OARC Privacy Committee Survey Result Report

OARC 36 2021, 29 November 2021

OARC and Data for Research and Analysis

- OARC encourages the **collection of various data sets**
  - **Data is made available to its members** for research and analysis
  - **Enabling research and analysis contributes directly to the core functions of OARC**
- Data Sharing Agreement (DSA) is an Appendix to Participation Agreement

Milestones for Privacy Committee include

- Produce report on current data **storage** and data **usage**
- **Conduct surveys** of the membership to inform decisions about direction

OARC data storage and usage survey

- Stocktaking of contribution and usage of data by OARC members
  - Link to survey sent to members mailing list
  - Survey ran: Tuesday, 9 to Monday, 22 November 2021
  - 2 part survey:

    16 responses total

    - **Part 1: organisations that contribute data** — **5 responses (at 21 orgs do contrib)**
    - **Part 2: individuals using data** — **11 responses (??)**

OARC data storage and usage survey

- Stocktaking of contribution and usage of data by OARC members
  - Link to survey sent to members mailing list
  - Survey ran: Tuesday, 9 to Monday, 22 November 2021
  - 2 part survey:

    16 responses total

    - **Part 1: organisations that contribute data** — **5 responses (at 21 orgs do contrib)**
    - **Part 2: individuals using data** — **11 responses (??)**

One problem is we don't know how many **people use data** - is this most of them??

Part 1: Data Contribution

Data Contribution: **Orgs that DO contribute (5)**

(Background: 21 orgs contributed 2021 DITL data, small sample)

1. **2/5** orgs modify data - anonymising host portion of network address

Data Contribution: **Orgs that DO contribute (5)**

(Background: 21 orgs contributed 2021 DITL data, small sample)

1. **2/5** orgs modify data - anonymising host portion of network address
2. Major concerns with current/future DSA
    - Compliance with data protection **(2/5)**
    - Data is underused **(2/5)**
    - Future use of cloud storage **(1/5)**

Data Contribution: **Orgs that DO contribute (5)**

(Background: 21 orgs contributed 2021 DITL data, small sample)

1. **2/5** orgs modify data - anonymising host portion of network address
2. Major concerns with current/future DSA
   - Compliance with data protection **(2/5)**
   - Data is underused **(2/5)**
   - Future use of cloud storage **(1/5)**
3. What changes to DSA would encourage future contribution?
   - Options on **where to store** data (OARC/cloud) **(2/5)**
   - Options on **how data is used** (temp cloud/anon) **(2/5)**
   - All said would continue to contribute with or without these changes

Data Contribution: **Orgs that DO contribute (5)**

(Background: 21 orgs contributed 2021 DITL data, small sample)

1. **2/5** orgs modify data - anonymising host portion of network address
2. Major concerns with current/future DSA
   - Compliance with data protection **(2/5)**
   - Data is underused **(2/5)**
   - Future use of cloud storage **(1/5)**
3. What changes to DSA would encourage future contribution?
   - Options on **where to store** data (OARC/cloud) **(2/5)**
   - Options on **how data is used** (temp cloud/anon) **(2/5)**
   - All said would continue to contribute with or without these changes

LIMITED CONCLUSION: Increase flexibility likely useful

Data Contribution: **Orgs that DO NOT contribute**

1. Major concerns with current/future DSA (5 responses)
   - Compliance with data protection **(3/5)**
   - Data is underused **(2/5)**
   - Data mis-use/exposure/correlation **(2/5)**
   - Future cloud storage of data **(0/5)**

2. What changes to DSA would encourage future contribution? (7 responses)
   - Options on where to store data (OARC/cloud) **(3/7)**
   - Options on how data is used (temp cloud/anon) **(5/7)**
   - Neither **(2/7)**

