sodestream tools: Exploring the Standards Development Process through analysing open data

Ryo Yanagida

sodestream.github.io

This work is funded by the UK Engineering and Physical Sciences Research Council, under grants EP/S033564/1 and EP/S036075/1.

Overview

sodestream project:

- Analysis of Standards Development Organisations (SDOs)
- Open, public data collection and analysis
- Built tools to carryout data collection and analysis

This talk's goal: present the tooling built over the years

- Showcase examples of our findings
  - Documents, authors, and messages
  - Social graph analysis
- Describe how we obtained these findings
- Ongoing, future work, and challenges

sodestream.github.io

Documents, authors, and messages

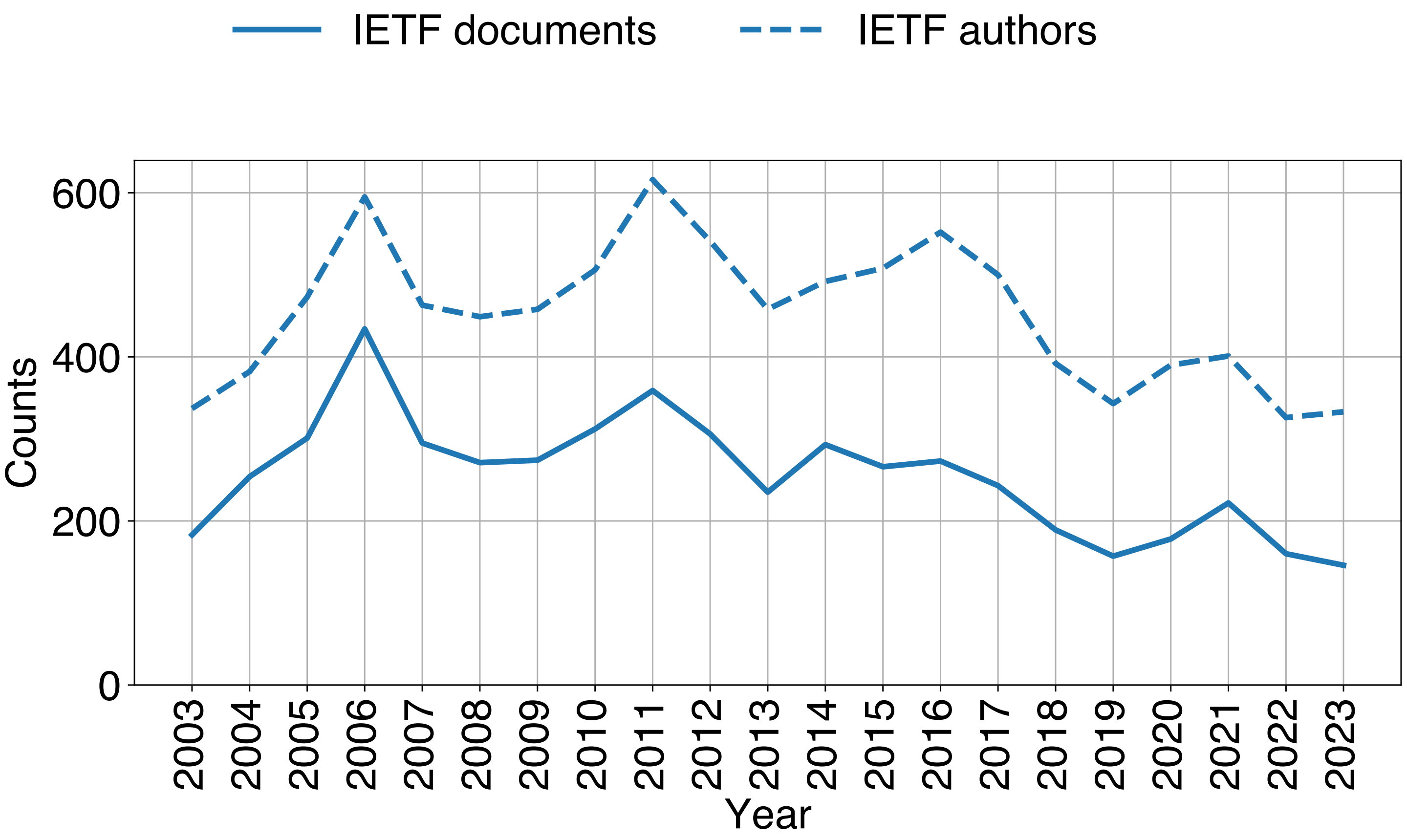
- We built tools to collect documents, authors, and messages data and to calculate stats
- Examples:
    - Annual document stats
    - Annual authors stats
    - Annual author affiliation stats
    - Annual messages stats
- These plots are useful to gauge big pictures around how the activity changed over the years

- These aren’t the only stats — please have a look at our publications for more!

  Please scan the QR Code at the end!

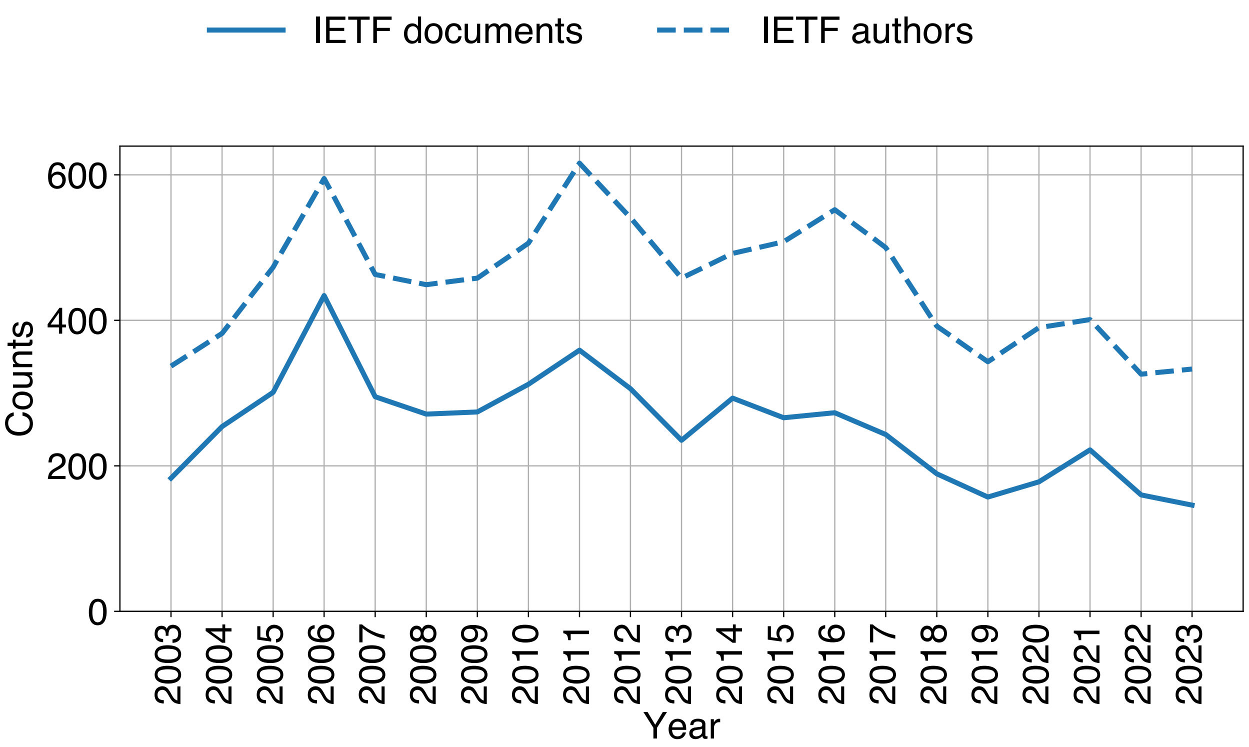
Documents, authors, and messages

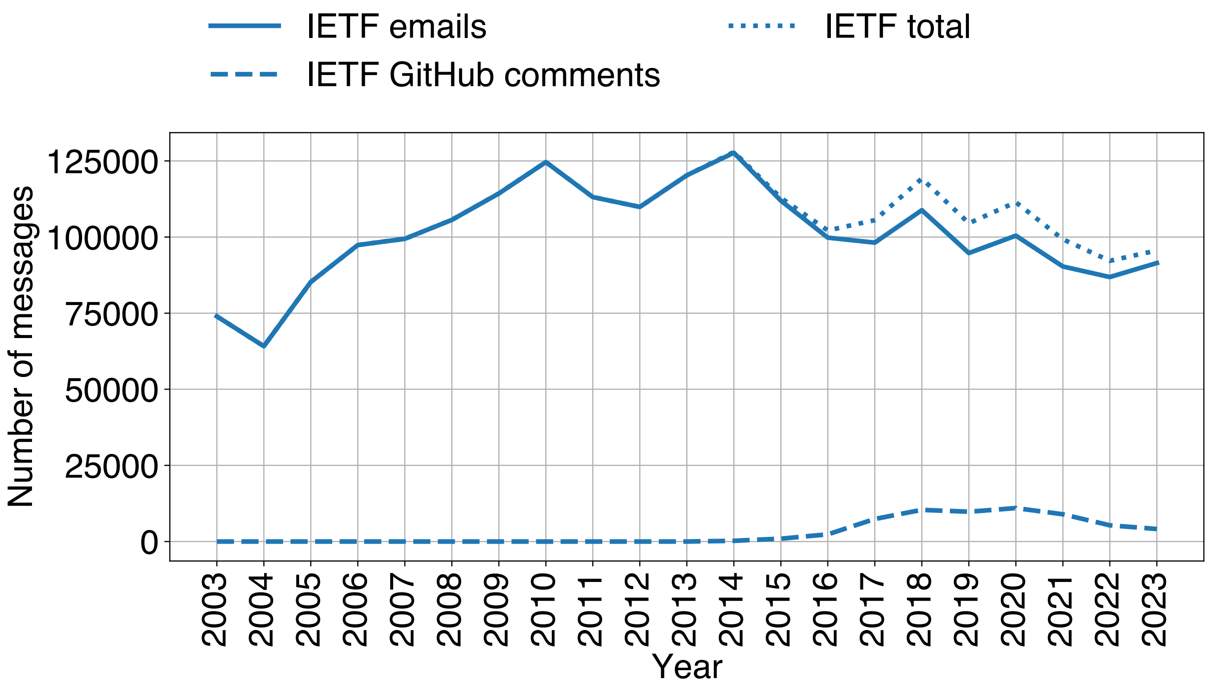
How many documents and authors do we have?



Documents, authors, and messages

Coordination cost per document is increasing!

