Workgroup: RASPRG
Internet-Draft:
draft-perkins-analysing-sdo-data-00
Published: 6 May 2026
Intended Status: Informational
Expires: 7 November 2026
Authors: C. Perkins
University of Glasgow
I. Castro
Queen Mary University of London
R. Yanagida
University of St Andrews
S. McQuistin
University of St Andrews

# Analysing Internet Standards Development Organisation Data

## Abstract

This document outlines some issues to consider when studying data relating to the Internet standards development ecosystem. It identifies observable components of standards development processes, proposes a taxonomy of possible measurements, and highlights methodological, interpretive, and ethical considerations. It is intended to support a range of uses, including monitoring standards development organisations (SDOs), evaluating the evolution of technical work, understanding technology deployment, and informing community, leadership, and governance discussions.

This document is submitted for consideration by the Research and Analysis of Standard-Setting Processes Research Group (RASPRG) in the IRTF. It is not an IETF product and is not a standard.

## About This Document

This note is to be removed before publishing as an RFC.

The latest revision of this draft can be found at https://csperkins.github.io/draft-analysing-sdo-data/draft-perkins-analysing-sdo-data.html. Status information for this document may be found at https://datatracker.ietf.org/doc/draft-perkins-analysing-sdo-data/.

Discussion of this document takes place on the RASPRG Research Group mailing list (mailto:rasprg@irtf.org), which is archived at https://mailarchive.ietf.org/arch/browse/rasprg/. Subscribe at https://www.ietf.org/mailman/listinfo/rasprg/.

Source for this draft and an issue tracker can be found at https://github.com/csperkins/draft-analysing-sdo-data.

## Status of This Memo

This Internet-Draft is submitted in full conformance with the provisions of BCP 78 and BCP 79.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF). Note that other groups may also distribute working documents as Internet-Drafts. The list of current Internet-Drafts is at https://datatracker.ietf.org/drafts/current/.

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress."

This Internet-Draft will expire on 7 November 2026.

## Copyright Notice

Copyright (c) 2026 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (https://trustee.ietf.org/license-info) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document.

## Table of Contents

1. Introduction
2. Standards Development as a Socio-Technical System
3. Analysing the IETF
   3.1. Datatracker
   3.2. RFC Editor
   3.3. Mailing List Archives
   3.4. Session Recordings
   3.5. Chat Archives
   3.6. GitHub
4. Analysing Other SDOs
   4.1. Data Availability Across SDOs
   4.2. Integrating Data Across SDOs
5. Data Processing
6. Ethics and Data Protection
7. Recommendations
   7.1. Recommendations for the IETF
   7.2. Recommendations for Researchers
8. Security Considerations
9. IANA Considerations
10. Informative References
Acknowledgments
Authors' Addresses

## 1. Introduction

Internet technologies are developed and standardised by a range of standards development organisations (SDOs), including the IETF, W3C,

IEEE, 3GPP, ITU-T, and others. The standards that these organisations produce underpin the interoperability and architectural evolution of the Internet and the Web.

Understanding how standards are developed, including who participates in the standards process, what collaborations occur during the development of standards, how the process is organised and governed, and the technical outputs, can support analysis of standards ecosystems. Such analysis can assist with monitoring standards development organisations, evaluating the evolution of technical work, understanding technology deployment, and ultimately be used to inform community leadership and governance discussions.

This document outlines considerations for studying data from the Internet standards development ecosystem. It aims to:

* identify observable components of the Internet standards development ecosystem;

* describe considerations for measuring and analysing the standards development process;

* provide a taxonomy of possible measurements and analytical approaches;

* highlight methodological, interpretive, and ethical considerations;

* illustrate the application of these methods to the IETF, given the availability of rich data about the IETF participants, documents, processes, and communication channels;

* discuss the relevance and limits of applying these methods to other SDOs and the extent to which differences in governance, transparency, and data availability affect such analysis; and

* encourage reproducible research practises and transparent analysis.

This document does not prescribe specific metrics, define evaluation criteria, or recommend approaches to comparative rankings of standards bodies, groups, or participants.

**TODO**: Cite prior work. This document currently does a very poor job of citing prior work. This omission will be remedied in future versions. While it is not intended that this document forms the basis for a comprehensive literature survey, if you know of relevant prior work that should be cited please contact the authors.

