

Browsing and Searching Metadata of TREC

Timo Breuer, Ellen M. Voorhees, Ian Soboroff

The 47th International ACM SIGIR Conference on
Research and Development in Information Retrieval

July 15, 2024, Washington D.C., USA

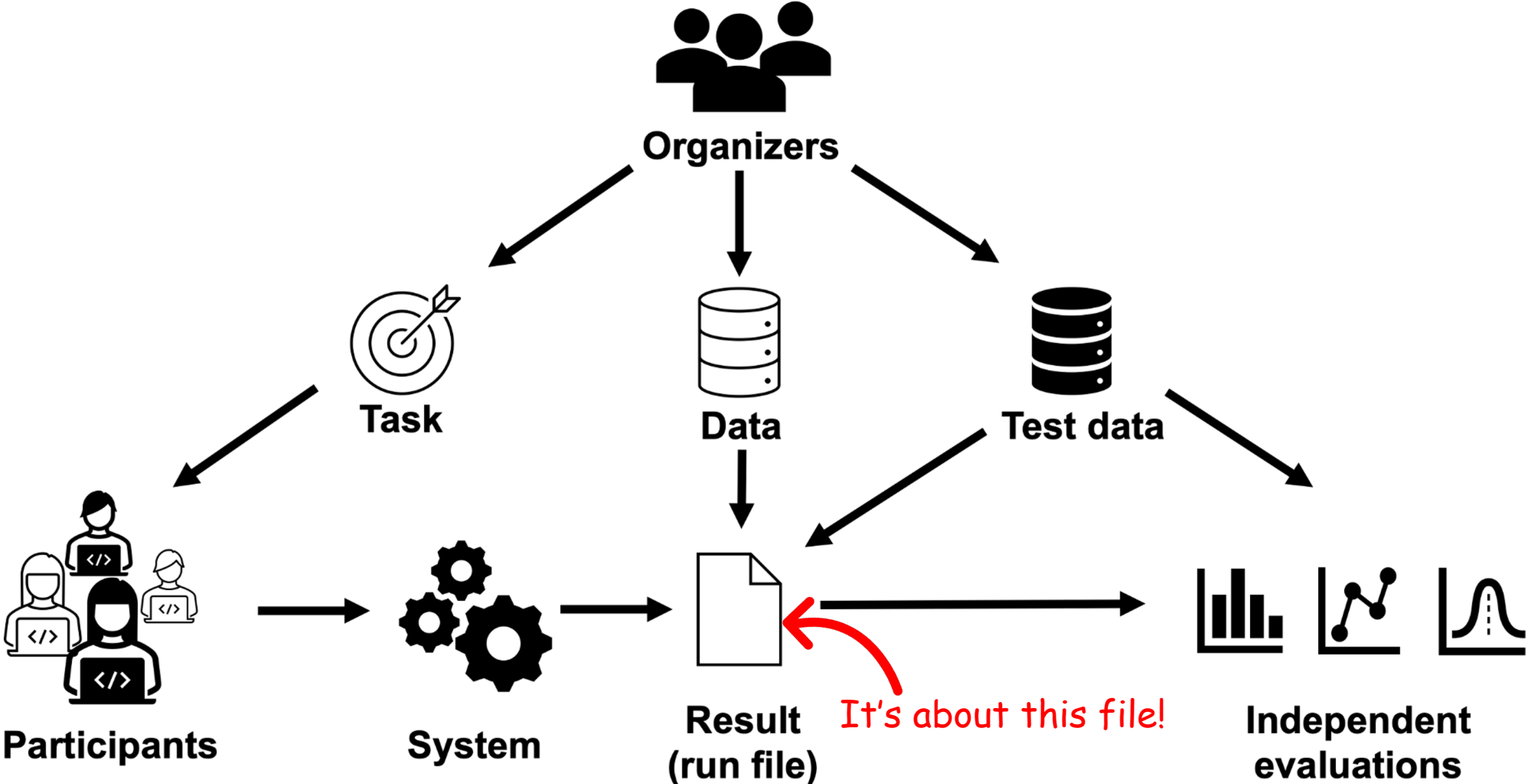


NATIONAL INSTITUTE OF
STANDARDS AND TECHNOLOGY
U.S. DEPARTMENT OF COMMERCE



SIGIR
2024
Washington, D.C.

