ICAConfPubs: Dataset of and Website for Past ICA Conference Papers

Anynomous

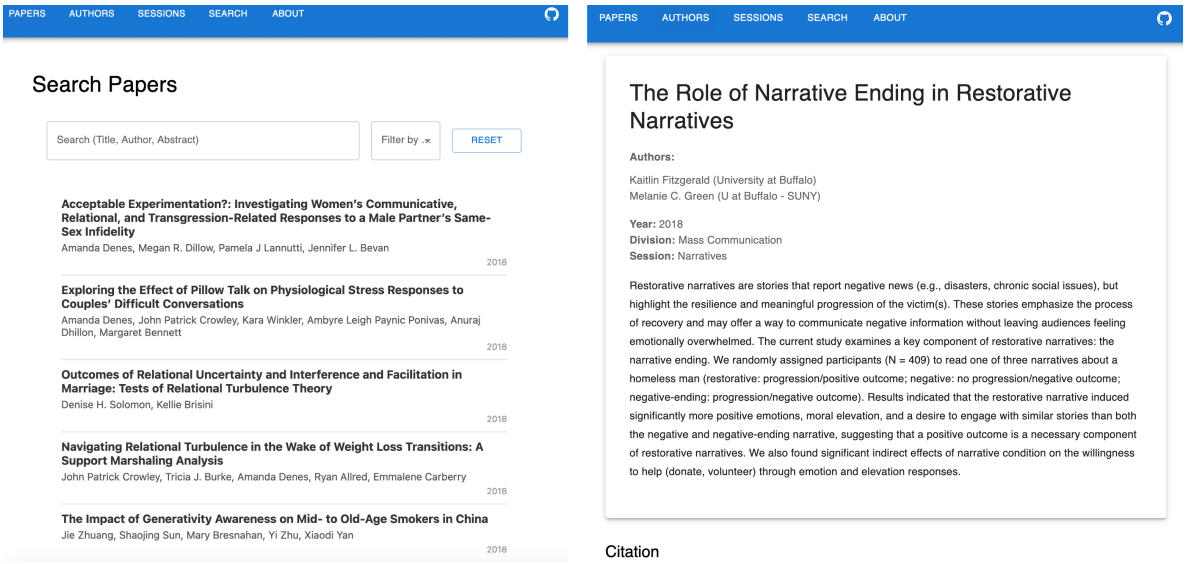


Figure 1: The interface of ICAConfPubs and paper details. The headers include papers, authors, and sessions. The authors page contains all authors and click on each author will show all papers authored by that person. The sessions page contains all sessions and clicking on each session shows all papers in that session. Clicking on each paper will navigate to a separate page for a paper.

Abstract

This paper presents a comprehensive dataset of the ICA annual conferences 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 tools can be accessed via a live website at https://icaconf.vercel.app, and the API is available at https://icaconf.onrender.com.

Index Terms: Communication, Publication Data, Dataset, Authors

1 Introduction

The year 2025 marks the 75th year of the Annual Conference of the International Communication Association (ICA). Each year, the annual conference welcomes submissions and scholars around the globe. As the largest and most prominent association in the field of Communication, the ICA annual conference serves as a venue for 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`. Those 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. Most importantly, journal papers are delayed. If we are able to make the ICA annual conference data public and update it regularly, scholars will find it easier to get the most recent ideas that are not found in journal publications.
- Circumvent publications biases. Publications might have biases [5]. Not all research projects end up being published. To circumvent publication bias, it is important to see the topics that are researched but not published. ICA annual conferences serve as a good starting point. Even though these presentations are not publications per se ready to be cited, they are still peer-reviewed. Each year, the acceptance rate is roughly 30%, ensuring the quality of papers presented each year.
- Enable large scientometric analysis. The ICA annual conference data over the past two to three decades are large. It contains over 30K papers, 20K authors and 5K sessions. This dataset is useful for large scale scientometric analysis. For example, 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 as well.

Motivated by these strengths, I obtained and processed all available data on ICA annual conferences. I also developed an API for the data and a user interface to help users navigate the data.

The contributions of this paper are as follows:

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

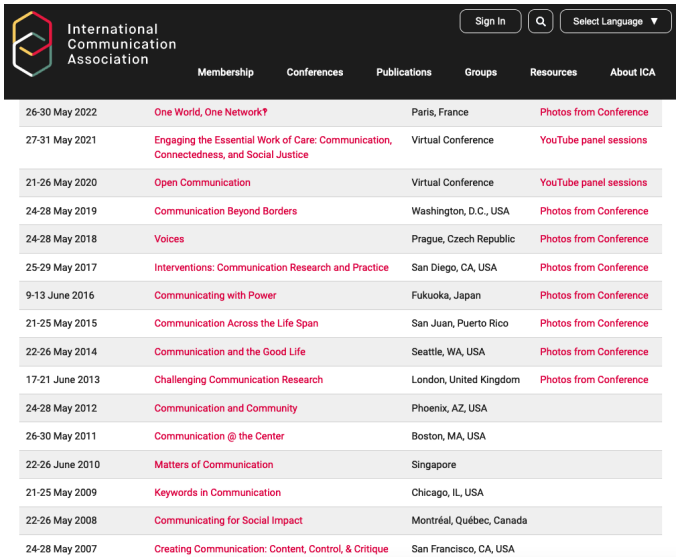
2 Related Work

Scientometric analysis is widely used by scholars in communication, such as in the works of [2] and [1], as well as by researchers in other fields, to gain a deeper understanding of academic landscapes. To support such analyses, several datasets [3] and interfaces [4] have been developed.

