cessda eric

Consortium of European Social Science Data Archives
European Research Infrastructure Consortium

CESSDA Data Catalogue Overview

June 2018

John Shepherdson
Chair, CESSDA Technical Working Group

CESSDA Data Catalogue (CDC)

Will launch in June 2018

Does not hold any data itself, or provide direct access to data

Automatically, frequently and periodically harvests metadata from Service Providers and other sources

Fully automated harvesting is crucial to ensure that the CESSDA Data Catalogue content is up to date

Researchers are directed to the source/service where the data may be obtained or analysed

See also https://www.cessda.eu/Research-Infrastructure/Products-Services

cessda eric

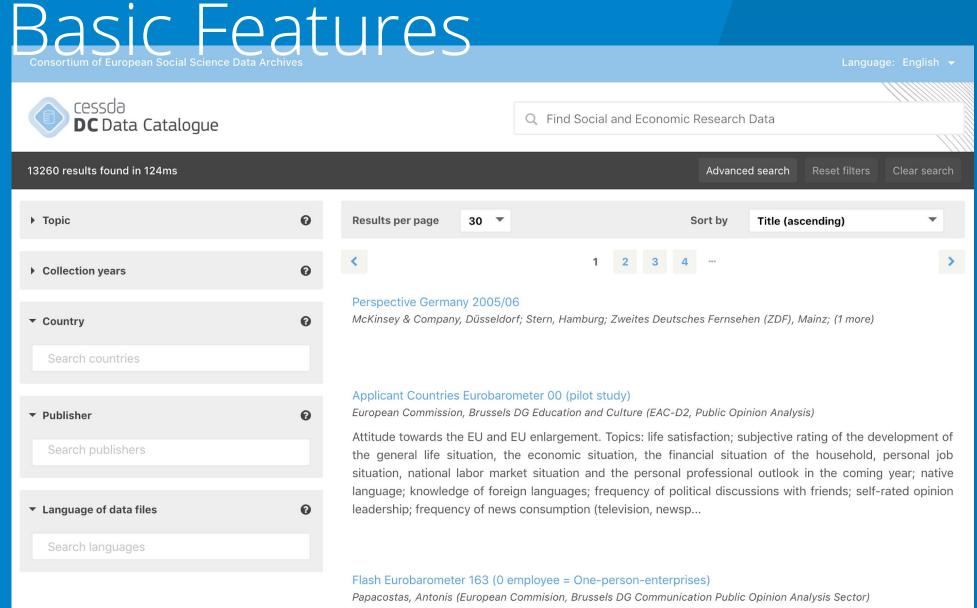# CESSDA Data Catalogue
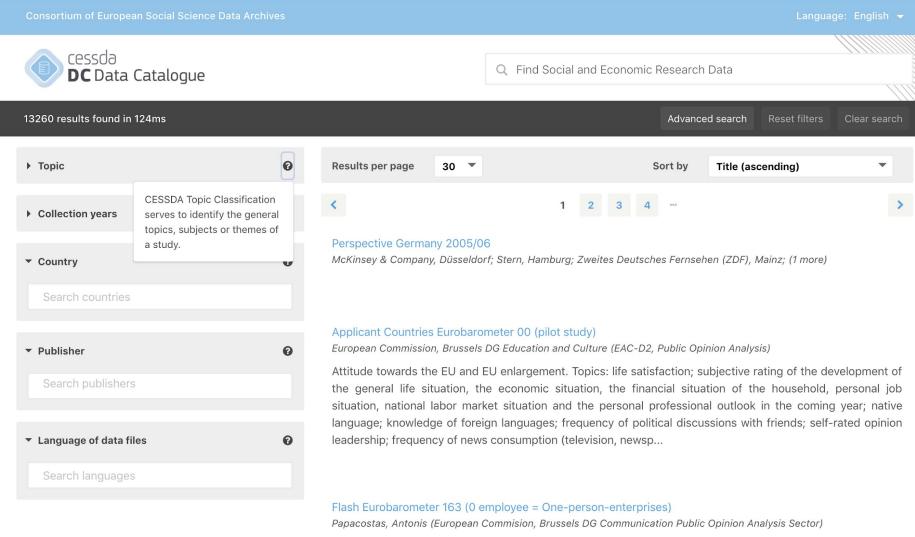
Circa 13,000 records to date:

