Statistical Commission
Fifty - Fifth Session
27 February – 1 March 2024

Background document
(Available in English only)

Item 3(d) of the provisional agenda
Items for discussion and decision: data science

Annotated Outline

# of the Playbook on integrating Data Science in the work of Statistical Offices

Prepared by United Nations Statistics Division

Table of Contents

Playbook on integrating Data Science in the work of Statistical Offices

Data Science Leaders Network

The Data Science Leaders Network (DSLN) was established by the Statistical Commission in 2022. DSLN reports is part of the UN Committee of Experts on Big Data and Data Science for Official Statistics (UNCEBD). As indicated in the Terms of Reference of DSLN (see Annex III of the 2023 report of UNCEBD to the Statistical Commission), the DSLN convenes data science leaders from national statistical offices, including the leaders of regional and sector hubs for big data, to enable strategic discussions on big data and data science issues, to share experiences and knowledge, to strengthen decision-making and leadership at national statistical offices and to accelerate research collaboration and technical partnerships. DSLN will provide input for the mandate of the UNCEBD in the formulation of an overall vision, as well as coordination, guidance, prioritization and direction in the area of data science for official statistics. The expected outputs are concrete strategies and roadmaps to achieve coherent and integrated programmes of work in this area.

DSLN conducted a first Sprint at the end of March 2023 in which it addressed the role of data science in official statistics, and more specifically also the role of the DSLN for the statistical community. It was concluded that DSLN should give guidance on how to promote a culture of innovation within NSOs and address institutional constraints and barriers to entry. DSLN should also give guidance on how to strengthen international collaboration in data science; and how to implement data science into the statistical production process, while taking into account the different levels of expertise of NSOs.

The 2nd Sprint of the DSLN took place on 23 and 24 January 2024 in Dubai, UAE, as part of the International Seminar on Data Science and Statistics for the Statistical and Transport Communities. This Sprint provided an opportunity to have an in-depth discussion with experts from NSOs and international organizations with the goal of developing a comprehensive outline of a "Data Science Playbook" for official statistics and kick-starting the process of drafting it.

Through the presentation of case studies, break-out sessions, and plenary discussions, participants delved into the nuances of integrating data science into National Statistical Offices (NSOs).

Why a Data Science “Playbook”?

The participants in the sprint agreed that the playbook should address the need for practical guidance for National Statistical Offices to take full advantage of the opportunities brought about by advances in the field of data science, with the aim of enhancing the production of official statistics through innovative data sources, technologies and methods. It will help NSOs leverage the transformative potential of data science in enhancing all aspects of the statistical business process, from data collection to analysis and dissemination.

In particular, the playbook will serve as a guide to integrate data science effectively into the operations of official statistics, enhancing capabilities for rapid response, data integration, and leveraging new tools and technologies that have the potential of transforming the way in which national statistical offices are organized.

Annotated Outline

Section 1: Leveraging basic tools of data science for immediate efficiency gains in NSO operations

1.1 Introduction to data science in NSOs

This chapter will explain why guidance on Data Science for Official Statistics is needed. Over the last 10 years – and especially during the COVID pandemic – statistical offices in both developed and developing countries have been given additional tasks to deliver timely indicators and timely insights into emerging issues. The advent of the Digital Age included the automatic generation of a continuous flow of digital data through sensors, mobile devices or 5G equipment. The community of official statistics was urged to explore how this wealth of data could be responsibly used to serve society. Data science emerged as a new discipline which could provide statistical offices with the tools to harness big data and other non-traditional data sources.

This introductory chapter will also provide a definition of data science within the context of official statistics, emphasizing its complementary nature to traditional statistical work. It will establishing a clear connection between data science and the Fundamental Principles of Official Statistics and the Generic Statistical Business Process Model. The chapter will also provide an overview of the foundational data science skills, technologies and methods that are most relevant to the work of official statisticians.

Additional recommendations from 2nd Sprint discussion:

- De-mystify data science in the context of official statistics.
- Distinguish between data engineering and data science roles.
- Define a number of good case studies by their ability to communicate effectively and serve as a proof of concept, providing a framework for experience sharing.

1.2 Automation and process optimization

This chapter will highlight the value of automation through the lens of Reproducible Analytic Pipelines (RAPs), in terms of enhanced efficiency, increased resilience and business continuity, as well as improved transparency, trust and quality of statistical outputs.

Through concrete examples, the chapter will illustrate how data science projects have allowed teams in National Statistical Offices to succeed in automating specific tasks and gain efficiencies by reducing the risk of errors as well as the time and cost involved in delivering statistical outputs.