## 2. Standards Development as a Socio-Technical System

Internet standards development can be understood as a socio-technical system in which technical artefacts, human participants,

organisational interests, and governance processes interact over time. Standards do not emerge solely from technical design choices, nor solely from institutional processes; rather, they arise through structured collaboration among individuals and organisations operating within formal and informal rules.

Technical outputs emerge from a socio-technical process in which engineering choices interact with expertise, incentives, organisational structures, review processes, historical precedent, deployment constraints, and the cultural norms and practices of the standards community. At the same time, the organisational and cultural context is not fixed: governance structures, working practices, and community norms evolve together over time and these changes in turn shape future participation and technical decision-making.

For analytical purposes, standards development ecosystems can be viewed as comprising several interacting components:

* **Participants:** Participants are the individuals who contribute to standards development. They may include engineers, researchers, operators, implementers, academics, independent contributors, civil society representatives, policy specialists, and others with relevant expertise or interests. Participation criteria differ across SDOs. Some use open participation, while others structure participation through organisational- or state-based membership, sometimes with additional exceptions or parallel open mechanisms.

  Participation models affect standards development by shaping both who is able to contribute, and how they are permitted to contribute. Open participation can broaden the pool of contributors and make it easier for individuals to join without prior institutional affiliation, which may increase diversity of experience and viewpoints. At the same time, openness does not eliminate all the barriers to participation. Effective participation may still depend on having sufficient time, funding, employer support, travel resources, and familiarity with the processes, tools, and norms of the community. Membership-based models may provide clearer institutional commitment and resourcing, but they can also limit participation to those acting through recognised organisations or membership categories.

* **Organisations:** Participants are often affiliated with organisations such as companies, academic institutions, governments, consultancies, or civil society groups. These organisations may provide forms of support including funding, staff time, technical expertise, or implementation experience.

  The relationship between participants and organisations is not equally visible across SDOs. In some models, participation is individual, and so any recorded affiliation may be incomplete, and reflect a specific contribution rather than the sustained view of the participant. In other models, where individuals participate on

behalf of a clearly indicated affiliation, the institutional link is clearer.

Even where affiliations are recorded, they may not fully describe the organisational context. A company may be a subsidiary of another company (or in the process of becoming so), and consultants or contractors may work for clients whose interests are not directly visible in participation records.

* **Technical Groups:** SDOs typically organise work through technical groups such as working groups, research groups, study group, committees, or similar bodies. These groups define scope, coordinate discussion, and develop technical outputs. They are not always organised as a single flat layer, with hierarchical and other structures in use.

  The number, names, and functions of these structures differ across organisations. In some cases, they reflect administrative oversight or broad technical areas; in others, they distinguish between different forms of technical development.

* **Artefacts:** Standards processes generate artefacts such as drafts, specifications, recommendations, reports, agendas, minutes, presentations, issue trackers, and final published standards. These artefacts provide an observable record of technical development. Revision histories, references, and relationships between documents may help reveal aspects such as participation dynamics, design iteration, and the evolution of the underlying technologies subject to standardisation.

  Different SDOs vary in how openly they make such information available and in how easily it can be accessed and reused. artefact availability can support the work of participants, researchers, and other observers, but collecting, maintaining, publishing, and organising this information also imposes costs on SDOs.

* **Collaboration Infrastructure:** Standards development requires communication among participants to propose work, discuss technical issues, review contributions, coordinate activity, resolve disagreements, and build support for possible outcomes. SDOs therefore rely on systems such as mailing lists, code repositories, and meetings to facilitate this debate.

  The mix of communication, collaboration, and coordination mechanisms differs across SDOs, often to support the other attributes described.

* **Governance Structures:** Standards bodies have formal governance structures, with charters specifying the scope of different activities, defined leadership roles, review and approval stages, appeals processes, voting rules, consensus procedures, and so on.

These structures define how work is initiated, scoped, reviewed, approved, and contested.

At the same time, influence is also exercised through reputation, recognised expertise, community norms, procedural familiarity, and control over agendas, drafting, or review capacity. Governance structures therefore shape how decisions are made, how priorities are established, how disagreements are managed, and, ultimately, how influence is distributed within standards development.

* **Standards Implementation and Deployment:** Implementation usually occurs outside the formal standards process, and may be voluntary by interested parties or mandated by policy in certain jurisdictions.

