizzazione strutturale. A questo proposito, lo standard ISO 12620:2019 descrive i meccanismi per documentare, armonizzare e gestire le categorie di dati contenute all'interno di una risorsa terminologica. In particolare, la norma sottolinea la necessità di aderire a delle specifiche di categorie di dati comuni per garantire la loro interoperabilità e riutilizzo. Per la loro armonizzazione, lo standard fa riferimento al repertorio DatCatInfo che raccoglie una lista documentata di tutte le categorie di dati disponibili. Inoltre, la norma stabilisce la possibilità per tutti gli sviluppatori di software di progettare il proprio repertorio specifico per una data risorsa terminologica. In questo senso, il progetto CAMEO prevede anche l'implementazione parallela di un repertorio contenente tutte le specifiche delle categorie di dati della risorsa. Al momento della redazione (settembre 2021), il repertorio è in fase di progettazione e fornisce, per ogni categoria di dato, le seguenti informazioni:

1. un *Persistent Unique Identifier* (PID), vale a dire un URL che fornisce un accesso web diretto alle specifiche delle categorie di dati;
2. un identificatore mnemonico unico e stabile;
3. il livello del meta-modello TMF (concetto, lingua e termine) al quale la categoria di dati è associata;
4. la tipologia di contenuto autorizzata[28];
5. una sua definizione chiara;

27 Come lavoro in corso, stiamo formalizzando delle linee guida per la gestione omogenea delle fonti.

28 Alcune categorie saranno di tipo "testo libero", come ad esempio la definizione del termine; altre categorie di dati, come la parte del discorso, saranno di tipo *picklist*.

6. altre spiegazioni e note informative;
7. alcuni esempi di utilizzo;
8. la traduzione del nome canonico della categoria di dati in tutte le lingue di lavoro della risorsa.

L'applicazione web relativa al repertorio verrà implementata utilizzando il pacchetto Shiny R (Chang 2015) e sarà liberamente disponibile online. La scelta di formulare il nostro repertorio di categorie di dati ci permette di disambiguare il significato delle categorie di dati e garantirne l'uniformità (Wright 2001). Gli utenti che consultano la risorsa possono quindi trovare tutte le informazioni necessarie per comprendere i dati forniti e le modalità di compilazione delle categorie. Infine, come suggerito dalla norma e per soddisfare la necessità di riutilizzo dei dati, l'utente potrà esportare le informazioni fornite nel repertorio nel formato Comma-Separated Values (CSV) ed eXtensible Markup Language (XML).

3.3 FORMATO DI IMPLEMENTAZIONE

Dopo aver definito il meta-modello strutturale della risorsa CAMEO, le categorie di dati che forniamo nel modello di scheda terminologica e le loro specifiche raccolte in un repertorio dedicato, descriviamo in questa sezione il formato di implementazione scelto per la nostra raccolta di dati terminologici. La sua implementazione si basa sullo standard ISO 30042:2019 relativo al formato TermBase eXchange (TBX). Questo documento definisce il quadro TBX, espresso nel linguaggio di marcatura XML, per l'analisi, la rappresentazione descrittiva e la diffusione di dati terminologici strutturati. In particolare, l'obiettivo principale del formato d'implementazione TBX è quello di garantire l'interoperabilità e la riusabilità dei dati terminologici in diverse applicazioni per la gestione della terminologia. Il formato TBX si basa su due componenti in interazione: una struttura di base che riflette il meta-modello TMF e un formalismo per definire i moduli TBX contenenti le categorie di dati contemplate nel database. La combinazione di questi due componenti definisce un dialetto specifico, vale a dire un linguaggio di marcatura XML conforme a TBX. Il sito web TBXinfo.net[29] raccomanda l'utilizzo di tre dialetti pubblici per lo scambio della terminologia:

— TBX-Core che comprende le categorie di dati: termine, data e note.
— TBX-Min che comprende, oltre alle categorie precedenti, le categorie di dati: parte del discorso e dominio.
— TBX-Basic che comprende, oltre alle categorie precedenti, le categorie di dati: contesto, definizione, riferimenti esterni, genere grammaticale, fonte, responsabilità e tipo di transazione[30].

