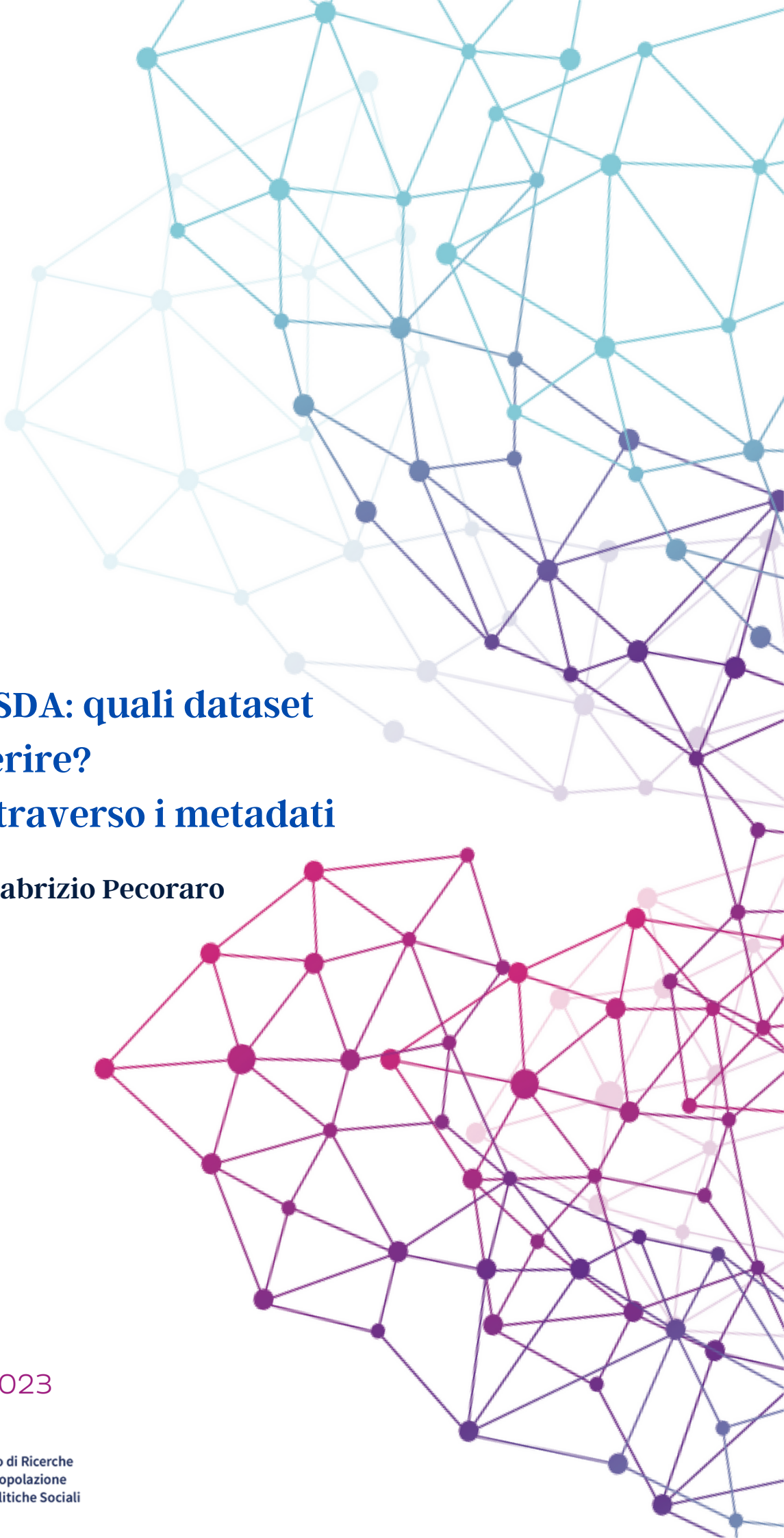
E-PUBLISHING
IRPPS



Il catalogo CESSDA: quali dataset si possono reperire? Un'indagine attraverso i metadati

Filippo Accordino, Fabrizio Pecoraro e Daniela Luzi

WORKING PAPER 141

DICEMBRE 2023

IRPPS Istituto di Ricerche sulla Popolazione e le Politiche Sociali

CNR – IRPPS

Il catalogo CESSDA: quali dataset si possono reperire?
Un'indagine attraverso i metadati

Filippo Accordino, Fabrizio Pecoraro e Daniela Luzi
2023, p. 47 IRPPS Working papers 141/2023

Sommario: Il Consortium of European Social Science Data Archive (CESSDA) è un'infrastruttura di ricerca che riunisce gli archivi europei di dati sulle scienze sociali. L'obiettivo di CESSDA è quello di condividere le esperienze e rafforzare la cooperazione, sensibilizzando la comunità scientifica alla condivisione dei dati della ricerca. Il CESSDA Data Catalogue rientra tra i principali strumenti erogati dall'infrastruttura e raccoglie tutti i metadati dei dataset custoditi nei singoli archivi. Attraverso un'interfaccia web, è possibile eseguire una ricerca dei dati tra tutti gli archivi aderenti. Il presente lavoro offre una panoramica generale sui dati reperibili attraverso il CESSDA Data Catalogue, fornendo un quadro introduttivo sull'ampiezza, le lingue utilizzate e i Paesi di riferimento. Successivamente, per ogni singolo archivio, sono approfonditi topic, modalità di raccolta dei dati e altre informazioni, valutando anche la disponibilità e la precisione dei metadati.

Parole chiave: CESSDA, Archivio dati, Catalogo dati, Metadati, Condivisione dati

CNR – IRPPS

The CESSDA Data Catalogue: which datasets are retrievable?
An investigation through metadata

Filippo Accordino, Fabrizio Pecoraro e Daniela Luzi
2023, p. 47 IRPPS Working papers 141/2023

Abstract: The Consortium of European Social Science Data Archive (CESSDA) is a research infrastructure bringing together social science data archives in Europe. The aim of CESSDA is to share experiences and strengthen cooperation, raising awareness among the scientific community about research data sharing. The CESSDA Data Catalogue is one of the main tools provided by the infrastructure. Through a web interface, it is possible to search for data among all participating archives. This work offers a general overview of the data retrievable through the CESSDA Data Catalogue, providing an introductory framework on the scope, languages used and geographical area. Subsequently, for each individual archive, topics, data collection methods and other information are explored in depth, also evaluating the availability and precision of the metadata.

Keywords: *CESSDA, Data archive, Data catalogue, Metadata, Data sharing*

Citare questo documento come segue:

Filippo Accordino, Fabrizio Pecoraro e Daniela Luzi (2023). *Il catalogo CESSDA: quali dataset si possono reperire? Un'indagine attraverso i metadati*. Roma: Consiglio Nazionale delle Ricerche – Istituto di Ricerche sulla Popolazione e le Politiche Sociali, (IRPPS Working papers n. 141/2023, p. 47).

CNR-IRPPS, via Palestro 32, 00185, Rome, Italy
E-MAIL: filippo.accordino@irpps.cnr.it

INDICE

Abbreviazioni

- API Application Programming Interface
- CDC CESSDA Data Catalogue
- CESSDA Consortium of European Social Science Data Archives
- CMM CESSDA Metadata Model
- CSV Comma separated value
- DAG Data Archiving Guide
- DASSI Data Archive for Social Sciences in Italy
- DDI Alliance Data Documentation Initiative Alliance
- DMEG CESSDA Data Management Expert Guide
- ERIC European Research Infrastructure Consortium
- ESFRI European Strategy Forum on Research Infrastructures
- FAIR Findability, Accessibility, Interoperability, and Reusability
- JSON JavaScript Object Notation
- ORCID Open Researcher and Contributor ID
- REST Representational State Transfer
- SQL Structured Query Language

Indice delle figure

Indice delle tabelle

1. Introduzione

Il Consortium of Social Science Data Archives (CESSDA) è un'infrastruttura di ricerca, operante nell'ambito delle scienze sociali, alla quale aderiscono numerosi Paesi europei. Nata nel 1976 da un primo nucleo di 7 Paesi e con nome originario "Council of European Social Science Data Archives", è oggi composta da 21 membri, 1 osservatore, 12 partners (vedi https://www.cessda.eu/About/Consortium-and-Partners/List-of-Service-Providers). Ogni Paese membro elegge e sostiene un proprio archivio nazionale, o *service provider*, che diventa parte dell'infrastruttura CESSDA. Lo scopo è unire le forze tra gli archivi nazionali che raccolgono dati afferenti alle scienze sociali, offrendo alla comunità scientifica l'opportunità di condurre ricerche di alta qualità, condividendo risorse ed expertise.

CESSDA, nel corso degli anni, ha rafforzato la propria struttura e il coinvolgimento dei governi e degli archivi nazionali, mutando nel 2013 la propria intitolazione da "Council" a "Consortium". Nel 2016 la sostenibilità di CESSDA come infrastruttura di ricerca viene riconosciuta attraverso l'attribuzione dell'ESFRI Landmark status (vedi https://www.esfri.eu/objectives-vision). L'European Strategy Forum on Research Infrastructures (ESFRI) è uno strumento strategico con l'obiettivo di sviluppare l'integrazione della ricerca scientifica in Europa.

Ruolo, scopi e organizzazione sono stati riconosciuti qualche anno più tardi, nel 2017, dalla Commissione Europea (Commission Implementing Decision (EU) 2017/995 del 9 June 2017), con l'attribuzione dello status di European Research Infrastructure Consortium (ERIC). La progressiva formalizzazione ha condotto a un crescente impegno dei Paesi coinvolti.

Il ruolo di CESSDA come consorzio si traduce nell'obiettivo di riunire depositi certificati di dati, incentivando la condivisione di expertise e promuovendo pratiche di condivisione e riuso (Pasquetto et al. 2017).

Le attività di CESSDA si fondano sul riconoscimento dell'importanza della condivisione dei dati, la cui crescente disponibilità favorisce importanti occasioni di sviluppo nella ricerca, nell'economia e nella società in generale. Il Consorzio aderisce a principi FAIR, un paradigma che sintetizza la necessità di rendere i dati reperibili, accessibili, interoperabili e riutilizzabili (Dekker, 2020; Hodson, Jones et al., 2018). CESSDA promuove la reperibilità dei dati attraverso i cataloghi, la qualità della metadatazione, il riuso. Le attività rivolte ai service provider comprendono la fornitura di risorse, servizi e formazione.

Il *service provider* italiano è DASSI (https://www.dassi-archive.it), un'infrastruttura di ricerca nata nel 2021 attraverso una Joint Research Unit (JRU) tra Il Consiglio Nazionale delle Ricerche e l'Università degli Studi di Milano-Bicocca. DASSI sta curando lo sviluppo del deposito di dati nazionale per l'Italia (Università degli Studi di Milano-Bicocca, 2021). Ad oggi, l'archivio italiano i cui dati confluiscono sul catalogo CESSDA (vedi par. 2) è UniData, un progetto che riunisce otto dipartimenti dell'Università degli Studi di Milano-Bicocca (CESSDA,

Data Archive Social Sciences Italy – DASSI, https://www.cessda.eu/About/Consortium-and-Partners/List-of-Service-Providers/Italy-spid1908), e la cui esperienza è di supporto alle attività di DASSI.

La produzione generale di dati in costante aumento, e la specifica necessità della ricerca scientifica di esporre e condividere i dati della ricerca aderendo a principi FAIR, incentiverà ulteriormente nei prossimi anni la crescita degli archivi nazionali e, di conseguenza, l'infrastruttura CESSDA che li riunisce.

Ad oggi CESSDA mette a disposizione alla comunità scientifica, e a chiunque sia interessato, un catalogo di libero accesso, il CESSDA Data Catalogue (CDC). Attraverso procedure di *harvesting*, il CDC raccoglie, dai cataloghi dei *service provider* nazionali, i metadati (Bradić-Martinović e Banović, 2021, p. 194) dei dataset depositati in ogni archivio. Tutti i dataset, di conseguenza, diventano rintracciabili attraverso un'unica interfaccia, con tutti i vantaggi che ne derivano. Il CDC è accessibile senza registrazione e con la possibilità di accedere ai dati di interesse individuati, depositati in uno degli archivi nazionali, attraverso un apposito link che rinvia alla risorsa.

2. Obiettivi del lavoro

Attraverso questo report si intende offrire una panoramica generale sui dati attualmente reperibili attraverso il CESSDA Data Catalogue (CDC). Obiettivi del lavoro svolto sono:

- Illustrare le caratteristiche dei dati condivisi, attraverso alcune informazioni ricavabili dalla metadatazione: riferimenti temporali, lingue impiegate, argomenti specifici all'interno dell'ampio ambito disciplinare delle scienze sociali, modalità di raccolta del dato, unità d'analisi osservata, tipi di produttori;
- Osservare eventuali differenze o affinità tra gli archivi nazionali che partecipano al CESSDA;
- Verificare completezza e qualità nella compilazione dei metadati.

La struttura del CDC sarà esposta nei paragrafi successivi, ripercorrendo il lavoro di estrazione dei dati e illustrando le informazioni ricavate.

Il CDC, nella sua forma attuale, è stato introdotto nel 2016 (cfr. CESSDA Strategy 2018-2022, p. 18). Attualmente è disponibile la versione 3.4.0, rilasciata in data 29/08/2023. È liberamente accessibile al sito https://datacatalogue.cessda.eu.

Come riportato nella descrizione (https://datacatalogue.cessda.eu/about):

> *The presented data includes a range of data types including quantitative, qualitative or mixed-modes data, covering both cross-sectional and longitudinal studies, as well as recently collected and historical data. The metadata (study descriptions) are presented in the language*

as originally provided by the organisations producing the metadata. Some publishers provide study descriptions both in English and in the local language, some only in either English or in the local language. Currently, about 75% of study descriptions are available in English.

3. Estrazione dati CDC

Sono stati scaricati e analizzati i metadati dei dataset disponibili nella sola lingua inglese, pari a un totale di 26.641 alla data del 21/02/2023.

Lo scaricamento è avvenuto tramite l'apposito servizio di interfaccia di programmazione delle applicazioni, o *Application Programming Interface (API)* reso disponibile da CESSDA. Le API REST (*Representational State Transfer*) sono funzionalità di un servizio web, accessibili da programmi esterni attraverso rete. Tramite un'apposita procedura di interrogazione, le API consentono di acquisire dati e risorse. La flessibilità concessa dal servizio permette di personalizzare l'interrogazione al fine di ottenere i dati desiderati. Lo scaricamento può essere gestito direttamente dagli stessi software adoperati per eseguire l'elaborazione del dato. Il servizio CESSDA Data Catalogue Search API è raggiungibile all'indirizzo web https://api.tech.cessda.eu, nel quale è disponibile anche l'apposita documentazione.

Tutto il processo di estrazione e il codice scritto in linguaggio di programmazione *R* sono riportati in appendice al par. 1.