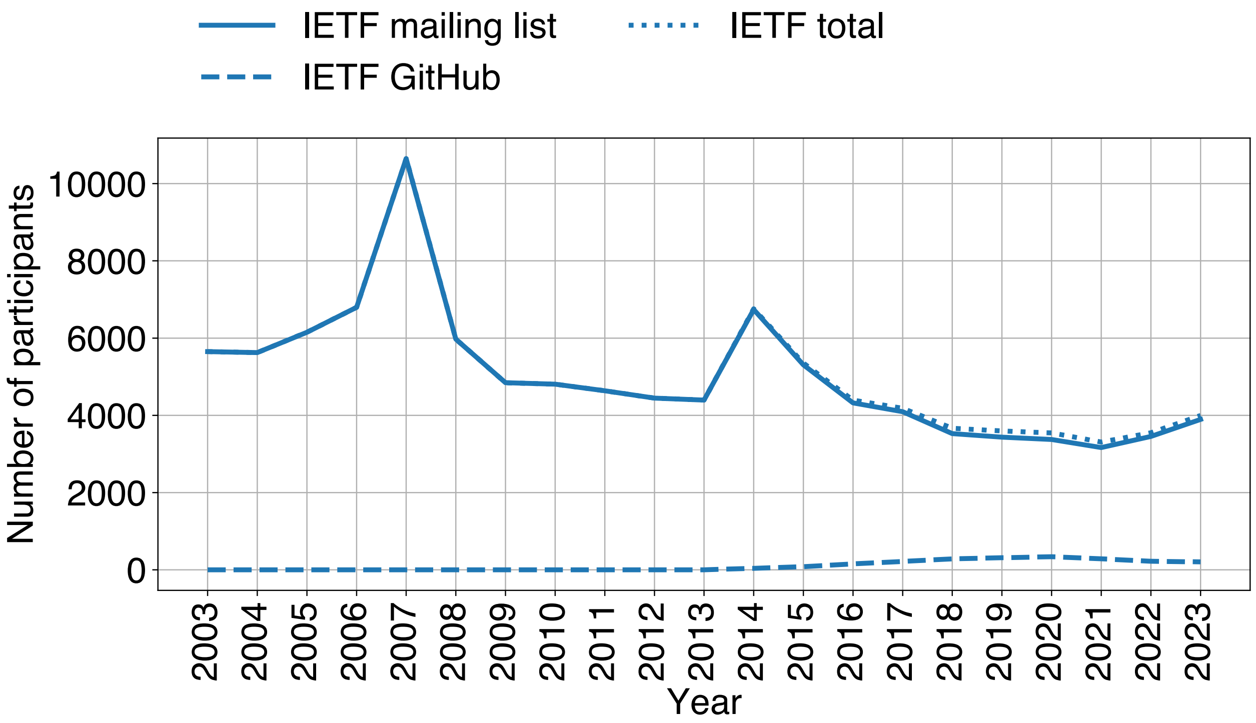


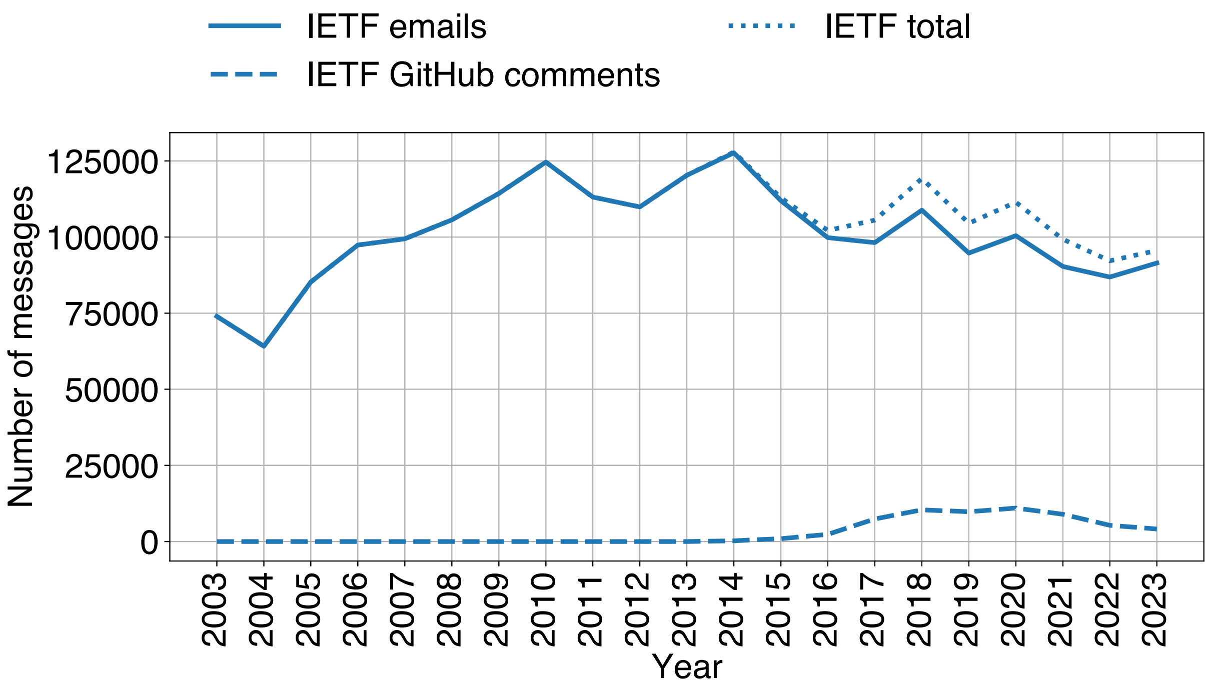


sodestream.github.io

Documents, authors, and messages

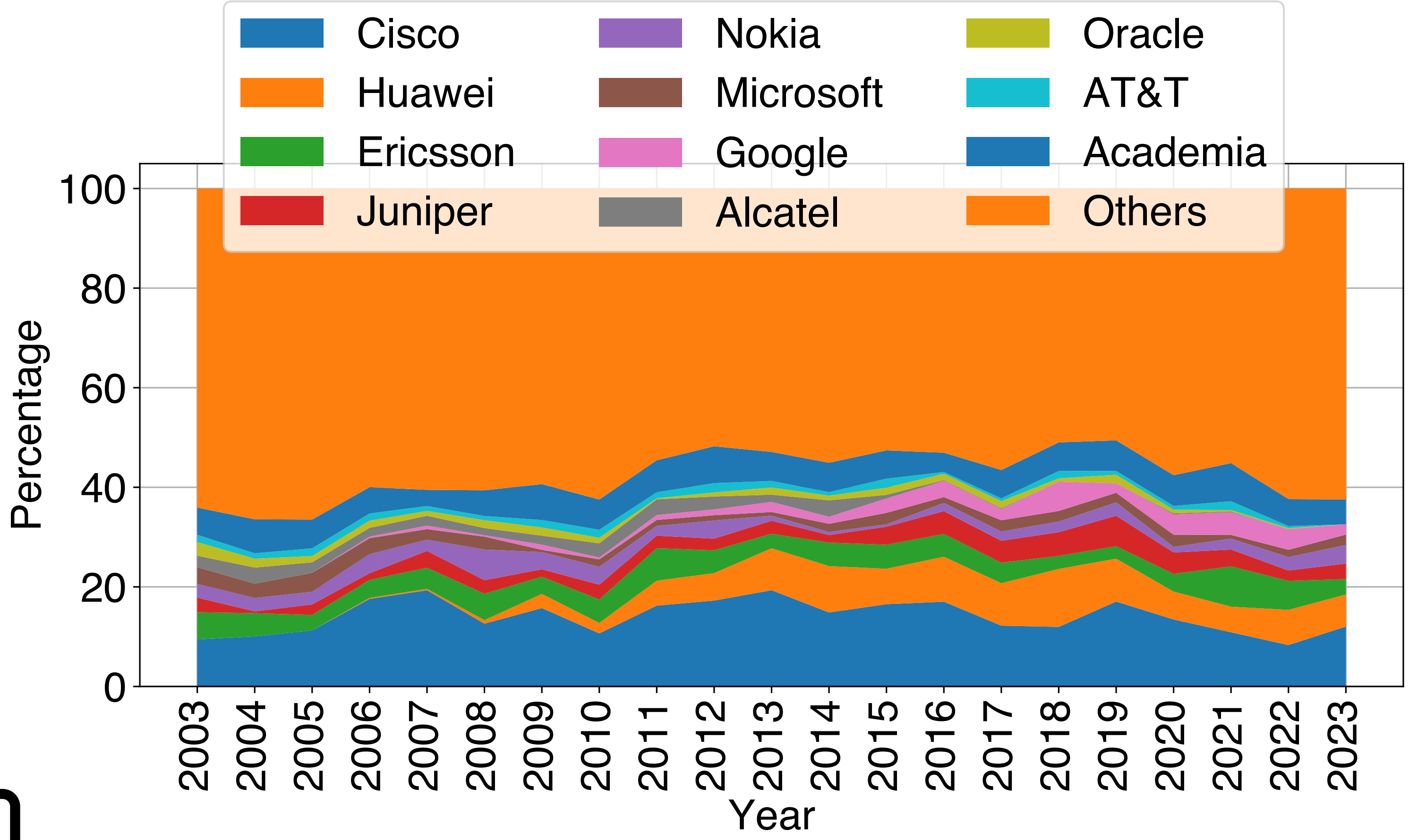
Use of the github is very small at the IETF





Documents, authors, and messages

The breakdown of affiliation has changed.



sodestream.github.io

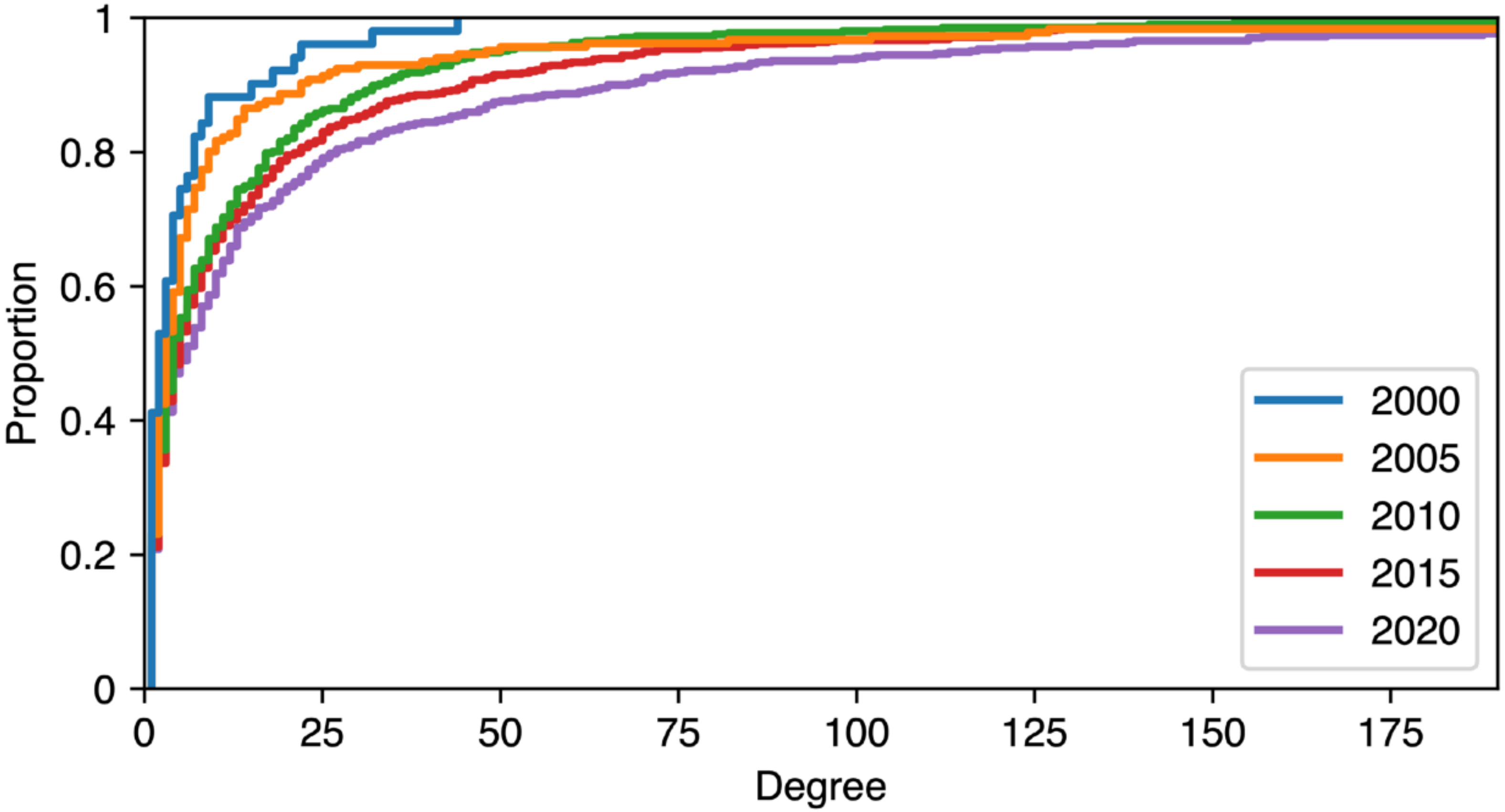
Social graph analysis

- We built tools to analyse how authors interacted over the years
- We constructed social graph
- This is useful to analyse the interaction — key part of understanding how people work

- Again: These aren't the only stats — please have a look at our publications for more!

