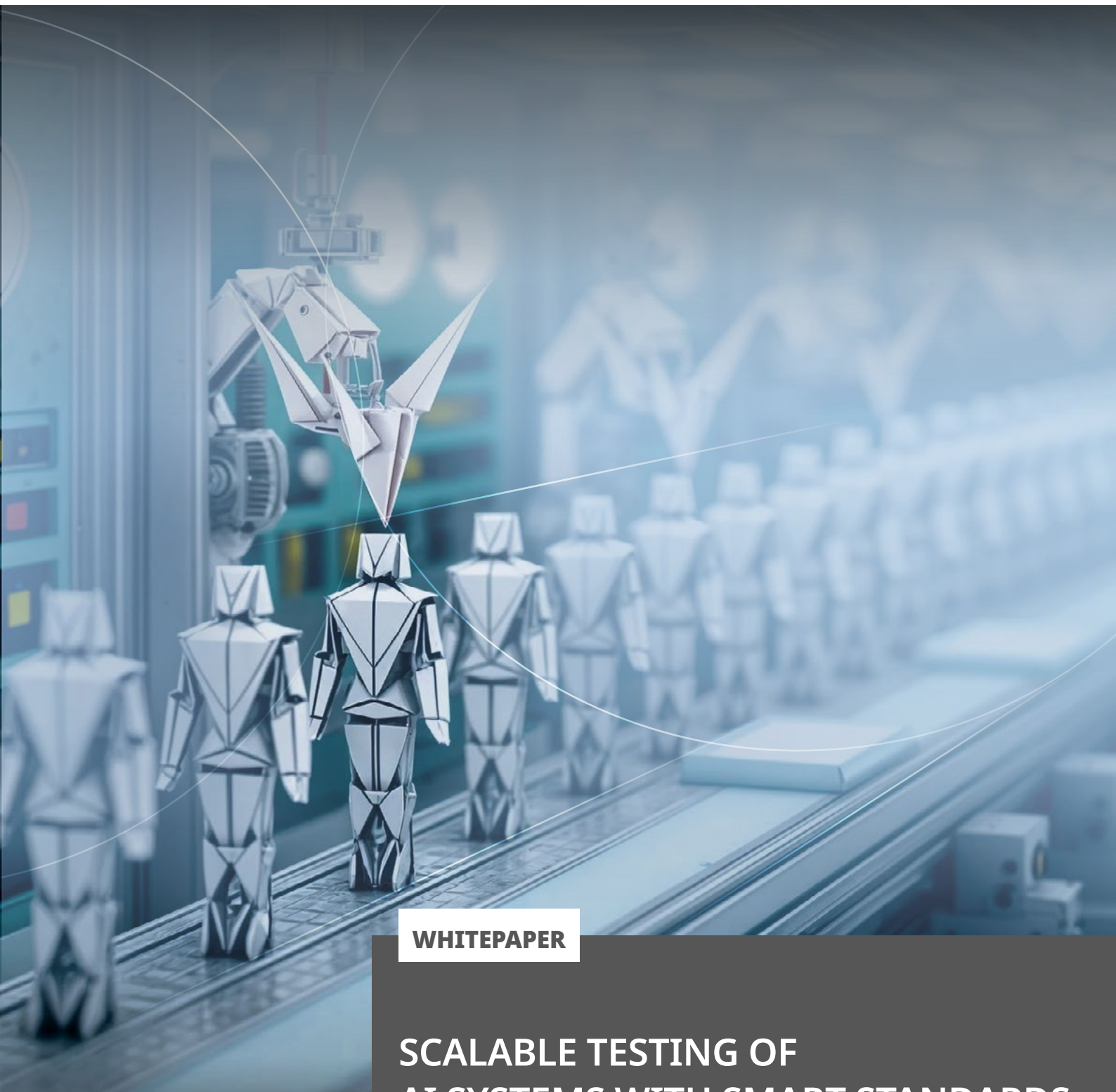




WHITEPAPER

**SCALABLE TESTING OF
AI SYSTEMS WITH SMART STANDARDS:
SCENARIOS FROM A USER PERSPECTIVE**

Authors

Dr. Philipp Heß, Raymond Puppen (DKE)

Contact

Janos Koschwitz (DKE)

Target audience & background

This white paper addresses users of standards in the field of AI, those involved in AI quality and testing, anyone interested in SMART standards, and other decision-makers who want to help shape a digital quality and testing infrastructure for AI systems.

It is the result of interviews and of discussions within the “Scalable AI Testing” (SKIP) group of the Digital Standards Initiative (IDiS) of DIN and DKE. SKIP ran from 2024 to 2025 in the context of the MISSION KI project (acatech, BMDV/ BMDS).

Acknowledgments

We thank all the experts who contributed to this white paper through interviews and technical discussions. Special thanks go to the participants in the “Scalable AI Testing” (SKIP) group and at the kick-off workshop in November 2024.

We would particularly like to thank the following experts for their contributions to the concept and the review of this whitepaper:

Martina Paul (Huawei)

Dr.-Ing. Marc P. Hauer (TÜV AI.Lab)

Achim Pascal Meyer (sphin-X e.V.)

Taras Holoyad (BNetzA)

as well as further contributors from DKE, DIN, QI-Digital and the AI Quality & Testing Hub (AIQ).

PUBLISHED BY



DIN e. V.

Burggrafenstraße 6
10787 Berlin, Germany
Phone: +49 30 2601-0
E-Mail: presse@din.de
Internet: www.din.de/en



VDE e. V.

DKE Deutsche Kommission Elektrotechnik Elektronik Informationstechnik
Merianstraße 28
63069 Offenbach am Main, Germany
Phone: +49 69 6308-0
Fax: +49 69 6308-9863
E-Mail: standardisierung@vde.com
Internet: www.dke.de/en

MANAGEMENT SUMMARY	4
IDIS – THE DIGITAL STANDARDS INITIATIVE OF DIN AND DKE	5
1 INTRODUCTION	8
2 BACKGROUND	12
2.1 Overview: testing artificial intelligence on the basis of standards	12
2.2 Scalable testing of AI systems	17
3 SCENARIOS FOR TESTING AI SYSTEMS WITH SMART STANDARDS	24
3.1 API-based delivery of machine-readable and machine-interpretable standards content	26
3.2 Reliably AI-generated test artifacts	29
3.3 Integration into an ecosystem for AI quality	31
3.4 Scalable generation of SMART content and operationalization content	38
4 CASE STUDY	41
4.1 Scenario 1: using machine-readable standards content via APIs	42
4.2 Scenario 2: generating AI tests	45
4.3 Scenario 3: integration into an ecosystem for AI quality	48
5 DISCUSSION	51
5.1 Evaluation of the scenarios	51
5.2 Relation to central concepts of digitizing standardization	53
5.3 Transferability	55
6 CONCLUSION	58
ABBREVIATIONS	60

MANAGEMENT SUMMARY

Need: testing AI systems is becoming a permanent task

As AI-based products and services grow in importance, so too do the requirements for safety, reliability and transparency, which take effect in practice through systematic testing on the basis of standards. The EU AI Act, together with other horizontal and sectoral legal acts, makes standards-based testing of AI systems an ongoing, integral part of development and operation.

Challenges: frequent testing, frequent updates, a complex standards landscape

Three factors drive the overall effort: the number of systems to be tested, how often they change through model updates and changes to data, and the complexity of an evolving legal and standards landscape. If innovation and quality assurance are to stay in balance, testing has to be automated as far as possible and the remaining manual activities accelerated by supplying precisely the information that is needed.

SMART standards enable automated use of standards content

SMART standards provide standards content as granular, machine-readable and machine-interpretable resources. They are a prerequisite for standards-based testing of AI systems to be made scalable. Combined with standardized testing tools, reference datasets, testing platforms and a data-driven quality infrastructure, recurring testing activities can be integrated into the ongoing development and operational processes of AI systems, partly automated and documented in a reproducible way.

From accessing standards via API to an ecosystem for AI quality

Three scenarios describe a step-by-step path from today's manual practice to an increasingly automated testing environment. The scenarios described in this white paper address the technical support and partial automation of testing activities. They are no substitute for expert judgment or the responsibility of testers, auditors or conformity assessment bodies (CABs). Requirements for independence, confidentiality, risk-based decision-making and other organizational conditions continue to depend on the

specific testing context and are not replaced by the technical concepts presented.

Scenario 1: API-based delivery of standards content

SMART standards are integrated into requirements management and development tools via software interfaces, so that standards content, together with unambiguous references, feeds directly into development and testing. Updates are detected automatically, affected requirements are flagged, and conformity reports are derived automatically from the documented test steps. This scenario could already be implemented with components available today.

Scenario 2: reliably AI-generated test artifacts

Building on scenario 1, an AI assistant takes control: it identifies the relevant requirements, proposes test plans and generates test code. Standards content such as limit values and thresholds is carried over from the standards unchanged and does not pass through the language model, so that it remains traceable to its clause in the standard. Full implementation requires new kinds of SMART content to be available.

Scenario 3: integration into an ecosystem for AI quality

Standards content, data, testing tools and testing services are available as sovereign resources in a shared data space. Certificates of conformity travel with the AI system as digital evidence. This scenario is designed for the long term and presupposes extensive agreement on the rules of such a data space.

IDIS – THE DIGITAL STANDARDS INITIATIVE OF DIN AND DKE

Background and objective

The German Institute for Standardization (DIN) and the German Commission for Electrical, Electronic and Information Technologies (DKE) have been driving the digital transformation of standardization since 2020 with the Digital Standards Initiative (IDiS). IDiS acts as the link between the development of standards and their digital application, in the form of an open community platform that brings together companies, associations, public authorities and research institutions and develops concrete requirements for the digital standard. IDiS was launched as a direct response to a fundamental gap: while industry was increasingly automating its processes and using digital requirements engineering systems, standards were still largely available in document-based formats unsuited to machine-supported processing.

SMART standards: the core of IDiS

At the center of these activities are SMART standards: standards whose content is made available digitally and is readable and usable for humans and machines alike. IDiS follows a staged maturity model (utility levels 2 to 5), ranging from initial digital provision of standards content through to fully automated, machine-interpretable application of standards.

Participants and working structure

IDiS brings together more than 130 participants from industry, business, associations and academia. The initiative's substantive work takes place in ad hoc groups and pilot projects, whose results feed into white papers and specifications. Twice a year the IDiS plenary meets for cross-group exchange. Coordination with CEN/CENELEC and ISO/IEC, as well as with national standardization organizations such as AFNOR, ensures that the work of IDiS ties in with international activities.

Digital services and tools from DIN and DKE

DIN and DKE are advancing the digitalization of their own processes and products. With the Online Standards Development Tool (OSD), ISO/IEC has had a shared platform since 2020; its broad rollout has been under way since 2023, and it is the default for new standardization projects from

2025. OSD enables standards to be drafted in a structured, XML-based way and thus creates the basis for enriching their content step by step to produce SMART content.

In addition, DIN and DKE are building a growing portfolio of digital services along the entire lifecycle of a standard, from drafting through quality assurance to application. These include AI-supported tools that assist standardization experts with research and editing, services that ensure terminological consistency across standards, and API-based access enabling companies to integrate standards and standards-related processes directly into their own systems. Together these building blocks form the bridge from traditional standardization processes to an end-to-end digital, machine-readable and application-oriented body of standards.

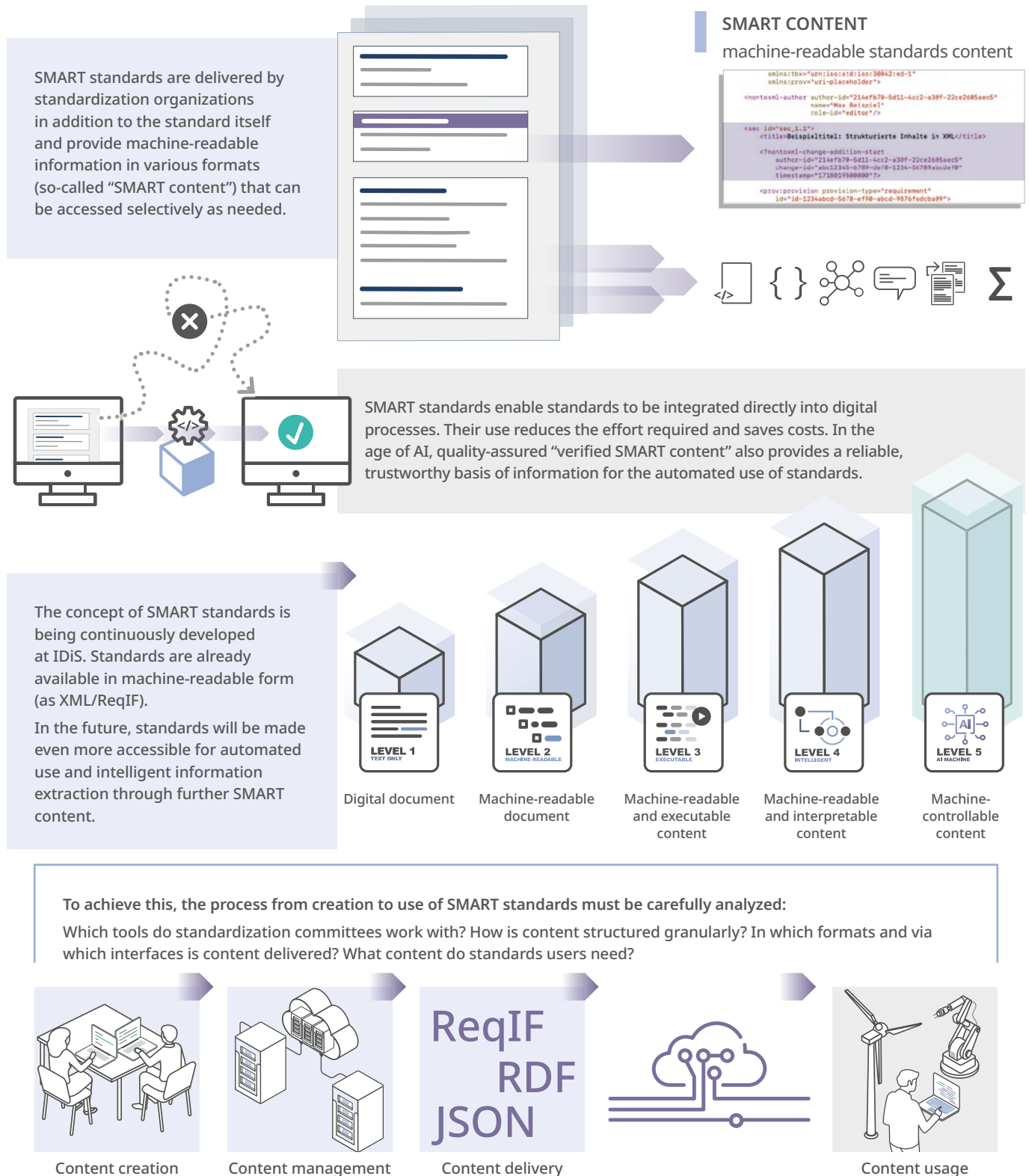
Key outputs

Several concrete results are already available. DIN DKE SPEC 99200 defines a uniform ReqIF interpretation for public standards and thus establishes a first standardized exchange format for machine-readable standards content, allowing companies to prepare their systems for SMART standards, including change management at the requirement level. Completed pilot projects have demonstrated, among other things, automated conformity testing and the integration of standards content into the Asset Administration Shell (AAS) for digital pre-certification processes.

Integration into the digital quality infrastructure

In the QI-Digital roadmap, SMART standards are positioned as a central building block of a digital quality infrastructure. They allow normative requirements to be provided in a structured, machine-processable form and thus create an important precondition for automated conformity testing and documentation, digital certificates and the linking of normative requirements with digital product information. Their use presupposes suitable data models, exchange formats and semantic descriptions. In this context, SMART standards can also play an important role in digital product data ecosystems such as the Digital Product Passport (DPP), by providing the normative reference for product information and compliance evidence linked in machine-readable form.

What are SMART standards?



Further information:

- IDiS Whitepaper 1. Scenarios for digitizing standardization and standards
- IDiS Whitepaper 2. Use cases of SMART Standards
- IDiS Whitepaper 3. The business benefits of SMART Standards



Figure 1

1

INTRODUCTION

1 INTRODUCTION

With projected growth rates of more than 30 percent per year, AI-based products and services are among the fastest-growing segments of the global economy. They are used in fields ranging from medicine and mobility to education, public administration and industry. As the social and economic significance of such systems grows, so does the responsibility to ensure that they are safe and reliable.

Standards are becoming increasingly important for the design and testing of AI systems. They codify knowledge developed in science and industry, establish a common technical language and create a framework intended to safeguard the quality, safety and reliability of AI systems.

Many AI standards cover not only technical but also societal requirements, for example with regard to non-discrimination, transparency, human oversight and the protection of safety, health and the environment. These aspects reflect policy objectives and guiding legal principles that go beyond purely technical quality characteristics.

As the product of extensive consensus-building processes, standards are increasingly being used internationally as an instrument of government regulation of AI, for instance by being referenced in the European AI Act. For many companies, the AI Act, and with it the application of the relevant AI standards, is likely to become highly important.

This also makes it ever more important to test AI systems for conformity with standards. It is foreseeable that testing AI will pose particular challenges for the developers and deployers concerned, above all because AI systems are non-deterministic, because they depend heavily on data quality, because complex models offer only limited explainability, and because learning systems change over their lifecycle.

New challenges also arise specifically for requirements management for AI systems and for products containing AI components. These include translating abstract normative requirements into verifiable technical criteria, continuously checking that requirements are still met when models or data change, and maintaining end-to-

end traceability of which standards content led to which technical requirements, which implementation and which test evidence.

This white paper is aimed at (potential) users of standards in the field of AI, at those working on AI quality and AI testing, and at anyone with a general interest in SMART standards. It sets out to give a clearer picture of the contribution that SMART standards can already make, and could make in the future, to improving the scalability of standards-based testing of AI¹. At the same time it is intended to serve as a basis for discussion on developing SMART standards further in the context of such testing.