3.1 Composizione dati CDC

Il CDC è corredato dalla CESSDA Data Catalogue User Guide, disponibile all'indirizzo https://datacatalogue.cessda.eu/documentation. La guida illustra l'interfaccia del CDC e le modalità di ricerca.

I dati esposti sul catalogo sono conformi al CESSDA Metadata Model v1.0 (CMM), come riportato alla pagina https://datacatalogue.cessda.eu/documentation/providing-oai-pmh.html al 17/02/2023, che in un'apposita porzione dello schema (campi "mapping information") riporta le etichette del catalogo equivalenti alla metadatazione CESSDA e le informazioni di corredo.

Il servizio CESSDA Data Catalogue Search API, raggiungibile all'indirizzo https://api.tech.cessda.eu, illustra uno schema di riferimento dei dati scaricabili, riportato in **Tabella 1**, dal quale è possibile cogliere la struttura e la gerarchia delle variabili.

Tabella 1. Variabili illustrate nel CESSDA Data Catalogue Search API (https://api.tech.cessda.eu)

Variabile nello schema	Descrizione
id	string - *example: UKDS__4929*
code	string - *example: UKDS*
creators	[string - *example: Fraser, R*]
dataCollectionPeriodStartdate	string - *example: 1984-01-01T00:00:00Z*
dataCollectionPeriodEnddate	string - *example: 1985-01-01T00:00:00Z*
dataCollectionYear	Integer - *example: 1984*
dataCollectionFreeTexts	[{ dataCollectionFreeText: string - *example: 1984* event: string - *example: start* }]
dataAccessFreeTexts	[string - *example: The depositor has specified that registration is required and ...*]
publicationYear	string - *example: 2005-04-11T00:00:00Z*
typeOfModeOfCollections	[{ vocab: string vocabUri: string id: string term: string - *example: Face-to-face interview* }]
keywords	[{ vocab: string - *example: YEAR* vocabUri: string id: string term: string - *example: 1945* }]
samplingProcedureFreeTexts	[string - *example: Purposive selection/case studies*]
classifications	[{ vocab: string - *example: CESSDA Topic Classification* vocabUri: string - *example: urn:ddi:int.cessda.cv:TopicClassification:4.1* id: string term: string - *example: LawCrimeAndLegalSystems.CrimeAndLawEnforcement* }]
abstract	string
titleStudy	string - *example: 1968: A Student Generation in Revolt, 1945-1985*
studyUrl	string - *example: http://doi.org/10.5255/UKDA-SN-4929-1*
studyNumber	string - *example: 4929*
typeOfTimeMethods	[{ vocab: string vocabUri: string id: string term: string - *example: Cross-sectional (one-time) study* }]
fileLanguages	[string - *example: en*]
publisher	[{ abbr: string - *example: UKDS* publisher: string - *example: UK Data Service* }]
studyAreaCountries	[{ abbr: string - *example: PT* searchField: string - *example: Portugal* country: string - *example: Great Britain* }]
unitTypes	[{ vocab: string vocabUri: string id: string term: string - *example: EventOrProcessOrActivity* }]
pidStudies	[{

	pid string - *example: APIS00066* }]
lastModified	string - *example: 2021-09-17T14:44:09Z*
isActive	boolean
langAvailableIn	[string - *example: en*]

Nella successiva **Tabella 2** sono riportate le variabili realmente ottenute a seguito dell'estrazione dal catalogo CESSDA e prima di qualsiasi operazione finalizzata all'estrazione delle liste annidate (nei linguaggi di programmazione, liste, ovvero elenchi indicizzati di oggetti, inserite all'interno di altre liste), che comporta la creazione di ulteriori variabili di livello inferiore, come meglio descritto dal successivo par. 3.2.

*Tabella 2. Variabili disponibili dopo aver eseguito la procedura di estrazione descritta. Quelle contrassegnate con * non risultano illustrate nello schema REST API e nel CMM*

N / variabile estratta	**N / variabile estratta**
1 * id	16 * publisherFilter
2 abstract	17 samplingProcedureFreeTexts
3 classifications	18 studyAreaCountries
4 * code	19 studyNumber
5 creators	20 studyUrl
6 dataCollectionPeriodStartdate	21 typeOfModeOfCollections
7 dataCollectionPeriodEnddate	22 titleStudy
8 * dataCollectionYear	23 typeOfTimeMethods
9 dataCollectionFreeTexts	24 typeOfSamplingProcedures
10 dataAccessFreeTexts	25 unitTypes
11 fileLanguages	26 * universe
12 keywords	27 * lastModified
13 pidStudies	28 * langAvailableIn
14 publicationYear	29 * studyXmlSourceUrl
15 publisher	30 * relatedPublications

Dalla descrizione dello schema in **Tabella 1** è possibile ricavare, nell'ordine, il nome delle colonne annidate estratte nelle fasi successive, permettendo una loro distinzione che si rende necessaria anche a causa dell'omonimia (ad esempio: vocab, vocabUri, id, term).

Da notare qualche disallineamento tra le variabili previste e descritte nello schema in **Tabella 1** e quelle estratte, in particolare la mancanza di alcune variabili sia nello schema sia nel CMM. Le variabili presenti nell'estrazione ma mancanti nello schema e/o non corrispondenti a un'etichetta CDC nel CMM sono descritte nella **Tabella 3.**

Tabella 3. Variabili non corrispondenti allo schema e/o al CMM

Manca nello schema	Manca nel CMM	Variabile	Note (interpretazione)
	x	id	Non è predisposta l'etichetta per il CDC nel CMM, ma è riportato nella metadatazione: Identifier of the study according to DDI 3.2 structure
	x	code	Corrisponde all'acronimo dell'archivio dal quale provengono i dati
	x	dataCollectionYear	Corrisponde al limite temporale minimo (anno) di raccolta (informazione riportata anche nel campo dataCollectionPeriodStartDate)
x	x	publisherFilter	Corrisponde all'acronimo del service provider
x	x	relatedPublications	Riferimenti bibliografici a pubblicazioni correlate (articoli scientifici)
x	x	universe	Non è predisposta l'etichetta per il CDC nel CMM, ma è riportato nella metadatazione: Description of the population of the study (freetext)
	x	lastModified	
	x	langAvailableIn	Lingua di disponibilità del dataset, sulla base di questo dato è suddiviso il CDC nelle undici lingue come disponibile nell'interfaccia sul sito web
	x	studyXmlSourceUrl	

3.2 Selezione e pulizia

A seguito dello scaricamento ed estrazione, si è proceduto a selezionare le variabili di interesse tenuto conto del CMM e a estrarre le liste annidate, ovvero le variabili di livello gerarchicamente inferiore, in modo da ricavare alcuni dataframe interrogabili in *R* tramite linguaggio *Structured Query Language (SQL)* ed esportabile in formato *comma separated value* (.CSV).

A tal fine sono stati predisposti una serie di script che di volta in volta hanno selezionato le porzioni di dataframe relative a ogni aspetto considerato, estraendo le variabili di livello inferiore.

Sono state ricavate alcune informazioni generali:

- Numero di dataset per archivio nazionale;
- Lingue in cui sono disponibili i dataset;
- Paesi ai quali si riferiscono i dataset;
- Periodo di riferimento dei dati (anno di inizio e di fine raccolta).

Per altre informazioni, l'estrazione e l'analisi dei dati è stata svolta su ogni singolo archivio, unendo i risultati successivamente al fine di procedere a comparazioni:

- Argomenti (topics) oggetto del dataset;
- Modalità di raccolta del dato: intervista, questionario, ecc.;
- Ambito temporale di riferimento dei dati: studi longitudinali, serie storiche, ecc.;
- Unità d'analisi dei dati: individui, famiglie, unità geografiche, ecc.

La gestione dei dati è stata eseguita quasi interamente in ambiente *R*. Le maggiori difficoltà riscontrate sono dovute a metadati compilati in modo errato anche quando è impiegato un vocabolario controllato, o refusi ed errori di altra natura. Ciò ha comportato un lungo e oneroso lavoro di pulizia nell'intento di minimizzare la perdita di informazione, prestando attenzione a non alterare la natura dei metadati.

4. Informazioni generali sul CDC

4.1 Ampiezza del catalogo

Gli archivi UKDS e GESIS, anche in virtù della loro longevità, compongono oltre la metà del totale di 26.641 dataset in lingua inglese. 9 archivi sui 16 totali non raggiungono ancora i 1.000 contributi.

*Tabella 4. Dimensioni degli archivi nazionali e altre informazioni. Nei casi contrassegnati con * la data di fondazione si riferisce a un precedente archivio, confluito in quello attuale. La data di fondazione, per alcuni archivi, non è reperibile nei rispettivi siti web*

Denominazione	Sigla	Paese	Sito web	Anno fondaz.	Totale al 21/02/2023
UK Data Service	UKDS	Regno Unito	https://ukdataservice.ac.uk/	1967	9.093
Leibniz Institute for the Social Sciences	GESIS	Germania	https://www.gesis.org/en/home	1960	5.478
Data Archiving and Networked Serviced	DANS	Paesi Bassi	https://dans.knaw.nl/nl/	2005	3.185
Norwegian Agency for Shared Services in Education and Research	NSD (Sikt)	Norvegia	https://sikt.no/en/about-sikt	1971*	2.515
Finnish Social Science Data Archive	FSD	Finlandia	https://www.fsd.tuni.fi/en	1999	1.764
The Austrian Social Science Data Archive	AUSSDA	Austria	aussda.at	2016	1.427
Danish National Archive	DNA	Danimarca	https://en.rigsarkivet.dk		1.220
Swedish National Data Service	SND	Svezia	https://snd.gu.se/en		748
Slovenian Social Science Data Archives	ADP	Slovenia	https://www.adp.fdv.uni-lj.si/eng/spoznaj/adp	1997	583
Swiss Centre of Expertise in the Social Sciences	FORS	Svizzera	https://forscenter.ch/data-services/	2008	272
UniData – Bicocca Data Archive	UniData	Italia	https://unidata.unimib.it	1999*	133
Social Data Network	SoDaNet	Grecia	https://sodanet.gr	1999	100
Social Sciences and Digital Humanities Archive	SODHA	Belgio	https://www.sodha.be		64
Portuguese Archive of Social Information	APIS	Portogallo	apis.ics.ulisboa.pt/en		36
Slovak archive of social data	SASD	Slovacchia	http://sasd.sav.sk/sk	2004	17
Center for socio-political data	ProgedoCDSP	Francia	https://cdsp.sciences-po.fr	2005	6
				TOTALE	**26.641**

Oltre la metà degli archivi sono stati creati a partire dal 1999. L'anno di fondazione di ogni archivio, indicato in **Tabella 4**, è stato ricavato dalle informazioni riportate nei rispettivi siti web, quando disponibili.

L'archivio UniData nasce nel 2015 come centro interdipartimentale dell'Università di Milano-Bicocca, raccogliendo in eredità il lavoro svolto dall'archivio ADPSS_Sociodata, nato nel 1999 (https://www.unidata.unimib.it/?page_id=615). L'archivio Sikt nasce nel 2022 sostituendo il precedente NSD, fondato nel 1971.

Il notevole contributo di alcuni archivi, in termini di numero di dataset resi disponibili, emerge dai grafici in **Figura 1** e **Figura 2**. In particolare, oltre un dataset su due dell'intero CDC appartiene all'archivio inglese (UKDS) o a quello tedesco (GESIS). Il terzo gradino del podio, in termini di quantità di contributi, spetta a DANS, con 3.185 dataset corrispondenti all'11,96% del totale. Tre archivi contribuiscono in una quota tra il 5% e il 10%. Tutti gli altri service provider raggiungono quote inferiori.

Figura 1. Numero di dataset pubblicati nei singoli archivi

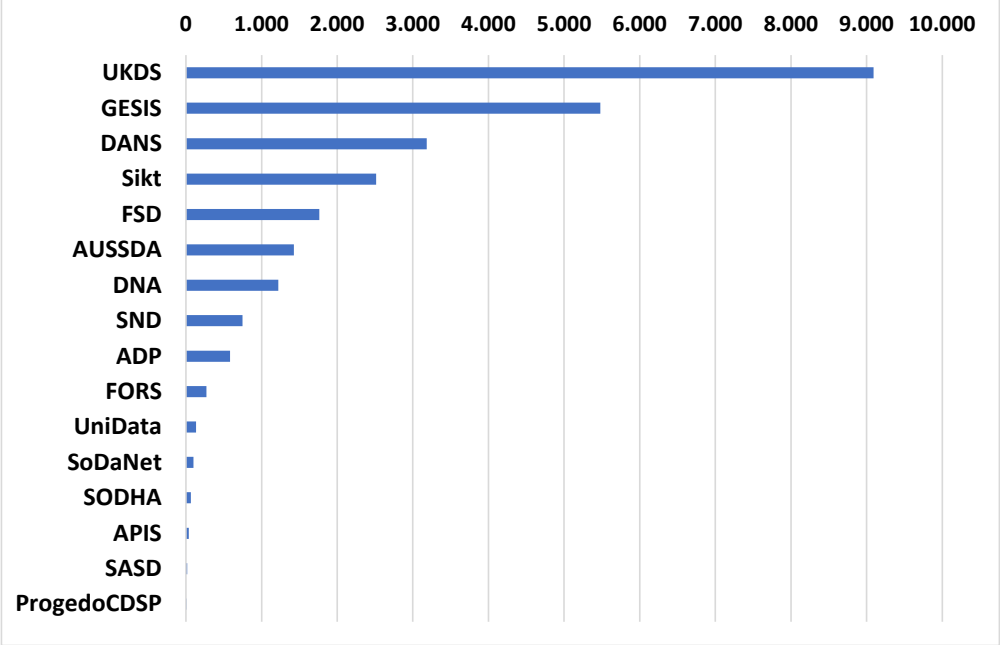
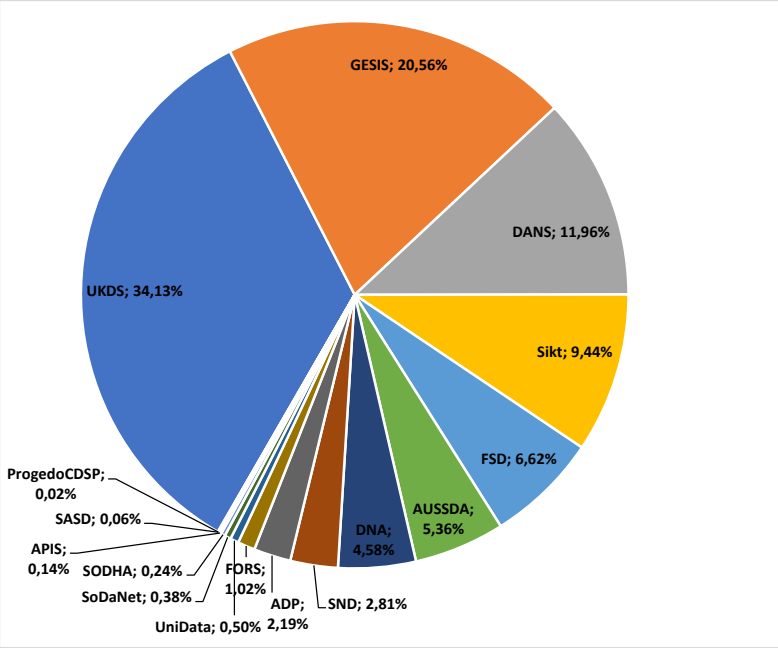


Figura 2. Contributo percentuale dei singoli archivi al CDC



4.2 Lingue utilizzate

Alla data del 21/02/2023, il CDC si compone in totale di 36.577 dataset, di cui 26.641 sono disponibili in lingua inglese. Ogni dataset può essere disponibile in una o più lingue tra le undici gestite dal catalogo.

