

subsetting data for montse

Devi Veytia

2025-11-18

Publications

File: mCDR_pubs, saved in mCDR_pubs.RData

Colnames:

- analysis_id : document id
- oro_type : type of mCDR (“CDR-BC” “CDR-BioPump” “CDR-OAE”)
- mean_prediction : relevance prediction from the LLM
- std_prediction : std deviation of relevance prediction from LLM

other columns contain bibliometric metadata from WOS or scopus

Legislation / policy

File: mCDR_leg, saved in mCDR_leg.RData

Note that there are more cdr oro_types than what was used in the paper – this is because some categories did not perform well with the LLM classifier and we could not get corresponding publication estimates.

Columns:

- search_id : the document ID as indexed in the source database
- Database: the name of the database where the document was retrieved from
- oro_type: the type of ORO
- matched_terms: terms matched in the document from the search query
- year: document year, parsed into standard format from Date
- Document.type: type of document as indexed by the database
- country_name: the country
- Binding.legislation: boolean indicator of whether the document is binding (ie legislation) or non-binding (ie policy)
- National.level: boolean indicator of whether the document is National level or not (e.g. agreement)
- Title: Document title
- Abstract: Document abstract
- Document.URL: URL of the document if available
- Date: Document publication date as indexed by the database (not standardised format)
- Keywords: Document keywords

Posts

file: mCDR_posts, saved in mCDR_posts.RData

columns:

- oro_type: type of ORO (note more types than in publication, see prev. section)
- year: post year
- variable_name: the variable long name and units
- y: the number of posts for the oro_type x year combination
- component: the name of the variable (e.g. public interest, public support)