[Data Demo] ICAConfPubs: Dataset of and Website for Past ICA Conference Papers (2003-2018)

Hongtao Hao[1], Xinyue Chen[2], Yanling Zhao[3], Jing Zhang[4]

[1]Department of Computer Sciences, The University of Wisconsin-Madison, USA
[2]School of Intelligence Science Technology, Peking University, China
[3]School of Communication, Northwestern University, USA
[4]Department of Communication and Media, the University of Michigan, USA

Abstract

This paper presents a comprehensive dataset of the past ICA annual conferences papers from 2003 to 2018, encompassing 27,466 papers, 21,038 authors, and 4,935 sessions. The dataset is available for download in both CSV and JSON formats. Additionally, an API has been developed to facilitate programmatic access, and an intuitive user interface enables users to navigate and explore the data easily. The dataset and API can be accessed via a live website at https://icaconf.vercel.app. Reproducible codes to obtain and process the data are available at https://gitlab.com/4peerreview/icaconfpubs.

Contents

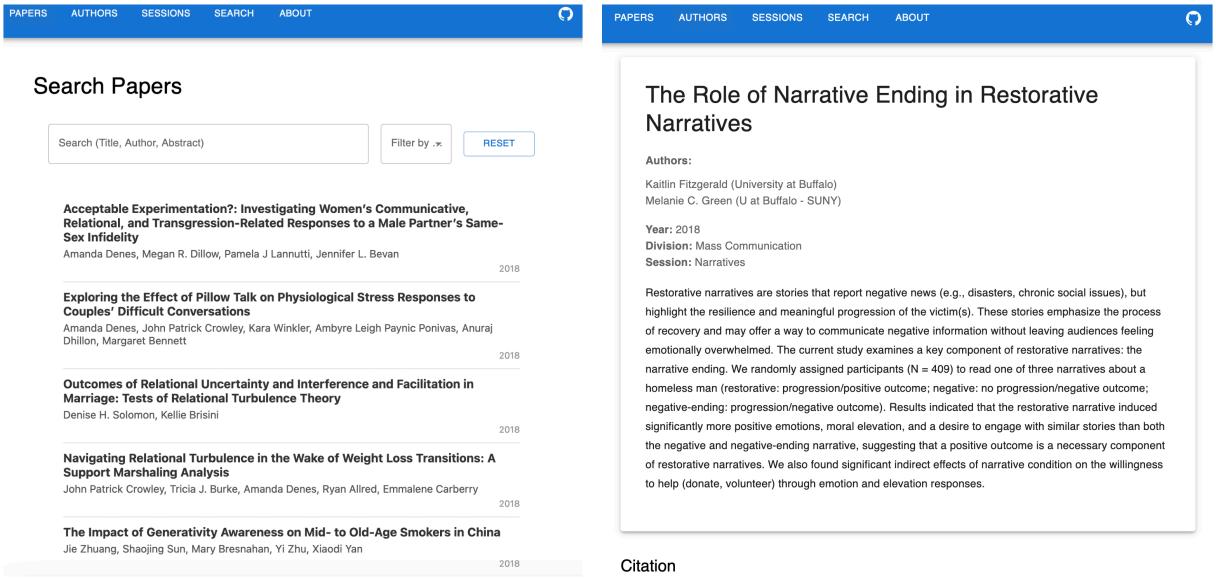


Figure 1: The interface of ICAConfPubs live website and paper details. The headers include papers, authors, and sessions. The authors page contains all authors and click on each author will show all papers authored by that person. The sessions page contains all sessions and clicking on each session will show all papers in that session. Clicking on each paper will navigate to a separate page for a paper where you can see information about authors, conference year, division, session, and abstract.

Introduction

The year 2026 marks the 76th Annual Conference of the International Communication Association (ICA). Each year, the annual conference welcomes submissions and scholars all around the globe. As the largest and most prominent association in the field of Communication, the ICA annual conference serves as a venue for worldwide scholars to present their most recent research progress and exchange ideas.

Unfortunately, all these data, i.e., submitted papers, participating scholars, and organized sessions, are not readily available to the public. Annual conferences from 2003 to 2018 have official websites that are powered by the conventions system by allacademic.com. Conference programs from 2019 and onward are only available in PDF formats to the public. Conferences prior to 2003 were not available. Even though online programs exist, they are complicatedly structured and not ready to organize and analyze for the public and scholars alike.