La lingua è il primo elemento che l'utente deve scegliere quando esegue una ricerca attraverso l'interfaccia web.

La lingua dovrebbe essere disponibile alla variabile *fileLanguages* (CMM no. 4.1.7), il cui contenuto è previsto sia standardizzato ISO 639-1 (Language code). In realtà, questo campo è solitamente riempito con testo libero e difficilmente utilizzabile, con molti errori, sinonimi e imprecisioni. Il dato corretto relativo alla lingua del dataset, corrispondente a quello ricavabile dall'interfaccia web del CDC, è riportato dal campo estratto *LangAvailableIn*, compilato in maniera standardizzata ISO 639-1, nonostante non sia descritto nel CMM né altrove.

Come illustrato dalla **Tabella 5** e dal grafico in **Figura 3,** i dataset sono disponibili in 10 diverse lingue: inglese, tedesco, olandese, danese, francese, finlandese, svedese, greco, sloveno, slovacco. Tabella e grafico mostrano tre informazioni, rispetto a ogni lingua: numero totale di dataset disponibili; quanti di questi sono disponibili anche in lingua inglese; dataset disponibili solo nella lingua di riferimento. 35 dataset sono contemporaneamente disponibili

in tre lingue: inglese, tedesco e francese. Il CDC è già predisposto per accogliere dataset anche in ceco. Il dato sulla lingua riproduce, in molti casi, le dimensioni dell'archivio nazionale quando è idioma ufficiale. I dati in olandese, greco, e quasi tutti quelli in francese, non sono disponibili in altre lingue. Complessivamente, come mostrato dal grafico a torta in **Figura 4**, solo il 27% dei dati non è disponibile in lingua inglese. La disponibilità di dati in lingua inglese in misura maggiore potrebbe ulteriormente incrementare le possibilità di acceso e riuso, favorendo la realizzazione di studi comparativi tra Paesi.

Tabella 5. Numero di dataset disponibili nelle varie lingue

	Numero dataset disponibili	In inglese	Solo una lingua
Inglese*	26.641	-	17.267
Tedesco*	7.352	5.535	1.782
Olandese	4.418	0	4.418
Danese	2.191	1.219	972
Francese*	2.097	50	2.012
Finlandese	1.801	1.764	37
Svedese	748	748	0
Greco	591	0	591
Sloveno	77	76	1
Slovacco	18	17	1
Ceco	0	0	0

Figura 3. Numero di dataset disponibili nelle varie lingue

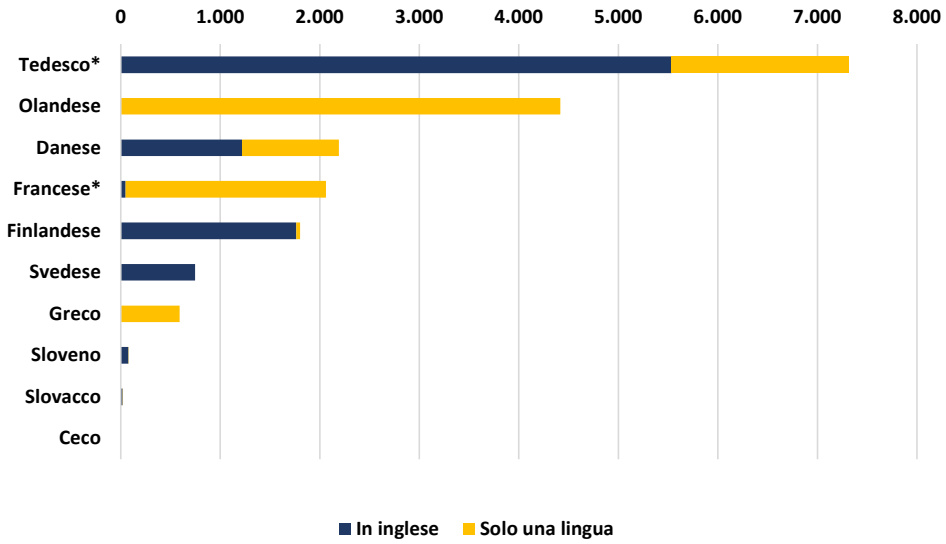
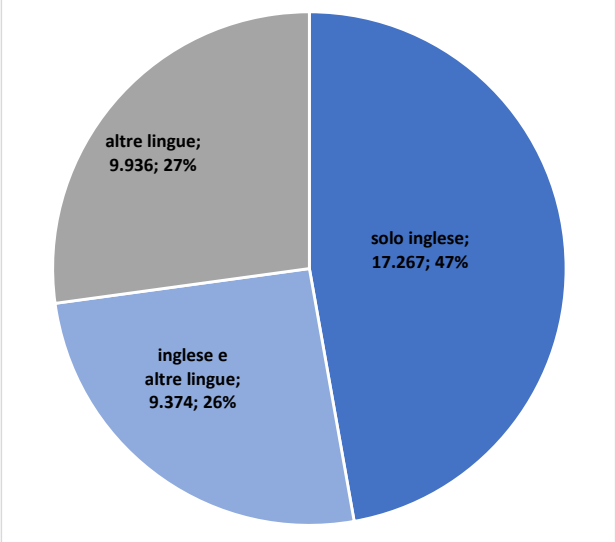


Figura 4. Partizione del catalogo rispetto alla lingua dei dataset



4.3 Countries

I dataset presenti su ogni archivio presente su CDC possono essere riferiti al Paese in cui ha sede, a questo insieme ad altri, oppure a Stati del tutto diversi.

L'informazione sui Paesi interessati è desumibile dall'elemento *studyAreaCountries*, corrispondente al n. 1.3.3 nel CMM, che descrive i paesi nei quali ha avuto luogo lo studio. Da questo si ricavano due variabili: country e abbr. *Country* è compilata con testo libero, *abbr* corrisponde alla codifica ISO 3166 2-letter code.

La compilazione di questo dato è opzionale nel CDC. La seguente **Tabella 6** illustra in che misura questo dato risulta compilato. Mentre è sempre disponibile la compilazione con testo libero (variabile *country*), il dato standardizzato della variabile *abbr* è completamente assente in cinque archivi. Il testo libero è difficilmente utilizzabile a causa dell'elevato numero di errori riscontrati nel dato presente, per i molti sinonimi riscontrati e a causa delle limitazioni tipiche nel raggruppare dati non standardizzati.

Tabella 6. Compilazione studyAreaCountries nelle due variabili "country" e "abbr"

code	totale	abbr	country
UKDS	9.093	nodata	1
GESIS	5.478	1	1
DANS	3.185	1	1
Sikt (NSD)	2.515	1	1
FSD	1.764	1	1
AUSSDA	1.427	1	1
DNA	1.220	dati errati	
SND	748	1	1
ADP	583	1	1
FORS	272	nodata	1
UniData	133	nodata	1
SoDaNet	100	nodata	1
SODHA	64	1	1
APIS	36	1	1
SASD	17	1	1
ProgedoCDSP	6	nodata	1

È stato fatto ricorso al dato ricavabile dalla variabile "abbr" data la standardizzazione, nonostante alcune compilazioni errate riscontrate e l'assenza totale negli archivi UKDS, FORS, UniData, SoDaNet e ProgedoCDSP. Per questi ultimi archivi, i dati sono stati integrati ricavandoli dal campo "country" e procedendo a una paziente pulizia e trasformazione in dato standardizzato ISO 3166 2-letter code attraverso join tabellare e pulizia manuale, tenendo conto di sinonimi, diverse lingue utilizzate per indicare i Paesi, e altro. L'archivio DNA riporta dati errati poiché il contenuto non riferibile chiaramente a Paesi (es: local, national, regional) ed è quindi stato escluso.

Eseguita questa prima operazione al fine di rendere omogenee le informazioni, è stato ricavato in che misura i singoli archivi dispongano di dati riferiti allo stesso Paese in cui hanno sede, o ad altri Stati. Ogni dataset può essere riferito a uno o più Paesi.

Come illustrato dal grafico in **Figura 5**, in riferimento all'intero CDC le collezioni si riferiscono allo stesso Paese nel 45% dei casi, mentre lo escludono nel 6%. Un quarto dei dataset è attinente solo a Paesi diversi da quello dell'archivio. Da notare la mancanza di questa informazione in ben il 24% dei dataset pubblicati. I dati di dettaglio sono riportati nella **Tabella 7**.

Tabella 7. Dataset disponibili suddivisi per archivio e in base all'attinenza con lo stesso Paese sede dell'archivio nazionale

	UKDS		GESIS		DANS		NSD		FSD		AUSSDA		DNA		SND	
	n	%	n	%	n	%	n	%	n	%	n	%	n	%	n	%
Stesso Paese	3.000	32,99	2.210	40,34	87	2,73	2.255	89,66	1.687	95,63	1.345	94,25	dati errati		564	75,40
Stesso Paese e altri	453	4,98	829	15,13	0	0,00	36	1,43	37	2,10	32	2,24			24	3,21
Altri Paesi	5.616	61,76	670	12,23	0	0,00	196	7,79	40	2,27	50	3,50			30	4,01
Non disponibile	24	0,26	1.769	32,29	3.098	97,27	28	1,11	0	0,00	0	0,00			130	17,38
TOT archivio	**9.093**		**5.478**		**3.185**		**2.515**		**1.764**		**1.427**		**1220**		**748**	

	ADP		FORS		UniData		SoDaNet		SODHA		APIS		SASD		ProgedoCDSP	
	n	%	n	%	n	%	n	%	n	%	n	%	n	%	n	%
Stesso Paese	486	83,36	0	0,00	129	96,99	56	56,00	14	21,88	33	91,67	15	88,24	5	83,33
Stesso Paese e altri	46	7,89	262	96,32	1	0,75	0	0,00	5	7,81	0	0,00	0	0,00	1	16,67
Altri Paesi	11	1,89	10	3,68	2	1,50	9	9,00	2	3,13	0	0,00	0	0,00	0	0,00
Non disponibile	40	6,86	0	0,00	1	0,75	35	35,00	43	67,19	3	8,33	2	11,76	0	0,00
TOT archivio	**583**		**272**		**133**		**100**		**64**		**36**		**17**		**6**	

Figura 5. Suddivisione del CDC in base all'attinenza dei dataset con l'archivio in cui sono ospitati

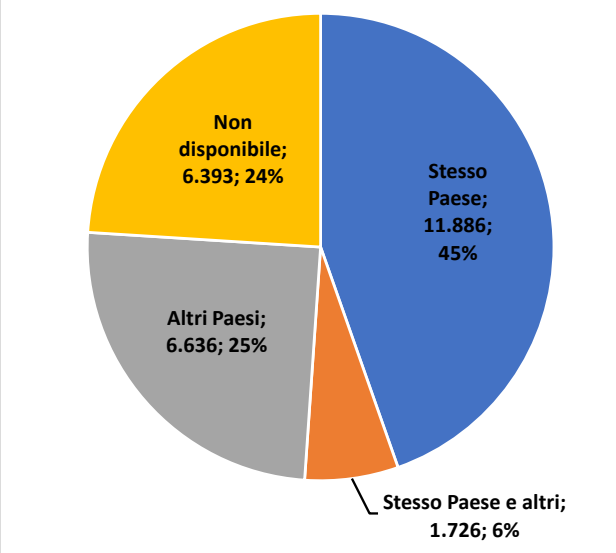
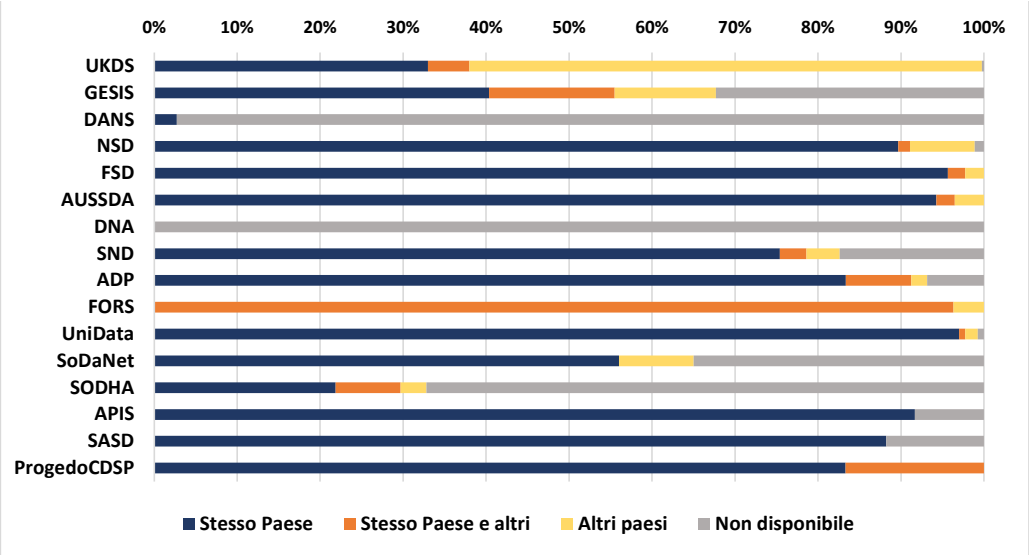


Figura 6. Paesi di riferimento dei dataset rispetto al singolo archivio



A seguito della pulizia dei dati, è stato ricavato quanti dataset corrispondano a ogni Paese identificato secondo la standardizzazione ISO 3166 2-letter code. Il grafico in **Figura 7** e la mappa in **Figura 8** illustrano quanti dataset del CDC si riferiscano, in percentuale, ai Paesi indicati. Sono stati considerati i Paesi interessati da almeno l'1% dei dataset del CDC.