Social graph analysis

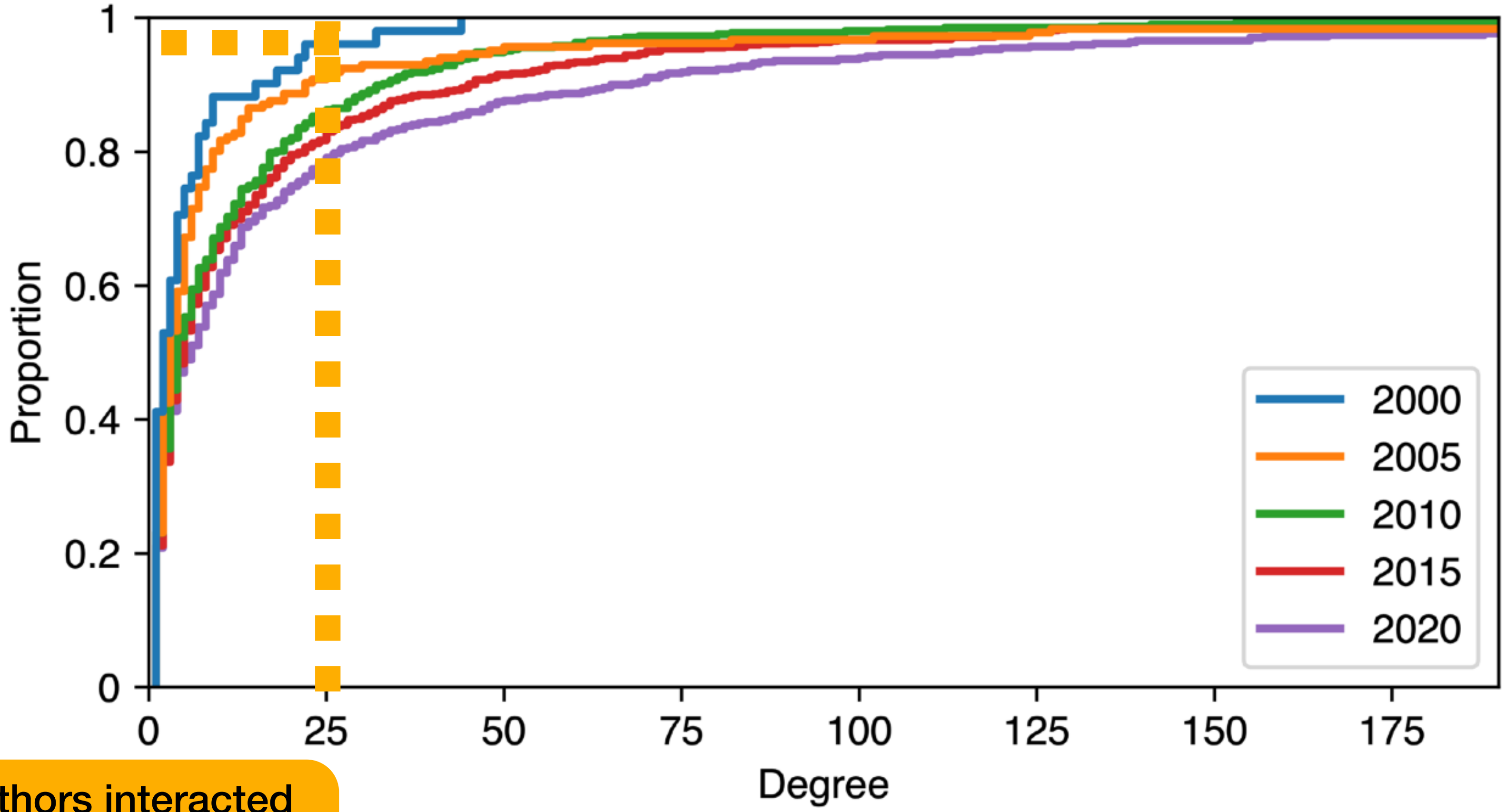
Authors’ interactions over the email seems to be increasing:



i.e. the number of authors interacted

Social graph analysis

Authors' interactions over the email seems to be increasing:

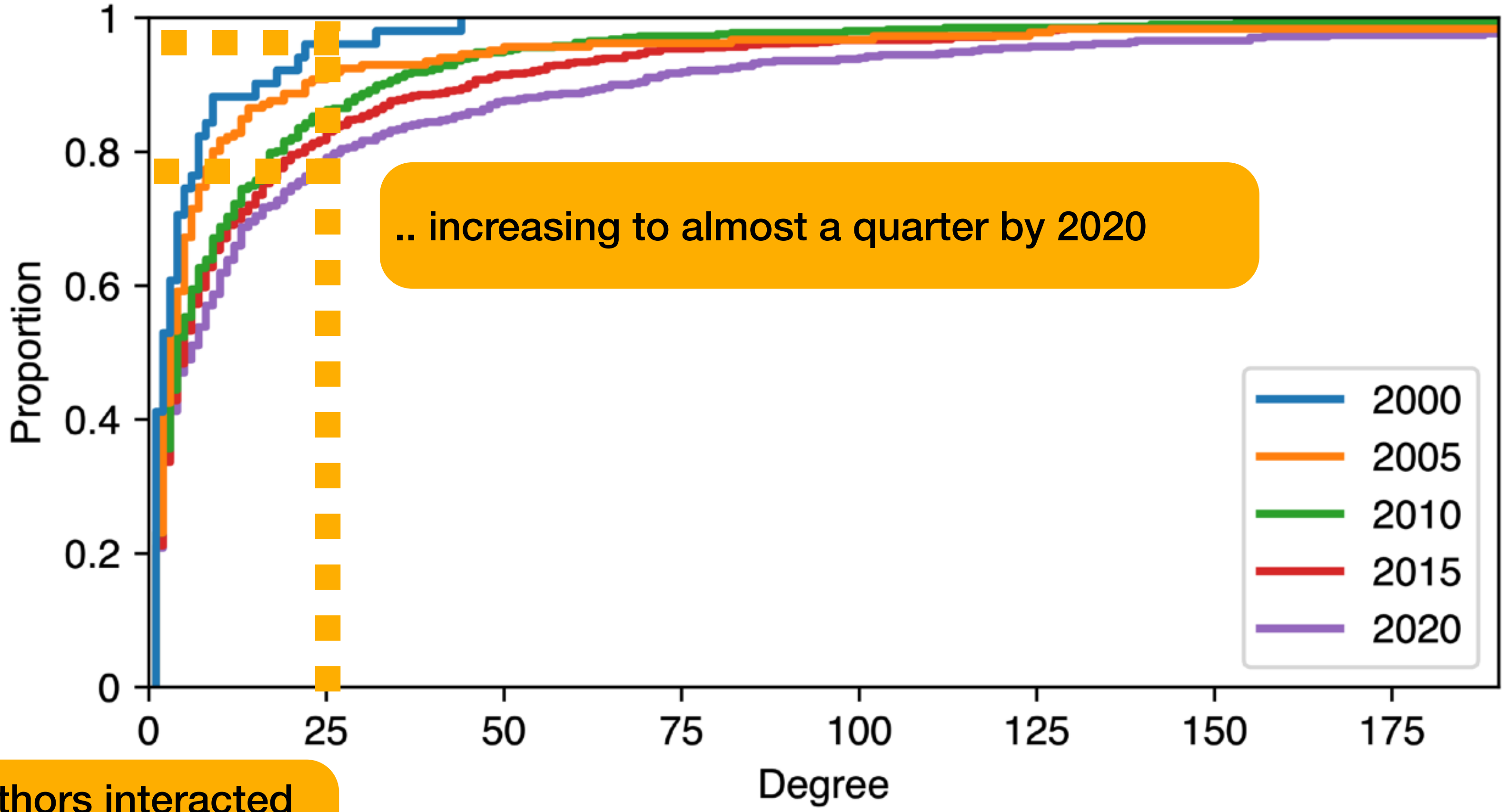


Only 5.5% of authors interacted with more than 25 people in 2000

i.e. the number of authors interacted

Social graph analysis

Authors' interactions over the email seems to be increasing:



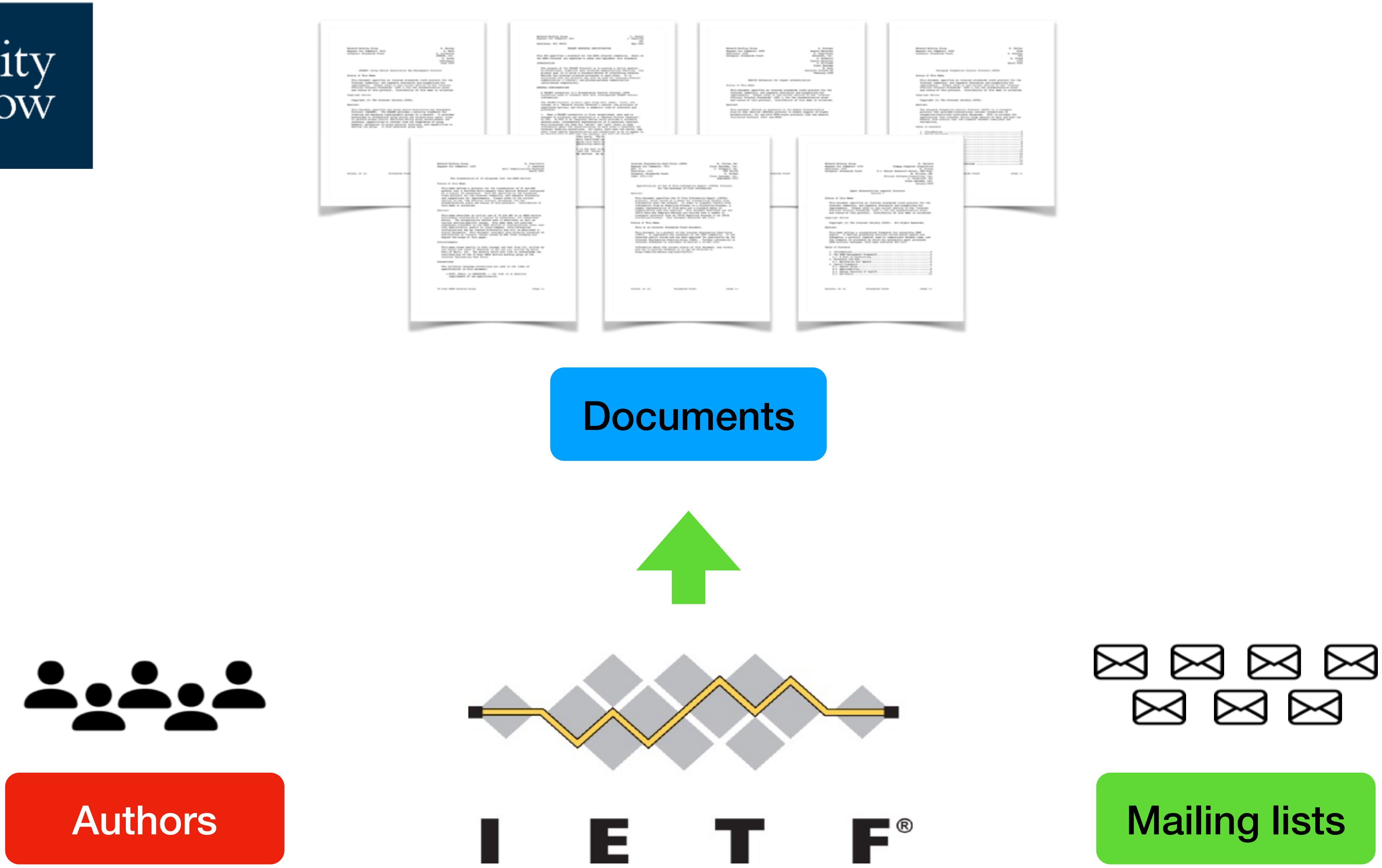
.. increasing to almost a quarter by 2020

Only 5.5% of authors interacted with more than 25 people in 2000

i.e. the number of authors interacted

How did we measure all of this?

