La geolinguistica digitale e le sfide lessicografiche nell'era delle *digital humanities*: l'esempio di VerbaAlpina

Beatrice Colcuc
Ludwig-Maximilians-Universität München
beatrice.colcuc@romanistik.uni-muenchen.de

Abstract

English. The increasing use of new technologies and the possibilities they offer have led to a change in the way in which research and data processing are conceived. Collaboration between projects and data exchange are modern practices, the rules of which have been summarized in an acronym (FAIR) containing the four basic principles of digital research. The project VerbaAlpina of Munich University investigates the idioms of the Alpine area and, since the beginning, has been carried out according to these principles. However, in order to act according to the recommendations and pursue its research objective, VerbaAlpina, as a multilingual project, had to match the lexical-semantic variation with the universality of concepts. To establish concepts in a universal way is a difficult task because, to represent a concept, it is nevertheless always necessary to use a certain language. For this reason, VerbaAlpina has started to apply the procedures provided by external projects such as Wikidata or the resources of the German National Library for language-independent labelling of concepts on the one hand and linguistic forms on the other. Thus, the paper reflects the challenges of modern lexicography and the possibilities of overcoming problems in the digital era exemplified by the project VerbaAlpina.

Italiano. L'utilizzo delle nuove tecnologie e le possibilità da esse offerte hanno condotto a un mutamento nel modo di concepire la ricerca e il trattamento dei dati. La collaborazione tra i progetti e lo scambio di dati sono pratiche, le cui regole sono state riassunte nell'acronimo FAIR indicante i quattro princìpi fondamentali della ricerca digitale. Il progetto VerbaAlpina dell'università di Monaco di Baviera si occupa dello studio degli idiomi dell'area alpina e, fin dalla sua concezione, è stato portato avanti secondo tali princìpi. Tuttavia, per poter agire secondo le raccomandazioni e perseguire il proprio obiettivo di ricerca, VerbaAlpina si è dovuto confrontare con la problematica lessicale-semantica relativa alla mancanza di universalità dei concetti. Fissare dei concetti in maniera universale è un arduo compito perché la riproduzione di una data idea si effettua sempre attraverso una determinata lingua. Per questo motivo, VerbaAlpina ha iniziato ad applicare procedure fornite da progetti esterni quali Wikidata oppure le risorse della Biblioteca Nazionale Tedesca per l'etichettatura dei dati indipendente dalla lingua. L'articolo vuole presentare una riflessione sulle sfide della lessicografia moderna e sulle possibilità di superamento delle problematiche nell'era delle digital humanities presentando l'esempio concreto del progetto VerbaAlpina.

1 Trattare i dati (lessicografici) nell'era delle *digital humanities*

La ricerca dell'era pre-digitale è stata contraddistinta da modalità di concezione progettuale fortemente individuali: la raccolta, l'analisi e l'illustrazione dei dati venivano effettuate da persone (o gruppi di persone) che operavano singolarmente. La comunicazione scientifica si compiva attraverso le pubblicazioni dei libri cartacei, i quali venivano conservati in luoghi circoscritti come ad esempio le biblioteche, e ciascuno studio rappresentava un progetto concluso in sé (cfr. Krefeld, 2016). Inoltre, molti dati rimanevano nelle mani degli stessi ricercatori che li avevano raccolti e, in generale, la comunità scientifica non compiva molti sforzi per mettere a disposizione del grande pubblico i dati provenienti dai diversi progetti scientifici.

Ormai, l'avvento delle nuove tecnologie non è più un fenomeno di recente datazione. Il passaggio dal cartaceo al digitale è avvenuto in maniera sempre più repentina, contribuendo al mutamento dell'approccio scientifico nei confronti delle modalità di ricerca. La rivoluzione digitale ha rinsaldato l'idea della condivisione e, allo stesso tempo, provveduto a fornire gli strumenti necessari per favorire e facilitare l'interscambio di dati come pure, più in generale, l'interazione tra diversi progetti. Ciononostante, la creazione di una rete di progetti e la condivisione dei relativi dati non sono fornite da una mera digitalizzazione degli strumenti di ricerca. La collaborazione rappresenta l'essenza primaria del fare scienza, poiché è sulla base delle conoscenze già presenti che si costruisce il progresso. Per operare in maniera collettiva nell'era digitale è però necessario che i dati di ricerca soddisfino alcuni requisiti fondamentali. In primo luogo i dati devono essere strutturati, descritti

ed eventualmente etichettati in maniera tale da prestarsi a essere maneggiati in sedi esterne al loro progetto originario e devono poter continuare a essere accessibili anche in un momento successivo all'eventuale chiusura del progetto. In questo contesto si inserisce l'idea di Web Semantico (e successivamente dei Linked Open Data), nata con l'obiettivo di rimodellare l'ambiente virtuale di internet. Numerosi sono i progetti che si collocano all'interno di questo pensiero: essi formano la cosiddetta *Linked Open Data Cloud* (cfr. Cyganiak e Jentzsch, 2007 -) e costituiscono, al contempo, una comunità interconnessa attraverso relazioni basate sui loro insiemi di dati (cfr. Bizer et al. 2009, 154).