Questi dialetti pubblici forniscono la documentazione di un insieme di categorie di dati che non copre l'integralità delle categorie previste nella scheda della risorsa CAMEO. In questo caso, la norma permette la formulazione di dialetti privati contenenti moduli specifici per la rappresentazione delle categorie di dati terminologici che non sono inclusi nei dialetti pubblici. In questo senso, per la documentazione di queste categorie di dati ci riferiamo a un dialetto privato

29 https://www.tbxinfo.net (consultato il 25/01/2022)
30 Per tipo di "transazione" intendiamo il tipo di operazione effettuato sulla scheda: ad esempio, "creazione" o "modifica".

denominato TBX-TriMED recentemente sviluppato dagli stessi autori, consultabile online[31] e attualmente sottoposto a validazione da parte del comitato di direzione TBX[32] (Vezzani e Di Nunzio 2020a, 2020b). Le schede terminologiche della risorsa CAMEO sono dunque progettate per essere consultabili e scaricabili in formato TBX direttamente dalla risorsa online. Questo permetterà sia ai redattori che ai traduttori tecnici di esportare e riutilizzare i dati terminologici relativi ai quattro sottodomini commerciali in differenti sistemi di gestione terminologica per la traduzione assistita e la redazione tecnica.

3.4 STATO ATTUALE DEL PROGETTO CAMEO

Come anticipato, il progetto CAMEO è stato ufficialmente avviato dall'Università di Padova nel gennaio 2021 e, al momento della redazione, la risorsa è sviluppata e messa a libera disposizione per consultazione. L'interfaccia web relativa al database è stata realizzata con il pacchetto Shiny R (Chang 2015) e permette la selezione della lingua di partenza e la lingua di arrivo tra le quattro lingue di lavoro del progetto. Come si può notare, le categorie di dati a corredo dei termini oggetto di analisi sono raggruppate in quattro sezioni per garantire una visualizzazione ottimale: caratteristiche formali, semantica, variazione e uso. I termini al momento contenuti nella risorsa sono pertinenti ai domini agroalimentare, cuoio e vetro e le informazioni linguistiche riportate all'interno del nuovo modello di scheda terminologica sono attualmente in fase di revisione e validazione da parte di esperti terminologi collaboratori del progetto.
Per quanto riguarda il popolamento della banca dati, i termini oggetto di analisi sono estratti semi-automaticamente tramite il software di gestione di corpora Sketch Engine (Kilgarriff *et al.* 2014) che costituisce lo strumento di supporto per la collezione del corpus di lavoro[33] sul quale stiamo conducendo la raccolta terminologica per i quattro domini commerciali. Trattandosi di settori identificati sulla base di indagini statistiche specifiche per la regione Veneto, il corpus di lavoro in lingua italiana è costituito da una collezione di documenti regionali per la descrizione dei prodotti locali maggiormente esportati all'estero e che necessitano, dunque, di una documentazione terminologica multilingue. In particolare, per la copertura terminologica del settore agroalimentare il corpus contiene la documentazione relativa ai prodotti alimentari e alle bevande leader di esportazione della regione. Nella fattispecie i documenti sono relativi a

1. la produzione di formaggi locali: la regione vanta una grande tradizione a livello caseario, tra cui otto Formaggi DOP (Asiago, Casatella Trevigiana, Grana Padano, Montasio, Monte Veronese, Provolone Valpadana, Taleggio, Piave)[34];

31 Il pacchetto di definizione del dialetto privato è disponibile al link del repertorio GitHub seguente: https://github.com/trimed-dialect-2020/ (consultato il 25/01/2022). Esso comprende la definizione del modulo *Trimed* in prosa, gli schemi del modulo e il formalismo TBXMD (*TBX Module Definition*).

32 https://www.tbxinfo.net/tbx-private-dialects/ (consultato il 25/01/2022)

33 In termini di grandezza, il corpus contiene il massimo delle parole (1 milione) messe a disposizione da Sketch Engine per gli utenti dell'Università di Padova.

34 https://www.regione.veneto.it/web/agricoltura-e-foreste/disciplinari-dop-igp-stg (consultato il 25/01/2022)

2. la produzione di vini locali: il Veneto vanta infatti ben 28 vini a denominazione di origine controllata (DOC) e 14 vini a denominazioni di origine controllata e garantita (DOCG), come ad esempio Soave, Valpolicella, Bardolino, Recioto, Prosecco[35].

