Data quality in European law enforcement and border control cooperation: Findings from survey research

Report

Author(s):
Leese, Matthias; Marugg, Fintan

Publication date:
2023-06

Permanent link:
https://doi.org/10.3929/ethz-b-000609825

Originally published in:
CURATE Report 1

Data quality in European law enforcement and border control cooperation

Findings from survey research

CURATE REPORT NO. 1

Data quality in European law enforcement and border control cooperation

Findings from survey research

CURATE Report No. 1

June 2023

Matthias Leese[1] & Fintan Marugg[2]

Abstract

This report presents findings from survey research on data quality and related institutional arrangements and processes, carried out between January and February 2023 among 65 national-level law enforcement and border control organizations in the European Union/Schengen area. Overall, findings suggest growing awareness of data quality issues and increasing professionalization of data quality practices. There is, however, considerable heterogeneity in approaches to data quality, raising concerns in light of the aggregation of data in EU-level systems where they are used for transnational law enforcement and border control purposes. From a practical perspective, findings indicate the need to further professionalize data quality as an everyday activity in law enforcement and border control. From a policy perspective, findings imply the need for further reform and harmonization, both on the national and the supranational level. From a research perspective, findings underline the need for more systematic academic engagement with data quality.

Suggested citation:

Leese M and Marugg F (2023) Data Quality in European Law Enforcement and Border Control Cooperation: Findings from Survey Research. CURATE Report No. 1, June 2023. doi: 10.3929/ethz-b-000609825.

[1] ETH Zurich, mleese@ethz.ch

[2] ETH Zurich, fintan.marugg@gess.ethz.ch

Table of Contents

Executive Summary

This report presents findings from survey research on data quality and related institutional arrangements and processes, carried out between January and February 2023 among 65 national-level law enforcement and border control organizations in the European Union/Schengen area. The aim of the survey was to systematically capture of the current state of the affairs in regard to data quality in European law enforcement and border control cooperation.

Survey responses indicate a general tendency towards awareness of data quality challenges and subsequently some degree of professionalization in terms of the structures and processes that are relevant for data quality. They do, however, also show that there is still a considerable amount of variation and fragmentation. This can be seen as cause for concern in regard to information sharing at the supranational level.

Implications can be drawn in regard to three dimensions.

1. From a *practical perspective*, findings indicate the need to further professionalize data quality as an everyday activity in law enforcement and border control agencies. Dedicated data governance frameworks are necessary, including the specification of data quality requirements for different use cases, the definition of roles and responsibilities, as well as the allocation of sufficient resources for the continuous implementation of data quality processes. Data quality measures, as survey responses indicate, should in this context not be understood as an exclusively technical issue, but include the role of human activities and their specific characteristics.

2. From a *policy perspective*, findings tease out the need for further reform and harmonization, both on the national and the supranational level. The need for reliable and trustworthy data becomes aggravated when data are entered into EU databases and shared across national boundaries. Currently, two major political initiatives at the European level can be witnessed in this regard: the Data Quality Roadmap initiated by the Justice and Home Affairs Council as well as the Commission's Implementing Decisions 2021/2224 and 2021/2225 that specify minimum quality requirements for data that are entered into EU-level systems. Especially the Roadmap that strives to harmonize standards and practices regarding data quality at the national level across all EU/Schengen countries is pertinent in the context of survey results.

3. From a *research perspective*, findings underline the need for more academic engagement with data quality. Notably, there remains a shortage of systematic social scientific inquiries concerning data quality and its surrounding policies and practices. Survey responses suggest that data quality needs to be conceptualized and studied as from a holistic perspective that takes into account the organizational embeddedness of data and the socio-technical composition of data practices.

Introduction

This is the first report from the CURATE Project.[3] It presents findings from survey research on data quality among 65 national-level law enforcement and border control organizations in the European Union/Schengen Area.

Scope of the project

CURATE is a five-year research project that has been selected for funding by the European Research Council (ERC) and is financed by the Swiss State Secretariat for Education, Research and Innovation (SERI).[4] Its main goals are to understand and theorize data quality and corresponding practices, and to assist in improving data quality in European law enforcement and border control cooperation. Inspired by reports about questionable data quality in EU-level databases and corresponding inaccurate security interventions (Council of the European Union, 2020a, 2020b), the project team investigates how data that end up in EU-level systems for law enforcement and border control are produced, cleaned, and consolidated at several critical junctions (local, regional, national, supranational).

In doing so, CURATE focuses on the ways in which data quality is "done" in everyday practice, i.e. how law enforcement and border control agencies interact with their data in order to make and keep them reliable and trustworthy, as well as fit for different tasks for which they are analyzed and used.

Scope of the survey

There is so far no systematic knowledge base (academic or otherwise) regarding questions of data quality and data quality practices in law enforcement and border control. To create an overview of the current state of affairs, the survey was set up as an exploratory inquiry, resulting in a diagnosis of the status quo from a cross-national perspective. Specific attention was given to how participating organizations assess the quality of the data that they produce and handle, where potential error sources in datasets are located, which institutional arrangements and processes regarding data quality are in place, and whether change can be observed in relation to these questions.

In the context of CURATE, survey results will additionally serve as a mapping of existing institutional arrangements, processes, and larger tendencies. Findings will inform and guide in-depth case study research during the next phase of the project.

[3] https://curate-project.com

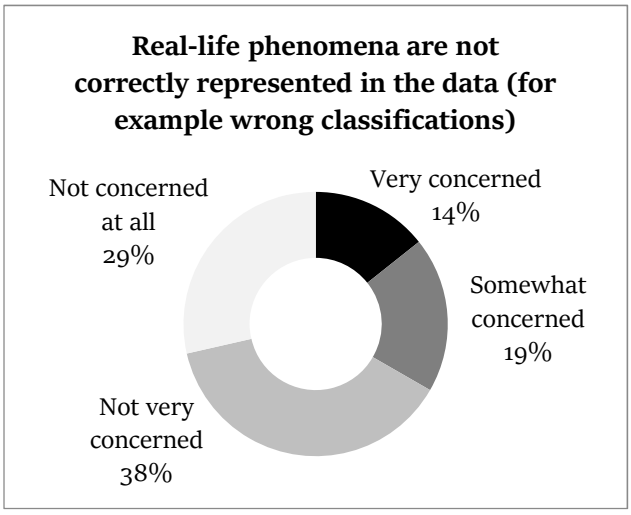
[4] Grant agreement MB22.00035

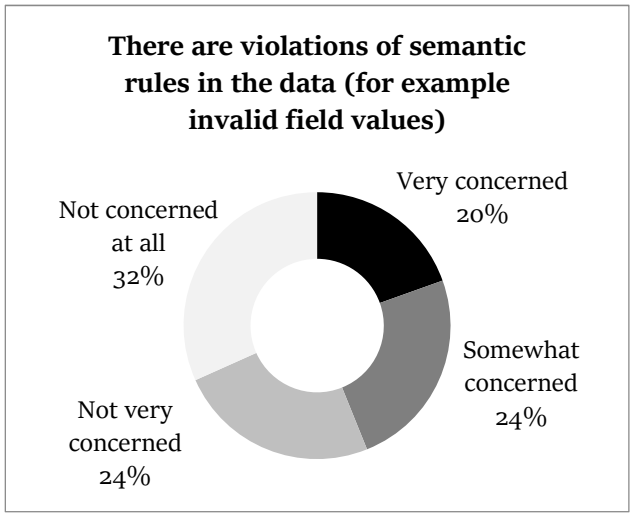
Findings

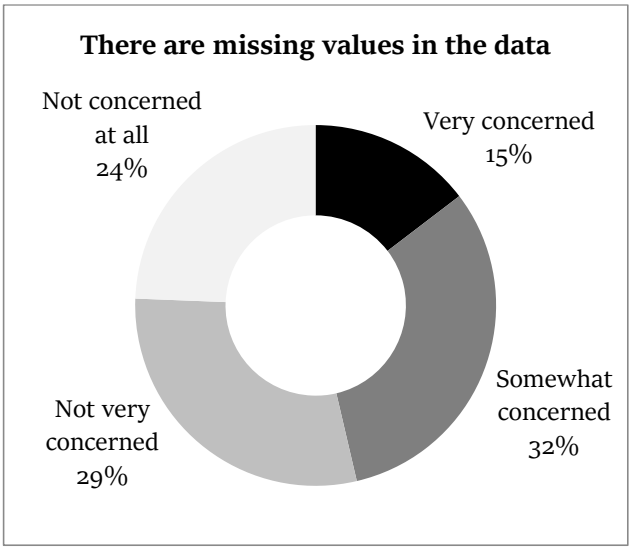
This section presents the findings of the five parts of the survey, featuring summaries and visualizations of the responses to questions on (1) the assessment of the current state of the art of data quality; (2) the assessment of the most pertinent error sources that potentially have a negative effect on the reliability and trustworthiness of datasets; (3) institutional arrangements regarding data quality; (4) established processes regarding data quality; and (5) change in relation to data quality.

