New Data Types in Data Management and Archiving

September 15th 2022

DOI: 10.5281/zenodo.7078856

cessda.eu

@CESSDA_Data

Outline

- Introduction, Brian Kleiner
- “Current capacities among CESSDA SPs for handling new data types”, Martin Vávra (CSDA)
- “Proposals for a More Coordinated Approach to Handling New Data Types Across CESSDA SPs”, Brian Kleiner (FORS)
- “Social Media Data Sharing in Social Research”, Yevhen Voronin (GESIS)
- “Social Science in the Embattled Digital Age: Adversarial Creation, Use and Sharing of New Data Types”, Pascal Jurgens (Johannes Gutenberg University Mainz)
- Break: 10 minutes
- The panel discussion

Current capacities among CESSDA service providers for handling new data types

Presenting: Martin Vávra, CSDA
martin.vavra@soc.cas.cz

15 September 2022, New Data Types in Data Management and Archiving

cessda.eu @CESSDA_Data

The presentation is based on work of CESSDA ERIC Agenda 21-24, Tasks 21-22 Widening & Outreach Pillar: Task 2 Survey on Researchers Needs and Widening the Perimeter of Data

Members of the team: Brian Kleiner, Dimitra Kondyli, Nikolas Klironomos, Libby Bishop, Tomas Čížek, Yana Leontiyeva, Martin Vávra

What are new data types?

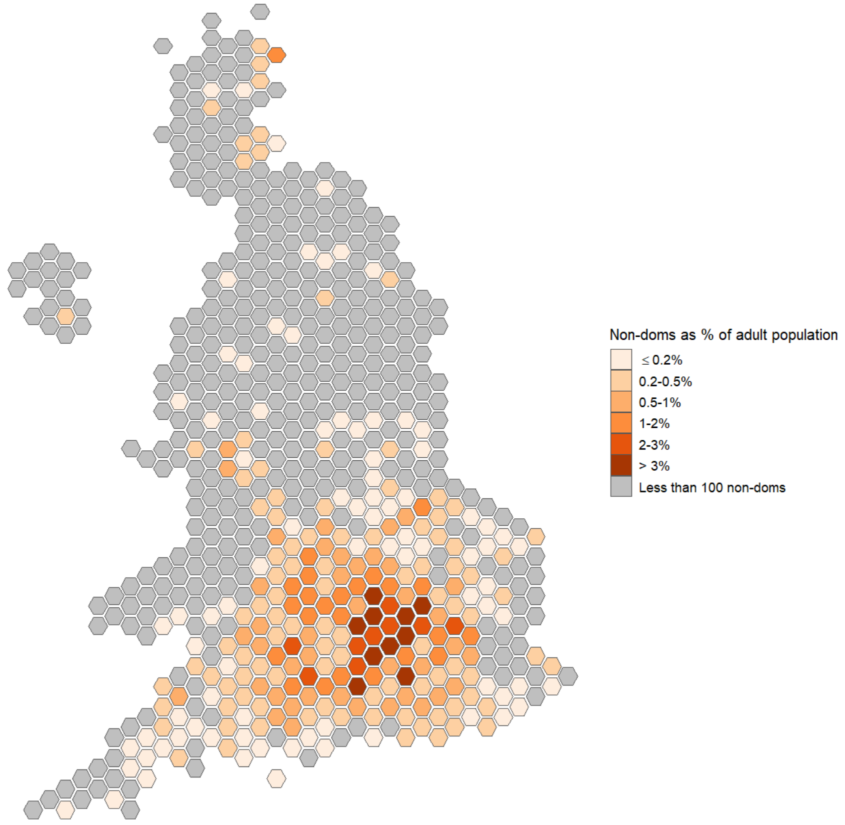
New data types can be defined by opposition to „traditional“ data types in social sciences. Those consist of data collected or created for research purposes (typically survey data, interviews). Existence of „new data types“ datasets as a resource for social sciences is a frequently side effect of public administration functions (e.g. tax records), commercial transactions (e.g. customer cards data), or internet-based activities (e.g. social media data).

„The rise of the internet and the mass digitization of administrative records and historical archives have unleashed an unprecedented amount of digital data in recent years. Unlike conventional datasets collected by social scientists, these new digital sources often provide rich detail about the evolution of social relationships across large populations as they unfold “.

Edelmann A, Wolff T, Montagne D, Bail CA. 2020. Computational Social Scienc and Sociology

A variety of new techniques are now available to analyze (large, complex datasets) new types of data.

New data are already applied in research – among many: e.g. Advani, A., Burgherr, D., Savage, M., & Summers, A. (2022); The UK’s global economic elite: a sociological analysis using tax data (No. 114607); London School of Economics and Political Science, LSE Library.



Types of new data

Category A: Government transactions	Individual tax records, Corporate tax records, Property tax records, Social security payments, Import/export records
Category B: Government and other registration records	Housing and land use registers, Educational registers, Criminal justice registers, Social security registers, Electoral registers, Population registers, Health system registers, Vehicle/driver registers, Membership registers
Category C: Commercial transactions	Store cards, Customer accounts, Other customer records
Category D: Internet usage	Search terms, Website interactions, Downloads, Social media data, Blogs; news sites
Category E: Tracking data	CCTV images, Traffic sensors, Mobile phone locations; GPS data
Category F: Satellite and aerial imagery	Visible light spectrum; Night-time visible radiation, Infrared; radar mapping
Category G: Health data	MRIs, ultrasounds, neuroimaging data, “patients records”, CT scans, x Rays
Category H: Other data types	All new data types other than those mentioned above.

OECD report: New Data for Understanding the Human Condition (2013), “Health data” added

Archiving and sharing new data types. Why is it important?

Basically for the same reason as with traditional data – archiving and sharing data improve effectiveness of research efforts:

- reduces costs of research
- improve comparability and reproducibility of research
- make collaboration easier

Further information on importance, conditions and barriers for new data types archiving and sharing:
Bishop, Libby, 2020. New Data Types in Social Science Research and Data Archives.
https://doi.org/10.5281/zenodo.3924177
Alongside the presentation, also a video introducing these issues is available.