    FSD

    GESIS

    Progedo
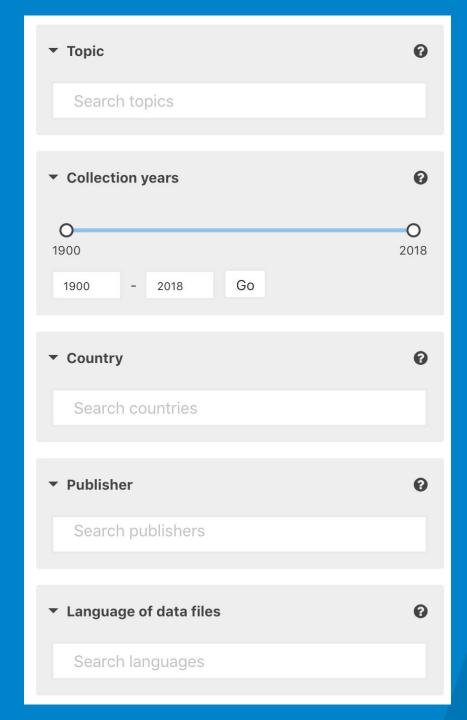
    UKDS

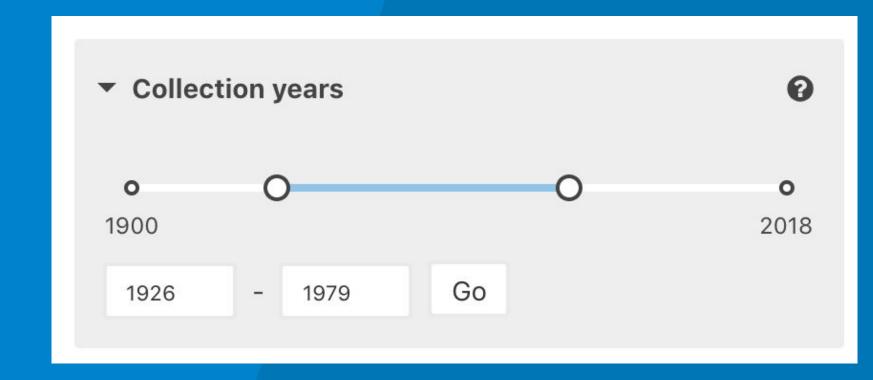Actively working with other Service Providers

Basic Features

Help - Tooltips

Topic ❓

Search topics

Collection years ❓

1900 — 2018

1900 – 2018 Go

Country ❓

Search countries

Publisher ❓

Search publishers

Language of data files ❓

Search languages

cessda eric

Collection years

1900   2018

| 1926 | – | 1979 | Go |

Topic

Search topics

Collection years

1900 ——————————————— 2018

| 1900 | – | 2018 | Go |

Country

Search countries

Publisher

Search publishers

Language of data files

Search languages

Filter summary   Advanced search   Reset filters   Clear search

cessda eric

Filter summary

The following filters have been applied to your search.

Topic: medical behavior

Collection years: 1942 - 2018

Country: England Finland

Language of data files: fi

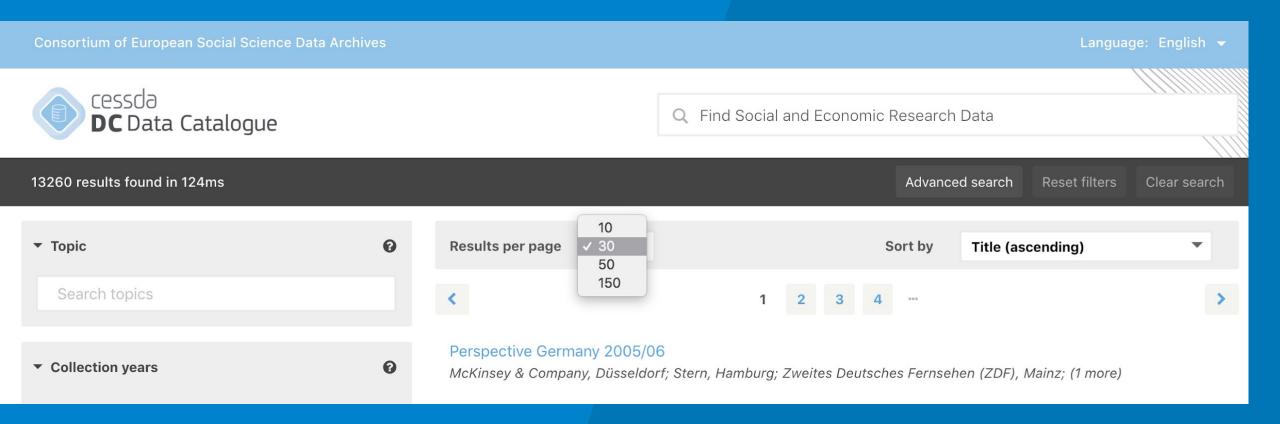Select a filter to remove it from this search.