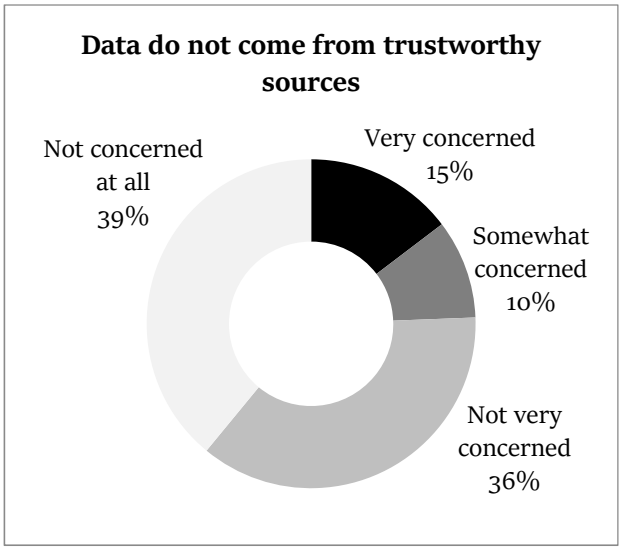
Quality assessment

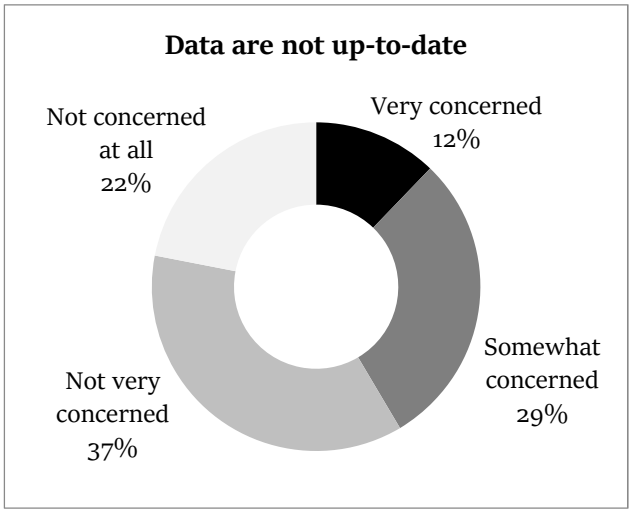
The first part of the questionnaire revolved around respondents' assessments of the data that their organization produced/handled. Based on the question *"When you think of the data that your organization enters into European databases, how concerned are you about the following?"*, several potentially relevant dimensions of data quality were presented and respondents were asked to indicate their level of concern about each of the dimensions on a four-point scale ("very concerned", "somewhat concerned", "not very concerned", and "not concerned at all").

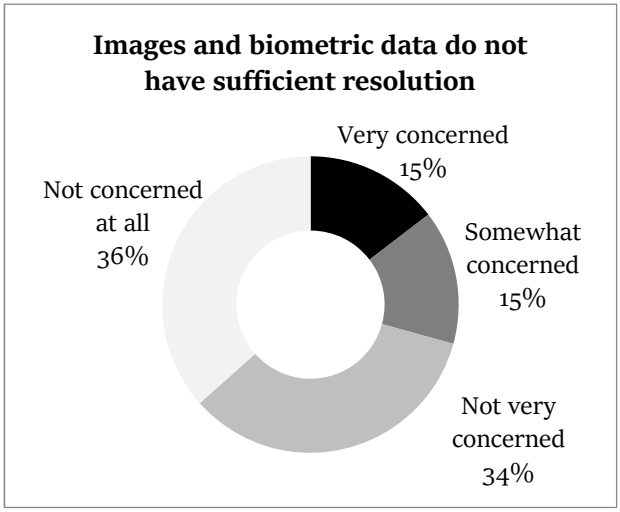








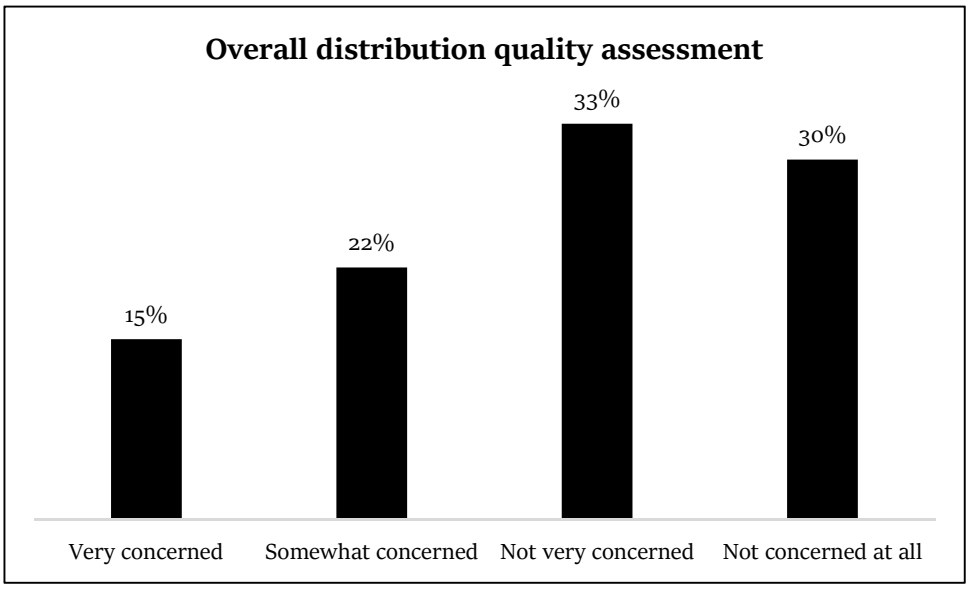




The responses indicate mixed assessments across all included dimensions of data quality. Respondents expressed particular confidence in the trustworthiness of data sources (combined 76%), in the correct representation of real-world phenomena (combined 67%), as well as in the timeliness of data (combined 59%).

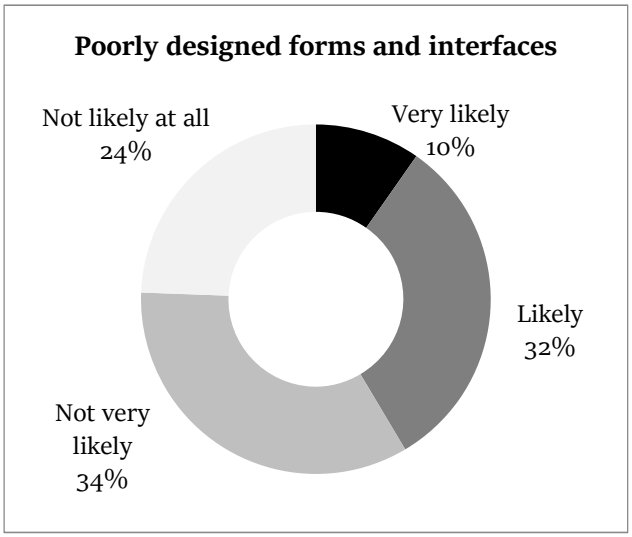
On the other hand, responses also indicate that in some cases, almost half of the respondents are "very concerned" or "somewhat concerned" about the quality of the data that their organization produces and handles. This is particularly true for missing values in the data (combined 47%), violations of semantic rules (combined 44%), and updates to data (combined 41%).