Data Contribution: **Orgs that DO NOT contribute**

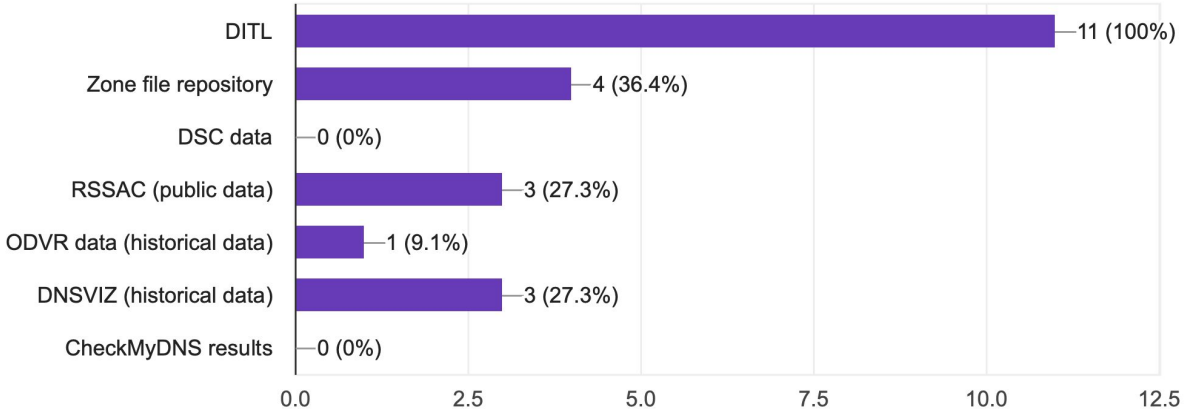
1. Major concerns with current/future DSA (5 responses)
   - Compliance with data protection **(3/5)**
   - Data is underused **(2/5)**
   - Data mis-use/exposure/correlation **(2/5)**
   - Future cloud storage of data **(0/5)**

2. What changes to DSA would encourage future contribution? (7 responses)
   - Options on where to store data (OARC/cloud) **(3/7)**
   - Options on how data is used (temp cloud/anon) **(5/7)**
   - Neither **(2/7)**

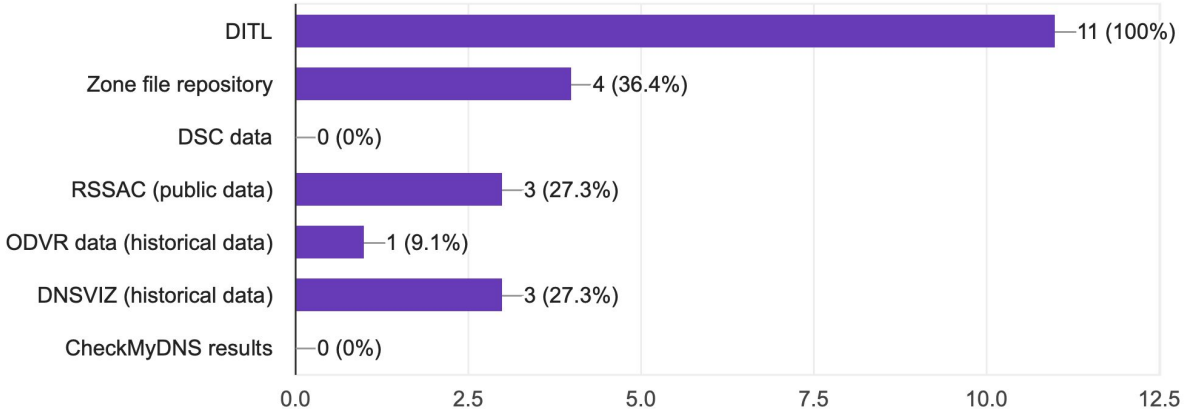
LIMITED CONCLUSION: Data compliance an issue, again increased flexibility to increase data usage likely useful