Methodological note – how we collected data

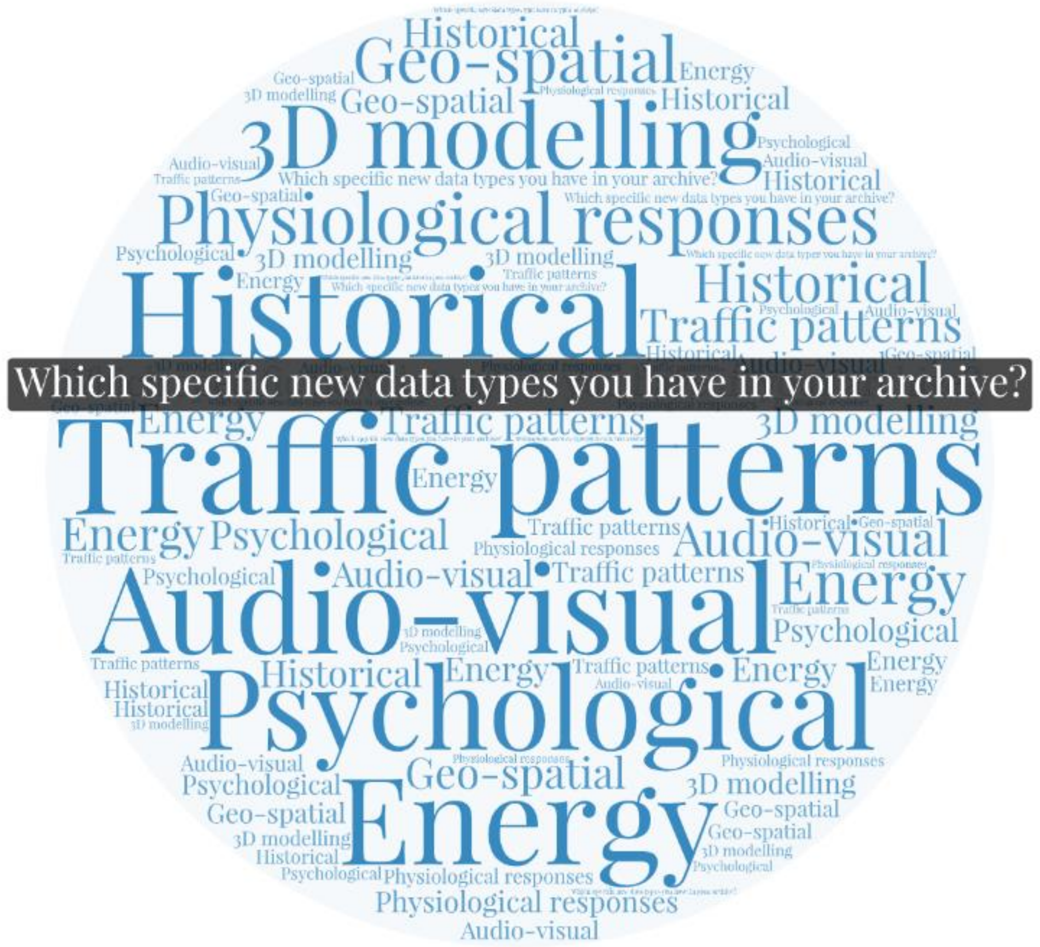
- Next slides are based on data from an online survey (very old data type) among European data repositories – both CESSDA members and CESSDA partner archives.
- Data were collected during April and May 2021.
- 24 completed questionnaires were received in total – 21 by CESSDA members (with only two member archives missing) and 3 by CESSDA partners.
- Questionnaire was semi-structured – closed questions for knowledge on new data types archived and problems faced/expected with archiving them, and open questions for more qualitative insights.

New data types in CESSDA archives

Type of the data	Number of archives keeping individual types of NTDs in collections
Category A: Government transactions	6
Category B: Government and other registration records	6
Category C: Commercial transactions	0
Category D: Internet usage	8
Category E: Tracking data	3
Category F: Satellite and aerial imagery	2
Category G: Health data	8
Category H: Other data types	8

N=24

Word clouds were created from corpuses consisting of text answers to respective open-ended questions.



Challenges related to archiving new data types

Type of data	How much individual types of data are perceived as „problematic‟ to be archived?
Category A: Government transactions	1,7
Category B: Government and other registration records	1,5
Category C: Commercial transactions	-
Category D: Internet usage	2,5
Category E: Tracking data	2,5
Category F: Satellite and aerial imagery	1,6
Category G: Health data	2,4

Mean value for N=24 archives. Mean was calculated from scale where 1 is "not at all" and 5 is "to a large extent"

Challenges related to archiving new data types

Type of issue	How much the issue is perceived as serious?
a. legal, ethical, and/or data protection issues	3
b. technical issues	2,7
c. available time and resources at the archive	2,4
d. available skills and know-how	2,4
e. archive's access to the data	1,7
f. cleaning the data	2,2
g. fitting to your existing metadata schema(s)	1,8

Mean value for N=24 archives. Mean was calculated from scale where 1 is "not at all" and 5 is "to a large extent"

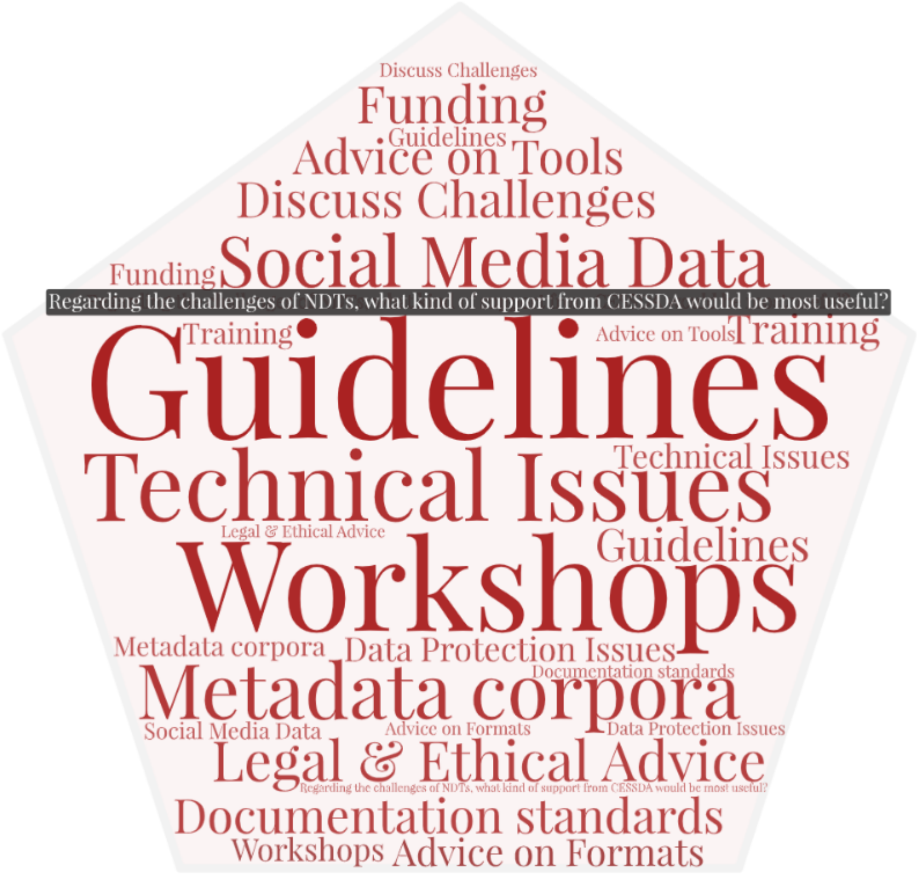
Most useful support from CESSDA regarding NDTs archiving

Channels of support:

- Guidelines and standards
- Training on data management
- Sharing best practices
- Online forum for knowledge exchange
- Coordinated CESSDA action

Areas of support:

- Standards and practices of archiving and disseminating data
- Technical solutions
- Legal and ethical issues



Ways to help SPs better handle NDTs

Most important ways according to SPs representatives are:

- Use cases or case studies for training and knowledge exchange
- Expert Seminars on the topic of NDTs
- Joint session with other research infrastructures or projects that have dealt with NDTs
- Sharing platforms to disseminate specific new data types (e.g., internet usage, tracking data, etc.) in order to create a common data management system
- Promoting the use and the availability of such NDTs within the CESSDA Data Catalogue, organising workshops on how to find/use such data
- Information for NDT users and producers (examples of good practice when using or working with NDT - basic requirements that need to be fulfilled, and which basic information to provide for archiving)
- Legal-licensing. Defining procedures for legal review have to be platform-specific.