The chapter will also provide advice on how to maintain trust in the outputs of processes that have been automated, by contrasting them with the results of traditional methods. It will propose specific metrics that can be used for measuring the impact of data science efforts geared towards process optimization, and discusses the importance of reproducibility, modularity, and version control in building trust and ensuring quality.

Additional recommendations from 2nd Sprint discussion:

- Prioritize organizational pain points and focus on quick wins that can demonstrate early success.
- Emphasize how efficiency gains allow for reallocating resources to tasks that add value, such as interpretation and communication.
- Reference the HLG MOOS ML Group framework for comparing ML efficiency with traditional methods.
- Identify opportunities for process reengineering to enhance efficiency and reduce human error.

1.3 Adopting new tools and technologies

This chapter will provide guidance on how to facilitate the adoption of data science tools and technologies by NSO teams, focusing on the value that data science projects bring to the production of official statistics.

The chapter will emphasize the use of tools and technologies that have immediate practical application, geared to the address the specific needs and priorities of NSOs. It will present criteria that National Statistical Offices can use to evaluate the suitability of new data science technologies and methods, focusing on the desired attributes of such tools.

In addition, the chapter will propose specific approaches that NSO teams can follow to facilitate the process of adopting new tools and technologies for the automation and optimization of statistical operations, including participation in platforms for sharing open-source software.

Additional recommendations from 2nd Sprint discussion:

- Give preference to tools and technologies that have broad, practical applicability.
- Aim to identify technologies that support the diverse needs of NSOs that have different levels of maturity.
- Implement parallel testing of traditional and new methods to ensure consistency in outcomes, acknowledging the challenges of time-series comparability and potential discrepancies.
- Recommend criteria for selecting tools, favoring either a modular or integrated approach based on project needs.

1.4 Cultural and process change management

This chapter will stress the importance of change management, and outline strategies such as mentoring and management of time-bound data science sprints. It will provide practical guidance on how to secure commitment of NSO staff time, as well as to effectively communicate the benefits of embracing data science innovations to empower teams to be more efficient and to achieve new results.

The chapter will also stress the need to enable horizontal data sharing across different organizational units, and the pursuit of user-centric data innovation. It will emphasize the need for senior sponsorship, breaking silos through cross-functional teams, and fostering an environment that allows for collaboration, experimentation, and learning from failure.

Additional recommendations from 2nd Sprint discussion:

- Advocate for cultural change through shared journeys
- Adopt an agile and iterative approach to project management.
- Encourage experimentation and accept failure as an opportunity to promote learning and experience sharing.

Section 2: Generating additional insights in response to emerging needs

2.1 Rapid response and data integration

The chapter will discuss the role of data science in enabling NSOs to respond more rapidly to emerging needs through efficient data management and integration practices.

This chapter will discuss the importance of having prerequisites in place such as RAPs, data lake infrastructure, common identifiers, geocoded data and data organized according to statistical standard classifications, so data science teams can have quickly access to analysis-ready data.

Additional recommendations from 2nd Sprint discussion:

- Clarify the definition and expectations of rapid response within the context of official statistics.
- Determine the NSO's role in crisis situations, balancing timeliness and quality in data provision.
- Ensure NSOs support crisis response effectively, with a focus on serving those managing the crisis.
- Develop agile infrastructure, such as a dedicated center of excellence, to enable swift NSO responses to emergencies.
- Recognize data integration as a distinct and critical topic for detailed exploration.
- Prioritize capacity development for data integration via common geographies and geocoding.
- Provide guidance on how to implement cross-domain interoperability standards and develop machine-actionable metadata.

2.2 Skill development and capacity building

This chapter will highlight the necessity of developing data science skills within the NSO through training, mentorship, secondments, and other types of on-the-job training. It will highlight the need for collaborative approaches to skill-building, as well as the importance of building new skills across the entire organization.

Additional recommendations from 2nd Sprint discussion:

- Prioritize investment in people over investment in software and technology.
- Build cross-functional data science teams with members who possess T-shaped skills—meaning they have deep expertise in a particular area of data science along with a broad base of general knowledge in other related disciplines.
- Prioritize soft skills, especially in communication, alongside data science and subject matter expertise, to make data insights accessible and understandable.
- Train middle managers on both the technical operation and the value proposition of the system.
- Utilize existing resources, such as the Big Data Maturity Matrix and examples like Statistics Poland's Data Science Academy for staff upskilling.
- Implement incentives to encourage ongoing learning and engagement with data science.
- Highlight mentorship as a key method for effective training and skill development. Make mentoring others and transfer of knowledge part of the job description of every data scientist.