Figura 7. Percentuale dei dataset riferiti ai singoli Paesi, in percentuale

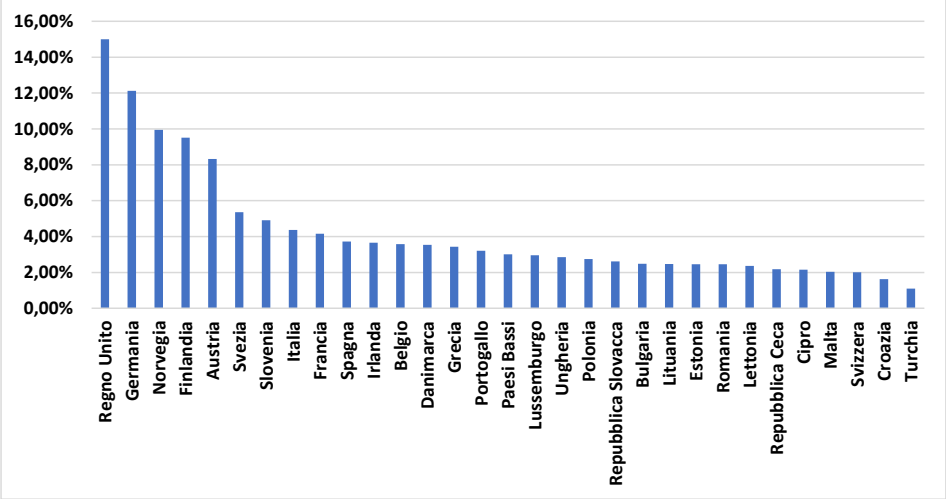
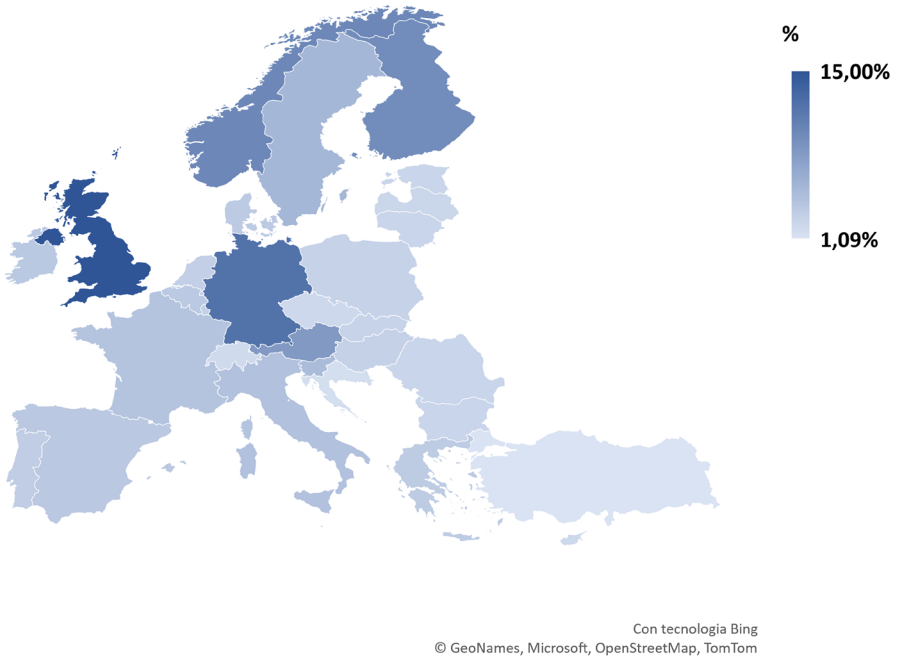


Figura 8. Percentuale dei dataset riferiti ai singoli Paesi, in percentuale



4.4 Principal Investigator

Il *Principal Investigator* identifica il ricercatore e/o l'organizzazione principalmente coinvolti nella produzione dei dati. L'informazione è ricavabile dalla variabile *creators* del CDC, corrispondente all'elemento *Principal Investigator Reference* (CMM, n. 1.1.7).

Si è inteso indagare in che misura i dati disponibili nel CDC siano riconducibili a ricercatori appartenenti a università, oppure pubblicati da altri enti di ricerca e organizzazioni di ogni tipo.

Il CDC non offre la possibilità di isolare i nomi dei ricercatori dalle rispettive affiliazioni, e il dato risulta disponibile in formati molto eterogenei, con l'affiliazione separata da virgole, parentesi o in diversi altri modi.

Si è quindi operato per distinguere tre gruppi sulla base dei nomi delle organizzazioni o delle affiliazioni riportate, isolando i *Principal Investigator*:

- riconducibili a Università;
- riconducibili a istituzioni, enti e organizzazioni diversi da Università;

- Principal Investigator non identificati, poiché nomi di persone non dotati di affiliazione o nomi di organizzazioni che non è stato possibile riconoscere attraverso la procedura seguita.

Il procedimento adottato e lo script eseguito in R sono riportati in appendice al par. 2.
Il risultato è riportato nella **Tabella 8** e nei successivi grafici in **Figura 9**, **Figura 10** e **Figura 11**.

Tabella 8. Principal Investigator distinti per service provider, secondo i tre gruppi individuati. NB: dati limitati agli archivi con oltre 1.000 creators

Archivio	Università	Organizzazione	Non identificati	Totale	Non identificati %
UKDS	9.173	5.725	868	15.766	5,51%
GESIS	7.161	6.266	1.803	15.230	11,84%
DANS	3.100	2.809	1.688	7.597	22,22%
FSD	1.780	1.230	443	3.453	12,83%
NSD	617	1.837	214	2.668	8,02%
AUSSDA	762	1.022	59	1.843	3,20%
ADP	942	620	155	1.717	9,03%
DNA	369	997	340	1.706	19,93%
SND	970	294	137	1.401	9,78%

Figura 9. Principal Investigator divisi nei tre gruppi individuati

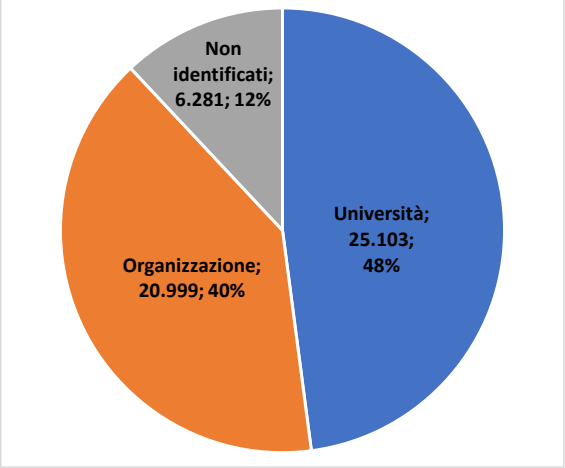


Figura 10. Principal Investigator distinti per service provider. NB: dati limitati agli archivi con oltre 1.000 creators

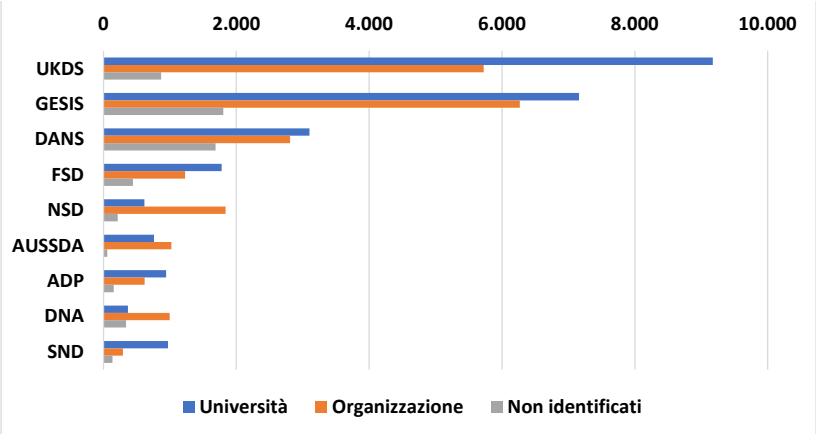
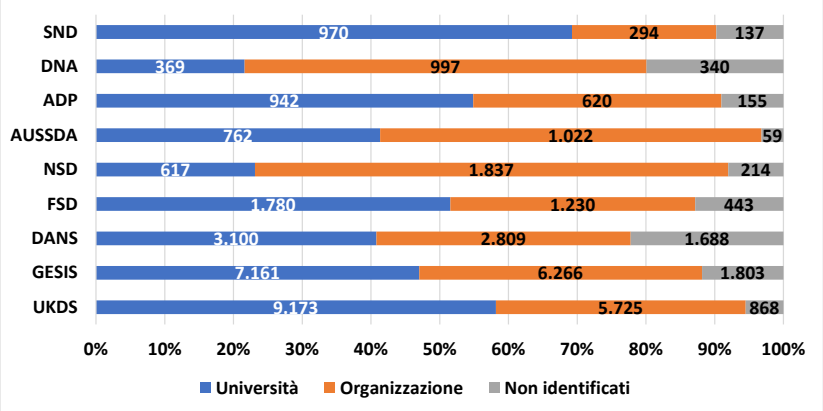


Figura 11. Percentuali dei tre gruppi individuati di Principal Investigator. NB: dati limitati agli archivi con oltre 1.000 creators



Tutti i risultati sono coerenti nel mostrare l'importante e complementare contributo alla condivisione dei dati sia da parte di Università, sia da parte di altri tipi di organizzazione. In percentuale, per gli archivi UKDS e SND prevale il contributo delle Università, mentre nel caso di DNA e NSD quello delle organizzazioni. Tra le organizzazioni impegnate nel conferire dati nei repository emergono, ovviamente, enti di ricerca e governativi, in grado di fornire dati statistici di particolare interesse.

I limiti dell'estrazione eseguita risiedono nella difficoltà di scindere in modo accurato i Principal Investigator nei due gruppi voluti. A tal proposito, sarebbe utile implementare nel CMM una variabile in grado di raccogliere un identificativo univoco dei ricercatori coinvolti,

ad esempio l'*Open Researcher and Contributor ID* (ORCID), o analogo identificativo per le organizzazioni qualora non si tratti di nomi di persone.

Infatti, l'estrazione così eseguita non consente un agevole raggruppamento secondo la distinzione voluta in università e organizzazioni, non essendo disponibile un identificatore per ogni ricercatore o ente, né il più semplice inserimento di questa informazione in un campo separato.

4.5 Data Collection Period

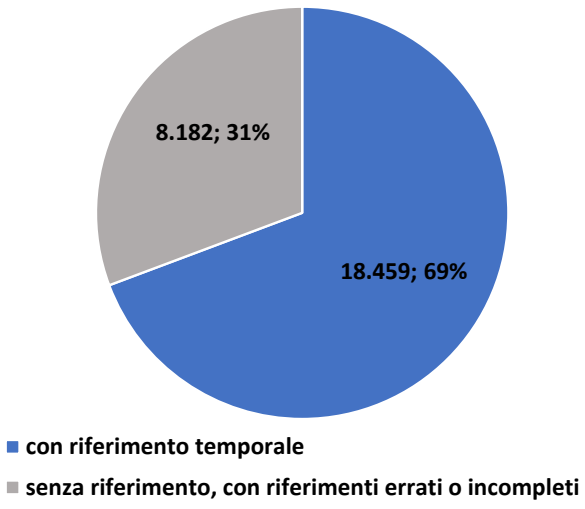
Il periodo in cui è avvenuta la raccolta dei dati (inizio e fine) ci consente di osservare gli anni di riferimento dei dataset disponibili. I campi di interesse sono:

- *dataCollectionPeriodStartdate* del CDC (CMM, no. 1.3.7.1, Data Collection Period Startdate (controlled): Start of data collection;
- *dataCollectionPeriodEnddate* del CDC (CMM, no. 1.3.7.2, Data Collection Period Enddate (controlled): End of data collection.

Un ulteriore campo che descrive il periodo di raccolta, non utilizzato, è *DataCollectionPeriod*, corrispondente a *Data Collection Period (freetext)* del CMM (n. 1.3.7.4), generalmente compilato con l'anno di inizio raccolta.

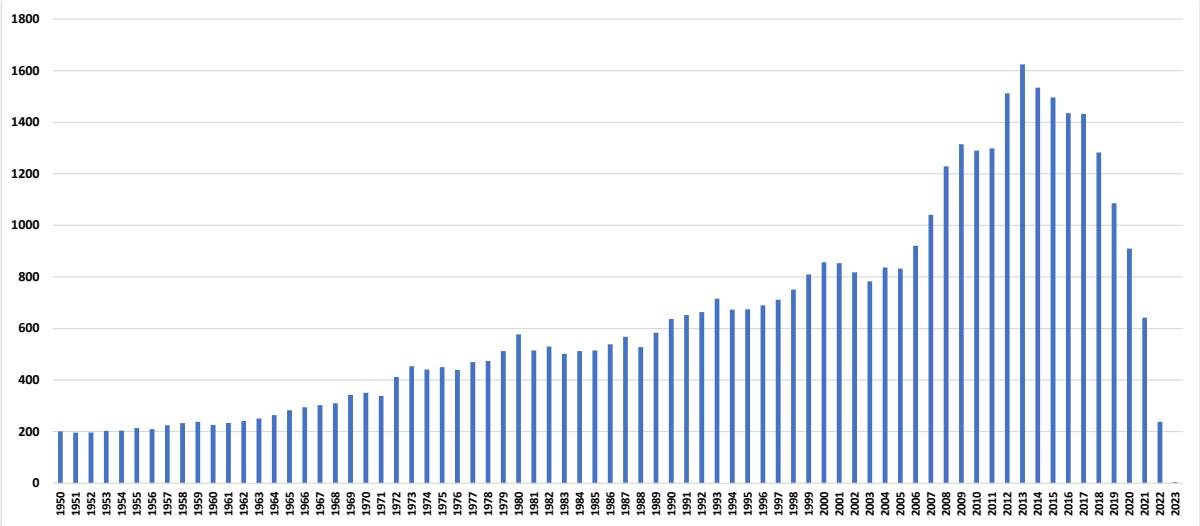
Nel CMM l'inserimento di queste due informazioni è raccomandato. Nonostante ciò, il 31% dei dati non detiene riferimenti temporali, o sono stati inseriti in maniera errata o incompleta (**Figura 12**).

Figura 12. Completezza dell'informazione relativa ai riferimenti temporali della raccolta dati



Nel seguente grafico in **Figura 13** è rappresentato il numero di dataset che si riferiscono a ogni singolo anno, nell'arco temporale tra il 1950 e il 2023.