Close

cessda eric

Uses Deep Links

https://datacatalogue-dev.cessda.eu/?publisher.publisher[0]=UK%20Data%20Service&publisher.publisher[1]=GESIS%20Data%20Archive%20for%20the%20Social%20Sciences&publisher.publisher[2]=Finnish%20Social%20Science%20Data%20Archive&publisher.publisher[3]=Centre%20de%20Donn%C3%A9es%20Socio-Politiques%20%28CDSP%29
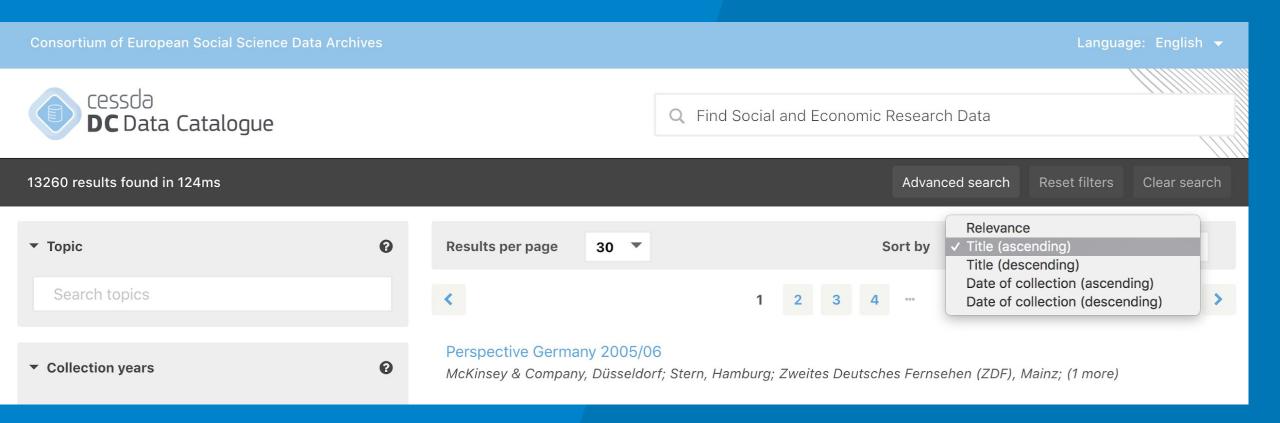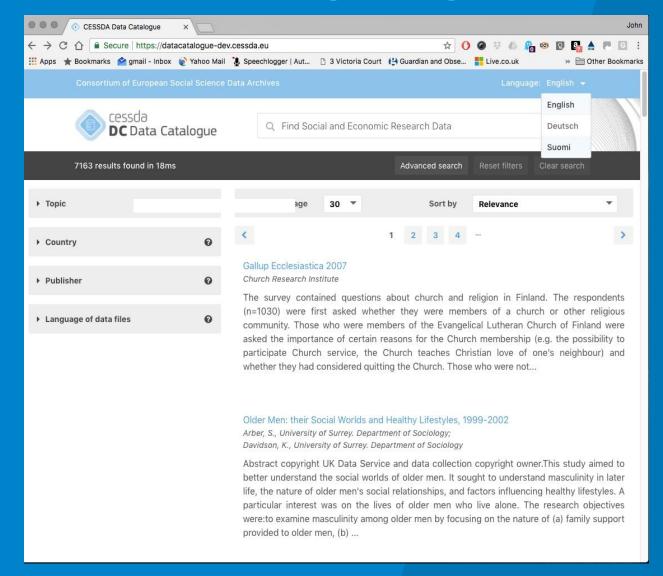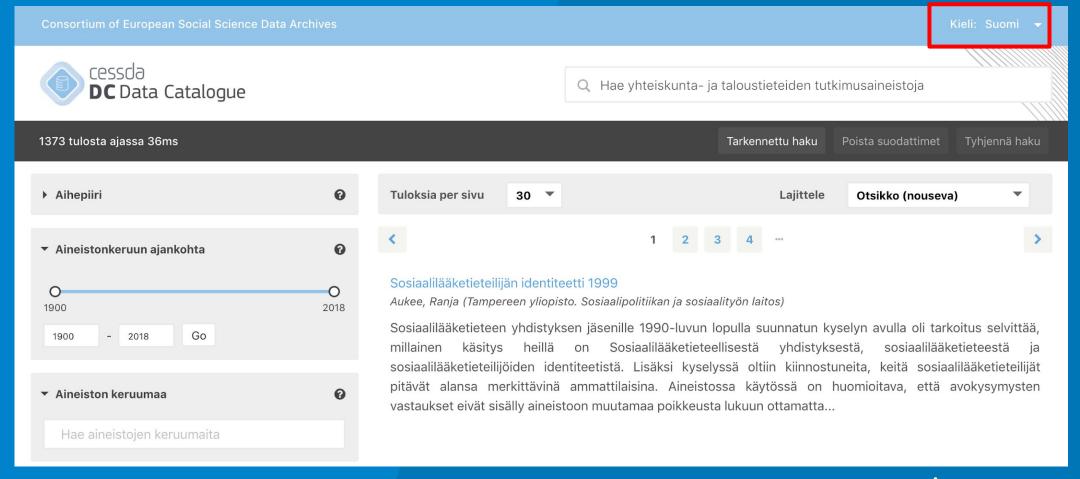
Results - Pagination