Part 2: Data Access and Usage

Data Access: **Individuals that DO access data (11)**

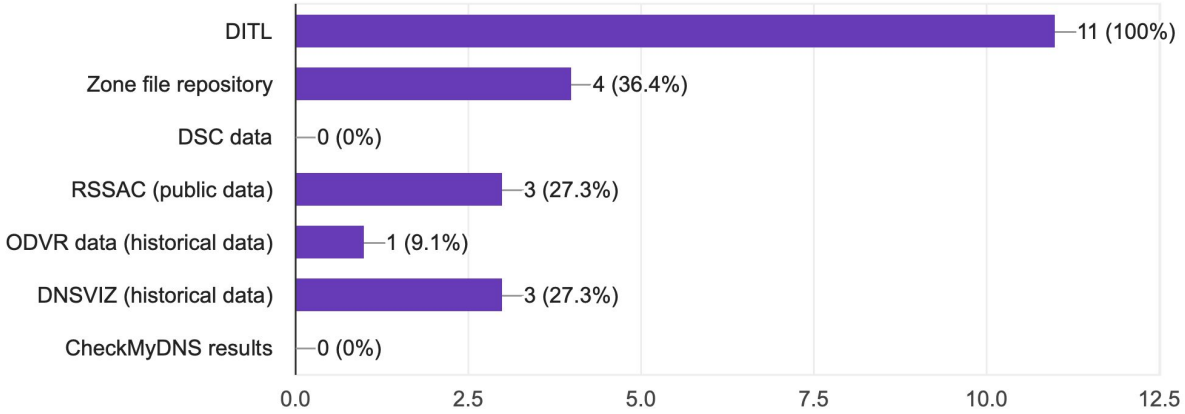


Data Access: **Individuals that DO access data (11)**



1. Mostly, data accessed on a monthly **(4/11)** to 1-2 times a year **(4/11)**
2. Data older than 1 year accessed reasonably regularly **(10/11)**
3. Data is ‘derived’ from OARC raw data reasonably regularly **(9/11)**
4. Derived data published/shared **(7/11)**
5. OARC server rented for analysis **(2/11)**

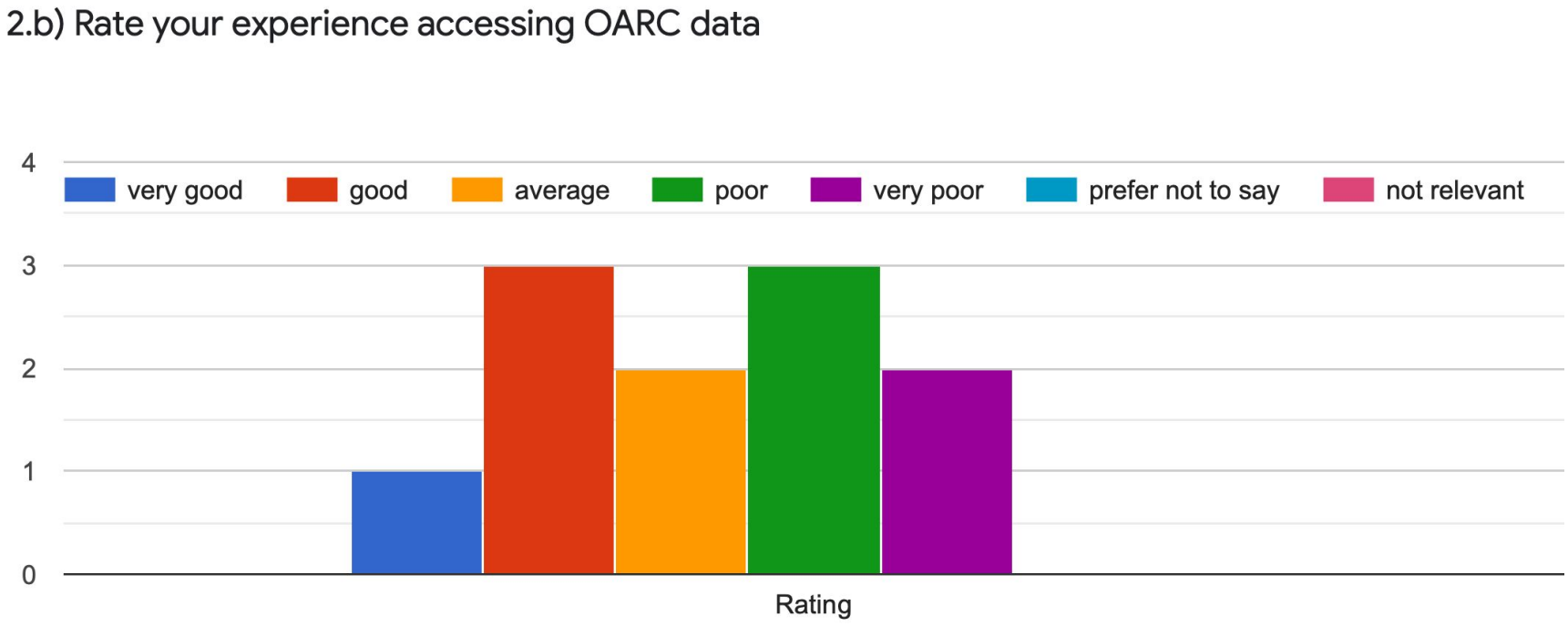
Data Access: **Individuals that DO access data (11)**