Priorities for CESSDA to support SPs

- CESSDA is important in areas where its status as a European infrastructure could make a difference – e.g. in communication with social media platforms.
  - To archive new data, the primary challenge is to establish what each platform (e.g., Twitter) permits. Not all platforms have been willing to enable research access. However, it is plausible that when data reuse does not harm commercial interests, they might be open to discussion. A single archive is only one voice, but CESSDA, as a European legal entity representing many funders, data creators, and data users as stakeholders, could be in a far stronger position.

- Perceived need to organise a working group with representatives from SPs to prepare depositor guidance for the most commonly used data (Twitter, Facebook, web scraped data). Outputs would be short, easy guides and case studies – CESSDA can be active here.

Priorities for CESSDA to support SPs with NDTs.

Knowledge exchange
Guidelines
Guidelines
Events
User guides
Metadata Standards
Technical advice
Metadata Standards
Legal & ethical advice
Technical advice
Legal & ethical advice
User guides
Training sessions
Training sessions
Social Media Data
Publications
Knowledge exchange
Knowledge exchange
Data management
Publications
Technical advice
Social Media Data
Data management
Training sessions
Events
Legal & ethical advice

Conclusions

- There are some „new data" deposited in CESSDA archives but not many – especially in smaller archives. This is despite the presumed quantities of datasets consisting of new data types being produced - but is this really surprising?
- CESSDA data archives - **general lack of experience in working with NDT**s, with the exception of some of the larger and longer-established social science data archives. Without data there is no experience, but without experience archives can be hesitant to try to acquire new data types.
- Reasons behind the situation can be quite complicated - e.g. lack of concrete type of data produced on national level, or that other organizations are used for archiving e.g. administrative data.
- Might be the case that archives are dissuasive regarding NDTs, **communicating in one way or another to researchers that they are not ready to accept certain non-traditional data types**.

Conclusions

- **Challenges faced by archives regarding NDTs -** many archives are confronted with **legal, ethical, and data protection issues** for many NDTs. Many of the new data types also were challenging with regard to the existing technical systems and metadata schemas of the archives. In addition, many SPs are faced with important resource, know-how, and skill challenges and gaps.

- At the same time, **archives share the recognition that NDTs are important** and that help and mutual assistance are needed in order to handle this.

Sources

- Advani, A., Burgherr, D., Savage, M., & Summers, A. 2022. *The UK's global economic elite: a sociological analysis using tax data* (No. 114607); London School of Economics and Political Science, LSE Library.

- Bishop, Libby, 2020. „New Data Types in Social Science Research and Data Archives". https://doi.org/10.5281/zenodo.3924177

- Edelmann, Achim, Tom Wolff, Danielle Montagne, Christopher A. Bail. 2020. „Computational Social Science and Sociology". *Annual Review of Sociology* 46:1, 61-81

- Kleiner, Brian et al. 2021. „Overview and summary of existing outputs (inside and outside of CESSDA) on New Data Types". Report D14 for CESSDA ERIC

Proposals for a More Coordinated Approach to Handling New Data Types Across CESSDA Service Providers

Brian Kleiner, FORS - Swiss Centre of Expertise in the Social Sciences
brian.kleiner@fors.unil.ch

Other members of the team: Dimitra Kondyli, Nikolas Klironomos, Martin Vávra, Libby Bishop, Tomas Cizek, Yana Leontiyeva

15 September 2022, New Data Types in Data Management and Archiving

cessda.eu

@CESSDA_Data

Key questions for data archives

- When will the NDT deluge arrive to data archives?
- Will digital data archives be ready for it?
- Where are the main gaps in capacity and skills going forward?
- How can we pool our resources to prepare?
- What mechanisms can we put into place that will allow us to adapt and respond?

Possible forms of coordination

- Forums for exchange of expertise
- Shared written materials
- Impact

Forums for exchange of expertise

- Channels for sharing expertise across SPs
- Periodic workshops to identify the available expertise and needs within SPs
- Establish a division of labour, for example regarding expertise in formats, metadata, or access schemes
- “Open hours” format, where people can come to exchange and learn about particular relevant topics
- Consider “eclectic affinities” with other CESSDA activities and projects, to maximise synergies

Shared written materials

- SPs could combine forces to create various forms of guidance and support for archives dealing with NDTs. This could include, for example:
  - handbooks,
  - checklists,
  - case studies,
  - etc.
- Documents could be published on the CESSDA Resource Directory, and could be referred to in the CESSDA Data Archiving Guide.

Impact

- Establish a common conceptual framework for NDTs
- Monitor SP NDT activities, data holdings, and objectives
- Detail the impact of particular capacity-building activities
- “Scenario planning” and SWOT analyses, so that SPs can be sure to be “future proof” in dealing with NDTs.

Other forms of collaboration?

Going forward, together: enacting ideas

- General capacity-building activities versus “responsive” actions (i.e., of the type “we have a problem – how do we fix it?”)
- Create new lines of communication across archives
- Small-scale pilot collaborations
- Reliance on CESSDA Resource Directory for written materials, including possibly a special directory
- All forms of collaboration should be modulated according to available resources from CESSDA and within service providers
- Collaboration with researchers and stakeholders

Social Media Data Sharing in Social Research

Results from online survey among authors of the social sciences using social media data for their research

Presenter: Voronin, Yevhen / GESIS
Authors: Akdeniz, Esra; Borschewski, Kerrin; Breuer, Johannes; Voronin, Yevhen / GESIS

15 September 2022, New Data Types in Data Management and Archiving
DOI: 10.5281/zenodo.7078856

cessda.eu @CESSDA_Data

Introduction

Increase in SMD use

One type of data that has recently seen a particular increase in use in the social sciences is social media data (SMD).

Important data sources in the social sciences.

(Breuer et al., 2021; Ledford, 2020; van Atteveldt and Peng, 2018; Weller and Kinder-Kurlanda, 2015).

Gap in data access

One precondition for research with SMD is that researchers can access them: primary and secondary data.

Access to SMD is associated with new challenges; gap in data access.

(Breuer et al., 2020; Weller and Kinder-Kurlanda, 2015)

Benefits of sharing

Sharing SMD with other researchers to be reused beyond the original collection purpose can reduce and prevent the inequality gap in data access.