University of Glasgow



Documents

Authors

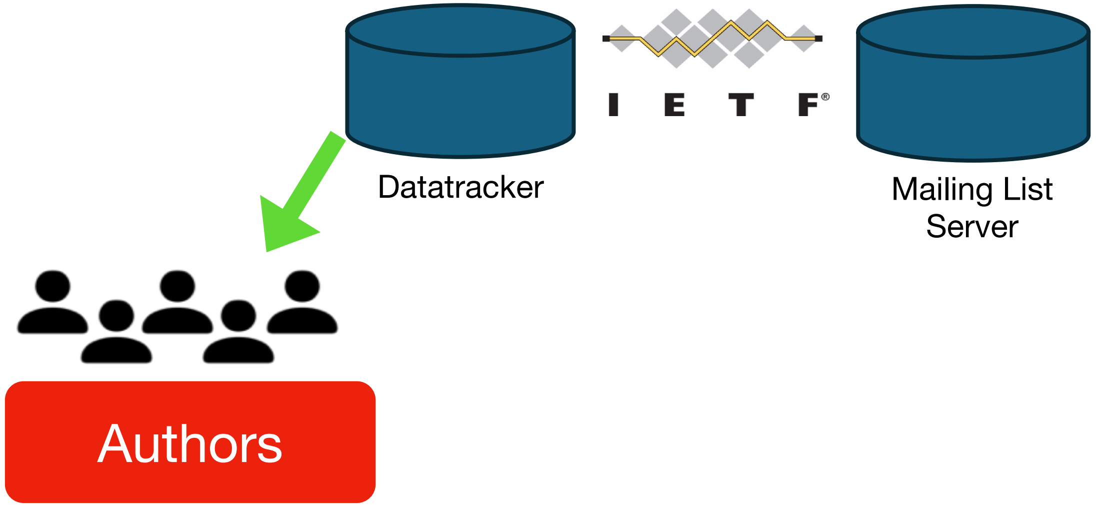
Mailing lists

sodestream.github.io

Datatracker

Mailing List Server

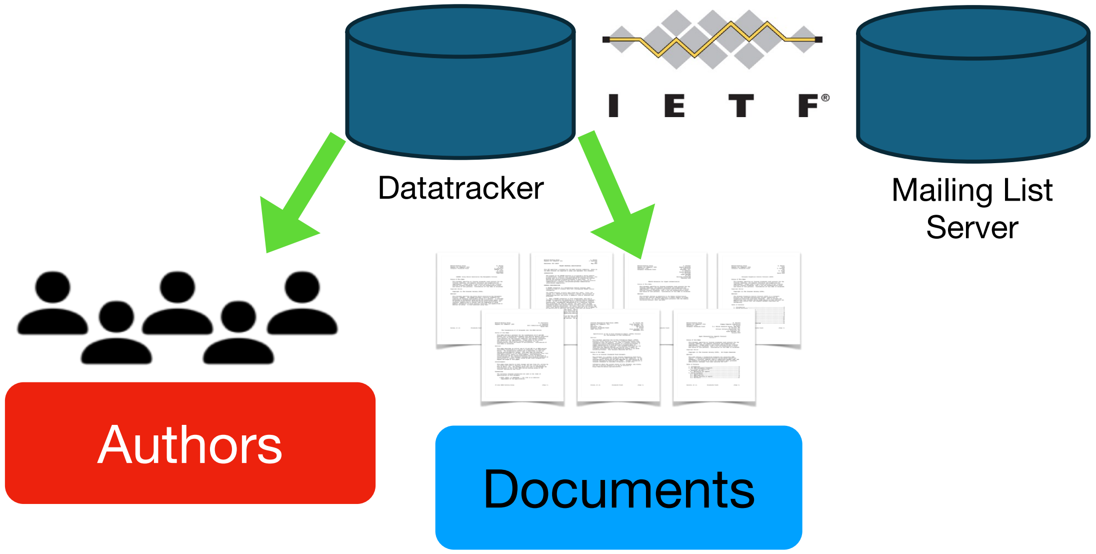
University of Glasgow

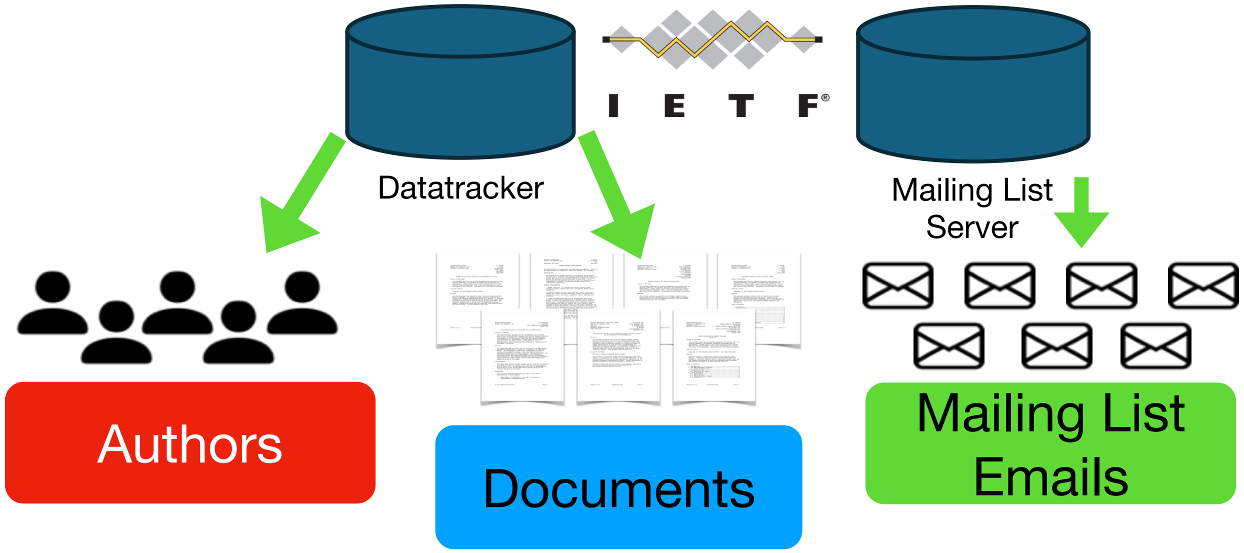


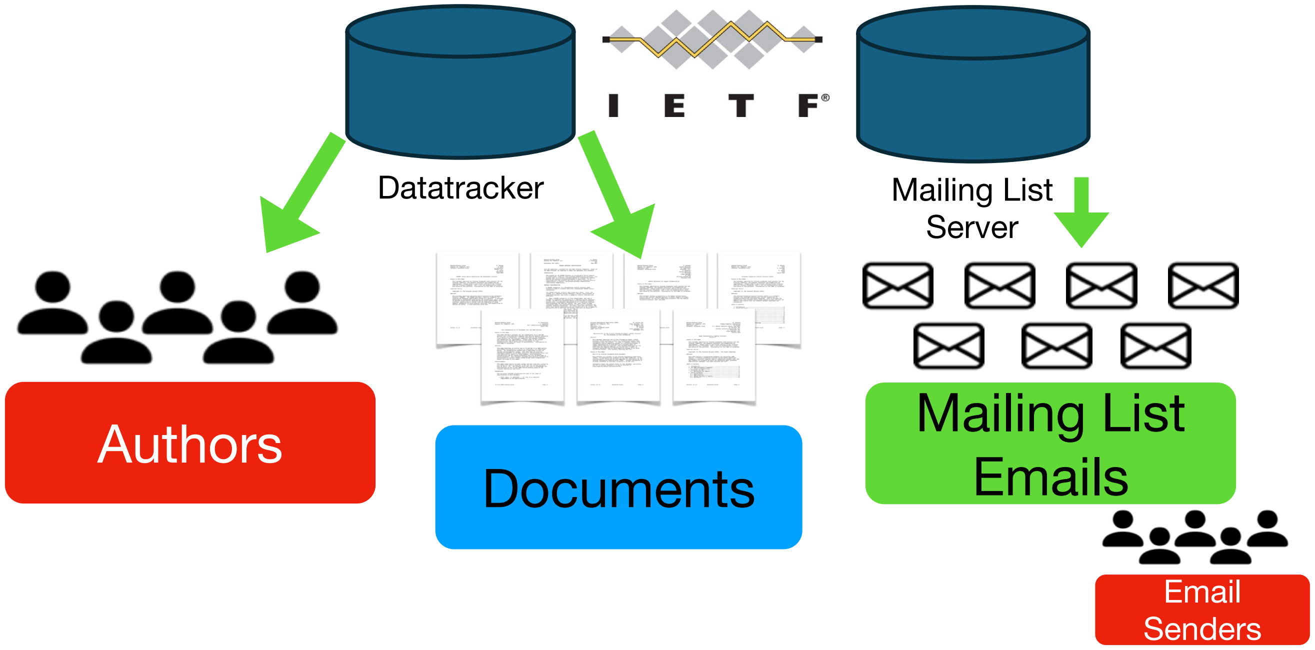
Datatracker

Mailing List Server

Authors


IETF
Datatracker
Mailing List Server
Authors
Documents


IETF
Datatracker
Mailing List Server
Authors
Documents
Mailing List Emails



IETF

Datatracker

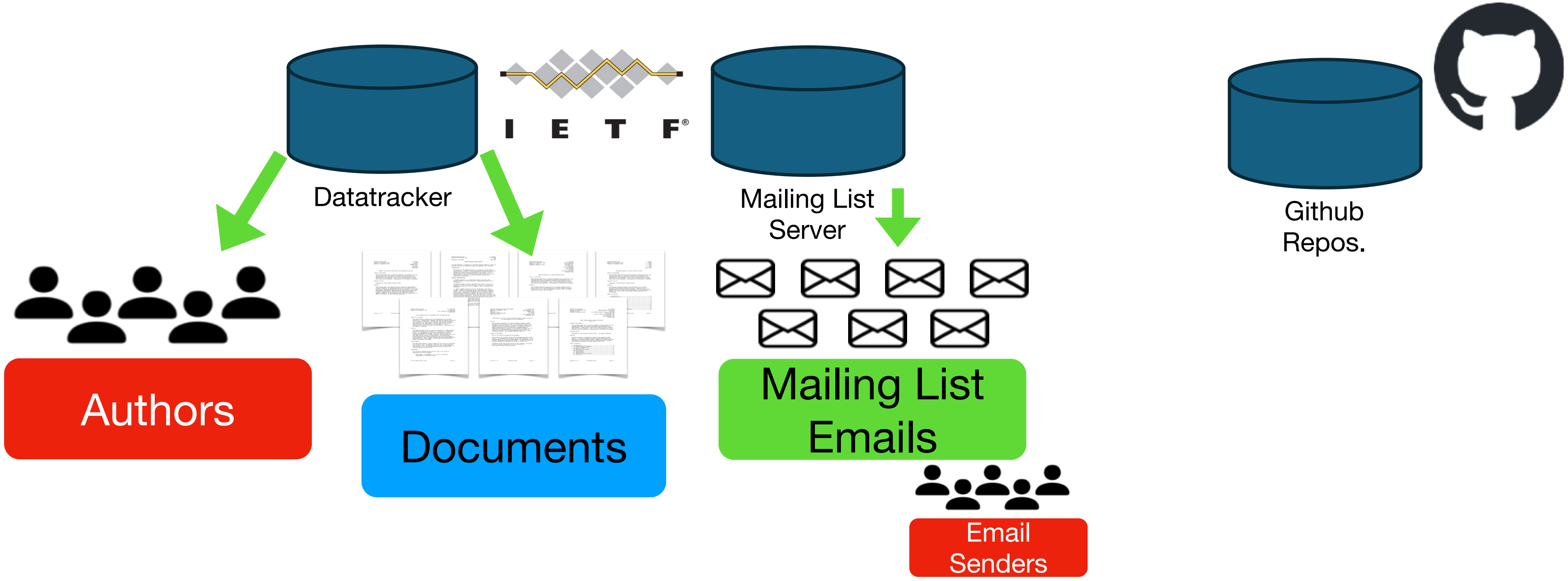
Mailing List Server

Authors

Documents

Mailing List Emails

Email Senders



Datatracker

IETF

Mailing List Server

Github Repos.