ICA annual conferences data are valuable and useful because they can:

- Inspire new research ideas. Right now, most communication literature comes from journal papers (searched mostly in Google Scholar). Findings from conferences may provide a new perspective and inspire new directions. More importantly, journal papers are delayed. If the ICA annual conference data is public and updated regularly, scholars will find it easier to get the most recent research ideas that are not found in journal publications.
- Circumvent publication biases. Publications might have biases (Sun & Pan, 2020). Not all research projects end up being published. For example, significant results or large effect sizes make studies more likely to be published (Sun & Pan, 2020). The publication bias suggests that we expand search strategies to include unpublished works such as conference papers. The papers of the ICA annual conference

papers serve as a good starting point. Although these presentations are not publications per se, ready to be cited, they are still peer-reviewed, ensuring the quality of the paper presented each year.
- Enable large scientometric analysis. The ICA annual conference data over the past two to three decades are large. It contains over 30,000 papers, 20,000 authors and 5,000 sessions. This dataset is useful for large scale scientometric analysis. For example, the conference papers dataset can be applied to study the topic evolution of communication studies in the past decades or to study academic collaboration or mobility within the field of Communication.
- Contribute to open science. If the data is publicly available, scholars from all other fields can use this dataset.
- Provide deeper insights. With these data, we can understand the diversity of communication scholars & research topics better. Right now, we only have access to journal data, but that is only part of communication scholars and communication research. To get a broader picture and a deeper understanding, we need data about the conference papers as well.

Motivated by these strengths, benefits, and potential contributions to the whole communication field and broader academia, we obtained and processed all available data on ICA annual conferences. An API for the data and a user interface to help users navigate the data were also developed.

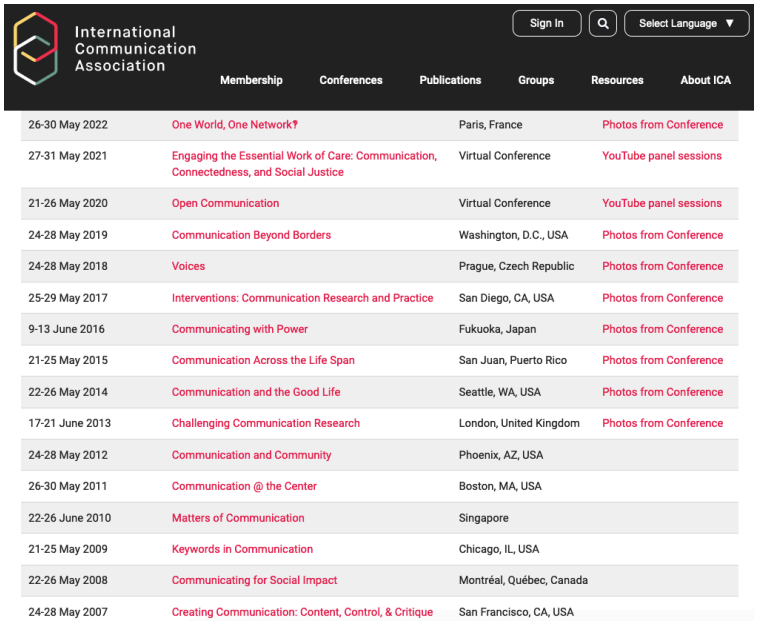
Related Work

Scientometric analysis is widely used by scholars in communication (Chakravartty et al., 2018; Freelon et al., 2023), as well as by researchers in other fields, to gain a deeper understanding of academic landscapes. To support such analyses, several datasets (Isenberg et al., 2017) and interfaces (Lange, 2024) have been developed.

This project draws inspiration from these examples, aiming to aggregate paper data and design an accessible interface specifically for the field of communication.

Data Collection and Processing

The official website of ICA provides information about all past annual conferences at https://www.icahdq.org/page/annual-conference.



26-30 May 2022	One World, One Network	Paris, France	Photos from Conference
27-31 May 2021	Engaging the Essential Work of Care: Communication, Connectedness, and Social Justice	Virtual Conference	YouTube panel sessions
21-26 May 2020	Open Communication	Virtual Conference	YouTube panel sessions
24-28 May 2019	Communication Beyond Borders	Washington, D.C., USA	Photos from Conference
24-28 May 2018	Voices	Prague, Czech Republic	Photos from Conference
25-29 May 2017	Interventions: Communication Research and Practice	San Diego, CA, USA	Photos from Conference
9-13 June 2016	Communicating with Power	Fukuoka, Japan	Photos from Conference
21-25 May 2015	Communication Across the Life Span	San Juan, Puerto Rico	Photos from Conference
22-26 May 2014	Communication and the Good Life	Seattle, WA, USA	Photos from Conference
17-21 June 2013	Challenging Communication Research	London, United Kingdom	Photos from Conference
24-28 May 2012	Communication and Community	Phoenix, AZ, USA	
26-30 May 2011	Communication @ the Center	Boston, MA, USA	
22-26 June 2010	Matters of Communication	Singapore	
21-25 May 2009	Keywords in Communication	Chicago, IL, USA	
22-26 May 2008	Communicating for Social Impact	Montréal, Québec, Canada	
24-28 May 2007	Creating Communication: Content, Control, & Critique	San Francisco, CA, USA	