The following questions are central:

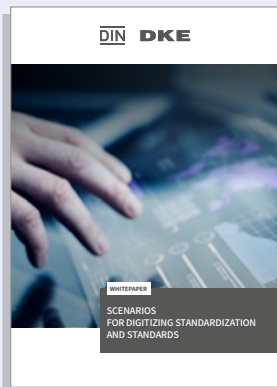
- What is the future need for scalable standards-based testing of AI?
- What contribution can SMART standards potentially make, and in which use cases?
- What functions must SMART standards perform in the future in order to do so?
- How can this be achieved? What has to be implemented technically, for example in standards development, and in which ecosystem should SMART standards be used for this purpose?
- What do the scenarios for using SMART standards in AI testing imply for the further development of SMART standards as a whole?

In terms of content and structure, this white paper builds on a series of IDiS publications and pilot projects² that have driven the conceptual and practical development of SMART standards forward in recent years.

In particular, white paper 1 “Scenarios for digitizing standardization and standards” (WP1), white paper 2 “Use Cases for SMART Standards” (WP2) and white paper 3 “The Business Oriented Benefit of SMART Standards in Standard Application Processes” (WP3) form an important basis.

¹ Scope: The detailed content of AI (testing) standards, the specific application and operationalization of relevant AI standards, and their conversion into SMART standards are not discussed case by case in this white paper.

² <https://www.dke.de/idis/pilotprojekte/roadmap>



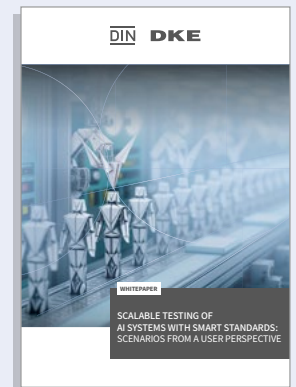
Whitepaper 1



Whitepaper 2



Whitepaper 3



Whitepaper 4

White paper 1 deals with scenarios for digitizing standardization and standards. The scenarios were described using the four value creation process phases (content creation, content management, content delivery and content usage) and explain which processes have to be adapted for a “digital standard”. The extended utility model describes five levels of digitized standards, beginning with digital but unstructured text documents (Word, PDF), moving through fully machine-interpretable content, and reaching a vision of standards that are automatically checked and revised by an AI. The scenarios in this white paper map directly onto that model and thus spell out, for AI testing, which contributions to which utility levels deliver which practical benefit.

White paper 2 deals with the applicability of SMART standards in the context of use cases. These describe in basic terms what will be possible when working with SMART standards in the creation, management, delivery and use of standards. A large number of use cases were consolidated into 11 Generic User Stories (GUS) in order to describe generic application scenarios for SMART standards. Several of these GUS have been applied to the “AI testing” use case in this white paper.

White paper 3 deals with the business benefit of SMART standards in the application of standards. It analyzes how digitized and machine-readable standards can be embedded in operational processes and what effects this has for companies. Using a benefit-oriented model, it sets out advantages in time, cost and quality along typical standards application processes. The white paper shows that SMART standards contribute to efficiency gains above all through

automation, reduced manual effort and lower error rates, and thus deliver measurable business value.

One key effect lies in breaking up the sequential nature of processes. In the PDF-based status quo, requirements from standards first have to be researched, then extracted manually and finally transferred into downstream systems before the actual development work can begin. Every step waits for the one before it. SMART standards decouple these dependencies. Because standards content is available as granular units with stable addresses, different teams and process steps can access the same normative content at the same time, without passing through a central extraction and transfer layer.

The result is a shift from serial to parallel processes, which white paper 3 quantifies using a value-added model: according to that model, 1.5 to 2 times more orders can be processed with the same staffing level, and the costs of standards-related activities fall by between 48% and 64%.

This white paper addresses standards-based testing of artificial intelligence using SMART standards and thus, in contrast to the white papers published so far, focuses on a single, very specific technology area.

In a field that is still taking shape, this presents a number of challenges. In some cases, assumptions have to be made about the future relevance of particular technologies and concepts, for example about the use of standardized AI testing tools, AI testing platforms, reference datasets or data spaces. Relevant practices such as requirements management for AI are also still being adapted and

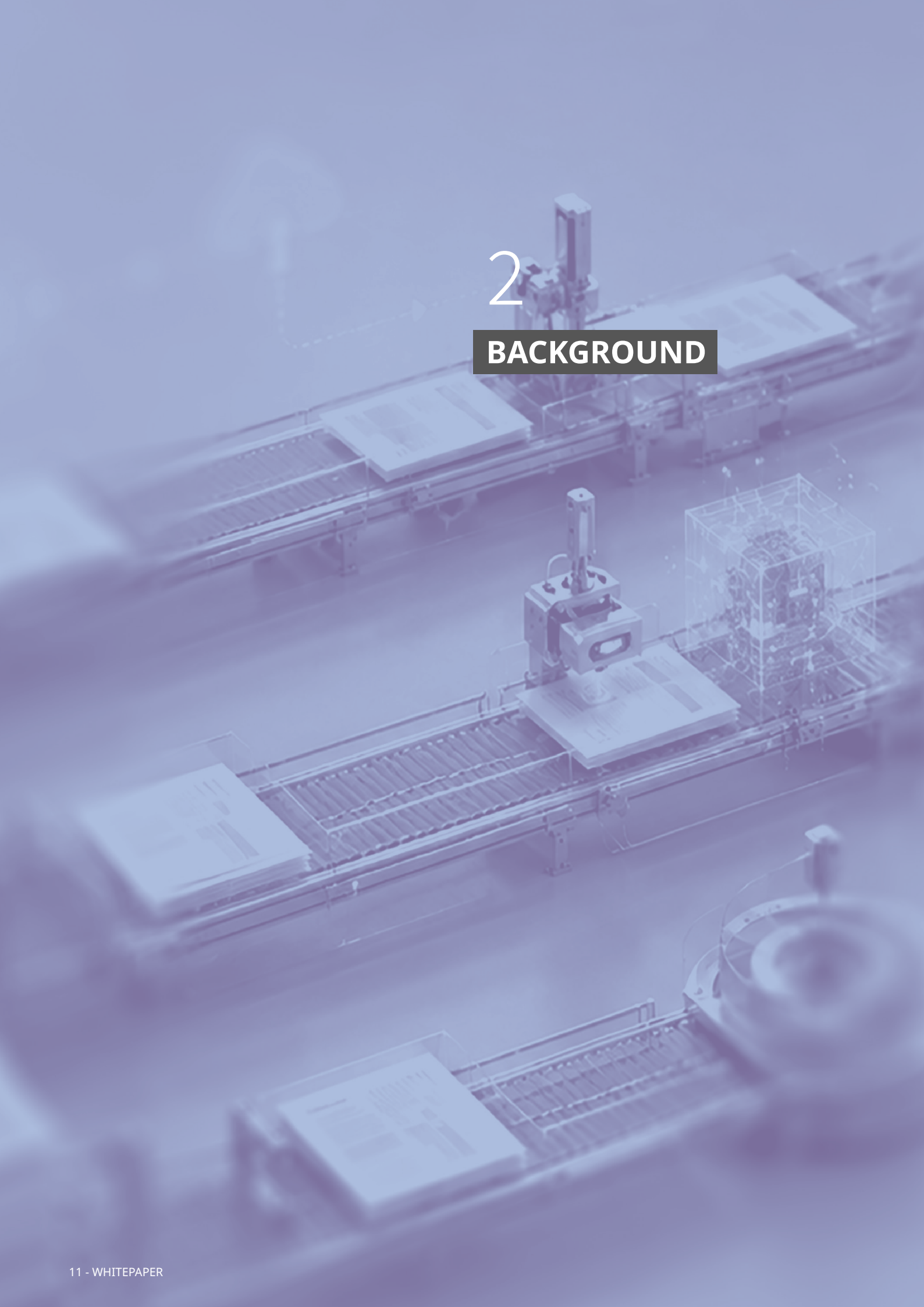
developed.³ In addition, at the time of publication the AI standards landscape is still at an early stage of development, in particular because some harmonized AI standards for the AI Act have yet to be published and because the shape of AI-relevant vertical standards is not yet clear. There is therefore a degree of uncertainty as to which standards content can be made available in which form via a SMART standard (“How specific will the requirements for AI in standards become?”).

These factors make it challenging to describe AI testing processes in technical detail and in a “future-proof” way, and to correctly assess the need for particular features of SMART standards. Yet it is precisely scenarios and proposals that offer an opportunity to establish a basis for discussion, one that brings the potential of SMART standards into the debate on developing the quality ecosystem (or quality infrastructure) for AI at an early stage.

The further development of the SMART standards concept can generally benefit from being connected to the field of artificial intelligence. Companies that apply AI standards and test AI typically have a high degree of digitalization and often work in agile development processes⁴. The need for SMART standards functionality is correspondingly high in this environment. Given the digital infrastructure and processes already in place, which frequently contain automated elements, it is also possible to design particularly far-reaching integration scenarios for SMART standards.

3 See for example: Habiba, U., Haug, M., Bogner, J., & Wagner, S. (2024). How mature is requirements engineering for AI-based systems? A systematic mapping study on practices, challenges, and future research directions. *Requirements Engineering*, 29(4), 567–600. <https://doi.org/10.1007/s00766-024-00432-3>

4 See white paper 3 for a discussion of the degree of digitalization of companies, of agile development processes (SCRUM) and of the use of SMART standards.

An aerial, high-angle view of a paper mill facility, showing multiple long conveyor belts and processing machines. The entire image is overlaid with a semi-transparent blue filter. A large, white number '2' is positioned in the upper-middle section of the image.

2

BACKGROUND

2 BACKGROUND

Testing determines whether, or to what extent, an artificial intelligence system meets the requirements placed on it. Such testing may involve activities such as (automated) testing of data and AI models or reviewing processes and organizational structures on the basis of documentation. “AI testing”⁵ can, for example, be a regular part of development and validation processes, or be carried out during an audit or a conformity assessment.

Particularly for providers or deployers of artificial intelligence whose application carries a high risk, such testing is essential given the clear regulatory requirements. At the same time it is becoming apparent that binding requirements are emerging for lower risk levels too, and that demand for independent testing is rising markedly in general. Voluntary testing not only helps to ensure a minimum level of quality; it also allows providers to make their technical and ethical responsibility visible and thereby strengthen trust among users, customers and supervisory authorities.

Approaches are therefore increasingly emerging that aim to enable systematic, transparent and efficient assessment of AI systems across all risk classes. This is where concepts of scalable testing come in, with the goal of reconciling quality assurance and the pace of innovation and making testing practicable in future AI ecosystems.

The following sections set out a classification of various types of standards-based testing. The aim is not to give a full account of any one regulatory context, such as conformity assessment under the EU AI Act. Rather, different application and testing contexts are considered in which AI systems are assessed against normative requirements. The types of testing presented serve as an analytical framework for structuring possible application scenarios for scalable testing with SMART standards. In practice these contexts frequently overlap and combine.

2.1 Overview: testing artificial intelligence on the basis of standards

While providers and deployers of AI systems set out a large number of their own requirements, it is standards that provide a public, systematic and often legally relevant compilation of operationalizable requirements for the properties of AI. Current standards for AI cover aspects such as risk and quality management, data governance, security, and, summarized under “trustworthiness”, transparency, human oversight and control, reliability and so on.⁶

Standards-based testing of AI can be carried out in different ways, depending on what is tested, who carries it out in which context, when it takes place and what goals it pursues.

Why (in which context) is testing carried out?

The goals of AI testing differ according to the application context. They range from internal quality assurance and risk minimization through requirements along the value chain to demonstrating legal compliance. Conformity assessments play an important role in regulated fields of application in particular. In the European legal framework, conformity is often demonstrated by implementing standards. For high-risk AI systems the AI Act provides specific conformity assessment procedures under Article 43 and Annexes VI and VII. Applying harmonized European standards can give rise to a presumption of conformity with the legal requirements addressed. Testing an AI system as part of a conformity assessment procedure can also play a role in complying with other, for example sectoral, legislation of relevance to AI, such as the Medical Device Regulation (MDR) or the new Machinery Regulation.⁷

Where AI systems are tested both under the AI Act and under other sectoral legislation such as the MDR, overlaps can arise between the respective conformity assessment procedures. As yet these procedures have not been fully

5 In what follows, AI testing is considered one component of AI assessment activities. The term refers in particular to testing AI systems (and not necessarily to testing using AI).

6 See for example: Standardization Roadmap on Artificial Intelligence, 2nd edition, p. 136 ff; standardization topics of JTC 21, <https://jtc21.eu/working-groups/>; MISSION KI quality standard, <https://github.com/mission-ki/standard/>; Fraunhofer IAIS AI assessment catalog, <https://www.iais.fraunhofer.de/de/publikationen/studien/2021/ki-pruefkatalog.html>; VDE SPEC 90012

7 <https://www.vde.com/topics-de/health/beratung/zulassung-von-ki-basierten-medizinprodukten-in-europa>

Testing artificial intelligence

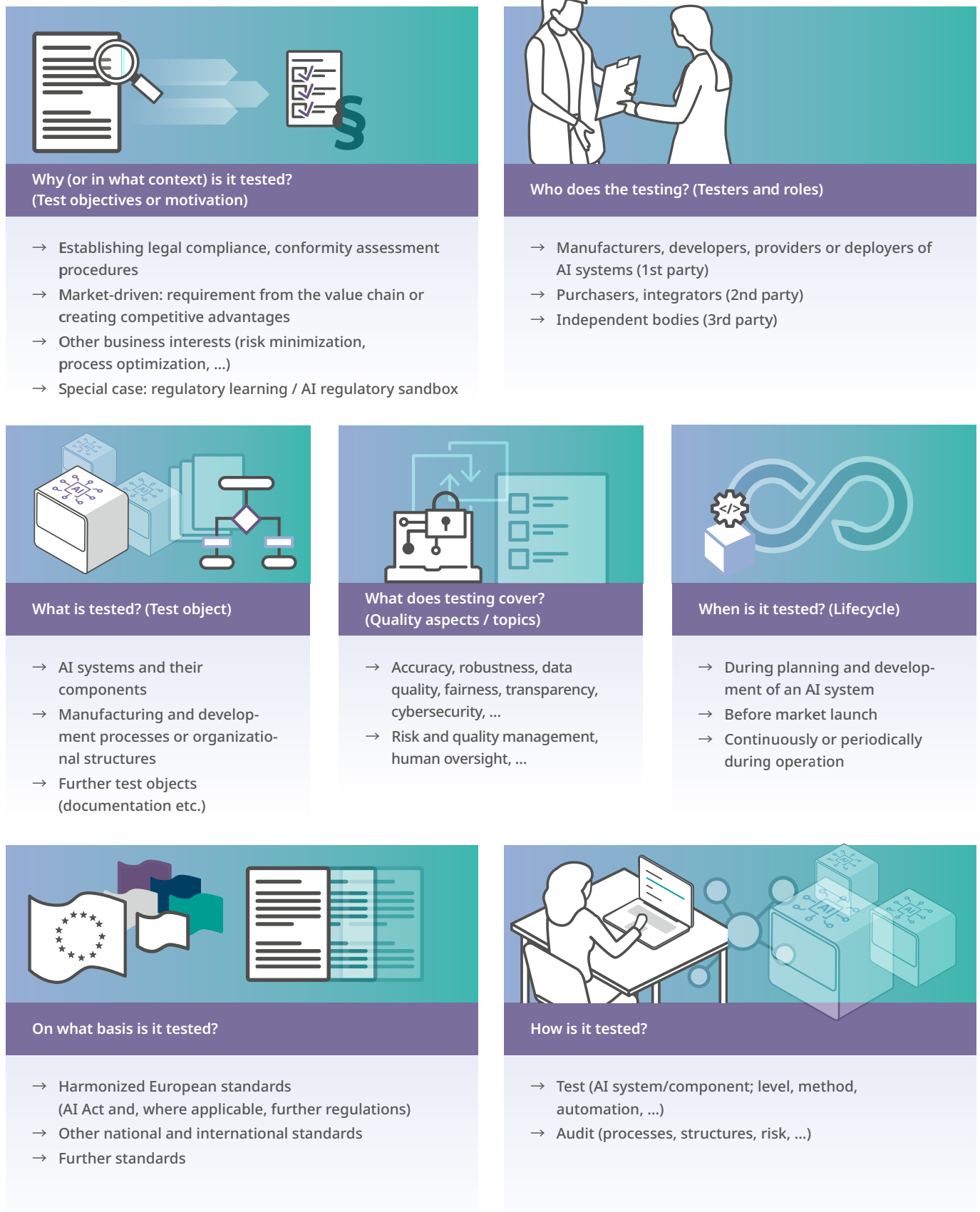


Figure 2

harmonized with one another. Exactly how these interfaces between the AI Act and sectoral legislation are shaped remains an open question in the ongoing development of the regulatory framework.

The reproducibility of standardized testing enabled by SMART standards could make an important contribution here: implemented and documented tests could be reused in different regulatory contexts without having to set up every procedure from scratch.

Besides testing carried out as part of legally required conformity assessments, there can also be various market-driven reasons. AI audits and certifications can involve different forms of testing and can be relevant for systems with lower risks under the AI Act. Requirements from the value chain, for example, can make testing necessary. Business partners, integrators or purchasers of AI components increasingly demand evidence of quality and safety. At the same time, a tested and, where applicable, certified AI system can offer competitive advantages: it signals quality characteristics such as transparency, reliability and responsibility, and strengthens the provider's position in the market.⁸ Finally, manufacturers, developers, providers and deployers use standards to test AI in order to minimize risks, maintain their own quality standards and improve internal development processes.

Testing AI systems in AI regulatory sandboxes is a special case.⁹ Here testing can be carried out under the guidance and supervision of the relevant authorities and with limited (legal) risk¹⁰. For authorities, the goal of such testing can be to test and improve AI regulation (regulatory learning), while companies may be able to carry out conformity assessments more easily or with less risk by testing in AI regulatory sandboxes.

When is testing carried out?

AI testing can take place at different phases of the system lifecycle, for example during the (re)planning or (re)design

of an AI system, or during (further) development for quality assurance and error prevention. The phases before and during deployment are also relevant, that is before market launch (as part of a conformity assessment procedure or a certification) and during operation in the form of continuous or periodic monitoring, for example when models are updated, as part of post-market monitoring (see also chapter 2.2).