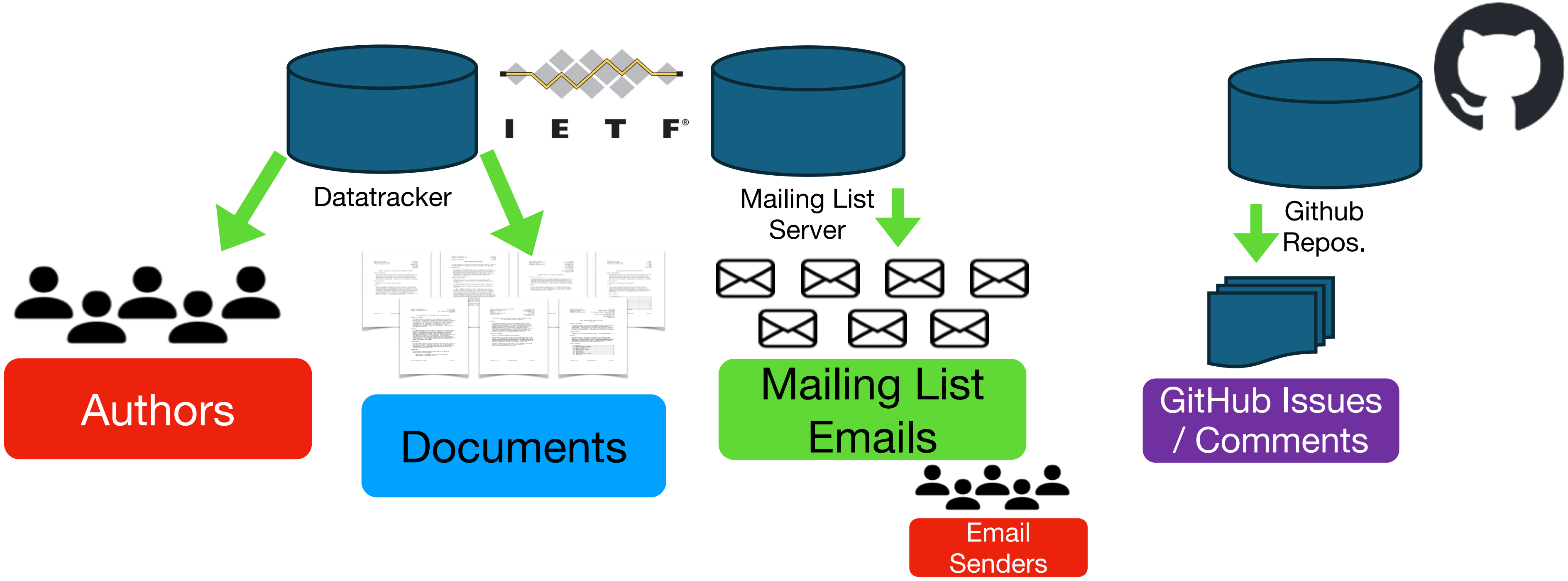
Authors

Documents

Mailing List Emails

Email Senders

University of Glasgow



Datatracker

Mailing List Server

Github Repos.

Authors

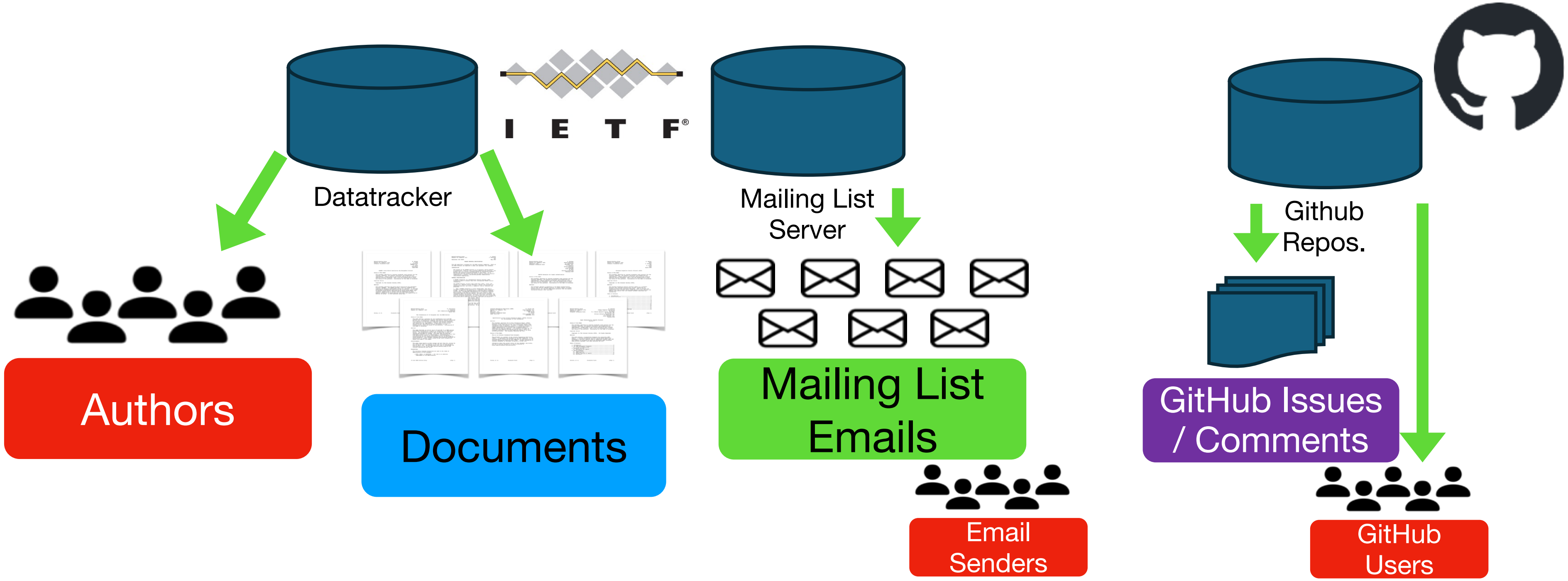
Documents

Mailing List Emails

Email Senders

GitHub Issues / Comments

sodestream.github.io



IETF

Datatracker

Authors

Documents

Mailing List Server

Mailing List Emails

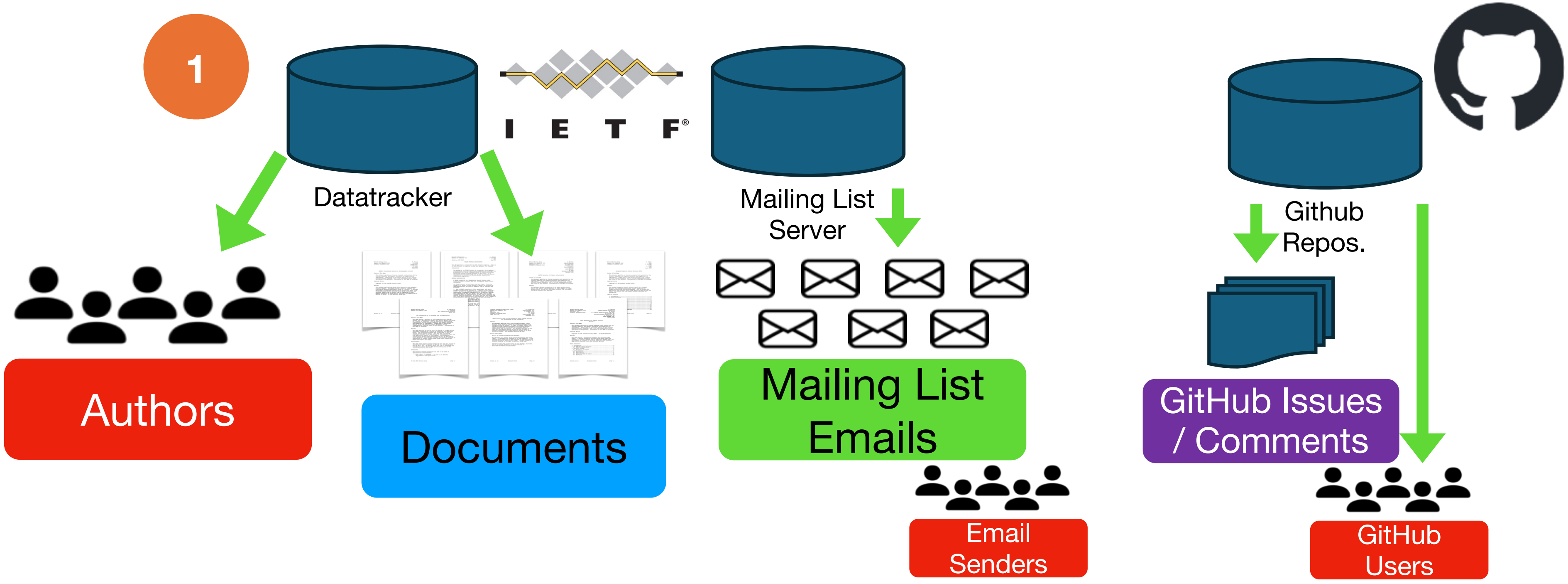
Email Senders

Github Repos.