Le esigenze di condivisione e connessione dei dati sul web sono inoltre state formulate in maniera esplicita nel 2016, quando un grande numero di ricercatori provenienti da diversi Paesi ha pubblicato le linee guida per la gestione moderna dei dati di ricerca (cfr. Wilkinson, Dumontier et al., 2016). Tali raccomandazioni sono state racchiuse nell'acronimo FAIR, una sigla che raccoglie i quattro princìpi fondamentali sui quali dovrebbero essere basate la comunicazione e la cooperazione scientifiche nell'era digitale. Secondo tali princìpi, i dati della ricerca dovrebbero essere rintracciabili (*findable*), accessibili (*accessible*), interoperabili (*interoperable*) e riutilizzabili (*reusable*). Per essere rintracciabili, i progetti ai quali i dati appartengono, devono essere reperibili attraverso portali centrali quali, ad esempio, i cataloghi delle biblioteche. I dati di ricerca che non sono soggetti ad alcuna restrizione giuridica (come potrebbero essere ad esempio i dati strettamente personali) devono essere messi a disposizione del grande pubblico e rinunciare, di conseguenza, al diritto d'autore. Al fine di poter essere interoperabili, inoltre, i dati devono essere innanzitutto scissi, successivamente strutturati ed essere descritti in maniera precisa. Il riutilizzo dei dati si rende infine possibile attraverso una corretta applicazione dei tre princìpi precedenti: si tratta di princìpi estensibili che non si lasciano mettere in contrapposizione l'un l'altro (cfr. Force11, 2011-2017; GoFair, 2011-2017; Lücke, 2018).

Le possibilità offerte dalla digitalizzazione in termini tecnici hanno altresì consentito di valutare da un nuovo punto di vista una delle questioni storiche relative al trattamento dei dati lessicografici. In linea di massima, le opere lessicografiche possono essere strutturate in maniera semasiologica (le parole di un dato idioma sono elencate seguite dal loro significato) oppure onomasiologica (si descrivono i significati e vi si collegano le diverse parole che ad essi conducono). Nell'era analogica tali opere erano strutturate secondo uno o secondo l'altro modo: in ambito romanzo, si ricordino, tra gli altri, l'atlante linguistico ed etnografico dell'Italia e della Svizzera meridionale (AIS), di stampo onomasiologico, oppure, il *Dicziunari Rumantsch Grischun* (DRG), strutturato, invece, in maniera semasiologica. Una concezione in entrambi i sensi non era possibile a causa di limiti pratici imposti dalle modalità di pubblicazione del passato, mentre oggigiorno, le possibilità fornite dalla digitalizzazione offrono nuovi strumenti e la concezione di un'opera lessicografica che vada in due direzioni è realizzabile. Ciononostante, benché l'aspetto tecnico non rappresenti più alcun problema alla messa in pratica di tale approccio bidirezionale, le sfide si aprono soprattutto dal punto di vista contenutistico, come si evincerà dai capitoli che seguiranno.

La compilazione digitale di opere lessicografiche quali i dizionari sembra essere oggi un'attività relativamente consolidata. Non sono rari i dizionari che offrono la possibilità di essere consultati in rete, anche se, alcuni di essi, non si sono mai realmente distanziati da una concezione cartacea. Tali opere, presentano ancora un grande margine di sviluppo e ampliamento per potersi definire digitalizzate in modo interoperabile secondo i gradi definiti in Lücke (2016). A titolo illustrativo, ma non esaustivo, si pensi alla versione online del *Romanisches Etymologisches Wörterbuch* (Meyer-Lübke, 1935): il contenuto dell'opera è messo a disposizione online in formato PDF, ma, ai sensi della digitalizzazione, in essa potrebbero essere implementate diverse funzioni, tra cui, ad esempio, la ricerca di singoli lemmi. Lo scopo di una lessicografia basata sul web non si limita alla mera presentazione virtuale di una determinata opera. Lo sviluppo digitale è bensì rappresentato da un più ampio tentativo di costituzione di reti lessicali e semantiche messe in relazione tra di loro. Progetti volti a fornire tali interconnessioni sono stati iniziati già verso la metà degli anni Ottanta. A titolo esemplificativo si pensi a *WordNet*, il database lessicale per la lingua inglese, ma anche a *EuroWordNet* nato negli anni Novanta come rete semantica per le lingue europee (cfr. Fellbaum 2006, 665, 669). Tali progetti costituiscono un primo tentativo di strutturare il materiale in maniera semantica e non solo lessicale come si è soliti fare nei dizionari cartacei e, soprattutto, si inseriscono nel panorama dei lavori relativi all'elaborazione del linguaggio naturale multilingue. In tempi più recenti si colloca la concezione di *BabelNet* (https://babelnet.org/), una rete semantica multilingue, automatizzata e di ampia copertura, ovvero un dizionario enciclopedico costituito unendo il contenuto lessicale di *WordNet* al sapere enciclopedico di *Wikipedia* attraverso processi automatizzati di integrazione dei contenuti di ambedue i database (cfr. Navigli e Ponzetto, 2010, 216).