  In many cases, publication of a standard does not by itself require implementation. Adoption may therefore vary widely: some standards are widely deployed, while others see limited or no implementation. Adoption may also be shaped by factors outside the standards process, including regulation, procurement, cost, and compatibility with existing systems.

  Data on implementation and operational use is often limited.

Measuring SDO activity is challenging. Observable metrics such as publication counts, message volume, attendance figures, authorship, or leadership roles can provide useful evidence, but each captures only part of the standards process. Analysis of artefacts and logs from the collaboration infrastructure (e.g., analysis of mailing list messages) can provide more detail and nuance, at the expense of additional complexity, but even these do not provide a complete view.

There are several reasons for this. One is that critical aspects of standards development are hard to observe directly. Influence, agenda setting, informal coordination, negotiation, and the practical exercise of power and authority may not be well represented by any single metric, or group of metrics, and are extremely challenging to infer from collaboration infrastructure logs.

Another reason is that the available data is often limited. Data availability and quality vary across SDOs. Different parts of the process are not equally observable, and even within a single SDO some information may be incomplete, difficult to access, inconsistently structured, or unavailable.

Combining multiple data sources introduces additional challenges. Observations from different parts of the process may not share stable identifiers, identifiers may change over time, and the same entity may appear in different forms across records. Voluntary declarations, non-standard terminology, and organisational changes such as mergers or acquisitions may further complicate linkage.

Metrics, artefacts, and other data sources may also differ in accuracy, representativeness, and relevance. Not all artefacts have the same significance, not all forms of participation have the same effect, and visible activity does not necessarily correspond to implementation, adoption, or wider impact. Measures should therefore be interpreted cautiously and, where possible, considered alongside complementary indicators.

## 3. Analysing the IETF

IETF participation is open with no formal membership. Individuals can participate by joining mailing lists, contributing to discussions, submitting Internet-Drafts, and attending meetings. Contributions ordinarily reflect the opinion of individual participants, and not necessarily their affiliation; exceptions to this norm exist for specific aspects such as draft authorship and intellectual property rights disclosures.

The IETF has a hierarchical group structure, with technical working groups (that have working group chairs) organised into distinct areas (that have area directors).

Reflecting its open participation model, much of the IETF's processes are publicly observable through open records and dedicated APIs. Mailing lists are a central forum for working group discussion, alongside meetings; some groups also use externally hosted repositories, for example on GitHub, to support drafting and issue discussion.

### 3.1. Datatracker

The IETF Datatracker (https://datatracker.ietf.org/) is the main source of day-to-day and historical data about the operation of the IETF. It can be accessed via the website or programmatically using a REST API and provides information about:

* Participants including names, email addresses, pronouns, biography, and photo, and external resources such as personal websites, GitHub usernames, and Orcid identifiers. The Datatracker maintains a record of the different names and email addresses used by individuals.

* Artefacts such as RFCs, Internet-drafts, agendas, blue sheets, working group charters, conflict reviews, shepherd write-ups, liaison statements, minutes, and presentation slides, including:

  - Metadata such as the title, name ("draft-ietf-..."), revision, date, state, and where appropriate abstract, working group, RFC number and publication stream, status on the standards track, area director, and document shepherd.

- Submissions (e.g., different revisions of internet-drafts) with document name, revision, date, title, abstract, authors, group, and metadata about documents the submission replaces.

- Authors with email address, affiliation, and country.

- Events such as state changes state, expiration, details of IESG processing, IETF last call, directorate reviews, IANA reviews, etc., with the document name, revision, date, and responsible person.

- Relationships including normative and informative references, and document replaced, updated, or obsoleted.

* Working groups, research groups, area, directorates, and leadership bodies such as the IESG, IRSG, and IAB, including the group name and acronym, group state, relationships between groups (e.g., working groups are organised in areas), the mailing list, charter text, milestones, and who occupies key roles in the group.

* IESG processing, including ballot positions, the text of comments and discusses, and scheduling of the IESG review.

* Directorate membership and directorate reviews, including the document, reviewer, outcome, data, and the review text.

* Meetings, including both plenary and interim meetings, with venues, dates, and times, details of what groups met in what time slots, and registration and attendance data.

