Data Veracity Assurance in Data Spaces

Bertalan Zoltán Péter[0000−0002−5577−1369], Nada Akel[0009−0004−0184−4093], László Gönczy[0000−0002−5317−2311], and Imre Kocsis[0000−0002−2792−3572]

Department of Artificial Intelligence and Systems Engineering
Budapest University of Technology and Economics,
Műegyetem rkp. 3, 1111 Budapest, Hungary
bpeter@edu.bme.hu

Abstract Data spaces are emerging as a transformative approach to data management, sharing, and governance. A data space ecosystem typically has entities that produce, process, and consume data as well as contract, consent, and identity management services. The aspect of data quality, or, more generally, data veracity, is often overlooked. However, it is becoming increasingly clear that for healthy, functioning data spaces, ensuring the veracity of the data shared will be an essential service.
In this paper, we introduce the novel notions of data veracity and data veracity assurance: a holistic notion of data quality in data spaces and a conceptual framework for its end-to-end assurance in data space service chains. Our approach is compatible with the leading data space frameworks, supports personal data sharing, and is being implemented as a set of open source building blocks. We define two major complementary regimens of data veracity assurance: presentations of trust-based Attestations of Veracity and the trustless checking of formal Proofs of Veracity. We employ verifiable credentials to bridge the gap between technical notions of data quality and legally binding contracts in a systematic way. Verifiable credentials also enable the application of our proposed framework in radically different trust settings, from closed, intermediary-orchestrated data spaces to fully decentralized ones and cross-data-space data exchange scenarios.

Keywords: data space · data veracity · data quality · verifiable credentials.

1 Introduction

In 2020, the European strategy for data set out the path to the creation of common European data spaces in key sectors, with the long-term goal of creating a single, interconnected market for data. There is now more or less consensus on the insight that the intents and obligations of exchanging data have to be captured in explicit *data contracts;* all the more so that this is, especially with the proliferation of AI model training, a more general trend in all kinds of cross-organizational data exchanges. While open data contract languages have been proposed, with an emphasis on describing the structure of the data and aspects

such as consent to data use in personal data scenarios, we believe that the fact that a data contract shall be a *symmetric relation* is not receiving enough attention. While the data provider has to have the right to prescribe the ways their data is used, and these obligations have to be enforced, the data consumer has rights, too: specifically, the right to receive data that meets their quality requirements for effective data use.

To that end, in this paper, we introduce the notion of *data veracity* in data spaces: a term borrowed from the classic language of Big Data processing and which expresses that data consumers in data spaces need data which is 'right' in a holistic sense [16]. We define data veracity assurance (DVA) as a key, and in its explicit formulation novel, function in data spaces. We interpret DVA as the set of mechanisms and procedures that ensure that parties comply with Veracity Level Agreements (VLAs) and enable remedial or punitive actions on breach of agreements. DVA and VLAs as terms intentionally rhyme with service level assurance and service level agreements (SLAs).

In our current technological milieu, we believe two fundamentally different approaches to VLA exist. The first is based on trust and direct property checking: data providers or trusted third parties can *attest* to the fulfilment of VLAs, and consumers can accept these attestations based on their trust in the provider or third party. We coin the term *Attestation of Veracity (AoV)* for such statements. Alternatively, key emerging technologies – such as proof-carrying data and zero-knowledge-proofs – enable novel, trustless models, where VLA fulfilment can be assessed by checking mathematically rigorous proofs on properties of the data. We call these statements *Proof of Veracity (PoV).*

In the EDGE-Skills EU project, we are designing and implementing DVA as an open-source building block, which is interoperable with the more extensive Prometheus-X data space ecosystem [14]. Importantly, we are designing VLAs, AoVs and PoVs as verifiable credentials (VCs), and their exchanges as verifiable presentations (VPs). While this approach has key technical and interoperability advantages, it is possibly even more important that it facilitates bridging the gap between legal agreements and technical assertions, setting the stage for VLAs as enforceable, e-IDAS-compliant Electronic Attestations of Attributes [9].