Barriers to sharing research data

resource-intensive

copyright

not enough credit

Terms of Service

lack of confidence and knowledge

informed consent

data protection laws

ethical challenges

lack of common standards

fear of getting scooped

revelation of errors

fear of misuse, misinterpretation

uncertainty about the value of the data

(Acord and Harley, 2013; Breuer et al., 2020; Breuer et al., 2021; Hemphill et al., 2021; Sloan et al., 2020; van Atteveldt et al., 2019; Weller and Kinder-Kurlanda, 2015; Zenk-Möltgen et al., 2018).

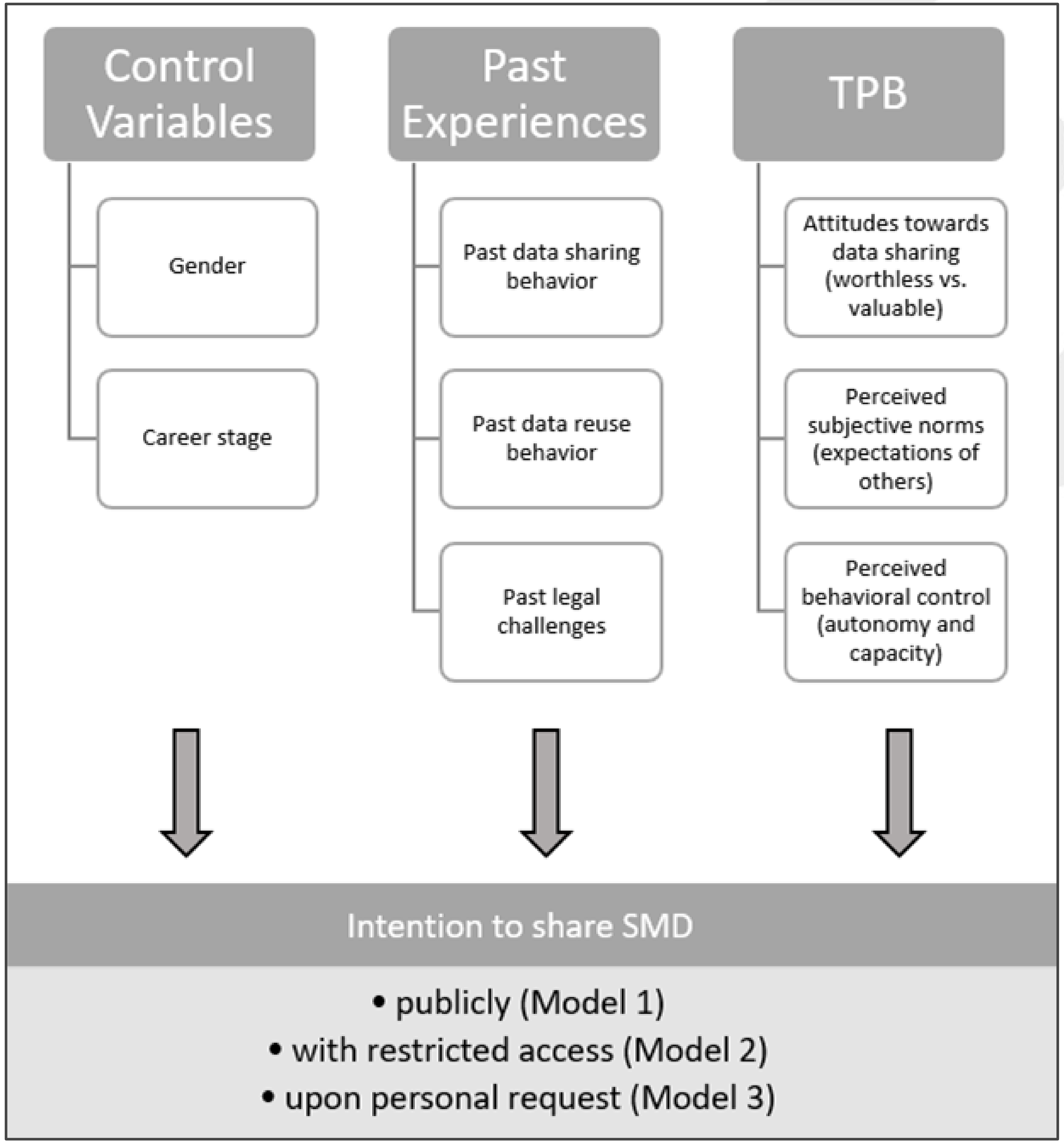
Theoretical model

Predicting intentions to share SMD.

Past experience:

- Data sharing as a two-way process;
- Working with SMD entails additional challenges.

- Theory of Planned Behaviour (TPB) (Ajzen, 1991)
  - Attitudes;
  - Subjective norms;
  - Behavioral control.


Control Variables
Gender
Career stage
Past Experiences
Past data sharing behavior
Past data reuse behavior
Past legal challenges
TPB
Attitudes towards data sharing (worthless vs. valuable)
Perceived subjective norms (expectations of others)
Perceived behavioral control (autonomy and capacity)
Intention to share SMD
• publicly (Model 1)
• with restricted access (Model 2)
• upon personal request (Model 3)

Data and methods (1)

Data from online survey among authors of the social sciences using SMD for their research and having published journal articles based on social media data between 2018 and 2021.

Version 1.1: https://doi.org/10.7802/2418

The questionnaire:

- a) data acquisition and use of secondary data,
- b) past data sharing behaviour,
- c) data sharing intentions,
- d) data documentation,
- e) use of other forms of data,
- f) personality and
- g) demography.

Theory of Planned Behavior (Icek Ajzen)

Data and methods (2)

1) Frequencies for closed-ended questions;
    i) Intentions to share SMD;
    ii) Used SMD platforms;
    iii) Used data acquisition methods;
    iv) Challenges for sharing SMD.

2) Frequencies for open-ended questions (ad hoc coding);
    i) Reasons to (not) share SMD.

3) Blockwise linear regression analysis per intentions to share data…
    i) Publicly,
    ii) Under controlled access,
    iii) Upon personal request.

Descriptives

Analytical sample = 249, *100%* (used SMD)

Female: 109 *(43.80%)*
Professor/Assistant professor/Associate professor: 127 *(51%)*

Used secondary data: 84 (*33.73%)*
Shared data: 94 *(37.75%)*
Used repositories or archives to share data: 44 *(17.67%)*

Intentions to share

Operationalization: TACT framework, 1-7 scale (extremely unlikely – extremely likely).

- "...Sharing (Action) SMD (Context) with others outside of your research team (Target) within the next 3 years (Time)..."

Three modes:

- publicly (with no restrictions),
- under controlled access (that regulates if and how data may be used by others),
- upon personal request (when being contacted directly by others).