2 *VerbaAlpina*: geolinguistica e lessicografia digitali

Mentre la digitalizzazione dei primi dizionari e corpora lessicali si colloca tra gli anni Ottanta e gli anni Novanta (cfr. Chiari 2012, 98), più recente risulta invece essere il passaggio dal cartaceo al digitale per quanto riguarda gli atlanti linguistici. Relativamente all'area alpino-romanza, diversi atlanti sono oggi disponibili in rete in formati PDF o JPG ma non hanno percorso tutte le tappe del passaggio dal cartaceo al digitale (cfr. Lücke, 2016; Knapp, 2017). È, a titolo esemplificativo, il caso di *NavigAIS* (Tisato, 2009-2018) all'interno del quale, nonostante la sua presenza su internet sia lodevole, potrebbero essere implementate diverse funzionalità, tra le quali ad esempio la rintracciabilità di forme attestate oppure una visualizzazione quantificata dei dati, la possibilità di consultare singoli gruppi di dati linguistici in prospettiva onomasiologica o semasiologica, come pure l'esportazione dei dati.
Alle esigenze della ricerca lessicografica e atlantistica in chiave moderna, cerca di dare una risposta il progetto *VerbaAlpina* dell'Università Ludwig-Maximilian di Monaco di Baviera, nato nel 2014 con l'intento di indagare lo spazio linguistico delle Alpi nella sua storica unità linguistico-culturale (cfr. Krefeld e Lücke, 2014 -). Fin dalla sua concezione, completamente digitale e pensata non solo sul web, ma per il web, il progetto VerbaAlpina ha operato nel pieno rispetto dei princìpi FAIR (che sarebbero stati formulati solamente due anni dopo) e promosso un'idea innovativa di lessicografia e atlantistica linguistica (cfr. Krefeld, 2018). Oltre all'aspetto linguistico, una parte consistente del progetto è specificatamente dedicata alla creazione di strumenti per la gestione dei dati di ricerca nei progetti digitali e pensati per il web.

2.1 Concezione e presentazione del progetto

Nucleo centrale dell'attività di VerbaAlpina è la raccolta strutturata di una precisa cornice semasiologica e onomasiologica, costituita dagli ambiti terminologici alpini, alla quale è possibile accedere attraverso una cartina interattiva. La ricerca prende in esame il lessico dialettale degli idiomi alpini, in modo particolare le parole relative agli ambiti della natura (flora, fauna, formazioni paesaggistiche), della cultura alpina storica (lavorazione del latte) e di quella corrente (turismo). I dati raccolti e analizzati da VerbaAlpina sono puramente dialettali, mentre i termini relativi alle lingue standard non sono presi in considerazione. Diversi sono gli scopi perseguiti da VerbaAlpina: in primo luogo il progetto intende documentare e analizzare in prospettiva linguistica e storico-etimologica la regione alpina, uno spazio fortemente frammentato per quanto riguarda le lingue e i dialetti ivi parlati. I confini dell'area di ricerca sono definiti dalla Convenzione delle Alpi (http://www.alpconv.org/), un trattato tra i Paesi del territorio alpino atto a promuovere e sviluppare questa area montana in diversi ambiti (cfr. Lücke 2018a).

I dati sono forniti dagli atlanti linguistici e dai dizionari relativi all'area di ricerca, analogici o digitali, pubblicati nel corso del tempo. In un primo momento, il materiale linguistico georeferenziato proveniente dalle fonti affronta un percorso di digitalizzazione attraverso un sistema di trascrizione basato esclusivamente sui caratteri ASCII (cfr. Krefeld e Lücke, 2016). In un secondo momento, il materiale trascritto viene sottoposto a una *tokenizzazione*, un processo che separa in singoli *token* (parole) il materiale trascritto in un momento precedente. L'interesse principale del progetto si esplica nella presentazione dei punti di coesione tra i diversi idiomi e le diverse famiglie linguistiche presenti sul territorio alpino soprattutto in prospettiva lessicologica. Per l'adempimento di tale scopo, il materiale linguistico viene raggruppato in tipi di base, ossia secondo la radice lessicale comune a diverse attestazioni che possono appartenere anche a diverse famiglie linguistiche[1] e in tipi morfolessicali, vale a dire in forme di un solo tipo di base, appartenenti a un'unica famiglia linguistica che presentano caratteristiche grammaticali comuni quali la parte del discorso, il genere e gli elementi di formazione delle parole (cfr. Krefeld e Lücke, 2016a). Ad esempio, il tipo di base latino *CASEU(M) 'formaggio' è presente sia in area linguistica germanica, sia in area romanza nelle forme deu. *Käse* e ita. *cacio*, i quali, a loro volta, rappresentano due tipi morfolessicali differenti.