On what basis is testing carried out?

Different bases with requirements for testing are used depending on the testing context and objective. A basic distinction can be drawn according to how legally binding these bases are:¹¹

As a rule, conformity assessments are carried out on the basis of the standards referenced by the legislation. These harmonized European standards (hEN) are developed by CEN, CENELEC or ETSI on the basis of standardization requests from the European Commission. Applying them is voluntary, and legal conformity can in principle also be demonstrated by AI testing that is not based on (these) standards but addresses the legal requirements in another way¹². However, since conformity assessments on the basis of harmonized standards confer a special "presumption of conformity" (with the legislation),¹³ standards-based AI testing is likely to be the most relevant form of testing in most cases, as it offers greater legal certainty and involves the least effort.

The central harmonized standards (hEN) for the AI Act are currently being developed. The first of them (prEN 18286, see info box) went through public enquiry at the end of 2025 and was published in 2026. Further core standards addressing individual requirements of the AI Act are being developed. In October 2025 CEN and CENELEC adopted special measures to ensure the availability of the harmonized standards by Q4 2026.

8 See for example the MISSION KI quality standard, <https://github.com/mission-ki/standard/>

9 E.g.: <https://epthinktank.eu/2026/04/01/ai-regulatory-sandboxes-state-of-play-and-implementation-challenges/>

10 cf. Art. 57 AI Act

11 Cf. Standardization Roadmap (2), p. 128 ff / Fig. 27

12 Annex VII AI Act

13 Art. 40 AI Act

AI STANDARDS AS A BASIS FOR TESTING (EXAMPLES)

Harmonized European standards (CEN/CENELEC):

- prEN 18286 - AI quality management system
- prEN 18228 - AI risk management: supports the requirements for the risk management system under Article 9 of the AI Act and is closely interlinked with the quality management system through Article 17
- prEN 18282 - Cybersecurity for AI systems: adds AI-specific security requirements and builds on existing standards such as the ISO/IEC 27000 series and ISO/IEC 15408 (Common Criteria)

Other international standards (ISO/IEC):

- ISO/IEC 42001 - Information technology - Artificial intelligence - Management system
- ISO/IEC 23894 - Information technology - Artificial intelligence - Guidance on risk management
- ISO/IEC 25059 - Software engineering - Systems and software Quality Requirements and Evaluation (SQuaRE) - Quality model for AI systems

In addition there are further national and international standards. These can serve as a basis for testing even without formal harmonization. They are often already available and are used in practice, for example while harmonized standards are still pending. Management system standards such as ISO/IEC 42001 address the process level; for product-related quality testing, ISO/IEC 25059 and the conformity assessment procedures built on it become the authoritative reference.

Finally, further standards such as consortium standards, frameworks and guidelines (e.g. NIST AI RMF, the MISSION KI quality standard, the Fraunhofer IAIS AI assessment catalog) play a complementary role, particularly where formal standards do not yet provide sufficiently operationalizable requirements.

Who does the testing?

Depending on testing goals and contexts, there are different roles and responsibilities in AI testing. For testing as part of formal conformity assessments in particular, these can be grouped into three levels. In the case of self-assessments (first-party assessment), testing is carried out by the manufacturer, developer, provider or deployer themselves, for example during the development phase or in ongoing operation, or as part of a legally required initial assessment (for example an “internal control” of high-risk AI systems¹⁴). The self-assessment considered here refers to testing activities... It therefore represents only one possible component of a more comprehensive conformity assessment or self-declaration.

External organizations such as purchasers or integrators test the quality of AI components or systems they have procured, for example to ensure that these are suitable for their own applications (second-party assessment). External, independent bodies (third-party assessment) carry out formal assessments and certification. These include legally required conformity assessment procedures, product certification and system audits.

The assessing party is determined by the regulatory context. Both first-party and third-party assessments can be legally required. For most high-risk systems the AI Act allows internal control by the provider (Art. 43 in conjunction with Annex VI), but for certain biometric systems it requires the involvement of a notified body (Annex VII). Second-party assessment is less often required by regulation, but in practice plays a growing role in securing AI components in supply chains.

Several roles are usually involved in AI testing, and exactly how these are defined varies according to the type of organization and the testing context. On the provider and deployer side, roles such as product owner, ML engineer, quality engineer, compliance manager or data scientist are typically involved. They are responsible for operationalizing requirements, implementing and running tests, and documenting test results. External testing is performed by accredited conformity assessment bodies (CABs), testing laboratories or specialized auditors. Finally, market surveillance authorities exercise downstream oversight once systems are in operation. Which of these roles are specifically responsible for AI testing, and how

¹⁴ Art. 43(1)(a) AI Act

they are embedded in organizations, is still being actively discussed, including in the context of ETSI TR 104 119, which gives recommendations on documentation and role responsibility for AI systems.

What is tested, and how?

AI testing can address various test objects for which rules have been laid down in AI standards. The central test object is the AI system in use, which usually means a system based on machine learning models. General definitions such as that of the AI Act¹⁵, however, also include other forms of AI.

Testing an AI system includes, for example, testing the system as a whole (end-to-end testing), or testing individual components such as (trained) models, the data used (training, test and validation data) and further system components and interfaces (code, APIs, infrastructure). Besides the object and level of testing, tests also differ in the degree to which they can be automated, in the level of access required to the AI model and (training) data, and in the types of test method used (e.g. black, white or gray box testing).¹⁶ A further, testable, artifact of an AI system is, for example, the accompanying documentation¹⁷, which covers technical descriptions or risk analyses.

Besides the AI system itself, the underlying development and operating processes can also be the object of testing (cf. process audits under ISO/IEC 17021). What is tested are the methods, procedures and organizational structures that ensure quality, safety and traceability during development. These include development and validation processes, the data management practices used, risk management measures and internal audit and release processes. Management systems that institutionalize these processes play a particular role here. Relevant examples are the AI management system according to ISO/IEC 42001, the quality management system according to ISO 9001 and the information security management system according to ISO/IEC 27001.

They create the organizational basis for reproducible, documented and verifiable AI development.

Furthermore relevant, according to ISO/IEC 17024, can be the assessment or certification of persons carrying out tests or responsible for development and monitoring. ISO/IEC 17025 further governs the assessment and accreditation of testing and calibration laboratories (conformity assessment bodies) in order to ensure their technical competence and independence. These types of testing are, however, not considered in this white paper.¹⁸

¹⁵ Art. 3 AI Act

¹⁶ see DIN SPEC 92006

¹⁷ Testing management systems is also largely carried out on the basis of documentation. To that extent the distinction is not clear-cut, depending on interpretation.

¹⁸ On this topic see for example: Kuch, S., & Kirmes, R. (2025). AI Certification: An Accreditation Perspective (Practitioner Track). In R. Gorge, E. Haedecke, M. Poretschkin, & A. Schmitz (Eds.), OASICs, Volume 126, SAIA 2024 (No. 14; Vol. 126, pp. 14:1-14:7). Schloss Dagstuhl – Leibniz-Zentrum für Informatik. <https://doi.org/10.4230/OASICs.SAIA.2024.14>

IN DETAIL: LEGAL REQUIREMENTS FOR RECURRING OR CONTINUOUS TESTING OF AI

For high-risk AI systems the AI Act establishes a series of obligations that distribute testing activities systematically across the lifecycle instead of concentrating them at the point of market launch. The risk management system (Art. 9) is expressly defined as a continuous, iterative process, the quality management system (Art. 17) prescribes testing, validation and change procedures across the entire lifecycle, and the required properties of accuracy, robustness and cybersecurity (Art. 15) are likewise formulated with reference to the lifecycle.

Operationally, this permanent task is made concrete above all through post-market monitoring (Art. 72). Providers have to collect performance data actively and systematically and compare it against the requirements of Chapter III, Section 2 of the AI Act in order to assess continuing conformity. The tight reporting deadlines for serious incidents under Art. 73 and the monitoring obligations of deployers under Art. 26(5) presuppose an equally continuous detection infrastructure.

Under Art. 43(4), changes to the system are subject to their own assessment logic: substantial modifications trigger a new conformity assessment, and the exception for learning

systems merely shifts the effort into continuous monitoring of a predetermined range of changes. There is also a quantitative dimension, since the AI Act covers eight high-risk fields of application. Companies with several products or model variants may therefore have to test not just one but many systems.

The AI Act is not the only piece of legislation that requires recurring testing. If an AI system is a safety component of an already regulated product, the lifecycle obligations of the sectoral frameworks (Medical Device and IVD Regulations, Machinery Regulation, Cyber Resilience Act) supplement those of the AI Act. Cross-cutting rules such as DORA, NIS2 and the GDPR increase obligations further by formulating requirements for continuous ICT risk management, ongoing cybersecurity measures and the ability to demonstrate compliance at any time. Management system standards such as ISO/IEC 42001 embed this permanent task in the organization through internal audits, management reviews and surveillance audits.

2.2 Scalable testing of AI systems

Preparing and carrying out testing of AI systems in the context of legislation, formal standards and other standards involves a series of complex activities, such as

- identifying the relevant quality dimensions, legislation and standards,
- carrying out risk assessments,
- operationalizing requirements,
- including investigating, defining and collecting specific metrics,
- and selecting software libraries or implementing (and verifying) code.

These activities are not necessarily specific to AI, but carrying them out differs from conventional software testing in several respects. AI systems mostly work non-

deterministically, and their behavior cannot be described completely by fixed rules. Results depend substantially on data, which can itself be a source of error, for example through bias, quality defects or drift during operation.

Unlike conventional software products, AI systems often lack complete and unambiguous specifications, so that properties such as accuracy, fairness or robustness can only be tested by means of metrics and approximations, and the resulting evaluations are subject to uncertainty. In addition, AI systems change over time, for example through retraining or changed contexts of use, so that one-off tests quickly lose their validity. The often limited explainability of many models also makes it harder to analyze the causes of faulty behavior.

Numerous legal provisions and requirements in standards demand regular or continuous testing and reporting on the conformity of AI systems (see info box). Testing of AI

systems therefore has to be carried out regularly and on a large scale.¹⁹ The amount of testing grows with three factors: the number of systems to be tested, how often they change through retraining, data updates or adjustments to dependent components, and the complexity of the standards landscape with its variety of information sources and testing aspects.

If innovation and quality assurance are to stay in balance, the effort per test must remain manageable in relation to the number of testing occasions. Scalability here means that testing is automated as far as possible and that the remaining manual activities are accelerated by making suitable information and tools easier to obtain.^{20,21}

Various technical concepts can improve the scalability of testing AI systems by automating processes, making them reproducible and reducing additional effort at interfaces. One example is embedding continuous, automated testing into MLOps frameworks.

While MLOps integration mainly addresses scaling over time and the parallel management of several system variants, a second class of measures concerns the comparability and reusability of tests across systems and testing occasions:

The operationalization of requirements from standards often leaves room for interpretation. Decisions have to be made concerning, for example, the precise choice of algorithms and metrics, the setting of (hyper)parameters, or the implementation or selection of libraries and validation datasets.

The more such decisions are standardized, that is, the more rules exist for operationalization, the easier it is to make tests comparable and reusable. This makes it easier to verify tests, compare test results and reuse them in different testing contexts, for example to draw on automated tests that have already been implemented when a conformity assessment is carried out for the market launch of an AI system.

An infrastructure for AI testing can include various components that contribute to the standardization and reproducibility of AI testing, such as AI testing tools, specialized testing services on corresponding technical platforms, standardized reference or test data, or semantic models of regulatory frameworks. Integration into the digital quality infrastructure also plays an important role: here, effort can be reduced and quality increased, for example through the use of digital, standardized test and conformity reports.

19 There are exceptions, for example in the case of “locked AI” systems that do not continue to learn after being placed on the market and therefore remain in an unchanging state. Their properties, which stay the same after assessment, fit better with procedures designed for static products.

20 Haedecke, E. G., Mock, M., Pintz, M. A., & Poretschkin, M. (2023). KI-Anwendungen systematisch prüfen und absichern. Prüfwerkzeuge und Prüfplattform zur Gestaltung vertrauenswürdiger Künstlicher Intelligenz.

21 Symposium on Scaling AI Assessments (SAIA 2024). Open Access Series in Informatics (OASICS), Volume 126, pp. 0:i-0:xii, Schloss Dagstuhl – Leibniz-Zentrum für Informatik (2025) <https://doi.org/10.4230/OASICS.SAIA.2024.0>

EXAMPLES OF CONCEPTS AND COMPONENTS FOR SCALABLE TESTING OF AI:

- 2.2.1 CABCA / MLOps
- 2.2.2 AI testing tools
- 2.2.3 Reference datasets
- 2.2.4. AI testing platforms
- 2.2.5. Semantic modeling
- 2.2.6 Standardized test reports
- 2.2.7 SMART standards

2.2.1 Continuous conformity testing in an MLOps framework

In the concept of continuous auditing-based conformity assessment of AI (CABCA)^{22,23,24} point-in-time audits are replaced by ongoing, automated checking of conformity. The core of the approach is the operationalization of abstract requirements, which are systematically translated into concrete, measurable and machine-readable criteria. On this basis, evidence is collected automatically and continuously from the AI system, for example from logs, performance indicators or data and model artifacts. The measurements collected are analyzed continuously and compared with defined thresholds. Deviations, new risks or non-conformities are detected promptly, documented and brought together into a current conformity status. Testing and monitoring extend from development and training through operation to post-market monitoring, with the depth and intensity of testing depending on the risk profile of the AI system.

Technically and organizationally, such a concept is ideally embedded in an MLOps (machine learning operations) framework, comprising the practices, processes and tools with which the entire lifecycle of AI models, from data preparation and training through deployment to operation and maintenance, is managed systematically.

In the context of AI testing, MLOps provides the operational infrastructure for continuous measurement and monitoring. The artifacts and evidence needed for continuous conformity testing are supplied by typical MLOps mechanisms such as versioning of models and data, automated tests, monitoring of model performance, drift detection and logging. Testing activities such as calculating and comparing metrics on the basis of data or AI models can be integrated directly into existing MLOps pipelines, so that testing takes place not as an additional manual step but as an integrated part of ongoing operation.

2.2.2 AI testing tools

(Standardized) AI testing tools²⁵ are designed to support transparent, traceable and repeatable testing processes. This presupposes that the way they work, and the metrics, inputs, parameters and outputs they use, are documented in a comprehensible form. Test objectives and test criteria are defined so that it is clear which properties of an AI system are tested under which conditions.

The central element of the concept is the substantive validity of the test methods used: the methodology and technical implementation of the tools and their interfaces are documented and given a sound rationale, so that the substance of the tests remains comprehensible. For the intended field of application, the aim is for the tools to work

22 Knoblauch, D., & Großmann, J. (2023). Towards a risk-based continuous auditing-based certification for Machine Learning. *The Review of Socionetwork Strategies*, 17(2), 255–273. <https://doi.org/10.1007/s12626-023-00148-w>

23 Minkinen, M., Laine, J., & Mäntymäki, M. (2022). Continuous Auditing of Artificial Intelligence: a Conceptualization and Assessment of Tools and Frameworks. In *Digital Society* (Vol. 1, Issue 3). Springer Science and Business Media LLC. <https://doi.org/10.1007/s44206-022-00022-2>

24 ETSI TS 104 008 V1.1.1 (2026-01)

25 DIN SPEC 92006

consistently, to have predictable run times and to handle error states transparently.

A further aspect is how measurement uncertainty and reproducibility are handled. Relevant sources of uncertainty are identified and, as far as possible, quantified. Non-deterministic influences and their contribution to the variance of results are documented in order to support the comparability of test results.

Finally, the concept covers systematic logging of relevant parameters, configurations and results, so that individual test runs can be reconstructed and integrated into recurring or automated testing processes.

2.2.3 Reference datasets

Standardized reference datasets²⁶ are another central element of a reproducible AI testing infrastructure. They make it possible to assess the performance of AI models under controlled, comparable conditions, regardless of who carries out the test. Reference datasets have to be clearly documented in terms of origin, composition, known biases and scope of validity. In the context of physical and technical testing, such datasets are comparable to certified reference materials in conventional metrology.

In Germany, the development of such an infrastructure is being driven by organizations such as the Physikalisch-Technische Bundesanstalt (PTB) and by initiatives such as QI-Digital. One concrete example is the pilot project “AI in medicine”, which is evaluating standardized test datasets for medical imaging AI. SMART standards support this by pointing to reference datasets via persistent identifiers and thus enabling automated integration into testing workflows. For tools, reference data and semantic models to work together in a consistent test workflow, an integrating technical and semantic layer is needed.

2.2.4 AI testing platforms

AI testing platforms are technical frameworks that enable testing processes for AI systems to be defined, executed and documented in a structured way. They provide an infrastructure in which test methods can be specified and executed reproducibly, regardless of their substantive focus. At the heart of this is the technical operationalization of test workflows: (standardized) testing tools, models, datasets and configurations are treated as clearly delimited components described in machine-readable form, whose interplay is defined explicitly. This provides mechanisms for integrating and executing any test method consistently.

In addition, a testing platform can provide mechanisms for versioning, logging and archiving. Every execution of a test workflow can be documented with unambiguous references to the components, parameters and artifacts used. This produces a traceable and repeatable test description that is independent of the substantive orientation of the individual tests. The platform thus acts as a generic, content-agnostic framework ensuring the formal structuring and technical consistency of AI testing processes.

Developers or operators of testing platforms can offer AI testing as a service, for example centrally (off-premise), where the party being tested provides its AI system or individual components for testing, or on-premise, on that party's own infrastructure. Companies can also implement the concept of an AI testing platform in-house by taking these central concepts into account when setting up an MLOps infrastructure.

Different actors can operate such platforms. Accredited testing bodies and conformity assessment bodies can run them as part of their service portfolio. Research institutions and national metrology institutes (such as PTB, including as part of QI-Digital) develop platforms that serve as references. Finally, platform building blocks can be available as open-source components and operated internally, for example within a company's own MLOps stack. In practice, hybrid arrangements are likely to emerge in which specialized platforms for particular fields of application (e.g. medical imaging, credit scoring) are integrated with further testing infrastructure.