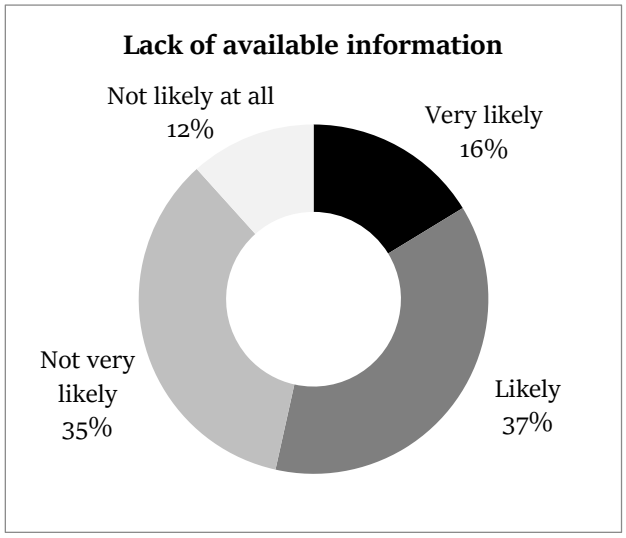
The overall distribution of assessments of data quality across all dimensions shows a general trend towards trust in the data that respondents' organizations produce and handle (combined 63% for "not very concerned" and "not concerned at all").

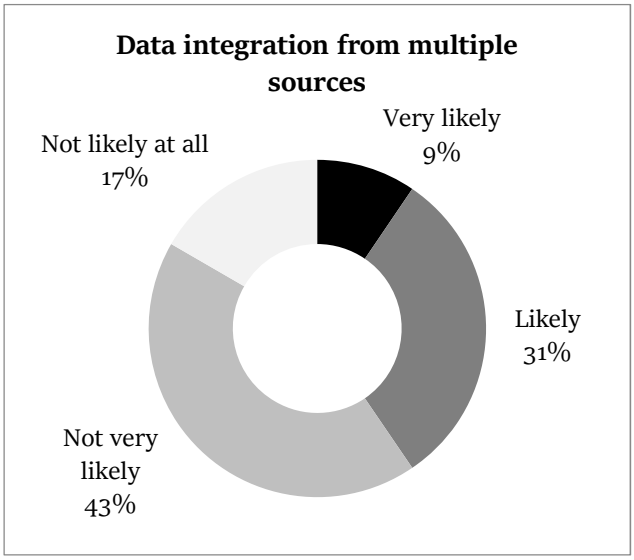


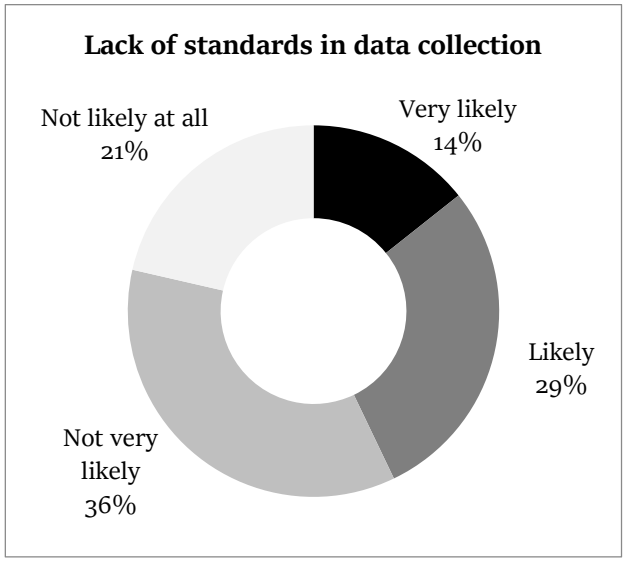
Error sources

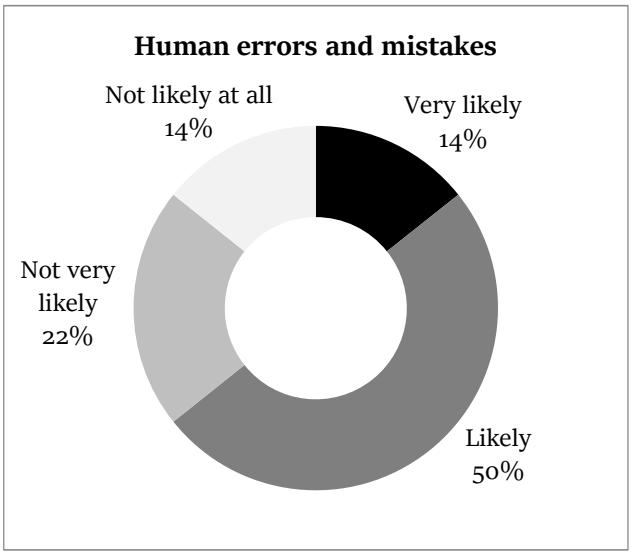
The second part of the questionnaire revolved around respondents' assessments of potential error sources impacting the data that their organization produces and handles. Based on the question *"In your opinion, how likely are the following issues to negatively affect the quality of the data that your organization enters into EU databases?"*, several potential error sources that can negatively affect data quality were presented and respondents were asked to assess how likely those would be to negatively affect the data produced and handled in their organization ("very likely", "likely", "not very likely", and "not likely at all").