I dati linguistici storici rilevati dagli atlanti e dai dizionari sono completati attraverso una piattaforma di crowdsourcing sviluppata all'interno del progetto (https://www.verba-alpina.gwi.uni-muenchen.de/en/?page_id=1741&db=191). La piattaforma si rivolge direttamente ai parlanti dei dialetti delle Alpi al fine di raccogliere materiale linguistico attuale e poter osservare lo spazio alpino anche in prospettiva diacronica. Una volta aperta la pagina, viene chiesto agli utenti di scegliere una lingua di navigazione tra quelle proposte (francese, italiano, sloveno, tedesco). Successivamente, vengono mostrate le istruzioni per l'utilizzo

[1]In molti casi non è dato sapere se la radice lessicale comune a diverse parole sia da ricollegare allo stesso sostrato linguistico oppure a un contatto linguistico più recente. Per questo motivo VerbaAlpina utilizza il termine "tipo di base", in quanto "etimo" si riferisce generalmente allo strato linguistico immediatamente precedente (cfr. Krefeld e Lücke 2016a).

della piattaforma e gli utenti sono invitati a inserire l'idioma alpino di cui essi sono i parlanti. Nel caso in cui un idioma non sia presente nella lista, gli utenti hanno la possibilità di segnalarlo direttamente alla redazione che provvederà a inserirlo. Innanzitutto, gli utenti sono chiamati a inserire il nome del comune di cui padroneggiano l'idioma. Cliccando sull'apposito campo "concetto", appare una lista con tutti i concetti esistenti nella banca dati di VerbaAlpina. Da qui, gli utenti possono scegliere per quali concetti inviare parole. I dati raccolti attraverso il crowdsourcing vengono trattati alla pari dei dati provenienti dai dizionari e dagli atlanti, con l'unica differenza che non sono sottoposti al processo di trascrizione. Per questi dati, la tokenizzazione avviene solamente qualora si tratti di un sintagma costituito da più elementi. La tipizzazione di queste parole avviene alla stregua dei dati raccolti dai dizionari e dagli atlanti linguistici. A livello di database, le singole attestazioni provenienti dal crowdsourcing ricevono un identificatore e sono collegate ai concetti di cui rappresentano le diverse forme dialettali. Successivamente al trattamento strutturato, i dati analizzati da VerbaAlpina possono essere visualizzati sulla cartina interattiva (https://www.verba-alpina.gwi.uni-muenchen.de/it/?page_id=27&db=191). Tramite l'utilizzo di filtri appropriati, i dati sono accessibili in prospettiva onomasiologica (si rappresentano tutte le attestazioni linguistiche collegate a un determinato concetto) oppure semasiologica (si rappresentano i concetti legati a un preciso tipo morfolessicale). Inoltre, la visualizzazione può essere impostata in modalità qualitativa, attraverso la quale è possibile evincere la distribuzione generale delle attestazioni linguistiche, oppure quantitativa, ossia indicante il numero di dati all'interno di una certa area. I dati possono essere visualizzati in prospettiva geografico-fisica oppure astratta: la prima mostra i dati distribuiti su una cartina fisica, attraverso la seconda, invece, i dati sono presentati su una mappa a nido d'ape.

Parallelamente all'attività linguistica, il progetto ha profuso un grande impegno nella gestione dei dati digitali con l'obiettivo ultimo di promuovere la sostenibilità e la durabilità del progetto anche dopo la sua chiusura definitiva. La descrizione di una parte delle attività che sono state intraprese in questo senso avviene nel corso del presente contributo. VerbaAlpina si impegna a utilizzare strumenti tecnologici adatti al web e applicabili al pensiero open source, come ad esempio *Wordpress* per la piattaforma centrale, *Leaflet* per la cartina interattiva e le banche dati relazionali *MySQL*. Dato che VerbaAlpina intende altresì fungere da creatore di nessi tra istituzioni e progetti già esistenti al fine di interscambiare, integrare e completare i dati linguistici riguardanti l'area alpina, per i diversi partner sono messe a disposizione banche dati all'interno delle quali i progetti cooperanti possono inserire i loro dati e collegarli così a quelli di VerbaAlpina.

2.2 Status quo

L'area di ricerca di VerbaAlpina prende in considerazione tutti gli idiomi parlati nell'arco alpino. Si tratta di una superficie di 190.000 km^2 comprendente alcune regioni di sei Paesi diversi (Austria, Francia, Germania, Italia, Slovenia e Svizzera) e due interi stati (Liechtenstein e Montecarlo) per un totale di quasi 6000 comuni, i quali rappresentano per VerbaAlpina le unità di georeferenziazione. Considerato che in linguistica l'unanimità di opinioni su una definizione unitaria di dialetto rappresenta ancora una visione remota (e, a dire il vero, di scarso interesse per la disciplina stessa), non è possibile indicare un numero, nemmeno approssimativo, di varietà alpine locali parlate dalla Francia alla Slovenia. Pur concedendo grande importanza all'aspetto locale delle varietà, l'analisi linguistica di VerbaAlpina si eleva al livello delle tre famiglie linguistiche che occupano il territorio alpino (romanza, germanica e slava). Per questo motivo, attraverso il processo di tipizzazione di cui sopra, il variegato materiale linguistico locale è raggruppato in tipi morfolessicali etichettati rispettivamente con le sigle ISO 639-5 relative alle famiglie linguistiche: *roa*. per romanzo, *gem*. per germanico e *sla*. per slavo (cfr. ISO 639-5). La base di conoscenza di VerbaAlpina racchiude ad oggi (novembre 2019) 55.407 stimoli (si tratta solitamente dei titoli delle carte degli atlanti di riferimento) ai quali sono collegate 165.521 attestazioni linguistiche, distribuite tra 3.989 concetti e riassunte in 9.556 tipi morfolessicali. Per quanto riguarda i dati provenienti dal crowdsourcing, si contano 1.065 informanti diversi e 15.249 parole totali inviate dagli utenti.