Figura 13. numero di dataset riferiti ai singoli anni tra il 1950 e il 2023

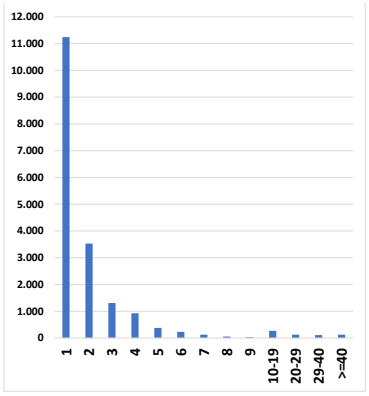


Rispetto ai dati successivi al 1950, la **Tabella 9** e il grafico in **Figura 14** illustrano la quantità di dataset per estensione del periodo di raccolta dei dati, in anni, conteggiato considerando l'anno di inizio raccolta e quello di fine. I dataset i cui dati sono stati raccolti all'interno dello stesso anno (*dataCollectionPeriodStartdate* coincidente con *dataCollectionPeriodEnddate)* sono 11.244 e rappresentano la porzione più consistente. Il dato diminuisce con l'aumentare dell'estensione riportata. Per i periodi superiori ai 9 anni, i dati per anno sono stati aggregati: 270 dataset hanno un'estensione del periodo di raccolta dei dati che varia tra 10 e 19 anni; 124 dataset superano i 40 anni di raccolta.

Tabella 9. Numero di dataset rispetto all'estensione temporale dello studio

n anni	n dataset
1	11.244
2	3.531
3	1.309
4	927
5	379
6	230
7	122
8	53
9	28
10-19	270
20-29	124
29-40	109
>=40	124

Figura 14. Numero di dataset rispetto all'estensione temporale dello studio



Nella **Tabella 10** sono stati riportati i limiti temporali complessivi dei dataset disponibili in ogni archivio nazionale afferente al CDC.

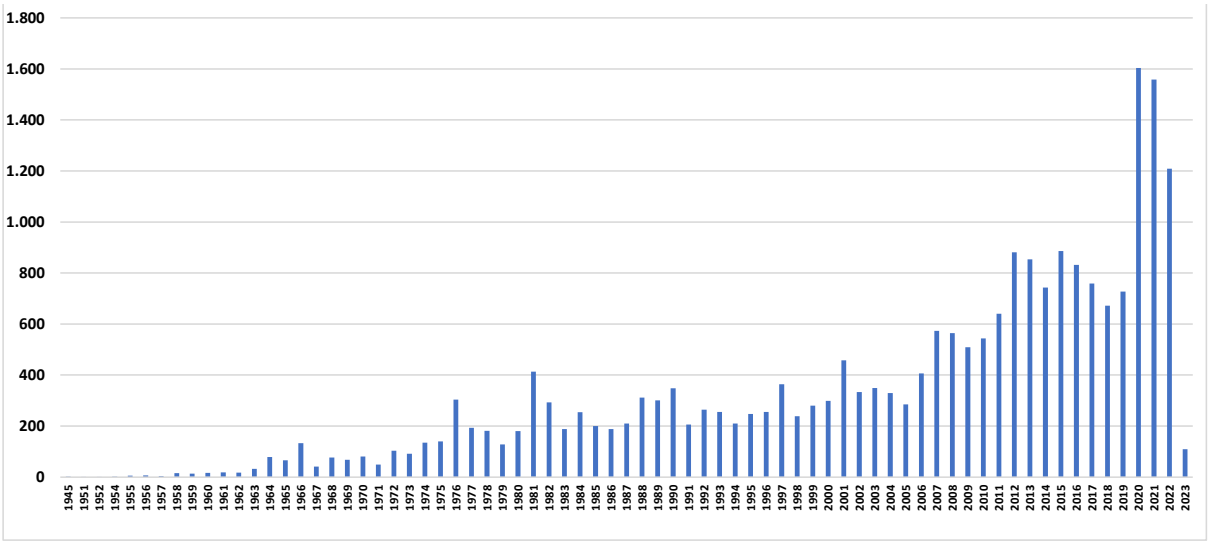
Interessante notare la presenza di 273 dataset il cui anno di fine raccolta è antecedente al 1950 (dataCollectionPeriodEnddate). Di questi, ben 87 riportano un limite temporale di fine raccolta precedente al 1900. Il dataset più antico, il cui periodo di raccolta si estende tra l'anno 1000 e il 1250, appartiene all'archivio GESIS, ed è costituito da dati su 370 documenti papali. Si intitola: "Parchments in Helmarshausen Gospels." (DOI: 10.4232/1.8061; abstract: «Parchment cut and document format in the papal office of the high middle ages. Data on 370 original papal documents in view of the parchments used whose composition enables source-critical references.»).

Tabella 10. Limiti temporali minimo e massimo dei dataset disponibili. Quattro archivi mettono a disposizione dati precedenti al Novecento

code	min.year	max.year
ADP	1962	2020
APIS	1999	2020
AUSSDA	1919	2022
DNA	1787	2011
FSD	1962	2022
GESIS	1000	2022
NSD	1893	2021
ProgedoCDSP	2007	2020
SASD	2002	2021
SND	1911	2022
SoDaNet	1987	2021
SODHA	2008	2022
UKDS	1826	2023
UniData	2001	2022

Un altro dato temporale interessante è l'anno di pubblicazione del dataset, disponibile attraverso la variabile *publicationYear*, corrispondente a *Publication Year (controlled [YYYY])* del CMM (n. 1.1.9). Il riferimento corrisponde all'anno in cui i dati dello studio sono diventati pubblicamente disponibili, o lo saranno qualora sia in atto un periodo di embargo. L'istogramma in **Figura 15** illustra il numero di dataset pubblicati per ogni anno a partire dal 1945.

Figura 15. Numero di dataset pubblicati ogni anno



Da notare che i primi anni riportati della serie precedono la data di fondazione dei primi archivi. Si tratta di dataset attribuiti agli archivi UKDS, fondato nel 1967, e GESIS, creato nel 1960. Nonostante la variabile, secondo quanto riportato nel CMM, sia da riferirsi all'anno di pubblicazione del dato, è possibile che questo dato sia stato compilato, in alcuni casi, riportando l'anno di pubblicazione dell'articolo scientifico o altra produzione editoriale a cui i dataset sono riferiti, oppure con l'anno di fine raccolta, o in altra maniera. Si tratta, evidentemente, di un errore. È sufficiente consultare il dataset più vecchio (publication year 1945) per cogliere il disallineamento. Si tratta di un dataset dal titolo "Decatur Study (Personal Influence)" (DOI: 10.4232/1.0241), consultabile nell'archivio GESIS e relativo ai famosi studi di Paul Lazarsfeld e Elihu Katz sul flusso di comunicazione a due stadi: i dati sono stati raccolti tra il giugno e l'agosto del 1945, le pubblicazioni correlate risalgono al 1955 e al 1962.

4.6 Altri metadati e vocabolari utilizzati

L'analisi della metadatazione del CDC è proseguita sulle informazioni il cui inserimento prevede l'uso di un vocabolario controllato: topics, modalità di raccolta del dato (*mode of collections*), dimensione temporale (*time methods*), unità d'analisi (*unit type*).

Ogni singolo metadato, nella procedura di estrazione, dà origine a quattro variabili illustrate nella seguente **Tabella 11**.

Tabella 11. Variabili di livello inferiore disponibili per i metadati

Nome	Descrizione
vocab	Nome del vocabolario utilizzato
vocabUri	URI del vocabolario
id	Value of the code: rimane uguale qualsiasi sia la lingua utilizzata
term	Termine descrittivo come riportato nel vocabolario, in una delle lingue in cui è codificato

L'analisi è stata condotta su ogni singolo archivio, rintracciando eventuali ricorsività circa completezza e correttezza dei dati inseriti, nonché dei vocabolari utilizzati.

A seguito dell'estrazione e analisi dei dati per singolo archivio, è seguito un confronto complessivo per quelli che adoperano i vocabolari controllati *CESSDA Topic Classification*, per i topic, e *DDI Alliance Controlled Vocabulary* relativamente a modalità di raccolta, unità d'analisi e dimensione temporale, ricorrendo alla variabile *term*.

Il *DDI Alliance Controlled Vocabulary* è costituito da un gruppo di vocabolari controllati per la produzione dei metadati nelle scienze sociali (https://ddialliance.org/controlled-vocabularies). Permettono di descrivere i metadati presenti nel CDC e nel CMM, ad esempio unità d'analisi, modalità di raccolta e altro. Creati dalla Data Documentation Initiative Alliance (DDI Alliance), costituita nel 2003 (https://ddialliance.org/about-the-alliance), sono stati adottati da numerosi repository oltre che da CESSDA. Tutti i vocabolari impiegati da CESSDA sono disponibili nell'apposita sezione *CESSDA Vocabulary Service*, disponibile all'indirizzo https://vocabularies.cessda.eu.

Nella maggior parte dei casi, come illustrato dalla **Tabella 12**, non risulta compilata la variabile *id*, corrispondente all'identificatore univoco del *term* o *code value* dei vocabolari utilizzati, che resta uguale per qualsiasi lingua impiegata ed è leggibile da macchina, requisito fondamentale previsto dai principi FAIR (Wilkinson et al., 2016, p. 3). La mancanza di questo dato disincentiva particolarmente qualsiasi analisi dei contenuti, poiché impedisce di operare in maniera agevole.

Tabella 12. Compilazione ID dei term presente (1) o non presente (0)

code	totale	ModeOfCollections.id	TimeMethods.id	UnitType.id	classifications.id
ADP	583	**1**	**1**	**1**	**1**
APIS	36	0	0	0	0
AUSSDA	1.427	0	0	0	0
DANS	3.185	0	0	0	0
DNA	1.220	**1**	**1**	**1**	0
FORS	272	0	0	0	0
FSD	1.764	**1**	**1**	**1**	**1**
GESIS	5.478	**1**	**1**	**1**	0
ProgedoCI	6	0	0	0	0
SASD	17	0	0	0	0
Sikt (NSD)	2.515	0	0	0	0
SND	748	0	0	0	0
SoDaNet	100	0	0	0	0
SODHA	64	0	0	0	0
UKDS	9.093	0	0	0	0
UniData	133	0	0	0	0
TOTALE	**26.641**				

Nel processo di analisi del contenuto delle variabili, si è tenuto conto di eventuali variazioni intercorse nelle differenti versioni dei vocabolari controllati.

Per i topics, gli archivi nazionali che adoperano vocabolari alternativi al CESSDA Topic CV o unitamente a questo sono: DANS, utilizza NARCIS-classification (https://www.narcis.nl/classification/Language/en); FSD utilizza OKM - Finnish Thesaurus and Ontology Service (https://finto.fi/okm-tieteenala/en); SND impiega in maniera complementare il CESSDA Topic CV e Standard för svensk indelning av forskningsämnen 2011 (https://www.scb.se/dokumentation/klassifikationer-och-standarder/standard-for-svensk-indelning-av-forskningsamnen).

Il prospetto in **Tabella 13** offre una panoramica sulla disponibilità del dato e sui vocabolari utilizzati. In verde sono evidenziati i dati estratti, analizzati e successivamente messi a confronto.

Tabella 13. Disponibilità dei dati (term) e vocabolari utilizzati relativamente a quattro metadati

code	totale	ModeOfCollections	TimeMethods	UnitType	classifications
UKDS	9.093	DDI Alliance CV	N.I.	N.I.	CESSDA CV
GESIS	5.478	DDI Alliance CV	DDI Alliance CV	DDI Alliance CV	CESSDA CV
DANS	3.185	N.D.	N.D.	N.D.	NARCIS-classification
Sikt (NSD)	2.515	DDI Alliance CV	N.I.	DDI Alliance CV	CESSDA CV
FSD	1.764	DDI Alliance CV	DDI Alliance CV	DDI Alliance CV	OKM
AUSSDA	1.427	DDI Alliance CV	DDI Alliance CV	DDI Alliance CV	CESSDA CV
DNA	1.220	DDA Data Collection Methodology	DDA Time Method	DDA Analysis Unit	CESSDA CV
SND	748	DDI Alliance CV	DDI Alliance CV	DDI Alliance CV	CESSDA CV + Standard för svensk indelning av forskningsämnen 2011
ADP	583	DDI Alliance CV	DDI Alliance CV	DDI Alliance CV	CESSDA CV
FORS	272	N.D.	N.D.	N.D.	CESSDA CV
UniData	133	DDI Alliance CV	DDI Alliance CV	DDI Alliance CV	CESSDA CV
SoDaNet	100	DDI Alliance CV	DDI Alliance CV	DDI Alliance CV	CESSDA CV
SODHA	64	N.I.	N.D.	N.I.	CESSDA CV
APIS	36	DDI Alliance CV	DDI Alliance CV	DDI Alliance CV	CESSDA CV
SASD	17	N.I.	N.I.	DDI Alliance CV	N.I.
ProgedoCDSP	6	DDI Alliance CV	DDI Alliance CV	DDI Alliance CV	CESSDA CV
TOTALE	**26.641**				

DDI Alliance CV o CESSDA CV
Parzialmente CESSDA CV e altro CV
Altro vocabolario (nome)
N.I.= testo libero o non identificato
N.D.= dato non disponibile

Numerosi errori nella compilazione di un dato, di norma dovrebbe essere standardizzato e compilato mediante l'uso di vocabolari controllati, hanno reso particolarmente oneroso svolgere l'attività di analisi. Tra le difficoltà riscontrate:

- la completa indisponibilità di dati per alcuni casi;
- la mancata informazione circa il vocabolario utilizzato;
- il mancato inserimento dell'id;
- l'uso di etichette riportate erroneamente, nonostante sia adoperato il CESSDA cv;
- l'uso di testo libero non rispondente a etichette da vocabolario ma a descrizioni.

Da questo scenario è derivata la necessità di procedere accuratamente alla verifica delle etichette e a operazioni di correzione, anche al fine di consentire join tabellari in vista del confronto tra archivi. Sono state verificate eventuali etichette esistenti nelle precedenti versioni del CESSDA CV e non più esistenti.

La correzione ha tenuto conto del contenuto riportato con meri errori materiali, di copiatura o simili, intervenendo in tutti quei casi in cui era interpretabile la volontà di utilizzare il CESSDA VC per i topic o il DDI Alliance VC per gli altri metadati. Infatti, sono frequenti i casi in cui è stato compilato il dato impiegando un vocabolario controllato, ma senza riportarne il nome nell'apposito campo.

4.7 Topics (Classifications)

L'elemento *classifications* illustra gli argomenti al quale lo studio si riferisce. Nell'interfaccia del CDC corrisponde a Topic, e nel CMM è riportato al n. 1.2.2, *Study topic (controlled)*.

Nonostante nella compilazione di questo dato vi sia l'obbligo di utilizzare il CESSDA CV for CESSDA Topic Classification (https://vocabularies.cessda.eu/vocabulary/TopicClassification), al momento nel CDC sono riportati anche termini provenienti da altri vocabolari, come riportato nel CMM (vedi colonna "Mapping information: Notes for DDI 2.5 Schema").

Il CESSDA CV è strutturato gerarchicamente in due livelli. I termini del secondo, ben più numerosi, sono raggruppati nel primo, riportato nel CV a carattere maiuscolo. Dopo aver raggruppato i termini impiegati, gli stessi sono stati ricondotti ai termini di primo livello. In seguito, anche questi sono stati raggruppati. Ad ogni dataset possono essere attribuiti uno o più topic.