2.3 Data science partnerships and collaboration

This chapter will provide guidance on how to build partnerships between the National Statistical Offices and external data providers, academia, and international organizations around data science projects and initiatives for the public good.

The chapter will draw lessons from real-world data science partnerships established by NSO, underscoring the value of learning from both successes and failures, and provide practical recommendations for fostering a culture of collaboration and shared learning.

Additional recommendations from 2nd Sprint discussion:

- Underscore the potential leadership role of NSOs in government-wide data science initiatives.
- Foster partnerships with academia and the private sector by offering mutual benefits such as data access and collaborative research opportunities instead of financial compensation.
- Encourage joint ventures by highlighting shared value creation, such as enhancing public knowledge and developing innovative solutions, to offset direct funding needs.

2.4 Quality frameworks

This chapter will emphasize the need for systematic quality assurance in rapid data science projects, providing practical recommendations to effectively assess the quality of new data sources, the fitness for purpose of the methods and algorithms used, as well as the quality of the statistical outputs obtained from them. Building on existing quality frameworks, the chapter will highlight the quality attributes that are particularly relevant to assess the outputs of innovative data science projects, such as explainability, reproducibility, comparability and auditability.

Additional recommendations from 2nd Sprint discussion:

- Promote openness and transparency of data quality assessments with a view to enhance trust, reproducibility and continuous improvement in the application of data science for the production of statistics.

2.5 Resource mobilization and leadership support

This chapter will provide guidance for securing leadership support and resources for data science projects conducted by statistical organizations in response to emerging needs. This includes recommendations such as making sure that such data science initiatives have clearly defined scope, and metrics that demonstrate their value to senior management.

Additional recommendations from 2nd Sprint discussion:

- Senior management should ensure staff can dedicate ample time to data science initiatives, allowing them to actively engage in projects and learn from experts, fostering a culture of continuous improvement and innovation.
- Propose metrics beyond cost efficiency to measure gains.
- Clearly identify beneficiaries based on key performance indicators (KPIs).

Section 3: Full transformation of official statistics through digitalization

3.1 Strategic leadership and support

This chapter will delve into the strategic leadership needed to navigate the full transformation of NSOs through data science, ensuring alignment with the overarching mission of the organization (as reflected, for instance, in the national statistical law or in the national strategy for the development of statistics), and ultimately with the Fundamental Principles of Official Statistics. It will highlight the need for strategic leadership to navigate the integration of data science and official statistics, ensuring full digital transformation of national statistical offices.

The chapter will also emphasize the importance of multi-layered leadership using the RACI framework to delineate responsibilities and empower contributors at all levels. This chapter makes the case for NSO leadership to embody the gradual transformation of statistical production processes through the systematic adoption of data science methods and practices.

The chapter also cover how strategic leadership is required to advocate for legal framework adjustments, improve communication of use cases, and foster collaborations between the NSO and other major players of the national data ecosystem, in alignment with the broader government digital agenda.

Additional recommendations from 2nd Sprint discussion:

- Incorporate change management principles from relevant literature.
- Ensure middle management effectively communicates change initiatives.
- Align strategic KPIs with transformation goals for measurable outcomes.
- Embed innovation as a key responsibility in job descriptions.
- Offer rewards and public recognition for innovative achievements beyond traditional tasks.
- Clearly define roles, responsibilities, and resource allocation to streamline transformation efforts.

3.2 Advanced data science tools and methods

This chapter will explore and provide guidance on the implementation of advanced data science tools and methods across all statistical production activities. It will present examples of advance data science use cases, from initial prototyping and experimentation and up to implementation into production.

Among other things, the chapter will highlight the need for modern practices such as DevOps, SecOps and ML Ops as advanced data science capabilities in NSOs.

Additional recommendations from 2nd Sprint discussion:

- Discuss the importance of open-source tools considering total cost of ownership.
- Promote sharing of codes and algorithms

3.3 Technology implementation and integration

This chapter will provide guidance on technology implementation and integration, emphasizing the need for a modern data architecture and IT infrastructure that supports the implementation of data science at scale. Topics to be covered include, among others, leading and contributing to open-source projects, metadata management, cybersecurity and privacy-enhancing technologies, as well as APIs, containerization and orchestration.