3 La lessicografia tradizionale e le sfide per il futuro nell'era delle *digital humanities*

Come è stato già accennato, solitamente, i dizionari classici sono strutturati in maniera semasiologica, ovvero il lessico ivi contenuto viene elencato partendo dall'unità lessicale (parola) alla quale sono collegati i diversi significati. La fortuna di questo modello di dizionario è da ricondurre essenzialmente a due ragioni: da un lato tali opere lessicografiche hanno il compito di raccogliere e illustrare il lessico appartenente a un dato idioma (e in questo senso fungono da ausili per la documentazione di una lingua); dall'altro lato, concepire un'opera

lessicografica in prospettiva semasiologica risulta essere un'operazione di più facile realizzazione. Per riprodurre una serie ordinata di segni che compongono un'unità lessicale si può contare su un sistema codificato e standardizzato di caratteri: la scrittura stessa. La difficoltà di creare un'opera lessicografica partendo dal contenuto concettuale (prospettiva onomasiologica) è invece molto più estesa. Il contenuto semantico dell'unità lessicale non può essere delineato così facilmente, né tantomeno può essere standardizzato. L'utilizzo di una determinata espressione, non predice nulla sulle caratteristiche intrinseche del concetto al quale è collegata la parola che si è cercato di riprodurre. Fondamentalmente, sia per quanto riguarda la riproduzione di una singola parola, sia per quanto riguarda la descrizione di un significato, si fa appello alla scrittura, ovvero, alla lingua. Tuttavia, tale ricorso alla lingua è problematico giacché è possibile utilizzare solamente un determinato idioma alla volta, mentre invece, alla luce di quanto detto poc'anzi e nell'ottica di una scienza interconnessa, sarebbe opportuno potersi riferire ai concetti indipendentemente dalla singola lingua. Il mero ricorso ai codici linguistici ostacola inoltre la condivisione dei dati e la loro connessione ad altri database, limitando in parte una più ampia collaborazione tra progetti scientifici.

4 L'approccio al problema sull'esempio di VerbaAlpina

Per l'accorpamento dei contenuti provenienti dai vari atlanti linguistici e dai dizionari, anche il progetto VerbaAlpina si è dovuto misurare con la suddetta questione. All'interno della banca dati relazionale che funge da base del progetto, è stata creata una tabella che racchiude i concetti (si tratta prevalentemente dei contenuti tematici delle mappe linguistiche). Le singole mappe degli atlanti sono quindi collegate al concetto appropriato corrispondente. Per quanto riguarda il contenuto semantico di un concetto, il minimo comune denominatore è rappresentato dall'insieme delle informazioni che compongono il dato concetto. Dal momento che VerbaAlpina tratta dati provenienti da diverse fonti esterne al progetto, la gestione e l'uniformizzazione degli stessi può essere concepita solamente attraverso una descrizione accurata dei dati che, nell'insieme, formano un singolo concetto. Tuttavia, il metodo di gestione appena descritto consente solo la comparabilità all'interno del progetto, mentre per collegare tra di loro diversi gruppi di dati, sarebbe auspicabile e necessaria una soluzione globale e indipendente dalla lingua.

La sfida della standardizzazione è stata intrapresa da lungo tempo all'interno delle biblioteche, dove l'esigenza dell'uniformità mira a creare uno standard ad esempio per la gestione dei dati relativi agli autori delle diverse pubblicazioni o per la realizzazione di differenti indicizzazioni tematiche. È da questa esigenza dell'ambito bibliotecario che nasce l'idea del cosiddetto *authority control*, ossia un sistema normato per la costituzione di un archivio (*authority file*) che possa contenere dati organizzati secondo uno stesso modello. In Germania, a partire dagli anni Ottanta del secolo scorso sono stati creati diversi sistemi per la standardizzazione dei dati relativi a persone (PND: *Personennamendatei*), enti (GKD: *Gemeinsame Körperschaftsdatei*) e voci (SWD: *Schlagwortdatei*). Inoltre, tra il 2009 e il 2013, la Biblioteca Nazionale Tedesca (*Deutsche Nationalbibliothek*) e altre associazioni bibliotecarie di lingua tedesca hanno intrapreso un'iniziativa volta a creare il cosiddetto GND (*Gemeinsame Normdatei*), un sistema di controllo di autorità che riassume tutti gli elenchi sopraccitati in un unico file. Oltre alle tradizionali entità quali organismi o persone, il GND raccoglie anche concetti.