1. Mostly, data accessed on a monthly **(4/11)** to 1-2 times a year **(4/11)**
2. Data older than 1 year accessed reasonably regularly **(10/11)**
3. Data is 'derived' from OARC raw data reasonably regularly **(9/11)**
4. Derived data published/shared **(7/11)**
5. OARC server rented for analysis **(2/11)**

CONCLUSION: DITL data is valuable and is used by those that do access it

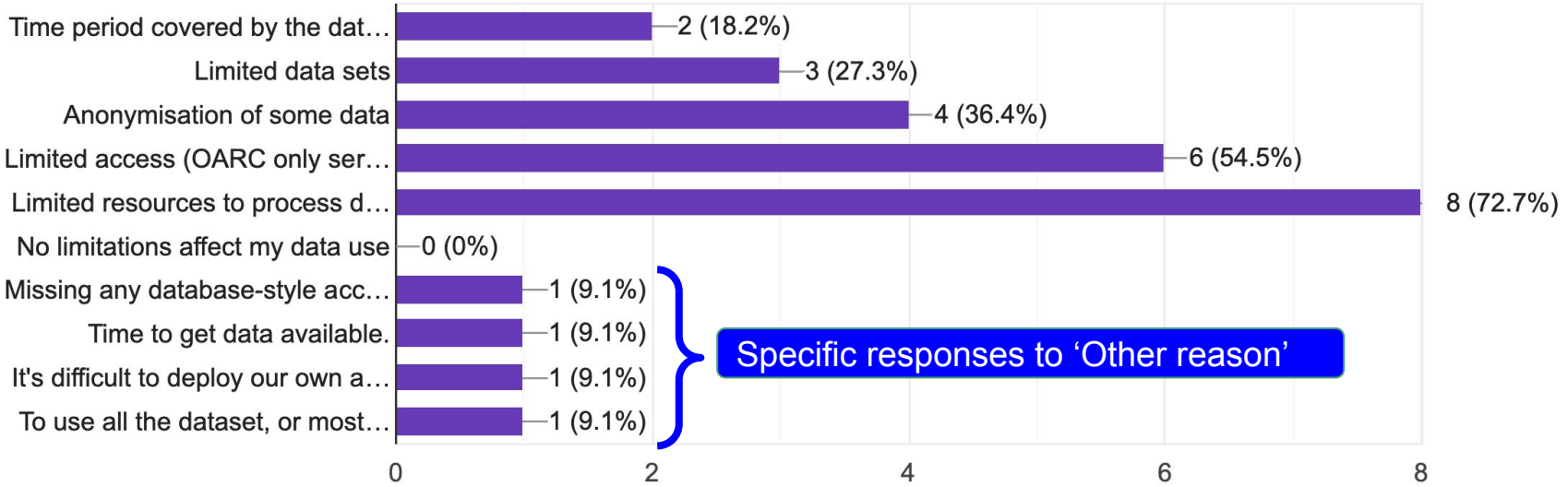
Data Access: **Individuals that DO access data (11)**