26 DIN SPEC 92007

2.2.5 Semantic modeling

One fundamental method for keeping AI testing scalable as the number and complexity of tests grow is the use of ontologies and knowledge graphs.

Ontologies create a formal, machine-readable structure for what is relevant when testing AI systems. This includes data types, model classes, training procedures, metrics, risks, assumptions and in particular requirements from standards and legislation.

This formalization allows objects relevant to testing to be standardized and made comparable. Instead of defining every test individually, testers can draw on reusable ontological building blocks that specify what has to be tested, for example data provenance, fairness criteria or robustness, and how these aspects relate to one another. In this way, for example, concepts and requirements from legislation can be mapped onto standards.²⁷

Concrete knowledge graphs enriched with specific knowledge build on ontologies and make it possible to link concrete information about data, models, pipelines and their lifecycle. Instead of working through isolated checklists, testing systems can analyze dependencies in a targeted way, for example which training data influenced a model, which pre-processing steps were applied, or which models depend on the same data sources. Requirements can be modeled formally and linked directly to model properties and data, for example to risk analyses or to the documentation of AI systems.²⁸ On this basis, automated queries can be formulated that identify inconsistencies or breaches of rules. New models or data flows are then integrated automatically into the existing knowledge graph and evaluated against known test rules.

2.2.6 Standardized test reports

A further building block of scalable AI testing is the automated generation of test result documentation in standardized formats. Test results, evidence and conformity status are no longer transferred into reports manually,

but exported directly from the testing systems. Relevant formats include digital certificates of conformity (D-CoCs), model cards or an “AI passport” containing structured, machine-readable descriptions of an AI system and its test results. In the future, such evidence may also be provided or referenced through digital product data ecosystems such as Digital Product Passports (DPPs).²⁹ Standardized reporting formats allow test results to be passed directly to supervisory authorities, notified bodies or downstream actors in the value chain, and thus reduce the effort of reporting and reusing evidence.

2.2.7 SMART standards

While the technical elements described so far address the execution and documentation of tests, a second dimension of effort concerns accessing the normative basis itself. So far, the path from a requirement in a standard to an executable AI test has involved manual effort that is sometimes considerable and prone to error. Relevant standards have to be identified and interpreted, requirements operationalized, suitable test methods selected and code written (see the figure on the next page). This effort is potentially repeated with every update of a standard and every change to the AI system.

The scenarios in the next chapter describe how SMART standards could help to reduce this effort and thus contribute to the scalability of AI testing.

27 Hernandez, J., Golpayegani, D., & Lewis, D. (2025). An open knowledge graph-based approach for mapping concepts and requirements between the EU AI act and international standards. *AI and Ethics*, 5(5), 4463–4474. <https://doi.org/10.1007/s43681-025-00708-6>

28 Golpayegani, D., Pandit, H. J., & Lewis, D. (2022). AIRO: An Ontology for Representing AI Risks Based on the Proposed EU AI Act and ISO Risk Management Standards. In *Studies on the Semantic Web*. IOS Press. <https://doi.org/10.3233/ssw220008>

29 <https://www.qi-digital.de/en/digital-certificates>

Standards-related activities in the testing of AI

Identify standards



- To define the framework for testing in conformity with standards, the applicable quality dimensions, legal requirements and relevant standards content first have to be identified.
- For example by analyzing the relationship between the legal basis and harmonized standards, tracing requirements across different standards, or identifying content in other documents (e.g. technical specifications, technical reports, guidelines) that concretizes abstract requirements.
- Results of manual research may have to be revised regularly in light of new publications and updates to standards.

Obtain and integrate content



- Once identified, standards and relevant contents within have to be obtained.
- Since many standards are not freely available, access rights first have to be acquired or demonstrated.
- Granular contents that are relevant for testing, such as requirements, definitions or formulas, have to be extracted and then transferred to and integrated into an organization's own systems (e.g. requirements management systems).
- The content has to be kept up to date with changes in the underlying documents.

Operationalize tests



- Functional and non-functional requirements have to be derived from the extracted standards content, and measurable criteria and indicators identified or selected.
- For tests that can be automated, testing services or code libraries have to be selected, parameters set, algorithms chosen and, where necessary, test code implemented.
- Relevant artifacts such as test data have to be obtained and integrated.

Document conformity



- Test results, the methods used and the versions of standards referenced have to be documented in such a way that auditors, notified bodies or supervisory authorities can follow and verify them.
- So far this documentation has been produced largely manually, often in proprietary formats and with redundant recording across reports, certificates and requirements management systems.
- Every time the AI system or the underlying standards are updated, this documentation has to be revised and synchronized across the various target systems.



Labor-intensive activities related to the use of standards can be simplified with SMART Standards.

→ Scenarios: chapter 3

3

SCENARIOS FOR TESTING AI SYSTEMS WITH SMART STANDARDS

3 SCENARIOS FOR TESTING AI SYSTEMS WITH SMART STANDARDS

SMART standards can make AI testing more scalable by making content that is relevant to testing available in machine-readable and machine-interpretable form, linking it to relevant resources and simplifying its integration into relevant processes. A testing environment that is enriched with SMART standards can:

1. simplify the identification of relevant standards and content,
2. provide information on changes to content and on the publication of new relevant content,
3. deliver standards content in a form suited to integration into digital processes,
4. make relevant content reliably usable in automated form (for example for processing by AI),
5. establish the link to further aids for operationalizing tests, and
6. make it possible to document test results automatically in a reproducible and interoperable way.

This can be achieved in particular through integration with the concepts and resources presented in chapter 2:

- Standards content can be made available more effectively for automated processing, so that it is readily accessible in MLOps processes and for the logic of continuous conformity testing.
- Where possible, a link can be established between SMART standards and quality-assured resources such as testing tools, reference data or entire testing services (platforms), in order to narrow the gap between standards and their operationalization.

From a user's perspective, the following technical and organizational scenarios show how this contribution could be established and employed. In the scenarios, which build on one another, the use of standards evolves from research-intensive, document-based work, through the integration of standards content via API, to AI-supported test management within an ecosystem for AI quality.

USAGE SCENARIOS:

3.1. API-based delivery of machine-readable and machine-interpretable standards content

In addition to the standard as a PDF, machine-readable SMART content is provided via API. Semantic content enables the intelligent identification of relevant standards and their linking to other data sources. Standards content can be transferred directly into information systems and kept up to date. Content is reliably available for operationalizing tests and generating test reports.

3.2. Reliably AI-generated test artifacts

AI testing is supported by an AI assistant which, on the basis of the testing context and the relevant standards, identifies the applicable requirements, proposes test plans, produces test reports and, where possible, generates executable test code. The reliability of the AI's use of standards can be increased by integrating testing tools and services and by using unambiguous SMART content.

3.3. Integration into an ecosystem for AI quality

In a shared infrastructure, SMART standards become integrated with resources that contribute to a more uniform operationalization of AI testing. Tests, testing actors and the required tools, data and services are integrated in a trustworthy way through a digital AI quality infrastructure.

CREATION SCENARIO:

3.4 Scalable generation of SMART content and operationalization content

To meet a growing need for various kinds of SMART content, AI is employed to help with content generation, supported by expert validation. (Referenced) content that supports the operationalization of the standard can be contributed from the ecosystem for AI quality.

Scalable AI testing with SMART standards

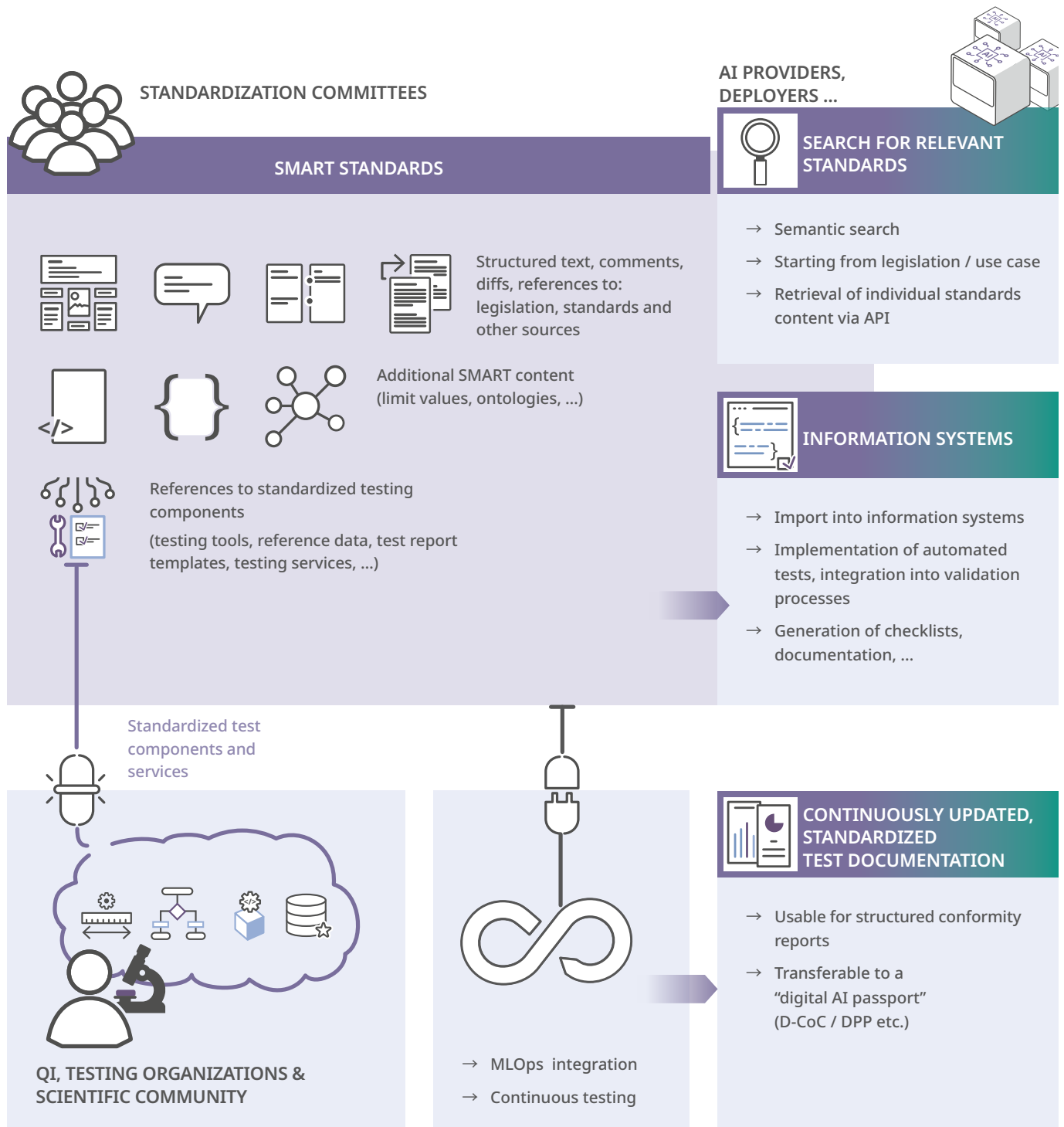


Figure 3

3.1 API-based delivery of machine-readable and machine-interpretable standards content

The first scenario focuses above all on access to individual items of standards content and their processing: instead of users obtaining standards manually as unstructured text documents through web portals, standardization organizations and their publishers provide them as structured SMART content via APIs.

Individual items of standards content, such as requirements, can be identified and extracted granularly and integrated into digital processes without manual pre-processing. Related content is integrated by resolving the references between standards.

Content is provided in versioned form; updated content can be obtained automatically. There is a “standards changelog”, and “diffs” can be displayed for different versions of a standard. For updates and new publications, needs analyses and procurement are simplified.

Individual components of a standard can be enriched with additional metadata such as thematic grouping or comments. Complex content is made available in specialized machine-readable formats, for example mathematical formulas in MathML (see Table 1). Important content that has so far been provided only in natural language can be made machine-readable through additional content, for example by listing values or value ranges explicitly (XML/JSON).

Semantic content such as ontologies, vocabularies and constraint models (e.g. in RDF, OWL, SHACL, SKOS) is delivered in addition to the actual standards content. It makes relationships between terms, requirements and test objects explicit and allows standards content to be linked automatically to other data sources (requirements management, knowledge graphs, test artifacts). Concrete use cases for this arise above all in the following scenarios.³⁰

This makes AI testing easier and more reliable to implement:

Identifying relevant standards intelligently

On the basis of the information about the system, the applicable legal requirements and harmonized standards are determined with system support, including the reference chain from legislation to harmonized standards and on to referenced standards. Relevant standards content (individual requirements, definitions, references) is obtained selectively. Semantic content also allows more complex relationships between standards and the organization's own testing context to be represented:

- Complex reports for identifying relevant standards are created in the organization's own information systems and updated automatically at regular intervals. For example: “show me all standards content from particular standardization organizations that is referenced directly or indirectly by the AI Act and carries the tag ‘data quality’”.
- Changes to standards or new editions are detected and displayed automatically, which replaces the need to search manually for new publications.

Integrating standards content automatically

Standards content identified as relevant is transferred automatically into local information systems as granular, machine-readable units. It can be integrated into requirements management, for example as overarching compliance requirements, as constraints for epics or as acceptance criteria for user stories. Every integrated requirement remains unambiguously linked to its source (standard, clause, version) and can be updated automatically.

- All requirements from selected standards that apply to an AI system are transferred into the backlog via API (Jira, Confluence, GitHub, GitLab, ...), including requirement text and type (must, shall, should) as well as references to standards and legislation (traceability). All requirements and other texts extracted from standards can be updated when a new version of a standard is published, and the changes can be displayed.

30 VDE SPEC 90010 V1.0, Guide to creating executable semantic standards (SemNorm)

Content offered (SMART content)		
SMART content	Description	Format examples
Standards text as granular units	Structured, classified textual components of a standard, e.g. individually addressable requirements with a unique ID, clause and version	XML (ReqIF)
Cross-references between standards and legislation	Explicit relations, e.g. AI Act → harmonized standard → further referenced standards	XML (ReqIF), RDF
Additional metadata	Tags, comments, ...	XML (ReqIF), RDF
Graphics		PNG, SVG, ...
Directly usable values	Numerical limit values, value ranges, thresholds, classification boundaries, reference ranges, units, categorical specifications, ...	XML, JSON, CSV, RDF
Mappings/translation tables	Correspondences between classification systems, between units or measurement methods, between versions, ...	XML, JSON, SKOS, RDF
Simple algorithmic and operational content	Formulas, test algorithms, decision trees, state machines, ...	MathML / LaTeX, DMN, BPMN, ...
Semantic information	Declarative semantics through ontologies and structured requirements, operational semantics through executable models	RDF, OWL, SHACL, MathML, DMN, BPMN, ...

Interfaces	
Interface	Description
Search API	Search in full text and metadata, semantic search
Content API (content as a service)	Retrieval of all content of a standard, or targeted retrieval of a clause or item of additional content by identifier, versioned
Change API	Notification of changes to standards; structured delta between versions.
RMS integration	Import of requirements and other text sections into relevant requirements management systems (Jira, Confluence, GitLab and others) via ReqIF (converter/wrapper), or where applicable directly via the CaaS API

Table 1: Digital delivery of SMART content for AI testing (technical implementation scenario)

Operationalizing standards content reliably and documenting tests in a structured way

Standards content relevant to testing is operationalized and integrated into CI/CD and MLOps pipelines as automated tests and monitoring checks. Results feed continuously into automatically generated conformity documentation.

- Standards content is provided directly in working environments, for example in notebooks or IDEs, where relevant, clauses are pointed out inline while validation code is being written. Requirements from standards are thus visible early and continuously, instead of being picked up again late in the project.
- Valid value ranges, parameter settings etc. for tests are integrated automatically from standards into

implementations, so that transfer errors from manual copying are avoided.

- Because SMART content and test implementations are linked, it is possible to trace at any time which tests relate to which standards content. In the event of deviations, test triggers can automatically start workflows such as retraining or alerts.
- Because tests and requirements from standards are unambiguously linked, evidence can be aggregated, versioned and evaluated with system support, as a basis for first-party conformity assessments or the preparation of external audits.

3.2 Reliably AI-generated test artifacts

The ability of generative AI systems to produce functional software code from structured input is currently changing the world of software development substantially. It also opens up new possibilities (and risks) for standards-based AI testing.

Using generative AI to support the implementation of standards-based AI testing as reliably as possible is a logical step toward scalability.³¹ In this scenario, machine-readable standards with unambiguous, quality-assured content are the input for an LLM agent that automatically generates checklists, test plans and test reports or directly executable test code, and that also carries out tests independently.

For popular AI standards, parts of this scenario can already be implemented today with available language models.³² The major challenge is that LLMs, as statistical language models, do not automatically offer the high degree of reliability that working with standards requires. Models generate plausible wording on the basis of training data, but have neither a verifiable knowledge base nor a defined context of use for the statements they reproduce.

An LLM could therefore paraphrase and distort requirements; it could use terms without referencing their standardized definition; and when in doubt it invents plausible-sounding passages of text or even values and formulas (“hallucination”).

Nonetheless, using AI to automate conformity checks on the basis of rule sets is a concept pursued in various regulated domains. There are corresponding research projects in areas such as the GDPR, financial regulation and building law.³³ Semantic modeling intended to contribute to automating

“compliance checks” is also emerging in the field of AI regulation.³⁴

To increase the quality of test artifacts generated from standards, it is worth considering the use of neurosymbolic AI, that is, a combination of a statistical and a rule-based approach.³⁵ Such an AI combines the flexibility and generalization capability of neural networks with the verifiability and formal guarantees of symbolic systems, and could deliver better results drawing on additional semantic information and further, deterministically used structured data from SMART standards.

Semantic information from SMART standards could feed into the generation of test artifacts in various ways. In line with a “knowledge infusion” concept³⁶, a specialized “standards LLM”³⁷ could be trained on the semantic information of all (relevant) standards in addition to natural-language standards content (“knowledge infused learning”). Another conceivable approach integrates the relevant semantic information into the generation process of a generic (foundation) language model as required (“knowledge infused inference” / “inference stage enhancement”).