Anche il database enciclopedico *Wikidata*, creato allo scopo di supportare *Wikipedia*, funziona attraverso il controllo di autorità (cfr. Wikidata a). In Wikidata, ogni concetto è registrato tramite un numero identificatore (ID) e descritto nel dettaglio attraverso relazioni gerarchiche. La concezione partecipativa di Wikidata ha permesso l'inserimento nella piattaforma di un numero considerevole di entità referenziabili. A partire dal 2018, tutti i concetti di VerbaAlpina sono stati connessi ai Q-ID (o *Q-item*) di Wikidata. Tale connessione attraverso gli identificatori permette di collegare con altre banche dati esterne le informazioni altrimenti gestite solo all'interno del progetto. In questo senso, le risorse interne, come ad esempio il database multimediale, possono essere collegate in maniera decentralizzata ed essere messe a disposizione di diversi progetti. Questa concezione permette ai dati di essere costantemente arricchiti di informazioni aggiuntive. Allo stesso modo, è possibile pensare anche alla presentazione dei contenuti in diverse lingue, in quanto Wikidata mette a disposizione le traduzioni delle denominazioni relative ai diversi concetti (o ai relativi Q-ID). La collaborazione tra Wikidata e VerbaAlpina prevede non solo una connessione attraverso l'applicazione dell'identificatore, ma anche una partecipazione attiva di quest'ultimo al database enciclopedico. VerbaAlpina dispone infatti di un account proprio sulla pagina di Wikidata al fine di mappare eventuali concetti ivi mancanti. Al momento (novembre 2019), 1000 concetti di VerbaAlpina sono muniti di un corrispettivo Q-ID. Una parte consistente di concetti contenuti nella banca dati di VerbaAlpina deve essere ancora elaborata e ogni entrata

corredata del rispettivo Q-ID, un'attività che è in costante aggiornamento. Al momento, è stata data priorità all'applicazione degli identificatori ai concetti di VerbaAlpina, in un secondo momento avverrà anche la mappatura di concetti mancanti su Wikidata. L'inserimento dei Q-ID di Wikidata nella banca dati di VerbaAlpina avviene in maniera del tutto manuale.

Le proposte sopraccitate si riferiscono a un tentativo di standardizzazione del contenuto semantico di un'ampia quantità di concetti. Dal momento che VerbaAlpina non si occupa solamente del trattamento di materiale semantico, ma anche di tipi morfolessicali, la creazione di un controllo di autorità applicabile anche a tali forme linguistiche sarebbe auspicabile per l'identificazione univoca del contenuto lessicale. Per un'etichettatura di questo genere, la situazione si presenta in maniera diversa: il GND non fornisce ancora un sistema mirato per la gestione dei dati in questo senso, mentre Wikidata offre la possibilità di creare attestazioni lessicali contenenti le informazioni legate al lemma stesso, alla lingua e alla categoria lessicale. Ogni attestazione lessicale è correlata a un identificatore L (L-ID) che viene generato automaticamente (cfr. Wikidata b). Le singole voci possono anche essere completate con informazioni quali genere e significato. Per varietà linguistiche di estensione meno ampia quali ad esempio il ladino dolomitico oppure il friulano, Wikidata richiede non solo l'inserimento del lemma vero e proprio, ma anche di indicare la cosiddetta *spelling variant*, ossia l'ortografia utilizzata per rappresentare un determinato lemma indicata mediante un codice linguistico. Ad esempio, se si desidera inserire il lemma *paurùns* (forma ladina per il concetto 'siero di latte dopo la prima separazione della materia solida'), viene richiesto di indicare secondo quale ortografia è stato inserito il lemma stesso (cfr. Wikidata c). Inoltre, è possibile anche aggiungere le informazioni relative al numero e al caso linguistico e collegarle al rispettivo concetto. Quest'ultimo sistema di referenziazione consente di collegare ai singoli lemmi anche le informazioni riguardanti le derivazioni o l'etimologia, una pratica che faciliterebbe, di conseguenza, il lavoro lessicografico. Tale modalità di etichettatura del materiale linguistico è stata implementata da VerbaAlpina solo recentemente, ma sarà portata avanti in maniera progressiva con le stesse modalità applicate per i concetti.

La connessione tra i concetti di VerbaAlpina e quelli di Wikidata attraverso l'applicazione di un identificatore non è fine a se stessa, ma si inserisce in un'ottica di condivisione più ampia in grado trovare punti di aggancio anche con il progetto *GeRDI* (*Generic Research Data Infrastructure*). Quest'ultimo nasce nel 2016 come aggregatore di dati allo scopo di offrire a tutti i progetti di ricerca in Germania la possibilità di archiviare, condividere e riutilizzare dati. GeRDI impiega Wikidata come base di conoscenza, realizzando così un sistema di consultazione di dati interdisciplinare e multilingue (cfr. Mutter, 2018).