The rest of this paper is structured as follows. In the next section, we give an overview of the related work in data spaces, data quality, and veracity assurance. Then, in section 3, we present the foundational conceptual model of our envisioned DVA building block, followed by the technical implementation of those concepts in section 4. Finally, in section 5, we elaborate on the legal implications of our work and conclude in section 6.

2 Related Work

While the holistic concept of data veracity is not mentioned, the idea of data quality has been included in some data space standards and reference models, even outside the data space world – these can also be used to define quality requirements in data spaces.

2.1 State-of-the-Art in Data Spaces

A data space is an interoperable framework based on common governance principles, standards, practices, and enabling services that enable trusted data transactions [3]. Data quality is a key issue in data space frameworks for several reasons. Incorrect data increases the effort required for data preparation. Poor data quality reduces trust between partners in a data ecosystem. Additionally, low-quality data harms inter-organizational business processes [1].

The International Data Spaces Association Reference Architecture (IDS-RAM) addresses the importance of data quality by enabling participants to assess data source quality through publicly available metadata [11]. The Data Spaces Support Center (DSSC) also supports data quality by offering value-added services [4]. Complementing this, Gaia-X [10] promotes open innovation through sovereign data sharing based on trust between all involved actors through the Gaia-X Data Product concept and Operational Model.

In the EDGE-Skills project, we followed the guidelines established by the International Data Spaces Association (IDSA) and Gaia-X for developing the Prometheus-X (PT-X) ecosystem [15].

2.2 Data Quality and Veracity Assurance

Data quality refers to the degree to which a set of inherent data characteristics fulfils requirements. ISO 8000-61 [12] serves as a process reference model for data quality management, enabling organizations to continuously improve data quality. It outlines the characteristics of information and data that determine its quality and offers approaches to managing, measuring, and enhancing that quality. Another standard related to data quality is the ISO 25000 family [13], which standardizes and unifies software product quality standards. Within this series, ISO/IEC 25012, titled *Data Quality Model,* and ISO/IEC 25024, which deals with measuring data quality, are particularly noteworthy.

Data contracts form a valuable approach to support data veracity in data spaces, defining an explicit agreement between data providers and consumers to help ensure the shared data behaves as expected and meets agreed standards. Existing data contract languages, such as the Data Contract Specification (DCS) [5] or the Open Data Contract Standard (ODCS) [2], already capture some fairly technical aspects of data quality; in contrast, meeting veracity requirements is a composite concept, ranging from the appropriate 'shape' of the data through statistical, timeliness, and sampling properties to semantic appropriateness at the domain and the intended business use level.

3 Conceptual Model

Please refer to Figure 1 for an abstract, comprehensive knowledge graph representation of all DVA-related concepts. The software manifestation of these concepts is presented later, in section 4. Due to the frequency of the terms 'data

provider' and 'data consumer', throughout this and the next section, we use P and C as their abbreviations.

Data veracity assurance starts with being able to describe data veracity *requirements* in a standardized way. We refer to the digital documents with such information as Veracity Level Agreements. A VLA targets a specific *data exchange* and consists of *veracity objectives.* In turn, each *objective* targets some *quality aspect* and has an *evaluation scheme* that describes how the objective must be checked.

Based on the requirements in the VLAs, P can create Attestations of Veracity (AoVs) and Proofs of Veracity (PoVs), which can later be verified by C. The key difference between AoVs and PoVs is that while the former is *trust-based,* the latter is *trustless* (in the sense that it comprises a formal cryptographic proof and P cannot generate a PoV for data that does not, in fact, fulfil the VLA).