* IPR disclosures including the document that the IPR relates to, the person making disclosure, details of the patent, and licensing terms.

The Datatracker has been developed over time, and this is reflected in the data that is available, with more recent data being significantly more complete than earlier data. Datatracker profiles are only required for a subset of IETF activities (e.g., draft submission, meeting registration), and so a number of active participants do not have a profile.

### 3.2. RFC Editor

The RFC Editor makes the RFC index available in machine readable form at https://www.rfc-editor.org/rfc-index.xml. The RFC index includes title, authors, publication date, status, abstract, publication stream, name of the precursor Internet-Draft, and the IETF area and working group that developed the RFC, if appropriate. This information is also available in the IETF Datatracker.

Information about RFC errata is available on the RFC Editor website at https://www.rfc-editor.org/errata.php. This data is also available in machine readable form.

### 3.3. Mailing List Archives

The IETF maintains public mail archives at https://mailarchive.ietf.org/ that are also available in machine readable form via IMAP from imap.ietf.org. The recent mail archives are essentially complete, but some historical lists that were not originally hosted on ietf.org are missing. Spam emails have largely, but not entirely, been removed from the archive. As of March 2026, the IETF mail archive contains approximately 3 million messages from almost 1400 mailing lists, around 40GB of data, with some messages dating back to the late 1980s.

The are significant data quality problems with older messages in the IETF mail archive, due to problems with the original messages rather than the archive, that make them difficult to process.

### 3.4. Session Recordings

The IETF makes video recordings of its plenary meetings available on YouTube (https://www.youtube.com/user/ietf). Audio recordings of IETF sessions from IETF 49 through to IETF 106 are available at https://get.ietf.org/archive/audio.

### 3.5. Chat Archives

The IETF makes chat logs available. Jabber was used prior to 2021, with archives at https://get.ietf.org/archive/jabber/. More recently, Zulip has been used accessible at zulip.ietf.org.

### 3.6. GitHub

Some IETF working groups, and some individuals, make extensive use of GitHub for document development and issue tracking. The IETF does not maintain a complete list of GitHub repositories associated with its work. The IETF Datatracker contains links to some repositories and user profiles.

Using the GitHub API, the following information is available:

* Information about GitHub users that contribute (e.g., username, email address, and other biography information).

* Contributions and changes, by way of Git commits, made by those users to documents.

* Discussion that takes place through comments and issues.

## 4. Analysing Other SDOs

Standards relevant to the Internet and the Web are also developed within the W3C, 3GPP, ITU-T, and others. Each organisation has its own governance model, participation structure, institutional culture,

and data availability. These differences affect both what can be observed, and how observations should be interpreted.

### 4.1. Data Availability Across SDOs

SDOs vary considerably in terms of the data that they make publicly available about their activities, and in how easily that data can be accessed and processed.

The W3C provides a REST API at https://api.w3.org, covering metadata about documents, participants, affiliations, and groups, and maintains a public mailing list archive. W3C groups make extensive use of GitHub for document development and issue tracking. The W3C operates under a membership model, in which participation is primarily through affiliated organisations. This affects how data about participants and their contributions should be interpreted, particularly when being compared to data from the IETF and other SDOs with individual participation models.

The ITU-T and 3GPP both operate under membership-based models where access to documents, meeting records, and contribution data is typically restricted to member organisations. Some ITU-T Recommendations are made publicly available after publication, while the 3GPP makes its specifications available at https://www.3gpp.org/specifications. The working documents, contributions, and meeting records are generally not accessible to non-members.

Differences in data availability mean that the methods applicable to the IETF, where rich longitudinal data is publicly available, may not be replicable across all SDOs. Any analyses should account for these availability differences.

### 4.2. Integrating Data Across SDOs

Efforts to understand the wider standardisation landscape requires combining data across multiple SDOs.

SDOs do not share common identifiers for participants, organisations, documents, or other metadata. An individual that participates across multiple SDOs may appear under different names, e-mail addresses, or usernames in the records of each SDO. Resolving these identifies requires suitable entity resolution mechanisms, and the risk of both incorrect matches (where two unrelated entities are linked together) and missed matches (where one entity has multiple, separate records in each SDO). The same risks apply to affiliations: companies may be recorded under different names, abbreviations, or subsidiary identities across SDOs.