La collezione di analisi per lo studio della terminologia tessile comprende i testi relativi alla preparazione, filatura, tessitura e finissaggio di fibre tessili, filati e tessuti, quali seta, cotone e lana. In particolare, la documentazione è raccolta a partire dalla Venice Textile Manufacturers[36] in qualità di rete di imprese venete per sostenere l'eccellenza del settore tessile locale. Per quanto riguarda la terminologia del cuoio, il corpus comprende la documentazione relativa alla: preparazione e concia del cuoio; fabbricazione di articoli da viaggio, borse, pelletteria e selleria; preparazione e tintura di pellicce; fabbricazione di calzature. Secondo i dati forniti dall'Unione Nazionale Industria Conciaria[37] (UNIC), il distretto veneto si configura infatti come uno dei maggiori distretti conciari del mondo, nonché il più importante in Italia per produzione e numero di addetti.
Infine, per lo studio della terminologia del vetro, il corpus raccoglie la documentazione relativa alla fabbricazione di vetro in tutte le sue forme (piano, cavo) e dei prodotti in vetro, ottenuti tramite qualsiasi processo. In particolare, i documenti sono pertinenti alla produzione del vetro artistico di Murano e del vetro del Veneziano, i quali rappresentano una vera e propria eccellenza locale capace di valere, da sola, circa un quarto di tutta la produzione del vetro artistico italiano, con una concentrazione non riscontrabile in nessun altro settore produttivo[38]. Questa documentazione costituisce dunque il corpus di lavoro sul quale stiamo attualmente conducendo il lavoro di collezione e analisi manuale della terminologia pertinente a uso di redattori / comunicatori e traduttori tecnici incaricati della descrizione dei prodotti commerciali.

4. Conclusioni

Sulla base di un'indagine preliminare volta a osservare l'eterogeneità strutturale e la ricchezza informativa delle risorse terminologiche attualmente disponibili per il dominio commerciale, in questo articolo abbiamo descritto il progetto CAMEO avente l'obiettivo di fornire una risorsa strutturalmente standardizzata e ricca di informazioni linguistiche per supportare il lavoro dei professionisti del testo commerciale. In particolare, abbiamo descritto una metodologia di normalizzazione strutturale basata sull'adozione degli ultimi standard ISO di riferimento in materia di gestione della terminologia al fine di rispondere ai principi FAIR per l'organizzazione ottimale di (meta)dati della ricerca terminologica. In questo senso, abbiamo presentato l'applicazione di questo paradigma alla risorsa CAMEO la quale:

35 https://www.regione.veneto.it/web/agricoltura-e-foreste/disciplinari-docg-doc-igt (consultato il 25/01/2022)

36 http://venicetextile.com (consultato il 25/01/2022)

37 https://unic.it (consultato il 25/01/2022)

38 https://www.regione.veneto.it/web/attivita-produttive/marchio-del-vetro-artistico-di-murano (consultato il 25/01/2022)

1. segue un meta-modello strutturale interoperabile (TMF);
2. permette l'accesso ai dati terminologici attraverso protocolli di comunicazione standard;
3. fornisce dei (meta)dati rigorosamente documentati, e quindi reperibili, attraverso un repertorio di categorie di dati;
4. garantisce il riutilizzo dei dati attraverso l'adozione del formato di riferimento TBX per lo scambio terminologico.

I passi presentati in questo lavoro coprono la maggior parte degli elementi chiave del processo di *FAIRification* dei dati, nel nostro caso dati terminologici. Sarebbe necessario un ultimo passo relativo alla definizione di un modello semantico (tipo Linguistic Linked Open Data[39]) per poter concludere definitivamente il lavoro. In questo senso, la nostra proposta contiene un modello concettuale (schema ER) che può essere utilizzato per ottenere il relativo schema a grafo e trasformare i dati ottenuti da XML a RDF in quanto modello di riferimento dei Linked Data.
Per concludere, tra le prospettive di lavoro futuro, auspichiamo l'arricchimento non solo delle lingue di lavoro attualmente contemplate nel database ma anche della terminologia analizzata tramite l'estensione dei sottodomini commerciali oggetto di indagine.

Bibliografia

— **Bowker, Lynne (2015)** "Terminology and translation". In Hendrik J. Kockaert e Frieda Steurs (a cura di) *Handbook of terminology*. Vol. 1. Amsterdam: John Benjamins, 304-323.

— **Calderan, Lisette e Millet, Jacques (a cura di) (2015)** *BIG DATA : nouvelles partitions de l'information. Actes du séminaire IST Inria, octobre 2014*. Louvain-la-Neuve: De Boeck Supérieur.