Il confronto è stato limitato solo gli archivi che adoperano il CESSDA CV. Sono rimasti esclusi, di conseguenza, DANS, FSD e SASD (vedi **Tabella 13**).

Il grafico in **Figura 16** illustra la percentuale d'uso di ogni singola etichetta sul totale dei dataset pubblicati su CDC in lingua inglese. Da notare il dato #N/D relativo al contenuto che non è stato possibile identificare perché errato o non conforme.

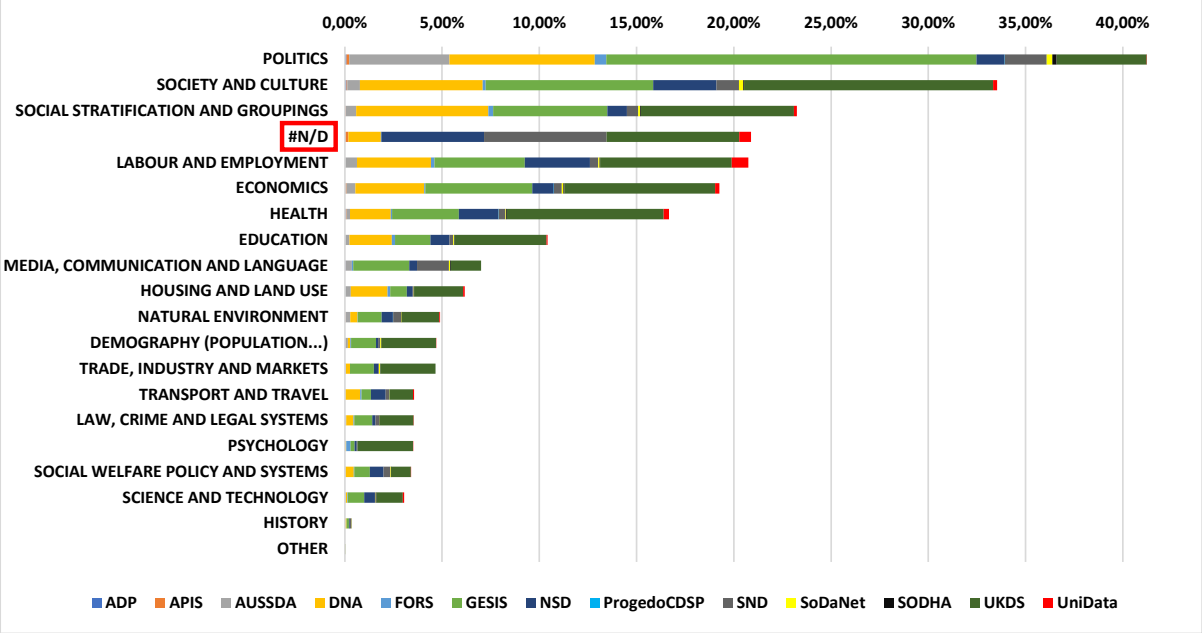
Le quattro tematiche di primo livello più ricorrenti sono:

- POLITICS: raggruppa, tra gli altri, dati su attitudini e orientamenti politici dei cittadini, o derivanti da survey in generale, ed anche su servizi e policy;
- SOCIETY AND CULTURE: categorie che comprende dataset su attitudini, orientamenti e attività culturali, uso del tempo, partecipazione;
- SOCIAL STRATIFICATION AND GROUPINGS: contiene dati su famiglia e matrimonio, mobilità sociale e occupazionale, diseguaglianze, questioni di genere;
- LABOUR AND EMPLOYMENT: condizioni di lavoro, occupazione, disoccupazione.

Ognuna delle quattro tematiche illustrate è attribuita a oltre il 20% dei dataset disponibili su CDC. Gli argomenti individuati ci offrono una tendenza di massima su quali ambiti delle scienze sociali siano più interessati dalla condivisione dei dati negli archivi nazionali. Le

tematiche più ricorrenti, di conseguenza, corrispondono agli interessi disciplinari che potrebbero maggiormente trovare nel CDC un'utile fonte dove reperire dati adatti al riuso.

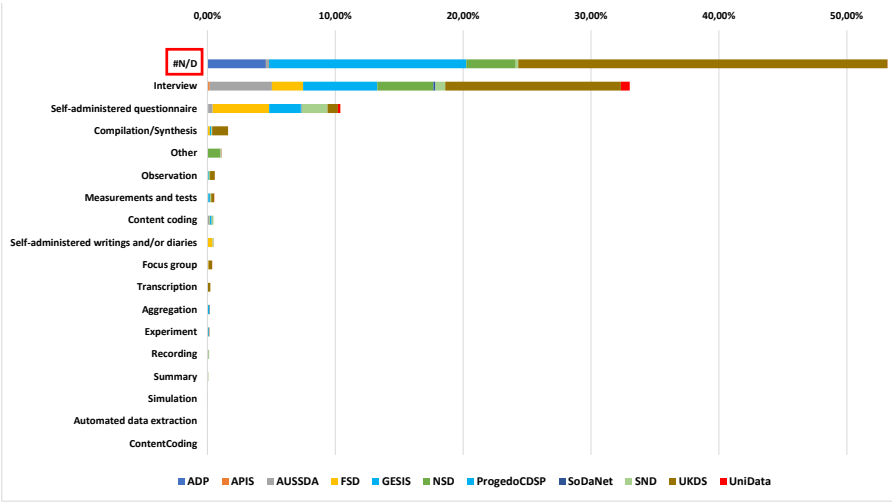
Figura 16. Argomenti dei dataset aggregati per topic di primo livello (CESSDA CV for CESSDA Topic)



4.8 Modalità di raccolta (*TypeOfModeOfCollections*)

Si è proceduto a raggruppare i termini utilizzati relativamente alle modalità di raccolta del dato, disponibili alla variabile *typeOfModeOfCollections* corrispondente a *Mode of Data Collection Vocabulary* (n. 1.3.6.4.2) nel CMM. In percentuale sul totale dei dataset disponibili su CDC in lingua inglese, le modalità di raccolta più ricorrenti risultano le interviste e il questionario (**Figura 17**). Da notare le basse percentuali dovute alla generale scarsa compilazione di questo dato, e alla presenza di di termini non identificati o compilazioni errate in oltre la metà dei casi.

Figura 17. Modalità di raccolta

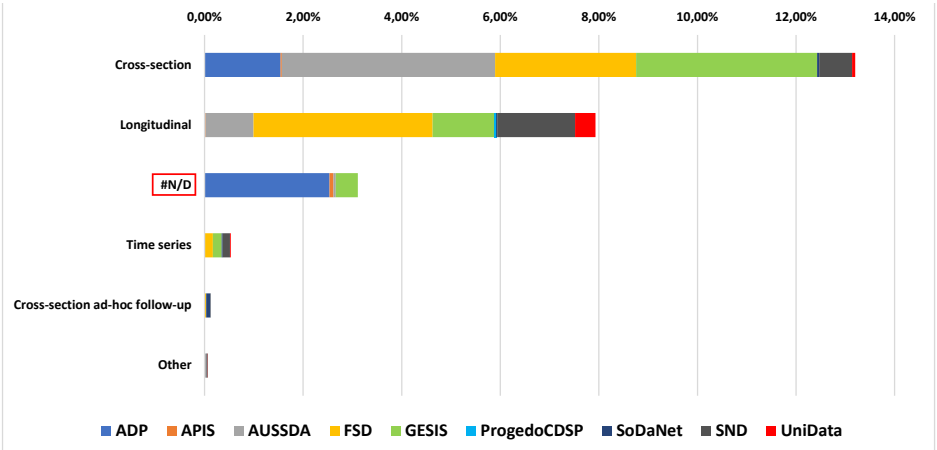


4.9 Dimensione temporale (*timeMethods*)

La dimensione temporale dei dataset è descritta dalla variabile *typeOfTimeMethods*, corrispondente a *Time Method (controlled)* nel CMM (n. 1.3.1). Il contenuto del metadato relativo alla dimensione temporale in molti casi non è completo o ben compilato (vedi **Tabella 14**). In particolare, il dato è assente o carente nei quattro archivi di più grandi dimensioni: UKDS, GESIS, DANS, NSD.

Longitudinal e Cross-section sono i tipi di dimensione temporale più ricorrenti (**Figura 18**).

Figura 18. Dimensione temporale

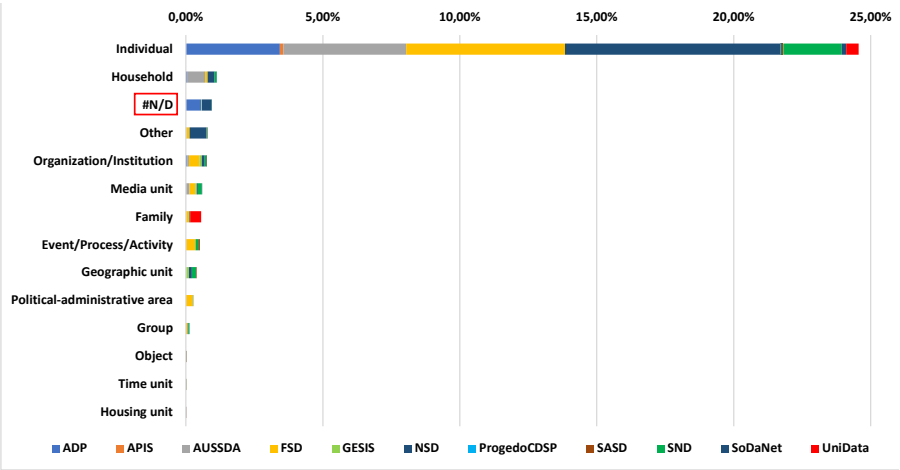


4.10 UnitType

Altro dato interessante ma non molto disponibile (vedi anche **Tabella 14** a p. 37) è l'unità di analisi utilizzata, disponibile alla variabile *UnitType* corrispondente a *Unit of Analysis (controlled)* del CMM (n. 1.3.5). La base individuale rappresenta l'attributo più ricorrente (**Figura 19**). Anche in questo caso la compilazione del dato è spesso assente, in particolare in quattro degli archivi con oltre mille contributi (UKDS, GESIS, DANS e DNA).

L'unità di analisi è sempre un dato particolarmente vincolante nell'ambito di qualsiasi ricerca. Per tale motivo, il suo inserimento nella metadatazione è particolarmente utile per avere ben chiaro l'ambito di indagine o per il ricercatore vincolato a una particolare unità di analisi per diversi motivi, e intenda selezionare di conseguenza solo ciò che interessa.

Figura 19. Unità d'analisi



4.11 Contenuto non rintracciabile

È stata ricavata, per archivio e metadato, la percentuale di contenuto non correttamente riferito a un termine dei vocabolari utilizzati sul totale del singolo archivio (*CESSDA Topic Classification CV* o *DDI Alliance CV*). I risultati sono illustrati nella **Tabella 13** e corrispondono alla categoria #N/D nei grafici precedenti (da **Figura 16** a **Figura 19**).

Da notare, relativamente al dato sulle modalità di raccolta (*Mode of Collection*), l'alta percentuale raggiunta dagli archivi UKDS, GESIS e NSD. Dal calcolo sono stati esclusi gli archivi per i quali non è stato possibile identificare il contenuto dei metadati, poiché errato, non riferito a un vocabolario controllato o contenente testo libero.

Occorre notare il caso particolare dell'archivio SND, che raggiunge una bassa percentuale in riferimento ai topic (*classifications*). Ciò è dovuto anche a un parziale utilizzo del *CESSDA Topic Classification CV* complementare al vocabolario *Standard för svensk indelning av forskningsämnen 2011*. Nella **Tabella 13** il dato è riportato in colore rosso.

Tabella 13. Percentuale di contenuto dei metadati non identificato. Il dato relativo all'archivio SND è in colore rosso perché il CESSDA Topic Classification CV è impiegato solo parzialmente

	ModeOfCollections	TimeMethods	UnitType	classifications
UKDS	63,01%	N.I.	N.I.	8,61%
GESIS	63,06%	8,12%	5,88%	0,00%
DANS	N.D.	N.D.	N.D.	
NSD	41,64%	N.I.	3,79%	23,57%
FSD	0,05%	0,00%	0,00%	
AUSSDA	4,15%	0,84%	0,21%	0,00%
DNA				4,34%
SND	5,84%	0,00%	0,13%	42,94%
ADP	96,21%	62,19%	13,70%	0,00%
FORS	N.D.	N.D.	N.D.	1,40%
UniData	0,41%	0,00%	0,00%	22,41%
SoDaNet	0,00%	0,00%	0,00%	0,00%
SODHA	N.I.	N.D.	N.I.	0,00%
APIS	2,50%	61,76%	0,00%	32,89%
SASD	N.I.	N.I.	0,00%	N.I.
ProgedoCDSP	0,00%	0,00%	0,00%	0,00%

Altro vocabolario
N.I.= testo libero o non identificato
N.D.= dato non disponibile

4.12 Completezza metadatazione

È stata calcolata la percentuale di compilazione dei principali metadati sul totale per ogni singolo archivio, come illustrato nella **Tabella 14**. La percentuale non attesta la qualità di compilazione del dato riscontrato ma indica in che misura è presente.

Dal totale dei dataset sono stati sottratti, per ogni metadato, quelli con dato mancante. Sono stati esclusi i metadati vuoti o riportanti testualmente 'NA' e 'NULL' (derivanti da indisponibilità iniziale così descritta o emersi nelle procedure di scaricamento e analisi). Di conseguenza, nell'eseguire l'interrogazione *SQL* è stato previsto per ogni archivio/metadato: != 'NULL'; != 'NA'; IS NOT NULL.

In seguito, è stata calcolata la percentuale dei rimanenti sul totale, rappresentante l'avvenuta compilazione del metadato.

La **Tabella 14** offre una visione complessiva dalla percentuale di compilazione dei metadati rispetto a ogni service provider. Da notare la maggiore completezza nella compilazione dei topic (*classifications*), mentre per gli altri metadati si riscontrano carenze diffuse. I metadati risultano particolarmente carenti in riferimento a quattro archivi: DANS; FORS; SoDaNet e SODHA. AUSSDA è l'unico archivio che raggiunge il 100% di compilazione in ogni metadato.