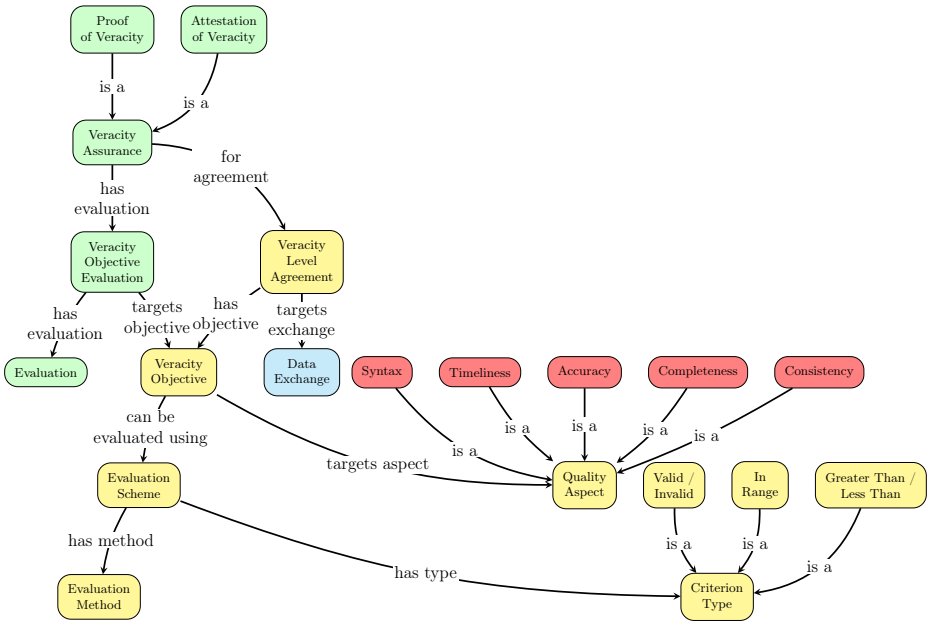

Figure 1. Conceptual model of DVA as a knowledge graph

3.1 Use Cases of the DVA Service

While DVA documents (AoVs and PoV) are created based on VLAs, the DVA building block is not actually the one responsible for the management of VLAs in a data space. Instead, VLAs are embedded in the *contracts.* However, to simplify the creation of VLAs, DVA provides predefined VLA *templates.* These templates are fragments that can be included in a VLA – for example, how many `NaN` values

a dataset contains and how to count them. These templates can be queried by the contract manager service in the data space, and the set of enabled templates can be updated by the data space orchestrator.

The rest of the use cases pertain to the generation and the checking of AoVs and PoVs. What these documents are exactly is described later in this section. P chooses whether to create an AoV or a PoV to attach to the data they produce (or, they may be contractually required to use one or the other). In the case of AoVs, we differentiate between self-attested and third-party versions.

C receives data from P together with AoVs or PoVs which they can *check*. Due to the trust-based nature of AoVs, C may wish to re-evaluate the data based on the VLA to confirm the findings of the attester. As we implement both AoVs and PoVs as VCs, a key part of checking these documents is checking the validity of VCs (which mainly means verifying cryptographic signatures).

Figure 2 shows the use cases of the DVA service associated with various data space actors.

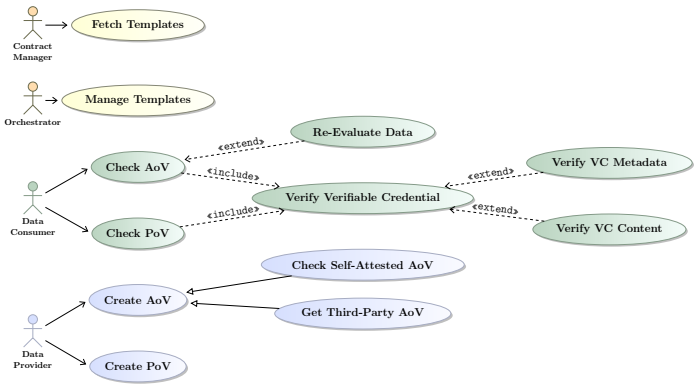

Figure 2. DVA use case diagram

Attestations of Veracity AoVs provide a *trust-based* solution to establish compliance without consumer-side checking. We distinguish two major categories of attestations. Third-party AoVs follow the typical trust-based claim attestation pattern; the usual concerns of the third party being trustworthy by C certainly apply here. Verifying these attestations will typically not go further than establishing the claim's validity as a valid and non-revoked VC.

We also allow for 'self-attestation' by P. Trust-wise, the additional assurance carried by self-attestations (note that a VLA is already a commitment by P) is that P can communicate partial or complete results of their veracity evaluation in such attestations. In general, this can be valuable for 'hard to compute, easy to verify' evaluations (e.g., NP-complete decision problems on the data); but in practice, we expect this mechanism to increase confidence in C through showing compliance for a sample of the data.