Figure 2: Past ICA annual conferences.

Online programs powered by `allacademic.com` exist for the years between 2003 and 2018. Conferences later on only provide programs in PDF format. These PDFs are very large and notoriously hard to parse. Therefore, the researcher decided to leave them alone and focus on online programs instead.

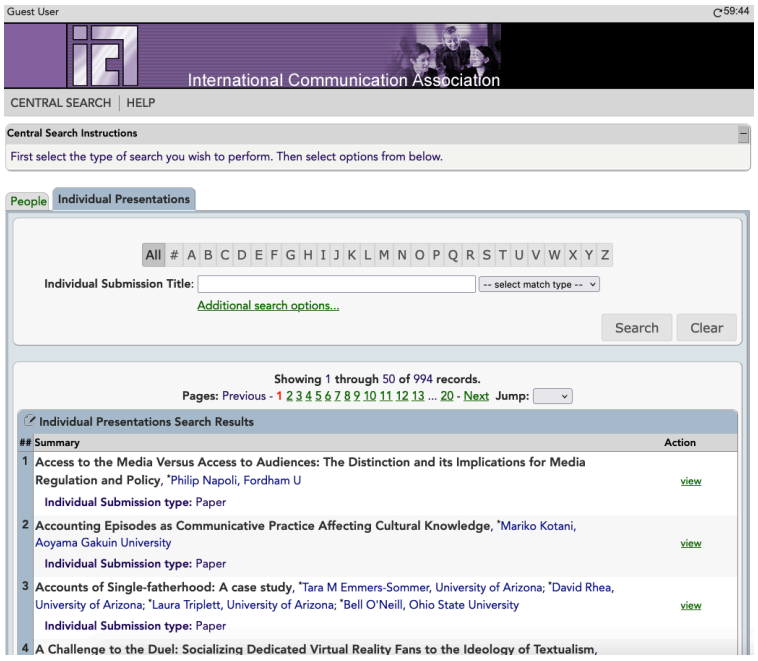


Figure 3: Online program example of ICA annual conference in 2003.

Figure 3 is an example of these online programs. These programs experienced three different periods:

- Years 2003-2004. In these two years, the online programs contain data about authors and papers. However, session and division information is not provided.
- Years 2005-2013. In these years, information about authors, sessions, and papers is provided.

- Years 2014-2018. The online programs changed their structures dramatically, and incorporated Interactive Papers.

The researcher dynamically collected the data using the Python package of `selenium`. After data wrangling, three datasets are the main outputs.

The paper data consists of the following columns:

- **Paper ID**: An identifier assigned to each paper, formatted as `year-index`.
- **Title**: The title of the conference paper.
- **Paper Type**: Indicates the type of presentation, either `Paper` or `Poster`. Note that the ICA website did not differentiate between these two types prior to 2014, so all presentations before 2014 are classified as `Paper`, although some may have originally been `Poster`.
- **Abstract**: The abstract of the paper.
- **Number of Authors**: The number of authors for this paper.
- **Year**: The year the paper was presented.
- **Session**: The title of the session in which the paper was presented.
- **Division/Unit**: The division or unit that organized the session.
- **Authors**: The authors of the paper.

The author data includes the following columns:

- **Paper ID**: An identifier assigned to each paper, formatted as `year-index`.
- **Paper Title**: The title of the conference paper.
- **Year**: The year the paper was presented.
- **Number of Authors**: The number of authors for this paper.
- **Author Position**: The position of this author (e.g., first author, co-author).
- **Author Name**: The name of the author.
- **Author Affiliation**: The affiliation of the author.

The session data contains the following columns:

- **Year**: The year the session occurred.
- **Session Type**: Specifies whether the session is a `paper session` or an `interactive paper session` (i.e., poster session).
- **Session Title**: The title of the session.
- **Division/Unit**: The division or unit organizing the session.
- **Chair Name**: The name of the session chair.
- **Chair Affiliation**: The affiliation of the session chair.