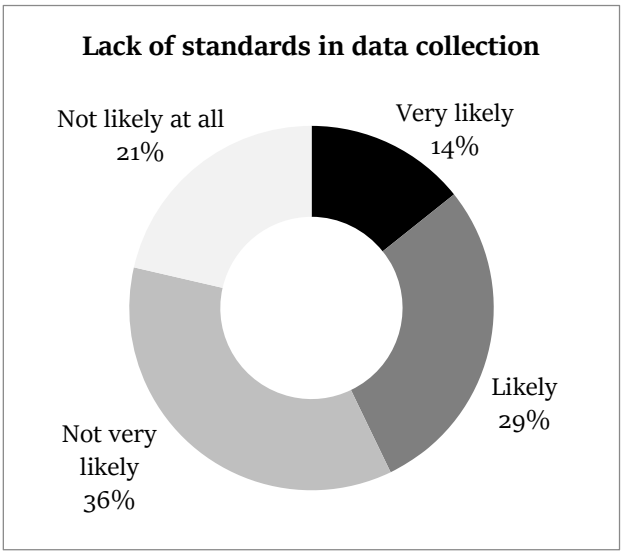


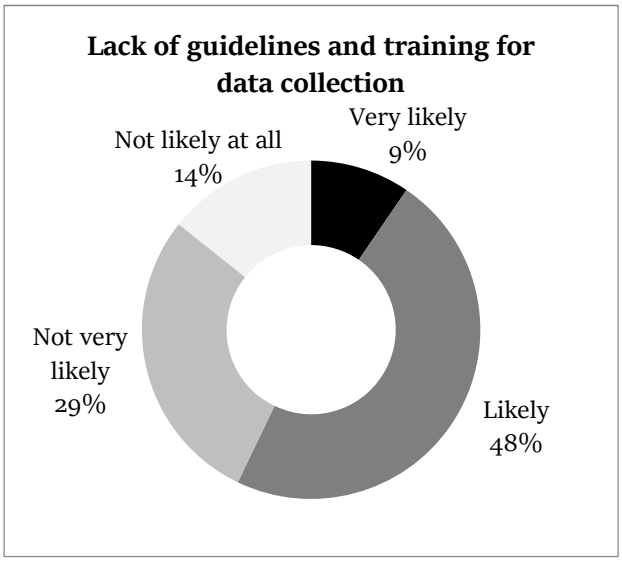




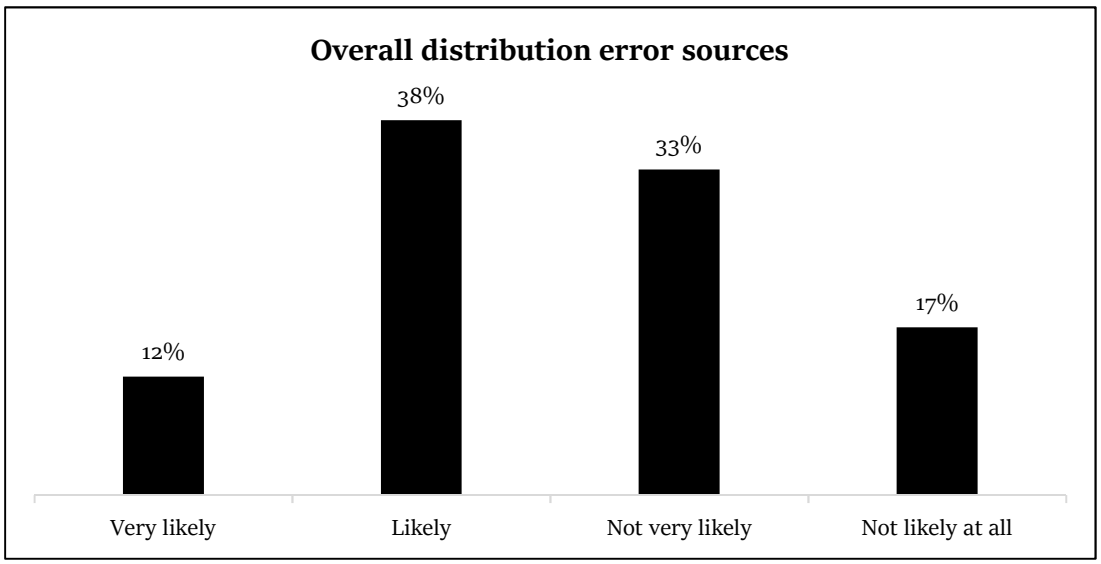








Similar to the previous block on data quality assessment, the response distributions in relation to potential error sources in data indicate a considerable degree of variation across participating organizations. A general trend can here identified around the two middle categories ("likely" and "not very likely", combined total between 65% and 74%), whereas the two extreme categories ("very likely" and "not likely at all") take up a much smaller amount of responses across all error dimensions (combined total between 24% and 35%).

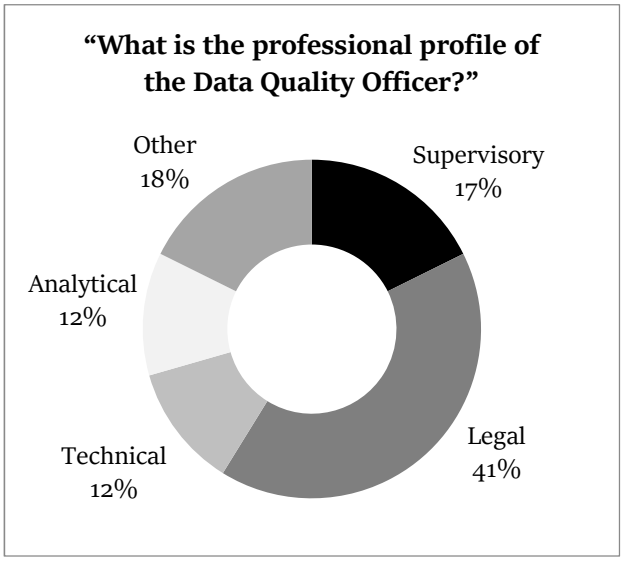


Structures

The third part of the questionnaire revolved around organizational structures. The goal of this part was to understand the status and organizational anchoring of data quality. Questions were set up in a way that presented different paths through this block, depending on the structures present in the respondents' organizations.

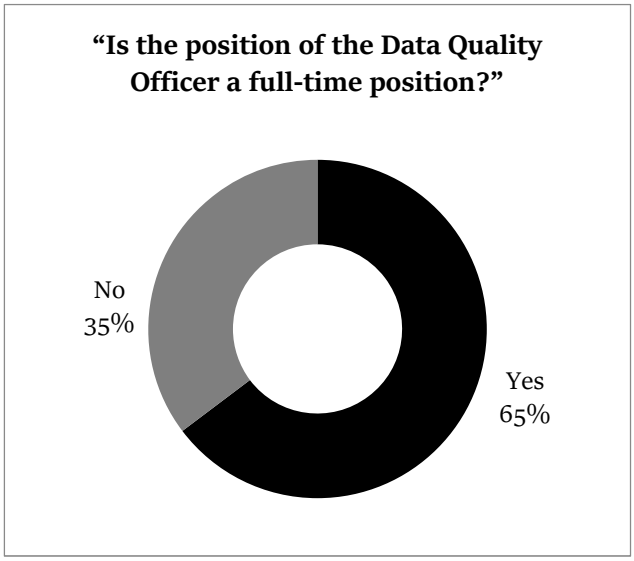
Respondents were first of all asked whether their organization had a dedicated data quality officer (i.e. a person who is formally responsible for data quality). 40% indicated that within their organization, there was such a position, whereas 24% indicated that there was no such position. A considerably large percentage of respondents (36%) indicated that they did not know whether such a position existed in their organization.

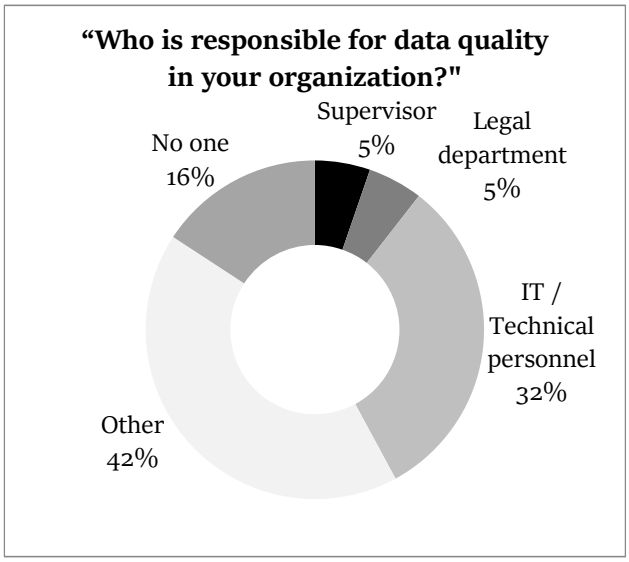
In case the question about the existence of a data quality officer position within the organization was answered with 'yes', respondents were asked about the professional profile of the data quality officer. 41% of respondents indicated that the position was set-up as a legal position. 18% of respondents indicated that the data quality officer in their organization had a supervisory role, 12% of respondents indicated an analytical profile, and another 12% indicated a technical profile. Finally, 18% of respondents indicated that the profile was different from the offered descriptions.



Respondents were further asked whether the position of the data quality officer was a full-time position (i.e. exclusively dedicated to data quality tasks) or whether it was a part-time position (i.e. part of a broader professional profile). 65% of respondents indicated that the former was the case, 35% indicated the latter.

Respondents who had indicated that there was no dedicated data quality officer in their organization were instead asked who was de facto responsible for data quality in their organization. Almost a third of respondents (32%) indicated that this was a technical task covered by IT personnel. 5% of respondents each indicated that the supervisor or the legal department were responsible for data quality. In 42% of the responses, none of the offered categories applied, and in 16% of the responses there were no specified responsibilities.