	N	mean	sd
intention to share publicly	222*	**3.26**	2.06
intention to share under controlled access	222*	**4.32**	1.81
intention to share only upon personal request	222*	**4.64**	1.81

*Listwise deletion

Overview: Origins of SMD used for research

Twitter
82.33%

Facebook
67.07%

YouTube
42.57%

Instagram
36.95%

Reddit
26.10%

Wikipedia
12.85%

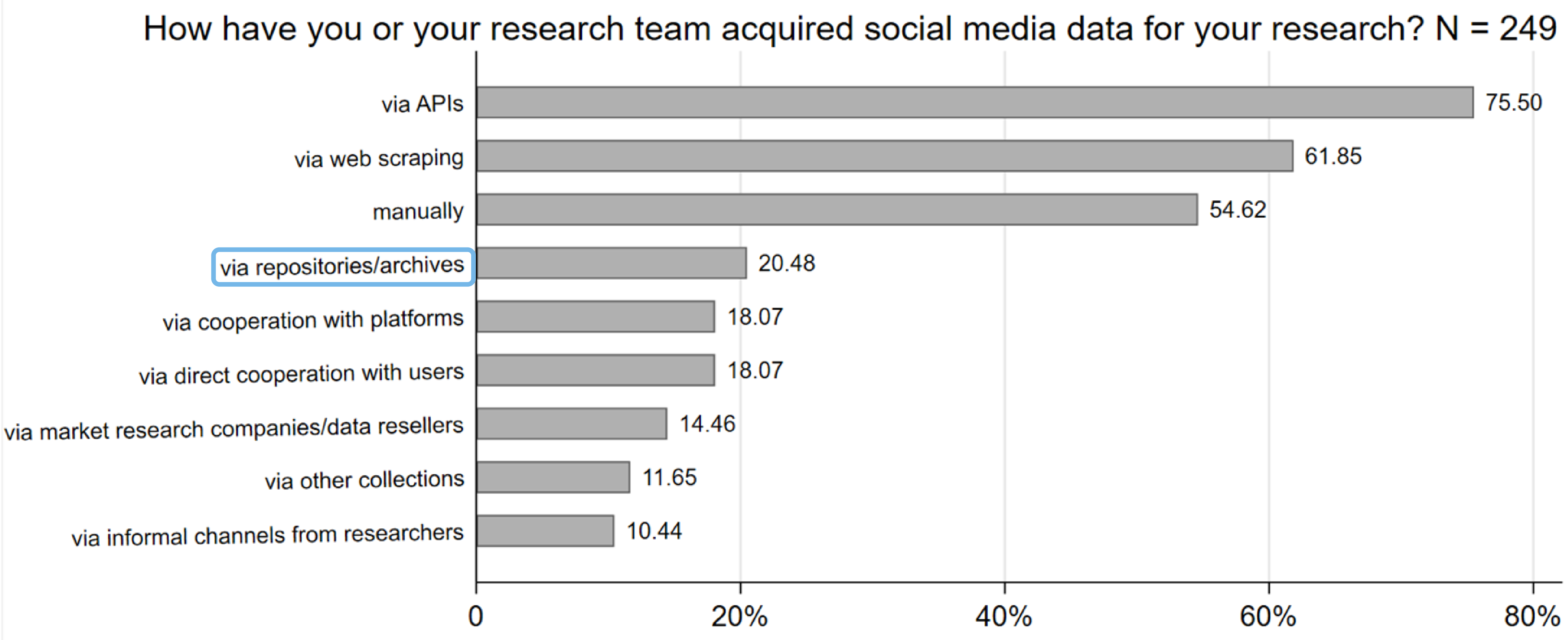
Telegram