GitHub Issues / Comments

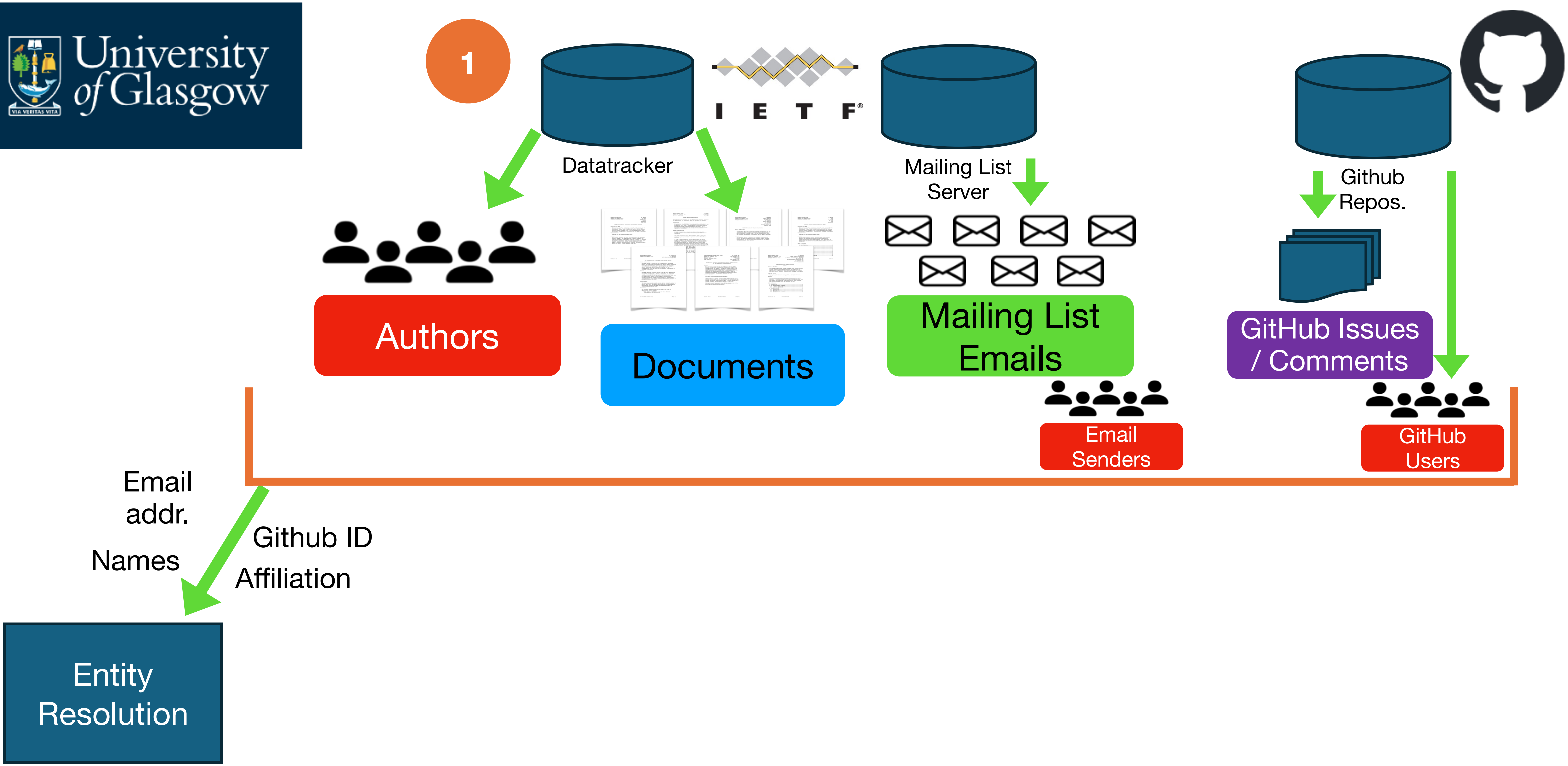
GitHub Users

University of Glasgow



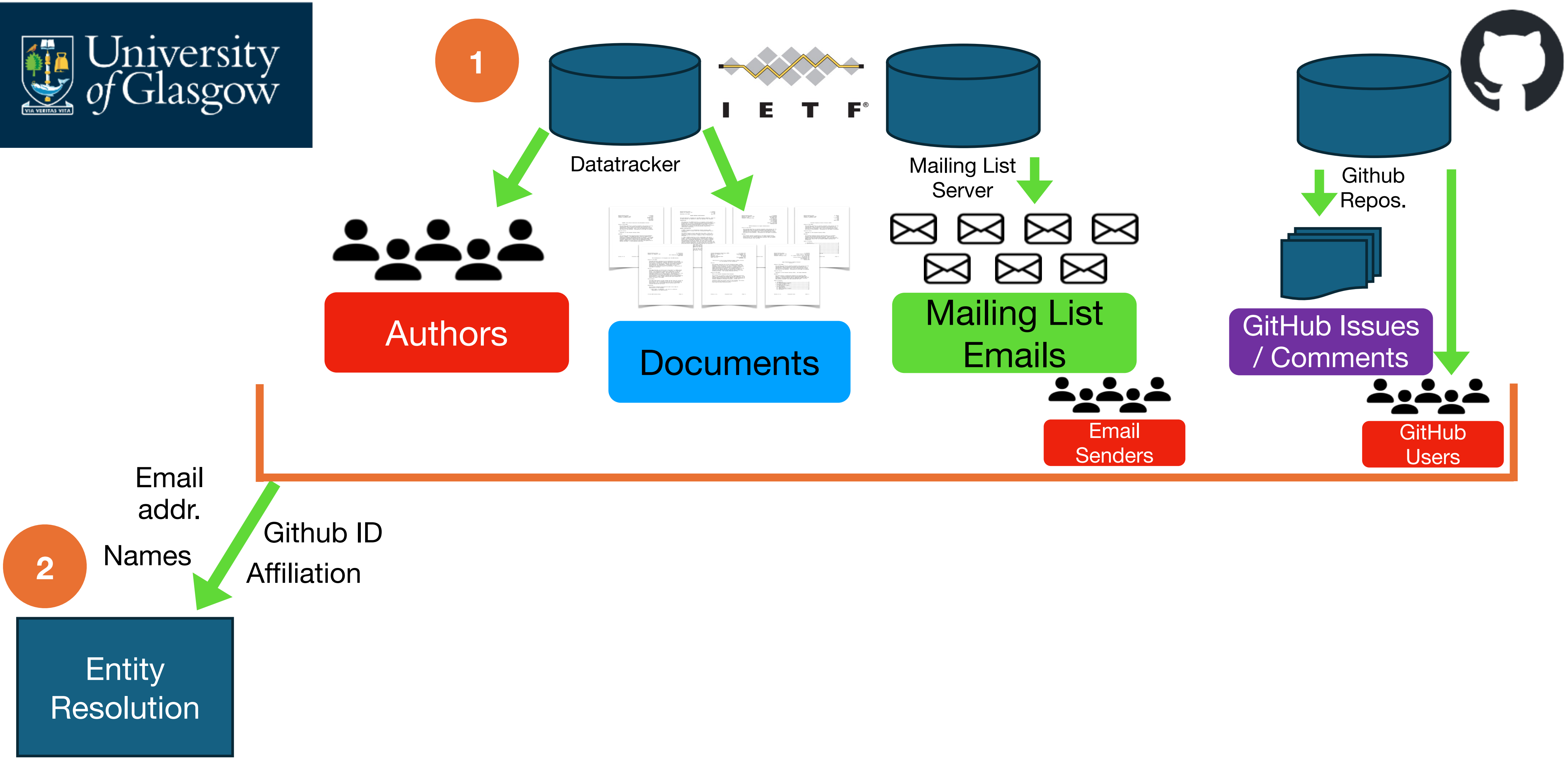
sodestream.github.io

University of Glasgow



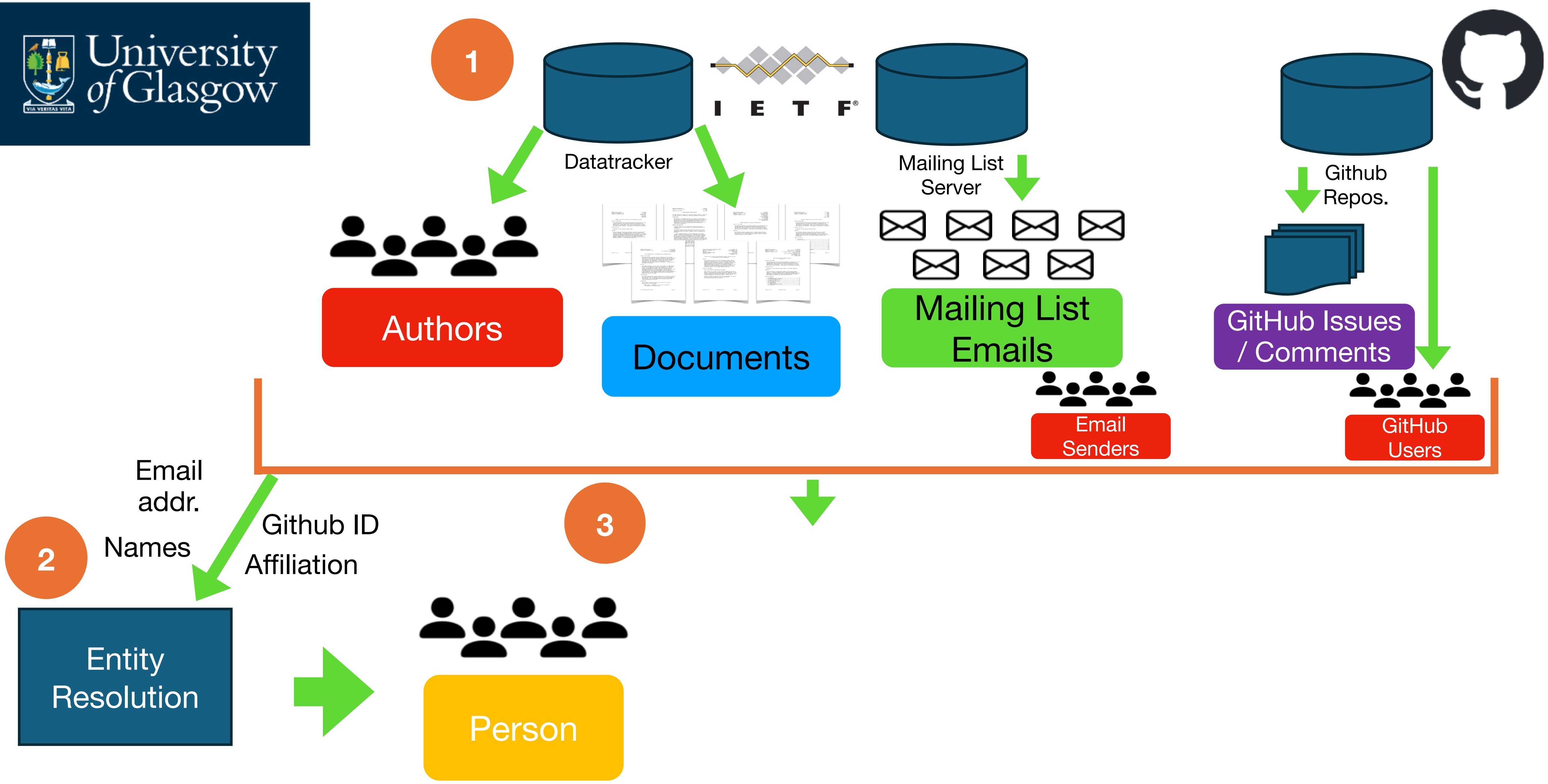
sodestream.github.io

University of Glasgow



sodestream.github.io

University of Glasgow



1

Datatracker

IETF

Mailing List Server

Github Repos.

Authors

Documents

Mailing List Emails

GitHub Issues / Comments

Email Senders

GitHub Users

Email addr.