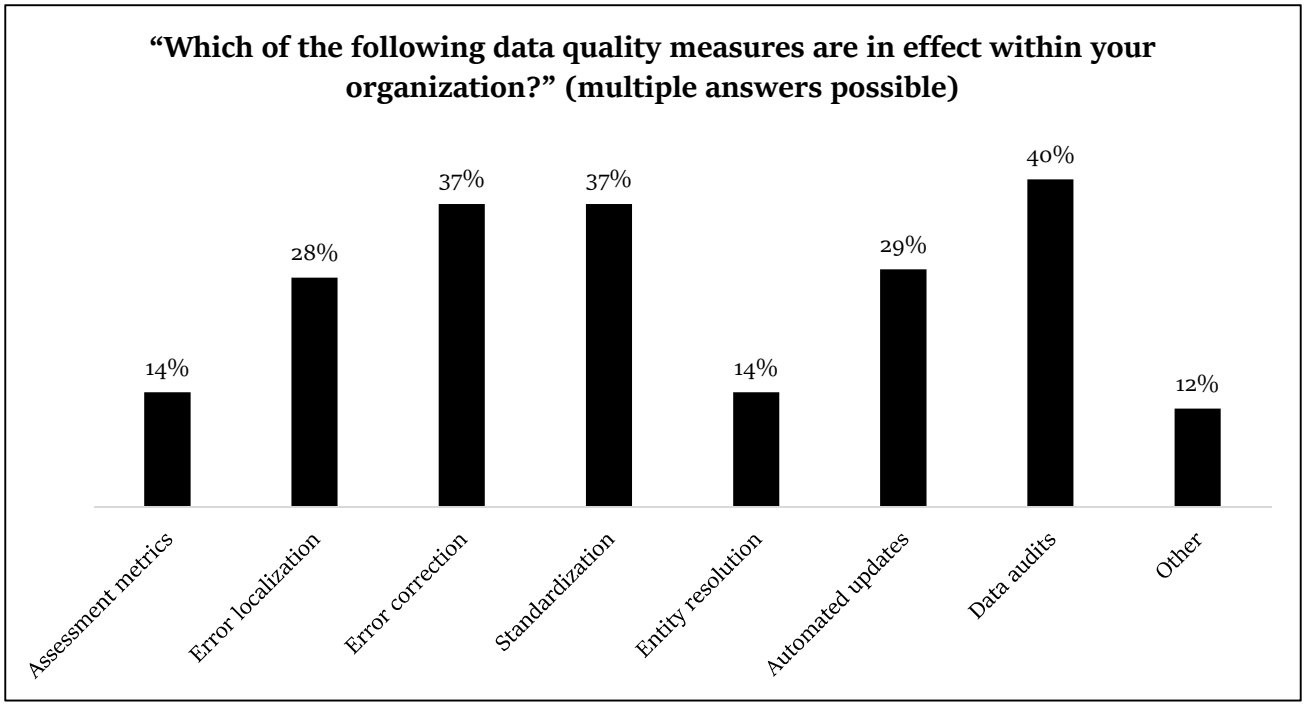
As a follow-up, respondents were additionally asked which other persons in their organization were additionally responsible for data quality. Free-text answers (no word limit) included case officers, liaison officers, control officers, a dedicated data and information management unit, and the office of the national commissioner.

Processes

The fourth part of the questionnaire revolved around data quality processes. The goal of this part was to understand the ways in which data quality policies are structured and how data governance and data quality practices are enacted within organizations.

Respondents were first asked whether their organization had a data quality policy in place. The large majority of respondents (71%) indicated that this was the case. 6% of respondents indicated that no such policy was in place, and 23% indicated that they did not know whether this was the case or not.

Next, respondents were asked what kinds of data quality measures were already in effect in their organization. As multiple answers were allowed, relative frequency counts are below offered for each measure.



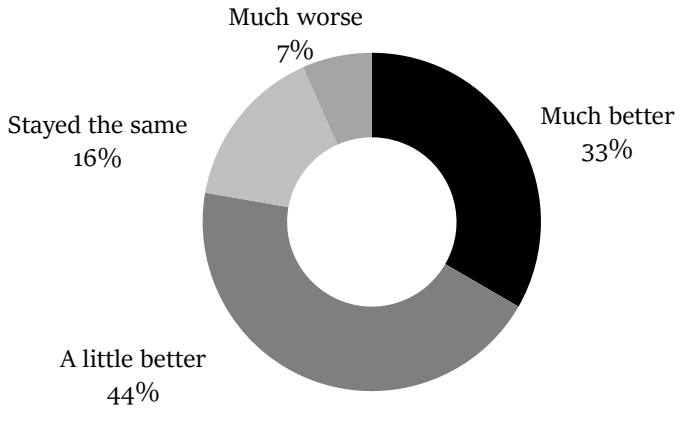
As a follow-up, respondents were asked which other data quality measures were (additionally) in effect within their organization. Free-text answers most frequently referred to manual reviews and checks. Other answers included verification processes, automated screening, and the use of checklists listing common errors.

Change

The fifth and final block of the questionnaire was concerned with change of data quality over time. The goal was to understand tendencies and trends in data quality and related processes and arrangements.

Respondents were first asked to assess whether over the past five years, the data quality in their organization had improved, stayed the same, or deteriorated. A majority of respondents indicated a positive assessment, i.e. that data quality had either become "much better" (33%) or "a little better" (44%). 16% assessed data quality as having "stayed the same", while only 7% indicated that data quality had become "much worse". No respondents indicated that data quality had become "a little worse".

"In your opinion, over the past five years, has the quality of the data that your organization transfers to EU databases become better or worse?"



To inquire about the reasons for change, respondents were as a follow-up question asked to explicate the reasons for their assessment. Free-text responses for the positive assessment categories ("much better" and "a little better") are in the following discussed together.

The most pertinent cluster of answers revolved around the theme of technical upgrades and new tools, both for quality assessment and quality improvement. Several respondents also highlighted the automation of diagnostics and verification processes as important aspects in the improvement of overall data quality.

A second theme throughout the responses revolved around the improvement of business processes, including general and use-case specific definitions of data quality, as well as institutional knowledge management. This also included the definition and/or clarification of roles and responsibilities in the area of data quality.

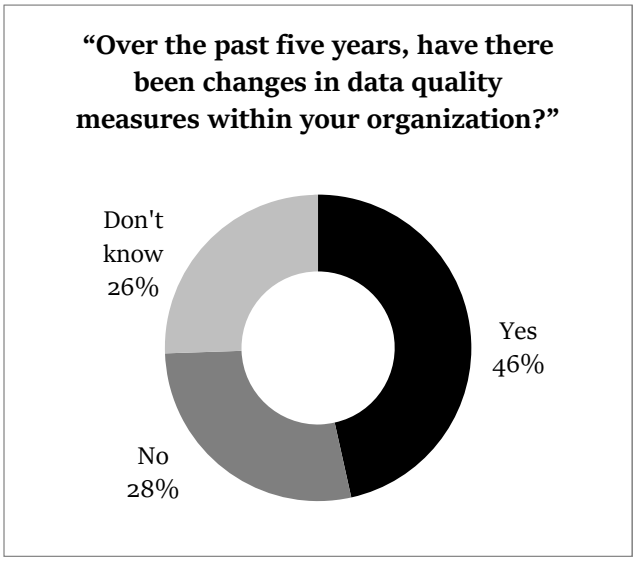
A third cluster emerged in relation to the allocation of more staff and training programs for staff. The latter include data production and data input contexts, but also quality control processes.

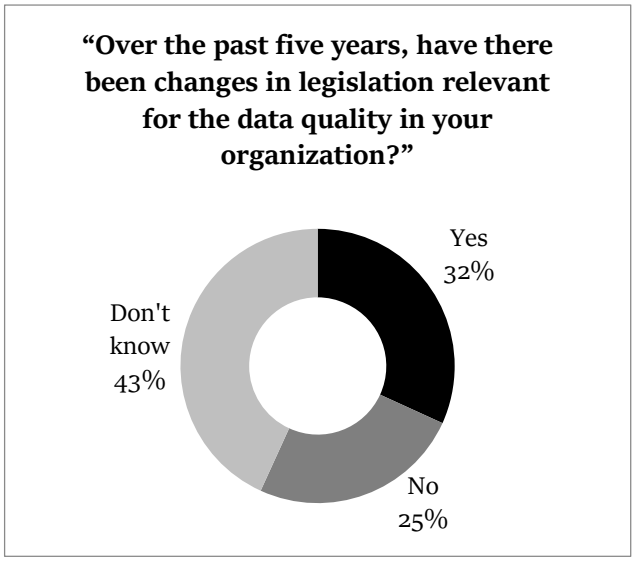
Finally, respondents pointed to the effects of changed legislative frameworks and regulations, as well as to better database integration (at both national and supranational level), including the use of biometrics for the de-duplication of records.

Respondents who indicated that data quality had stayed the same in their organization over the past five years highlighted how improvement efforts were undercut by staff shortages and difficulties to recruit qualified personnel, as well as unchanged top-down data quality standards imposed by EU-level databases.

For those respondents who indicated that data quality had deteriorated in their organization, no free-text answers were recorded.

Respondents were subsequently asked whether over that same time period, changes in the data quality measures in their organization had occurred. Almost half of the respondents (47%) indicated that this had been the case, whereas 28% indicated that this had not been the case and 26% did not know. For those respondents indicating that changes in data quality measures had been undertaken, a follow-up free-text field asked them to specify these changes.