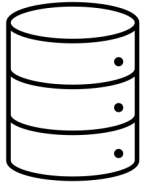
Common task framework or “How to TREC?”



Motivation

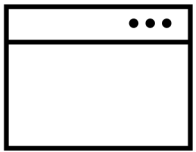


Our contributions



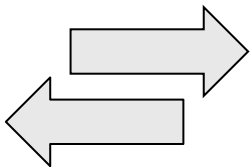
Database

Metadata covering over 30 years of TREC experiments for a sustainable reuse of submissions and better data provenance.



Metadata browser

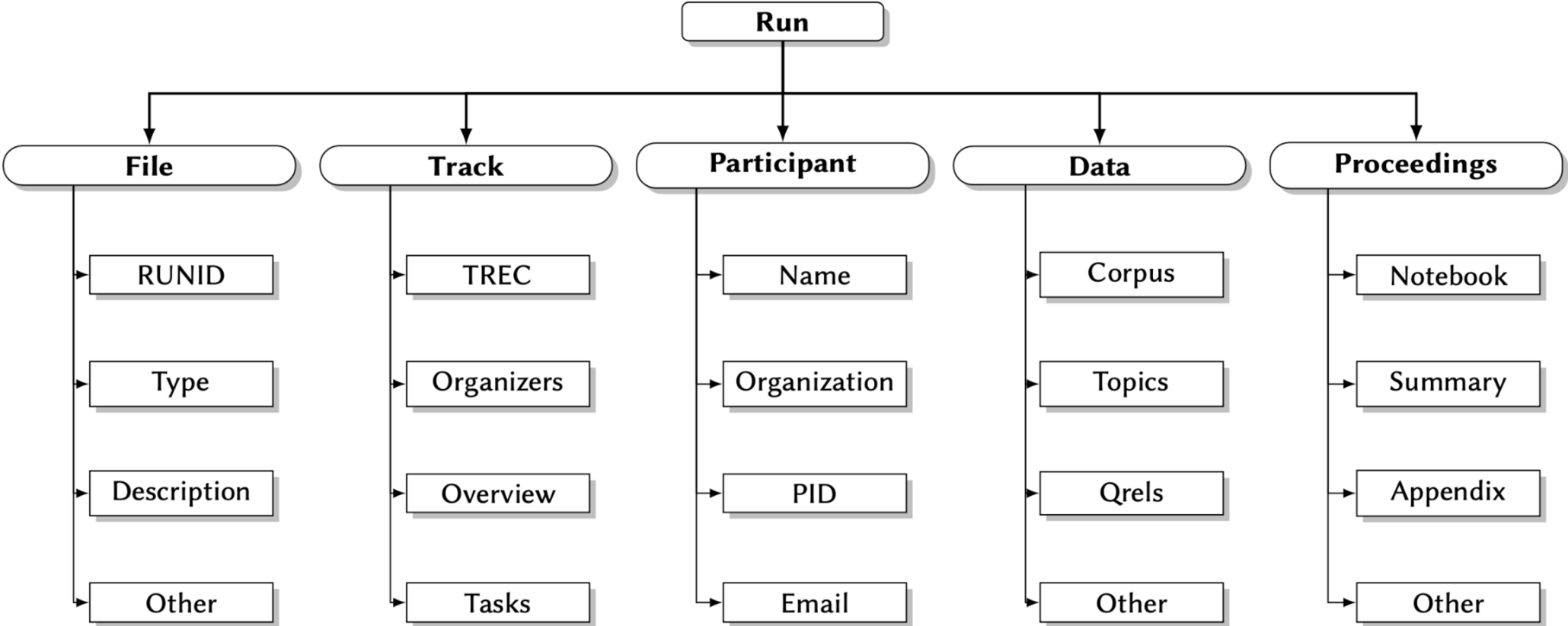
A new web-based browsing interface to TREC metadata, making TREC resources easier to find and browsable.