12.05%

TikTok
8.84%

Other
24.50%

Images: Unsplash, Canva

Overview: Data acquisition methods used



Results: Shared

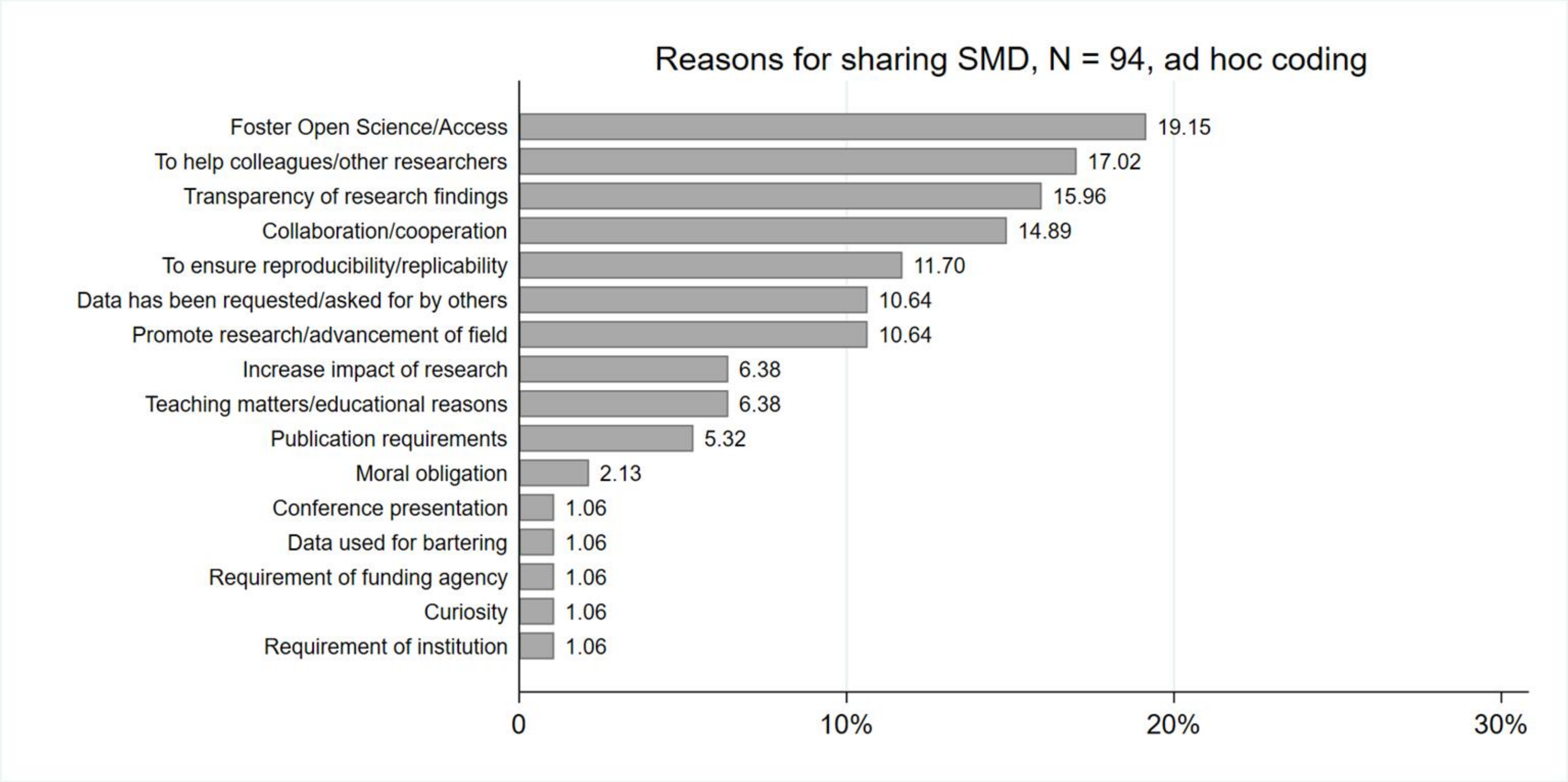


Image created with STATA

Results: Not shared

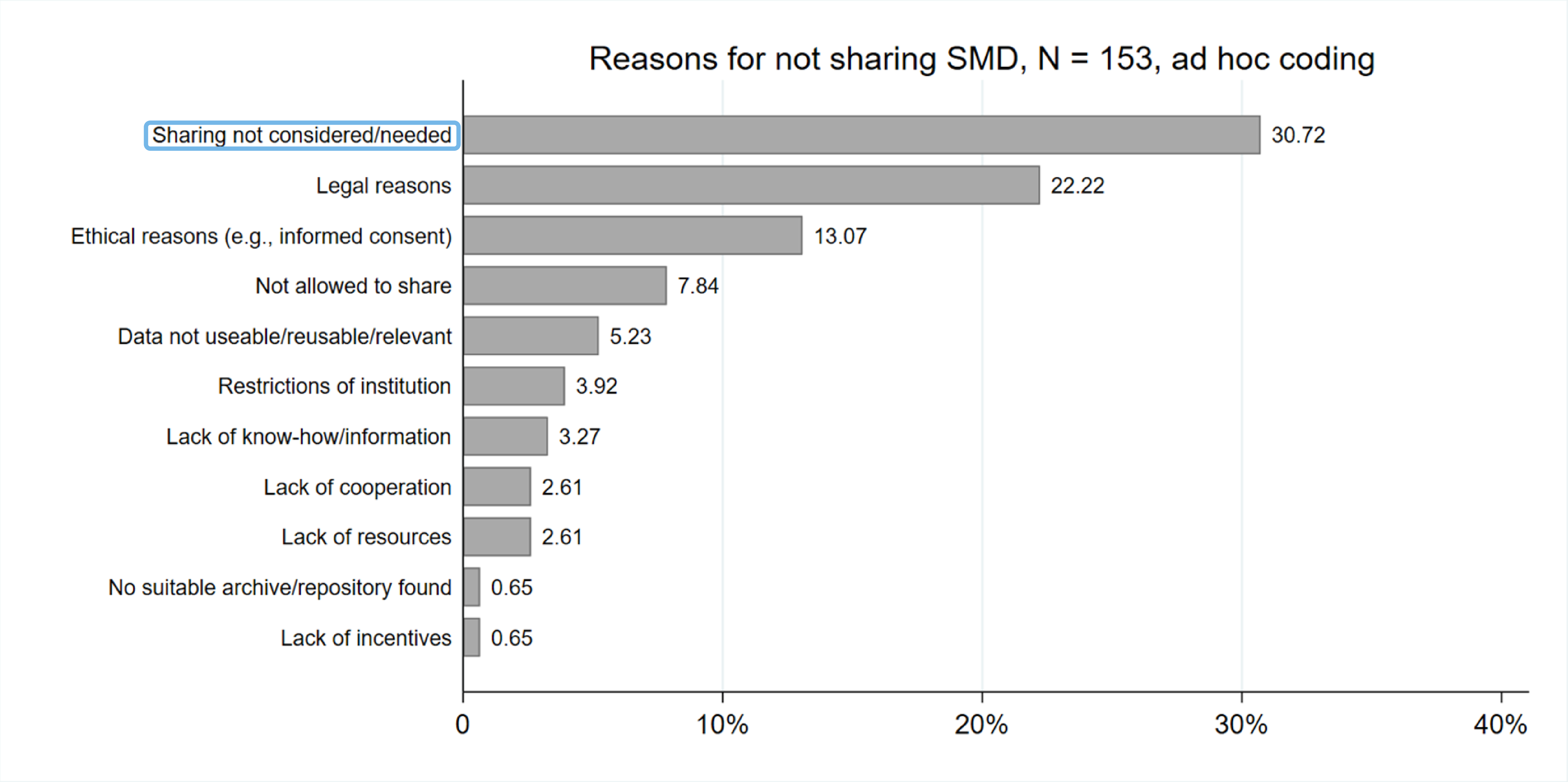


Image created with STATA

Results: Challenges

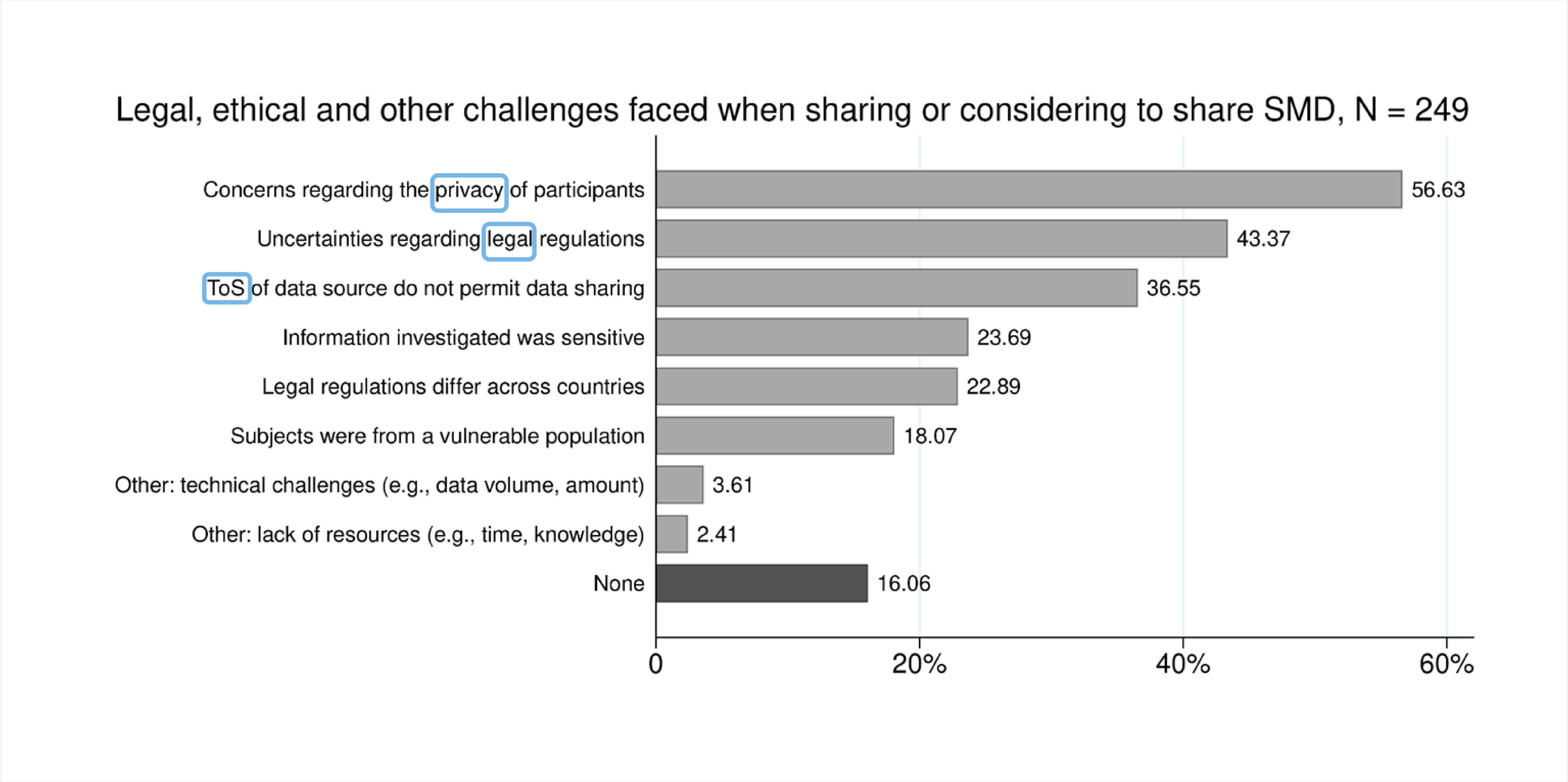
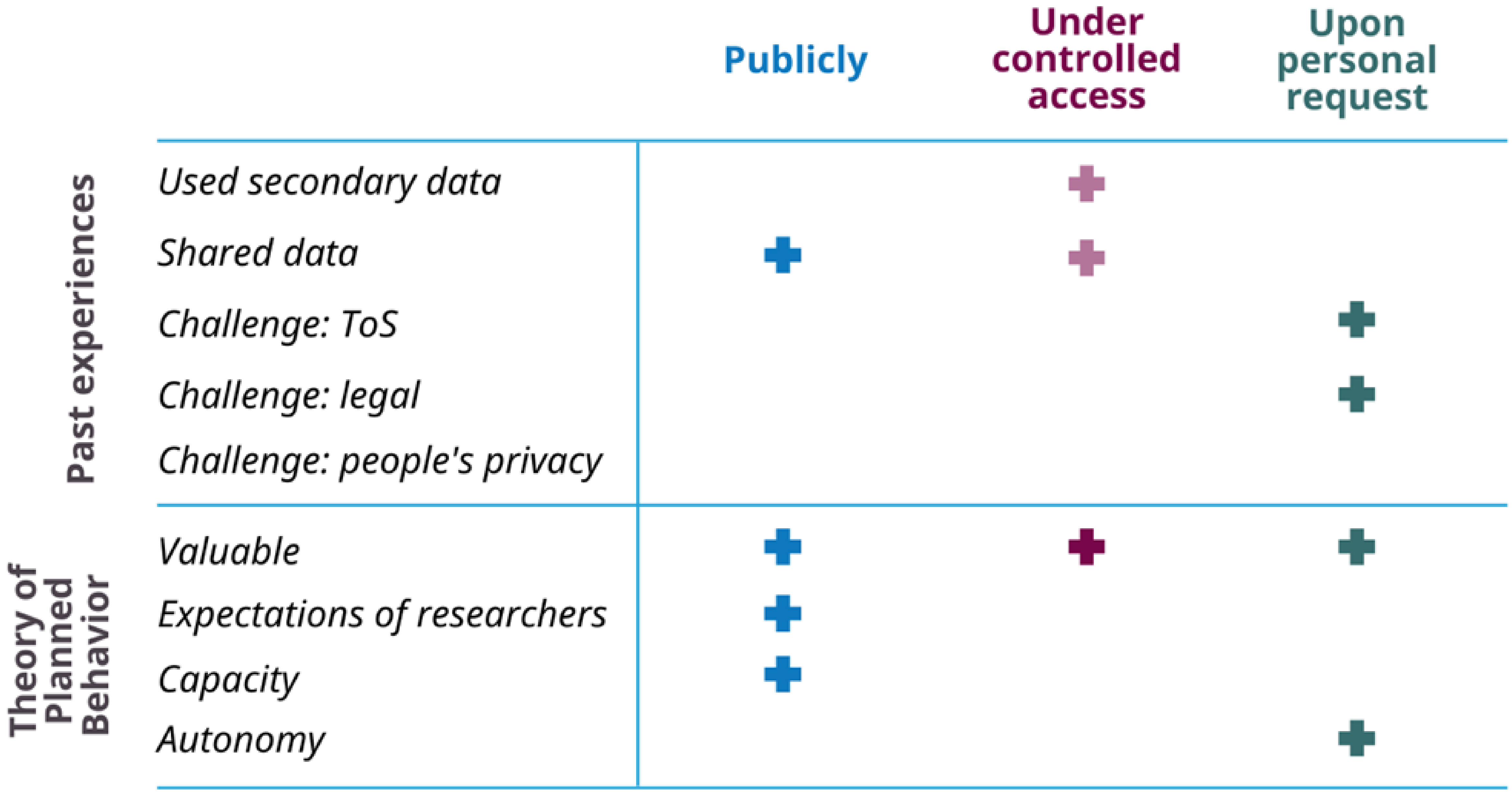