Standards developed within one organisation may reference, build upon, or be coordinated with work at another SDO, but these relationships are not captured in any shared record. Reconstructing these relationships requires either manual effort, or natural

language processing of document content, introducing the risk of errors.

SDOs operate on different timescales and with different process structures. Comparing activity across organisations at a given point in time may not reflect equivalent stages of development.

Finally, differences in governance and participation models affect which comparisons are meaningful. Data analyses, and the interpretation of them, must consider that apparent differences between SDOs may reflect structural factors (e.g., open vs. membership-based participation) rather than substantive differences in behaviour or outcomes.

## 5. Data Processing

Significant processing effort is required to clean, normalise, and link data records before they can be processed.

The same individual participant may appear across each of the data sources with different identifiers, including names, e-mail addresses, usernames. These identifiers may change over time. Entity resolution (using exact and heuristic matching) is feasible in many instances, but requires careful validation to prevent the introduction of errors into later analyses. Entity resolution across organisations is similarly challenging, where companies may be subsidiaries of another, might merge or be acquired, or, given the unstructured nature of the dataset, appear under different names (to illustrate the scope of the entity resolution problem note that, as of May 2026, there are 282 variants of the name "Huawei" in the IETF Datatracker). Information external to the Datatracker, and other data sources, is often needed to process organisational data.

Participants may have more than one affiliation, including across the lifetime of a particular contribution (e.g., an Internet-Draft). Affiliation data is only recorded for a subset of activities, and may need to be inferred (e.g., from corporate domain names) in other cases. As a result, affiliation data, where recorded, indicates the participant's affiliation at moment in time for a particular contribution, making it difficult to form a continuous history.

Document life cycles are non-linear, and documents might pass through multiple working groups, by replaced or updated by later drafts, and change authorship over time.

Working group leadership is difficult to reconstruct: knowing who chaired a working group during a particular period, or which area a given group belonged to at a given time, requires the reconstruction of a timeline from historical event records held in the Datatracker. These records can be incomplete or inconsistently formatted.

E-mail metadata and message content presents a number of challenges. A significant number of messages contain malformed or archaic header

fields that break widely used email processing libraries and need correction. Mail clients perform the threading of messages in different ways, with the separation between new and quoted text becoming unclear. Natural language processing of message content requires contextualisation, with informal conventions, technical vocabulary, and the use of acronyms (all of which may evolve over time) presenting challenges that are unique to the dataset.

As noted, the quality of the dataset degrades significantly for historical records. Data that was not gathered by the Datatracker at the time, or that has been subject to partial backfilling later, must be treated with caution, both in terms of data processing and later analyses.

## 6. Ethics and Data Protection

Data is made available by the IETF, and other Internet SDOs, subject to their particular privacy and data protection policies and terms of use. For the IETF, these are described at https://www.ietf.org/privacy-statement/; other SDOs will have their own policies.

The available data includes considerable amounts of personal data that is potentially sensitive and subject to legal restrictions on processing and use in many jurisdictions (e.g., the GDPR in Europe). Researchers must ensure that their use of such data conforms to any applicable regulations. It is important to note that the regulations that apply to research use of such data may differ from those that apply to the IETF, or other SDOs, with regards to their use of the data as part of the standards process.

Researchers must ensure that their research, in particular research that involves personal data from the IETF or other SDOs, is conducted ethically and with respect for persons, in careful consideration of the risks and benefits of the work, taking care to ensure that those who bear the risk also gain some benefit, and with respect for the law and public interest. Researchers should consult with their organisation's Institutional Review Board, Research Ethics Committee, or similar, prior to conducting research that might raise ethical concerns, and are referred to the guidance in the Menlo Report [MENLO], the Belmont Report [BELMONT], and the ACM Policy on Research Involving Human Participants and Subjects [ACM] for further discussion of issues around ethical conduct of research.

Researchers are reminded that while data may be public, the implications of that data are not always well-known. For example, data that can be collected from the IETF Datatracker makes it possible to derive measures of the effectiveness of individuals in certain roles that, if presented out of context, might be considered sensitive. It is inappropriate to publish data about specific individuals without their explicit consent.