31 The FAICE project (Framework for AI-Enabled Collaborative Engineering, [prostep ivip](https://www.prostep.org/projekte/framework-for-ai-enabled-collaborative-engineering-faice)) is pursuing a related approach outside AI testing. It is developing a “standards chatbot” intended to provide AI-supported, granular access to standards content.

See <https://www.prostep.org/projekte/framework-for-ai-enabled-collaborative-engineering-faice>

32 Example: generating a checklist in accordance with ISO 42001 and checking it against local documentation using Claude / ChatGPT etc.

33 E.g.: Using Large Language Models for the Interpretation of Building Regulations. (2024). Journal of Engineering, Project, and Production Management. <https://doi.org/10.32738/jepm-2024-0035>

34 E.g.: AIRO (AI Risk Ontology), <https://delaramglp.github.io/airo/>

35 Garcez, A. d'Avila, & Lamb, L. C. (2023). Neurosymbolic AI: the 3rd wave. Artificial Intelligence Review, 56(11), 12387–12406. <https://doi.org/10.1007/s10462-023-10448-w>

36 Pan, S., Luo, L., Wang, Y., Chen, C., Wang, J., & Wu, X. (2024). Unifying Large Language Models and Knowledge Graphs: A Roadmap. IEEE Transactions on Knowledge and Data Engineering, 36(7), 3580–3599. <https://doi.org/10.1109/tkde.2024.3352100>

37 There was a corresponding pilot project at IDiS, although it initially used neither semantic information nor neurosymbolic AI.

Broadly speaking, two usage scenarios can be derived:

a) A central “standards LLM” trained on SMART standards
A neurosymbolic AI model trained on all available SMART content is offered centrally (e.g. by one or more standardization organizations), for example as a service (e.g. as a chatbot and/or via API). The content to be tested and other context-relevant content from the user organization are made available to the AI at run time for inference.

And/or:

b) Generic LLM + SMART standards RAG
A user organization uses a generic (possibly neurosymbolic) language model equipped with the local context data of the system to be tested. At run time the model is enriched via RAG/MCP mechanisms with relevant SMART standards provided by standardization organizations, and can use semantic content for inference.

Besides semantic information, in both cases the structured additional content provided by SMART standards is also used when generating test artifacts:

Integrating structured additional content as “ground truth”

Precisely defined values, mappings and clearly defined algorithmic and other operational content (see Table 1) are adopted by the AI as structured data from SMART standards. There is no need to extract them from the text or reconstruct them from training knowledge. The model takes the values directly and uses them to generate test artifacts. For example: inserting limit values directly, implementing metrics from mathematical formulas. Values are adopted from the standard unchanged and remain traceable back to the relevant clause. In this way “normative truth” and generation logic are separated, so that “hallucinations” can be reduced at this point.

Referencing further testing resources and tools

It is foreseeable that standards and the corresponding additional content (such as operational semantic information) will not always be able to offer information sufficient for fully operationalizing a test. Standards might, for example, describe test concepts that can be implemented in different ways for very specific testing contexts. Precise implementation instructions (e.g. the

choice of code library for implemented standard algorithms) go beyond the level of detail of a standard, but can be decisive for the eventual test results and have an impact on comparability, reproducibility and correctness.

Such uncertainties can and should be flagged by the AI for quality assessment by a human. An automated quality assessment could, however, also be included implicitly through the use of externally quality-assured resources that offer ready-made implementations for particularly complex, ambiguous or under-specified requirements:

For specialized tests such as statistical tests for fairness, robustness tests or security analyses, standardized and where applicable certified testing tools (cf. DIN SPEC 92006) can be referenced as services directly in the SMART content (see also the next scenario).

Application scenario: AI-generated test content

The scenario presented enables a range of use cases based on SMART standards. The focus is always on using an AI in the context of AI testing to process information from standards, generate test artifacts (code, checklists, test plans) and carry out tests on that basis.

Responsibility for the quality of the tests remains with humans. Generated artifacts are reviewed and approved by qualified specialists before they are used in production. AI accelerates implementation without replacing expert judgment.

- Through chatbot-style interaction, testers state which AI system is to be tested and against which standards. An AI identifies all testable requirements and generates a structured test plan in which references to the underlying clauses are given for every testing activity.
- Requirements that can be tested automatically (e.g. through defined thresholds, metrics, mandatory fields in documentation) are distinguished from those that require human judgment.
- For requirements that can be tested automatically, test results are proposed (if they can be derived from the context or from local documentation and code), or test code is generated (and where applicable executed).
- Requirements that cannot be tested automatically are checked through question and answer, or by including further documents in the chat context.

- The test plan can be integrated into the local infrastructure for regular execution and connected to an AI agent: automated tests are integrated into the MLOps framework. Documentation on tests to be carried out manually is stored in an information system. An AI agent reminds testers by e-mail or ticket system at the relevant times to carry out the test.

3.3 Integration into an ecosystem for AI quality

In scenarios 1 and 2, companies access a wide range of resources from different providers: SMART standards from standardization organizations, testing tools and reference datasets from testing institutes and research institutions, accreditation credentials from the responsible bodies, and further content from the AI development value chain. A large number of open-source tools and datasets for testing AI systems already exist today. The OECD Catalogue of Tools & Metrics for Trustworthy AI alone lists several hundred entries.³⁸

If such resources are used for testing conformity with standards, central quality assurance by qualified bodies can offer several advantages:

- **Operationalization and reference to standards.** When standards are operationalized, the choice of algorithms, libraries, parameters, thresholds and validation datasets remains a decision for the user, as does the reference to specific clauses. A qualified body could assess and validate open-source libraries and other implementations for testing against a particular standard and bind the result to the resource as a signed, machine-readable suitability attestation. Users can then determine directly from the SMART standard which tools in which versions are suitable for a requirement, instead of having to make that assessment themselves.
- **Promoting the ecosystem.** With such central verification, open-source projects and tool providers

gain an incentive to align their implementations with established quality requirements and to apply for a corresponding assessment. This strengthens the community and prevents testing tools from having to be redeveloped in parallel in different organizations.

- **Comparability of test results.** Without agreed tool versions, default parameters and reference data, identical input can produce divergent results. Conformity assessments and mutual recognition, however, presuppose reliable comparability.
- **Reliability of the tools themselves.** Testing tools are software and can be faulty. A faulty fairness metric or a numerically unstable robustness test produces systematically wrong statements. In safety-critical fields of application their reliability is no less important than that of the systems being tested.
- **Protection against data contamination.** If reference and test datasets are made publicly available, they find their way into later models through training data and lose their value as independent test resources. Central management allows test data to be kept available in a targeted way and to be made accessible only through controlled mechanisms.
- **Efficiency on the user side.** Without reliable recommendations, every company has to build up its own capacity for researching, assessing and validating external tools, which overburdens smaller providers in particular.

³⁸ <https://oecd.ai/en/catalogue/tools>

Assessment by qualified bodies turns this individual burden into a collective task.

- **Accreditation and supervisory conformity.** Accredited conformity assessment bodies have to demonstrate the suitability of their test resources; without central quality assurance every body carries this burden alone. A shared assessment also facilitates mutual recognition across sectors and national borders.

The digital quality infrastructure is already responding to this situation with concrete initiatives. The QI-Digital roadmap formulates the vision of “trustworthy AI through testing platforms and an AI passport”, in which AI systems are executed in secured environments of a testing platform or integrated through defined interfaces, with automated comparison against reference values. Initial preparatory work on the platform, reference data and methodological basis has already been done.³⁹ These initiatives address the central quality assurance of tools, data and services as part of the ongoing transformation of the quality infrastructure.

This is where the present scenario comes in, extending the central building blocks of the QI-Digital vision to include the controlled, trustworthy brokering of these resources between many providers and consumers.

Working out the details raises a number of open questions that can be answered only through a coordinated approach among those involved:

- How can tools, datasets and services be made discoverable and comparable across provider boundaries?
- Who takes on the quality and suitability assessment, and how are recommendations by qualified bodies made visible and verifiable?
- How is the link between tool and standard made explicit and kept up to date when standards are updated?
- Under which access conditions, licenses and billing models are resources provided, including where their use is subject to a charge?

- How are reference and test datasets provided in such a way that their integrity is maintained over time despite repeated use?

A data space as shared infrastructure

Since individual agreements between the various actors would involve a great deal of effort, it makes sense here to use a data space as a shared infrastructure in which discoverability, trust, access control and billing are governed uniformly for all participants. **Table 2** gives a structured overview of the roles that could be involved. **Table 3** lists various resources relevant in the ecosystem.

Three central principles of this scenario are:

- **Data sovereignty.** Every actor involved in the QI retains control over the resources it provides, without having to transfer them to a central platform. Standardization organizations can themselves enforce licenses and access models for standards content. Providers of AI testing tools can offer paid licenses or restrict access to accredited bodies only. AI manufacturers can protect model weights and proprietary implementations. Providers of reference datasets can safeguard them against contamination.
- **Federated discoverability.** SMART standards from standardization organizations, AI testing tools and reference datasets, and testing services from testing institutes remain with their respective providers and are made discoverable through a distributed catalog. The reference to specific clauses and AI application domains can be carried along in machine-readable form.
- **Trust through verifiable identities.** Accreditations, certifications by conformity assessment bodies and suitability assessments of AI testing tools by qualified bodies are held as signed verifiable credentials. Users can thereby establish whether a tool recommendation really comes from the stated body for the stated standard, and access to certified testing services can be limited to accredited bodies.

39 https://www.qi-digital.de/fileadmin/user_upload/website/Roadmap/QID_Roadmap_V1.0_20251008.pdf. Among other things, the roadmap cites the ongoing development by PTB of a platform for the objective testing and quality assessment of AI algorithms, and the PTB pilot project “Artificial intelligence in medicine” with work on synthetic reference data and use-case-specific reference datasets. As a methodological basis it references the METRIC framework, cf. Schwabe, D., Becker, K., Seyferth, M., Klauf, A., Schäffter, T. (2024): The METRIC-framework for assessing data quality for trustworthy AI in medicine: a systematic review. npj Digital Medicine 7, 203, <https://doi.org/10.1038/s41746-024-01196-4>.

Role	Actors (German examples)	Contribution to the ecosystem
Standardization organizations	DIN, DKE	Provision of SMART standards; source of normatively anchored vocabularies for the semantic description of tools, datasets and attestations
Providers of testing tools and reference data	Research institutions, open-source communities, commercial providers	Development and provision of AI testing tools and reference datasets; linking to clauses through catalog metadata
Accreditation bodies	DAkkS	Accreditation of conformity assessment bodies; issuing of verifiable accreditation credentials
Qualified bodies	Bodies with demonstrated technical competence in assessing AI testing tools	Suitability attestation of tools for particular clauses; issuing of signed verifiable credentials
Conformity assessment bodies	Accredited and notified bodies, and cross-cutting governance projects	Provision of compliance catalogs and assessment workflows, carrying out conformity assessments using quality-assured tools; operation of certified testing services, where applicable in compute-to-data mode; harmonization of the necessary services through a domain-specific “rulebook”
AI providers	Companies that develop AI systems and place them on the market	Provision of the models to be tested (locally or in compute-to-data mode); consumption of tools, reference data and attestations from the data space
AI users	Organizations that deploy AI systems	Selection and procurement of conformant AI components on the basis of the references resolved in the data space
Supervisory authorities	Market surveillance and data protection authorities	Spot checks; verification of attestations and certificates of conformity
Platform operators	Consortium of a data space based on the Quality-X concept framework	Operation of the federated infrastructure including connectors, catalog, trust and vocabulary services, with the involvement of QI institutions

Table 2: Possible actors and roles in an ecosystem for AI quality

Building blocks and extensions for AI testing in the Quality-X concept

With Quality-X, a data space concept specifically for the quality infrastructure is already emerging.⁴⁰ Quality-X considers central building blocks to which an ecosystem for quality-assured AI testing could connect, such as connectors for sovereign exchange (for example Eclipse EDC), a federated catalog for discovering distributed resources (for example on the basis of DCAT-AP), shared vocabularies for their semantic description, and identity and trust services for verifying actors and certificates.

Working these building blocks out in detail for AI testing lies beyond the scope of this white paper. Some starting points for SMART standards can already be identified, however:

- Connectors enable computations without sensitive resources having to be shared. In compute-to-data mode, model weights remain with the manufacturer and black box test datasets with the qualified body. Only the test request and the signed result are exchanged. A requirement anchored in a SMART standard thus leads directly to a test execution whose provenance, tool version and input data are logged at the connector.
- Federated catalog: for AI testing tools and reference datasets, machine-readable references to clauses of standards and to AI application domains have to be stored (additional metadata in the DCAT-AP vocabulary), so that suitable resources can be resolved from a specific requirement.⁴¹ In concrete terms, starting from a robustness requirement in an AI-related SMART standard, a user could resolve both suitable testing tools and matching reference datasets, each with a reference to the corresponding suitability assessment and the application domain covered.
- In the case of the shared vocabularies, SMART standards could be used both for content referenced in the data space and as a source of normatively anchored terms

(for example for AI test procedures, quality dimensions, application domains) with which tools, datasets and attestations are described uniformly. In this way the same term defined in a standard would also be used in the catalog entry of a testing tool and in the credential of a suitability assessment, without a separate translation layer between standard and data space having to be maintained.

- The organizational and legal requirements arising from SMART standards and data space processes, as well as restrictions in confidential computing, compute-to-data or secure processing environments (SPE), could be unified using the Open Digital Rights Language⁴², for example.
- In the case of the trust services, the familiar verification of accreditations and certificates of conformity, for which a machine-readable format is already being developed within QI-Digital in the form of the Digital Certificate of Conformity (D-CoC), would be joined by the need to carry suitability attestations for AI testing tools in machine-readable form as well (as signed verifiable credentials held with the tool in the catalog). Before use, a consumer could check automatically whether a tool has been assessed by a qualified body for the applicable standard in the version at hand, without having to gather this evidence manually from separate sources.
- Beyond today's functionality, tools themselves would have to be discoverable, obtainable and executable, increasingly also in agent-based form (with interfaces such as the Model Context Protocol). An agent-based testing tool could read the applicable requirements from a SMART standard, resolve matching sub-tools and datasets in the catalog and initiate the test at the location of the data, without these steps having to be chained manually. The marketplace envisaged in Quality-X would be an obvious starting point. The technical implementation would have to be clarified in detail.

40 Quality-X: federated platform for the quality infrastructure, QI-Digital roadmap, chapter 2.4.2; additional information at <https://www.qi-digital.de/digitale-qi/quality-x>.

41 With Daseen, the MISSION KI project has already produced a proof of concept for a federated dataset catalog to which an extension covering tools, suitability attestations and references to standards could connect: <https://github.com/Mission-KI/Dataset-Search-Engine>

42 <https://www.w3.org/TR/odrl-model/>

AI-supported generation of SMART content

MANUAL STANDARDS DEVELOPMENT & QUALITY ASSURANCE OF SMART CONTENT

AUTOMATIC GENERATION OF SMART CONTENT DRAFTS

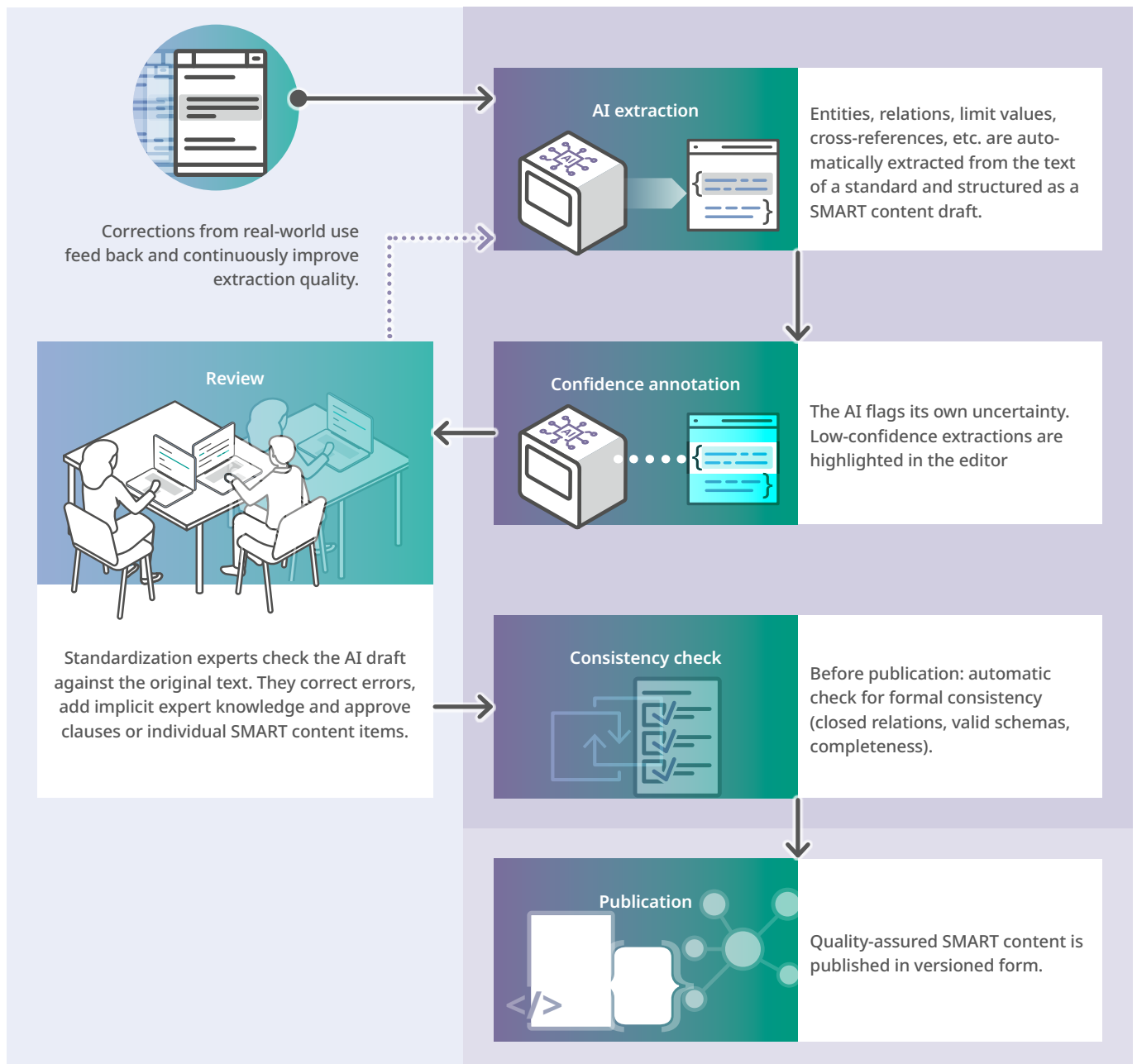


Figure 4

→ In combination with domain-specific data spaces⁴³, a data space of the digital quality infrastructure for AI testing could be adapted to the requirements of the respective domain. A credential issued in one data space could be verified and recognized directly in another, without the issuing body having to be involved again. Compliance approvals and test evidence would feed into a shared workflow across data spaces as verified credentials, without introducing manual handovers at the boundaries of the infrastructures involved. This would create a common ecosystem of specialized data spaces. The Dataspace Protocol⁴⁴, for example, could be used here for the bilateral negotiation of usage rights and compliance requirements between decentralized participants. Verifiable credentials⁴⁵ and the Decentralized Claims Protocol⁴⁶ could form the basis for organizational interoperability.