Proofs of Veracity PoVs establish compliance through cryptographic, not trust-based approaches – when required and feasible. Such proofs are *sound,* meaning that a cheating P cannot forge a PoV for a piece of data that does not adhere to the VLA's requirements. (Mathematically and succinctly) verifiable zero-knowledge as well as non-zero-knowledge proofs on data have been an emerging field of mathematics in the last two decades, with increasingly rapid development in the last few years. However, as algorithms, standards, software frameworks, and use cases are still evolving, our DVA building block provides a highly extensible framework for PoVs, driven by its use cases.

3.2 Architecture and Behaviour

The DVA building block is a distributed service in a data space – each participant runs their own instance. Normally, DVA service instances are not expected to engage in direct communication with each other; rather, it is the data space connector components that communicate (possibly through a data intermediary), and they in turn interact with DVA components during data exchanges. An overview of our architectural model with the possible interactions between the components can be found in Figure 3. We are considering an additional communication channel that connects DVA instances directly (represented by a dashed grey line in the figure), which would likely be implemented by a distributed ledger, but this is currently left for future work.

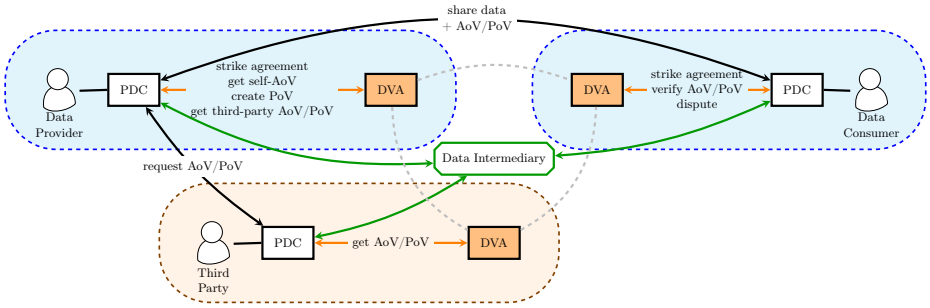

Figure 3. DVA architecture

4 Building Block Design

In section 3, we have introduced the fundamental concepts of our DVA approach – we now present their implementation and technical manifestation.

4.1 VLAs as ODCS documents

Out of the competing data contract standards, we chose to build on ODCS [2], as it seems to be the most mature and to be most in line with our needs in

terms of description of veracity (quality) requirements. In the latest version of ODCS (3.0.2), data quality requirements can be encoded as one of four different requirement descriptors: human-readable *text,* as an element of a predefined attribute *library,* an *SQL* query, or a *custom* description. The last option opens the possibility of describing requirements in various data quality engines and platforms, such as Great Expectations, the Soda Data Quality Platform, and the Monte Carlo Data Observability Platform. At the current stage of development, DVA supports requirements defined as Great Expectations checks, jq queries, and JSON Schema checks.

Great Expectations[1] is supported because it is a versatile, open-source data quality engine. The key limitation of Great Expectations is that it is designed to work with tabular data. However, for generic data space use cases, data may come in various other forms; for example, JSON (or other types of semi-structured) documents, video or audio streams, arbitrary binary data, etc. In the envisioned use case scenarios of PT-X, most of the data is represented by JSON files. To harness the power of Great Expectations in these cases, we have implemented a crude data mapping function in the DVA building block that, given a set of JSONPath – column name pairs, generates tabular data (Pandas data frames) from JSONs files.

The ubiquity of semi-structured data (primarily, JSONs) prompted us to also support the definition of data quality checks that can operate on these data directly. jq is a well-known query language for JSON that can be used in VLAs for simple binary-result checks; e.g., for a check `$.score.raw >= $.score.min`, the input document `{ "score": { "raw": 4, "min": 0 }}` would pass, but `{ "score": { "raw": 0, "min": 0 }}` would fail. Furthermore, as it can be an elementary syntactic requirement that a given JSON document is valid according to some schema, DVA also supports a custom data quality engine referred to as `schema` that can be parameterized with the URI of a JSON Schema (such as https://spec.example.com/xapi-schema.json).