Wikidata rappresenta agli occhi di VerbaAlpina una piattaforma centrale attraverso la quale quest'ultimo può mettersi in relazione con altri progetti collegati analogamente alla stessa base di conoscenza. Un esempio potrebbe essere il dizionario enciclopedico BabelNet, anch'esso connesso a Wikidata. La connessione diretta tra VerbaAlpina e Babelnet sarebbe interessante per quanto riguarda l'identificazione dei lemmi, ma, quest'ultimo non dispone ancora di numeri univoci per questo tipo di materiale lessicale, né raccoglie dati dialettali, centrali invece per l'attività di VerbaAlpina. Ad ogni modo, benché non in maniera diretta, VerbaAlpina e Babelnet dispongono entrambe di Wikidata come progetto di collaborazione comune.

5 Prospettive e attività future

Facendo di nuovo riferimento ai princìpi FAIR menzionati all'inizio, organizzare e gestire i dati linguistici nel quadro dei sistemi di identificazione descritti poc'anzi, rappresenta un passo importante verso il rispetto di questi princìpi. I dati strutturati si presentano non solo più accessibili da parte di altri progetti e più facilmente reperibili grazie al collegamento in rete, ma la standardizzazione dei riferimenti esatti ne favorisce anche il trattamento dal punto di vista dell'interoperabilità. VerbaAlpina non è leale a questi princìpi solamente per quanto riguarda l'interoperabilità dei dati, ma anche relativamente alla loro rintracciabilità (*f*) attraverso l'inserimento del progetto nei cataloghi della biblioteca universitaria dell'università di Monaco di Baviera. VerbaAlpina sposa in toto l'idea di libero accesso alla conoscenza come bene comune e utilizza solamente licenze Creative-Commons (CC) rinunciando, di conseguenza, al diritto d'autore e permettendo l'utilizzo dei dati con la sola restrizione dell'obbligo di citazione (cfr. Lücke, 2016a); il riutilizzo dei dati è reso possibile attraverso la loro esportazione tramite un'interfaccia di programmazione di un'applicazione (eng. *Application Programming Interface*; API), la quale permette l'accesso all'intero dataset di VerbaAlpina. Una documentazione e spiegazione dettagliata è consultabile al seguente indirizzo: https://www.verba-alpina.gwi.uni-muenchen.de/?page_id=8844&db=191. Attraverso tale API i dati di VerbaAlpina sono accessibili meccanicamente (*machine readable*) e possono essere scaricati, modificati ed elaborati ulteriormente. La connessione dei dati con altri dataset è garantita attraverso lo schema di metadati di DataCite

(https://datacite.org), la quale si trova ancora in fase di sviluppo. Nonostante l'accesso meccanico ai dati dall'esterno sia già possibile mediante l'API e la connessione dei dati con altri dataset attraverso i metadati, le procedure per inserire i dati nella nuvola dei Linguistic Linked Open Data (https://linguistic-lod.org/) sono in fase di avvio allo scopo di creare un'ulteriore connessione tra progetti e contribuire in questo senso all'idea di web strutturato. Operare nell'era delle *digital humanities* significa creare conoscenza interconnessa, condivisibile, accessibile, una conoscenza più ampia e coesa. Equivale a creare strumenti e a metterli a disposizione non solo della comunità scientifica, ma anche del grande pubblico. Si tratta di un'amplificazione dell'originale pensiero umanista: creare sapere, renderlo accessibile e diffonderlo affinché l'umanità possa accrescere le proprie conoscenze.

Bibliografia

AIS = Karl Jaberg e Jakob Jud. 1928-1940. *Sprach- und Sachatlas Italiens und der Südschweiz*. 8. vol. Riniger&Co, Zofingen.

API = VerbaAlpina. 2014 -. API Dokumentation. https://www.verba-alpina.gwi.uni-muenchen.de/?page_id=8844&&db=191 [accesso 29/11/2019].

BabelNet = BabelNet | The largest multilingual encyclopedic dictionary and semantic network. https://babelnet.org/ [accesso 25/11/2019].

Cartina interattiva = VerbaAlpina. 2014 -. Cartina interattiva. https://www.verba-alpina.gwi.uni-muenchen.de/it/?page_id=27&db=191 [accesso 25/11/2019]

Christian Bizer, Tom Heath, e Tim Berners-Lee. 2009. Linked data - the story so far. *International Journal of Semantic Web and Information System* 5(3):1-22. doi:10.4018/jswis.2009081901 [accesso 12/11/2019].

Christiane Fellbaum. 2006. Wordnet and wordnets. In Keith Brown, ed. by., *Encyclopedia of Language and Linguistics*, Elsevier, Oxford: 665-670.

Christina Mutter. 2018. Wikidata. *VerbaAlpina-it 19/1* (creazione: 18/1), *Metodologia*. https://doi.org/10.5282/verbaalpina?urlappend=%3Fpage_id%3D493%26db%3D191%26letter%3DW%23105

Convenzione delle Alpi. 1995-. https://www.alpconv.org/it/home/ [accesso 29/11/2019].

Crowdsourcing = VerbaAlpina. 2014 -. Crowdsourcing-Tool. https://www.verba-alpina.gwi.unimuenchen.de/en/?page_id=1741&db=191 [accesso 26/11/2019].