AI testing locally or in the data space

Two complementary scenarios are conceivable for executing AI testing tools in the ecosystem, differing in where the computation takes place: at the user's location or at the location of the resource to be protected. Both scenarios have their place and depend on the protection needs of the resources involved.

In the local scenario, a testing tool would be obtained from the data space and executed in the user's infrastructure. This scenario would be the natural choice where neither the models tested nor the test data are particularly in need of protection, for example where a manufacturer tests its own models against publicly available reference data or provides proof of conformity in a controlled internal environment. The early phase of conformity testing, in which a tool is first

tried out with in-house material, would also be suited to local execution.

In the compute-to-data scenario, data and model remain at their place of origin and the testing tool is executed there. Only the test request and the signed result are exchanged. This scenario allows a symmetrical application of data sovereignty that is particularly relevant in AI testing. Model weights remain with the manufacturer, so that the AI tested does not have to leave its infrastructure. Black box test datasets remain with the qualified body, which protects their integrity against repeated use and contamination. Certified testing services run by accredited bodies could accordingly be operated in such a way that their internal logic cannot be read out from within the data space. In the end the user would receive a traceable result without gaining access to the tool's internals or to the reference data.

Both scenarios could be bundled in dedicated testing platforms operated by qualified bodies or provided as hostable open-source software. From the point of view of the data space, such platforms would be addressable through the same connectors, catalog entries and suitability attestations as individual tools.

Concrete preparatory work for compute-to-data mode can already be found in existing data space initiatives. The Extended Dataset Profile Service (EDPS) is a service executed at the location of the data, where it produces standardized data quality profiles and makes them discoverable through the federated search engine Daseen, without the original data leaving its environment.⁴⁷ There are also open-source prototypes of AI testing tools executed on-premise that check data exchange processes against requirements from extensive rule sets.⁴⁸

43 such as the European Health Data Space (EHDS) or the Sphin-X data space; Sphin-X: one data space for health, <https://sphin-x.de/>

44 ISO/IEC DIS 26450

45 <https://www.w3.org/TR/vc-overview/>

46 <https://github.com/eclipse-dataspace-dcp/decentralized-claims-protocol>

47 Extended Dataset Profile Service (EDPS) and the federated search engine Daseen, developed within the MISSION KI initiative as a compute-to-data service. Available as open source at <https://github.com/Mission-KI/Dataset-Search-Engine>; additional description at <https://mission-ki.de/de/datenraeume-vernetzen/data-set-search-engine>.

48 Compliance Monitor, proof of concept of an AI-based on-premise testing system for data transactions under the requirements of the Data Act, Data Governance Act and DORA, developed by Noxtua (formerly Xayn) within the MISSION KI initiative. Language model and documentation available at <https://huggingface.co/ACATECH/ncos>; further information at <https://mission-ki.de/de/datenraeume-vernetzen/compliance-monitor>.

Resource	Description	Format/standard (examples)
SMART standards	Machine-readable standards content with requirements, test specifications and definitions	XML (ReqIF), further formats to follow
Domain-specific vocabularies and ontologies	Concept systems for AI test procedures, quality dimensions and application domains, derived from SMART standards or supplementing them	RDF, OWL, SKOS
AI testing tools	Software for automated testing of AI systems against requirements from standards; lists of compliance requirements for specific usage scenarios	Open-source repositories, container images (OCI); catalog description in DCAT-AP / JSON-LD
Reference datasets	Curated datasets for training, validation and testing, including standardized data quality profiles	Data formats customary in the domain (CSV, Parquet etc.); dataset profiles per the EDPS schema
Black box testing services	Testing services that contain protected test data or protected tool logic and are addressed in compute-to-data mode (testing platforms)	API endpoints with OpenAPI specification; called through data space connectors
Suitability attestations	Qualified bodies' assessments of a tool's suitability for particular clauses of standards	Verifiable credentials in JSON-LD
Accreditation credentials	Accreditation certificates for conformity assessment bodies	Verifiable credentials in JWT & JSON-LD
Certificates of conformity	Results of completed conformity assessments; carried in a DPP or "AI passport" throughout the product lifecycle	Digital Certificate of Conformity (D-CoC), signed JSON-LD documents, DPP data models
Asset Administration Shell	Description of physically embedded AI components with references to vocabularies from standards and classification systems via semanticIds	AAS submodels in JSON, XML or AASX; semanticIds according to IEC CDD and ECLASS

Table 3: Resources and formats in a data space for AI quality (format details are examples and not a definitive specification)

3.4 Scalable generation of SMART content and operationalization content

The scenarios described above presuppose high-quality SMART content and supplementary operationalization content, the range of which is considerable. Creating this entire body of content purely manually would not be scalable. To keep this scope manageable, AI-supported generation with human validation could be combined with creation shared among several actors with different areas of competence.

AI-supported generation with human validation

The approach proposed follows the human-in-the-loop principle. AI handles the laborious initial extraction; standardization experts assure quality (see Figure 5). The basis is an editor such as the Online Standards Development Tool (OSD)⁴⁹, supplemented by AI support and validation. The TermAnalyzer developed by DKE for the quality checking and semantic harmonization of terms in standards can also be used here.

AI models are suited above all to tasks that are highly structured and large in volume, such as extracting entities and terms from the text of a standard, identifying explicit cross-references, deriving constraints from requirement statements, showing changes in content between versions of a standard, and checking consistency across several standards. Human judgment remains necessary wherever expert judgment is central, in particular when resolving normative ambiguity, adding implicit expert knowledge, deciding borderline cases, judging the impact of changes to standards, and setting priorities where standards compete.

The same human-in-the-loop principle could potentially be applied to generating ontologies and knowledge graphs. LLMs can extract concept hierarchies, relations and formal constraints from the text of a standard and express them as a structured draft graph, typically in hybrid pipelines combining LLMs with domain-specific embeddings, retrieval-augmented generation and fine-tuning.⁵⁰

A step-by-step approach with an increasing degree of formalization makes sense. At the first stage, terms, definitions and synonyms are produced with little validation effort, comparable to familiar terminology work. The second stage adds relations between requirements such as “requires”, “refines” or “refers_to” as RDF/OWL, whose relation types require expert judgment. At the third stage, formal test rules are generated, whose validation requires testing expertise, since silent errors can have a major impact on all downstream scenarios. At every stage the AI supplies a draft including a confidence annotation, on the basis of which standardization experts make decisions.

Who creates which content?

The previous section outlines extensive new SMART content. In terms of capacity and areas of competence, it is likely to be difficult for standardization committees to contribute this content on their own. Some of this content is not standards content in the conventional sense anyway, but operationalization and interpretation, for which other actors are better equipped. How such a division of labor could be organized in concrete terms is an open question whose answer presupposes extensive agreement.

One possible layer model distinguishes four levels, with the “normative core” at the first level. Standardization committees first provide granular standards content, definitions of terms and cross-references as a binding basis. At a second level, research and metrology institutions could contribute to operationalization, for example with scientifically grounded test catalogs, SHACL shapes or reference datasets. Their contribution to trust would rest on scientific methodology, peer review and institutional responsibility. At a third level, sector-specific elaboration is provided by trade associations, industry consortia and research projects, which could contribute interpretation guidelines or mappings between standards. A fourth, tool and application level brings together contributions from (AI testing) tool manufacturers, open-source communities and users, who supply tool adapters, reference implementations and libraries.

49 <https://www.iec.ch/online-standards-development>

50 Cf. Babaei Giglou, H., D'Souza, J., Mihindukulasooriya, N., & Auer, S. (2025), “LLMs4OL 2025 Overview: The 2nd Large Language Models for Ontology Learning Challenge”, Open Conference Proceedings 6, <https://doi.org/10.52825/ocp.v6i.2913>. Pilot projects on the machine interpretability of industry standards such as Gjerver et al. (2025), “Machine-interpretable Engineering Design Standards for Valve Specification”, IEEE Access, <https://doi.org/10.1109/ACCESS.2025.3643528>, demonstrate the concrete application to technical standards.

In such a model, testing organizations would be less the primary producers of operationalization content than users, employing the content from the upper levels when carrying out accredited testing and giving back their own contributions, for example as certificates of conformity (D-CoC) and accreditation-related verifiable credentials.

The infrastructure for such a shared model would be the data space described in Scenario 3. Every item of content would carry the identity of its publisher, a trust class attested in machine-readable form as a verifiable credential and an explicit reference to the version of the standard it relates to. If a new version of a standard appeared, dependent content could be flagged automatically as possibly out of date, comparable to dependencies in a package manager.

A key risk of the model would be fragmentation, if several actors published competing operationalizations for the same standard. A curating role could be taken on by research and metrology institutions with established trust status, which would recommend quality-assured “layer 2 content” transparently (without making it “normative”). Implementing such a model in practice would presuppose extensive agreement on roles, interfaces and the basis of trust, and therefore represents a long-term development goal.

The background of the page features a light blue, low-poly illustration. In the foreground, three stylized human figures are shown from behind, walking away from the viewer. Above them, a large, sharp, upward-pointing arrow is composed of several geometric shapes. The overall aesthetic is modern and minimalist.

4

CASE STUDY

4 CASE STUDY

The scenarios described above become more concrete in the following example of a company that implements them step by step. This chapter traces the path of an AI company from its first steps with SMART standards to a testing environment embedded in a digital quality infrastructure, and thus makes tangible what the conceptual scenarios can mean in practice.⁵¹



THE COMPANY

A European company develops and operates an AI-based credit scoring system offered to banks and consumer credit providers as SaaS. The system assigns credit applications to credit rating classes on the basis of socio-demographic and financial characteristics and thus supports decisions on granting credit and setting conditions. Under Annex III of the AI Act it therefore falls into the high-risk category. The company has to meet the requirements of Chapter III Section 2 of the AI Act, in particular appropriate risk management (Art. 9) and a minimum level of accuracy, robustness and cybersecurity (Art. 15).

The company practices agile, SCRUM-style software development. Requirements are maintained in GitLab and operationalized through epics and user stories. Developing the AI system consists largely of software development and data science processes. The MLOps infrastructure is implemented as a GitLab CI/CD pipeline which, for ML-specific tasks, integrates an MLOps framework (such as MLflow)

as an experiment tracking and model registry backend. Test engineers store test code as CI jobs. Training runs, model versions and test results are logged and versioned in the MLOps framework. New model versions are delivered in release cycles after automated testing. Responsibility for analyzing and implementing requirements from standards lies with a product owner together with the test engineers.

A central quality dimension for the system is data drift, that is, the change in the statistical properties of production input data compared with the training or reference dataset. In a credit scoring system, data drift arises for example from economic shifts, new groups of applicants or changed consumption habits. If it is not monitored, there is a risk of creeping distortions of credit decisions that impair both robustness (AI Act Art. 15) and risk management (Art. 9).

⁵¹ This example is an adapted version of the example in ETSI TS 104 008 V1.1, A.3

THE (FICTIONAL) STANDARDS

Three fictional standards interlock for this quality dimension (see also Figure 6):

- **EN 123456:2026** “Data quality of AI systems”, a harmonized standard under the AI Act with horizontal, abstract requirements. It requires: “The data-drift should be constantly monitored and should not exceed a moderate level.”
- **ISO/IEC 98761:2026** “Data quality metrics”, an international standard that defines the Population Stability Index (PSI) as an indicator of data drift, gives the formula for calculating it and refers to a registered reference implementation.
- **ISO/IEC 98768:2026** “Quality of AI applications for credit scoring”, an international standard of a vertical, concrete nature. It specifies the drift thresholds for the credit scoring application context: “stable” at $PSI < 0.1$, “moderate” at $PSI < 0.2$, “significant” from $PSI \geq 0.2$.

4.1 Scenario 1: using machine-readable standards content via APIs

Identifying and obtaining standards

In the concept phase the company has to establish which standards and standards content are relevant for its AI system. Instead of searching manually for PDFs in web shops and working through them, the company uses the SMART standards interface of a standards provider (e.g. a REST API, cf. VDE CaaS).

An intelligent search with semantic functions allows the research to be linked to the system context: intended purpose (credit assessment), risk class (high risk under Annex III of the AI Act) and sector (financial services). The starting point is the reference chain from the AI Act: Art. 9 and Art. 15 refer to harmonized standards, which in turn refer to further standards. The search service returns the relevant harmonized standard EN 123456:2026 as a hit and automatically follows its cross-references to ISO/IEC 98761:2026. A semantic search in the catalog of vertical standards identifies ISO/IEC 98768:2026 as the relevant specification for credit scoring.

Integrating standards content into requirements management

The standards content identified is transferred into GitLab as granular, machine-readable units via ReqIF import⁵², instead of being filed as a PDF and paraphrased manually. Three linked items are created for the requirement considered here:

- From EN 123456:2026, requirement Req-5-2 (“The data-drift should be constantly monitored and should not exceed a moderate level.”) as a GitLab issue in the epic “Data quality under EN 123456”, with the label `compliance::EN-123456` and a reference to the AI Act articles it addresses (Art. 9 and Art. 15).
- From ISO/IEC 98761:2026, the definition of the PSI: the calculation formula (in MathML, a standardized markup format for mathematical formulas) and a reference to a registered reference implementation with a persistent identifier and semantic versioning. Both are linked in the issue.
- From ISO/IEC 98768:2026, the context-specific threshold profile as a JSON file stored in the repository. It links the numerical PSI thresholds to the abstract drift levels from EN 123456 and later serves directly as configuration for the test.

⁵² Note: GitLab does not yet offer a ReqIF import, but has documented this as a work item: https://gitlab.com/groups/gitlab-org/-/work_items/21272

Fictional standards and the associated SMART content

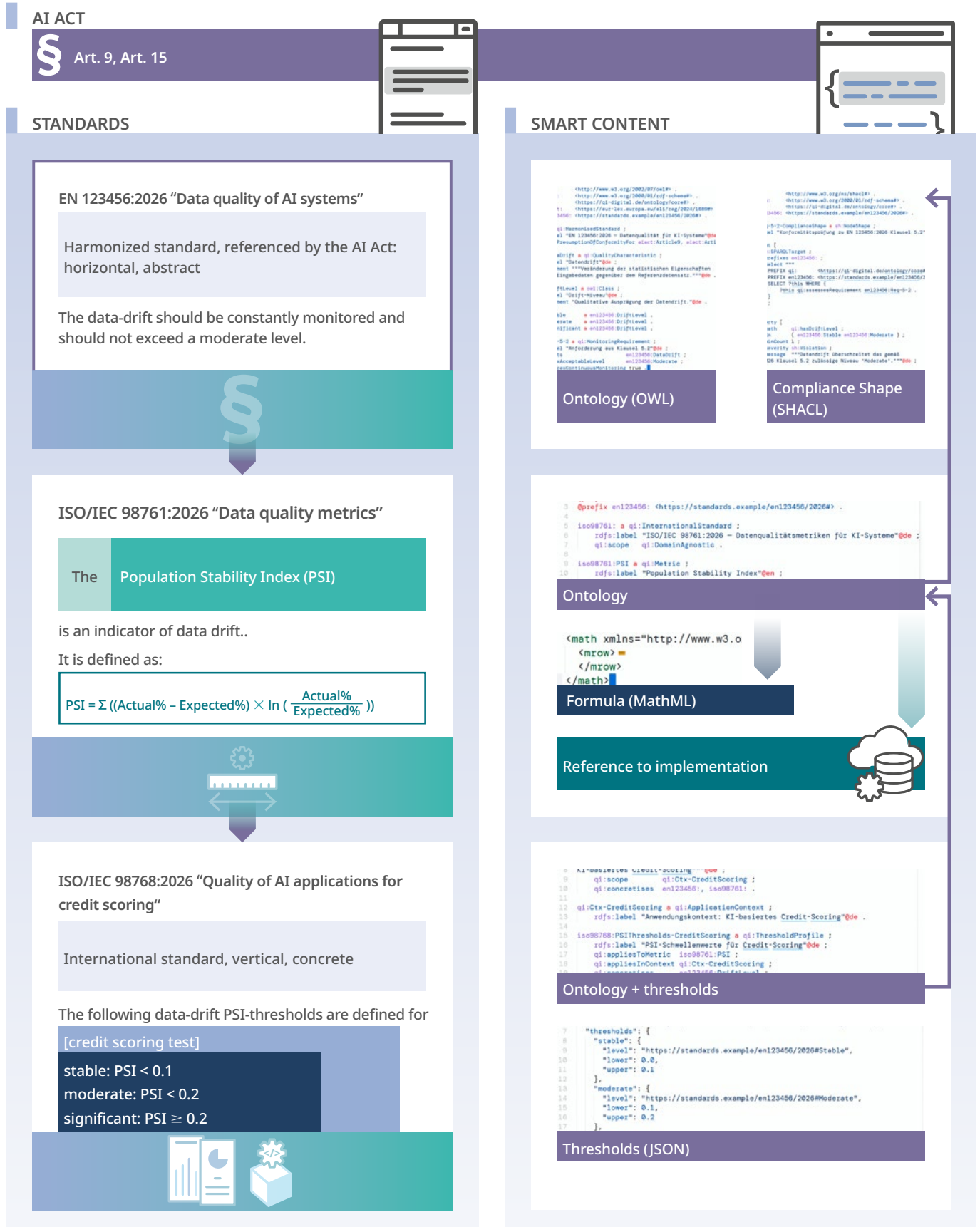


Figure 5

In a structured metadata field, each issue carries the persistent identifier of the SMART content, for example a URN that unambiguously addresses the standard, clause and version (in the example: ``urn:smartstd:en123456:2026:req-5-2``). The source is thus documented and can be evaluated automatically through the GitLab API.