The most pertinent cross-cutting themes in the answers revolved around verification process and reports, as well as regular audits. A second major theme concerned the establishment of dedicated data governance frameworks, including clearly defined responsibilities, processes, and workflows for a number of issues such as technical solutions, collaboration with other units/organizations, and legal aspects such as retention times and corresponding deletion deadlines. A third theme throughout the answers concerned technical improvements, including the implementation of specialized software for data profiling and metadata production, allowing for better diagnostics and improved follow-up measures. Moreover, respondents highlighted revisions of classification systems and interfaces, potentially resulting in less error during data production and data input.

Finally, respondents were asked whether over the past five years there had been changes in relevant legislation that had an impact on data quality in their organization. Almost a third of respondents (32%) indicated that this was indeed the case, whereas 25% indicated that this was not the case and 43% indicated that they did not know. For those respondents indicating that changes in legislation had been relevant for data quality in their organization, a follow-up free-text field asked them to specify these changes.

Responses primarily highlighted the top-down effects of international regulations, as they became implemented at the national level. Among those legal frameworks, the EU General Data Protection Regulation (GDPR), and to a lesser extend the Data Protection Law Enforcement Directive (DPLED) were most frequently mentioned.

Best practice recommendations

The questionnaire concluded with an open question, asking respondents "*Can you provide best practice recommendations for data quality?*" For answers, a free-text field without word limit was provided. In the following, the most pertinent clusters across all responses are summarized.

The recommendation that was put forward most frequently concerned regular audits and assessments of data quality, as well as correction and improvement measures based on the results of such audits and assessments.

This was followed by recommendations concerning data sources, and in this context most notably the production contexts of data. Respondents in this context highlighted that manual input of data should be limited as far as possible, and that manual input should be validated in automated ways. Moreover, they suggested to limit unstructured data (e.g., free-text descriptions), resulting in more readily machine-readable and analyzable datasets. Finally, respondents pointed out that working conditions for staff concerned with data production and data input should be improved, especially in light of workload and adequate technical equipment.

Another prevalent theme revolved around the question of clearly defined responsibilities for data quality and, more generally speaking, a holistic governance approach that includes clearly specified roles and processes.

A final major theme in respondents' answers concerned improved database integration and questions of harmonization. Especially in regard to the latter point, several answers pointed to the need for a review and streamlining of classification systems, making data structures more easily understandable and usable, as well as compatible.

Discussion

Overall, survey findings fall in line with expectations based on literature reviews, previous research experiences, as well as based on the pre-testing phase and personal exchanges with law enforcement and border control professionals over multiple years. There are several pertinent insights from the responses provided by national-level law enforcement and border control agencies.

Firstly, there is considerable variance in regard to the *assessment* of the quality of the data that law enforcement and border control organizations produce and handle. While a general tendency towards a favorable assessment of data quality can be noted, responses show mixed levels of trust in data across all relevant dimensions of data quality covered by the questionnaire. These mixed assessments correspond closely with the variance in institutional arrangements and processes discussed further below.

Secondly, in regard to the potential *error sources* that negatively impact data quality, similar variance can be observed. A clearly discernible trend can, however, be diagnosed in relation to those error sources that are caused by humans, i.e. errors during data generation, classification of phenomena, etc. Moreover, it becomes clear that respondents additionally relate such human errors to shortcomings in technical infrastructures, i.e. databases and user interfaces. Finally, respondents also relate error sources to poor or insufficient guidelines for data generation and handling, as well as to a lack of proper training. These assessments correspond closely with literature that has identified frontline data collection/generation as a major concern for overall data quality in law enforcement and border control organizations.

Thirdly, in regard to the *structures* relevant for data governance and data quality control procedures, major fragmentation can be witnessed in relation to the prioritization of data quality as well as the professional domains within which it is located. A considerable part of respondents indicated that their organizations do not have a dedicated data quality officer. Additionally, it can be assumed that in those cases where no knowledge about the existence of a data quality officer was available, no such position was existent either. In organizations without clearly defined professional roles, responses indicate that data quality tasks in practice fall mainly within the scope of IT/technical personnel. On the contrary, in those organizations that indicated that they had a dedicated data quality officer, the professional profile of this role is overwhelmingly legal. This arguably points to a conceptualization of data quality primarily as a question of compliance with legal regulations such as the GDPR and the DPLED, for example regarding modes of data storage and retention times, forms of processing, or access and rectification rights for data subjects. Overall, responses on data quality structures indicate considerable variation not only in regard of the existence of such structures in the first place, but also in regard to their design and disciplinary scope.

Fourthly, in regard to *processes* survey responses indicate that data quality measures are widely in place. The concrete tools and activities used to ensure data quality do, however, vary widely. Generally speaking, there is a trend towards automation (assessment, correction) that can complement manual review and validation practices.

Fifthly, in regard to *change* survey responses indicate that data quality has over the past five years become a somewhat dynamic area in law enforcement and border control contexts. Overall, results

show that responding organizations consider the data that they produce and handle more reliable and put more trust in them to be fit for the tasks that they are used for. This tendency falls in line with wider (political) trends towards an awareness and acknowledgement of potentially unreliable data in law enforcement and border control – with the willingness to address these shortcomings. This also corresponds with the findings on processes and data quality measures that are already in place in surveyed organizations. The acknowledgement that attention and resources need to be devoted to the management of data in a systematic fashion speaks to the fact that data governance and data quality are increasingly becoming professionalized.

Finally, *best practice recommendations* given by survey respondents highlight that although there is change towards more systematic and structured approaches to data quality, there remains potential for further improvement. Overall, recommendations for best practices correspond closely with literature on data quality in security organizations and other public agencies, highlighting how the human element of data quality, as well as human-computer interfaces should be addressed. They also correspond closely with the responses to the questionnaire as detailed above.

In *summary*, survey responses indicate considerable variation in the assessment of data quality itself, as well as in terms of potential error sources. Responses do, however, suggest that human errors and the lack of proper training and guidelines for data collection stand out as major concerns for reliable and trustworthy data. Further, survey results indicate a high degree of variation in terms of institutional arrangements and processes surrounding data quality. Organizational approaches to data quality vary widely, including the question whether data quality is seen as a legal, technical, or practical issue. Similarly, there is little unanimity in regard to existing policies and data quality measures.

Overall, findings suggest growing awareness of data quality issues and increasing professionalization of data quality practices. Yet, there is considerable heterogeneity in approaches to data quality, raising concerns in light of the aggregation of data in EU-level systems where they are used for transnational law enforcement and border control purposes.

Limitations of the study

While survey research has been deemed a viable strategy in organizational research to cover and assess phenomena that cannot be observed directly (Bartlett, 2005: 98), the approach has limitations that should be kept in mind.

Firstly, law enforcement and border control agencies tend to show a considerable degree of organizational variance, i.e. they differ in size, structure, tasks, budget, and not least culture. The reasons for such variance are usually found in historical developments and/or national traditions, as well as forms of political organization. Variations in data quality structures (e.g., institutional arrangements and responsibilities, processes, etc.) should thus also be understood against the background of organizational variance.

Secondly, a general issue in survey research is social desirability bias, i.e. the fact that respondents tend to give answers they believe are expected from them based on social conventions, thus distorting survey results. Social desirability bias is a particular problem in sensitive contexts, for example when

it comes to taboo topics or discriminatory attitudes where self-representation might considerably differ from actual beliefs (Krumpal, 2013). In regard to data quality, it cannot be ruled out that social desirability bias played a role in answering the questionnaire, for example when it comes to the assessment of the quality of the data that organizations produce and handle. Given the societal implications of "bad" data in security contexts, respondents might have painted a more optimistic picture of data quality than is actually the case – which would correspond with the generally positive attitudes towards data quality expressed throughout the survey.

Finally, some variance can arguably also be contributed to the fact that, in all likelihood, respondents with different professional profiles and positions responded to the questionnaire. While this is a fairly common issue in survey research on organizations, it does, however, at the same time indicate a limited comparability between responses from different organizations. Given the lack of clear responsibilities for data quality and data governance that became apparent from the survey data, there might thus have been situations where different persons with different professional perspectives would have provided different answers.