DataCite: https://datacite.org/index.html [accesso 28/11/2019].

DRG = Dicziunari Rumantsch Grischun. 1939-2013. Institut dal Dicziunari Rumantsch Grischun, Coira.

Force11, ed. by. 2011-2017. *The Fair Data Principles*. https://www.force11.org/group/fairgroup/fairprinciples [accesso 03/09/2019].

GeRDI = Generic Research Data Infrastructure. 2016 -. https://www.gerdi-project.eu/ [accesso 14/11/2019].

GoFair, ed. by. 2011-2017. *FAIR Principles*. https://www.go-fair.org/fair-principles. [accesso 03/09/2019].

Graziano Tisato, ed. by. 2009-2018. NavigAIS. AIS Digital Atlas and Navigation Software, Padova, Istituto di Scienze e Tecnologie della Cognizione (ISTC) - Consiglio Nazionale delle ricerche (CNR). Versione 1.47. http://www3.pd.istc.cnr.it/navigais/ [accesso 20/11/2019].

Isabella Chiari. 2012. Il dato empirico in lessicografia: dizionari tradizionali e collaborativi a confronto. *Bollettino di Italianistica*. Per Tullio De Mauro II, pp. 94-125.

ISO 639-5 = Library of the Congress Registration Authority. 2009. Codes for the representation of Names of Languages – Part 5: Alpha-3 code for language families and groups. https://www.loc.gov/standards/iso639-5/index.html [accesso 17/11/2019].

Katharina Knapp. 2017. Elenco dei siti atlantistici e lessici dialettali online. https://www.kit.gwi.uni-muenchen.de/?p=12110 [accesso 17/11/2019].

Linguistic Linked Open Data: https://linguistic-lod.org/ [accesso 28/11/2019].

Mark Wilkinson, Michel Dumontier et al. 2016. The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data 3. https://www.nature.com/articles/sdata201618 [accesso 05/09/2019].

Richard Cyganiak und Anja Jentzsch. 2007 - . Linked Open Data Cloud. https://lod-cloud.net/ [accesso 31/10/2019].

Roberto Navigli e Simone Paolo Ponzetto. 2010. Babelnet: building a very large multilingual semantic network. *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pp. 216-225. Association for Computational Linguistics. https://www.aclweb.org/anthology/P10-1023/ [accesso 28/11/2019].

Stephan Lücke. 2016. Digitalizzazione. *VerbaAlpina-it* 19/1 (creazione 16/1), *Metodologia*. https://doi.org/10.5282/verba-alpina?urlappend=%3Fpage_id%3D493%26db%3D191%26letter%3DD%2315

Stephan Lücke. 2018. FAIR-Prinzipien. *VerbaAlpina-it* 19/1 (creazione 18/2), *Metodologia*. https://doi.org/10.5282/verba-alpina?urlappend=%2Fit%3Fpage_id%3D21%26db%3D191%26letter%3DP%23128

Thomas Krefeld e Stephan Lücke, ed. by. 2014 –. *VerbaAlpina. Der alpine Kulturraum im Spiegel seiner Mehrsprachigkeit*. München, online. http://dx.doi.org/10.5282/verba-alpina

Thomas Krefeld e Stephan Lücke. 2016. Codice Beta. *VerbaAlpina-it* 19/1 (creazione 16/1), *Metodologia*. https://doi.org/10.5282/verba-alpina?urlappend=%3Fpage_id%3D493%26db%3D191%26letter%3DB%237

Thomas Krefeld e Stephan Lücke. 2016a. Tipizzazione. *VerbaAlpina-it* 19/1 (creazione 16/1), *Metodologia*. https://doi.org/10.5282/verba-alpina?urlappend=%2Fit%3Fpage_id%3D21%26db%3D191%26letter%3DT%2358

Thomas Krefeld. 2016. Comunicazione scientifica nel web. *VerbaAlpina-it* 19/1 (creazione 16/1), *Metodologia*. https://doi.org/10.5282/verba-alpina?urlappend=%2Fit%3Fpage_id%3D21%26db%3D191%26letter%3DC%2362

Thomas Krefeld. 2018. I principi FAIR nel progetto VerbaAlpina, ossia il trasferimento della geolinguistica alle Digital Humanities. *VerbaAlpina-de 19/1* (creazione 18/2). https://www.verba-alpina.gwi.uni-muenchen.de/?p=8212&db=191-rf1-8212

Wikidata a. Introduction. https://www.wikidata.org/wiki/Wikidata:Introduction [accesso 05/09/2019].

Wikidata b. Lexikographische Daten/Dokumentation. https://www.wikidata.org/wiki/Wikidata:Lexicographical_data/Documentation/de [accesso 05/09/2019].

Wikidata c. https://www.wikidata.org/wiki/Wikidata:Lexicographical_data/Glossary [accesso 17/11/2019].

Wilhelm Meyer-Lübke. 1935. Romanisches Etymologisches Wörterbuch, versione online. urn:nbn:de:bvb:355-ubr07799-0 [accesso 17/11/2019].