Results - Sort Order

Select UI Language

Select UI Language

cessda
DC Data Catalogue

🔍 "UK-Data-Service__4167"

Advanced search   Reset filters   Clear search

‹ Back   </> View JSON   ⬈ Go to study

Study title
Effects of Television News in British General Elections, 1997

Creator
Sanders, D., University of Essex. Department of Government
Norris, P., Harvard University. John F. Kennedy School of Government

Study Persistent Identifier
4167 (UKDA)
10.5255/UKDA-SN-4167-1 (datacite)

Abstract
Abstract copyright UK Data Service and data collection copyright owner.

The research used an experimental design to assess the extent to which UK voters' political perceptions are shaped by the content of television news coverage. The research involved:
1. administering a pre-test questionnaire to a randomly selected sample of 1125 respondents;
2. randomly assigning respondents to a series of groups;
3. exposing each group of respondents to a distinctive selection of video news items; and
4. administering a post-test questionnaire to each respondent.
The purpose of the research was to establish the extent to which changes between pre-test and post-test responses varied according to the type of video footage that respondents had seen.
Three core sets of hypotheses were tested:

cessda eric

Similar Results

View JSON representation

Metadata Issues

Many and various, including:

- Leading and trailing spaces or non-printing characters
- Capitalisation
- Missing/incorrect language tags
- Date and time formats
- Alternative accepted spellings (e.g. Surinam, Suriname*)

*State in South America, formally referred to as Republic of Suriname

cessda eric

Technical Details

- Java-based
- Spring Boot framework
- Elastic search
- UI - Javascript & Searchkit
- Docker containers running in Kubernetes clusters on Google Cloud Platform

cessda eric

Technical Details

- Uses CESSDA's Open Source Metadata Harvester
- Supports CESSDA Metadata Model v1
- Repository Handlers
  - adapter for OAI-PMH (DDI XML 2.5)
  - adapter for NESSTAR (DDI XML 1.2)