These data are available to be downloaded in `CSV` format. There are 27,466 papers authored by 21,038 scholars. These papers are presented in 4,935 sessions.

Topic Modeling

[Jiye Sun will be responsible for writing this part.]

Blablabla

API Development

Aggregation and Data Structure

The strength of data in `CSV` is that they are lightweight and easy to parse and analyze. However, the drawbacks are obvious. First, it is not friendly for development. In web development, pagination is preferred when the data is very large. However, it is not easily ready with `CSV`. Second, `CSV` data makes it hard to use nested structures. For example, for each paper, it is hard to combine paper data with author data. Suppose I want to add all the author names and author affiliations, it is hard to implement in `CSV`.

Therefore, the researcher decided to aggregate all three datasets, i.e., `papers.csv`, `authors.csv`, and `sessions.csv` into one single data in `JSON` format, which is optimal for nested data structure.

This aggregated data is `papers.json`. Its structure is as follows:

```
class Authorship(BaseModel):
    position: Optional[int] = None
    author_name: Optional[str] = None
    author_affiliation: Optional[str] = None

class SessionInfo(BaseModel):
    session: str
    session_type: Optional[str] = None
    chair_name: Optional[str] = None
    chair_affiliation: Optional[str] = None
    division: Optional[str] = None
    years: List[int] = []
    paper_count: Optional[int] = None
    session_id: Optional[str] = None

class Paper(BaseModel):
    paper_id: str
    title: str
    paper_type: str
    abstract: Optional[str] = None
    number_of_authors: int
    year: int
    session: Optional[str] = None
    division: Optional[str] = None
    authorships: Optional[List[Authorship]] = None
    author_names: Optional[List[str]] = None
    session_info: Optional[SessionInfo] = None
```

As can be seen, it combines both the author and session data nicely.

To facilitate analysis, the researcher also aggregated the author data and session data:

```
class Author(BaseModel):
    author_name: str
    attend_count: int
    paper_count: int
    paper_ids: Optional[List[str]] = None
    affiliations: Optional[List[str]] = None
    affiliation_history: Optional[str] = None
```

```
    years_attended: Optional[List[int]] = None

class Session(BaseModel):
    session: str
    session_type: Optional[str] = None
    chair_name: Optional[str] = None
    chair_affiliation: Optional[str] = None
    division: Optional[str] = None
    years: Optional[List[int]] = []
    paper_count: Optional[int] = None
    session_id: str
```

The data is served on MongoDB.

API

To make the data easier to use for web development, we developed an API using `Next.js`. Below, we explain the most important endpoints.

Retrieve Papers

Endpoint: `GET /papers`

Description: Retrieves a list of papers with optional filters for searching by various fields.

Parameters:

- `page` (int): Page number for pagination (default: 1).
- `limit` (int): Number of items per page (default: 100).
- `paper_id` (str): Unique ID assigned to the paper.
- `title_contains` (str): Keyword to search within the paper title.
- `paper_type` (str): Type of presentation, either `Paper` or `Poster`.
- `abstract_contains` (str): Keyword to search within the paper abstract.
- `number_of_authors` (int): Number of authors.
- `session_contains` (str): Keyword to search within the session title.
- `year` (int): The year the paper was presented.
- `session` (str): The session title.
- `division` (str): Division or Unit that organized the session.
- `has_author` (str): Author name appearing in the paper.
- `first_author` (str): First author of the paper.
- `last_author` (str): Last author of the paper.
- `session_id` (str): Exact match for the session ID.

Example Request:

```
GET /papers?title_contains=communication&year=2003
```

Example Response:

```
[
    {'paper_id': '2003-0155',
     'title': 'Computer-Mediated Communication,...',
     'paper_type': 'Paper',
     'abstract': 'The present study ...',
     'number_of_authors': 2,
```

```
    'year': 2003,
    'session': None,
    'division': None,
    'authorships': [{'position': 0,
                     'author_name': 'Jo Anna Madrid',
                     'author_affiliation': 'Rio Hondo College'},
                    {'position': 1,
                     'author_name': 'Richard L. Wiseman',
                     'author_affiliation': 'California ...'}],
    'author_names': ['Jo Anna Madrid', ...],
    'session_info': None},
   ...
]
```

Retrieve a Paper by ID

Endpoint: `GET /papers/{paper_id}`

Description: Retrieves detailed information for a specific paper by its unique `paper_id`.