— **Cardillo, Elena, Folino, Antonietta, Guaglianone, Maria T., Iozzi, Francesca, e Taverniti, Maria (2011)** "Terminologia specialistica per l'artigianato orafo e tessile calabrese". In Maria Teresa Zanola e Maria Francesca Bonadonna (a cura di) *Terminologie specialistiche e prodotti terminologici"*. Milano: EDUCatt, 31-45.

— **Chang, Winston (2015)** "Shiny: Web Application Framework for R". In *R package version 0.11*.

— **Chen, Peter Pin-Shan (1976)** "The entity-relationship model—toward a unified view of data". In *ACM transactions on database systems (TODS)*, n. 1(1), 9-36. https://dl.acm.org/doi/pdf/10.1145/320434.320440 (consultato il 25/01/2022).

— **Chessa, Francesca, De Giovanni, Cosimo e Zanola, Maria Teresa (2015)** *La terminologia dell'agroalimentare*. Milano: FrancoAngeli.

39 https://linguistic-lod.org/ (consultato il 25/01/2022)

— **Cimiano, Philipp, Chiarcos, Christian, McCrae, John P. e Gracia, Jorge (2020)** *Linguistic Linked Data: Representation, Generation and Applications*. Cham: Springer.

— **Corti, Louise, Van den Eynden, Veerle, Bishop, Libby e Woollard, Matthew (2019)** *Managing and sharing research data: a guide to good practice*. Los Angeles: SAGE.

— **De Jong, Franciska, Maegaard, Bente, De Smedt, Koenraad, Fišer, Darja e Van Uytvanck, Dieter (2018)** "CLARIN: towards FAIR and responsible data science using language resources". In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation* (LREC 2018), 3259-3264. https://aclanthology.org/L18-1515.pdf (consultato il 25/01/2022).

— **De Matos, David M., Ribeiro, Ricardo e Mamede, Nuno J. (2004)** "Rethinking reusable resources". In *Proceedings of the Fourth International Conference on Language Resources and Evaluation* (LREC 2014), 357-360. http://www.lrec-conf.org/proceedings/lrec2004/pdf/395.pdf (consultato il 25/01/2022).

— **Eaker, Chris (2016)** "What could possibly go wrong? The impact of poor data management". In *The Medical Library Association's Guide to Data Management for Librarians*. Lanham (Maryland): Rowman and Littlefield.

— **Erkimbaev, Adilbek O., Zitserman, Vladimir Y., Kobzev, Georgii A. e Kosinov, Andrey V. (2019)** "Curation of digital scientific data". In *Scientific and Technical Information Processing*, n. 46 (3), 192-203.

— **Forkel, Robert, List, Johann-Mattis, Greenhill, Simon J., Rzymski, Christoph, Bank, Sebastian, Cysouw, Michael, Hammarström, Harald, Haspelmath, Martin, Kaiping, Gereon A. e Gray, Russell D. (2018)** "Crosslinguistic data formats, advancing data sharing and re-use in comparative linguistics". In *Scientific Data*, n. 5(1), 1-10.

— **Gartner, Erwin, Dianat, Katharina, Benvenuto, Luca, Castelluccio, Markus, Malossini, Giorgio, Schiavon, Luigino e Soligo, Stefano (2012)** *Le parole della vite e della frutta: piccolo vocabolario dei termini per operatori. Italiano – English – Deutsch*. Legnaro: Veneto Agricoltura. https://www.venetoagricoltura.org/upload/pubblicazioni/glossario%20interreg/Glossario%20Italiano.pdf (consultato il 25/01/2022).

— **ISO 639-1 (2002)** *Language codes*. Geneva: International Organization for Standardization.

— **ISO 12620 (2019)** *Management of terminology resources – Data category specifications*. Geneva: International Organization for Standardization.

— **ISO 16642 (2017)** *Computer applications in terminology – Terminological markup framework*. Geneva: International Organization for Standardization.

— **ISO 30042 (2019)** *Management of terminology resources – TermBase eXchange (TBX)*. Geneva: International Organization for Standardization.

— **Kilgarriff, Adam, Baisa, Vít, Bušta, Jan, Jakubíček, Miloš, Kovář, Vojtěch, Michelfeit, Jan, Rychlý, Pavel e Suchomel, Vít (2014)** "The Sketch Engine: ten years on". In *Lexicography,* n. 1(1), 7-36. https://link.springer.com/article/10.1007/s40607-014-0009-9 (consultato il 25/01/2022).