cessda eric

Moving over to using the OSMH implementation. Leveraging repository(ies) handlers for harvesting DDI-C/L data directly from Service providers
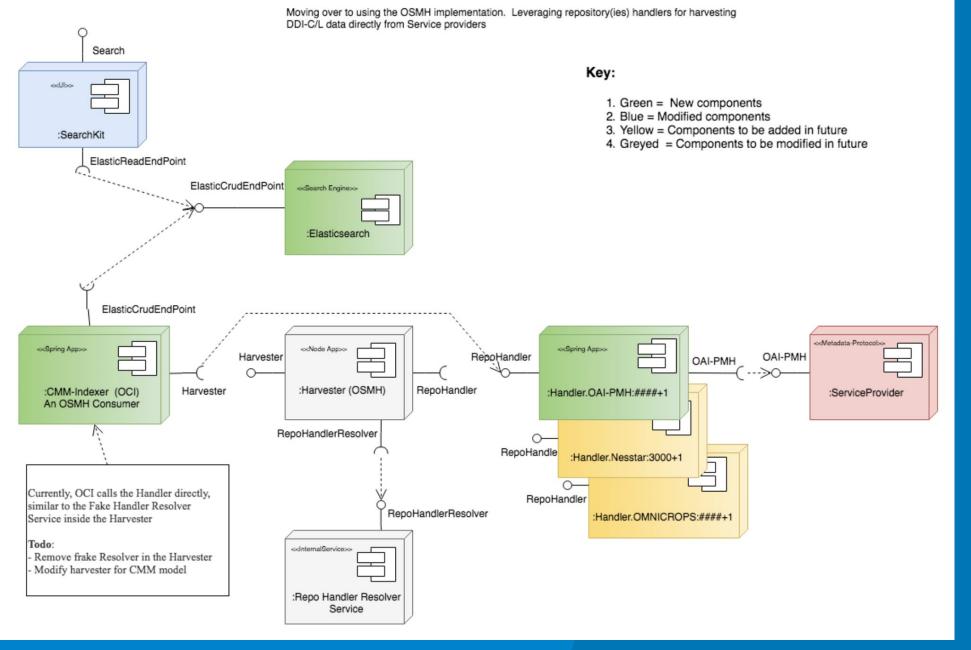
Search

<<UI>>

:SearchKit

ElasticReadEndPoint

ElasticCrudEndPoint

<<Search Engine>>

:Elasticsearch

Key:

1. Green = New components
2. Blue = Modified components
3. Yellow = Components to be added in future
4. Greyed = Components to be modified in future

ElasticCrudEndPoint

<<Spring App>>

:CMM-Indexer (OCI)
An OSMH Consumer

Harvester

Harvester

<<Node App>>

:Harvester (OSMH)

RepoHandler

RepoHandler

RepoHandlerResolver

RepoHandler

<<Spring App>>

:Handler.OAI-PMH:####+1

RepoHandle

:Handler.Nesstar:3000+1

RepoHandler

:Handler.OMNICROPS:####+1

OAI-PMH

OAI-PMH

<<Metadata-Protocol>>

:ServiceProvider

Currently, OCI calls the Handler directly, similar to the Fake Handler Resolver Service inside the Harvester

Todo:
- Remove frake Resolver in the Harvester
- Modify harvester for CMM model

RepoHandlerResolver

<<InternalService>>

:Repo Handler Resolver
Service

cessda eric

Version 1 (December 2017)

Fixed and upgraded Open Source Metadata Harvester

- Crawls endpoints for XML metadata and ingests transformed data into Elasticsearch

Leveraging a Microservice architecture with Spring Boot, Java 8 Stream for a declarative sets of business instructions:

- Harvester Microservice: Sole responsibility is to harvest and store XML metadata

Version 1(December 2017)

Indexer Microservice: An ETL tool responsible for extraction of harvested metadata, transformation with named-entity recognition (NER) enrichment and ingestion into Elasticsearch.

Spring Boot Admin App: To monitor and manage the above Microservice's logs, configuration, JMX exported beans, fail fast feature toggles, etc.

Search UI: Worked with a Frontend developer to customise SearchKit for FullText Search with advanced search query and result filtering.

cessda eric

Version 2 (June 2018)

Similar functionality to Version 1 - total rebuild to handle currently and future XML Schemas.

- Adopted a pluggable Microservice architecture style with multiple Repository Handlers, exposed as Restful APIs.
- Harvester responsible for discovering handlers and handing over requests to a given handler to process its XML schema specifics.
- Consumer Service responsible for orchestrating harvesting of any XML Metadata endpoint and ingestion of output into Elasticsearch.

cessda eric

Version 3 (Date TBC)

- More metadata cleaning, transformation, disambiguation
- Support for CESSDA Metadata Model v2
- Harvest from other sources
- Repository Handler adapter for OAI-PMH (DDI XML 3.2)
- Make metadata harvestable by other aggregators

- Identify relationships among all research entities

  publications, data, funding, researchers, organisations and data sources

Thanks for listening

Any Questions?

cessda eric