Finally, we note that researchers must take care to avoid disruption to the Internet standards process. In part, this requires that they

consult with the operations staff in the IETF LLC, or other SDOs, to ensure their data access does not cause operational difficulties (e.g., overload of servers that might disrupt an ongoing meeting). More broadly, researchers should ensure that any results that might be considered sensitive or disruptive are responsibly disclosed to the affected parties prior to publication. The effective operation of the Internet standards process directly affects critical global infrastructure, and researchers should be mindful of this when presenting results.

## 7. Recommendations

Analysis of standards development data is useful to support transparency and provide insight into the health, structure, and evolution of standards ecosystems, including patterns of participation, collaboration, concentration, and the development of technologies. It can inform discussions within SDOs and provide indicators of how technical work progresses over time. It can also inform broader Internet governance questions, such as how decision-making is structured, how participation is distributed, and the extent of centralisation in these processes [RFC9518], and can be useful to external stakeholders, including regulators, policy makers, and civil society, seeking to understand how standards are developed and governed.

Analysis of standards development is constrained by what can be observed. Important aspects of the process such as informal discussion, trust, institutional memory, cultural norms, and the exercise of influence may be only partially visible. In addition, the available data is often incomplete, inconsistently structured, and shaped by changes in tools and processes over time, with historical records in particular being sparse or unreliable.

As a result, analyses based on these data provide only a partial view of the process. Quantitative metrics such as message volume, authorship, participation counts, or leadership roles can be useful indicators, but do not directly capture influence, authority, or impact. They should therefore be interpreted with care and in context, rather than in isolation.

Where data is derived or reconstructed (e.g., via entity resolution, affiliation inference, or automated extraction) it is important to retain a clear link to the original sources. The provenance of such transformations should be documented, and derived data should be distinguishable from primary records. This allows results to be checked and, where necessary, corrected.

SDOs can support analysis of their processes by ensuring that the data they produce remains consistent, well-structured, and accessible over time. This includes maintaining clear, timestamped documentation of artefacts and processes, recording changes and their implications, and using consistent data formats and identifiers. Providing structured access to data, for example through stable and well-

documented APIs can be especially helpful. When introducing changes to tools, processes, or working practises, it is important to consider how these affect what is recorded and how it can be analysed. Where changes introduce discontinuities these should be clearly documented, including their scope and implications, so that their impact on the data can be understood and accounted for in subsequent analysis.

Comparisons across standards development organisations require particular care. Differences in governance, participation models, and transparency affect both what is observable and how it should be interpreted. Apparent differences between organisations may reflect these structural factors rather than substantive differences in behaviour or outcomes.

Finally, although much of the data used in this type of analysis is publicly available, its use still raises ethical questions. Analyses can have implications for individuals and organisations, especially if results are presented without sufficient context. Researchers should take care in how findings are reported, particularly where they relate to identifiable participants.

### 7.1. Recommendations for the IETF

* **Preserving a centralised and stable data access:** The Datatracker provides a central interface for structured data about IETF activity. Maintaining this role, including stable identifiers, consistent schemas, and well-documented APIs, supports reproducible and longitudinal analysis. Where data is maintained across multiple systems, stable references to authoritative sources help ensure consistency and integration.

* **Data quality and consistency:** The data reflects changes in tools and practices over time, which can make it harder to interpret, especially for older records. Common data such as events, roles, group metadata, and document states may be inconsistent across time. Where possible, these differences should be made consistent or clearly documented.

* **Historical data and backfilling:** Historical data may be incomplete. Where records can be reconstructed with confidence, backfilling can improve coverage. Backfilled data should be clearly identified, and its provenance documented.

* **Provenance of derived data:** Where data is derived from primary sources (e.g., extraction from archival material), the relationship between source and derived data should be explicit. Original artefacts should be retained where possible, and derived records clearly distinguished to allow validation and correction.

* **Error reporting and correction:** Datasets will contain errors, particularly in historical or reconstructed records. Providing a

transparent mechanism for reporting and correcting errors, along with maintaining a record of changes, improves reliability.

* **Separation of primary and inferred data:** Some data useful for analysis (e.g., identity resolution, affiliation inference) involves interpretation. Such data should be distinguishable from primary records, with clear documentation of how it was produced.

  - **TODO:** is this done by the IETF, by the researchers, or both?

* **Impact of process and tooling changes:** Changes to tools and working practises affect what is recorded and how it can be analysed. Where such changes introduce differences in data structure or coverage (e.g., adoption of different collaboration platforms), these should be documented clearly, including their scope and implications, to preserve comparability across groups and over time.