Image created with STATA

Results: Predicting intentions to share SMD



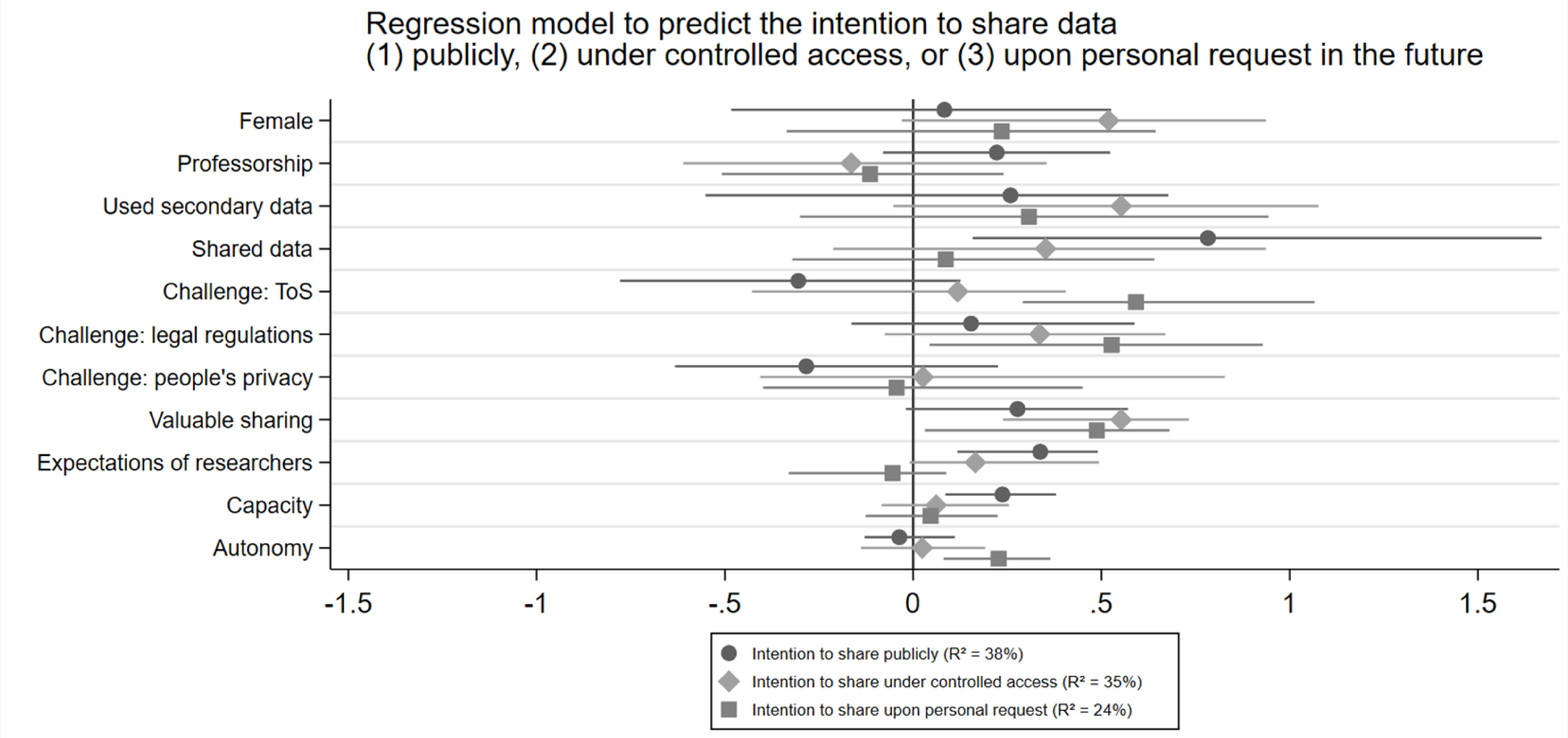
		Publicly	Under controlled access	Upon personal request
Past experiences	*Used secondary data*		+	
	Shared data	+	+	
	Challenge: ToS			+
	Challenge: legal			+
	Challenge: people's privacy			
Theory of Planned Behavior	*Valuable*	+	+	+
	Expectations of researchers	+		
	Capacity	+		
	Autonomy			+