Conclusions

The aim of this survey was to come up with first systematic diagnosis of the current state of the affairs in regard to data quality in law enforcement and border control cooperation in the EU/Schengen area. Survey results on the one hand indicate a general tendency towards awareness of data quality challenges and subsequently some degree of professionalization in terms of the structures and processes that are relevant for data quality. On the other hand, results do, however, also indicate that a considerable amount of variation and fragmentation still in place. This can be seen as cause for concern in regard to information sharing at the supranational level.

Implications from the findings can be drawn in regard to three dimensions.

Firstly, from a *practical perspective*, survey results indicate the need to further professionalize data quality as an everyday activity in law enforcement and border control agencies. This includes the specification of data quality requirements for different use cases, the definition of roles and responsibilities, as well as the allocation of sufficient resources for the continuous implementation of data quality processes. Data quality measures, as survey responses indicate, should in this context not exclusively be conceived of as a technical issue, but explicitly include the role of human activities and human-computer interactions.

Secondly, from a *policy perspective*, survey results tease out the need for further reform and harmonization, both on the national and the supranational level. The need for reliable and trustworthy data becomes aggravated when data are entered into EU databases and shared across national boundaries. Currently, two major political initiatives at the European level can be witnessed in this regard: the data quality roadmap initiated by the Justice and Home Affairs Council as well as the Commission's Implementing Decisions 2021/2224 and 2021/2225 that specify minimum quality requirements for data that are entered into EU-level systems. Especially the former initiative that strives to harmonize standards and practices regarding data quality at the national level across all EU/Schengen countries is pertinent in the context of survey results.

Thirdly, from a *research perspective*, survey results underline the pressing need for more academic engagement with data quality. Notably, while literature from statistics and computer science has dealt with aspects of data properties and databases and literature from business management has dealt with concrete assessment and improvement methods for data quality, there remains a shortage of social scientific study of data quality. Survey results highlight the importance of data quality for governance, in this case specifically for law enforcement and border control tasks. Just as well, they indicate the importance of studying data quality from a holistic perspective that takes into account the organizational embeddedness of data and the socio-technical composition of data practices.

Annex: Methodology

This annex explicates the methodological considerations that informed the design, implementation, sampling, roll-out, and analysis of the survey.

Survey design

Survey and questionnaire design requires careful planning, including the definition of the research question and the scope of the survey, as well as considerable preparatory work that precedes actual data collection and analysis (Diekmann, 2007: 192).

Theoretical background

A first step in survey design usually consists of theoretical work that explicates the relations between questions and theoretical concepts, as well as empirical expectations (Porst, 2014: 16; Reinecke, 2014: 602). This theoretical background can then at a later point in time inform decisions on what information is needed to address the research question, what specific questions should be asked, and in which sequence (Bradburn et al., 2004: 315-6). Three distinct bodies of literature were considered particularly relevant in the context of CURATE.

Firstly, based on literature from criminology (e.g., Kusak, 2022; Leese, 2023; O'Connor et al., 2021), border studies (e.g., Forti, 2022; Pollozek, 2020; van Rossem, 2021), and EU internal security scholarship (e.g., Bigo, 2021; Leese, 2022; Levi and Wall, 2004), a general understanding of the production and handling of data within security organizations was derived. This literature suggests that data quality in law enforcement and border control is often considered questionable, for example due to the lack of available and/or reliable information vis-à-vis crime and (irregular) migration, but also due to human errors in data production and handling, as well as technical issues in both database infrastructures and user interfaces. Additionally, this literature suggests that law enforcement and border control agencies are highly diverse in their internal organization, resulting in fragmented data quality requirements as well as variation in approaches to data governance and related processes.

Secondly, based on literature from computer science and statistics (e.g., Batini et al., 2009; Batini and Scannapieca, 2006; Wang et al., 2002), relevant dimensions of data quality in security contexts were identified. This literature highlights that data quality is a multi-faceted and relational concept. This means that data quality can be assessed in regard to different properties of datasets, for example concerning their completeness, their internal integrity, or their timeliness. While there is little unanimity when it comes to number of data quality dimensions and their definitions, it is important to notice that data quality assessments are usually carried out in accordance with the quality requirements that have been defined for particular use cases.

Thirdly, based on business management literature (e.g., Loshin, 2001; McGilvray, 2008; Morbey, 2011), practical approaches to the improvement of data quality were identified. This literature highlights the importance of clearly defined institutional arrangements and processes if data quality is to be ensured at the required level. Moreover, this literature suggests that responsibilities for data quality should be clearly defined and allocated and sufficient resources should be budgeted. Finally,

this literature emphasizes that data quality efforts should not be considered as a one-off activity but that they require continued execution and adjustments.

Scope of the questionnaire

Building on this broad interdisciplinary theoretical and conceptual basis, the questionnaire was designed to cover five major themes: (1) the assessment of the current state of the art of data quality, i.e. how organizations think about the quality of the data that they produce and handle; (2) the assessment of the most pertinent error sources that potentially have a negative effect on the reliability and trustworthiness of datasets; (3) institutional arrangements regarding data quality, i.e. whether there are dedicated structures and how these structures are designed; (4) established processes regarding data quality; i.e. which tools and measures are already in use; and (5) change in relation to data quality, i.e. whether there are positive or negative trends in data quality and whether institutional arrangements and processes are in flux. Additionally, respondents were at the end of the questionnaire asked to give best practice recommendations for data quality.

Pre-testing

Pre-testing is considered an essential step in questionnaire design and survey preparation (Bradburn et al., 2004: 317). It serves to ensure that all relevant aspects necessary for the research goals are included and that they are presented and asked in a way that avoids ambiguity and misunderstandings which cannot be rectified once the survey has been rolled out (Reinecke, 2014: 614).

Therefore, the initial version of the questionnaire was between August and October of 2022 shared with eight contacts from law enforcement and border control agencies. Respondents were asked to give feedback on the understandability of the questions, whether all relevant aspects concerning data quality were covered from a practical perspective, and whether the wording and the time to complete the survey corresponded well with their professional vocabulary and working environment.

Moreover, the initial version of the questionnaire was shared with eight academic colleagues whose research is concerned with the role of data and databases in the field of security. Respondents were asked to give feedback on whether the most relevant aspects of data quality had been considered from an academic point of view.

Based on responses from the pre-testing phase, the final questionnaire was revised for consistency and precision, as well as shortened.

Technical implementation

The questionnaire was implemented via Qualtrics XM,[5] a professional service that is data protection compliant and allows, among other things, for the design and management of online surveys. For management purposes, a unique access link to the questionnaire was created for each organization that was invited to take part in the survey. This allowed for respondents to resume the questionnaire from the last position in case it was not completed on the first attempt. Furthermore, it allowed

[5] https://www.qualtrics.com/

tracking whether the organization had opened and/or completed the questionnaire and enabled individualized follow-up/reminder emails.

Language

A major challenge in cross-national survey research concerns the question of language. Translation of a questionnaire into different languages offers understandability for respondents who only speak their native language. However, translation also inevitably causes variation and might thus interfere with the comparability of responses from different countries (Brace, 2008: 208).

Due to practical considerations, English was selected as uniform language for the entire survey. English is the standard language in international contexts and it was expected that survey respondents would be highly educated professionals who have proficient command of the English language.

Sampling and roll-out

Identification of relevant organizations

Relevant organizations to be invited to participate in the survey were identified via desk research based on whether they provide data for one (or more) of the following EU-level systems for law enforcement and border control cooperation: the Europol Information System (EIS), the Schengen Information System (SIS II), and the European Asylum Dactyloscopy Database (Eurodac). Overall, 191 relevant organizations were identified.

The focus on national-level agencies was chosen in accordance with data sharing mechanisms in EU internal security matters, notably the fact that member states retain legal ownership of the data that they submit to EU-level databases and remain responsible for data quality.

Based on the different domains covered by the three systems (police matters, judicial cooperation, border control, asylum and migration management), for each country it was expected to identify multiple relevant organizations producing and submitting data (e.g., police, migration and asylum authorities, border guard, coastguard, etc.). In many countries, the number of relevant organizations was further multiplied by federalist structures.

Despite these considerations, the overall number of relevant organizations was still considered low enough to be covered by a census approach (N=population) rather than through sampling. Census approaches are not uncommon in the study of particular organization types and are considered a viable way to create a comprehensive picture of particular aspects of organizational practices (Passmore and Baker, 2005: 53).