Web interface / REST API to the metadata

REST API for principled and systematic search access to the TREC metadata and resources.

Metadata scheme



Browsing TREC metadata

The screenshot shows the TREC Browser website interface. On the left is a navigation menu with links for TREC Browser, Home, and TREC-8 through TREC-COVID. The main content area features the title "Text REtrieval Conference (TREC)", an illustration of a woman and a child with a large pile of papers, and a link to "Proceedings | Data | trec.nist.gov". Below this is a timeline chart showing the evolution of TREC tasks from 1999 to 2023. The tasks are categorized into various domains such as Personal documents, Retrieval in a domain, Answers, not documents, Corporate repositories, Efficiency and web search, Beyond text, Language focus, Human-in-the-loop, Streaming text, and Static text. The right side of the page contains a "Table of contents" with links to various TREC tracks and tasks.

Year	Task
1999	Static text
2000	Static text
2001	Static text
2002	Static text
2003	Static text
2004	Static text
2005	Static text
2006	Static text
2007	Static text
2008	Static text
2009	Static text
2010	Static text
2011	Static text
2012	Static text
2013	Static text
2014	Static text
2015	Static text
2016	Static text
2017	Static text
2018	Static text
2019	Static text
2020	Static text
2021	Static text
2022	Static text
2023	Static text

<https://pages.nist.gov/trec-browser>



Searching TREC metadata

GET `trec/api/v1/trec26/core/WaterlooCormack/WCrobust04`

```
{
  'appendix_url': 'https://trec.nist.gov/pubs/trec26/appendices/core/WCrobust04.pdf',
  'data': {
    'corpus': 'https://catalog.ldc.upenn.edu/LDC2008T19',
    'qrels': 'https://trec.nist.gov/data/core/qrels.txt',
    'topics': 'https://trec.nist.gov/data/core/core_nist.txt'
  },
  'date': '6/8/2017',
  'description': 'Logistic Regression (Sofia-ML, Cornell TF-IDF features, from Total Recall BMI) trained on Robust 04 qrels.',
  'input_url': 'https://trec.nist.gov/results/trec26/core/input.WCrobust04.gz',
  'md5': '33768045401d9c6aa1205c3a9042cc63',
  'participant': {
    'name': 'Gordon V. Cormack',
    'organization': 'University of Waterloo',
    'pid': 'WaterlooCormack'
  },
  'publication': {
    'author': 'Maura R. Grossman, Gordon V. Cormack',
    'bibtex': '...',
    'biburl': 'https://dblp.org/rec/conf/trec/GrossmanC17a.bib',
    'key': 'DBLP:conf/trec/GrossmanC17a',
    'title': 'MRG_UWaterloo and WaterlooCormack Participation in the TREC 2017 Common Core Track',
    'url': 'https://trec.nist.gov/pubs/trec26/papers/WaterlooCormack-CC.pdf'
  },
  'runid': 'WCrobust04',
  'summary_url': 'https://trec.nist.gov/results/trec26/core/summary.trec_eval.WCrobust04',
  'task': 'main',
  'track': 'core',
  'trec': 'trec26',
  'type': 'automatic',
  'year': 2017
}
```

We embrace the community's engagement!

The screenshot shows the GitHub repository page for `usnistgov/trec-browser`. The repository is public and has 2 watchers, 0 forks, and 0 stars. The main content area displays the README for the "Metadata Browser of the Text Retrieval Conference". The README includes a note for reviewers and an overview table.

Directory / file	Description
browser/	This directory contains all the resources required to build and run the TREC



<https://github.com/usnistgov/trec-browser>

Thank you for your attention!