Control variables: female, professorship. Blockwise linear regression models (one per data sharing type; 1st block - control; 2nd block - past experiences, 3rd block - TPB). Bootstrapped standard errors, listwise solution. N = 177.

"+" indicates a positive B coefficient in the second and/or third block with p <= 0.05.

Image created with Canva

Results: Predicting intentions to share SMD



Bootstrapped standard errors, listwise solution. N = 177

Image created with STATA

Discussion

The reasons for sharing SMD: idealistic/altruistic, self-serving, and compliance motives.

The reasons that prevent researchers from sharing SMD: legal and/or ethical challenges, lack of resources/available repositories/knowledge, lack of value, benefit, and usefulness.

Predicting intentions, both past experiences and TPB components (attitudes, subjective norms, and perceived behavioral control) play a role here. Depends on the mode of sharing.

References (i)

Acord, S. K., & Harley, D. (2013). Credit, time, and personality: The human challenges to sharing scholarly work using Web 2.0. New Media & Society, 15(3), 379–397. https://doi.org/10.1177/1461444812465140

Ajzen, I. (1991). The theory of planned behavior. Organizational Behavior and Human Decision Processes, 50(2), 179–211. https://doi.org/10.1016/0749-5978(91)90020-T

Akdeniz, E., Borschewski, K., Breuer, J., & Voronin, Y. (2022a). Survey—Accessing, (re)using, and sharing social media data in academia (1.1) [Data set]. GESIS Data Archive. https://doi.org/10.7802/2418

Akdeniz, E., Borschewski, K., Breuer, J., & Voronin, Y. (2022b). D11 Dataset from the survey on researchers' needs published. https://doi.org/10.5281/zenodo.5554492

Atteveldt, W. van, Strycharz, J., Trilling, D., & Welbers, K. (2019). Computational Communication Science| Toward Open Computational Communication Science: A Practical Road Map for Reusable Data and Code. International Journal of Communication, 13(0), 20.

Breuer, J., Al Baghal, T., Sloan, L., Bishop, L., Kondyli, D., & Linardis, A. (2021). Informed consent for linking survey and social media data—Differences between platforms and data types. IASSIST Quarterly, 45(1). https://doi.org/10.29173/iq988

Breuer, J., Bishop, L., & Kinder-Kurlanda, K. (2020). The practical and ethical challenges in acquiring and sharing digital trace data: Negotiating public-private partnerships. New Media & Society, 22(11), 2058–2080. https://doi.org/10.1177/1461444820924622

References (ii)

Breuer, J., Borschewski, K., Bishop, L., Vávra, M., Štebe, J., Strapcova, K., & Hegedűs, P. (2021). Archiving Social Media Data: A guide for archivists and researchers. https://doi.org/10.5281/zenodo.5041072

Hemphill, L., Hedstrom, M. L., & Leonard, S. H. (2021). Saving social media data: Understanding data management practices among social media researchers and their implications for archives. Journal of the Association for Information Science and Technology, 72(1), 97–109. https://doi.org/10.1002/asi.24368

Ledford, H. (2020). How Facebook, Twitter and other data troves are revolutionizing social science. Nature, 582(7812), 328–330. https://doi.org/10.1038/d41586-020-01747-1

Sloan, L., Jessop, C., Al Baghal, T., & Williams, M. (2020). Linking Survey and Twitter Data: Informed Consent, Disclosure, Security, and Archiving. Journal of Empirical Research on Human Research Ethics, 15(1–2), 63–76. https://doi.org/10.1177/1556264619853447

Weller, K., & Kinder-Kurlanda, K. (2015). Uncovering the Challenges in Collection, Sharing and Documentation: The Hidden Data of Social Media Research? Proceedings of the International AAAI Conference on Web and Social Media, 9(4), 28–37.

Zenk-Möltgen, W., Akdeniz, E., Katsanidou, A., Naßhoven, V., & Balaban, E. (2018). Factors influencing the data sharing behavior of researchers in sociology and political science. Journal of Documentation, 74(5), 1053–1073. https://doi.org/10.1108/JD-09-2017-0126

“Social Science in the Embattled Digital Age: Adversarial Creation, Use and Sharing of New Data Types”

Presenter: Pascal Jurgens / Johannes Gutenberg University Mainz

15 September 2022

cessda.eu

@CESSDA_Data

Key **Antagonistic Challenges** in **Data Collection**, **Archival** and **Sharing**

Challenge	Desideratum	Restriction	Antagonist
Data Collection	Document & understand societal and political reality	• terms and conditions • copyright • feasibility • soft and hard threats	• tech companies (good/bad faith) • copyright holders • totalitarian regimes
Data Archival	Create durable archives of societal and political reality	• volume of data • technical access & preservation	• funders • technical designers, implementers and standards bodies
Data Sharing	Enable transparency, reproducibility, replicability, facilitate research and boost coherence	• copyright • data protection (GDPR) • commercial TOS • individual liability	• copyright holders • Legislators • funding agencies • learned societies • universities / legal staff

Panel discussion

CESSDA Webinar: New Data Types in Data Management and Archiving

Thank you!

cessda.eu

@CESSDA_Data