Tabella 14. Percentuali di compilazione dei metadati

code	totale	classifications	country	abbr	dataCollectionPeriodStartdate	dataCollectionPeriodEnddate	typeOfModeOfCollections.term	typeOfTimeMethods.term	unitTypes.term
UKDS	9093	96,62	99,73	0,00	87,31	74,23	93,38	68,48	99,87
GESIS	5478	83,75	67,71	67,71	96,13	72,44	93,63	26,78	1,22
DANS	3185	100,00	2,73	2,73	0,00	0,00	0,00	0,00	0,00
NSD	2515	99,40	99,64	98,89	96,82	96,74	98,25	98,61	99,72
FSD	1764	100,00	100,00	100,00	100,00	80,90	99,77	100,00	100,00
AUSSDA	1427	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00
DNA	1220	89,18	99,67	0,00	100,00	99,51	99,02	98,85	0,00
SND	748	100,00	82,75	82,62	78,74	74,33	72,99	82,62	92,38
ADP	583	11,84	95,71	93,14	94,17	79,76	99,66	99,49	98,63
FORS	272	96,32	100,00	0,00	0,00	0,00	0,00	0,00	0,00
UniData	133	100,00	99,25	0,00	88,72	88,72	100,00	100,00	100,00
SoDaNet	100	81,00	65,00	0,00	41,00	40,00	48,00	47,00	51,00
SODHA	64	98,44	35,94	32,81	10,94	10,94	10,94	3,13	7,81
APIS	36	100,00	91,67	91,67	88,89	88,89	88,89	94,44	97,22
SASD	17	100,00	100,00	88,24	94,12	94,12	100,00	76,47	100,00
ProgedoCDSP	6	100,00	100,00	0,00	100,00	100,00	100,00	100,00	83,33

Complessivamente, la carenza di compilazione rispetto ad alcuni metadati è inversamente proporzionale alla possibilità di reperimento e riutilizzo, soprattutto se si intende eseguire una ricerca condotta su più archivi. Come è possibile notare tenendo conto delle caratteristiche generali degli archivi, la presenza di una metadatazione più completa non dipende dalle dimensioni né dall'età dell'archivio.

Occorre anche considerare che la compilazione della maggior parte dei metadati è facoltativa secondo quanto previsto dal CMM, seppur raccomandata, e che la mancanza dei dati nel CDC potrebbe essere dovuta a problemi di *harvesting* dei dati presso l'archivio originale.

Una metadatazione completa e accurata, aderente a principi FAIR, favorisce le pratiche di riuso e accessibilità del dato, oltre a renderlo reperibile se si intende considerare ambiti tematici o altri aspetti specifici.

5. Conclusioni

Le pratiche di condivisione dei dati nelle scienze sociali rivestono un ruolo fondamentale nello sviluppo odierno e futuro delle opportunità di ricerca. Oggi, con l'aumento esponenziale

della produzione di dati, la ricerca attinge in misura sempre crescente a fonti che si discostano da quelle tradizionali: *social media*, documenti condivisi in rete, servizi erogati tramite applicazioni su *smartphone*, dati ricavati dai più diversi sensori e dispositivi sempre presenti nella vita quotidiana. Le nuove possibilità della ricerca transitano anche attraverso una necessaria cultura dell'Open Data e delle pratiche di condivisione. Per questi motivi, risulta preziosissimo il lavoro svolto dagli archivi di dati e da CESSDA, le cui attività formative e di erogazione di servizi devono trovare un crescente consenso all'interno della comunità scientifica.

La cultura dell'Open Data dovrà necessariamente coinvolgere maggiormente anche l'editoria scientifica. Un modo per incentivare le pratiche di condivisione, infatti, potrebbe essere quello di conferire un valore aggiunto alle proposte ricevute degli autori quando corredate dalla pubblicazione dei dati prodotti nei repository certificati.

Come premesso nell'introduzione, dopo un lungo percorso storico iniziato negli anni Settanta, oggi gran parte dei Paesi europei sono dotati di un proprio archivio nazionale di riferimento, e coinvolti nell'ambizioso progetto di CESSDA di riunire tutti i dati conferiti in un unico catalogo. Nel lavoro proposto, si è cercato di descrivere lo stato dell'arte del CDC, strumento chiave dell'infrastruttura. Da quanto illustrato sin qui, emerge tutto il potenziale presente e futuro. Il CDC permette il reperimento di dati su diversi campi delle scienze sociali, favorendo la possibilità di realizzare studi comparati, e illustrando le sensibilità tematiche scientifiche presenti nei vari archivi europei. Non è di secondaria importanza l'obiettivo, garantito dai service provider aderenti a CESSDA, di assicurare la conservazione a lungo termine dei dati depositati, assicurandone accesso e riuso anche rispetto agli scenari tecnologici futuri.

Il raggiungimento di questi e altri obiettivi può essere migliorato superando le criticità sulla metadatazione rilevate nel CDC allo stato attuale. Occorre incentivare sempre di più le pratiche di condivisione del dato fornendo gli strumenti e le competenze necessarie per garantire la qualità del processo, in modo da assicurare la maggiore adesione possibile ai principi FAIR. Per tale motivo, i singoli service provider, unitamente alla struttura CESSDA in generale, sono costantemente impegnati nel promuovere attività di formazione, con particolare interesse alla gestione del dato e alla metadatazione. I materiali diffusi da CESSDA, quali la *Data Archiving Guide (DAG)* e la *CESSDA Data Management Expert Guide (DMEG)*, unitamente a slides, video e altri materiali disponibili sulle piattaforme, sono preziosissimi nel favorire pratiche di condivisione sempre più consapevoli dalle quali possa derivare il riuso dei dati della ricerca. È importante che strategie di outreach mirate e incentivi alla condivisione incontrino l'interesse di tutta la comunità scientifica.

Nonostante la ricchezza dei dataset raccolti dal CDC, le difficoltà riscontrate nel lavoro proposto risiedono in una metadatazione spesso carente o, fatto ancor più grave, errata. Ad esempio, la mancanza degli ID dei *term*, corrispondenti a *code* nei vocabolari controllati,

nonostante siano impiegati i vocabolari controllati, rende difficoltoso, o a volte impossibile, il reperimento accurato dei dati rispetto a qualsiasi finalità di ricerca, e impedisce che gli stessi siano leggibili da macchina. Stesso problema riscontrato anche rispetto ad altri metadati, quando il contenuto non rispecchia fedelmente ciò che è stabilito dal CMM e dai vocabolari.

Oltre a incentivare la formazione sulle pratiche di condivisione, risulta necessario, di conseguenza, una maggiore attenzione da parte degli archivi nazionali, relativamente alle fasi di acquisizione dei dataset. Un controllo efficace nella metadatazione di quanto condiviso consentirebbe l'abbattimento quasi totale di tutte le criticità in termini di errori, consentendo di archiviare i dataset in modo accurato e realmente utile in previsione del riutilizzo.

6. Bibliografia e riferimenti

Bischoff, Frank M. 1998. Parchments in Helmarshausen Gospels. GESIS Data Archive, Cologne. ZA8061 Data file Version 1.0.0, https://doi.org/10.4232/1.8061.

Banović Jelena, and Bradić-Martinović Aleksnadra. *Meta-data specification for the description of social science data resources – CESSDA Metadata Model*. Paper presented at Sinteza 2021 - International Scientific Conference on Information Technology and Data Related Research, Singidunum University, Belgrade, Serbia, 2021. doi:10.15308/Sinteza-2021-193-199.

CESSDA. 2018. *CESSDA Strategy 2018-2022*.

CESSDA. 2019. CMM CESSDA Metadata Model, 1.0 (15 November 2019), DOI: 10.5281/zenodo.3543756.

CESSDA Controlled Vocabulary for CESSDA Topic Classification, Accessed January 31, 2023. https://vocabularies.cessda.eu/vocabulary/TopicClassification.

CESSDA Data Catalogue Search API, Accessed January 31, 2023. https://api.tech.cessda.eu.

CESSDA Data Catalogue User Guide, vers. 3.2.0, Accessed April 06, 2023. https://datacatalogue.cessda.eu/documentation.

CESSDA DC Data Catalogue, Accessed January 31, 2023. https://datacatalogue.cessda.eu.

CESSDA Vocabulary Service, Accessed January 31, 2023. https://vocabularies.cessda.eu.

CESSDA, *CESSDA Data Archiving Guide*, Accessed April 06, 2023. https://dag.cessda.eu.

CESSDA, Data Archive Social Sciences Italy – DASSI, Accessed April 06, 2023. https://www.cessda.eu/About/Consortium-and-Partners/List-of-Service-Providers/Italy-spid1908.

CESSDA, *Data Management Expert Guide*, Accessed April 06, 2023. https://dmeg.cessda.eu.

CESSDA, DDI Alliance Controlled Vocabulary for Analysis Unit, Accessed April 06, 2023. https://vocabularies.cessda.eu/vocabulary/AnalysisUnit.

CESSDA, DDI Alliance Controlled Vocabulary for Mode Of Collection, Accessed April 06, 2023. https://vocabularies.cessda.eu/vocabulary/ModeOfCollection?lang=en.

CESSDA, DDI Alliance Controlled Vocabulary for Time Method, Accessed April 06, 2023. https://vocabularies.cessda.eu/vocabulary/TimeMethod?lang=en.

CESSDA, List of Service providers, Accessed April 06, 2023. https://www.cessda.eu/About/Consortium-and-Partners/List-of-Service-Providers.

DASSI, Data Archive for Social Sciences in Italy, Accessed November 28, 2023. https://www.dassi-archive.it.

Data Documentation Initiative Alliance, Accessed April 06, 2023. https://ddialliance.org/about-the-alliance.

Dekker Ron. 2020. "Social data: CESSDA best practices" in *Data Intelligence*, 2, 220–229. DOI: 10.1162/dint_a_00044.

DDI Alliance Controlled Vocabulary, Accessed April 06, 2023. https://ddialliance.org/controlled-vocabularies.

ESFRI, Objectives & Vision, Accessed April 06, 2023. https://www.esfri.eu/objectives-vision.

Finnish Thesaurus and Ontology Service, Accessed April 06, 2023. https://finto.fi/okm-tieteenala/en.

Hodson, Jones et al. 2018. Turning FAIR data into reality. Interim report of the European Commission Expert Group on FAIR data. https://doi.org/10.5281/zenodo.1285272.

Katz, Elihu, and Lazarsfeld, Paul F. 1945. Decatur Study (Personal Influence). GESIS Data Archive, Cologne. ZA0241 Data file Version 1.0.0, https://doi.org/10.4232/1.0241.

NARCIS-classification, Accessed April 06, 2023. https://www.narcis.nl/classification/Language/en.

OKM - Finnish Thesaurus and Ontology Service, Accessed April 06, 2023. https://finto.fi/okm-tieteenala/en.

Pasquetto, Irene V., Randles, Bernadette M., Borgman, Christine L., 2017, "On the Reuse of Scientific Data". *Data Sci. J.* Vol. 16. DOI: 10.5334/dsj-2017-008.

Standard för svensk indelning av forskningsämnen 2011, Accessed April 06, 2023. https://www.scb.se/dokumentation/klassifikationer-och-standarder/standard-for-svensk-indelning-av-forskningsamnen.

Università degli Studi di Milano-Bicocca (2021), *Bicocca sede del coordinamento DASSI Data Archive Social Sciences Italy*, Accessed April 13, 2023. https://www.unimib.it/comunicazione/orientamento-comunicazione-eventi/comunicazione-istituzionale-e-redazione-web/focus-bicocca/bicocca-sede-del-coordinamento-dassi-data-archive-social-sciences-italy.

Wilkinson, M., and Dumontier, M., Aalbersberg, I. et al. 2016. "The FAIR Guiding Principles for scientific data management and stewardship". *Sci Data* 3, 160018. DOI: 10.1038/sdata.2016.18.

Appendice

1. Scaricamento dati dal CESSDA Data Catalogue

L'accesso al servizio, e la successiva gestione ed elaborazione dei dati, sono stati operati impiegando il software open source *R*, con l'ausilio di alcuni pacchetti appositamente installati, ovvero librerie di codice attraverso le quali accedere alle funzionalità necessarie, tra le quali: *dplyr, httr, jsonlite, tibble, tidyquery, tidyverse, xlsx*.

CDC rende disponibile lo scaricamento dei dati in formato *JavaScript Object Notation* (JSON). Si tratta di un formato di testo comprensibile sia da uomo sia da macchina, particolarmente adatto allo scambio dei dati e alla gestione attraverso linguaggi di programmazione. I dati sono strutturati in maniera gerarchica offrendo la possibilità di organizzare le informazioni su più livelli.

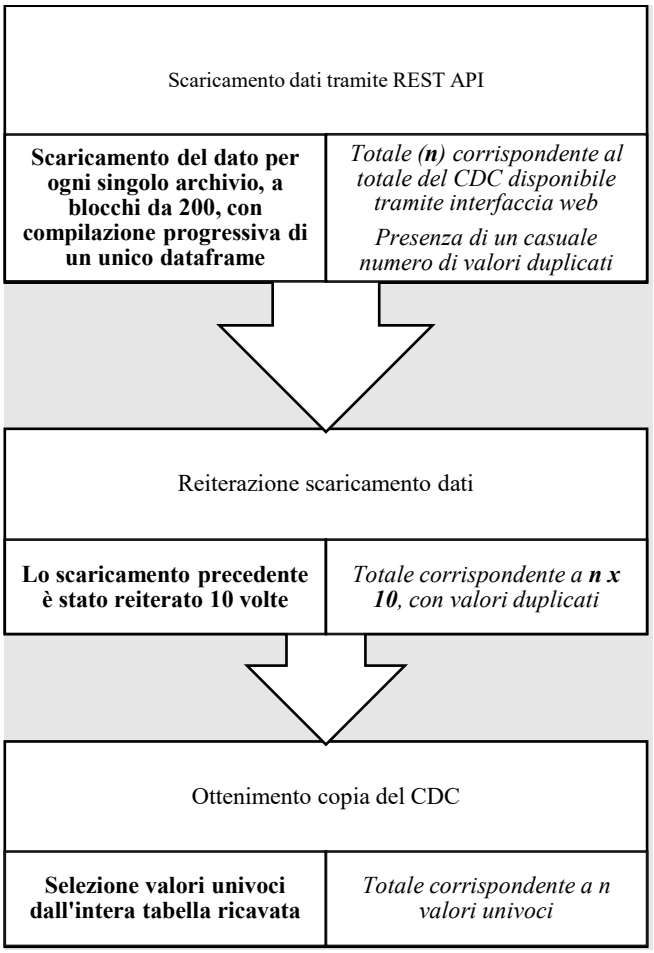
Lo scaricamento dei dati dal CDC è possibile a blocchi di massimo 200 dataset per volta (*limit = 200*), con la possibilità di stabilire il range temporale di riferimento, la lingua e altre caratteristiche dell'interrogazione. Attraverso la predisposizione di uno script per lo scaricamento basato su cicli *for* in R, è stata automatizzata la procedura, ottenendo la compilazione progressiva di un unico *dataframe*. Il CDC non offre la possibilità di eseguire una singola estrazione che ecceda i 10.000 record totali. Di conseguenza, sono stati estratti i dati di ogni singolo archivio, procedendo successivamente alla loro unione.

