Open Data Contract Standard (ODCS) as integral part of the Universal Metadata Layer

***Paula Hartung**, Statistics Iceland,* paula.hartung@hagstofa.is

***Sigurður Jónsson**, Statistics Iceland*, sigurdur.jonsson@hagstofa.is

Abstract

Statistics Iceland is adopting the Open Data Contract Standard (ODCS) [1] as a foundational metadata layer to modernize and streamline its data operations. This paper outlines how ODCS is being integrated into the institution's technical ecosystem, including data-ingestion pipelines, SQL Server environments, the data build tool (DBT) [2], and GitLab/GitHub–based version control workflows.

ODCS is a machine readable, version controlled, code-as-documentation approach to defining data contracts that aligns closely with both current and future data infrastructure needs at Statistics Iceland. By defining all incoming datasets at the point of collection and possibly in collaboration with data providers according to their structure, content, and semantics, ODCS supports systematic measurement and improvement of data quality.

The implementation of ODCS within a contract-first approach can facilitate several key benefits: enhanced data transparency, improved data quality, improved connectivity between datasets, enhanced knowledge transfer, centralized metadata governance, and increased opportunities for automation.

The paper describes the implementation timeline, the challenges encountered along the way and the tools that have proven most effective. Furthermore, it will be demonstrated how register data can be processed for statistical systems.

Keywords: ODCS, Dataflow, Metadata, Digital Transformation, Data Governance

1. Introduction

National statistical offices (NSOs) face growing pressure to modernize data infrastructure to produce higher-quality outputs more efficiently and with greater transparency. A persistent challenge is that knowledge about datasets tends to reside with specific staff rather than being embedded in systems: a fragile arrangement, vulnerable to turnover and organizational change.[3] Without a shared metadata layer, digital transformation risks becoming a fragmented effort, where modern tooling is applied to data that remains poorly understood and inconsistently described.[4]

Data contracts are formal agreements that define the expectations and standards for data quality, format, structure, semantics, and usage, serving as a governance tool to ensure data meets specific quality criteria irrespective of its source.[6] The Open Data Standard (ODCS) provides a machine-readable, version-controlled, open framework for formalizing exactly these agreements.[7] Adopted within a contract-first approach, ODCS has the potential to serve as the backbone of a universal metadata layer: a shared, structured description of every dataset entering an organization's systems.

Statistics Iceland is gathering experience on this journey. The concept and technical architecture have been established, foundational tooling is in place, and the first data contracts have been drafted. This paper describes the rationale, the technical ecosystem, early observations, and the challenges encountered — contributing to the growing NSO conversation about how open standards and modern data engineering can drive sustainable digital transformation.

2. The Open Data Contract Standard and the contract-first approach

The Open Data Contract Standard (ODCS) is a machine-readable, version-controlled specification written in YAML format. It is a formal agreement between data producers and consumers that describes a dataset's structure, content, semantics, and quality expectations, see Table 1.[1] Its governance under the Linux Foundation ensures long-term vendor neutrality and community-driven development.[7] The contract-first approach places the definition of a dataset — its structure, semantics, and quality rules — at the center of the ingestion process. The core principle is that no data flows into Statistics Iceland data pipelines without being formally defined. Rather than describing data retrospectively after it

Section	Content
Fundamentals	Identifier, name, version, owner, description, domain
Schema	Logical and physical data model; field names, types, descriptions
Data Quality Rules	Completeness, validity, uniqueness, and referential integrity checks
Service-Level Agreements	Freshness, latency, and availability expectations
Server / Infrastructure	Physical location, format, access method
Custom Properties	Organization-specific extensions (e.g., statistical classifications)

Table 1: Open Data Contract Standard (ODCS) sections and their content.

has already entered the system, a contract is written to establish expectations upfront. The contract serves as authoritative definition before data enters production systems.

This approach provides structural stability: by decoupling data flows from changes at the source, updates to how a data provider delivers data can be absorbed at the contract level, without cascading disruptions to downstream statistical processes. This builds the base for cross-dataset consistency and automation.