The chapter will also provide guidance to help data science teams within NSOs collaborate closely with their IT department, to meet the unique infrastructure needs of data science projects in the context of official statistics. This includes, among other aspects, providing guidance to help navigate the complexities of cloud migration initiatives, avoiding vendor lock-in, and ensuring compliance with organization-wide data governance policies.

Additional recommendations from 2nd Sprint discussion:

- Establish a clear pathway for transitioning AI prototypes into production.
- Create a secure environment to safely experiment with restricted tools.
- Engage the entire organization in discussions about optimal integration strategies and necessary platforms and tools.

3.4 Data quality management and security

This chapter will reinforce the need for robust data quality management and security practices as NSOs increasingly rely on innovative, big data sources and data science analytical methods. Building on existing data quality frameworks, the chapter will provide additional guidance on topics such as responsible and ethical data practices, and identify areas where data quality frameworks may need to be further developed.

Additional recommendations from 2nd Sprint discussion:

- Ensure cloud migration includes clear data control measures to counter vendor lock-in, drawing lessons from Canada's experience.
- Evaluate and plan for data portability as part of cloud migration to maintain flexibility and control over data.

Section 4: Cross-sectional themes

4.1 Cultural change

This chapter will address the cultural shifts required to integrate data science into established statistical practices, promoting an environment of continuous organizational learning and adaptation. The chapter will reinforce the importance of cultural shifts that enable the mainstreaming of new data science practices.

The chapter will reiterate the necessity of cultural change for the successful integration of data science, emphasizing collaboration and user-centric approaches. It will provide practical recommendations on how to address resistance to change, including the provision of clear communication about the benefits of NSO modernization and the introduction of new data science practices in the production of official statistics.

The chapter will explore how creating a community of practice around practical use cases can help demonstrate value and mitigate risk aversion inherent in official statistics.

Additional recommendations from 2nd Sprint discussion:

- Address tension between compliance and innovation by aligning incentives of NSO staff with overarching organizational goals and mission.
- Enhance transparency and education around new data sources, methods and technologies to overcome fear and enable informed decision-making.

4.2 Institutional arrangements

This chapter addresses the importance of establishing institutional frameworks to bolster data science initiatives in national statistical offices. It outlines how structuring teams, projects, and partnerships can facilitate collaboration and align incentives among data scientists, statisticians, and other stakeholders within and outside the NSO, leading to a coherence data science strategy for official statistics. Additionally, the chapter explores the development of supportive arrangements like the formalization of cross-functional teams to drive data science innovation beyond traditional organizational boundaries.

Additional recommendations from 2nd Sprint discussion:

- The playbook must acknowledge the diversity of governance systems across NSOs, emphasizing that a one-size-fits-all approach is ineffective.
- International collaboration, particularly under UN frameworks, is crucial for leveraging global knowledge and resources.

Take-aways from the sprint and the way forward

Participants in the spring reiterated the expectation that the playbook will play an important role in guiding NSOs through the adoption and integration of data science for the production of official statistics. Many of them highlighted the iterative nature of process that will develop playbook, and encouraged contributors to collaborate to ensure that the playbook is updated on an ongoing basis, continuously introducing improvements based on NSO experiences and the evolving data science landscape. Therefore, the Data Science Playbook for official statistics should serve as a living document, continuously updated to reflect new insights, technologies, and practices that emerge from the NSOs' ongoing transformation journey.

The discussions underscored the necessity of integrating data science into NSO operations not as a substitute but as a complement to traditional statistics. Participants in the sprint called for a balanced approach between leveraging new tools and technologies and managing cultural and process changes, aiming for a comprehensive transformation that enhances efficiency, quality, and the relevance of official statistics.

Further, participants in the sprint underlined the critical need for NSOs to evolve in response to the digital age's challenges and opportunities. This evolution requires strategic leadership, advanced tools and methods, robust technology infrastructure, and adaptive data quality management practices.

An editorial board will be established to oversee the development and iteration of the playbook, ensuring it remains relevant and effective for NSOs globally. The initial draft will maintain the current structure, with a subsequent review to refine and possibly reorganize the content based on feedback. The playbook should be delivered in a collaborative platform like GitHub or GitLab to promote reproducibility, versioning, and international collaboration.

Additional recommendations from 2nd Sprint discussion:

- Selecting universal use cases, like the Consumer Price Index (CPI) or migration statistics, can provide relevant guidance for all countries
- The playbook should present a positive outlook, focusing on opportunities for using data science in official statistics and providing practical solutions to challenges.
- Consider incorporating in the playbook a dedicated chapter on how to effectively integrate geospatial and statistical data and methods as part of new data science workflows.