4.2 AoVs and PoVs as VCs

It is clear that AoVs must support *authentication* as their recipients must be able to verify that they originate from a trusted entity (which may be the data provider or a trusted third party). This may not strictly be required for PoVs as a proof of some properties can be considered valid regardless of who created it. However, for simplicity, we handle PoVs and AoVs in the same way in this matter and represent them as digitally signed documents.

An identity management system is required to authenticate an entity reliably. As data spaces are inherently distributed (and, in many cases, decentralized), we believe a decentralized identity solution benefits them. In recent years, 'self-sovereign' approaches have been emerging – with rapidly maturing standards – as a means to handle identities of arbitrary entities and to represent 'claims' that are cryptographically verifiable. Concretely, there are recent W3C recommendations

[1] greatexpectations.io

for decentralized identifiers (DIDs) [18] and for verifiable credentials (VCs) [17]. While VCs were designed with distributed ledgers in mind, they are often well-applicable in contexts without one.

We chose to represent AoVs and PoVs as W3C verifiable credentials (VCs) as these documents fundamentally represent *claims* (e.g., 'the data being sent fulfils the VLA') and VCs offer a standardized, portable way to encode these claims. Employing DID solves the issue of how to authenticate the entities referenced in these claims, such as their issuer. The usage of VCs also implies the necessity of a *verifiable data registry* in the data space, which can be implemented by a blockchain (although it can also be a centralized service).

4.3 Internal Software Architecture

The DVA building block is a distributed service that comprises multiple internal components, primarily, an *API* and a *processing* module. The former exposes a REST API and also services some basic requests, such as the management – meaning CRUD operations – of VLA templates. For more complex operations, such as the generation of AoVs and PoVs and their verification, the API module posts a message to a broker (a RabbitMQ instance).

These messages are handled on a FIFO basis by the *processing* module. The key reasons for the separation of these two modules are to offer better scalability, follow a separation of concerns principle, and to be able to efficiently utilize Python-based data quality engines such as Great Expectations (the API module was implemented in Kotlin). An inherent result of this architecture is that most requests are processed asynchronously. The requesting party receives a response as a HTTP request to a callback address specified in their initial request. Figure 4 shows an overview of DVA's software components with an interaction flow when generating a new AoV.

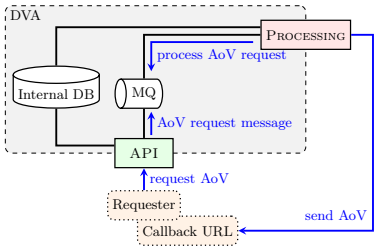

Figure 4. DVA internal software architecture

5 Legal Considerations

The DVA building block establishes trust and legal compliance in data exchange scenarios and must adhere to several recent European regulations that govern

data access, control, and use. It supports the objectives of both the Data Governance Act [7] and the Data Act [6] by ensuring that data is accurate and trustworthy, which enhances trust in voluntary data sharing and contributes to making data more accessible and usable, encouraging data-driven innovation, increasing data availability, and incentivizing the production of high-quality data.

Additionally, DVA supports the goals of the EU AI Act [8] – in particular, when AI systems involve model training, high-quality data helps the high-risk AI systems perform as intended and avoid outcomes such as discrimination that the law prohibits. By supporting accurate and reliable training, validation, and testing data sets, DVA contributes to effective data governance and management, ensuring that data sets are relevant, representative, error-free, and complete for the system's intended purpose.

6 Conclusion

In this work, we have introduced the concept of data veracity assurance to the context of data spaces and the accompanying ideas of Veracity Level Agreement (VLA) for the description of the veracity requirements and Attestation of Veracity and Proof of Veracity for their assurance. We also presented our technical implementation of these concepts within a European data space currently under construction: the Prometheus-X ecosystem. Connecting legal contracts with technical checks using VCs makes data sharing safer and more enforceable.

While our envisioned building block has applications even outside the context of data spaces (especially when supported by blockchains or other distributed ledger technologies), this approach lays the groundwork for building trustworthy data spaces and advancing the vision of a unified European data market.