Updates are detected by a GitLab scheduled pipeline that queries the SMART standards API daily for every imported identifier. If the API returns a newer version of a piece of standards content (for example because EN 123456 has been published as version 1.1), the pipeline opens a merge request marking the textual and structural changes. The responsible test engineers are assigned automatically through GitLab and review the delta before the updated requirement is adopted into ongoing development.

In Jupyter notebooks in the GitLab repository (for example via the GitLab Web IDE) and in IDEs such as VS Code, the SMART content is provided inline. A SMART standards extension displays the relevant clause of the standard, the formula and the thresholds directly in the editor while validation code is being written.

Operationalizing in the MLOps pipeline

The test engineers derive the test implementation from the SMART content linked in GitLab. The drift test itself runs in two steps:

- Calculation and classification: the PSI value is calculated for a slice of production data against the reference dataset and then assigned to a drift level (“stable”, “moderate” or “significant”) using the threshold profile from ISO/IEC 98768.
- Conformity check: the level determined is compared against the requirement from EN 123456, that is, it is checked that it does not go beyond “moderate”. EN 123456 provides this check as a formal, machine-executable rule (held in the SMART content of the standard as a SHACL⁵³ constraint).

A short Python function handles the implementation: it calls the referenced PSI library from ISO/IEC 98761 via its persistent identifier, loads the thresholds directly from

the JSON profile and returns the drift level. The formal conformity check itself is validated with a standard library (such as pyshacl) against the rule held in the standard.

The test is set up as a job in the GitLab CI/CD pipeline. It runs on every model release and is additionally executed continuously against production data via a scheduled pipeline. In the job, the PSI value calculated, the drift level determined and the result of the conformity check are assigned via the tracking API of the MLOps framework to the respective model version in the model registry, and are thus traceable in versioned form. If the “moderate” level is exceeded, the pipeline raises an alert and triggers a retraining workflow.

Because the test and the standards content are unambiguously linked through the identifiers maintained in GitLab, and test results are bound to the model version through the MLOps framework, it is possible at any time to trace which test covers which requirement for which model version.

Generating conformity documentation

The conformity documentation is derived automatically from the existing versioned artifacts; a downstream manual step is no longer needed. Evidence comes from three sources:

- GitLab supplies the normative reference. For every compliance label (“compliance::EN-123456”) the associated issues can be listed with their persistent identifiers, including the version of the standard referenced.
- The MLOps framework supplies the test results. Every run is linked through the model version to the respective requirement from the standard and contains the PSI value, drift level and conformity result.
- Git supplies the state of the code. The commit hash of the underlying pipeline definition and test implementation is part of every run in the MLOps framework.

A dedicated CI job (“compliance-report” in the example) brings these sources together and produces a structured conformity report. For each requirement the report contains the chain of clause, requirement identifier, test

53 SHACL is a language standardized by W3C for describing such conformity rules: <https://www.w3.org/TR/shacl/>

implementation used including its version, data slice, test result and conformity testing. It is produced both as a JSON file for further machine processing and as a rendered HTML or PDF version for human readers, and is archived as a GitLab release artifact and in the MLOps framework.

Because all components of the report are versioned and persistently referenced, reproducibility is assured. The standards content is addressed through its URNs (for example `urn:smartstd:en123456:2026:req-5-2`), the

test code through the Git commit hash, and the model version and test results through the run ID of the MLOps framework. The corresponding report can therefore be regenerated at the push of a button for any past model version, for example when an auditor asks about a particular release or it has to be established retrospectively which version of the standard was tested against.

4.2 Scenario 2: generating AI tests



THE COMPANY (...CONTINUED)

The company from scenario 1 uses SMART standards through specialized tools. In the next step a local chatbot takes over. A test engineer interacts with the chatbot in their own IDE (for example VS Code with a chatbot extension). The chatbot accesses the SMART standards API via MCP

and bases its answers on the machine-readable standards content using RAG. Local context sources are connected via MCP as well, in particular the GitLab project with its issues, wikis and repository content.

Determining requirements for a specific context

The test engineer can bring in the system context in two ways. A short description in natural language is enough for a quick search, for example “AI-based credit scoring system for consumer credit, used in the EU. The relevant requirements are those of the AI Act on data quality and robustness.”

For a well-founded requirements analysis, for example as part of a formal conformity assessment, it is more effective to give the chatbot access to a selection of relevant documentation artifacts, such as the system description, architecture documentation, model card and risk classification. From these documents the chatbot extracts the system properties relevant for identifying standards

(intended purpose, risk class, sector, type of processing) and uses them as a structured basis for the search.

Both approaches can be combined. In addition, the chatbot reads the local context from GitLab (existing issues, architecture documentation in wikis, model cards and code in the repository) and matches it semantically against the standards content available in the SMART standards catalog.

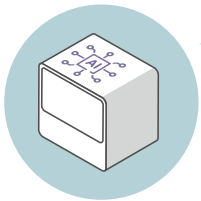
At the term level, local terms are matched against the ontology of the standard. The terms “model drift” or “distribution shift” in the internal documentation map onto the term “data-drift” from EN 123456. At the structural level, the system properties (high-risk class, credit scoring

Support from a chatbot for a standards-based test (simplified example)



TEST ENGINEER

We're developing an AI-based credit scoring system for consumer loans (EU). Which AI Act requirements for data quality and robustness are relevant, and how do we test them?



CHATBOT

```
//Research harmonized standards via the SMART Standards API, read local
context from GitLab (model card, architecture, issues), match terms against
the standards ontology.
```

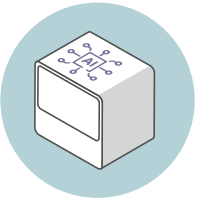
Three standards apply:

- EN 1236:2026 (harmonized): continuous drift monitoring, level ≤ "moderate"
- ISO/IEC 98761:2026: PSI as a drift indicator (formula + reference implementation)
- ISO/IEC 98768:2026: PSI thresholds for credit scoring



TEST ENGINEER

Please create a test plan and test code for the drift requirement.



CHATBOT

```
// Derive test plan from the standard content, generate test code with PSI
calculation, threshold classification, and SHACL conformity check.
```

Test plan:

- Metric: PSI on production data sample vs. reference dataset
- Classification: threshold profile from ISO/IEC 98768
- Conformity: SHACL constraint from EN 123456
- Trigger: every release and daily, alert at "significant"

```
psi    = psi_lib.compute(window, reference)    # ISO/IEC 98761
level  = classify(psi, thresholds_cs)          # ISO/IEC 98768
validate_shacl(level, en123456_constraint)    # EN 123456
mlflow_client.log_metric("psi", psi)
```

Should I set up the test as a CI job in the GitLab pipeline and configure logging in the MLOps framework?

Figure 6

application context) are matched against the scopes of the standards. The result is an annotated list of relevant requirements, such as the drift requirement from EN 123456:2026, the associated metric from ISO/IEC 98761:2026 and the context-specific thresholds from ISO/IEC 98768:2026, each with a reference to the clause and a justification of its relevance.

Proposing a test plan

From the list of requirements the chatbot generates a structured test plan:

- What is tested: PSI as an indicator of data drift in the production data slice against the reference dataset.
- How it is tested: calculation according to the MathML formula from ISO/IEC 98761:2026, comparison against the JSON threshold profile from ISO/IEC 98768:2026.
- When it is tested: on every model release and continuously in monitoring (in line with “constantly monitored” from EN 123456:2026).
- Which actions are triggered: an alert at “significant”, a retraining trigger on repeated “moderate”.

Requirements that can be tested automatically are distinguished from those that require human judgment. The test engineer reviews the test plan, adds any missing aspects and approves it.

Generating test code

For the testing activities that can be performed automatically, the chatbot generates executable test code, which is stored in the Git repository and run as a job in the GitLab CI/CD pipeline. The trustworthiness of the generation rests on a clear separation between SMART content (limit values, calculation rules) and the language model.

- The PSI calculation is linked to the referenced implementation from ISO/IEC 98761:2026; the library is included via its persistent identifier rather than reconstructed by the model.
- The thresholds from the JSON profile in ISO/IEC 98768:2026 are carried over directly into the code.

- The formal test rule from EN 123456:2026 (see operationalization in scenario 1) serves as the specification for the conformity check.

The language model supplies the test logic and the embedding into the local toolchain, including the logging of test results in the MLOps framework. The values from the standards are firmly anchored and traceable back to their clause. The generated code is checked automatically for syntax and validated against the test rule of the standard, in order to ensure that the requirement is correctly represented. The chatbot files the result in the GitLab repository as a merge request, which the test engineer reviews and merges.

Producing reports on request

The same chatbot enables ad hoc reporting from natural language. The test engineer asks, for example, “Produce an overview of the data quality test results under EN 123456 for the last three releases”, and the chatbot uses MCP to retrieve the relevant runs from the MLOps framework and the corresponding GitLab issues, then produces the report in the same structured format as the CI job from scenario 1.

Alongside the standardized reports at fixed trigger points (scenario 1), scenario 2 adds ad hoc analyses for specific audits, internal reviews or queries from conformity assessment bodies, without requiring separate implementation effort. The chain of evidence remains deterministic, since the language model only interprets the query and designs the layout, while the supporting values come unchanged from the versioned source systems.

Quality assurance and responsibility

The test engineer remains responsible for the quality of the test. The role shifts from writing test code to targeted review of the generated artifacts. For safety-critical requirements a manual review step is mandatory. The system keeps a complete log of which chatbot output is based on which normative source, as a precondition for traceability under AI Act Art. 12.

4.3 Scenario 3: integration into an ecosystem for AI quality



THE COMPANY (...CONTINUED)

The company from scenarios 1 and 2 has expanded its portfolio. Alongside the existing credit scoring system it now develops further AI-based services, for example for assessing the risk of business customers and for detecting fraud in payment transactions. With the expanded portfolio the number and variety of tests grow, and it becomes clear that not all the resources needed can sensibly be provided within the company itself. For certain tests, suitable reference datasets are not available, or the tools are so specialized that developing them in-house does not pay off, or the required independence of the test calls for an accredited external body.

The company therefore joins a federated quality infrastructure platform. It connects step by step: first a data space connector (Eclipse EDC) is deployed as a container service in the existing infrastructure and linked (via REST API or via MCP) to the GitLab CI/CD pipeline and the chatbot from scenario 2. In parallel the company registers as a participant and receives the associated verifiable identity. It can then obtain tools and datasets in the data space, commission testing services and make its own conformity evidence visible. The following sections describe typical applications.

Obtaining testing tools from the data space

Scenario 1 assumed that the PSI library referenced in ISO/IEC 98761 was readily available. Scenario 3 makes the assured provision of such a library concrete. On connecting to the data space, the company resolves the identifier held in the SMART content against the federated catalog. The search returns two implementations, each carrying a suitability attestation from a qualified body for ISO/IEC 98761:2026. The attestations are held as signed verifiable credentials on the tool entry in the catalog and name the tool version tested, the clauses covered and the issuing body.

The company makes its selection on the basis of the license conditions and the version binding to the model versions used in the GitLab pipeline. The tool is obtained as a container image via the connector. The associated verifiable credential is transferred with it and verified locally in the CI/CD pipeline before the image is executed. The GitLab pipeline from scenario 1 is adapted so that the library no longer comes from a public package repository but from the data space, with a verifiable link to the attested suitability.

Testing the model against a protected reference dataset

A further requirement for the credit scoring system arises from the industry-wide specification of the AI Act for the financial sector: the drift test is to be carried out not only against the company's own reference dataset, but additionally against a demographically balanced reference dataset maintained for the sector by a qualified body. This reference dataset is deliberately not made publicly available, so that it does not find its way into later training data and lose its value as an independent test resource.

The data space allows this test to be carried out in compute-to-data mode. The company makes the inference function of its model available as a service via the connector; the qualified body calls this service with data points from the protected reference dataset, calculates the PSI values and drift levels with the attested PSI library from the previous section, and returns only the signed result (PSI value, drift level, conformity testing). Model weights do not leave the company; the reference dataset does not leave the qualified body.

The logic of the test itself remains the same as in scenario 1; only the place of execution and the protection of the resources involved are distributed differently. The company thus obtains an additional, independently produced piece of evidence for the drift test against the independent reference dataset. The result is bound to the respective model version in the MLOps framework and feeds into the conformity report from scenario 1. For the qualified body the dataset remains fully under its control, without the need for a bilaterally negotiated data transfer agreement with every customer.

Certificate of conformity as a D-CoC in the “AI passport”

Before the credit scoring system is placed on the market, it undergoes an external conformity assessment by an accredited body, and that body issues the result as a Digital Certificate of Conformity (D-CoC). Unlike the PDF certificate customary today, the D-CoC is a machine-readable document held as a signed verifiable credential, containing the versions of the standards tested against, the model version tested, the test procedures used, the issuing accredited body and its accreditation credential in machine-readable form.

The D-CoC is included in the credit scoring system’s “AI passport”, which accompanies the system throughout its lifecycle and bundles both model and conformity information. Before concluding a contract, customers can check automatically whether a valid D-CoC exists, whether it comes from an accredited body and whether it relates to the model version actually in use. The underlying chain of trust is verifiable end to end through the data space: the D-CoC refers to the accreditation of the issuing body (DAkKS accreditation credential), to the suitability attestations of the testing tools used and to the versions of the underlying SMART standards.

This chaining brings the vision set out in the QI-Digital roadmap into effect: conformity is no longer handed over as an isolated certificate, but carried along the value chain as a traceable web of standard, tool, test execution and issuing body.

Further use cases

Beyond the applications described in detail, connecting to the data space opens up further possibilities, some of which are outlined briefly here:

- **Update of a standard:** If a new version of ISO/IEC 98768 is published, the pipeline from scenario 1 not only checks the textual diff of the thresholds, but also verifies in the data space whether the PSI library in use still carries a valid suitability attestation for the new version. If not, this is flagged in the merge request and the company can switch to an attested variant at an early stage.
- **Commissioning a specialized testing service:** For a rare but regulatorily required bias test across protected groups, the company commissions a testing service from an accredited conformity assessment body via the marketplace. The service runs as compute-to-data, and the signed result feeds into the conformity report as additional evidence. The company does not have to operate the corresponding test setup itself.
- **Audit request from market surveillance:** A supervisory authority asks retrospectively for evidence of the testing of a particular model version. All components are verifiably traceable through the data space references: which tool version with which suitability attestation, which reference dataset with which provenance, which certificate of conformity from which accredited body. The authority checks the verifiable credentials directly, without having to ask the company to supply additional documents.
- **Agent-based end-to-end testing:** The chatbot from scenario 2 is extended with a data space connection. Starting from a new requirement, it not only resolves the SMART content but also finds suitable tools in the marketplace, checks suitability attestations and initiates the test. The mechanisms from scenarios 2 and 3 thus combine into a continuous workflow.
- **Providing its own tool in the marketplace:** In the course of day-to-day operation the company has developed a credit-scoring-specific drift variant that delivers more stable results for its domain than the generic PSI library. The company makes it available as an open-source tool in the marketplace and has it attested by a qualified body for ISO/IEC 98768. In doing so the company moves from being purely a consumer to being a contributor in the ecosystem.

5

DISCUSSION

5 DISCUSSION

5.1 Evaluation of the scenarios

The scenarios described in chapter 3 represent different stages of development for the use of SMART standards in testing AI systems. Together they pursue the goal of making testing more scalable, more reproducible and more efficient. They differ in the maturity of SMART content required, in the technical infrastructure and in the organizational effort involved. From a user perspective, the question is less whether individual test steps can be fully automated, and more what practical benefit SMART standards deliver when AI tests are carried out, and what preconditions have to be created for this.

Opportunities and benefits

A key benefit of the scenarios presented lies in reducing manual activities throughout the testing process. Testing AI systems typically involves a large number of recurring tasks, among them analyzing normative requirements, drawing up test plans, deriving test criteria, documenting results and providing evidence to customers, auditors or supervisory authorities. These activities often account for a substantial share of the total effort of a test.

Scenario 1 alone allows normative requirements to be fed directly into requirements management, development and testing systems. This reduces manual handovers between systems, allows requirements to be referenced unambiguously and enables changes to standards to be tracked in a targeted way. The use of granularly referenceable standards content in particular simplifies versioning and the traceability of requirements across the entire testing process.

SMART standards also promote the reuse of test artifacts. Test plans, test cases, checklists or conformity evidence do not have to be created from scratch for every test, but can build on existing artifacts. This increases the reproducibility and consistency of AI testing and makes it easier to compare test results across organizations and use cases.

A further benefit lies in easing the burden of test preparation. Identifying relevant requirements, drawing up test checklists and assigning requirements to test steps can be supported considerably by SMART standards. Changes

SMART standards can already make the testing of AI systems more scalable today.

The information needed to implement various use cases is often already available in ReqIF format:

- Import of granular standards content (requirements, definitions, cross-references) into requirements management, with an unambiguous link to the standard, clause and version via persistent identifiers.
- Mapping of the reference chain between legislation, harmonized standards and referenced standards as explicit relations, including linking imported requirements to the AI Act articles they address.
- Enrichment of the content with metadata such as requirement type (shall, should, may), compliance labels or thematic groupings.
- Automated detection of new versions of a standard by comparing identifiers, diff display between versions, and triggering of review workflows in requirements management.
- Aggregation of requirements, source references and compliance labels for conformity documentation, and connection to test results from CI/CD and MLOps frameworks via the persistent identifiers.

to standards can be detected automatically and mapped onto existing test processes, which lowers the maintenance effort.