### 7.2. Recommendations for Researchers

Analysis of standards development data requires careful handling of both the data and its interpretation. The following practises can improve the robustness and reproducibility of such work:

* **Care in Datatracker use:** When using the Datatracker, it is preferable to download a local snapshot of the data, while respecting any access limits, and perform analysis on that copy. This avoids repeated queries to the live API.

* **Use versioned data snapshots:** The underlying datasets evolve over time. Analyses should be based on well-defined snapshots rather than live data, so that results can be reproduced and compared.

* **Document data processing steps:** Significant processing is often required before analysis, including cleaning, normalisation, and entity resolution. These steps can materially affect results and should be clearly documented, including any assumptions or heuristics used.

* **Handle identity and affiliation data with care:** Participants may appear under multiple identifiers, and affiliations may be incomplete, ambiguous, or change over time. Methods used to resolve identities or infer affiliations should be validated where possible and treated as approximations.

* **Account for incomplete and inconsistent data:** Not all aspects of the standards process are equally observable, and available data may be incomplete or inconsistent, particularly for historical records. Analyses should account for these limitations and avoid over-interpreting gaps or trends.

* **Be cautious in interpreting metrics:** Common metrics such as message volume, authorship, or participation counts do not

directly capture influence, authority, or impact. Results should be interpreted in context and, where possible, supported by complementary evidence.

* **Consider the impact of tooling and process changes:** Changes in tools or working practises (e.g., use of different collaboration platforms) can affect what is recorded and how it is structured. These changes should be considered when interpreting longitudinal trends or comparing across groups.

* **Engage with the community:** Data alone provides an incomplete view of the standards process. Engagement with participants or domain experts can help interpret results and identify factors that are not visible in the data.

* **Support reproducibility and reuse:** Where possible, researchers should share datasets, code, and methods, subject to applicable policies and privacy considerations. This reduces duplication of effort and improves the reliability of results.

* **Contribute improvements where appropriate:** Effort spent cleaning or structuring data may be of broader value. Where feasible, contributing corrections or improvements back to shared data sources can benefit the wider community.

* **Consider ethical implications:** As discussed in the Ethics and Data Protection section, analysis may have implications for individuals or organisations. Care should be taken in how results are presented, particularly where they may be sensitive or open to misinterpretation.

## 8. Security Considerations

Research into the operation of the Internet standards development ecosystem does not directly affect the security of the Internet. Effective operation of the Internet standards process is, however, critical to the security of the network, and researchers studying the development of Internet standards must consider potential security implications of their results and ensure that any such implications are responsibly disclosed to the relevant SDO. Examples might include, but are not limited to, research that discovers attempts to subvert or disrupt the operation of the standards process.

## 9. IANA Considerations

This document has no IANA actions.

## 10. Informative References

**[ACM]** ACM Publications Board, "ACM Publications Policy on Research Involving Human Participants and Subjects", n.d., <https://www.acm.org/publications/policies/research-involving-human-participants-and-subjects>.

**[BELMONT]** National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research, "The Belmont Report - Ethical Principles and Guidelines for the Protection of Human Subjects of Research", n.d., <https://www.hhs.gov/ohrp/regulations-and-policy/belmont-report/>.

**[MENLO]** US Department of Homeland Security Science and Technology Directorate, "The Menlo Report - Ethical Principles Guiding Information and Communication Technology Research", August 2012, <https://www.dhs.gov/sites/default/files/publications/CSD-MenloPrinciplesCORE-20120803_1.pdf>.

**[RFC9518]** M. Nottingham, "Centralization, Decentralization, and Internet Standards", December 2023, <https://datatracker.ietf.org/doc/html/rfc9518>.

## Acknowledgments

This document builds on work funded, in part, by the UK Engineering and Physical Sciences Research Council under grants EP/S033564/1 and EP/S036075/1.

## Authors' Addresses

Colin Perkins
University of Glasgow

Email: csp@csperkins.org

Ignacio Castro
Queen Mary University of London

Email: i.castro@qmul.ac.uk

Ryo Yanagida
University of St Andrews

Email: ryo@htonl.net

Stephen McQuistin
University of St Andrews

Email: sjm55@st-andrews.ac.uk