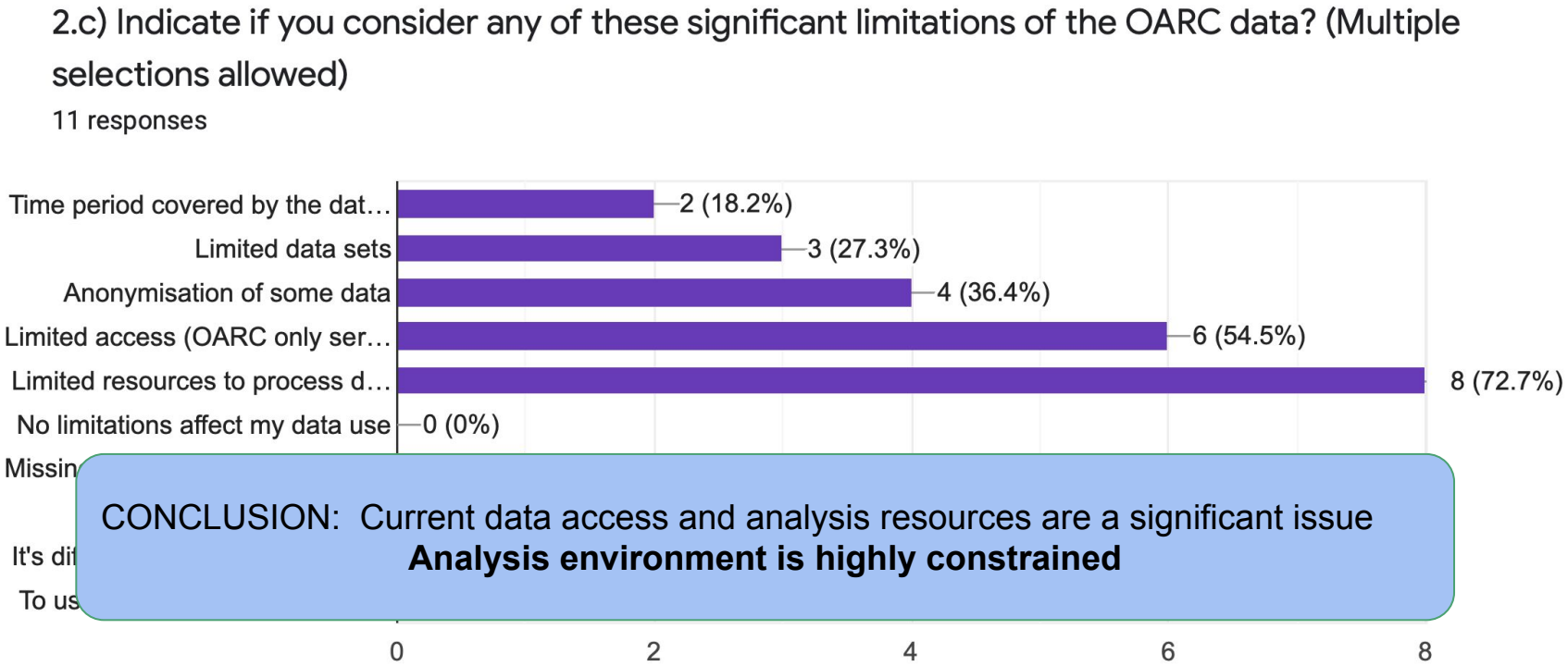
Data Access: **Individuals that DO access data (11)**

2.c) Indicate if you consider any of these significant limitations of the OARC data? (Multiple selections allowed)

11 responses



Data Access: **Individuals that DO access data (11)**



CONCLUSION: Current data access and analysis resources are a significant issue
Analysis environment is highly constrained

Data Access: **Individuals that DO access data (11)**

1. Would **full anonymisation** of IP addresses be a blocker to any of your analysis? YES: **(9/11)**
    - ASN, Prefix and Geolocation are heavily used in analysis
    - If no consistent mapping in one data set, source ID impossible
    - If no consistent mapping between data sets, correlation impossible

2. Would **pseudo anonymisation** of IP addresses be a blocker to any of your analysis? YES: **(6/11)**
    - Accept some privacy may be needed
    - If limited to /24 and maps 1:1 from real to anon IP across all the data sets, it would be mostly usable