Browsing and Searching Metadata of TREC

Timo Breuer, Ellen M. Voorhees, Ian Soboroff

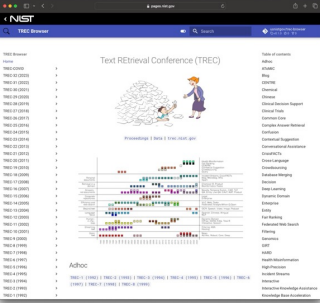
Contributions

- Database covering over 30 years of TREC experiments for a better data provenance
- Metadata browser making TREC resources browsable and more accessible
- Web interface / REST API for principled and systematic search access to TREC metadata and resources

Metadata scheme


```
graph TD
    Run --> File
    Run --> Track
    Run --> Participant
    Run --> Data
    Run --> Proceedings
    File --> RUNID
    File --> Type
    File --> Description
    File --> Other
    Track --> TREC
    Track --> Organizers
    Track --> Overview
    Track --> Tasks
    Participant --> Name
    Participant --> Organization
    Participant --> PID
    Participant --> Email
    Data --> Corpus
    Data --> Topics
    Data --> Qrels
    Data --> Other
    Proceedings --> Notebook
    Proceedings --> Summary
    Proceedings --> Appendix
    Proceedings --> Other
```


Browsing metadata on the web



Searching metadata with the API

```
curl -X GET 'https://trec.nist.gov/api/v1/trec26/ooze/WaterlooCormack/MCrobust04'
{"appendix_url": "https://trec.nist.gov/pubs/trec26-appendix/ooze/MCrobust04.pdf", "data": {"corpus": "https://catalog.lbc.upenn.edu/JC000013", "qrels": "https://trec.nist.gov/data/ooze/qrels.txt", "topics": "https://trec.nist.gov/data/ooze/ooze_topics.txt"}, "date": "6/8/2017", "description": "Logistic Regression (Sofia-M. Cornell TP-SDP Feature, from Total Recall 2011) trained on Robust-04 qrels.", "appendix_url": "https://trec.nist.gov/pubs/trec26-appendix/ooze/WaterlooCormack/MCrobust04.pdf", "name": "Gordon V. Cormack", "organization": "University of Waterloo", "pid": "WaterlooCormack", "publication": {"author": "Mauro S. Grossman, Gordon V. Cormack", "bibcite": "https://dblp.org/rec/conf/ooze/ooze2017a.bib", "key": "Ooze.conf/ooze2017a", "title": "MLL: Mutation and WaterlooCormack Participation in the TREC 2017 Common Core Task", "url": "https://trec.nist.gov/pubs/trec26-appendix/WaterlooCormack-CC.pdf"}, "nameid": "MCrobust04", "summary_url": "https://trec.nist.gov/pubs/trec26-appendix/ooze/summary_trec_url_MCrobust04", "task": "eval", "track": "core", "year": "trec26", "type": "Automatic", "year": 2017}}
```

 **TREC Browser**
pages.nist.gov/trec-browser/

 **GitHub**
github.com/usnistgov/trec-browser

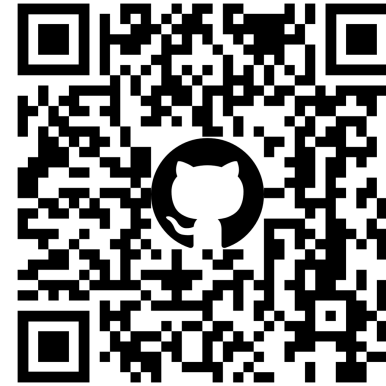
NIST NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY
U.S. DEPARTMENT OF COMMERCE

SIGIR 2024
Washington, D.C.

Poster & Demo Session

When:
Wednesday, July 17th, 2024
14:00-17:30

Where:
Congressional Ballroom & Senate



<https://github.com/usnistgov/trec-browser>



<https://pages.nist.gov/trec-browser/>