Names

Github ID

Affiliation

2

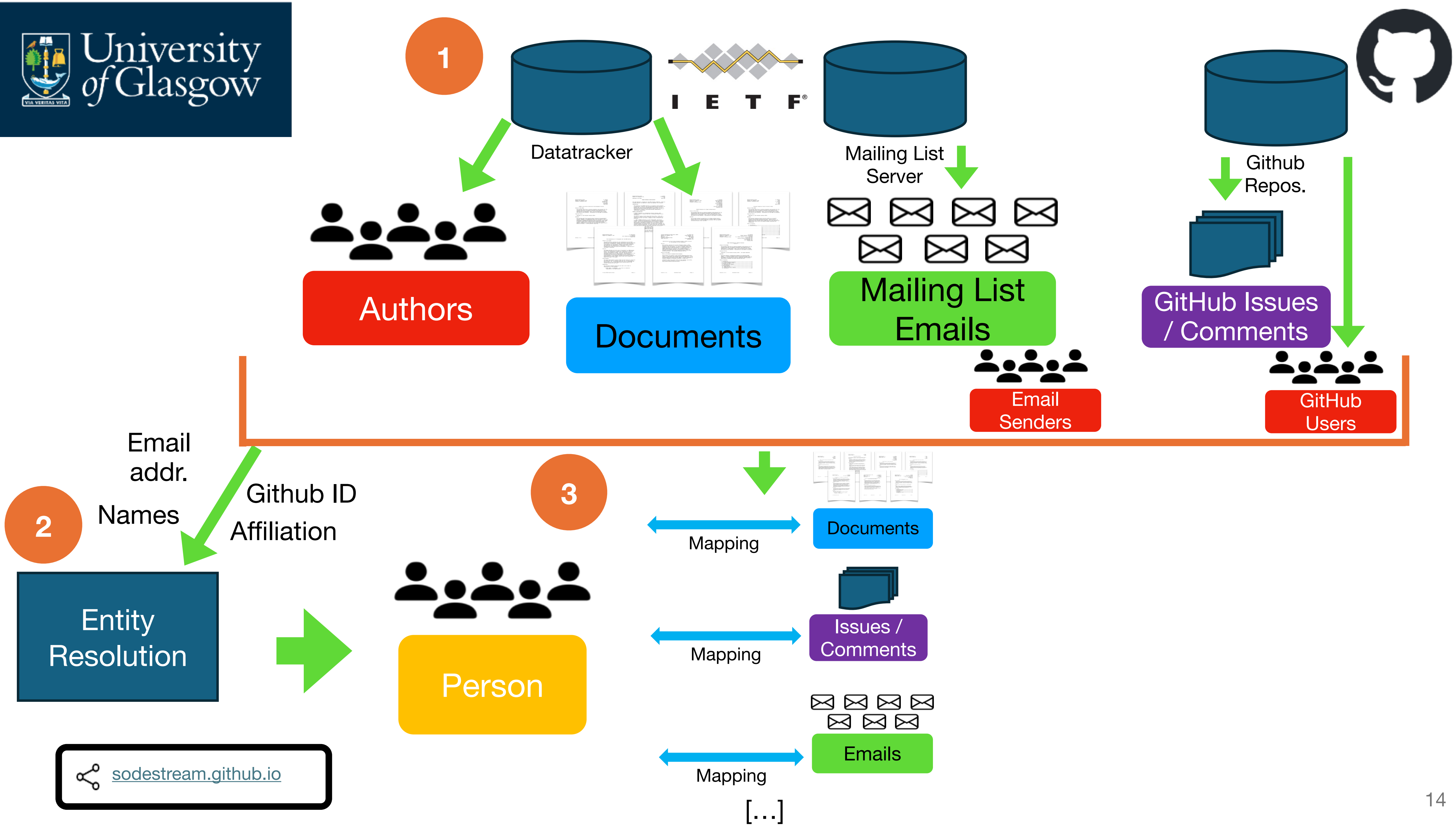
3

Entity Resolution

Person

sodestream.github.io

University of Glasgow



1

Datatracker

IETF

Mailing List Server

Github Repos.

Authors

Documents

Mailing List Emails

GitHub Issues / Comments

Email Senders

GitHub Users

2

Email addr.

Names

Github ID

Affiliation

Entity Resolution

3

Person

Mapping

Documents

Mapping

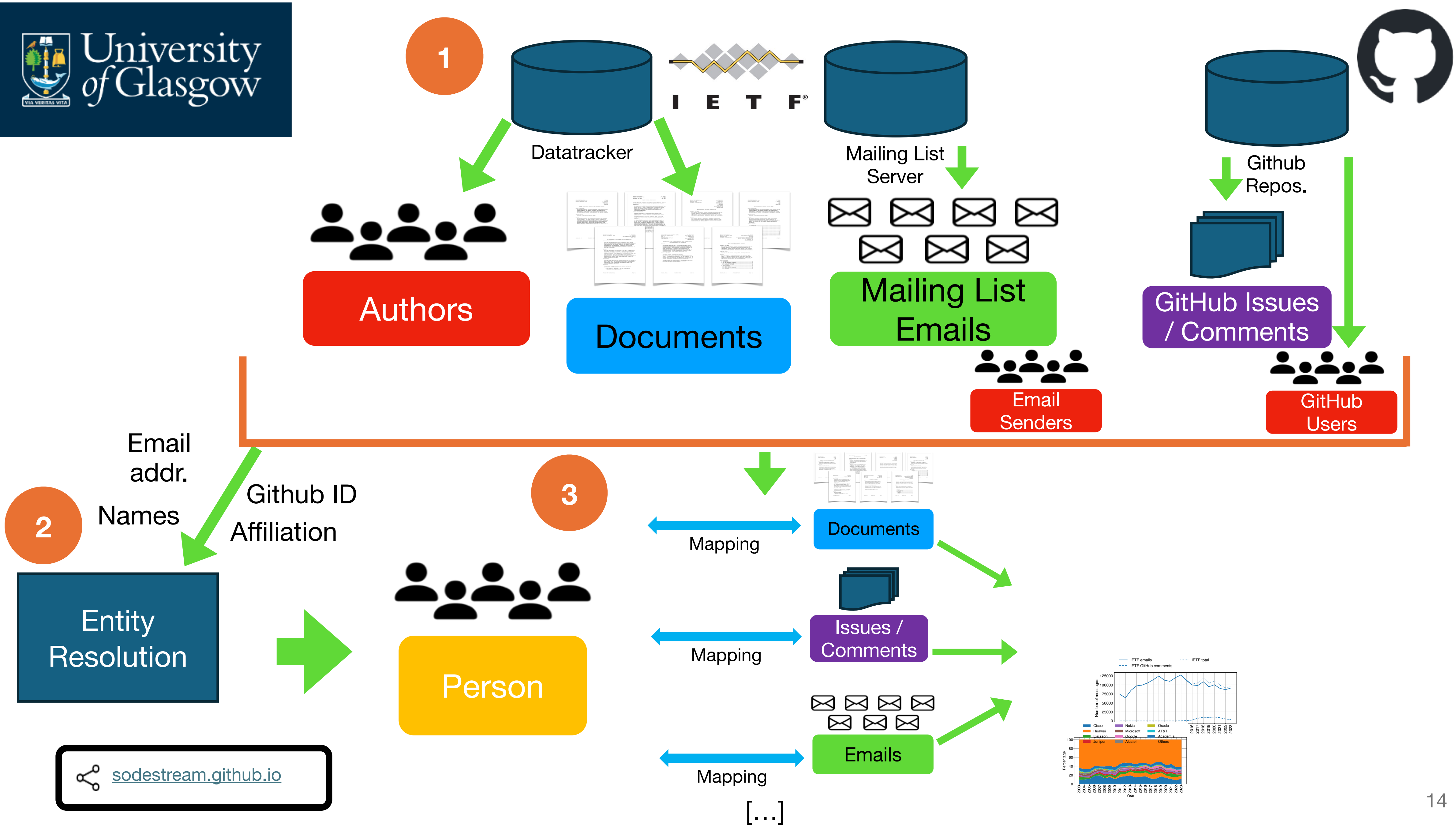
Issues / Comments

Mapping

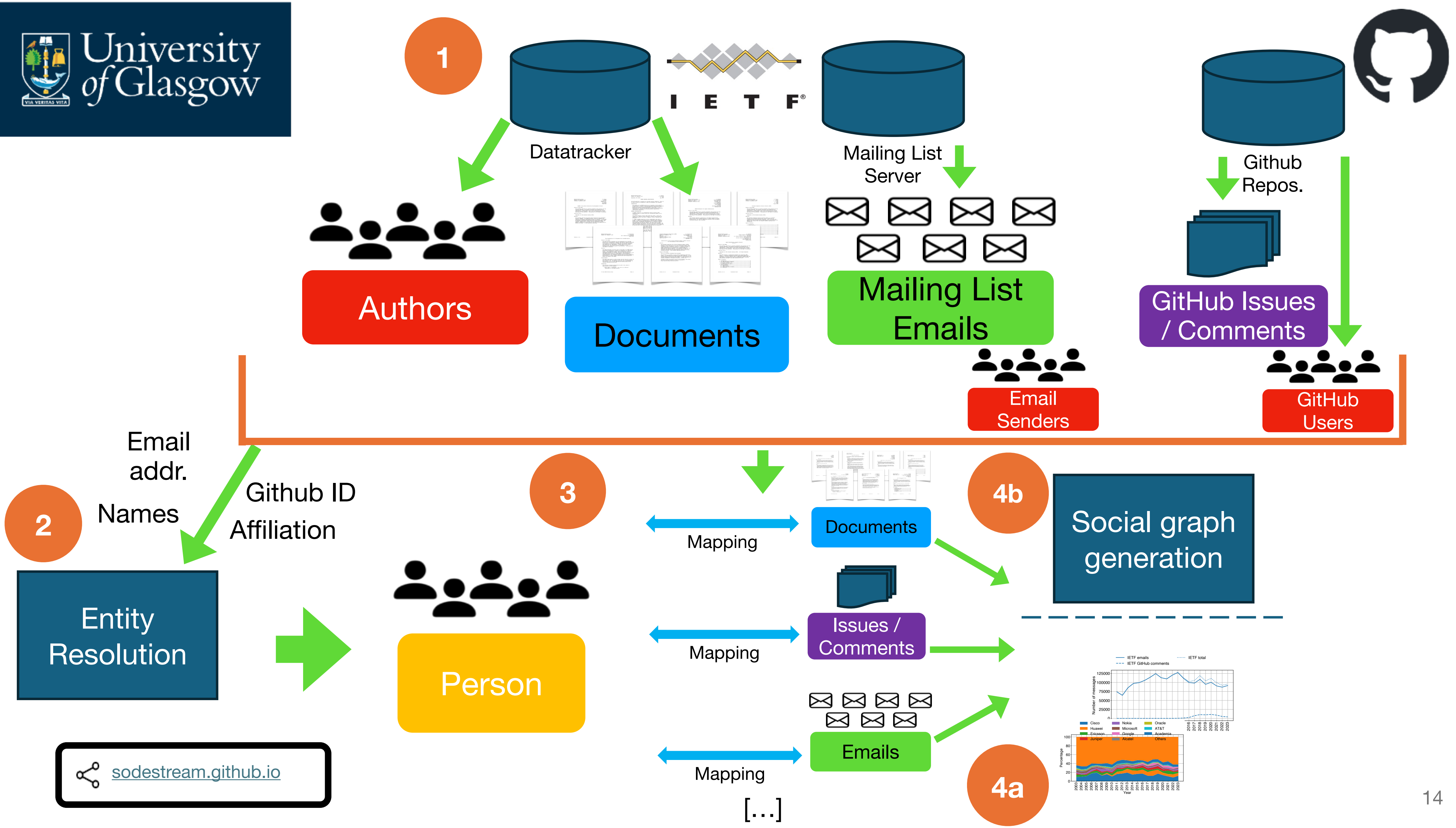
Emails

[…]