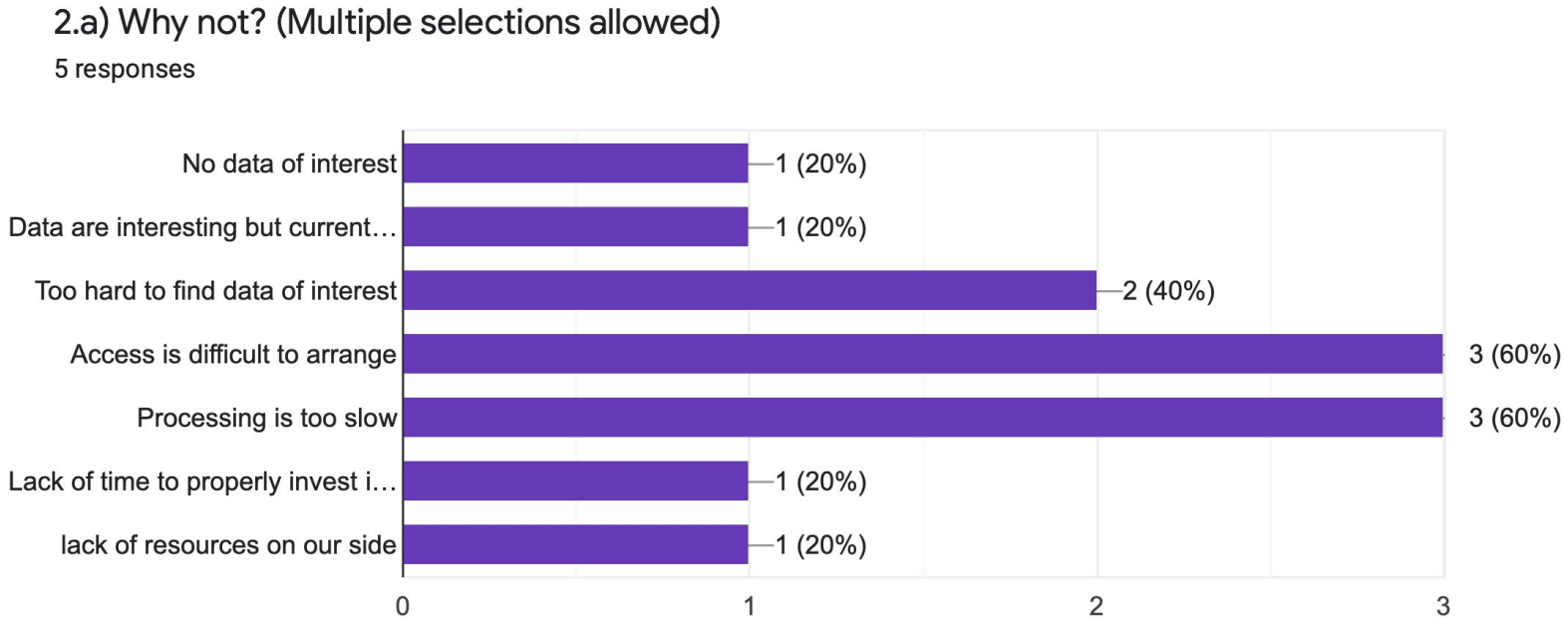
Data Access: **Individuals that DO access data (11)**

1. Would **full anonymisation** of IP addresses be a blocker to any of your analysis? YES: **(9/11)**
    - ASN, Prefix and Geolocation are heavily used in analysis
    - If no consistent mapping in one data set, source ID impossible
    - If no consistent mapping between data sets, correlation impossible