Parameters:

- `paper_id` (str): The unique ID of the paper.

Example Request:

```
GET /papers/2003-001
```

Example Response:

```
[{'paper_id': '2003-0001',
  'title': 'Access to ...',
  'paper_type': 'Paper',
  'abstract': 'When the issue ...',
  'number_of_authors': 1,
  'year': 2003,
  'session': None,
  'division': None,
  'authorships': [{'position': 0,
                   'author_name': 'Philip Napoli',
                   'author_affiliation': 'Fordham U'}],
  'author_names': ['Philip Napoli'],
  'session_info': None}]
```

Retrieve Authors

Endpoint: `GET /authors`

Description: Retrieves a list of authors with optional filters to search by various fields.

Parameters:

- `page` (int): Page number for pagination (default: 1).
- `limit` (int): Number of items per page (default: 50).
- `author_name` (str): Exact name of the author.
- `min_attend_count` (int): Minimum number of times the author attended.
- `min_paper_count` (int): Minimum number of papers the author has.
- `affiliation_contains` (str): Keyword to search within author affiliations.

- `year_attended` (int): Specific year the author attended.

Example Request:

```
GET /authors?author_name=Eiri Elvestad
```

Example Response:

```
[{'author_name': 'Eiri Elvestad',
  'attend_count': 2,
  'paper_count': 2,
  'paper_ids': ['2013-1099', '2018-0124'],
  'affiliations': ['Vestfold U College', 'U ...'],
  'affiliation_history': 'Vestfold U College -> U ...',
  'years_attended': [2013, 2018]}]
```

Retrieve Sessions

Endpoint: `GET /sessions`

Description: Retrieves a list of sessions with optional filters for specific session attributes.

Parameters:

- `page` (int): Page number for pagination (default: 1).
- `limit` (int): Number of items per page (default: 50).
- `session` (str): Exact name of the session.
- `session_type` (str): Type of the session, e.g., `Paper Session`.
- `chair_name` (str): Name of the session chair.
- `chair_affiliation` (str): Affiliation of the session chair.
- `division` (str): Division or unit organizing the session.
- `year` (int): Specific year the session was held.
- `paper_count` (int): Number of papers in the session.

Example Request:

```
GET /sessions?paper_count=10
```

Example Response:

```
[
    {'session': 'Best Student Papers in Public Relations',
     'session_type': 'Paper Session',
     'chair_name': 'Chiara Valentini',
     'chair_affiliation': 'Aarhus U',
     'division': 'Public Relations',
     'years': [2014, 2015],
     'paper_count': 10,
     'session_id': '005831a276e8'
    },
    ...
]
```

This can be easily done in Python. For example:

```
import requests
```

```
base_url = ""

# Parameters for the request
params = {
    "title_contains": "communication",
    "year": 2003
}

# Make the request
response = requests.get(f"{base_url}/papers", params=params)

# Check the response
if response.status_code == 200:
    papers = response.json()
    print("Papers retrieved:", papers)
else:
    print("Failed to retrieve papers:", response.status_code, response.text)
```

User Interface

Clearly, the data is very large with around 30,000 papers and 20,000 authors. To make the data more easily available to the public and interested scholars, the researcher developed a web application using `React`.

Figure 1 shows the basic interface of this web app. The left panel shows all the 27,466 papers and the right panel shows the details of each paper.

The interface also includes authors as in Figure 5 and sessions as in Figure 6. The Authors page shows all the 21,038 authors. Clicking on the author will show all the papers by that author. Similarly, the Sessions page shows all the 4,935 sessions and clicking on each session will go to the page that presents all the papers in that session.

In the paper panel, filtering through the year is enabled.