sodestream.github.io

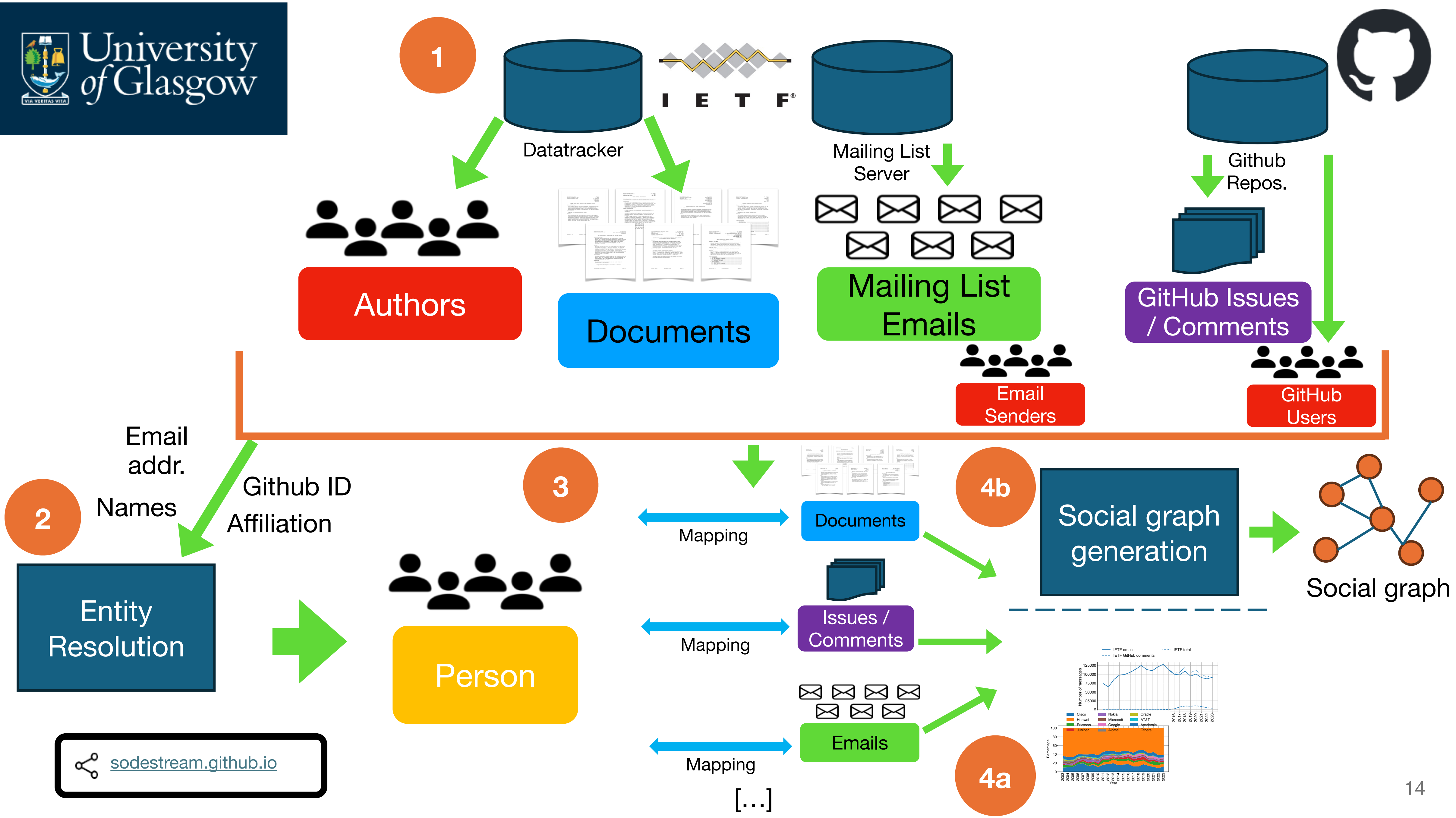


sodestream.github.io



sodestream.github.io

University of Glasgow



1

Datatracker

IETF

Mailing List Server

Github Repos.

Authors

Documents

Mailing List Emails

GitHub Issues / Comments

Email Senders

GitHub Users

Email addr.

Names

Github ID

Affiliation

2

Entity Resolution

Person

3

Mapping

Documents

Mapping

Issues / Comments

Mapping

Emails

[…]

4b

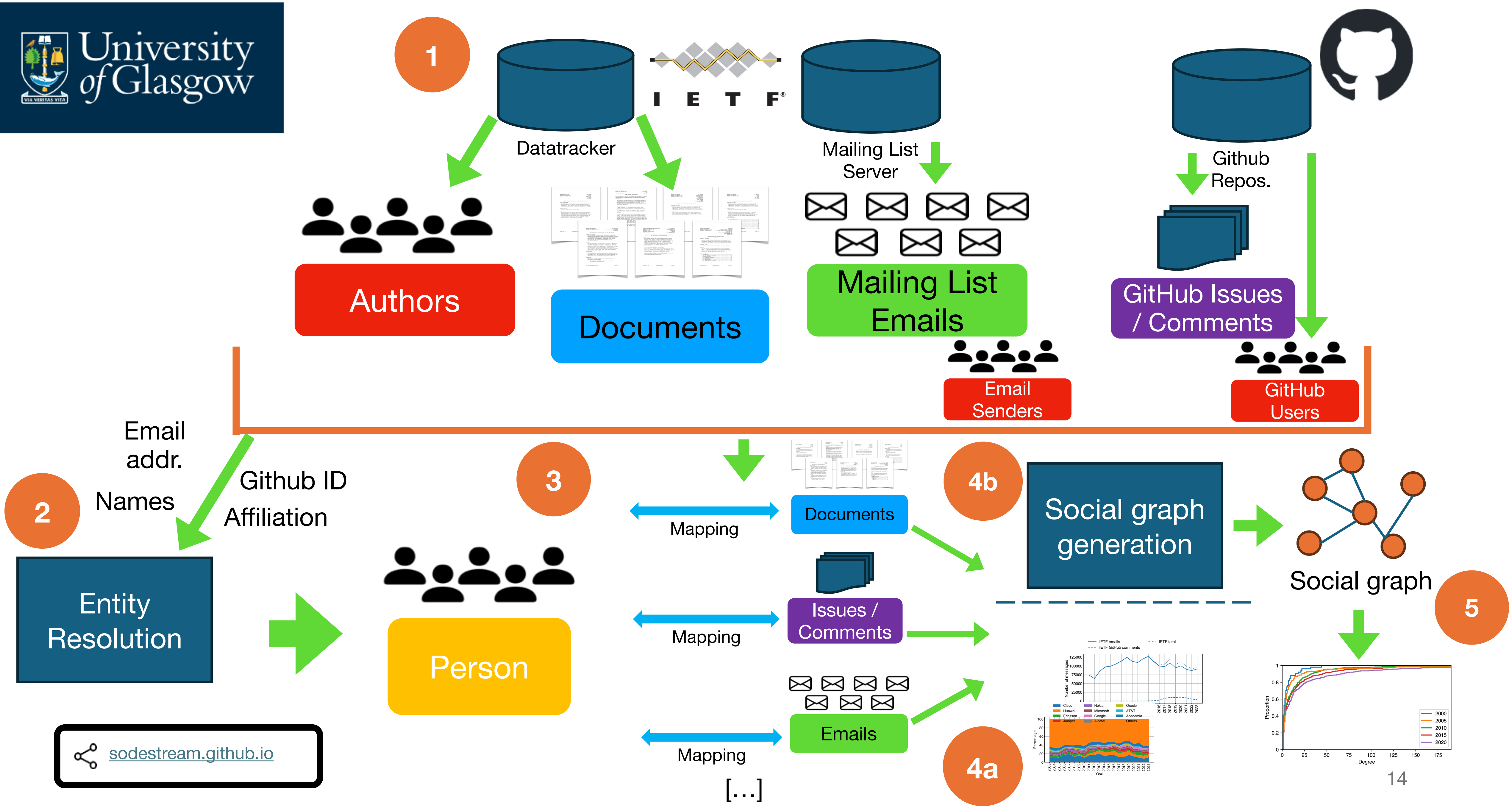
Social graph generation

Social graph

4a

sodestream.github.io

University of Glasgow



1

Datatracker

IETF

Mailing List Server

Github Repos.

Authors

Documents

Mailing List Emails

GitHub Issues / Comments

Email Senders

GitHub Users

Email addr.

Names

Github ID

Affiliation

2

Entity Resolution

Person

3

Mapping

Documents

Mapping

Issues / Comments

Mapping

Emails

[…]

4b

Social graph generation

Social graph

5

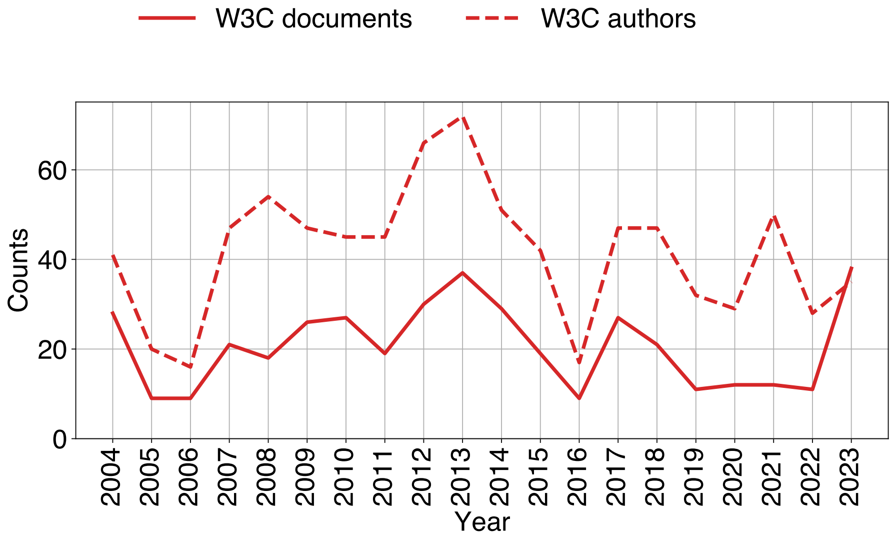
4a

sodestream.github.io

Ongoing / future work

- Branching out to other SDOs — e.g. W3C
- Looking at how people interact with multiple SDOs
  - Moving from one to the other
  - Topics / areas of activities
  - Amount of activities
- And more…

e.g.



sodestream.github.io

Challenges and next steps

sodestream project
https://sodestream.github.io

sodestream github org:
https://github.com/sodestream

- Data is messy
- Entity resolution — i.e. matching individuals while their contacts / name changes over time is challenging
- Mapping different information across SDOs are challenging
  - Different format of information
  - Different amount of data
- These are common to any public data analysis like this — **Please talk to us!**
- We have done some work on entity resolution — **would this RG perhaps be interested if we were to share such code?**

“Characterising the IETF through the lens of RFC deployment”

Stephen McQuistin, Mladen Karan, Prashant Khare, Colin Perkins, Gareth Tyson, Matthew Purver, Patrick Healey, Waleed Iqbal, Junaid Qadir, Ignacio Castro.

ACM IMC 2021

References

“Power and Vulnerability: Managing Sensitive Language in Organisational Communication”. Patrick Healey, Prashant Khare, Gareth Tyson, Mladen Karan, Ignacio Castro, Ravi Shekhar, Stephen McQuistin, Colin Perkins, Matthew Purver. Frontiers in Psychology, section Psychology of Language, 2024. https://sodestream.github.io/paper-power.html

“Temporal Network Analysis of Email Communication Patterns in a Long Standing Hierarchy” Matthew Russell Barnes, Mladen Karan, Stephen McQuistin, Colin Perkins, Gareth Tyson, Matthew Purver, Ignacio Castro, Richard G. Clegg. Proceedings of the 18TH International AAAI Conference on Web and Social Media 2024. https://sodestream.github.io/paper-temporal.html

“Tracing Linguistic Markers of Influence in a Large Online Organisation” Prashant Khare, Ravi Shekhar, Mladen Karan, Stephen McQuistin, Colin Perkins, Ignacio Castro, Gareth Tyson, Patrick G.T. Healey, Matthew Purver. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics, 2023. https://sodestream.github.io/paper-ling-markers.html

“LEDA: a Large-Organization Email-Based Decision-Dialogue-Act Analysis Dataset” Mladen Karan, Prashant Khare, Ravi Shekhar, Stephen McQuistin, Colin Perkins, Ignacio Castro, Gareth Tyson, Patrick G.T. Healey, Matthew Purver. In Findings of the Association for Computational Linguistics: ACL 2023. https://sodestream.github.io/paper-leda.html

“Errare humanum est: What do RFC Errata say about Internet Standards?” Stephen McQuistin, Mladen Karan, Prashant Khare, Colin Perkins, Matthew Purver, Patrick Healey, Ignacio Castro, and Gareth Tyson. In Proceedings of the 7th Network Traffic Measurement and Analysis Conference (TMA) (pp. 1-9). IEEE.2023. https://sodestream.github.io/paper-errare-humanum-est.html

“The Web We Weave: Untangling the Social Graph of the IETF” Prashant Khare, Mladen Karan, Stephen McQuistin, Colin Perkins, Gareth Tyson, Matthew Purver, Patrick Healey, Ignacio Castro. https://sodestream.github.io/paper-the-web-we-weave-untangling-the-social-graph-of-the-ietf.html

“Characterising the IETF Through the Lens of RFC Deployment” Stephen McQuistin, Mladen Karan, Prashant Khare, Colin Perkins, Gareth Tyson, Matthew Purver, Patrick Healey, Waleed Iqbal, Junaid Qadir, Ignacio Castro https://sodestream.github.io/paper-characterising-the-ietf-through-the-lens-of-rfc-deployment.html

“Mitigating Topic Bias when Detecting Decisions in Dialogue” Mladen Karan, Prashant Khare, Patrick Healey, and Matthew Purver. In Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue (pp. 542-547), 2021. https://sodestream.github.io/paper-mitigating-topic-bias-when-detecting-decisions-in-dialogue.html