3. Contract lifecycle

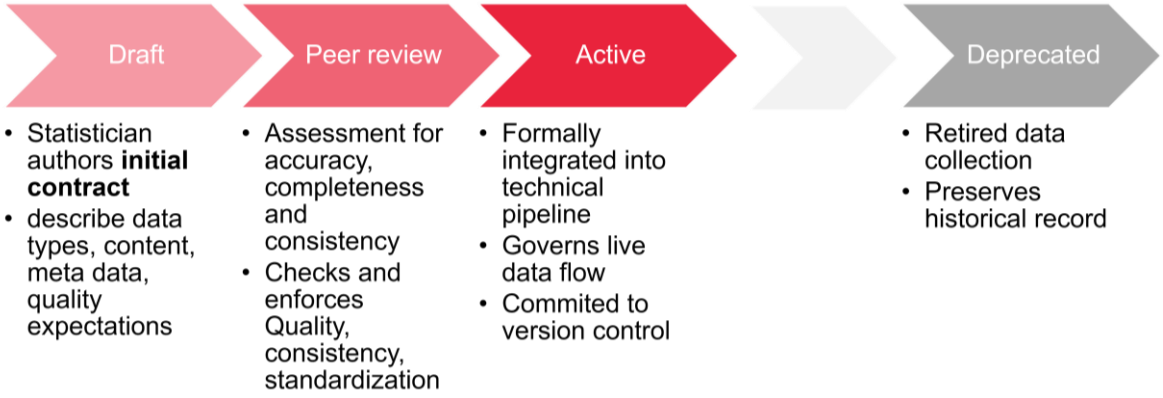

Figure 1: The four-stage lifecycle of data contracts at Statistics Iceland, with a planned fifth stage for external provider collaboration.

The lifecycle of a contract at Statistics Iceland follows a defined sequence of stages designed to ensure quality, consistency, and traceability, as shown in Figure 1. Currently, contracts are internal documents which are authored and maintained by Statistics Iceland's own staff. Collaboration with external data providers on contract definitions is an intended future direction, with the potential to align data deliveries at the point of collection.

4. Technical Ecosystem

The contract-first methodology is supported by a set of interconnected tools that translate the formal contract definitions into operational data pipelines, see Figure 2.

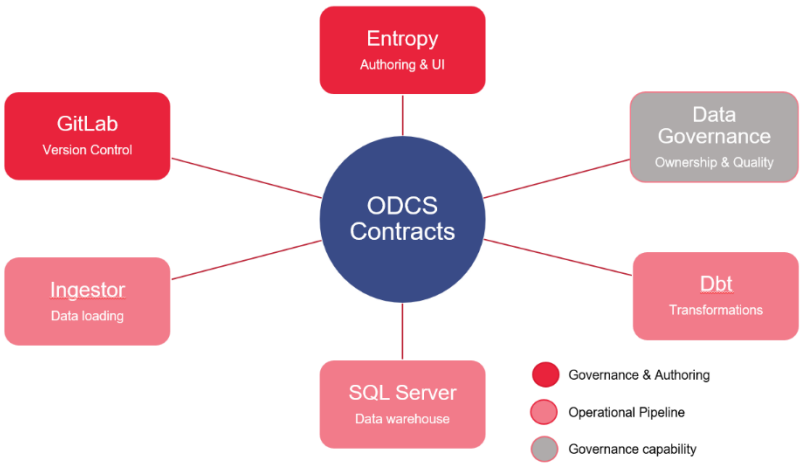

Figure 2: The technical ecosystem centered around the ODCS contracts, including Governance of the contracts in red, Operational pipeline derived from them in light red.

Contracts are authored and managed through Entropy Data [9]. From there data transformation models and automated quality tests are generated directly from the contract definitions, ensuring that downstream processing logic stays aligned with the formal specification. All active or historical contracts are stored as YAML files in a GitLab repository, providing full version control, audit history, and a structured peer review process through merge requests. This treats contracts as code — versioned, reviewable, and traceable over time.

Incoming data is loaded into data pipelines through a dedicated ingestion tool (Ingestor) that uses multiple collection adapters. The contract defines what to expect. The Ingestor validates incoming data against these expectations prior to loading.

Data transformation within the data pipelines is handled by DBT (data build tool),[2] whose models and tests are generated from the contract specifications, creating a direct and maintained link between the formal data definition and the transformation logic applied to the data.