— **McLure, Merinda, Level, Allison V., Cranston, Catherine L., Oehlerts, Beth e Culbertson, Mike (2014)** "Data curation: a study of researcher practices and needs". In *Libraries and the Academy*, n. 14(2), 139-164.

— **Moretti, Cesare (2002)** *Glossario del vetro veneziano: dal Trecento al Novecento*. Venezia: Marsilio.

— **Muzii, Luigi (1995)** *La redazione dei documenti tecnici. Dalla progettazione alla realizzazione. Con esempi di documenti e istruzioni per l'uso degli strumenti informatici*. Milano: FrancoAngeli.

— **Naldi, Maurizio (2014)** *Traduzione automatica e traduzione assistita*. Bologna: Società Editrice Esculapio.

— **Palmer, Carole, Weber, Nicholas M., Renear, Allen H. e Muñoz, Trevor (2013)** "Foundations of data curation: The pedagogy and practice of 'purposeful work' with research data". In *Archives Journal*, n. 3. https://www.archivejournal.net/essays/foundations-of-data-curation-the-pedagogy-and-practice-of-purposeful-work-with-research-data/ (cosultato il 25/01/2022).

— **Piccardo, Giuseppina (2020)** "CeRTeM - Centro di Ricerca in Terminologia Multilingue dell'Università di Genova". In *Publifarum*, n. 12. https://www.publifarum.farum.it/index.php/publifarum/article/view/328 (consultato il 25/01/2022).

— **Romary, Laurent (2001)** "An abstract model for the representation of multilingual terminological data: TMF-terminological markup framework". In *TAMA (Terminology in Advanced Microcomputer Applications) 2001*. https://hal.inria.fr/inria-00100405/document (consultato il 25/01/2022).

— **Scarpa, Federica (2001)** *La traduzione specializzata: lingue speciali e mediazione linguistica*. Milano: Hoepli.

— **Sistema Statistico Regionale (2019)** "Interscambio commerciale con l'estero". https://statistica.regione.veneto.it/jsp/commercionuovo.jsp?D1=2019+III%B0&D8=_totven&D7=999-MONDO&-D0=1&B1=Visualizza (consultato il 25/01/2022).

— **Sole 24 Ore (2018)** "Al Veneto il primato delle esportazioni". https://www.ilsole24ore.com/art/al-veneto-primato-esportazioni-AEzpHsdE?refresh_ce=1 (consultato il 25/01/2022).

— **Van Campenhoudt, Marc (2017)** "Standardised modelling and interchange of lexical data in specialised language". In *Revue française de linguistique appliquée,* n. 22 (1), 41-60.

— **Vezzani, Federica e Di Nunzio, Giorgio Maria (2020a)** "Methodology for the standardization of terminological resources: design of TriMED database to support multi-register medical communication". In *Terminology,* n. 26 (2), 266-298. https://doi.org/10.1075/term.00053.vez (consultato il 25/01/2022).

— **Vezzani, Federica e Di Nunzio, Giorgio Maria (2020b)** "On the Formal Standardization of Terminology Resources: The Case Study of TriMED". In *Proceedings of the 12th Language Resources and Evaluation Conference* (LREC2020), 4903-4910. https://aclanthology.org/2020.lrec-1.603.pdf (consultato il 25/01/2022).

— **Warburton, Kara (2015)** "Managing terminology in commercial environments". In Hendrik J. Kockaert e Frieda Steurs (a cura di) *Handbook of terminology*. Vol. 1. Amsterdam: John Benjamins, 359-391.

— **Wilkinson, Mark D., Dumontier, Michel, Aalbersberg, IJsbrand J., Appleton, Gabrielle, Axton, Myles, Baak, Arie, Blomberg, Niklas, Boiten, Jan-Willem, Da Silva Santos, Luiz B., Bourne, Philip E. *et al.* (2016)** "The FAIR guiding principles for scientific data management and stewardship". In *Scientific data*, n. 3. https://www.nature.com/articles/sdata201618.pdf (consultato il 25/01/2022).

— **Wright, Sue Ellen (2001)** "Data categories for terminology management". In Sue Ellen Wright e Gerhard Budin (a cura di) *Handbook of Terminology Management*. Vol. 2. Amsterdam: John Benjamins, 552-571.