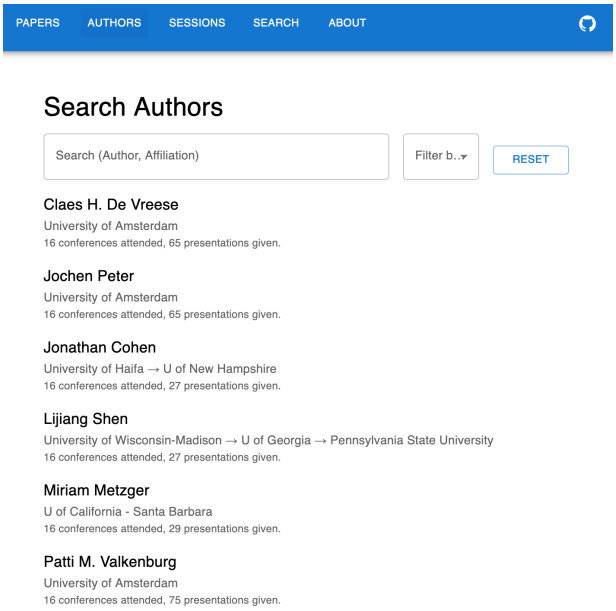


Figure 5: Author Page

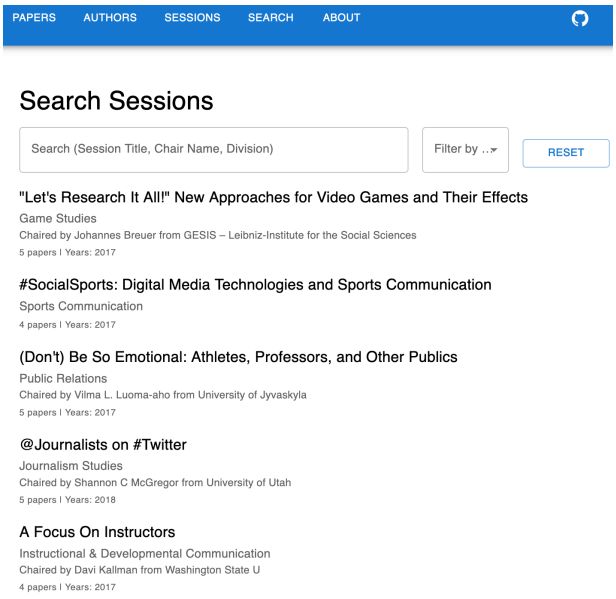


Figure 6: Session Page

Figure 4: Screenshots of the Author and Session Pages in the user interface.

Discussion and Contribution

This paper innovatively demonstrates a large and comprehensive dataset of the past ICA annual conferences papers from 2003 to 2018, encompassing 27,466 papers, 21,038 authors, and 4,935 sessions. The contributions of this paper are as follows:

- Contributed a single dataset that aggregated all past ICA conference papers, authors, and sessions, which, to our knowledge, is the first that work in this direction.
- Developed an API to allow other developers and scholars to utilize the data more easily.
- Designed an interface to allow easier navigation through, and better presentation of, the dataset.

The dataset is open to the public via a live website interface and API. Scholars in the communication field or other fields can easily use this dataset to get deeper insights into how communication research developed over the past years, the research topics/focus/subfields evolution trend, get inspiration from past conference papers, or conduct more comprehensive large-scale scientometric analysis.

Limitations

There are many improvements to be done. For example:

- Data from years after 2019 and before 2003 are not available yet.
- More filtering logic should be allowed in the interface.
- Deduplication of the author names, distinguishing between two different authors who share the same name, is almost impossible. Usually, we can rely on disciplines to deduplicate authors. However, in this dataset, all authors belong to Communication. Also, many contributors are students who move more frequently between institutions, which makes the deduplication task even harder.
- We did not deduplicate affiliations.
- In the interface, searching only allows exact matches. Users might want to search any relevant papers even though the exact search term does not appear in the abstract or title.

Accessibility

The dataset and tools can be accessed via:

- Live website at https://icaconf.vercel.app/.
- Reproducible codes to obtain and process the data are available at https://gitlab.com/4peerreview/icaconfpubs.

Bibliography

Chakravartty, P., Kuo, R., Grubbs, V., & McIlwain, C. (2018). # communicationsowhite. *Journal of Communication*, *68*(2), 254–266.

Freelon, D., Pruden, M. L., & Malmer, D. (2023). # politicalcommunicationsowhite: Race and politics in nine communication journals, 1991-2021. *Political Communication*, *40*(4), 377–395.

Isenberg, P., Heimerl, F., Koch, S., Isenberg, T., Xu, P., Stolper, C., Sedlmair, M., Chen, J., Möller, T., & Stasko, J. (2017). vispubdata.org: A Metadata Collection about IEEE Visualization (VIS) Publications. *IEEE Transactions on Visualization and Computer Graphics*, *23*. https://doi.org/https://doi.org/10.1109/TVCG.2016.2615308

Lange, D. (2024,). *Vispubs.com: A Visualization Publications Repository*. https://doi.org/10.31219/osf.io/dg3p2

Sun, Y., & Pan, Z. (2020). Not published is not perished: Addressing publication bias in meta-analytic studies in communication. *Human Communication Research*, *46*(2–3), 300–321.