Scalability is not the same as full automation. Many requirements in AI standards continue to call for expert judgment, contextual knowledge and human decisions. The scenarios presented therefore do not primarily aim to automate testing completely, but to provide technical support for recurring activities in the preparation, execution and documentation of testing. Automatically supplying the relevant standards content, supporting the analysis of standards, generating test artifacts or producing structured evidence can in itself considerably reduce the effort involved in AI testing, without automating responsibility for the test verdict.

Scenario 2 in particular also opens up new possibilities for using generative AI in testing processes. Language models are already being used today to analyze standards, produce audit documentation or support conformity assessments. SMART content can help make this use more reliable and more traceable by providing normative content in a structured, versioned and unambiguously referenceable form.

Preconditions and challenges

Implementing the scenarios described requires normative content to be made available in a suitable form. The text of a standard alone is often not enough. Technical support for AI testing additionally requires structured requirements, metadata, reference information, test methodologies, semantic descriptions and suitable exchange formats.

Operationalizing normative requirements is a particular challenge. Many requirements in AI standards are deliberately formulated in terms of principles and contain few thresholds that can be automated directly. For technical use, they have to be formalized in part, linked to test methods or supplemented with additional information. SMART standards create an important basis for this, but they do not replace the expert interpretation of the requirements.

The semantic enrichment of normative content is also becoming more important. If test support is to be generated automatically from the requirements in standards, then terms, relationships and requirements have to be described

in a machine-interpretable way. This calls for additional expertise in areas such as knowledge modeling, ontology engineering and the formal description of normative content.

At the same time, linguistic precision becomes more important for automated processing. Unclear or ambiguous wording makes standards harder to use technically and increases the effort of validating the results produced. Consistently applying authoring rules therefore becomes an important precondition for the scalability of standards-based AI testing.

Human validation is particularly important here. Test artifacts, evaluations or recommendations generated automatically still have to be reviewed by qualified experts. Regardless of the degree of automation, responsibility for testing decisions remains with the testers, auditors or conformity assessment bodies concerned.

Scenarios 2 and 3 in particular also require additional technical and organizational infrastructure. Besides suitable authoring and validation tools, procedures are needed for quality assurance of SMART content, reference data, testing tools and semantic models. For many of these tasks there are as yet neither established processes nor clearly defined institutional responsibilities.

Risks

Despite their potential, the approaches presented also carry risks. One major risk is automation bias. Suggestions, evaluations or test results generated automatically could be adopted without review, even though AI-based systems in particular can still produce erroneous or misleading results. At the same time there is a danger that the effort required for adequate human validation offsets part of the expected efficiency gains.

A further risk lies in narrowing the understanding of testing to those requirements that lend themselves particularly well to formalization or automation. With AI systems especially, key properties such as robustness, fairness, transparency, human oversight or appropriateness often depend on the application context. A strong focus on automatable test criteria could push requirements that are hard to operationalize, but that are nonetheless relevant, into the background.

Increasing formalization also raises the risk of systemic error propagation. Errors in SMART content, ontologies, reference implementations or test artifacts can potentially affect a large number of tests and thus have far-reaching consequences.

Governance questions arise as well. The more testing builds on shared content, tools and services, the harder it can become to assign responsibility. This applies in particular to cooperation between standardization organizations, tool vendors, testing bodies and users. As the division of labor increases, there is a danger of responsibility becoming diffused along the entire testing value chain.

In the ecosystem scenario there is the additional risk of fragmentation into several competing and non-interoperable solutions. Different data models, interfaces or governance approaches could limit the intended economies of scale and the reusability of test artifacts.

Finally, the established basic principles of trustworthy testing must be maintained in digitally supported testing processes too. These include confidentiality, data protection, independence and impartiality. Access to sensitive models, data holdings or documentation therefore requires appropriate technical and organizational safeguards. Digital tools used in testing AI systems must not compromise these fundamental requirements.

5.2 Relation to central concepts of digitizing standardization

From serial to parallel processes

IDI's white paper 3 quantifies the business benefit of SMART standards in standards application processes. Its value-added model shows that SMART standards can decouple and parallelize activities that today have to run serially (researching, extracting, transferring, implementing and checking a standard), with measurable effects on time, cost and quality.

The three scenarios make this change concrete at three levels. Scenario 1 removes the central extraction and transfer bottleneck: product owners, developers and MLOps engineers access the same granular standards content at the same time. Scenario 2 transfers the principle into the ongoing MLOps cycle; an AI agent replaces sequential individual steps with concurrent orchestration, and changes to standards trigger a targeted reassessment without a complete test cycle having to be set up again. This makes the CABCA principle described in chapter 2 technically feasible. Scenario 3 raises parallelization to the level of the ecosystem: standardization committees, testing organizations, industry associations and tool vendors contribute content independently of one another and are coordinated through versioning and trust credentials.

Semantic integration

IDI's white paper 1 describes the transformation from document-based to machine-interpretable standards using a utility model with five maturity levels. Its conceptual core is a change in the object of reference: at the lower levels (levels 1 to 2) the document is at the center; it is available digitally but has to be evaluated manually. At the higher levels (levels 3 to 5) the focus shifts to the granular standards content itself, which is uniquely identified, linked through ontologies and enriched with execution information. Semantic integration then means more than machine readability: standards content fits seamlessly into existing processing chains and is treated there like native data.

The three scenarios map different utility levels onto the AI testing use case. Scenario 1 presupposes levels 3 to 4 and makes tool-supported use of granular content that is addressable via a catalog. Scenario 2 requires level 4 with formally encoded constraints and testability annotations,

on the basis of which an AI agent generates test artifacts deterministically. Scenario 3 widens the perspective to the ecosystem: the standard not only becomes machine-interpretable but is also operated as a sovereign, governable resource in a data space. The scenarios can thus be understood as a use-case-specific concretization of the utility model: they show which investments in which utility level yield which measurable benefit.

SMART standards as a single source of truth

In current practice, requirements from standards often exist in several parallel representations: in the original document, in internal paraphrases in requirements management, in test specifications, in test reports. Each of these representations is created manually and is rarely updated in full when the standard changes, which leads to divergent states of the same standard across different systems. SMART standards with verified content reverse this pattern: the standard itself is the single authoritative source, which all downstream artifacts reference directly via persistent identifiers.

This property takes on additional significance in combination with generative AI: SMART standards act as a verified knowledge base (ground truth) against which AI output is checked and in which it is grounded. Instead of relying on training knowledge, language models take values, definitions and requirements directly from the SMART content. “Hallucinations” or outdated training data are thus ruled out at the critical point, and the generation of test artifacts remains traceable back to the relevant clause of the standard.

The scenarios implement this principle step by step. In scenario 1 the standard becomes the single source of truth for requirements management: every imported requirement retains its link to clause and version, and changes to the standard automatically flag the requirements affected. Scenario 2 extends the principle to test artifacts generated by AI: limit values and the like are taken deterministically from the SMART content rather than produced by LLM inference; the normative values do not pass through the language model and remain traceable back to their clause. Scenario 3 adds cryptographically verifiable trust credentials (verifiable credentials) to this technical identity; these attest, in machine-readable form, the origin, accreditation status and level of trust of every item of standards content and every test artifact.

Integration into the digital quality infrastructure

The QI-Digital roadmap describes the paradigm shift from a document-based to a data-driven quality infrastructure: standards, test certificates, accreditation credentials and certificates of conformity are no longer sent as documents, but exchanged as machine-readable resources in a sovereign data space. In this architecture SMART standards form the normative foundation on which automated conformity testing, the Digital Certificate of Conformity (D-CoC), Digital Calibration Certificates (DCC), the Digital Product Passport (DPP) and a dedicated “AI passport” are built.

The scenarios in this white paper therefore do not describe an isolated AI testing ecosystem, but concrete use cases within this digital QI. The infrastructure components from scenario 1 (semantic search, content delivery, query APIs) correspond to those that QI-Digital envisions for SMART standards in general; the test reports and certificates of conformity generated in scenario 2 are the same building blocks that are needed for the DPP and the AI passport; the data space from scenario 3 can be integrated into the Quality-X concept.

The ecosystem outlined in scenario 3 supplies the intermediate products that come together in a DPP or an AI passport: quality-assured testing tools, traceable test executions and verifiable attestations of suitability and conformity. Anyone building the infrastructure for scalable AI testing is at the same time helping to build the digital QI as a whole, and can use the scenarios as a testbed: an area with clearly defined actors, defined artifacts and available technologies, in which transitional concepts can be tried out in practice. Initial trial environments already exist, for example the additive manufacturing living lab of the QI-Digital initiative, in which SMART standards, digital certificates and data space components are tested together.

5.3 Transferability

Physical AI

The scenarios focus on the use of standards for testing software-based AI systems in MLOps environments. A growing field of increasing practical relevance, however, is that of AI systems integrated into physical products, such as embedded AI in machinery, autonomous systems or AI-supported sensor technology in medical devices. Testing such systems requires physical testing under real or simulated environmental conditions in addition to software tests.

Digital twins offer a promising approach here: the virtual representation of a physical system allows tests in simulated environments before hardware is available, and enables continuous validation in operation. Particularly relevant here is the Asset Administration Shell (AAS), which, as a standardized digital representation of physical assets, can also contain AI components, training data and test results as submodels. In a completed pilot project, IDiS has already tested the integration of standards content into AAS submodels.

For the scenarios described in this white paper, this means an extension in two dimensions. First, SMART standards could supply AAS submodel references directly, so that normative requirements are transferred automatically into the digital representation of the physical product. Second, testing platforms (scenario 2) would have to be capable of orchestrating not only software tests but also hardware-in-the-loop tests, with physical test environments integrated via standardized interfaces. Scenario 3 would then also cover physical test benches and calibration services as subscribable resources in the data space. The conceptual architecture remains the same; the interfaces to the physical world would be the additional integration task.

Further fields of application

As discussed, the three scenarios in this white paper are only partly specific to AI; at their core they propose ways of implementing central transformation concepts for standardization and the application of standards. They

describe how machine-readable standards content can be integrated into iterative development and operating processes in order to make testing more scalable, more reproducible and easier to automate. This basic structure can be transferred to a large number of further testing domains in which standards play a central role.

Examples with high transferability are the testing of medical devices (MDR, IEC 62304), functional safety in automation technology (IEC 61508, ISO 26262), cybersecurity (IEC 62443), as well as energy systems and smart grids, where requirements from standards increasingly have to be embedded in digital processes. The same basic problem recurs in all these areas: standards exist as unstructured documents, have to be translated manually into requirements and embedded into testing processes that have to be repeated regularly because the system under test changes frequently.

The technical infrastructure from the scenarios (SMART content, semantic search, machine-readable constraints, MCP interfaces, data space governance and so on) would be largely identical in these areas. What changes are the domain-specific standards and ontologies. This white paper can therefore serve as a template for similar scenarios in other standardization domains. Conversely, sector-specific differentiation, for example separate scenarios for medical devices or cybersecurity, would be a valuable next step building on the foundation laid here.

Developments at the EU level

The scenarios outlined here also connect to policy developments at the EU level. The final report of Workstream 14 of the High-Level Forum on European Standardisation⁵⁴ pursues an approach to data interoperability in which harmonized European standards give technical substance to the requirements of the Data Act and take regulatory effect through the presumption of conformity. Recommendation R.11 of the report also calls for a European framework for trustworthy ontologies, which is to govern the selection, maintenance and trustworthiness of semantic models. A comparable governance question arises for SMART standards, whose reliability depends in the

⁵⁴ European Commission, High-Level Forum on European Standardisation, Workstream 14 (2024): Final Report on Data Interoperability. Endorsed Version, April 2024. Available online at <https://ec.europa.eu/docsroom/documents/58914>

long run on the maintenance and institutional anchoring of the underlying schemas.

The report also states explicitly that AI models gain in reliability through the integration of knowledge graphs and ontologies, and that “hallucinations” pose a major risk in safety-critical domains such as medicine. The separation of stochastic and deterministic functions proposed in scenario 2, in which the language model handles language-processing tasks while the SMART standards infrastructure provides the normative anchoring, can be understood as a domain-specific implementation of this principle for standards-based conformity testing.

6

CONCLUSION

6 CONCLUSION

As artificial intelligence grows in importance, so do the requirements placed on aspects such as safety, reliability and transparency. In practice, these requirements take effect through systematic testing on the basis of standards, and the volume of that testing can increasingly exceed the testing capacity available. The right answer is not to lower these requirements; instead, tools and concepts are needed that make them easier and more reliable to meet. SMART standards are a central building block here, because they make standards content available as machine-readable resources and thus form the foundation of scalable, standards-based testing of AI.

The scenarios described here set out a realistic, step-by-step path toward a digital quality infrastructure for AI, in which standards, testing tools and test data are available as sovereign, machine-readable resources, allowing reliable AI testing to be partly automated.

Four transformation concepts underpin the scenarios and reach beyond the AI testing use case: the move from serial to parallel processes, semantic integration along the IDiS utility model, establishing the standard as the single source of truth, and integrating all of this in the digital quality infrastructure. These concepts are not specific to AI and can be transferred to other testing domains in which standards-based tests have to be carried out iteratively and in large numbers.

The technical feasibility of the scenarios rests on existing standards, but implementing them in full requires that the corresponding standards content, such as ontologies, be built up step by step. Since creating this content manually for the entire relevant body of standards would be extremely labor-intensive, AI-supported generation of SMART content based on the human-in-the-loop principle plays a key role. Language models handle the laborious initial extraction from the text of the standard, and standardization experts validate the results, so that what emerges in the end is always verified content offering the reliability the scenarios require.

Besides the SMART content generated from standards, scalable testing of AI systems requires further operationalization elements such as testing tools and reference datasets, which have to be linked together reliably so that tests can run automatically. Because no

single actor can provide this breadth of content and resources, the white paper describes a possible division of labor across different roles. The data space from scenario 3 connects these contributions into a consistent test workflow while remaining open to extensions such as testing physically embedded AI systems via Asset Administration Shells and hardware-in-the-loop testing.

The scenarios in this white paper were developed as a proposal from the perspective of (potential) users and discussed within IDiS. They do not necessarily reflect the strategic plans of the standardization organizations involved, nor can they simply be read as an implementation roadmap. Their purpose is to show in concrete terms, at an early stage, how SMART standards can support the scalable testing of AI systems, and thereby to create a basis for discussion on the further development of SMART standards.

Against this background, the scenarios described here are not aimed at replacing established testing actors, but at relieving the pressure on them and strengthening their work. Testing organizations, consultancies, auditors and conformity assessment bodies contribute expertise that no technical infrastructure can replace. They interpret standards for specific application contexts, assess borderline cases and bear responsibility for the test verdict. SMART standards and the infrastructure described in the scenarios are an offering to these actors. They are a toolbox for automating recurring, formalizable activities, leaving more capacity for the tasks that require expert judgment and human responsibility. The services outlined in the scenarios, such as operating testing platforms, maintaining operationalization content or providing quality-assured test data, are at the same time new business areas that build on the existing expertise of these actors.

Several developments have to be advanced in parallel for the scenarios described to become viable in practice. The technical concepts and tools behind SMART standards must be developed further and made concrete. Verified SMART content has to be built up for key (AI) standards. This involves establishing AI-supported generation processes with human validation, but also other technical and organizational projects, such as piloting data spaces in which standards content, testing tools and

reference data are made available as sovereign resources. Organizationally, the roles of the actors involved have to be consolidated in line with the proposed division of tasks, and mechanisms for curation and trust-building have to be established. In parallel, the concepts should be transferred to further testing domains in which comparable scaling problems arise.

NEXT STEPS

This white paper does not describe a finished concept; it is an invitation to help shape one. Developing SMART standards further for the testing of AI systems is a joint task that calls for technical know-how from AI development just as much as for expertise in standardization, testing and quality infrastructure.

IDiS is the central platform for everyone who wants to play an active part in shaping SMART standards. Participation is free of charge and open to companies, associations, public authorities and research institutions. Its substantive work takes place in thematic ad hoc groups and in pilot projects in which SMART standards are trialed in specific application contexts; twice a year the IDiS plenary meets for cross-group exchange.

Anyone wishing to actively shape the scenarios described in this white paper, and in particular the ongoing and planned work on SMART content, ontologies and data spaces, will find IDiS the right place to do so. Alongside the ad hoc groups, IDiS also initiates and supports such pilot projects.

Anyone wishing to contribute their own use cases, concepts or ideas can approach IDiS with a project proposal.

GET INVOLVED NOW:

dke.de/idis/mitwirken

Current projects and the IDiS roadmap:

dke.de/idis/pilotprojekte

ABBREVIATIONS

Term	Meaning
AAS	Asset Administration Shell
AI Act	Artificial Intelligence Act
API	Application Programming Interface
CaaS	Content as a Service
CABCA	Continuous Auditing Based Conformity Assessment
CI/CD	Continuous Integration / Continuous Delivery
DCAT-AP	Data Catalog Vocabulary, Application Profile
DCC	Digital Calibration Certificate
D-CoC	Digital Certificate of Conformity
DevOps	Development Operations
DORA	Digital Operational Resilience Act
DPP	Digital Product Passport
EDC	Eclipse Dataspace Components
EDPS	Extended Dataset Profile Service
IDE	Integrated Development Environment
LLM	Large Language Model
MCP	Model Context Protocol
MDR	Medical Device Regulation
ML	Machine Learning
MLOps	Machine Learning Operations
OSD	Online Standards Development
OWL	Web Ontology Language

Term	Meaning
QI	Quality Infrastructure
RAG	Retrieval-Augmented Generation
RDF	Resource Description Framework
ReqIF	Requirements Interchange Format
RMS	Requirements Management System
SaaS	Software as a Service
SHACL	Shapes Constraint Language
VC	Verifiable Credentials



DIN e. V.

Burggrafenstraße 6
10787 Berlin, Germany
Phone: +49 30 2601-0
E-Mail: presse@din.de
Internet: www.din.de/en



VDE e. V.

DKE Deutsche Kommission Elektrotechnik Elektronik Informationstechnik

Merianstraße 28
63069 Offenbach am Main, Germany
Phone: +49 69 6308-0
Fax: +49 69 6308-9863
E-Mail: standardisierung@vde.com
Internet: www.dke.de/en

August 2026