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

3 Data Collection and Processing

The official website of ICA provided 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, I decided to leave them alone and focus on online programs instead.

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

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

- 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.

I 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.

4 API Development

4.1 Aggregation and Data Structure

The strength of data is 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, I 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, I 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.

4.2 API

To make the data easier to use for web development, I developed an API using FastAPI. Official documentation is available at https://ica-conf.onrender.com/docs. This API is hosted on render.com.

Below, I will explain the most important endpoints.

4.2.1 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},
 ...
]
```

4.2.2 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}]
```

4.2.3 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]}]
```

4.2.4 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:

```python
import requests

base_url = "https://ica-conf.onrender.com/"

# 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)
```

5 User Interface

Clearly, the data is very large with around 30K papers and 20K authors. To make the data more easily available to the public and interested scholars, I 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.

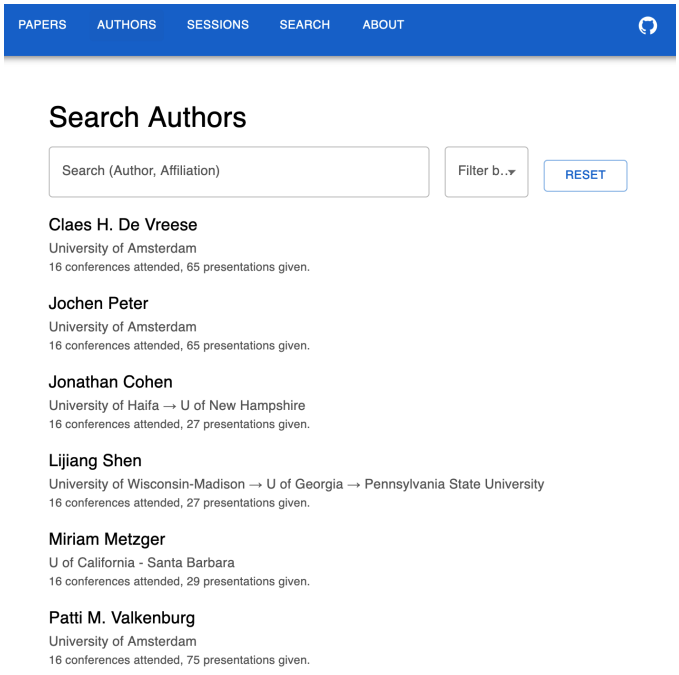


Figure 4: Author Page

The interface also includes authors as in Figure 4 and sessions as in 5. 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.

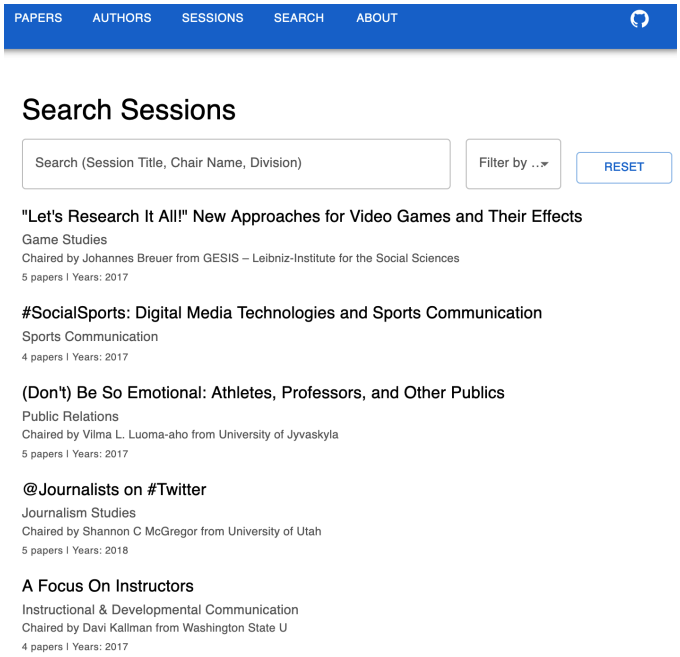


Figure 5: Session Page

6 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.
- In the data, I did not finish deduplication of the author names and 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.

References

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

[2] D. Freelon, M. L. Pruden, and D. Malmer. # politicalcommunicationsowhite: Race and politics in nine communication journals, 1991-2021. *Political Communication*, 40(4):377–395, 2023. 2

[3] P. Isenberg, F. Heimerl, S. Koch, T. Isenberg, P. Xu, C. Stolper, M. Sedlmair, J. Chen, T. Möller, and J. Stasko. vispubdata.org: A Metadata Collection about IEEE Visualization (VIS) Publications. *IEEE Transactions on Visualization and Computer Graphics*, 23, 2017. To appear. doi: 10.1109/TVCG.2016.2615308 2

[4] D. Lange. Vispubs. com: A visualization publications repository. 2

[5] Y. Sun and Z. Pan. Not published is not perished: Addressing publication bias in meta-analytic studies in communication. *Human Communication Research*, 46(2-3):300–321, 2020. 1