5. Population Register: A Pilot Case

This section describes the process of drafting an ODCS contract for the population register. This foundational dataset contains personal identifiers, name, gender, and birth and death dates that underpin a wide range of statistical outputs. It arrives as two tables: a registry and a disappearance table (Tables 2).

registry	
SSN	String
Name	String
BirthDate	Date
Gender	Integer

disappearance	
SSN	String
Name	String
Gender	Integer
DeathDate	Date

Tables 2: Overview of two tables that are content of the Data Contract "Individuals Example".

These tables contain largely overlapping data, differing in one key date column summarized in Table 3. A simple join in the data warehouse connects them after ingestion. A notable advantage of this structure is resilience to upstream change: if the provider modifies their data model, the impact is contained and traceable at the contract level.

Field names	Data Type	Description	Table(s)
SSN	varchar(10)	Unique Social Security Number, 10 digits	Both
Name	varchar(100)	Legal name of Individual	Both
Gender	integer	Gender code; 1= male, 2= female, 3= boy under 18 years, 4 = girl under 18 years, 7 = gender neutral/other, 8= gender neutral under 18 years	Both
BirthDate	date	Date of birth	registry
DeathDate	date	Date of death (nullable)	disappearance

Table 3: Population register field summary (across two tables).

The contracting process began with a statistician authoring a draft: defining the expected field names, data types, and descriptions alongside data quality rules covering completeness and validity of key identifiers.

A minimal example contract is publicly available on GitHub.[8] Once finalized and activated, the contract will inform the Ingestor configuration, provide the basis for DBT model generation, and serve as the authoritative reference for anyone working with population register data in the future.

6. Pilot Scope and Implementation Timeline

Initial contract development took several weeks as the technical environment was being established. As tooling matured, the time required to draft and implement a contract has decreased significantly — data can now be collected within days of contracts being finalized. Peer review remains a bottleneck, though it serves a

necessary function: enforcing consistent naming conventions and data type standardization across all contracts.

The pilot has progressed from concept and architecture through tooling integration to active contracts governing the first base datasets. The next phase extends contracting to derived datasets and, in time, to collaboration with external data providers.

7. Challenges

Several challenges have emerged during the pilot., see Figure 3.

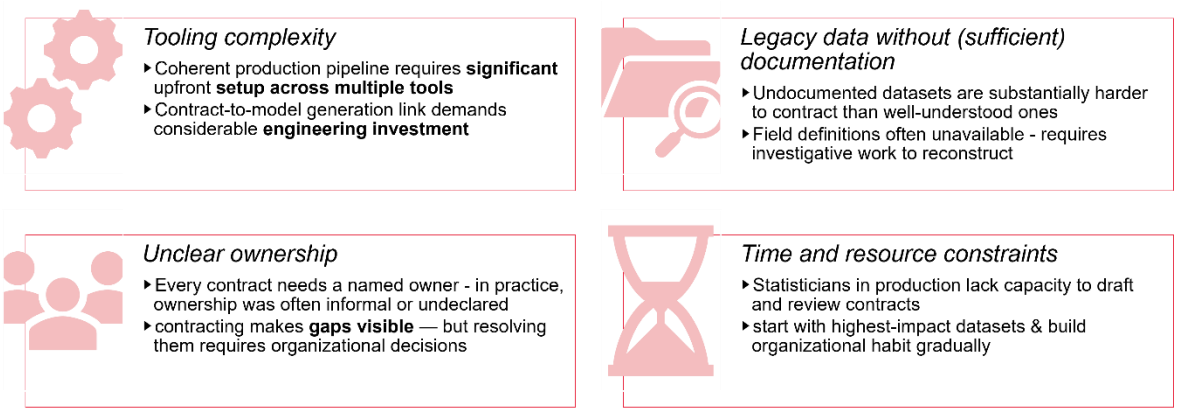

Figure 2: Selection of faced challenges. Tooling complexity and time constraints are ongoing challenges: the integration of multiple components demands significant upfront engineering investment, and statisticians in production need dedicated capacity for contract authoring and review. Therefore, rollout must be sequenced carefully: starting with the most impactful reference datasets and building organizational habit gradually, rather than attempting comprehensive coverage at once. Data ownership, by contrast, is being progressively resolved: the contracting process makes unclear ownership visible, which then enables the organizational decisions needed to address it.