Le scelte tecniche operate derivano da una serie di test, svolti al fine di individuare la soluzione migliore per ottenere tutti i dati di un catalogo che, nonostante i servizi API predisposti, non è organizzato in previsione di un'estrazione di tale portata.

Da un primo ciclo di estrazione si è pervenuto a un risultato composto da un notevole numero casuale di valori duplicati, nonostante il totale dei dati estratti corrispondesse a quello disponibile dall'interfaccia web del CDC. Alcuni dati, di conseguenza, risultavano mancanti. Ciò è dovuto alla specifica architettura del servizio API, che permette lo scaricamento del dato impostando un record di partenza e la quantità da estrarre (attraverso i parametri *limit*= e *offset*= all'interno dell'interrogazione), senza assicurare che si tratti di dati univoci.

Per ovviare all'inconveniente e recuperare i dati mancanti, sostituiti dai duplicati, è stata adottata la soluzione di reiterare il ciclo 10 volte, ottenendo un'estrazione moltiplicata in tale misura rispetto ai dati disponibili. Al termine, sono stati selezionati i valori univoci basati sulla variabile ID del CDC (vedi par. 3.1), ottenendo una copia fedele del catalogo. Il processo eseguito è illustrato nello schema in **Figura A1**.

Figura A20. Schema del processo di estrazione dei dati eseguito



Al fine di accertare il corretto scaricamento dei dati, sono stati eseguiti confronti tra dati scaricati tramite API e ricerca attraverso l'interfaccia web del catalogo disponibile sul sito, con particolare attenzione alla corrispondenza del totale generale e rispetto a ogni singolo service provider.

Nel successivo riquadro, è riportato lo script creato ed eseguito per l'estrazione del catalogo.

```
##attivazione pacchetti utili
library(httr)
library(tibble)
library(dplyr)
library(tidyverse)
library(jsonlite)

all_variable<-c()
df = data.frame(matrix(nrow = 0, ncol = 0))
'colnames(df) = colonne'

start_value <- 0
end_value <- 9400

#lista contenente gli archivi pubblicati su CESSDA (publishers)
publishers <- list("Austrian%20Social%20Science%20Data%20Archive%20%28AUSSDA%29",
                   "DANS-KNAW",
                   "Danish%20National%20Archives%20%28DNA%29",
                   "FORS%20-%20Swiss%20Centre%20of%20Expertise%20in%20the%20Social%20Sciences",
                   "Finnish%20Social%20Science%20Data%20Archive%20%28FSD%29",
                   "GESIS%20-%20Leibniz%20Institute%20for%20the%20Social%20Sciences",
                   "PROGEDO%3A%20Center%20for%20Socio-Political%20Data%20%28CDSP%29",
                   "Portuguese%20Archive%20of%20Social%20Information%20%28APIS%29",
                   "Sikt%20-%20Norwegian%20Agency%20for%20Shared%20Services%20in%20Education%20and%20Research",
                   "NSD%20-%20Norwegian%20Centre%20for%20Research%20",
                   "Slovak%20Archive%20of%20Social%20Data%20%28SASD%29",
                   "Slovenian%20Social%20Science%20Data%20Archives%20%28ADP%29",
                   "SoDaNet%20-%20Greek%20Research%20Infrastructure%20for%20Social%20Science",
                   "Social%20Sciences%20and%20Digital%20Humanities%20Archive%20%28SODHA%29",
                   "Swedish%20National%20Data%20Service%20%28SND%29",
                   "UK%20Data%20Service",
                   "UniData%20-%20Bicocca%20Data%20Archive")

#ciclo for per l'estrazione dei dati CDC
for(i in 1:10) {
for(p in publishers) {
  for(i in seq(start_value, end_value, by = 200)) {
    string = paste("https://datacatalogue.cessda.eu/api/DataSets/v2/search?publishers=",p,"&limit=200&offset=",i,"&metadataLanguage=en")
    res1 = GET(gsub(" ", "", string))
    CDC1 = fromJSON(rawToChar(res1$content))
    if (as.matrix(CDC1$ResultsCount$retrieved) > 0) {
      CDC1_as_tibble <- as_tibble(CDC1$Results)
      df <- bind_rows(df, CDC1_as_tibble)
      'df[nrow(df) + 1,] <- CDC1_as_tibble
        variable <- as.matrix(CDC1$Results$id)
        all_variable <- append(all_variable, variable)'
    }
  }
}
}

#al termine, per trovare righe univoche basate su id usiamo:
df_distinct <- distinct(df, id, .keep_all = TRUE)

#rimuoviamo df per risparmiare spazio
remove(df)

#df_distinct diventa il nuovo df
df <- df_distinct
```

2. Principal investigator

L'estrazione è stata eseguita isolando i singoli *creators* inseriti, e confrontandoli con una serie di parole chiave attraverso linguaggio SQL. Nel linguaggio SQL, l'operatore LIKE permette di confrontare il contenuto di una stringa, anche in maniera parziale attraverso l'inserimento del segno di percentuale al fine di sostituire una parte dei caratteri. Adoperando le parole chiave scelte è stato possibile individuare il tipo di affiliazione.

Nel primo insieme, per isolare i nomi di organizzazioni o le affiliazioni relative a università, è stata imposta la condizione che ogni singolo dato contenesse '%univer%' (per includere tutte le varianti nelle varie lingue: università, universidad, university, ecc), '%fakulteta%', '%college%', '%school%'.

Nel secondo gruppo, sono stati esclusi i dati contenenti le stringhe precedenti e il confronto è stato eseguito adoperando 44 parole chiave, relative a concetti tipicamente riferiti a organizzazioni (ad esempio, '%institu%' , '%research%', '%forschun%', '%agency%', '%analys%', ecc.) o acronimi e nomi di enti esistenti (ad esempio: '%MORI%','%sozialwissenschaftliche studiengesellschaft%','%sonar%','%union%','%Konrad-Adenauer-Stiftung%','%usia%',
'%marplan%','%emnid%','%forsa%','%zuma%','%tarki%','%romanian academy%').

Il processo è stato eseguito sia complessivamente, sia per singolo service provider. Nel riquadro successivo è riportato la porzione di processo diestrazione in R con linguaggio SQL, relativo all'estrazione complessiva, in riferimento alla creazione dei tre gruppi (università, organizzazioni, non identificati).

```
cr_totale_university <-
  query("SELECT * FROM cr_totale_col_elenco_group_by
        WHERE LOWER(creators) LIKE '%univer%'
        OR LOWER(creators) LIKE '%fakulteta%'
        OR LOWER(creators) LIKE '%college%'
        OR LOWER(creators) LIKE '%school%'
        OR LOWER(creators) LIKE '%hochschule%'")

cr_totale_organization <-
  query("SELECT * FROM cr_totale_col_elenco_group_by
                WHERE LOWER(creators) NOT LIKE '%univer%'
                AND LOWER(creators) NOT LIKE '%fakulteta%'
                AND LOWER(creators) NOT LIKE '%college%'
                AND LOWER(creators) NOT LIKE '%school%'
                AND LOWER(creators) NOT LIKE '%hochschule%'
                AND (LOWER(creators) LIKE '%institu%'
                OR LOWER(creators) LIKE '%research%'
                OR LOWER(creators) LIKE '%forschun%'
                OR LOWER(creators) LIKE '%agency%'
                OR LOWER(creators) LIKE '%analys%'
                OR LOWER(creators) LIKE '%europe%'
                OR LOWER(creators) LIKE '%statisti%'
                OR LOWER(creators) LIKE '%commission%'
                OR LOWER(creators) LIKE '%education%'
                OR LOWER(creators) LIKE '%ation%'
```

```
            OR LOWER(creators) LIKE '%service%'
            OR LOWER(creators) LIKE '%department%'
            OR LOWER(creators) LIKE '%board%'
            OR LOWER(creators) LIKE '%association%'
            OR LOWER(creators) LIKE '%office%'
            OR LOWER(creators) LIKE '%network%'
            OR LOWER(creators) LIKE '%centr%'
            OR LOWER(creators) LIKE '%center%'
            OR LOWER(creators) LIKE '%bureau%'
            OR LOWER(creators) LIKE '%archiv%'
            OR LOWER(creators) LIKE '%istat%'
            OR LOWER(creators) LIKE '%building%'
            OR LOWER(creators) LIKE '%economic%'
            OR LOWER(creators) LIKE '%histor%'
            OR LOWER(creators) LIKE '%council%'
            OR LOWER(creators) LIKE '%trust%'
            OR LOWER(creators) LIKE '%ipsos%'
            OR LOWER(creators) LIKE '%gfk%'
            OR LOWER(creators) LIKE '%garr%'
            OR LOWER(creators) LIKE '%cern%'
            OR LOWER(creators) LIKE '%luiss%'
            OR LOWER(creators) LIKE '%eurostat%'
            OR LOWER(creators) LIKE '%oxfam%'
            OR LOWER(creators) LIKE '%tns gallup%'
            OR LOWER(creators) LIKE '%ministry%'
            OR LOWER(creators) LIKE '%library%'
            OR LOWER(creators) LIKE '%plattform%'
            OR LOWER(creators) LIKE '%laboratorio%'
            OR LOWER(creators) LIKE '%census%'
            OR LOWER(creators) LIKE '%kommission%'
            OR LOWER(creators) LIKE '%europ%'
            OR LOWER(creators) LIKE '%eurisko%'
            OR LOWER(creators) LIKE '%gesis%'
            OR LOWER(creators) LIKE '%wissen%'
            OR LOWER(creators) LIKE '%govern%'
            OR LOWER(creators) LIKE '%hospital%'
            OR LOWER(creators) LIKE '%social%'
            OR LOWER(creators) LIKE '%observa%'
            OR LOWER(creators) LIKE '%election%'
            OR LOWER(creators) LIKE '%politi%'
            OR creators LIKE '%MORI%'
            OR LOWER(creators) LIKE '%sozialwissenschaftliche studiengesellschaft%'
            OR LOWER(creators) LIKE '%sonar%'
            OR LOWER(creators) LIKE '%union%'
            OR LOWER(creators) LIKE '%Konrad-Adenauer-Stiftung%'
            OR LOWER(creators) LIKE '%klinik%'
            OR LOWER(creators) LIKE '%usia%'
            OR LOWER(creators) LIKE '%divo%'
            OR LOWER(creators) LIKE '%marplan%'
            OR LOWER(creators) LIKE '%emnid%'
            OR LOWER(creators) LIKE '%forsa%'
            OR LOWER(creators) LIKE '%zuma%'
            OR LOWER(creators) LIKE '%tarki%'
            OR LOWER(creators) LIKE '%romanian academy%'
            )
            ")

cr_totale_NOTorgNOTuni <-
  query("SELECT * FROM cr_totale_col_elenco_group_by
            WHERE LOWER(creators) NOT LIKE '%univer%'
```

```
AND LOWER(creators) NOT LIKE '%fakulteta%'
AND LOWER(creators) NOT LIKE '%college%'
AND LOWER(creators) NOT LIKE '%school%'
AND LOWER(creators) NOT LIKE '%hochschule%'
AND LOWER(creators) NOT LIKE '%institu%'
AND LOWER(creators) NOT LIKE '%research%'
AND LOWER(creators) NOT LIKE '%forschun%'
AND LOWER(creators) NOT LIKE '%agency%'
AND LOWER(creators) NOT LIKE '%analys%'
AND LOWER(creators) NOT LIKE '%europe%'
AND LOWER(creators) NOT LIKE '%statisti%'
AND LOWER(creators) NOT LIKE '%commission%'
AND LOWER(creators) NOT LIKE '%education%'
AND LOWER(creators) NOT LIKE '%ation%'
AND LOWER(creators) NOT LIKE '%service%'
AND LOWER(creators) NOT LIKE '%department%'
AND LOWER(creators) NOT LIKE '%board%'
AND LOWER(creators) NOT LIKE '%association%'
AND LOWER(creators) NOT LIKE '%office%'
AND LOWER(creators) NOT LIKE '%network%'
AND LOWER(creators) NOT LIKE '%centr%'
AND LOWER(creators) NOT LIKE '%center%'
AND LOWER(creators) NOT LIKE '%bureau%'
AND LOWER(creators) NOT LIKE '%archiv%'
AND LOWER(creators) NOT LIKE '%istat%'
AND LOWER(creators) NOT LIKE '%building%'
AND LOWER(creators) NOT LIKE '%economic%'
AND LOWER(creators) NOT LIKE '%histor%'
AND LOWER(creators) NOT LIKE '%council%'
AND LOWER(creators) NOT LIKE '%trust%'
AND LOWER(creators) NOT LIKE '%ipsos%'
AND LOWER(creators) NOT LIKE '%gfk%'
AND LOWER(creators) NOT LIKE '%garr%'
AND LOWER(creators) NOT LIKE '%cern%'
AND LOWER(creators) NOT LIKE '%luiss%'
AND LOWER(creators) NOT LIKE '%eurostat%'
AND LOWER(creators) NOT LIKE '%oxfam%'
AND LOWER(creators) NOT LIKE '%tns gallup%'
AND LOWER(creators) NOT LIKE '%ministry%'
AND LOWER(creators) NOT LIKE '%library%'
AND LOWER(creators) NOT LIKE '%plattform%'
AND LOWER(creators) NOT LIKE '%laboratorio%'
AND LOWER(creators) NOT LIKE '%census%'
AND LOWER(creators) NOT LIKE '%kommission%'
AND LOWER(creators) NOT LIKE '%europ%'
AND LOWER(creators) NOT LIKE '%eurisko%'
AND LOWER(creators) NOT LIKE '%gesis%'
AND LOWER(creators) NOT LIKE '%wissen%'
AND LOWER(creators) NOT LIKE '%govern%'
AND LOWER(creators) NOT LIKE '%hospital%'
AND LOWER(creators) NOT LIKE '%social%'
AND LOWER(creators) NOT LIKE '%observa%'
AND LOWER(creators) NOT LIKE '%election%'
AND LOWER(creators) NOT LIKE '%politi%'
AND creators NOT LIKE '%MORI%'
AND LOWER(creators) NOT LIKE '%sozialwissenschaftliche studiengesellschaft%'
AND LOWER(creators) NOT LIKE '%sonar%'
AND LOWER(creators) NOT LIKE '%union%'
AND LOWER(creators) NOT LIKE '%Konrad-Adenauer-Stiftung%'
AND LOWER(creators) NOT LIKE '%klinik%'
```

```
AND LOWER(creators) NOT LIKE '%usia%'
AND LOWER(creators) NOT LIKE '%divo%'
AND LOWER(creators) NOT LIKE '%marplan%'
AND LOWER(creators) NOT LIKE '%emnid%'
AND LOWER(creators) NOT LIKE '%forsa%'
AND LOWER(creators) NOT LIKE '%zuma%'
AND LOWER(creators) NOT LIKE '%tarki%'
AND LOWER(creators) NOT LIKE '%romanian academy%'
")
```