Acknowledgments. The work of Bertalan Zoltán Péter was partially created under, and financed through, the Cooperation Agreement between the Hungarian National Bank (MNB) and the Budapest University of Technology and Economics (BME) in the Digitisation, artificial intelligence and data age workgroup.

This work has been partially supported by the European Union in the frame of the project European Dataspace for Growth and Education – Skills, short: EDGE Skills (101123471)

Disclosure of Interests. The authors state no competing interests.

References

1. Altendeitering, M., Dübler, S., Guggenberger, T.M.: Data Quality in Data Ecosystems: Towards a Design Theory. In: AMCIS (2022)
2. Bitol, Open Data Contract Standard, (2024). https://bitol-io.github.io/open-data-contract-standard/latest/. Accessed: 2025-03-31.
3. CEN-CENELEC Workshop Agreement, CWA 18125:2024 - Trusted Data Transaction. Tech. rep., Accessed: 2025-01-21. CEN-CENELEC (2024). https://www.cencenelec.eu/media/CEN-CENELEC/CWAs/RI/2024/cwa18125_2024.pdf

4. Data Spaces Support Centre, Value-Added Services, (2024). https://dssc.eu/space/BVE/357076468/Value-Added+Services. Accessed: 2025-03-31.
5. DataContract.com, Data Contract Specification, https://datacontract.com/. Accessed: 2025-03-31.
6. European Commission, Proposal for a Regulation of the European Parliament and of the Council on harmonised rules on fair access to and use of data (Data Act), https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex:52022PC0068 (2022). COM/2022/68 final.
7. European Commission, Regulation (EU) 2022/868 of the European Parliament and of the Council of 30 May 2022 on European data governance (Data Governance Act). Official Journal of the European Union **L 152** (2022)
8. European Commission, Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). Official Journal of the European Union **L** (2024)
9. European Commission, Regulation (EU) No 910/2014 of the European Parliament and of the Council of 23 July 2014 on electronic identification and trust services for electronic transactions in the internal market and repealing Directive 1999/93/EC. Official Journal of the European Union **L 257** (2014)
10. Gaia-X European Association for Data and Cloud, Gaia-X Architecture Document - Data Product Conceptual Model, https://docs.gaia-x.eu/technical-committee/architecture-document/latest/component_details/#data-product-conceptual-model. Accessed: 2025-03-31.
11. International Data Spaces Association, 4.3.7 Data Quality — IDS Reference Architecture Model, (2022). https://docs.internationaldataspaces.org/ids-knowledgebase/ids-ram-4/perspectives-of-the-reference-architecture-model/4_perspectives/4_3_governance_perspective/4_3_7_data_quality. Accessed: 2025-03-31.
12. International Organization for Standardization, *ISO 8000-61:2016: Data quality — Part 61: Data quality management: Process reference model*, (2016). https://www.iso.org/standard/63086.html.
13. International Organization for Standardization, *ISO/IEC 25000:2014: Systems and software engineering – Systems and software Quality Requirements and Evaluation (SQuaRE) – Guide to SQuaRE*, (2014). https://www.iso.org/standard/64764.html.
14. Péter, B.Z., Gönczy, L., Kocsis, I.: Data Veracity Assurance BB. Design Document, https://prometheus-x-association.github.io/docs/data-veracity/design-document.html. Prometheus-X (2025)
15. Prometheus-X Association, Prometheus-X: Building Blocks for Data Spaces, (2025). https://prometheus-x.org/. Accessed: March 31, 2025.
16. Reimer, A.P., Madigan, E.A.: Veracity in big data: How good is good enough. Health Informatics Journal **25**(4), 1290–1298 (2019). https://doi.org/10.1177/1460458217744369
17. Sporny, M., Longley, D., Chadwick, D., Herman, I.: Verifiable Credentials Data Model v2.0. W3C Proposed Recommendation, https://www.w3.org/TR/vc-data-model-2.0/. W3C (2025)
18. Sporny, M., Longley, D., Sabadello, M., Reed, D., Steele, O., Allen, C.: Decentralized Identifiers (DIDs) v1.0. W3C Recommendation, https://www.w3.org/TR/did-1.0/. W3C (2022)