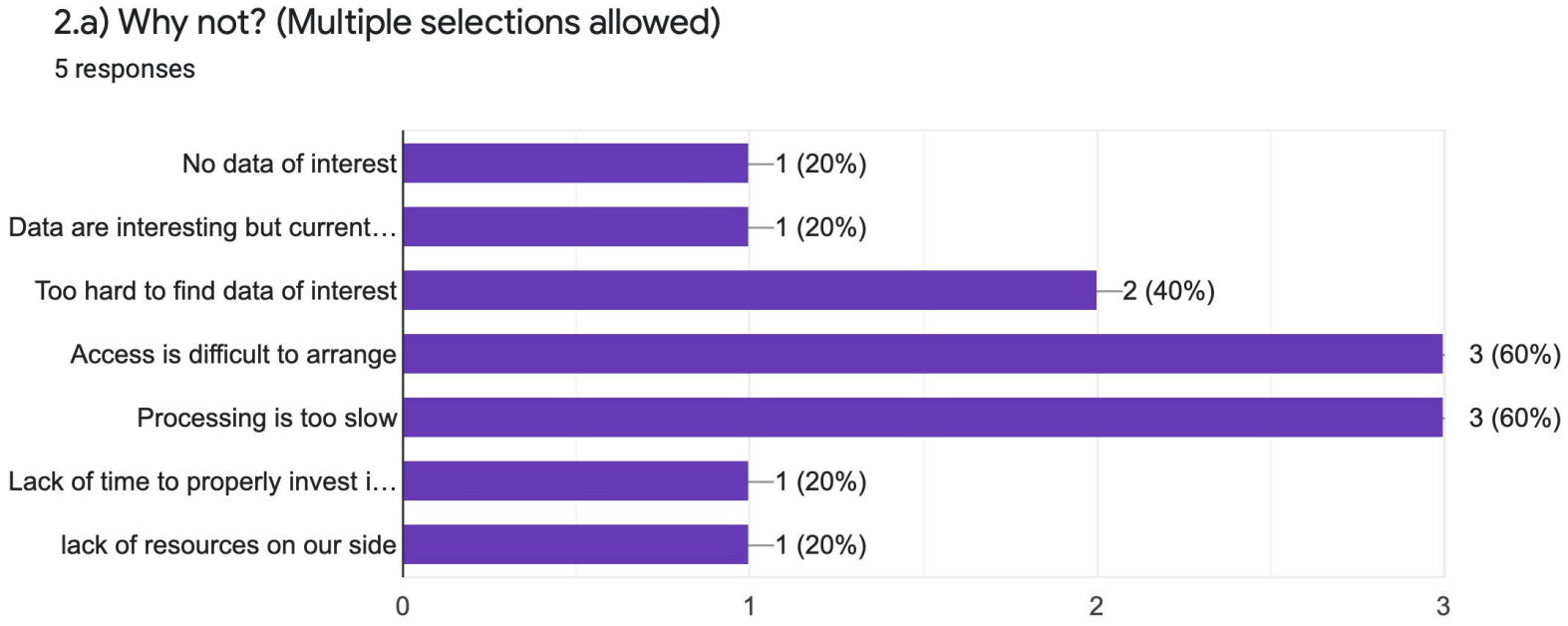
2. Would **pseudo anonymisation** of IP addresses be a blocker to any of your analysis? YES: **(6/11)**
    - Accept some privacy may be needed
    - If limited to /24 and maps 1:1 from real to anon IP across all the data sets, it would be mostly usable

CONCLUSION: A form of pseudo anonymisation could be applied to address privacy concerns

Data Access: **Individuals that DO NOT access data (5)**



Data Access: **Individuals that DO NOT access data (5)**



CONCLUSION: Again, data access and processing resources are an issue

Key conclusions

- **Data contribution** sample is small
  - whilst data compliance is an issue,
  - no strong objections to increasing the flexibility of storage or data access if it increases data usage

- **Data access and usage** - hard to know sample size but
  - Respondents see value in data (particularly DITL and zone data)
  - But... feel limited by current access model, processing resources and lack of data catalogue
  - Pseudo anonymisation might be usable, but with care and will still hamper some analysis

Next steps - goals are more data and more analysis!

- Gathre more responses?
  - Re-run survey to gather more data?
  - Missed question asking respondents to identify themselves! Please contact us if willing!!
- How to improve data contribution?
  - Direct outreach to 21 orgs that contribute data - open to change of DSA
- How to improve data usage?
  - Create a data catalogue and increase awareness of data available to academia/etc.
  - Create a stronger user community - share code and tools and understand numbers of users better
  - Actively review options to ease barriers to data usage in constrained environment
- Generate report to stimulate community discussion

Questions?