8. Results and Conclusions

8.1. Early observations from the pilot

The most immediate benefit has been in collaboration. Previously, communication between engineers and statisticians was largely informal: engineers interpreted data deliveries as best they could; statisticians described needs in ways that were not always technically precise. The contract-first approach replaces this with a shared, structured framework and communication vehicle: statisticians express requirements formally, engineers work from a clear specification rather than institutional memory.[3] Peer review surfaces ambiguities before they reach downstream processes.

A second gain is institutional knowledge retention. Metadata is often unevenly distributed — maintained by individuals rather than embedded in systems. [3,4] Contracts make this knowledge explicit, version-controlled, and machine-readable. A natural byproduct is a shared data glossary: as field definitions are agreed upon across contracts, a consistent institutional vocabulary emerges organically.

8.2. Metadata as the foundation of digital transformation

Digital modernization is often framed as a technology challenge. Statistics Iceland's experience confirms that technology, while necessary, is not sufficient.[4] Without a shared metadata layer, modern tooling risks being applied to data that remains inconsistently described. Placing formal data definitions at the center of the process, before ingestion and ideally before first delivery, addresses this directly.

NSO Data arrives from a variety of sources (registers, administrative sources, survey instruments, companies …) and follow their own conventions. Without a shared metadata framework, reconciling these requires costly, case-by-case effort.[5] A contract-first approach shifts harmonization upstream, building the data glossary as a byproduct.

The organizational dimension is equally critical.[10] Unclear ownership, undocumented legacy data, and time constraints (see 7. Challenges) are not technical failures. They are governance gaps that ODCS makes visible and forces organizations to resolve. The contract becomes a productive forcing function: writing them forces decisions and documentation that would not otherwise exist.

8.3. Conclusion and next steps

Statistics Iceland is at pilot stage: architecture established, tooling in place, first contracts being finalized. The observations reported here already confirm that digital transformation is as much an organizational challenge as a technical one. ODCS are a crucial step to building a coherent Metadata layer and maintaining harmonization, consistency and standardization throughout the whole institution.

The contract-first approach shifts data governance from reactive to proactive. Next steps ahead are: activate contracts for core reference datasets, implement derived datasets, and begin external provider collaboration.

9. Appendix: List of references

[1] Bitol (2025). Open Data Contract Standard (ODCS) v3.1.0. LF AI & Data Foundation. https://bitol-io.github.io/open-data-contract-standard/v3.1.0/

[2] dbt Labs (2024). dbt — Data Build Tool. https://www.getdbt.com/

[3] UNECE/HLG-MOS (2016). Handbook on Managing Statistical Data and Metadata: Chapter 14.3. https://projects.officialstatistics.org/hb-mgnt-org-nss/

[4] Simeoni, F. et al. (2016). How Metadata Systems Support the Modernization of Statistical Offices. Proceedings of the International Conference on Establishment Surveys (ICES-V). American Statistical Association. https://ww2.amstat.org/meetings/ices/2016/proceedings/042_ices15Final00156.pdf

[5] UNECE/HLG (2024). Data Governance Framework for Statistical Interoperability (DAFI). United Nations Economic Commission for Europe. https://unece.org/sites/default/files/2024-03/HLG2023 DAFI Final_0.pdf

[6] ThoughtSpot (2024). What are data contracts? https://www.thoughtspot.com/data-trends/data-governance/data-contracts

[7] Entropy Data (2023). Open Data Contract Standard — Overview. https://www.entropy-data.com/learn/open-data-contract-standard

[8] Hartung, P. & Jónsson, S. (2025). Minimal Example ODCS Contract for Population Register. GitHub. https://github.com/surfing-in-the-sun/odcs_population_example

[9] Entropy Data (2024). Entropy — Data Contract Platform. https://www.entropy-data.com

[10] Wasser, J., Kumara, I., Monsieur, G., Van Den Heuvel, W.J. & Tamburri, D.A. (2026). Data Contracts in Data Mesh: A Systematic Gray Literature Review. In: Shishkov, B. (ed.) Business Modeling and Software Design. BMSD 2025. Lecture Notes in Business