Selection of respondents

A considerable challenge is often posed by the identification of persons competent to answer a questionnaire within organizations. Bloom and van Reenen (2010: 105) in this regard suggest to retrieve information from organizations' websites or to contact them via phone to ask who is responsible for a particular task within the organization. While this might be a viable strategy for

industry or business contexts, it is, however, more difficult to realize in the security domain. Law enforcement and border control agencies usually do not provide information about their staff online, which means that individual responsibilities and contact details cannot be retrieved from official websites. Moreover, many of them do not provide telephone contact details, meaning the only available contact information is often an email address for the press or media office that serves as single point of contact for inquiries from the outside.

It was thus decided to distribute survey invitations to the official contact point of each organization, coupled with the request to internally forward the access link to the questionnaire to the relevant division/responsible person within the organization. While this approach yielded good results in terms of the overall response rate (see below), it also meant that there was no way to retrace internal communication and forwarding processes. In other words, it was in most cases not possible to know who eventually completed the questionnaire.

Tracking

The initial batch of survey invitations was sent out between 17 and 20 January 2023. Participation in the survey was tracked via tokens that allowed access to questionnaires only via the personalized link included in the invitation email.

For non-completed questionnaires, a first reminder was sent out one week after the initial invitation. A second reminder was sent out after another week, this time specifying a deadline (15 February) for the completion of the questionnaire. Eventually, a final reminder was sent out four weeks after the original invitation, extending the final deadline to 28 February.

Response rate

Out of the originally identified population of 191 relevant organizations, 20 either indicated that they did not consider themselves competent to give information on data quality or could not be reached due to the lack of contact information/a functioning email address for the official point of contact. Out of the remaining 171 organizations, 68 completed the questionnaire. Out of those 68, three did not agree to the mandatory consent question at the beginning of the questionnaire. Overall, participation has thus resulted in 65 out of 171 possible responses, equaling a response rate of 38 percent.

Responses were recorded from organizations in the following 27 countries: Austria, Belgium, Bulgaria, Croatia, Cyprus, Czechia, Denmark, Estonia, Germany, Greece, Hungary, Iceland, Ireland, Latvia, Liechtenstein, Lithuania, Luxembourg, Malta, Norway, Poland, Portugal, Romania, Slovakia, Slovenia, Spain, Sweden, and Switzerland.

Analysis

Survey data were prepared and analyzed using the statistical software package R.[6] In line with the overall goal to explore the current state of affairs in regard to data quality in European law enforcement and border control cooperation, for the quantitative parts of the survey descriptive

[6] https://www.r-project.org/

statistical methods were chosen. Descriptive statistics are key in adequately describing datasets and understanding the information that data provide in regard to the research question (Agresti, 2018: 4; Ott and Longnecker, 2016: 62). Notably, descriptive statistics can be used to visualize data and make them easily accessible and comprehensible (Mohr et al., 2022: 15).

More advanced descriptive statistics such as means, ranges, or standard deviations were not considered suitable given the limited number of cases in the dataset. Just as well, inferential statistical methods were not used for the same reason.

The non-structured parts of the dataset (resulting from free-text responses) were analyzed by means of qualitative data analysis. Specifically, using the computer-assisted analysis software MaxQDA,[7] text elements were clustered into meaningful categories across the responses from participating organizations.

[7] https://www.maxqda.com/

References

Agresti A (2018) *Statistical Methods for the Social Sciences, Fifth Edition*. Boston: Pearson.

Bartlett K R (2005) Survey Research in Organizations. In Swanson R A & Holton E F (eds.) *Research in Organizations: Foundations and Methods of Inquiry*. San Francisco: Berrett-Koehler Publishers, 97-114.

Batini C, Cappiello C, Francalanci C and Maurino A (2009) Methodologies for Data Quality Assessment and Improvement. *ACM Computing Surveys* 41(3): Article 16.

Batini C and Scannapieca M (2006) *Data Quality: Concepts, Methodologies and Techniques*. Berlin/Heidelberg: Springer.

Bigo D (2021) Interoperability: A Political Technology for the Datafication of the Field of EU Internal Security? In Bigo D, Diez T, Fanoulis E, Rosamond B & Stivachtis Y A (eds.) *The Routledge Handbook of Critical European Studies*. London/New York: Routledge, 400-417.

Bloom N and van Reenen J (2010) New Approaches to Surveying Organizations. *American Economic Review: Papers & Proceedings* 100(May): 105-109.

Brace I (2008) *Questionnaire Design: How to Plan, Structure and Write Survey Material for Effective Market Research. 2nd Edition*. London/Philadelphia: Kogan Page.

Bradburn N, Sudman S and Wansink B (2004) *Asking Questions: The Definitive Guide to Questionnaire Design - For Market Research, Political Polls, and Social and Health Questionnaires*. San Francisco: Jossey-Bass.

Council of the European Union (2020a) 10422/20: Questionnaire Preparing the First European Data Quality Day (EDQD), 9 September. Brussels.

Council of the European Union (2020b) 11825/20: European Data Quality Day - Summary of the Replies, 11 November. Brussels.

Diekmann A (2007) *Empirische Sozialforschung: Grundlagen, Methoden, Anwendungen*. Rowohlt: Hamburg.

Forti M (2022) Data Quality Issues in EU Digital Procedures of Border Control. *Diritti Fondamentali* (3): 230-244.

Krumpal I (2013) Determinants of Social Desirability Bias in Sensitive Surveys: A Literature Review. *Quality & Quantity* 47(4): 2025-2047.

Kusak M (2022) Quality of Data Sets That Feed AI and Big Data Applications for Law Enforcement. *ERA Forum* online first: 10.1007/s12027-022-00719-4.

Leese M (2022) Fixing State Vision: Interoperability, Biometrics, and Identity Management in the EU. *Geopolitics* 27(1): 113-133.

Leese M (2023) Enacting Criminal Futures: Data Practices and Crime Prevention. *Policing and Society* 33(3): 333-347.

Levi M and Wall D S (2004) Technologies, Security, and Privacy in the Post-9/11 European Information Society. *Journal of Law and Society* 31(2): 194-220.

Loshin D (2001) *Enterprise Knowledge Management: The Data Quality Approach*. Burlington: Morgan Kaufmann.

McGilvray D (2008) *Executing Data Quality Projects: Ten Steps to Quality Data and Trusted Information*. San Francisco: Morgan Kaufmann Publishers.

Mohr D L, Wilson W J and Freund R J (2022) *Statistical Methods, Fourth Edition*. London/San Diego/Cambridge/Oxford: Academic Press.

Morbey G (2011) *Data Quality for Decision Makers: A Dialog Between a Board Member and a DQ Expert*. Wiesbaden: Springer Gabler.

O'Connor C D, Ng J, Hill D and Frederick T (2021) Thinking About Police Data: Analysts' Perceptions of Data Quality in Canadian Policing. *The Police Journal* online first: 10.1177/0032258X211021461.

Ott R L and Longnecker M (2016) *An Introduction to Statistical Methods & Data Analysis, Seventh Edition*. Boston: Cengage Learning.

Passmore D L and Baker R M (2005) Sampling Strategies and Power Analysis. In Swanson R A & Holton E F (eds.) *Research in Organizations: Foundations and Methods of Inquiry.* San Francisco: Berrett-Koehler Publishers, 45-56.

Pollozek S (2020) Turbulences of Speeding Up Data Circulation: Frontex and its Crooked Temporalities of 'Real-Time' Border Control. *Mobilities* 15(5): 677-693.

Porst R (2014) *Fragebogen: Ein Arbeitsbuch. 4th Edition.* Wiesbaden: Springer VS.

Reinecke J (2014) Grundlagen der standardisierten Befragung. In Baur N & Blasius J (eds.) *Handbuch Methoden der empirischen Sozialforschung.* Wiesbaden: Springer VS, 601-618.

van Rossem W (2021) Data Quality in Data Infrastructures of Migration and Border Control. Paper presented at the 8th STS Italia Conference, Trento, Italy.

Wang R Y, Ziad M and Lee Y W (2002) *Data Quality.* New York/Boston/Dordrecht/London/Moscow: Kluwer Academic Publishers.